Sequencing data formats and quality control

Vladimir Mikryukov

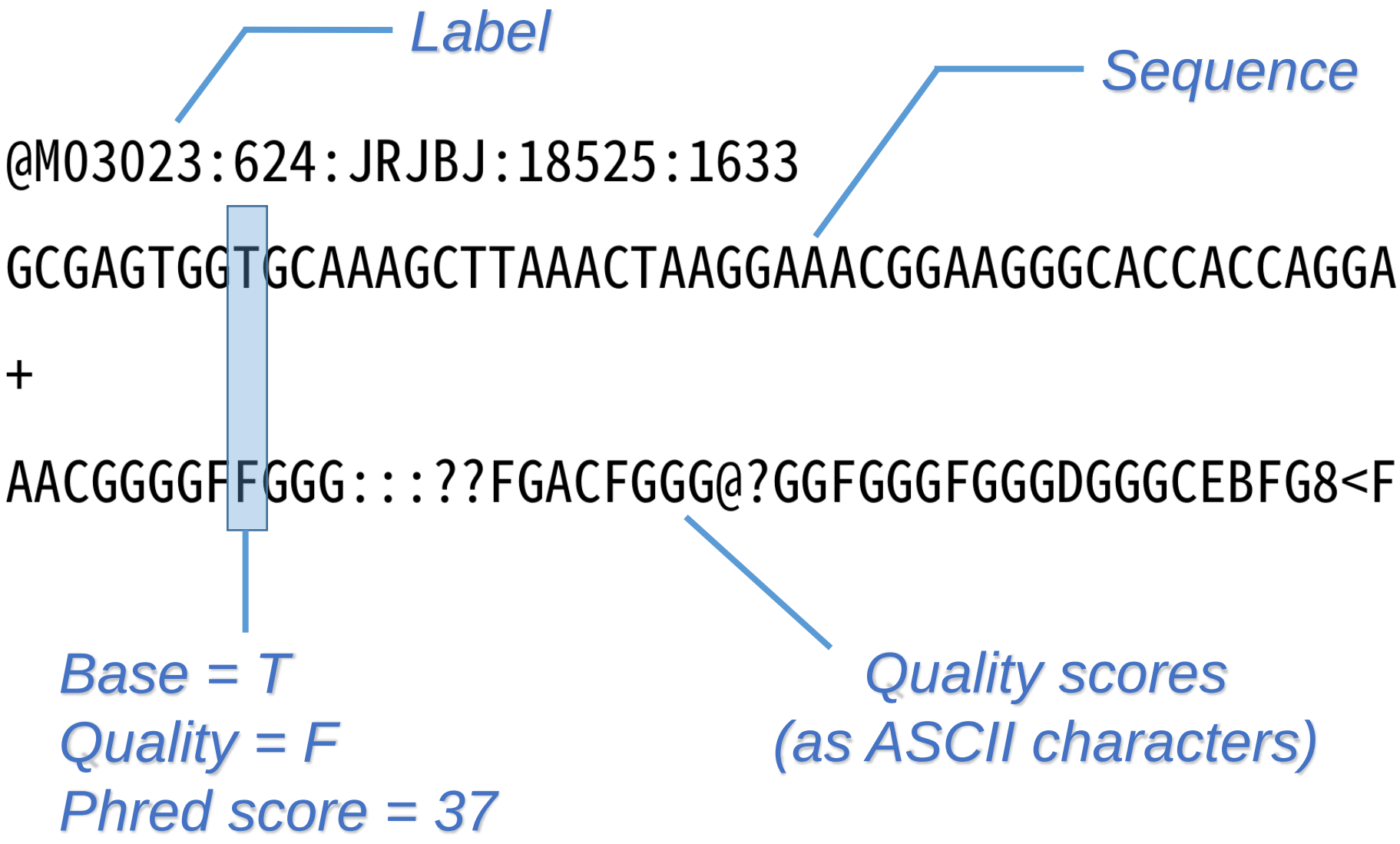
UT International Summer University,
Tartu, August 01-05 2022

FASTA


```
>SeqID1
GGGTGGACGGTTTATCCACCATTGGCAGCGGGTATTGCTCA
>SeqID2
GGGTGGACGGTTTATCCACCATTGGCAGCGGGTATTGCTCA
>SeqID3
GGGTGGACGGTTTATCCACCATTGGCAGCGGGTATTGCTCA
GTGAGTCATCGAATCTTTGAACGCACATTGCGCCCTCTGGT
```

```
>SeqID1;tax=d:Eukaryota,p:Arthropoda,c:Entognatha
GGGTGGACGGTTTATCCACCATTGGCAGCGGGTATTGCTCA
>SeqID2;tax=d:Eukaryota,p:Arthropoda,c:Insecta
GGGTGGACGGTTTATCCACCATTGGCAGCGGGTATTGCTCA
>SeqID3;tax=d:Eukaryota,p:Arthropoda,c:Entognatha
GGGTGGACGGTTTATCCACCATTGGCAGCGGGTATTGCTCA
```

FASTQ



FASTQ

```
@M02459:45:000000000-ATN9N:1:1101:19250:1081
GTGAATCATCGAATCTTTGAACGCACATTGCGCCCCTTGGTATTCCGAGGGGCATGCCTGTTCGAGCGNNANNAC
+
CCCCCGGGGGGCGFGGGG,CFFGGGGFGGGGCGGGGGGGGEGGGGFGGGGGGGGGGGGGGGGGGEGGG##:##:9
@M02459:45:000000000-ATN9N:1:1101:21592:1097
GTGAGTCATCGAATCTTTGAACGCACATTGCGCCTCCTGGTATTCCGGGAGGCATGCCTGTTCGAGCGTCANTAA
+
CCCCCGGGGGGCGGGGGG9CFFGGGGGGGGGGGG7,,CFGGGGGGG8+6+@CFGGGGGGGGGGGCFGGFGG#:CC
@M02459:45:000000000-ATN9N:1:1101:10733:1108
GTGAGTCATCGAATCTTTGAACGCACATTGCGCCCTCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTCANTAT
+
CCCCCGGGGGGGGGGGGG9CEGGGGGGGGDFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFG,FEGGGC#:BF
@M02459:45:000000000-ATN9N:1:1101:20537:1091
GTGAACCATCGAATCTTTGAACGCATCTTGCGCTCCTTGGTATTCCGAGGAGCATGCCTGTTTGAGTGNTCNNTA
+
CCCCCFGGGGGGGGGGGG9CEFGGGCCFFGG+F@F@EFGG;6;E,E@@@77C,C,,6CCACF,@,:5,#,,##::
@M02459:45:000000000-ATN9N:1:1101:9029:1114
GTGAATCATCGAATCTTTGAACGCACATTGCGCCCGGTGGTATTCCGCCGGGCATGCCTGTTCGAGCGTCATTAG
+
CCCCCGGGGGGGGGGFFG,CFF@FFGCFGGGGGGGGGCFGGFGGGFGGGCEGGGGFFFGGGGGGDFGGFEEGGGF
@M02459:45:000000000-ATN9N:1:1101:11753:1122
GTGAATCATCGAATCTTTGAACGCACCTTGCGCCCTTTGGCTATTCCGAAGGGCATGCCTGTTTGAGTATCATGA
+
CCCCCGGGGGGGGGGGGG,CEFGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGG
@M02459:45:000000000-ATN9N:1:1101:21505:1122
GTGAATCATCGAATCTTTGAACGCACCTTGCGCCCTCTGGTATTCCGGAGGGCATGCCTGTTTGAGTGTCATGTA
+
CCCCCFGGGGGGGGGGGF9CFFGGGGGGGGGGG@FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFFGGGGG
```

FASTQ (paired-end data)

*_R1.fastq

```
@A00459:120:HHFYJDRXY:1:2101:1515:1016 1:N:0:AGTCAGGT+GTGGTTAC
TNTACGGCGTTCTTCATCGATGGGAGAACCAAGAGATCCGTTGTCAAAAGTTGTATTTTAGTTCTGCCCG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:1642:1016 1:N:0:AGTCAGGT+GTGGTTAC
ANGCGAACTAATCGACCAGCGGAGGGATCATTAATGAATAAACTCGGTGGATTGTTGCTGGCTCTTAGGA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFF
@A00459:120:HHFYJDRXY:1:2101:1325:1031 1:N:0:AGTCAGGT+GTGGTTAC
TNTGCTGCGTTCTTCATCGATGGGAGAGCCAAGAGATCCGTTGCTGAAAGTTGTATAGGTTTAATGACTG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF::
@A00459:120:HHFYJDRXY:1:2101:1362:1031 1:N:0:AGTCAGGT+GTGGTTAC
ANCAGCTCAGATCGACCAGCGGAGGGATCATTATTGAATAAACCTGGCGTGGTTGTAGCTGGTCCTCTCG
+
F#FFFFFFFFFFFFFFFFFFF:FFFFFF,FFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:2374:1031 1:N:0:GGTCAGGT+GTGGTTAC
ANGCATCGCACTCGACCAGCGGAGGGATCATTAATGAATAAACTTGGTAGATTGTTGCTGGCTTCTAGGA
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

*_R2.fastq

```
@A00459:120:HHFYJDRXY:1:2101:1515:1016 2:N:0:AGTCAGGT+GTGGTTAC
ACACAGTCCTGACGACCAGCGGAGGGATCATTAGTGAATACACCGGGGAATCCACTCTGTGGGCCCCAACC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:1642:1016 2:N:0:AGTCAGGT+GTGGTTAC
TTCGCTACGTTCTTCATCGATGCGAGAGCCAAGAGATCCGTTGTTGAAAGTTGTATAAGTTTTAAAGCCTA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:1325:1031 2:N:0:AGTCAGGT+GTGGTTAC
ACGTAACCACGTCGACCAGCGGAGGGATCATTATTGAATACGAATCGGTTCTGATGCTGGCCCTTCACCGG
+
:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFF:FFFFFF:FFFFFFF
@A00459:120:HHFYJDRXY:1:2101:1362:1031 2:N:0:AGTCAGGT+GTGGTTAC
TTCGCTACGTTCTTCATCGATGCGAGAGCCAAGAGATCCGTTGCTGAAAGTTGTATATAGTTTATAGGCAC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:2374:1031 2:N:0:GGTCAGGT+GTGGTTAC
TTTGCGGCGTTCTTCATCGATGCGAGAGCCAAGAGATCCGTTGTTGAAAGTTGTATTAAGTTTTATAAGGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFF::F
```

Interleaved

```
@A00459:120:HHFYJDRXY:1:2101:1515:1016 1:N:0:AGTCAGGT+GTGGTTAC
TNTACGGCGTTCTTCATCGATGGGAGAACCAAGAGATCCGTTGTCAAAAGTTGTATTTTAGTTCTGCCCGA
+
F!FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:1515:1016 2:N:0:AGTCAGGT+GTGGTTAC
ACACAGTCCTGACGACCAGCGGAGGGATCATTAGTGAATACACCGGGGAATCCACTCTGTGGGCCCCAACC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00459:120:HHFYJDRXY:1:2101:1642:1016 1:N:0:AGTCAGGT+GTGGTTAC
ANGCGAACTAATCGACCAGCGGAGGGATCATTAATGAATAAACTCGGTGGATTGTTGCTGGCTCTTAGGAG
```

File extensions

***.fasta, *.fa**
***.fastq, *.fq**

*.fasta**.gz**, *.fa**.gz**
*.fastq**.gz**, *.fq**.gz**

File extensions

*.fastq.**gz**

File extensions

***.fasta, *.fa**
***.fastq, *.fq**

} Uncompressed data

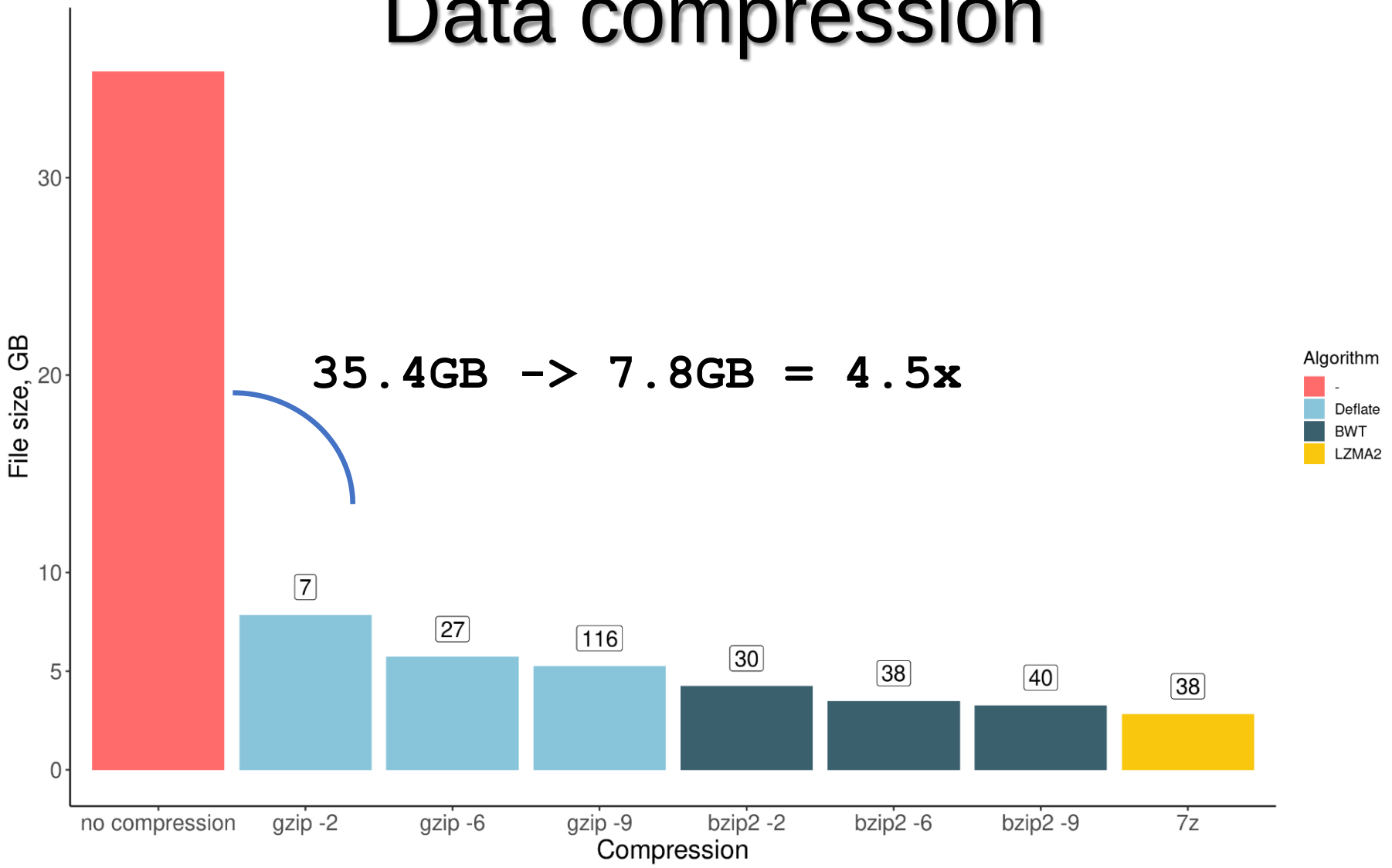
*.fasta.**gz**, *.fa.**gz**
*.fastq.**gz**, *.fq.**gz**

} gzip-compressed

*.fasta.**bz2**, *.fa.**bz2**
*.fastq.**bz2**, *.fq.**bz2**

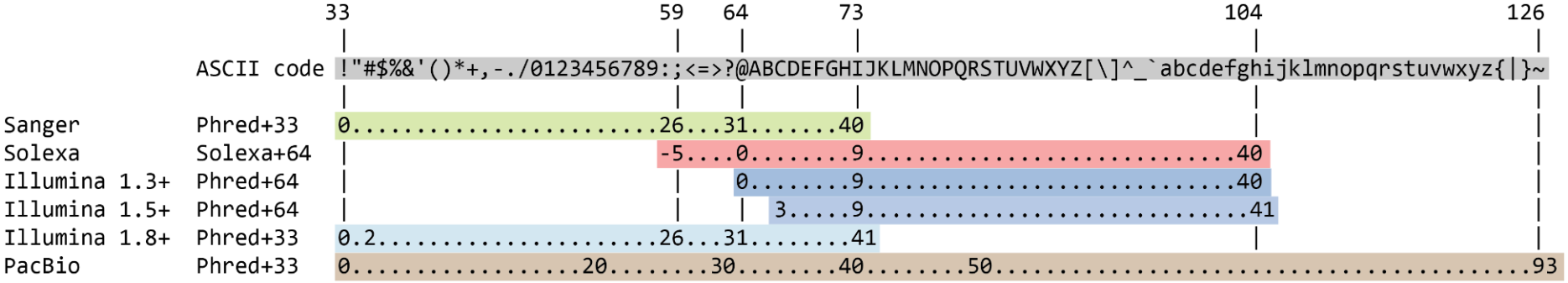
} bzip2-compressed

Data compression



Input = NovaSeq data, 67.9M sequences

Quality scores



Phred quality scores

$$Q = -10 * \log_{10} P \qquad P = 10^{-Q/10}$$

ASCII	Q	P	Chance that the base is wrong (Error)	Accuracy (1 - Error)
$	3	0.50	50%	50%
+	10	0.10	10%	90%
.	13	0.05	5%	95%
5	20	0.01	1%	99%
:	25	0.00316	0.316%	99.7%
?	30	0.00100	0.1%	99.9%
I	40	0.00010	0.01%	99.99%
S	50	0.00001	0.001%	99.999%
]	60	0.000001	0.0001%	99.9999%
~	93	0.000000001	0.0000001%	99.9999999%

} PacBio HiFi Reads

FASTQE🤔

Bin	Emoji
N	🚫
2-9	💀
10–19	💩
20–24	⚠️
25–29	😄
30–34	😆
35–39	😎
≥ 40	😍

```
@Good_Sequence
N T G G C C C C G G G T C G A C G T G G C
+
J J J H H G G G F F F E E C B = ; ; : : 9
😍😍😍😎😎😎😎😎😎😎😎😎😎😆😆😄😄😄😄😄⚠️
```

```
@Bad_Sequence
N T G G C C C C G G G T C G A C G T G G C
+
! 9 9 9 3 3 2 2 2 1 1 . . - - # $ % & % (
🚫⚠️⚠️⚠️💩💩💩💩💩💩💩💩💩💩💩💀💀💀💀💀💀
```

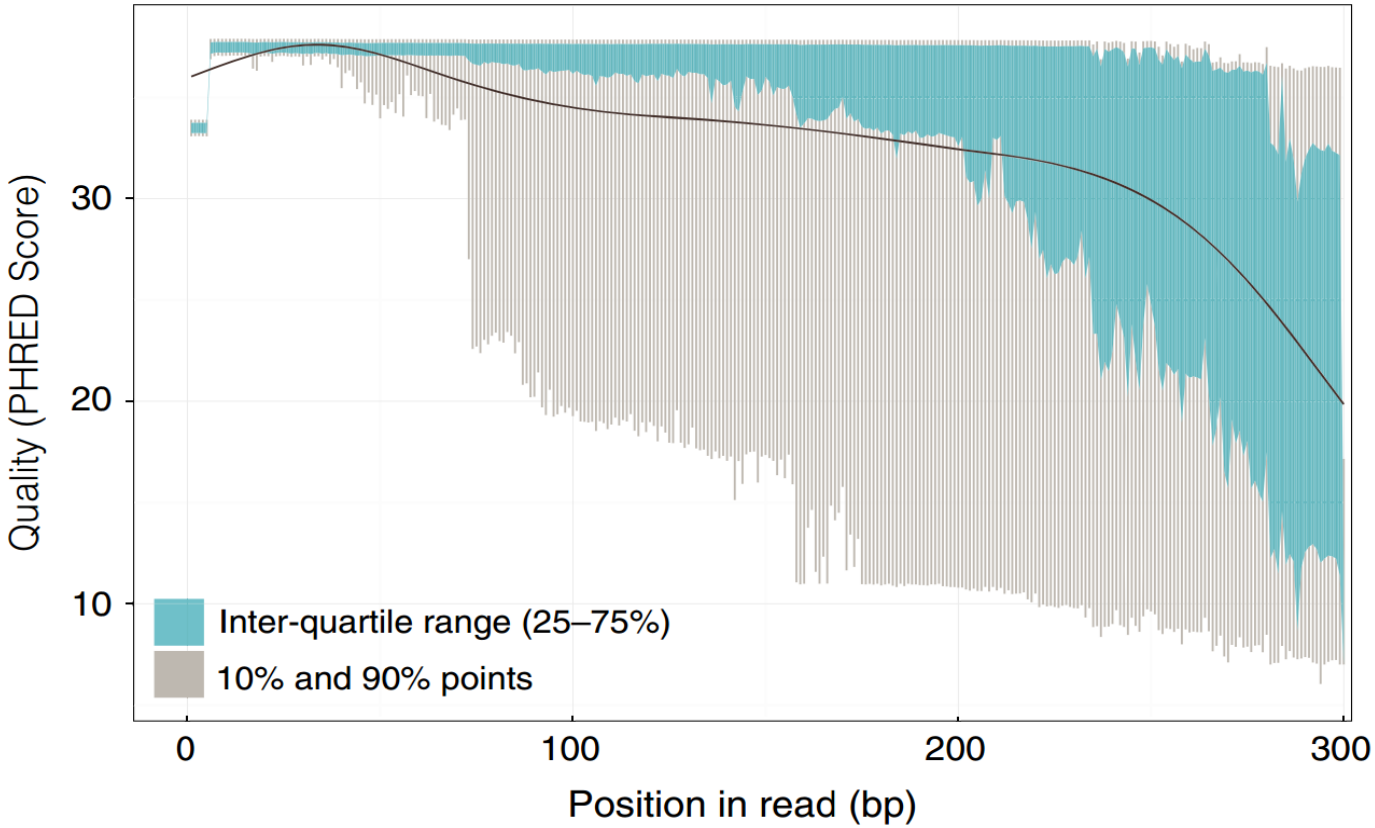
https://github.com/fastqe/fastqe

Quality Control

99.9%

99%

90%



Gweon et al. (2015) DOI:10.1111/2041-210X.12399

Binned quality

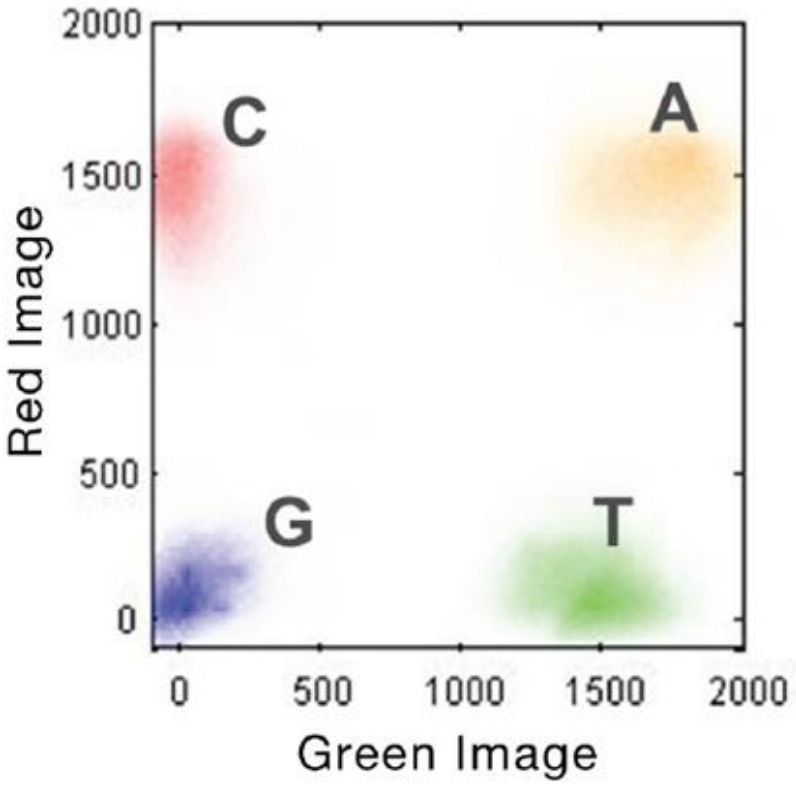
- Illumina 2-Channel SBS Technology (NovaSeq / NextSeq / MiniSeq / iSeq)

2, **11, 25**, and 37

2, **12, 23**, and 37

Poly-G tails

```
@A00551:123:H2MKG
...TCGCAGGTAGCGCGCCGCCGGGGGGGGGG
+
...:FF::F:FF,::F:,:,:,FFFFFFFFFF
```



Illumina 2-channel technology

Imortance of quality control

Garbage In, Garbage Out

Your analysis is as good as your data!

Image source: https://serendipics.blogspot.com

Quality filtering

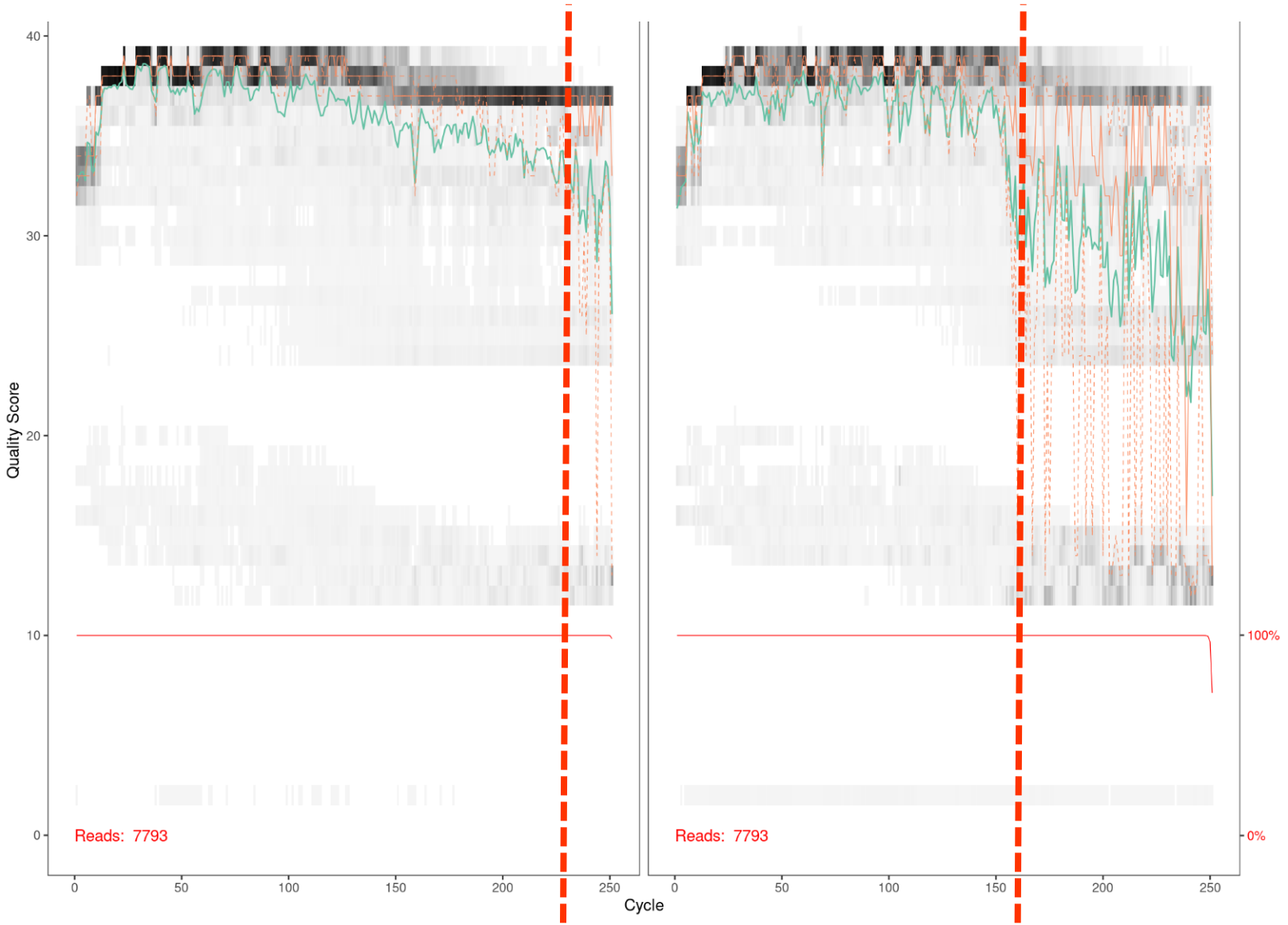
- Number of ambiguous (N) nucleotides
- Average quality score
- Maximum number of expected errors (MaxEE)
- Maximum expected errors as a percentage of read length (MEEP)
- Number / percentage of unqualified bases

$$(10 + 20) / 2 \neq 15$$
$$(10 + 20) / 2 = 12.6$$

$$\mathrm{EE} = \sum_{i} \mathrm{P}_{i}$$

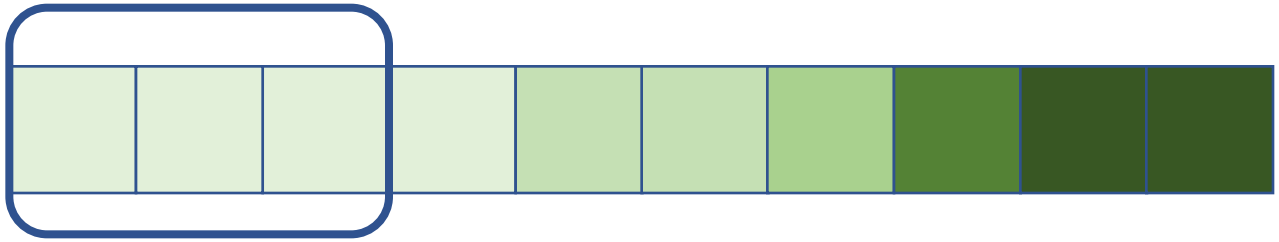
Edgar & Flyvbjerg (2015)
DOI:10.1093/bioinformatics/btv401

Trimming

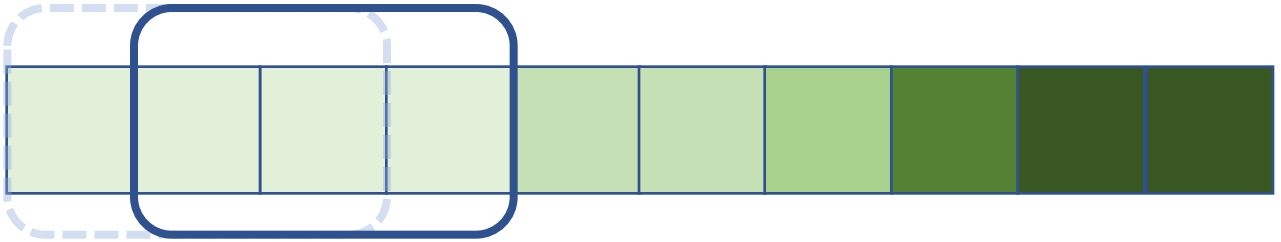


Sliding window trimming

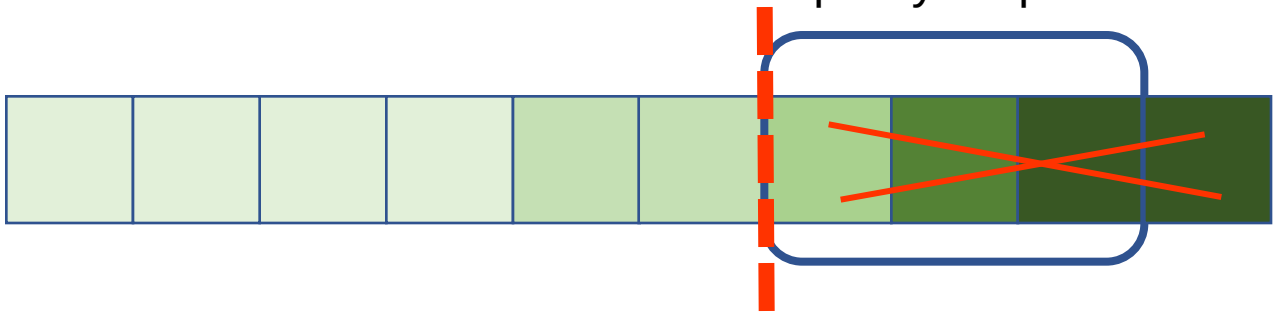
Window (size = 3)



Slide one bp forward, esimate average quality in the window

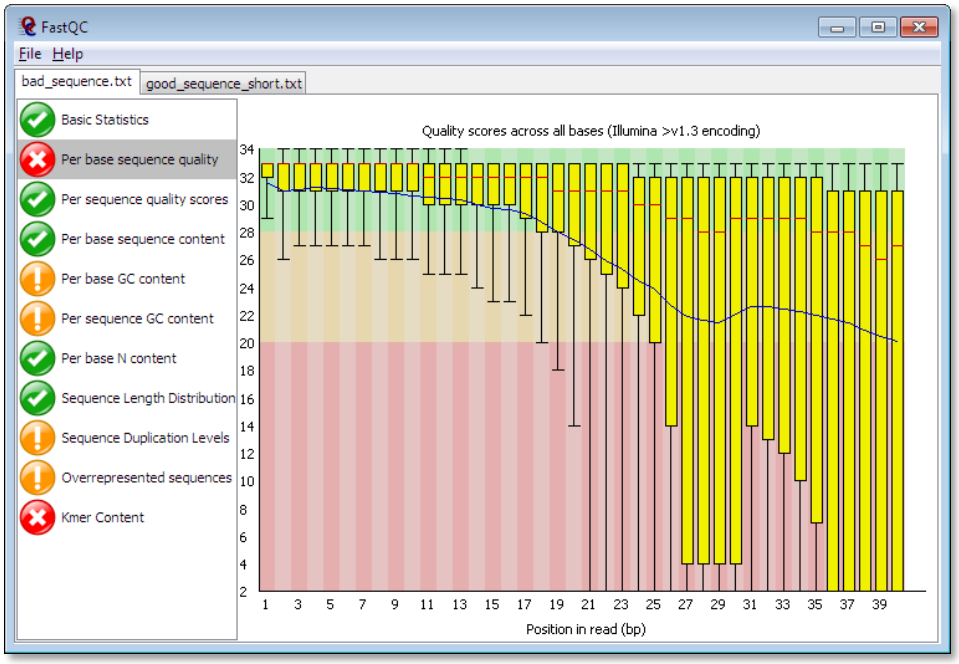


Cut the leftmost position in the window
where the quality drops below the threshold



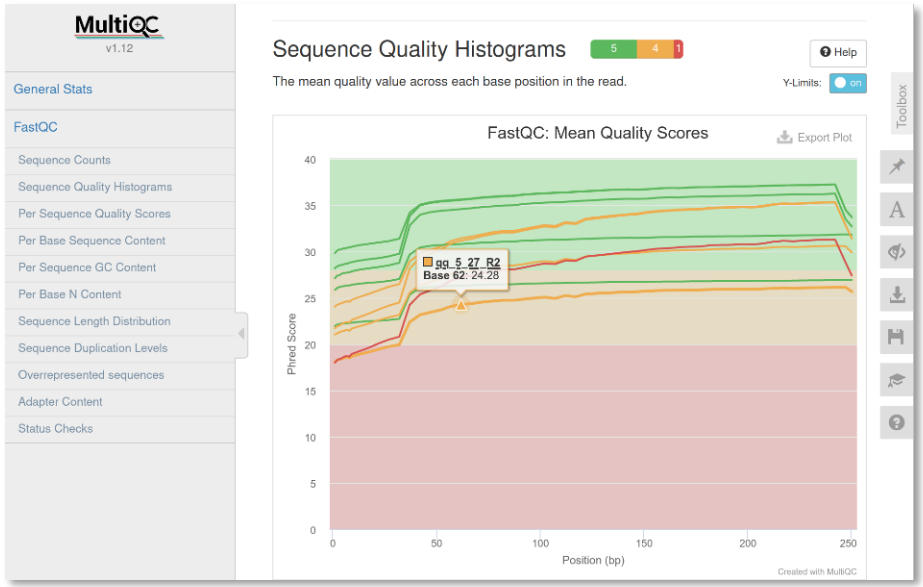
Quality check

FastQC



Andrews et al. (2010)

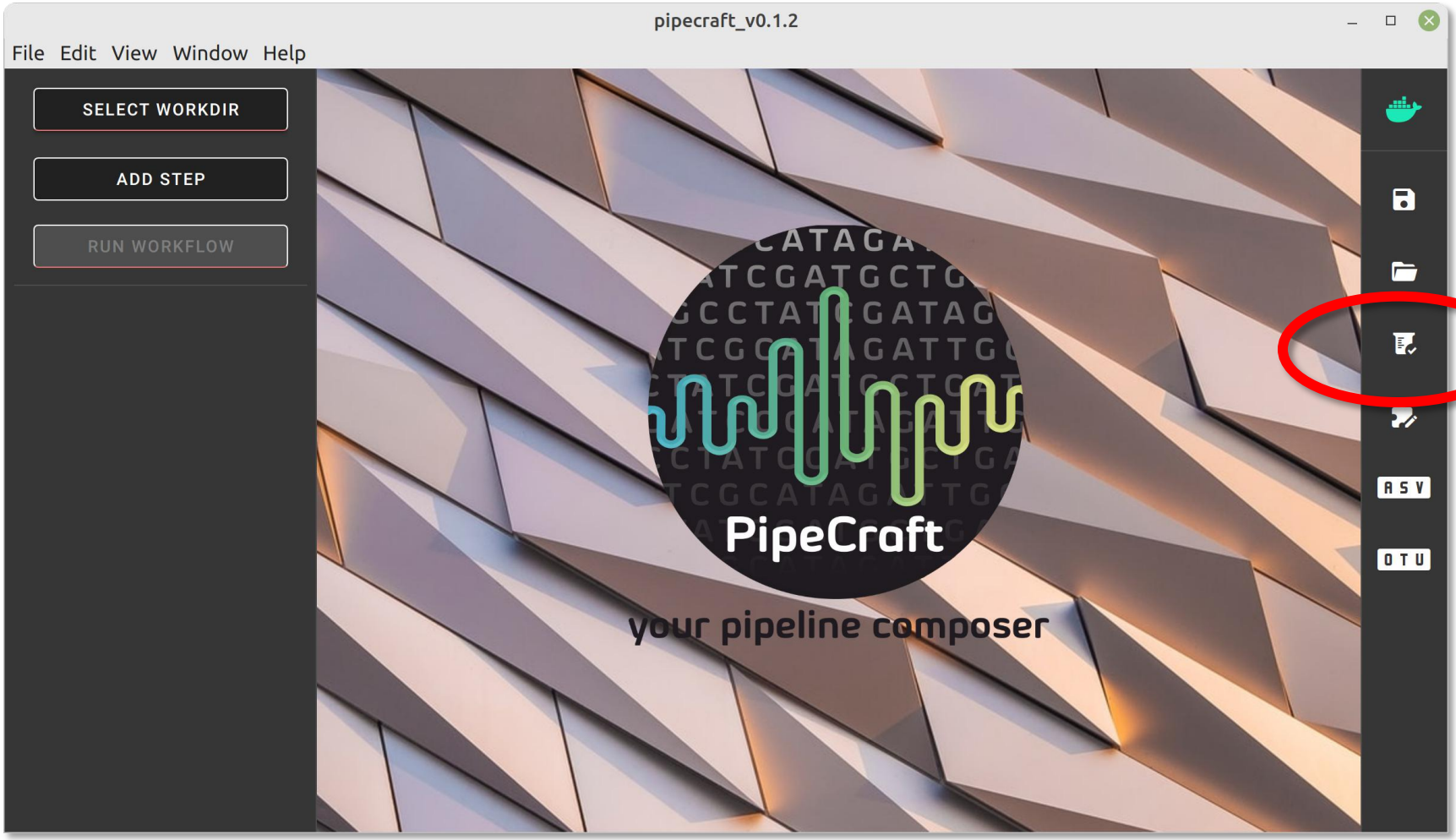
MultiQC



Ewels et al. (2016)
DOI:10.1093/bioinformatics/btw354

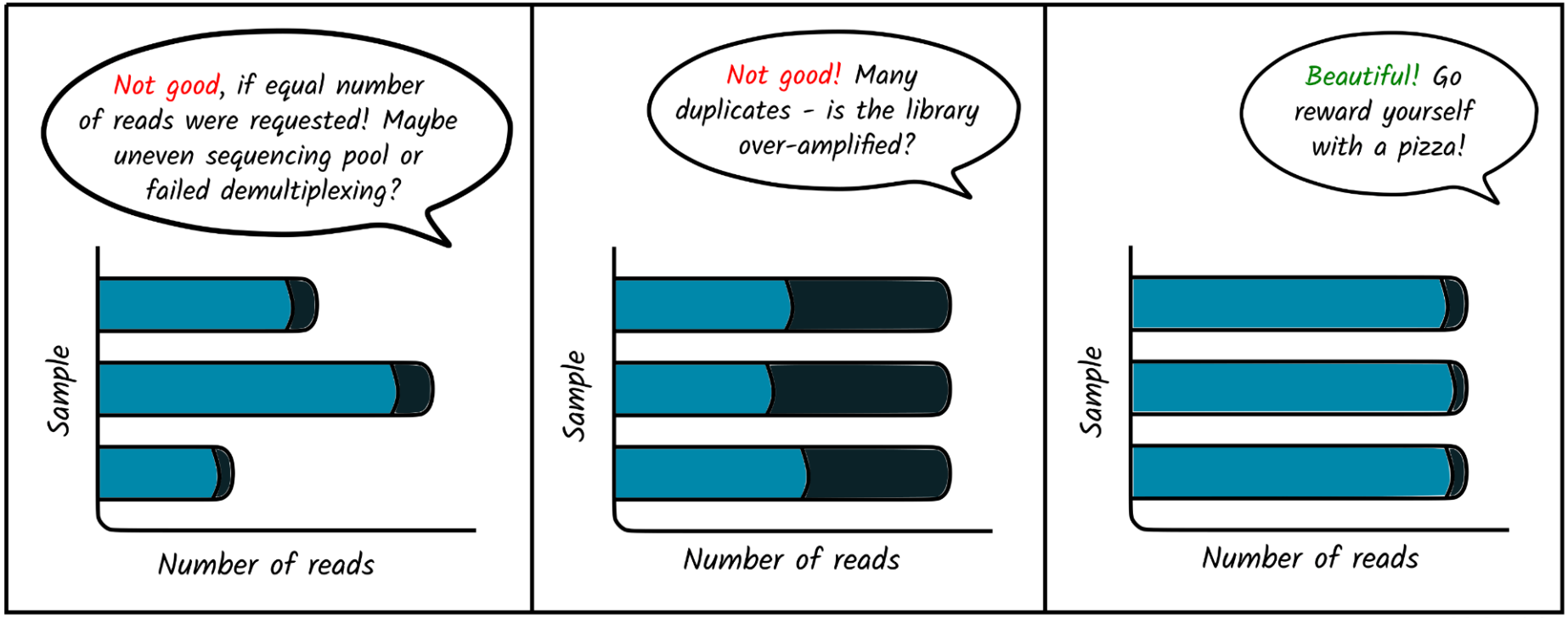
https://multiqc.info/
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Quality check in PipeCraft2



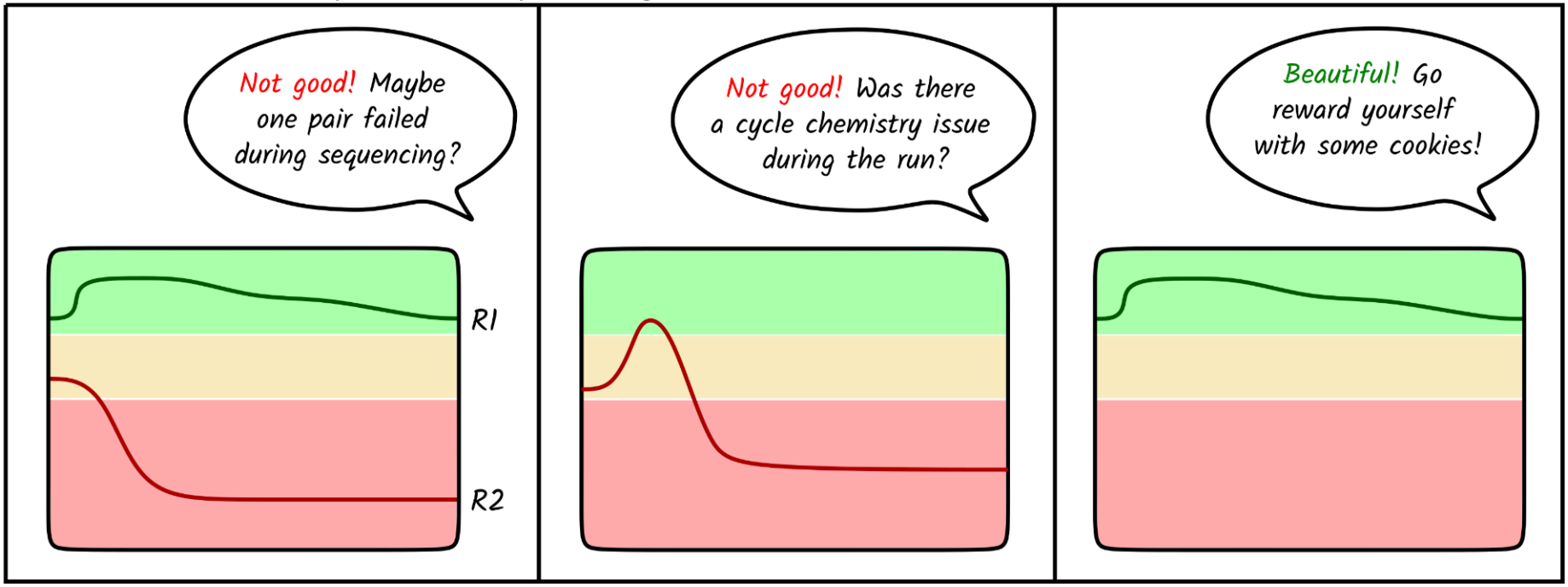
FASTQC - Sequence count

ZandraSelina



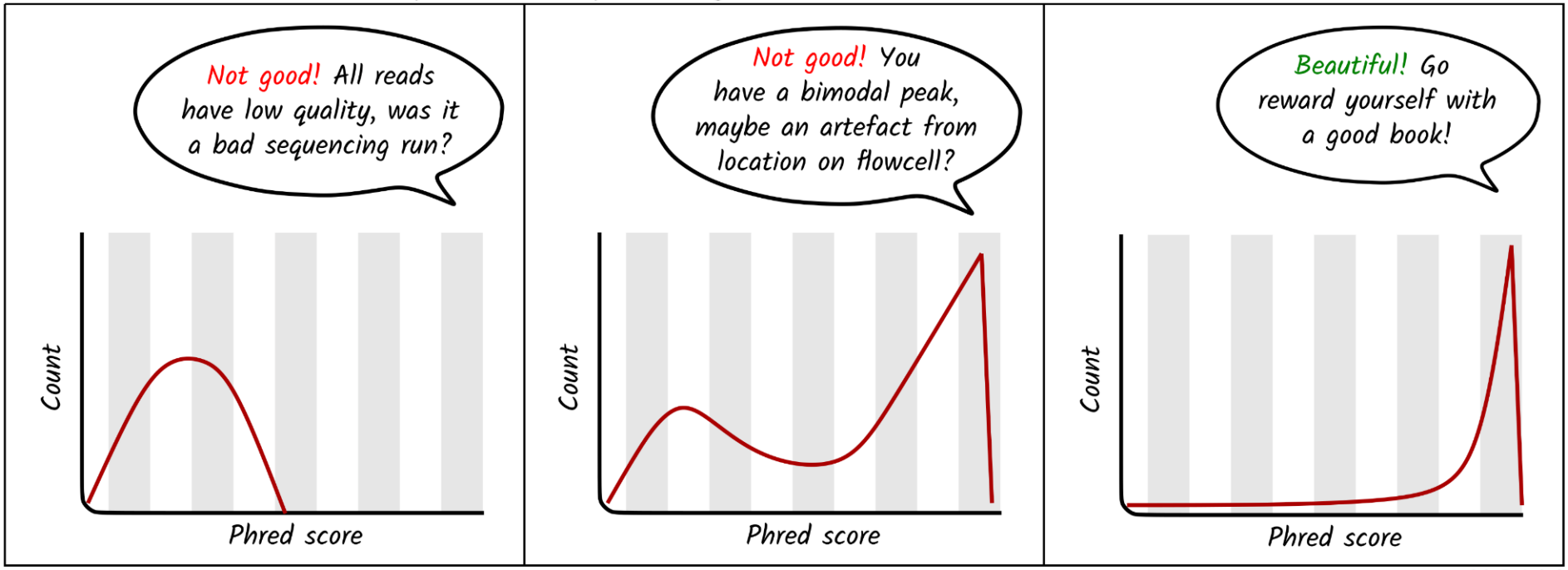
FASTQC - Sequence quality

ZandraSelina



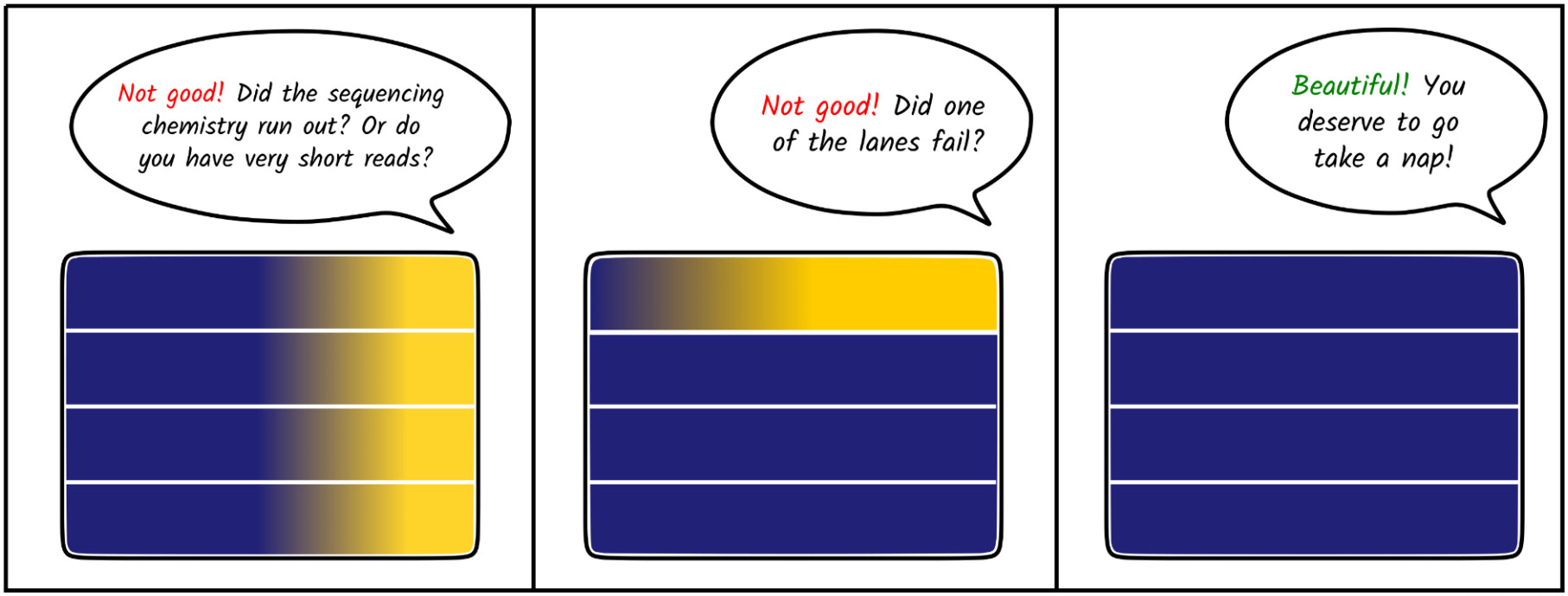
FASTQC - Per sequence quality score

ZandraSelina



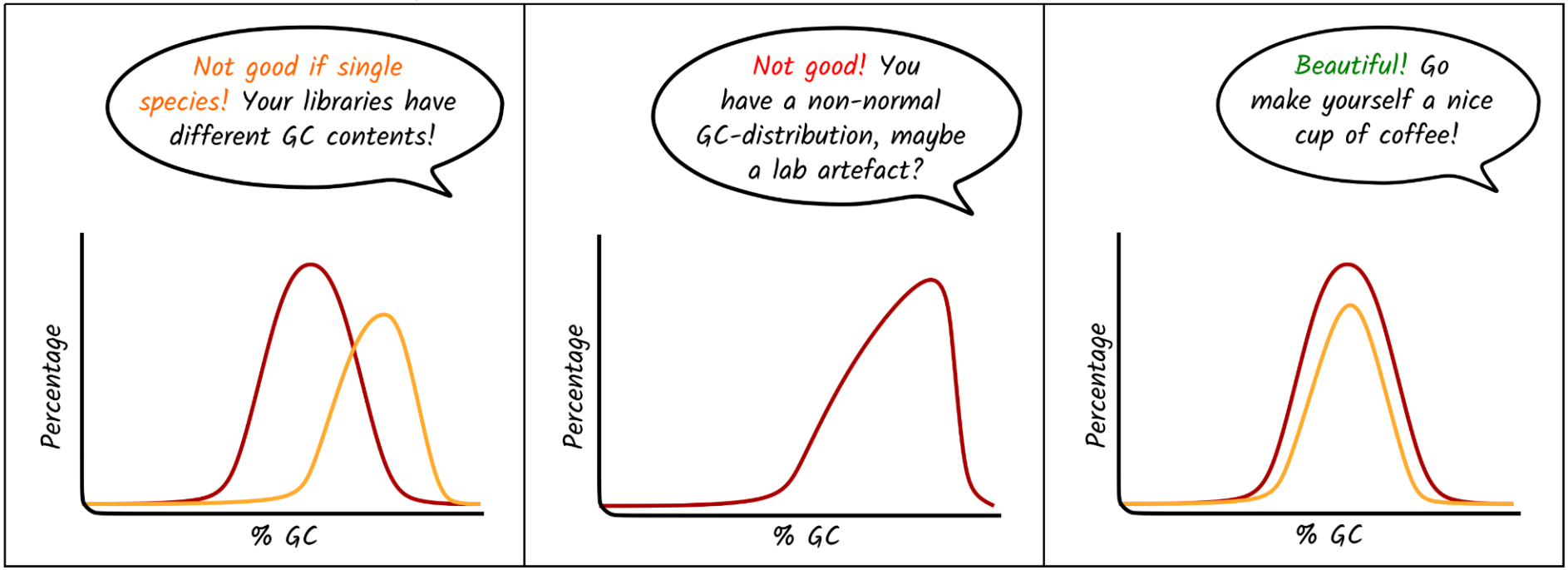
FASTQC - Per base sequence content

ZandraSelina



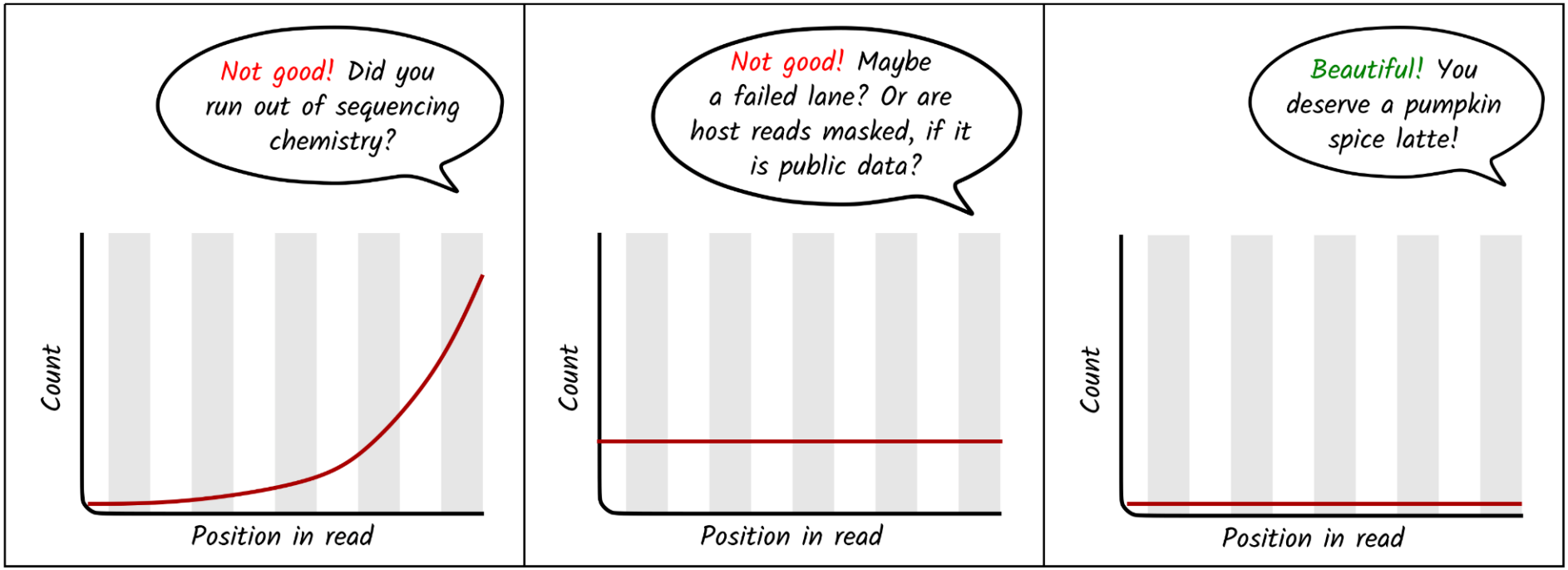
FASTQC - Per sequence GC-content

ZandraSelina



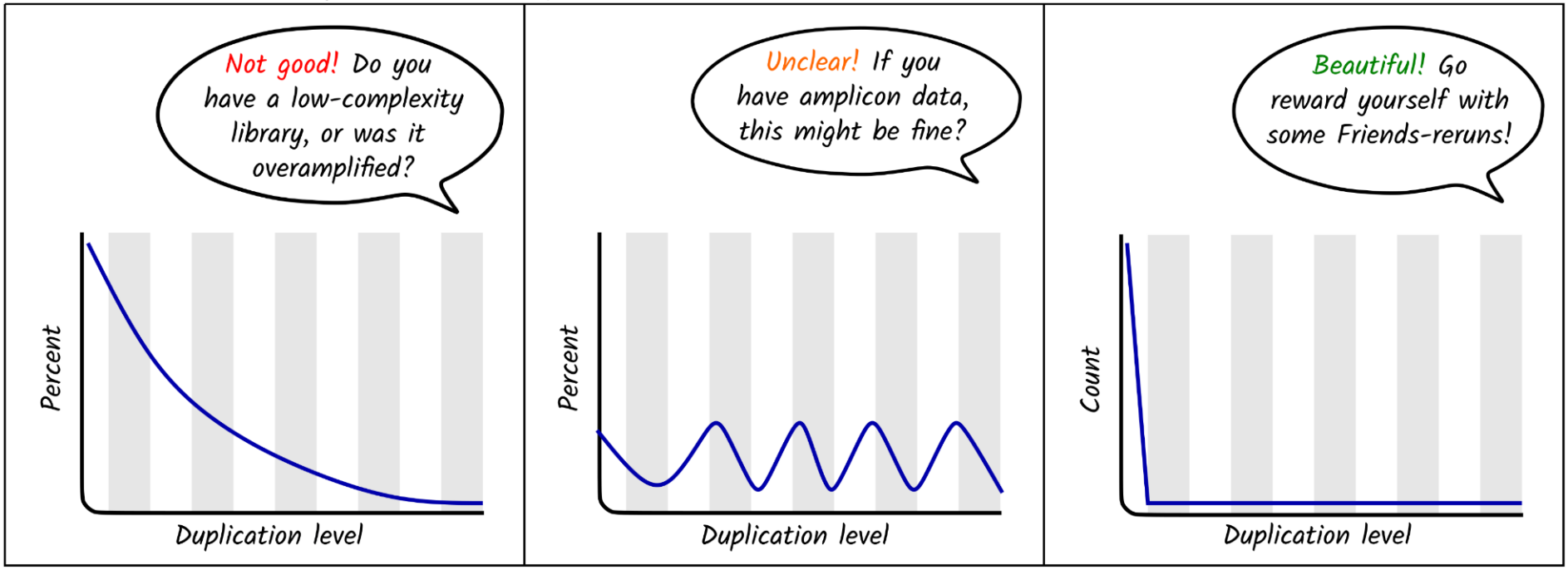
FASTQC - Per base N content

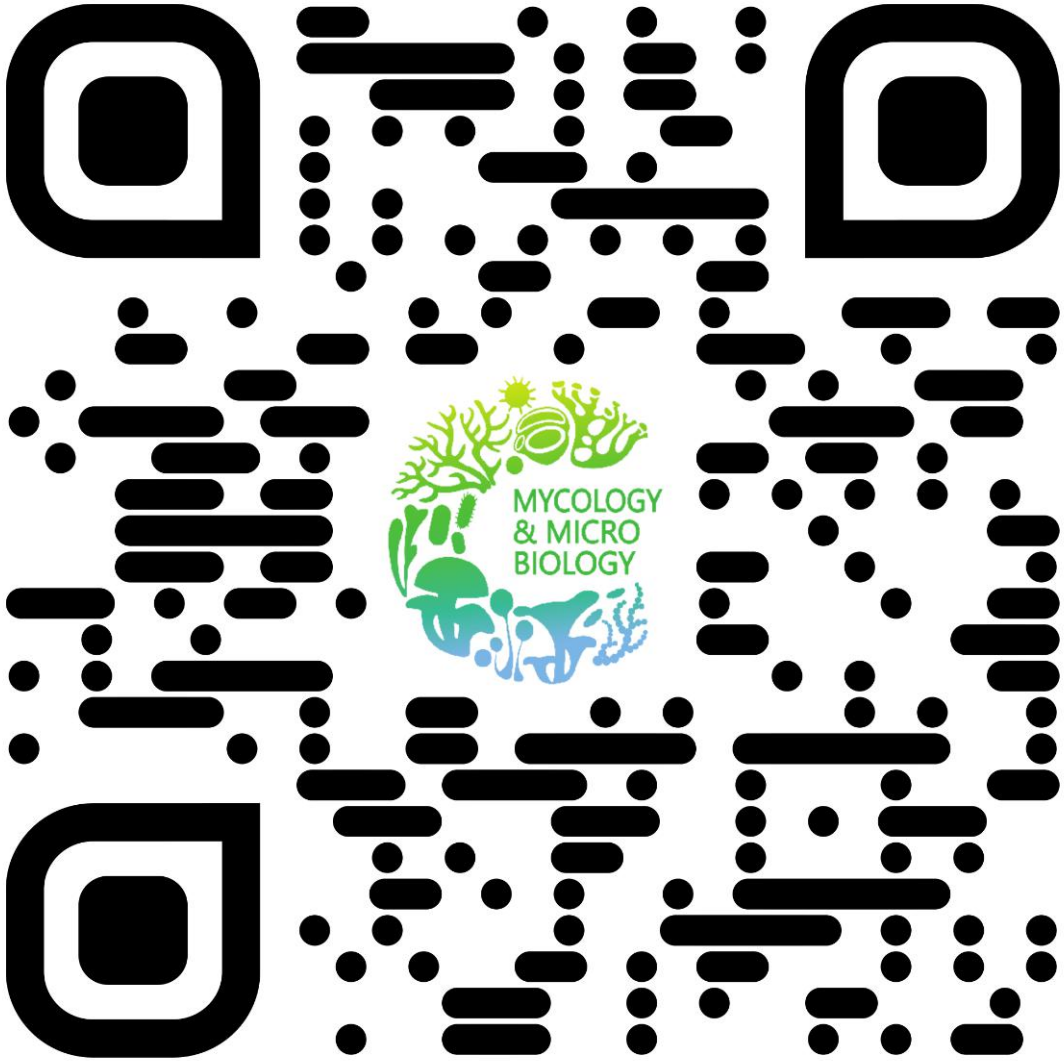
ZandraSelina



FASTQC - Sequence duplication levels

ZandraSelina





https://mycology-microbiology-center.github.io/Metabarcoding2022/

https://t.ly/dvZS

main **Metabarcoding2022** / **data** /

vmikk qc

..

DB

MiSeq_Fungi

QC

Set01

Set02

Set03

README.md

main **Metabarcoding2022** / data / **QC** /

vmikk add data for QC

..

MiSeq_R1.fq.gz

MiSeq_R2.fq.gz

NovaSeq_R1.fq.gz

NovaSeq_R2.fq.gz

PacBio.fq.gz

https://github.com/Mycology-Microbiology-Center/Metabarcoding2022/tree/main/data