NGS file formats and quality control

Dr. Sreenu Vattipally

MRC-University of Glasgow Centre for Virus Research

Sequence file formats: FastA

- Ubiquitous sequence format
- Originally developed for Fast Align program
- Can store protein or nucleotide sequences
- Each entry has two sections
- Header: Single line. Always starts with ">"
- Sequence: Follows header. Single or multiple lines

Ex:

```
>gi1097381|NC_19817181|Sequence name1
ATGCCGATTGCGATCGCGGGCGCGATCGATCGCGATCGTCGCATCG
ATTTTTAGGGATGCACATCATCGATCGGGATCGATTCAGACTCGAT
>gi1294382|NC_12814182|Sequence name2
TTTAGCGATCGATGGATCGATCGATCGATCGATCGCAGCGATCGAT
ATTGCGACGATCGGGATCGATCGCGCATCGATCGATCGCGATCGAT
```

Sequence file formats: FastQ

- Universal next generation sequence format
- Typically contains millions of reads
- Each read has four rows
- 1^{st}: Name: starts with "@"
- 2^{nd}: Sequence
- 3^{rd}: Optional comment field, starts with "+"
- 4^{th}: Quality scores

The length of 2^{nd} and 4^{th} lines is ALWAYS equal

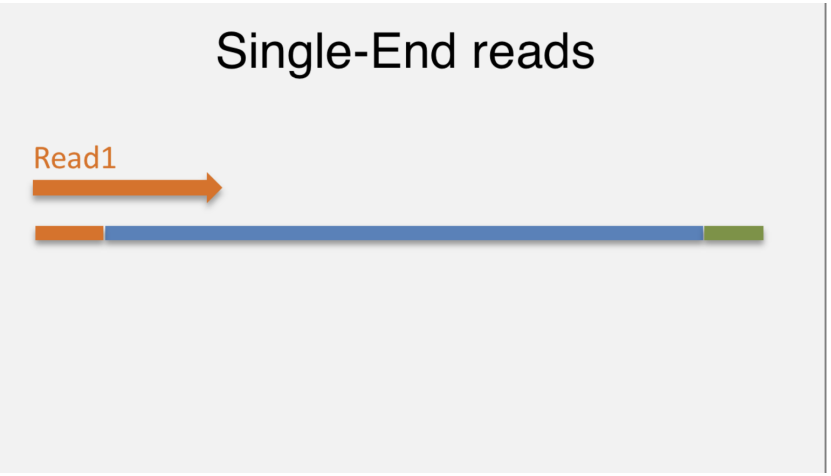
Sequence file formats: FastQ

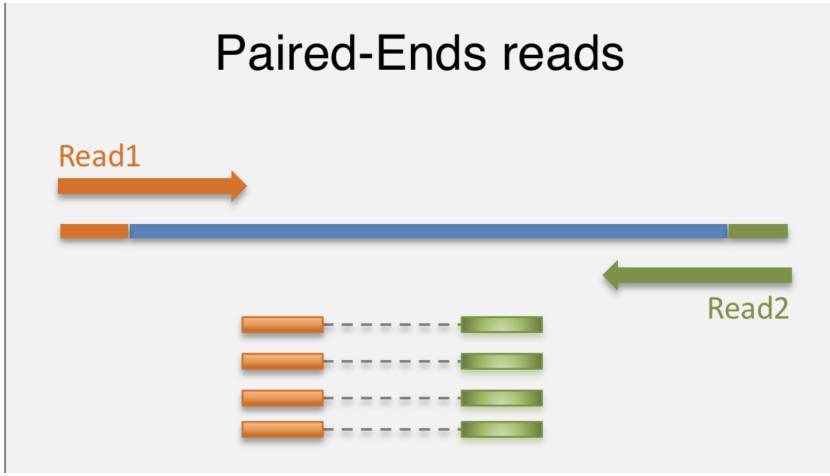
```
@Sequence name/id
TGCAGCGATCGAATGCGATTCGATCGATCGAT
+
BBFFIIBBIIIGGADDAIIABBAIIBBFIIII
```

FastQ: Paired-end

- Sequenced from both ends
- Stored in two FastQ files
- Can be stored in interleaved FastQ

Sequence file formats: FastQ





The order in the fastq files corresponds to pairing of the reads.

Sequence Quality Scores

Phred Quality Scores

- 1st used in Phred Program
- Each nucleotide assigned a base calling score
- It is the probability of an error
- Q = -10 $\log_{10}$ P
- P = 10 $^{-Q/10}$
- In FastQ file it is converted to Q + 33 ASCII value

Sequence Quality Scores

Phred	Sym	Error
0	!	1.00000
1	"	0.79433
2	#	0.63096
3	$	0.50119
4	%	0.39811
5	&	0.31623
6	'	0.25119
7	(	0.19953
8	)	0.15849
9	*	0.12589
10	+	0.10000
11	,	0.07943
12	-	0.06310
13	.	0.05012
14	/	0.03981
15	0	0.03162
16	1	0.02512
17	2	0.01995
18	3	0.01585
19	4	0.01259
20	5	0.01000

Phred	Sym	Error
21	6	0.00794
22	7	0.00631
23	8	0.00501
24	9	0.00398
25	:	0.00316
26	;	0.00251
27	<	0.00200
28	=	0.00158
29	>	0.00126
30	?	0.00100
31	@	0.00079
32	A	0.00063
33	B	0.00050
34	C	0.00040
35	D	0.00032
36	E	0.00025
37	F	0.00020
38	G	0.00016
39	H	0.00013
40	I	0.00010

Sequence Quality Scores

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                            104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Phred score

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
```

Q	Error	Accuracy
0	1 in 1	0%
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

Sequence file formats: SAM

Sequence Alignment/Map (SAM) format

- Tab-delimited generic alignment format
- Supports short and long reads
- Independent of sequencing platforms
- Flexible style
- Compact size
- Has two sections: Header and Alignments

Compressed SAM files are called BAM files (binary SAM)

Sequence file formats: SAM

SAM format

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

The corresponding SAM format is:[1]

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Sequence file formats: SAM

Sections: Header

- Starts with @
- @ Follows a two letter TAG name
- Header fields are pairs of TYPE:VALUE
- In the header, each line is tab-delimited

Ex:

```
@HD VN:1.5 SO:coordinate
@SQ SN:Ebola Genome LN:18400
```

ToDo

How do we get header information from a sam file?

Sequence file formats: SAM

Sections: Alignment

- Should start with non"@" character
- Should have 11 mandatory fields
- Fields are TAB-Delimited
- Can have optional fields

ToDo

How do we get alignment information from a sam file?

Sequence file formats: SAM

Alignment Section: Mandatory fields

1. QNAME: Query name
2. FLAG: Bitwise flag
3. RNAME: Reference name
4. POS: 1-based mapping position
5. MAPQ: Mapping quality
6. CIGAR: Extended CIGAR string
7. RNEXT: Reference name of next read
8. PNEXT: Position of the next reads
9. TLEN: Template length
10. SEQ: Read sequence
11. QUAL: Read quality

Sequence file formats: SAM

SAM FLAG explained

1. 2^0: 1: Template having multiple reads (paired-end)
2. 2^1: 2: Each read properly aligned according to the aligner
3. 2^2: 4: Read unmapped
4. 2^3: 8: Next read is unmapped
5. 2^4: 16: Read being reverse complemented
6. 2^5: 32: Next read is reverse complemented
7. 2^6: 64: First read in the template
8. 2^7: 128: Last read in the template
9. 2^8: 256: Secondary alignment
10. 2^9: 512: Not passing quality controls
11. 2^{10}:1024: PCR or optical duplicate
12. 2^{11}: 2048: Supplementary alignment

Sequence file formats: SAM

CIGAR values

- M: match/mismatch
- I: Insertion
- D: deletion
- S: Softclip
- H: Hardclip
- P: Padding
- N: Skip

Ex:
CIGAR: 3S5M1D7M1I4M2S
AGCATATGGATTTGCG-ATGCTC
GATATATG-ATTTGCGGATGCAA

HTS: Quality check using FastQC

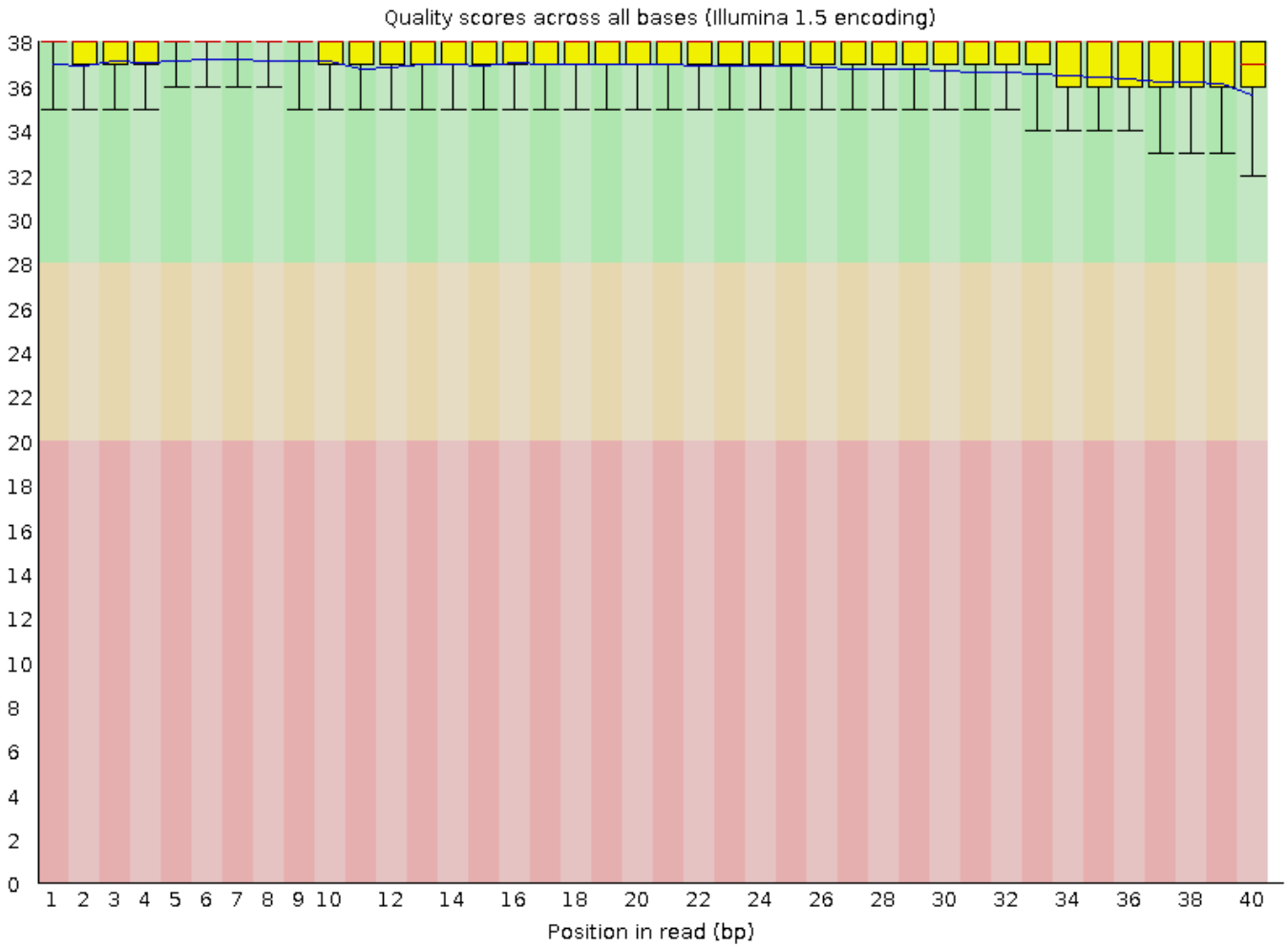
- Developed by Babraham Bioinformatics
- Works with FastQ, SAM and BAM files
- Runs interactively or command line
- www.bioinformatics.babraham.ac.uk/projects/fastqc/

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

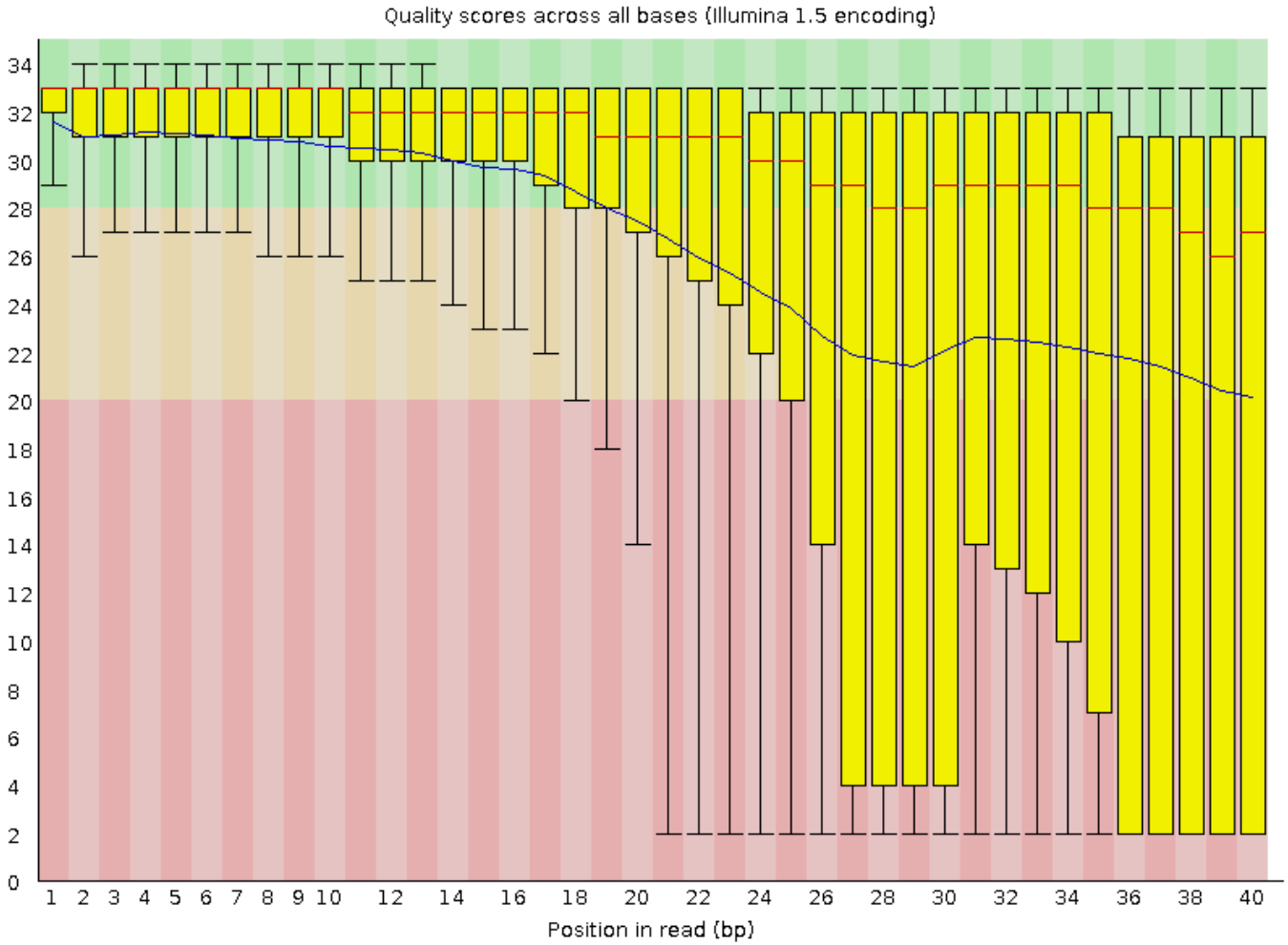
HTS: Quality check using FastQC

Good data



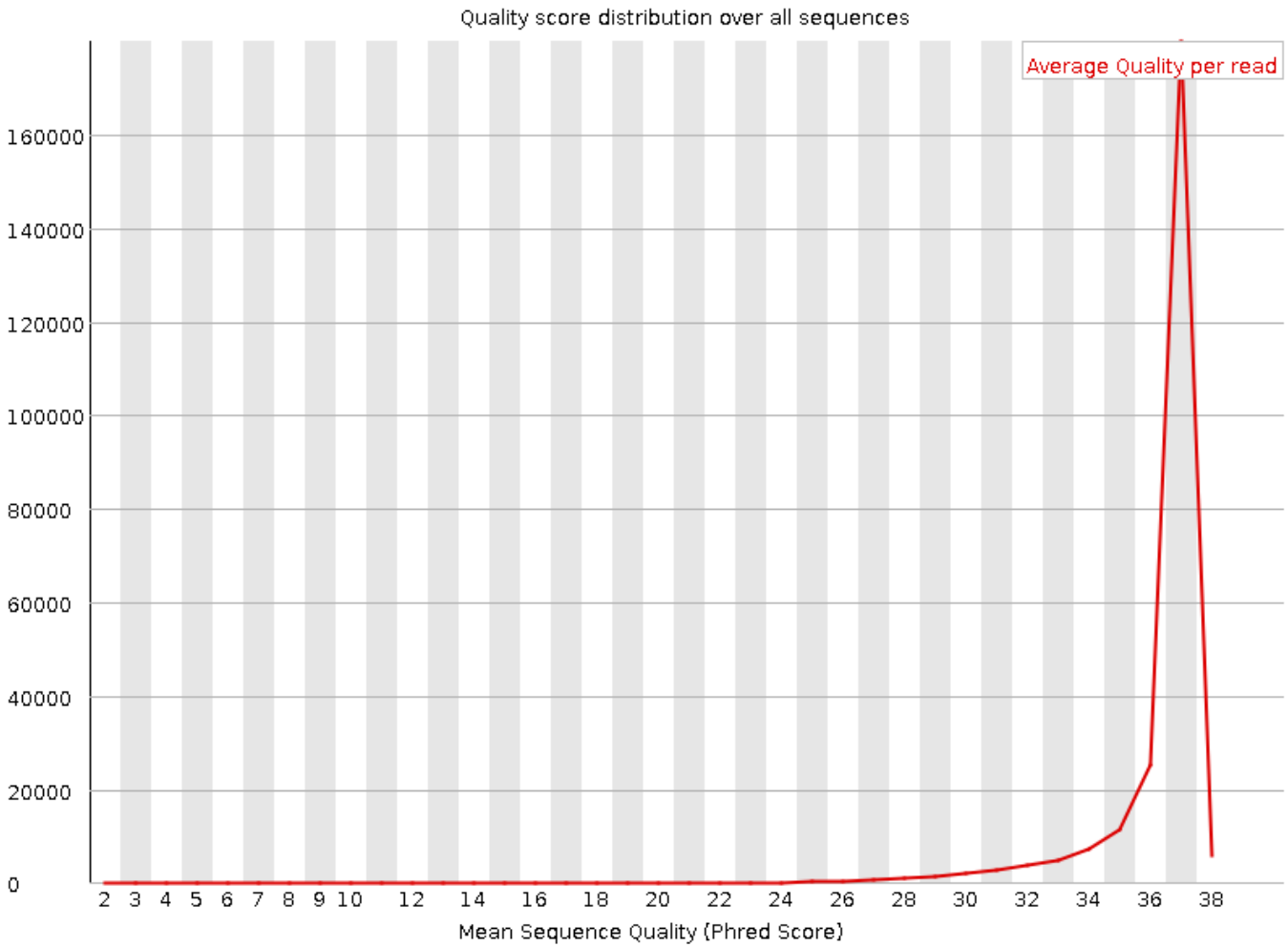
HTS: Quality check using FastQC

Bad data



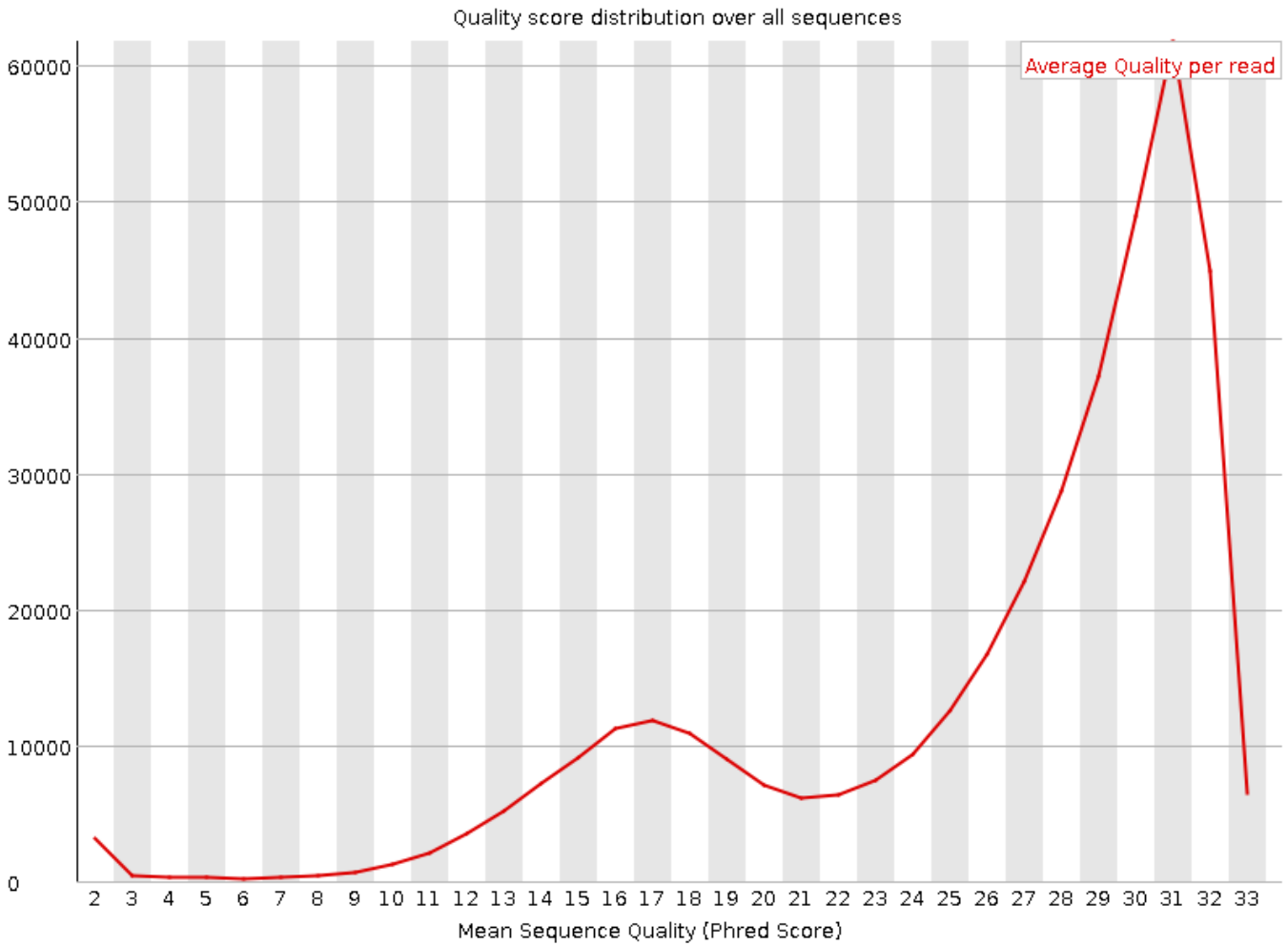
HTS: Quality check using FastQC

Good data



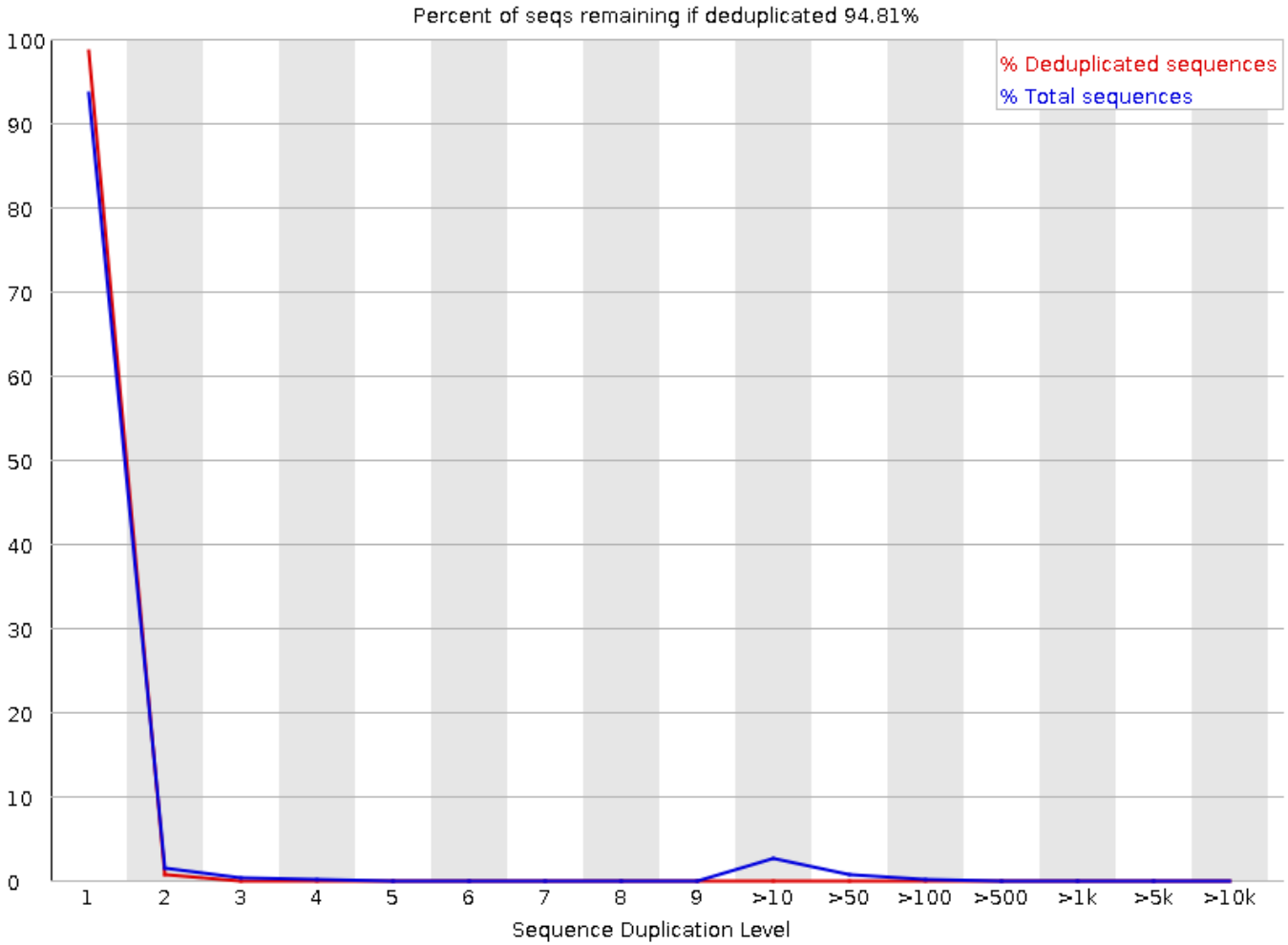
HTS: Quality check using FastQC

Bad data



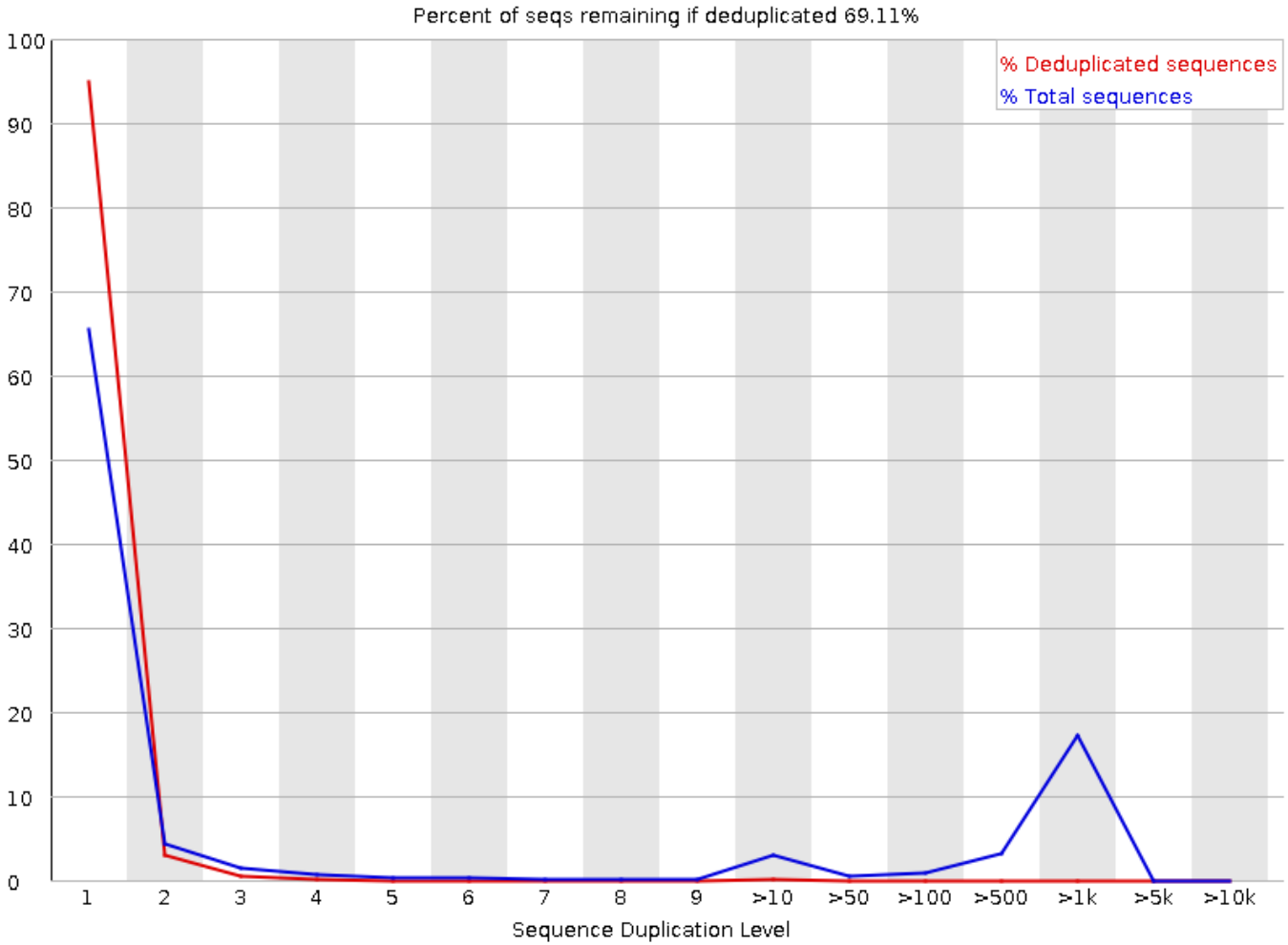
HTS: Quality check using FastQC

Good data



HTS: Quality check using FastQC

Bad data



Adapter and Low Quality Trimming

Trim Galore

- Developed by Babraham Bioinformatics
- Removes adapters
- Trims low quality reads
- Removes short sequences
- Accepts FastQ or compressed FastQ
- www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Adapter and Low Quality Trimming

Trimmomatic

- Developed in Java by USADEL Labs
- Trims low quality reads
- Filters short sequences
- `http://www.usadellab.org/cms/?page=trimmomatic`

Adapter and Low Quality Trimming

Prinseq

- Developed by San Diego State University
- Trims low quality reads
- Removes short sequences
- Accepts FastA, FastA+Qual and FastQ
- Removes duplicates
- `http://prinseq.sourceforge.net/index.html`