Session 2.2 – Quality assessment and read preprocessing

Sarai Varona

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

28 Junio - 2 Julio 2021, 3ª Edición
Programa Formación Continua, ISCIII

Step in the process

Fastq files

Preprocessing

Alignment

Postprocessing: Genome analysis…

Raw output files format

Illumina

.fastq

454 .sff

SOLiD

.fasta

.qual

Nanopore

FAST5

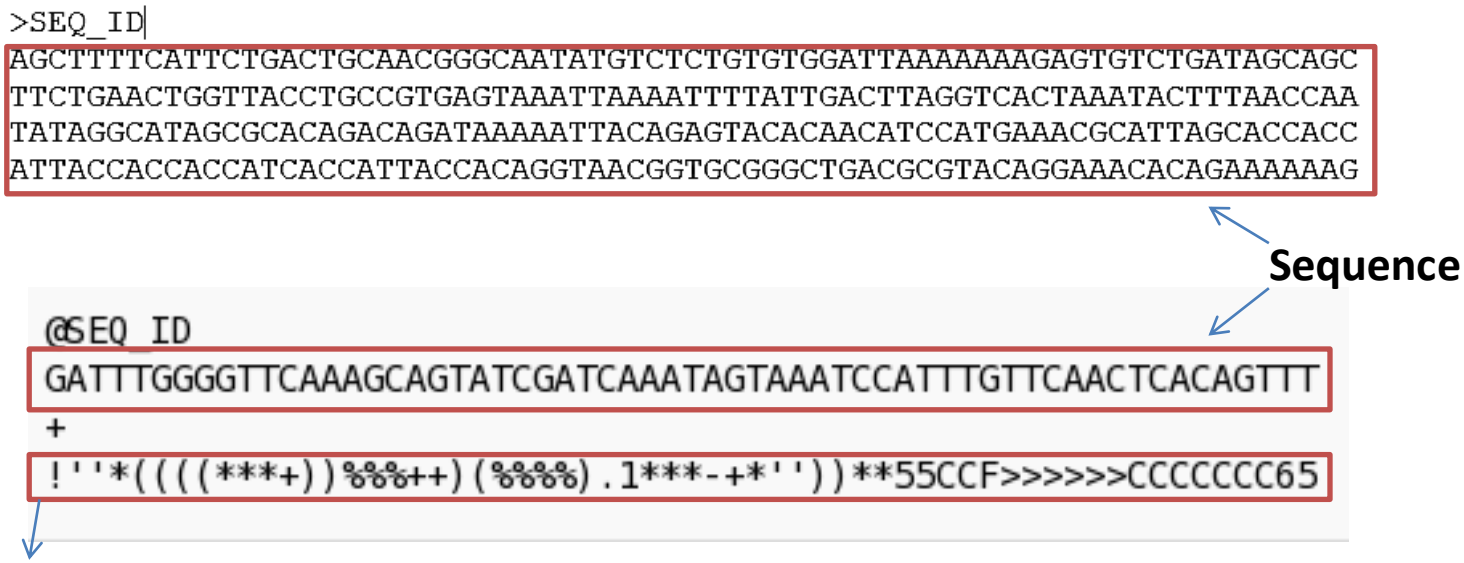
PacBio RSII

Bax.h5

fasta

FASTQ format

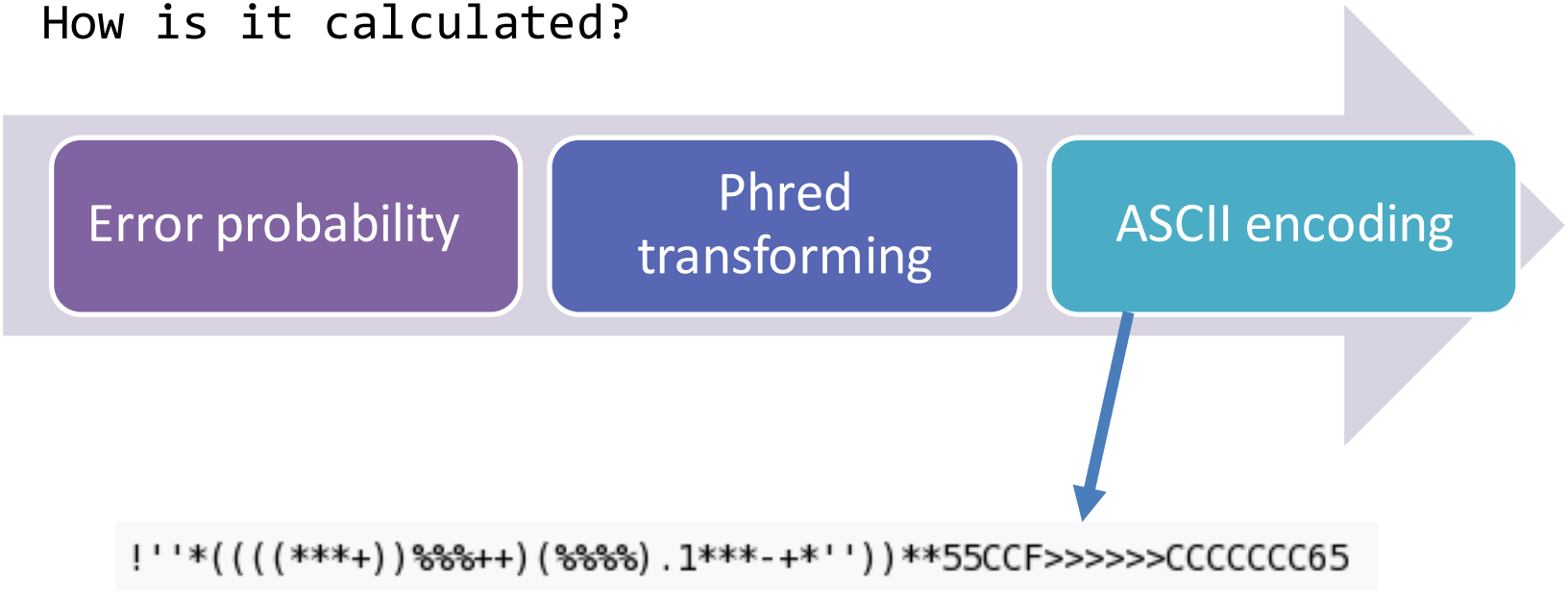
- Is a FASTA file with quality information
- Within HTS, FASTA contain genomes y FASTQ reads



Quality: must be 1 bit

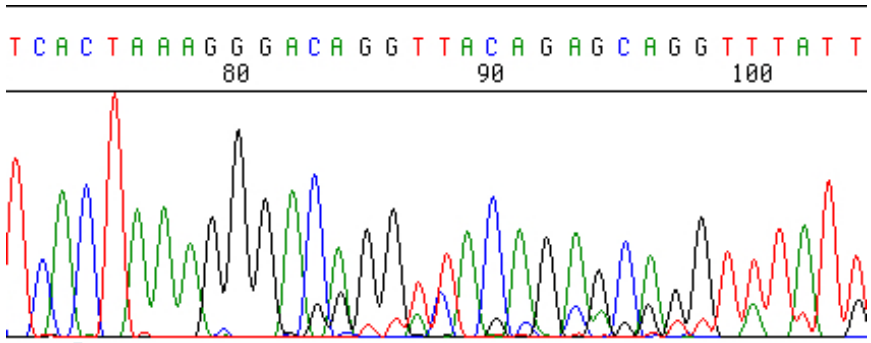
FASTQ format

- Each base has an assigned quality score
  - Sequencing quality scores measure the probability that a base is called incorrectly
- How is it calculated?

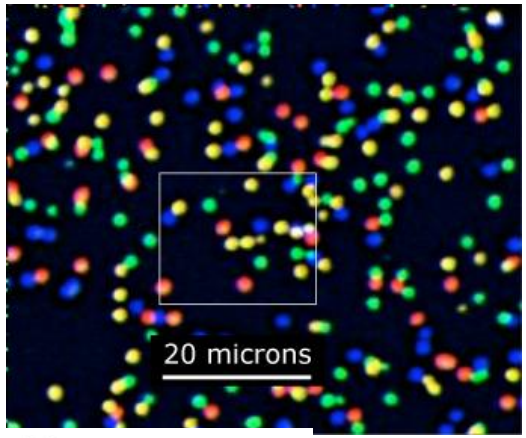


Phred quality and error probability

- **Light intensity** is used to calculate the error probabilities
- Convert error probability into Phred score quality - Ewing B, Green P. (1998)
- Phred originated as an algorithmic approach that considered Sanger sequencing metrics, such as **peak resolution and shape**



wellcome sanger institute



illumina®

Phred quality and error probability

- Convert error probability into Phred score quality - in real time on Illumina platforms
- Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)
- Phred quality range between 0-40 for Sanger and Illumina 1.8+

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Phred quality and error probability

- Convert Phred quality score into ASCII, a compact form, which uses only 1 byte per quality value

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

- Phred+33 (Sanger and current Illumina). 0 Phred quality correspond to decimal 33, which is the symbol !

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

- Phred+64 (Solexa and Illumina 1.3-1.5)

Phred quality and error probability

- **Phred 33 example**

```
@HWI-ST731_6:1:1101:1322:1938#1@0/1
NTGACAAAGGGCTAATATCCAGAATCTACAAAGAACTTAAACAAATGTATAAGAATAAAAGTATAGTGCTAACAAT
+
#1:BDDADFDFDD@F>BGFIIIB@CFHIHICAGBC9CBCBGGIGCFF??>GGHFHIGGEGI<FECGDE=FHCHEG=
```

P=0.0001 → Q=-10*log10(0.0001)= 40 → ASCIII 33+40 = 73 → I

P=0.001 → Q=-10*log10(0.001)= 30 → ASCIII 33+30 = 63 → ?

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

FASTQ format

Illumina read header

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

```
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:2458:1027 1:N:0:ACAGTG
AGAAAAAACCTTGGANGGAAAAAAATCAGACATTTTCTAGAGGTGGAAGGCAAACTGAACAAAGAAATAATTCACA
+
DGGGEDHHHHGGGFE#CBACBCA<?HHHHBHHHHHHHHDHHHHEHEFEGGGGGG/GGDDDGHFHGFCHFHHEHEH8
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3082:1029 1:N:0:ACAGTG
GGTAATACAGACTGANATGATCAAAGGCATGCTGGAAACAAACCTATTAAAGATAAGCTTGGATCAAGCTTTCATT
+
B:B:?BB/:=55177#55877<775EDD>E=B?BBBBGGGDDAG@G>GGGGGG@)EEEEBEG>GGGGGGGAAA?<D
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3185:1033 1:N:0:ACAGTG
TCTGGGACATTGCTCNTGGCTGGGAGTCACCTGTCTGGGACATTGCTCAGGGCTGGGAGACACGTGTTGGAGGGAC
+
BC??A66;)74781<#7??;452.27'64(8,851DDG8GB?##############################
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3268:1033 1:N:0:ACAGTG
ATTCAAATTAGAAGANAGTTGATCGTTCTTCATGATGCCCAAAAATTTCACTGAGAAAACCCTTTTTTAAGCCCAC
+
IIIIIIIIIIFFFFE#ABACFEEFFIIGIIIFIHE@BIIIIIIIIIHHIIFIIF>HHIHIFGDIIIIIIGFHIEGH
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3400:1035 1:N:0:ACAGTG
TCCTGCTTTAGGAGANTCCTCATGCTCTGACAGGATGCTCTCTATGTGAGTTGAGCTGGTCTTCTCACTTTTATAG
+
IIIIIHIHIIGGEGG#AACA@?=?BHHIIIIIHHIHIIIHIHIHHGIHIHGHGIGIHGEGGGGHG@EFGGCEFAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3962:1033 1:N:0:ACAGTG
CCACCAACACAGTCTNCACCTTCTGTTGCTGGTGATAGATTTTTGCACCTTTCCATCCTCCAGGTTTCAAAATAGC
+
HHFHHDHDHH>C?CA#EEEE>?A?>HHDGHEGBGBCEEEEGHHF8HEHEEHECH,=>>==EAEE>BEBBAEAACAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:4491:1028 1:N:0:ACAGTG
AGAGAGAGAGAGAGANAGAGGACTCTGGAGATGCCGAAGCACAAGCCTGCAAGAGTCCCAGCAAAGAAAATAAAAA
+
GADGGEGGEGBBB?B#@=@@72;64GGGFGB>GGGBDG<DBGB<DA??/?##########################
```

ASCII-coded (0-40):

- "!"#$%" lowest quality
- "FGHI" highest quality

Sequencing quality assessment

- To asses quality, software uses **Phred per-base quality** score is used
- Is the **first quality control step** after sequencing. There should be one after every step of the analysis
- After quality assessment user can know how **reliable** are their datasets
- QC will determine the next **filtering** step
- Filtering decisions will **impact** directly in **further analysis**
- Many other steps also use this quality as variable in their **algorithms**

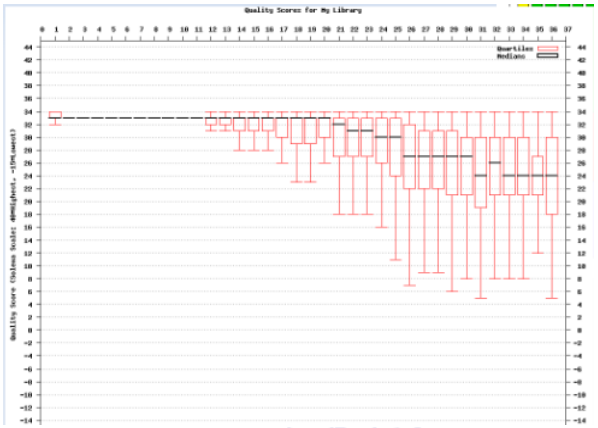
Sequencing quality assessment: Artifacts

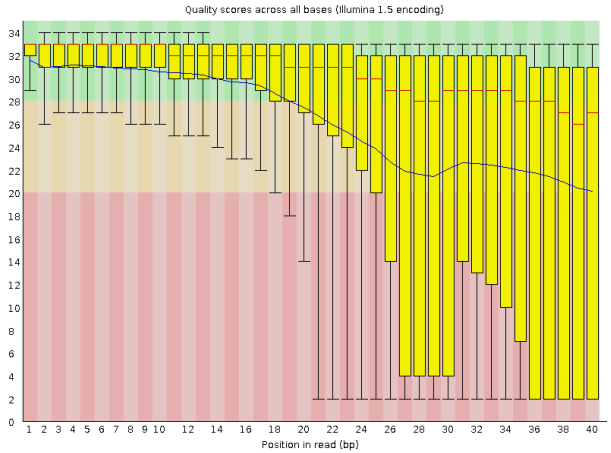
HTS methods are bounded by their technical and theoretical limitations and sequencing errors cannot be completely eliminated (Hadigol M, Khiabanian H. 2018)

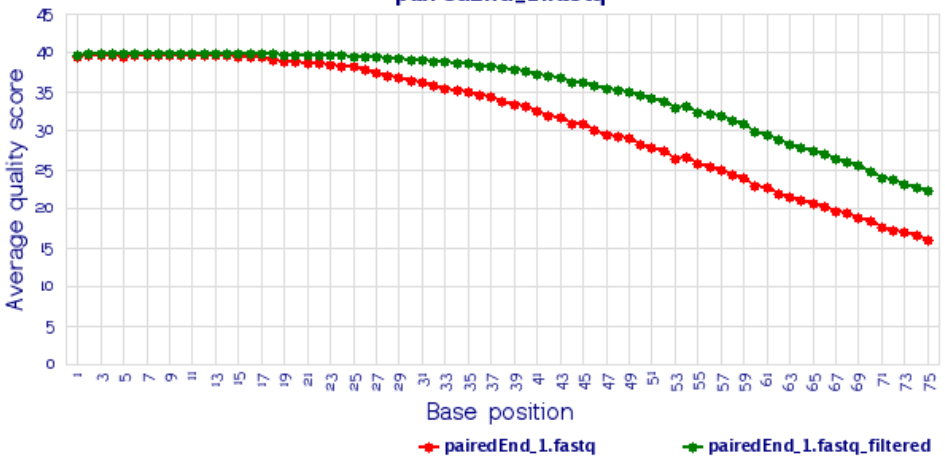
- **Artifacts in library preparation**
  - Remaining adapters
  - High rate of duplicates
  - GC regions bias
  - Polymerase error rate
  - DNA damage during breakdown
- **Artifacts during secuencing**
  - Low quality in sequence ends(Phasing: cluster loose sync)
  - Complication in certain regions:
    - Repetitions
    - Homopolymers
    - High CG content

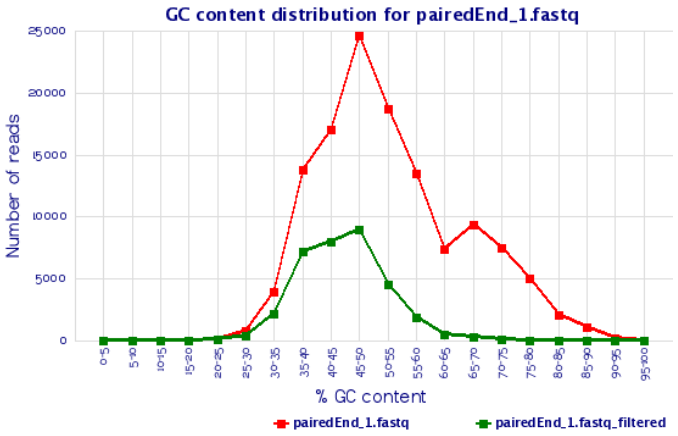
Sequencing quality assessment

- FastQC, fastx-toolkit, sfftools, NGSQCToolkit, etc…



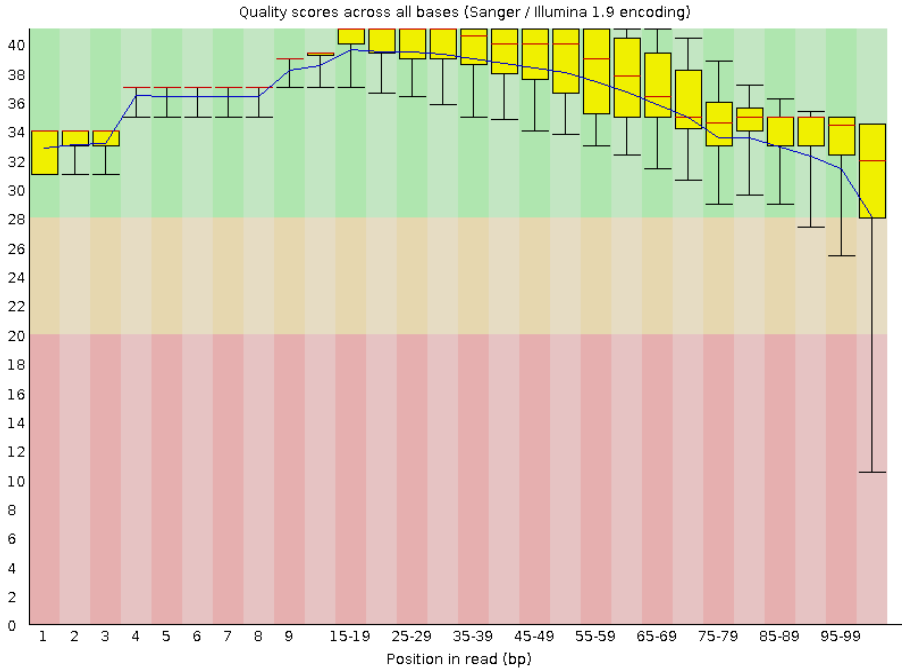




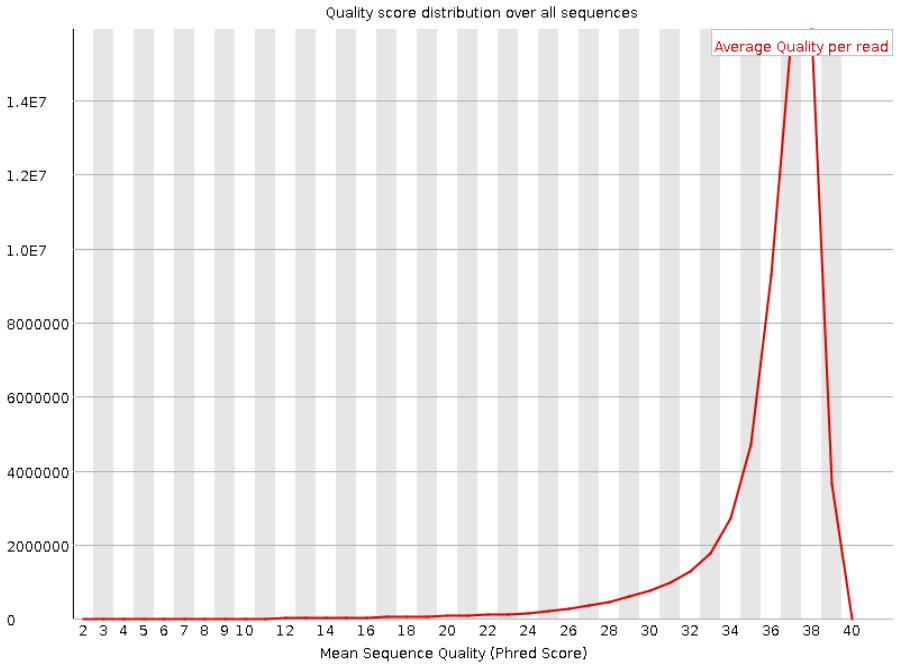


Sequencing quality assessment: FastQC

Per base sequence quality



Per sequence quality scores



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

Sequencing quality assessment: fastp

- Fastp

fastp report

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	paired end (149 cycles + 149 cycles)
mean length before filtering:	116bp, 116bp
mean length after filtering:	117bp, 117bp
duplication rate:	1.704150%
Insert size peak:	95
Detected read1 adapter:	CACCTAAGTTGGCGTATACGCGTAATATATCTGGGTTTTCTACAAAATCATACCAGTCCT
Detected read2 adapter:	CACCTAAGTTGGCGTATACGCGTAATATATCTGGGTTTTCTACAAAATCATACCAGTCCT

Before filtering

total reads:	1.296756 M
total bases:	151.424921 M
Q20 bases:	143.112834 M (94.510754%)
Q30 bases:	137.905419 M (91.071812%)
GC content:	40.410939%

After filtering

total reads:	854.250000 K
total bases:	100.537720 M
Q20 bases:	99.598139 M (99.065444%)
Q30 bases:	97.968091 M (97.444115%)
GC content:	39.665634%

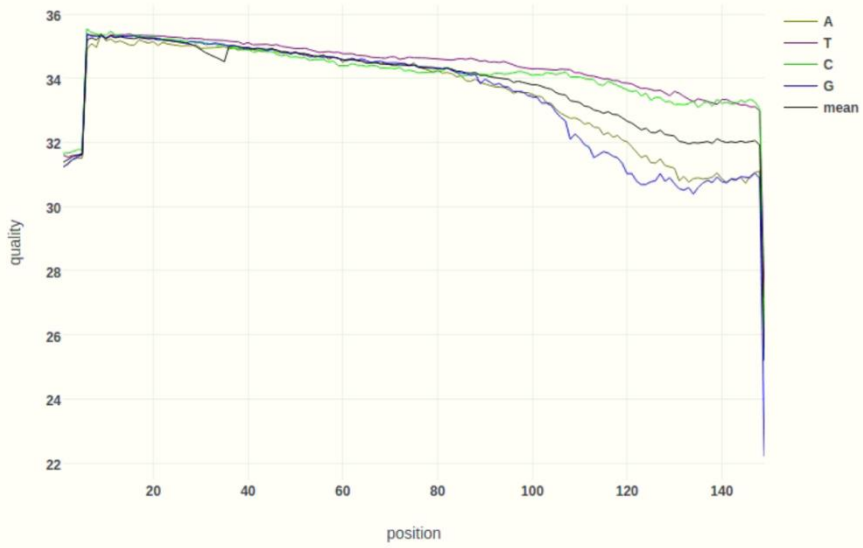
Filtering result

reads passed filters:	854.250000 K (65.875924%)
reads with low quality:	352.272000 K (27.165635%)
reads with too many N:	84 (0.006478%)
reads too short:	90.150000 K (6.951963%)

Before filtering

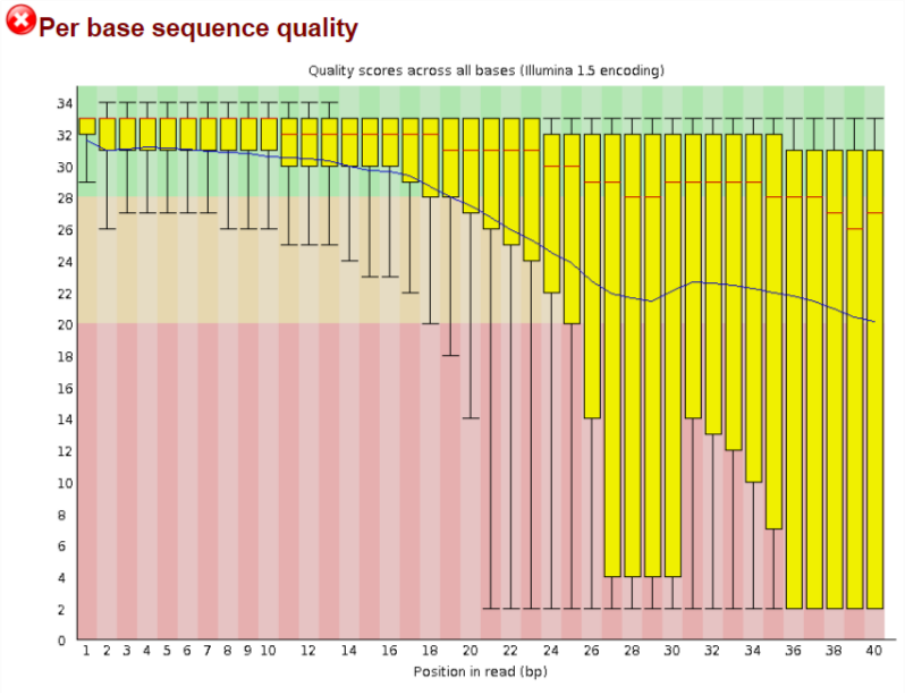
Before filtering: read1: quality

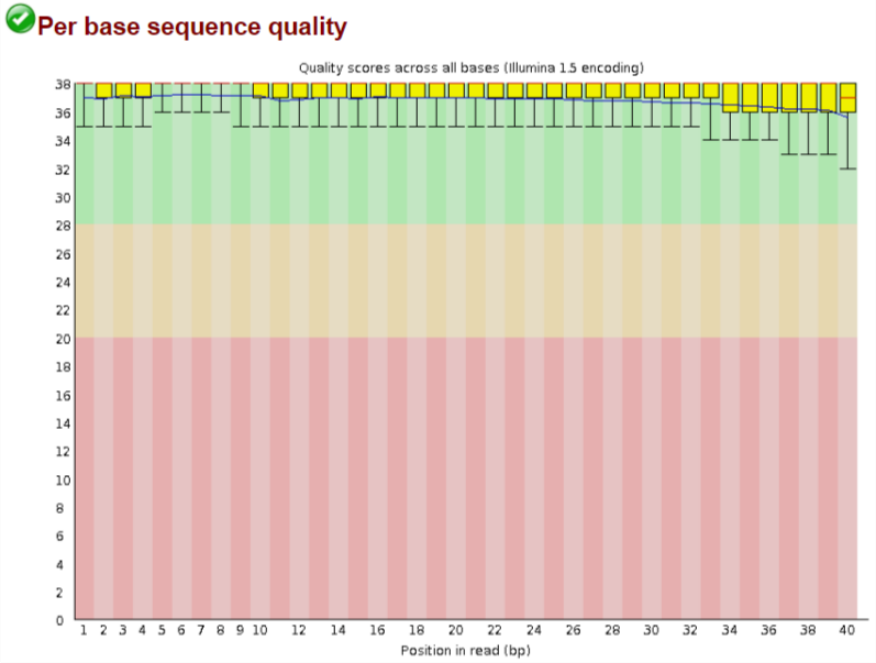
Value of each position will be shown on mouse over.



FastQC: Per base sequence quality

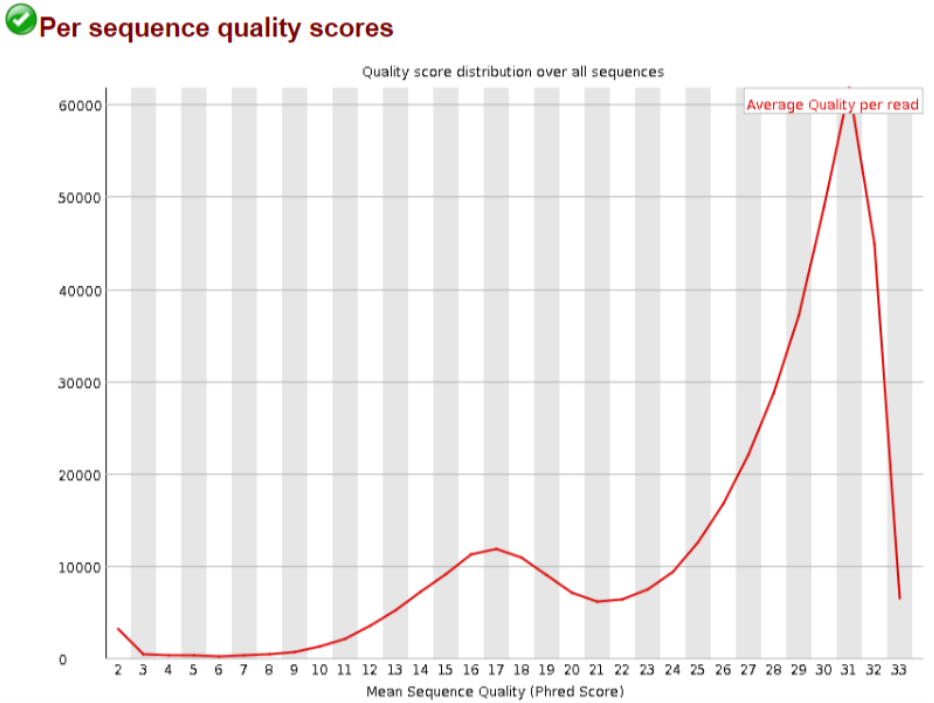
- Overview of the range of quality values across all bases at each position in the FastQ file
- **Median**, **inter-quartile range** **(25-75%), 10-90% points,** **mean quality**

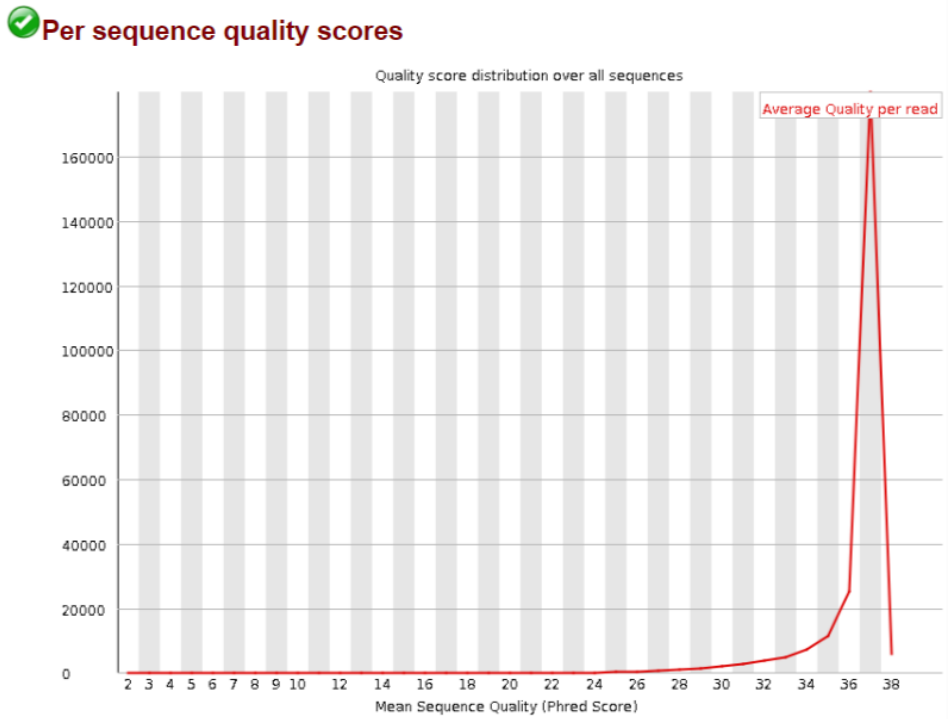




FastQC: Per sequence quality score

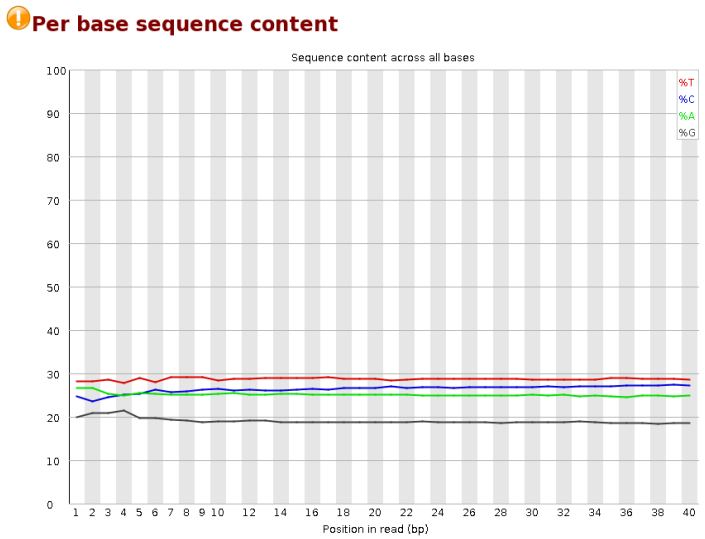
- Number of sequences with the same mean quality

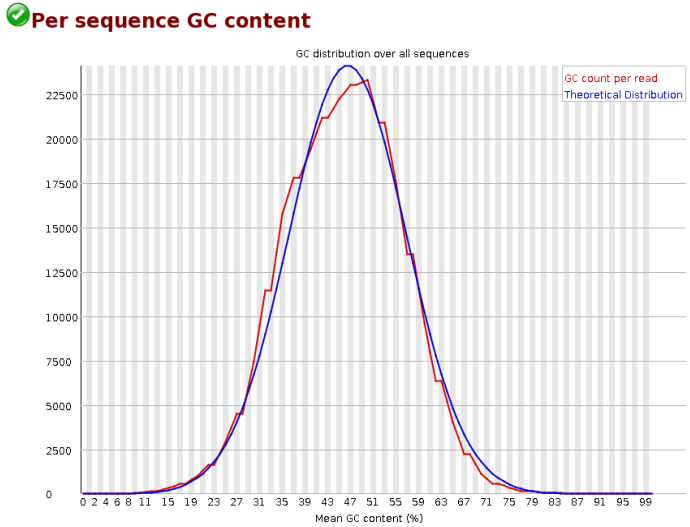




FastQC: Nucleotide related errors

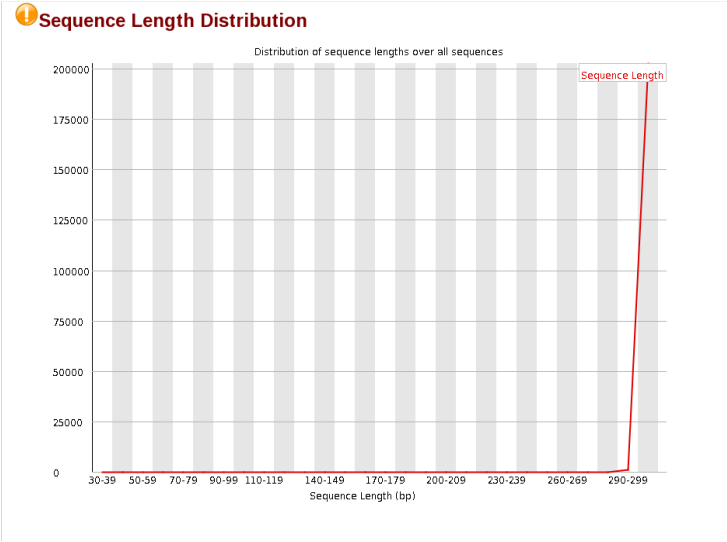
- How expected nucleotide distribution deviates from expected
  - Per base sequence content
  - Per base GC content
  - Per sequence GC content
  - Per base N content

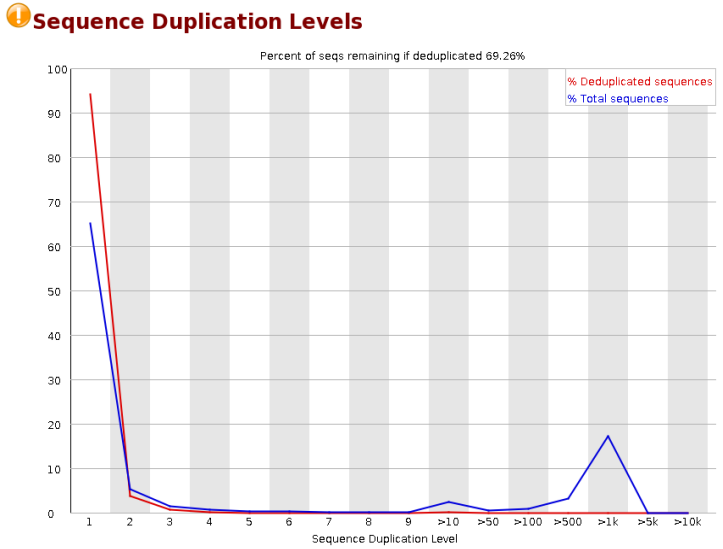




FastQC: Sequence related errors

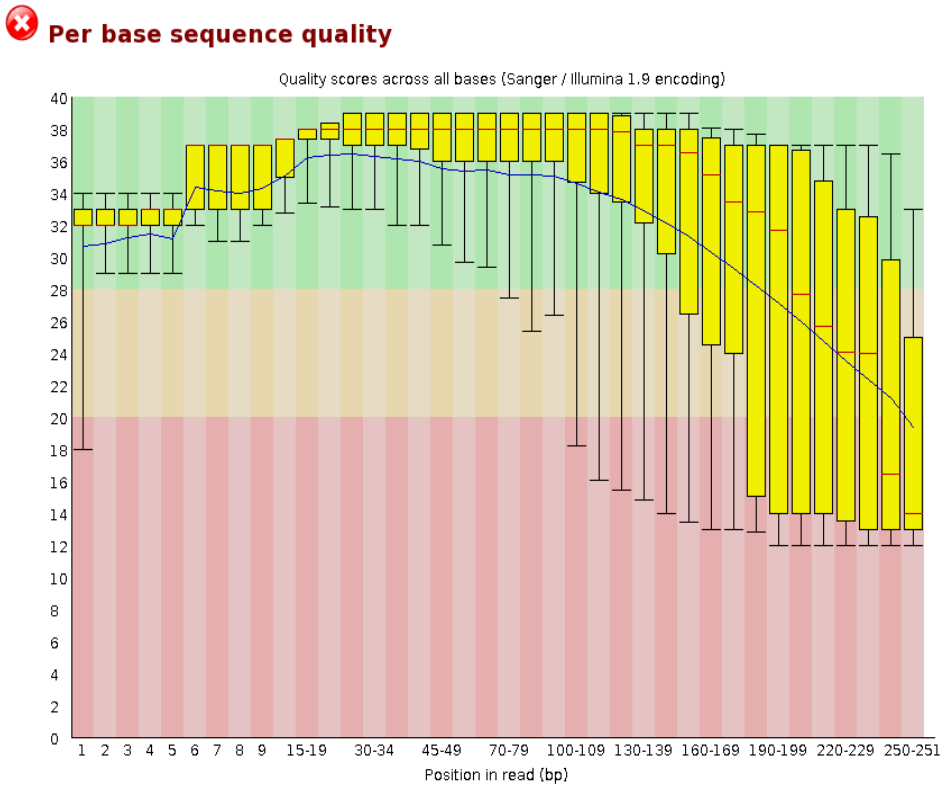
- How expected nucleotide distribution deviates from expected
  - Sequence Length Distribution - Fragments
  - Sequence Duplication Levels
  - Overrepresented sequences
  - Adapter Content

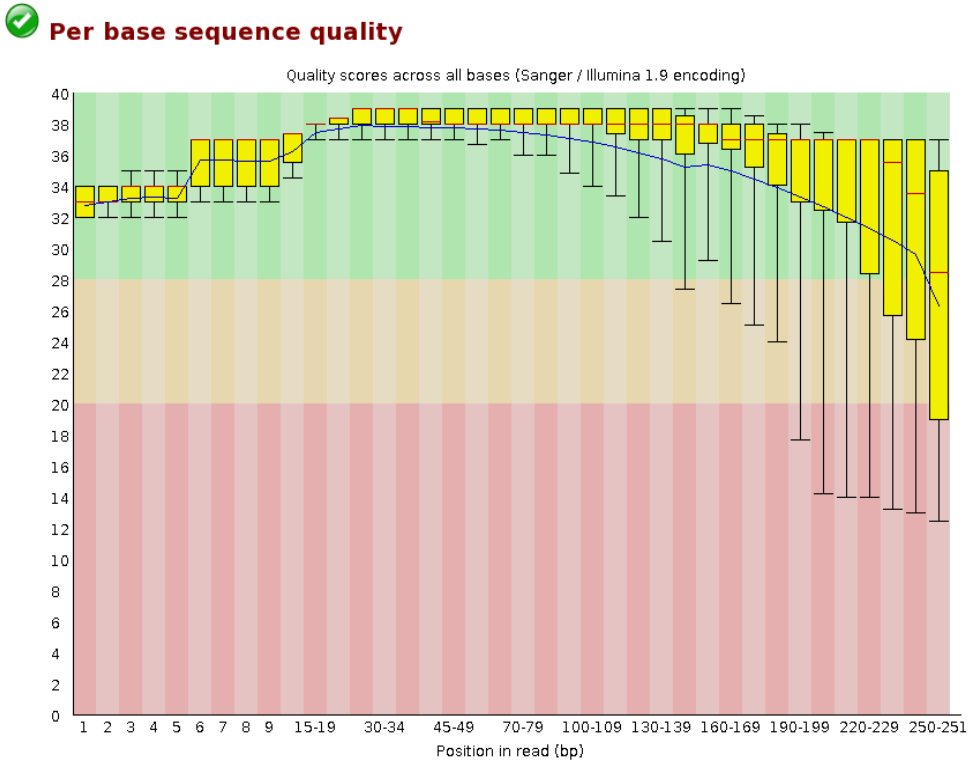




FastQC: Per base sequence quality

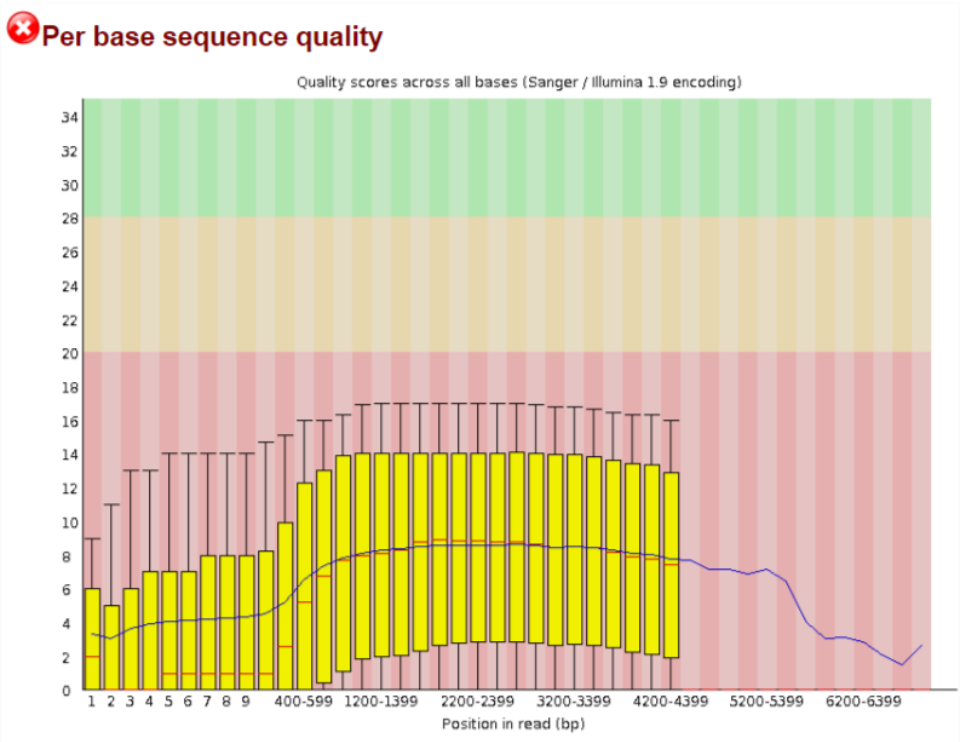
- **Miseq assymetry**

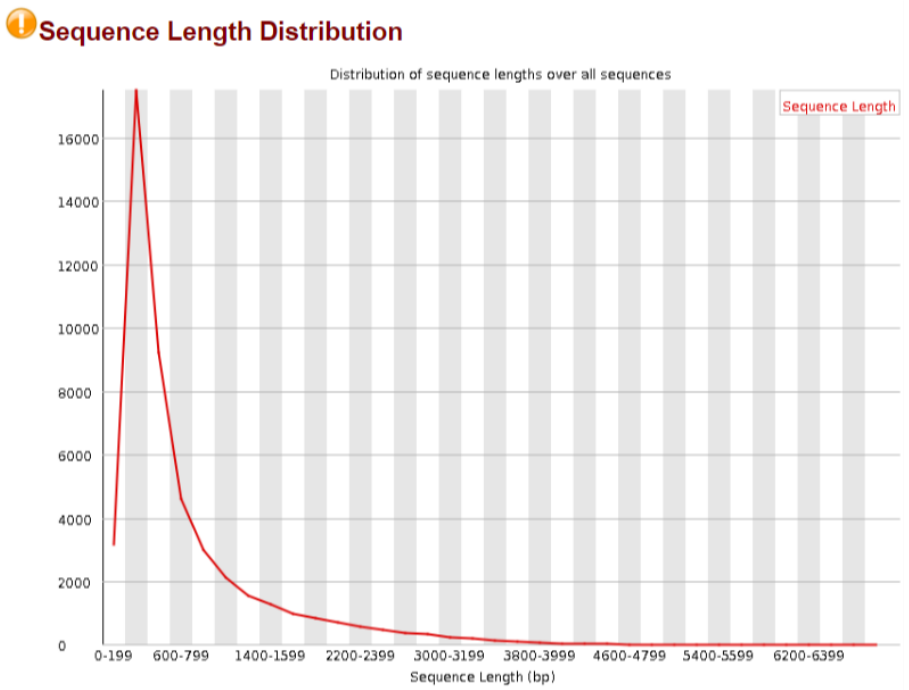




FastQC: Per base sequence quality

- **SMRT PacBio**





Sequence filtering

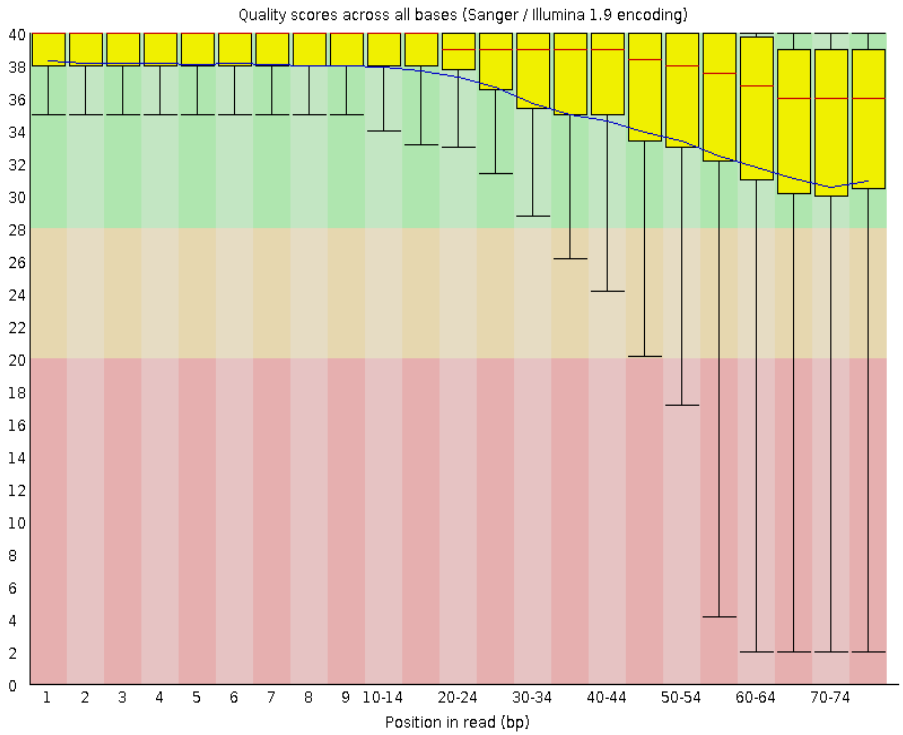
- **Remove residual adapters**
  - Depending on used library

- **Filtering parameters**
  - Quality filtering
    - Overall mean quality
    - Local mean quality
      - Sequence end
      - Sliding window
  - Size filtering
    - Overall sequence size
    - Remaining sequence size after filtering

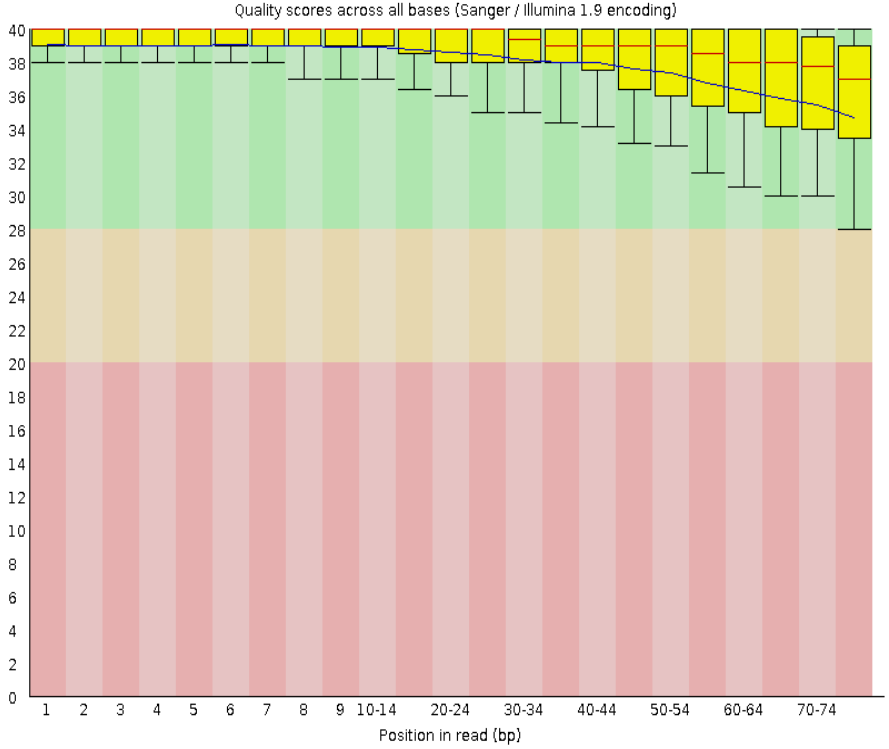
Sequence filtering

- **Example of quality filtering**

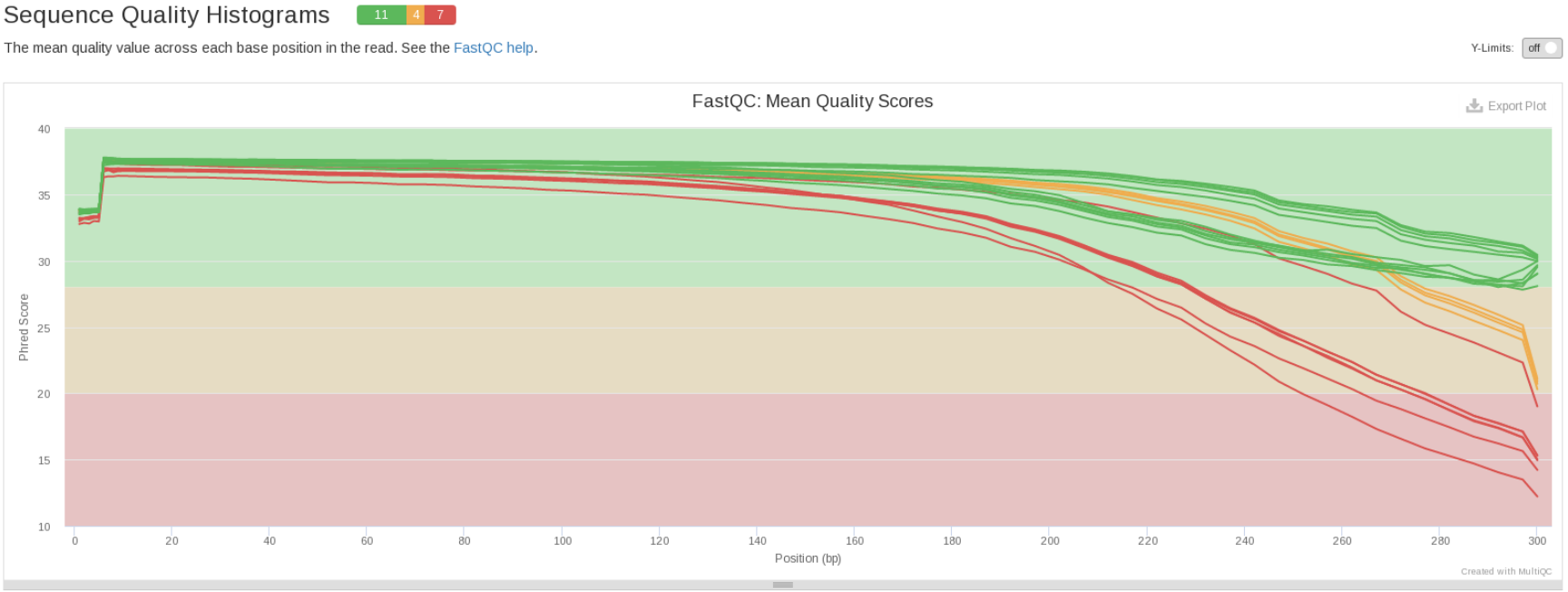
Per base sequence quality



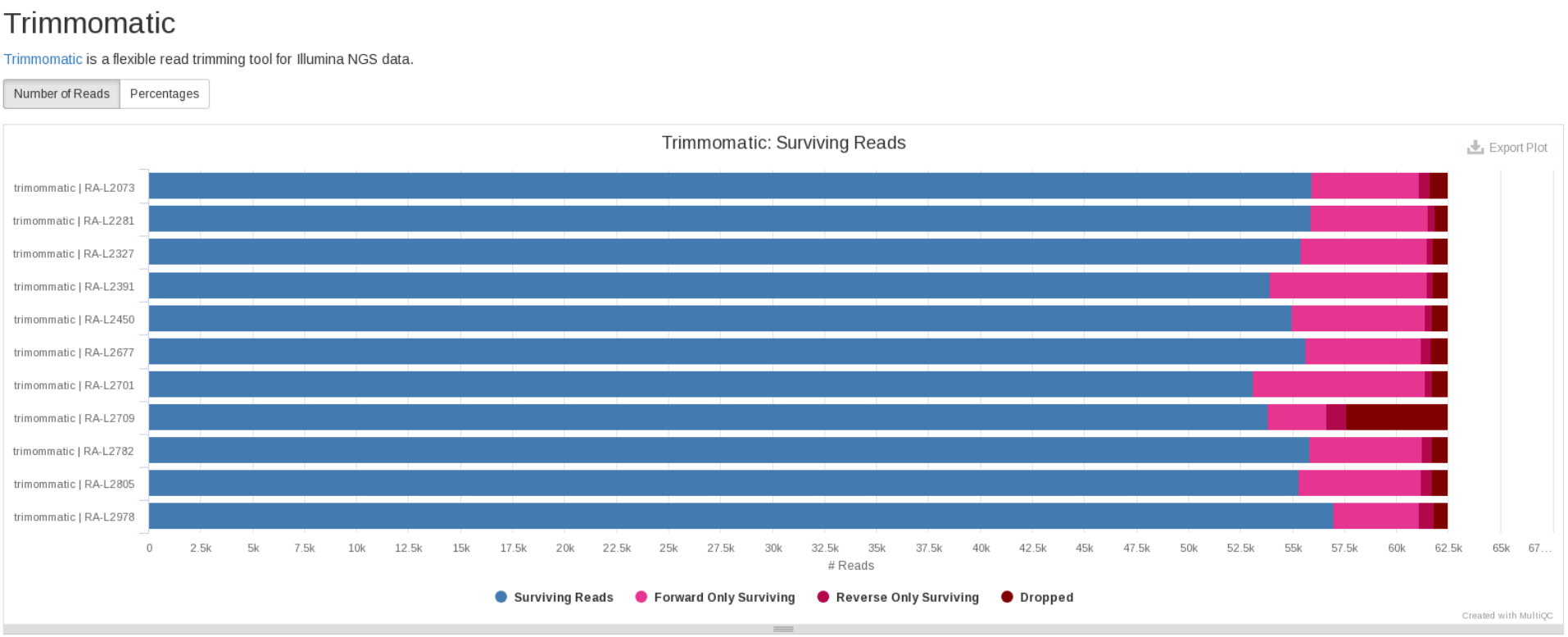
Per base sequence quality



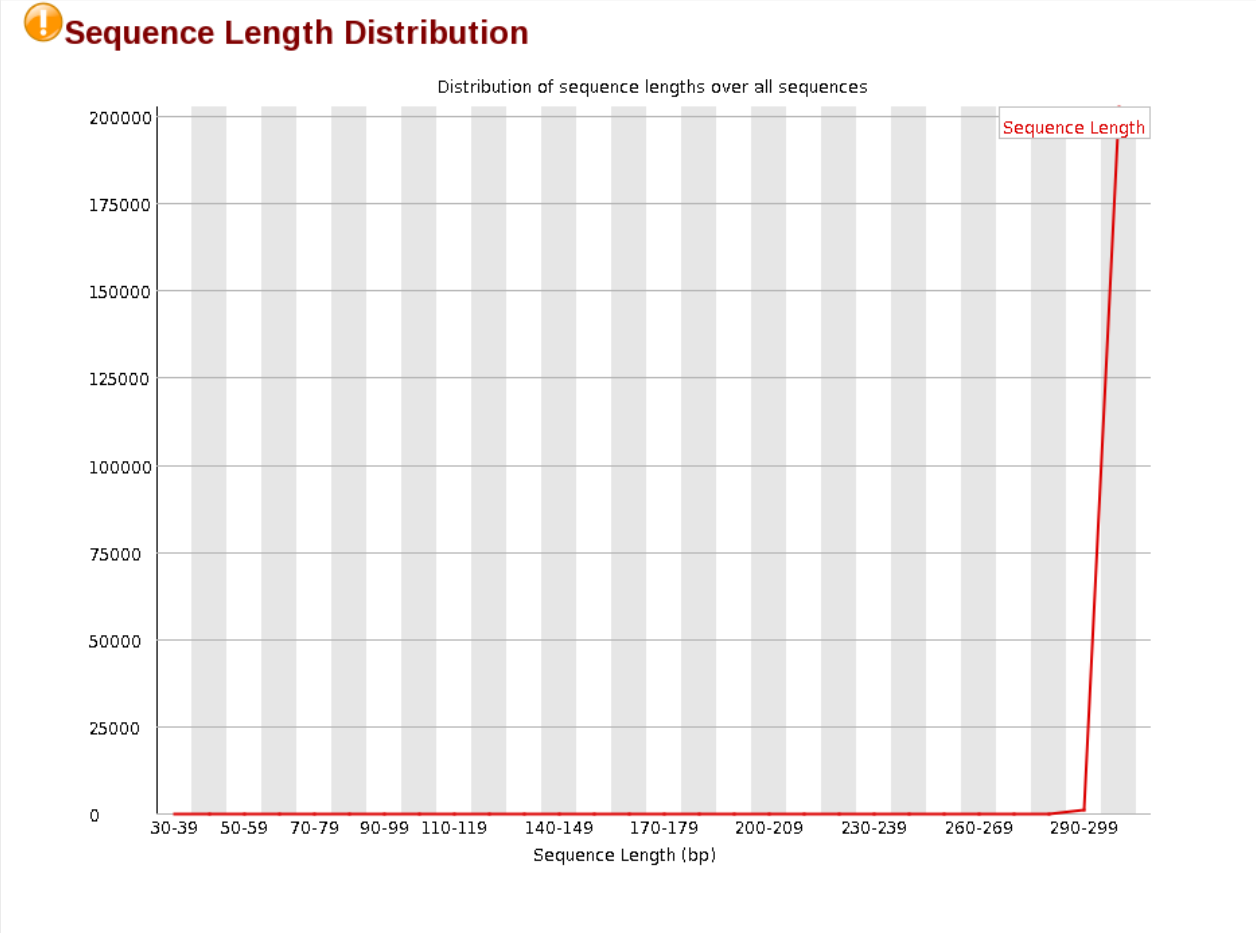
Sequence filtering: stats with MultiQC



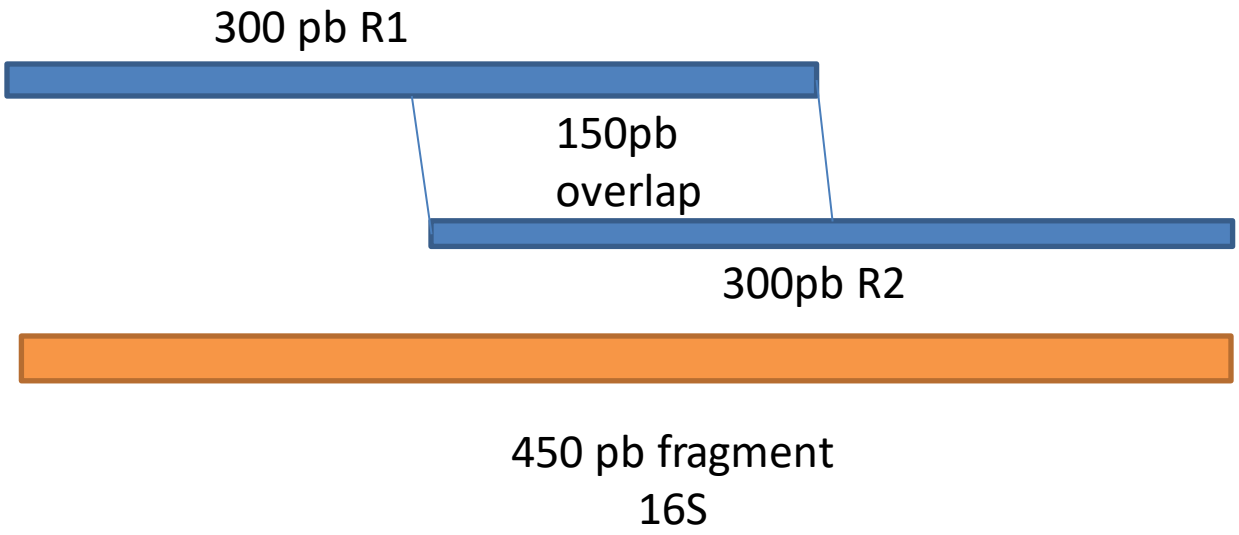
Sequence filtering: stats with MultiQC



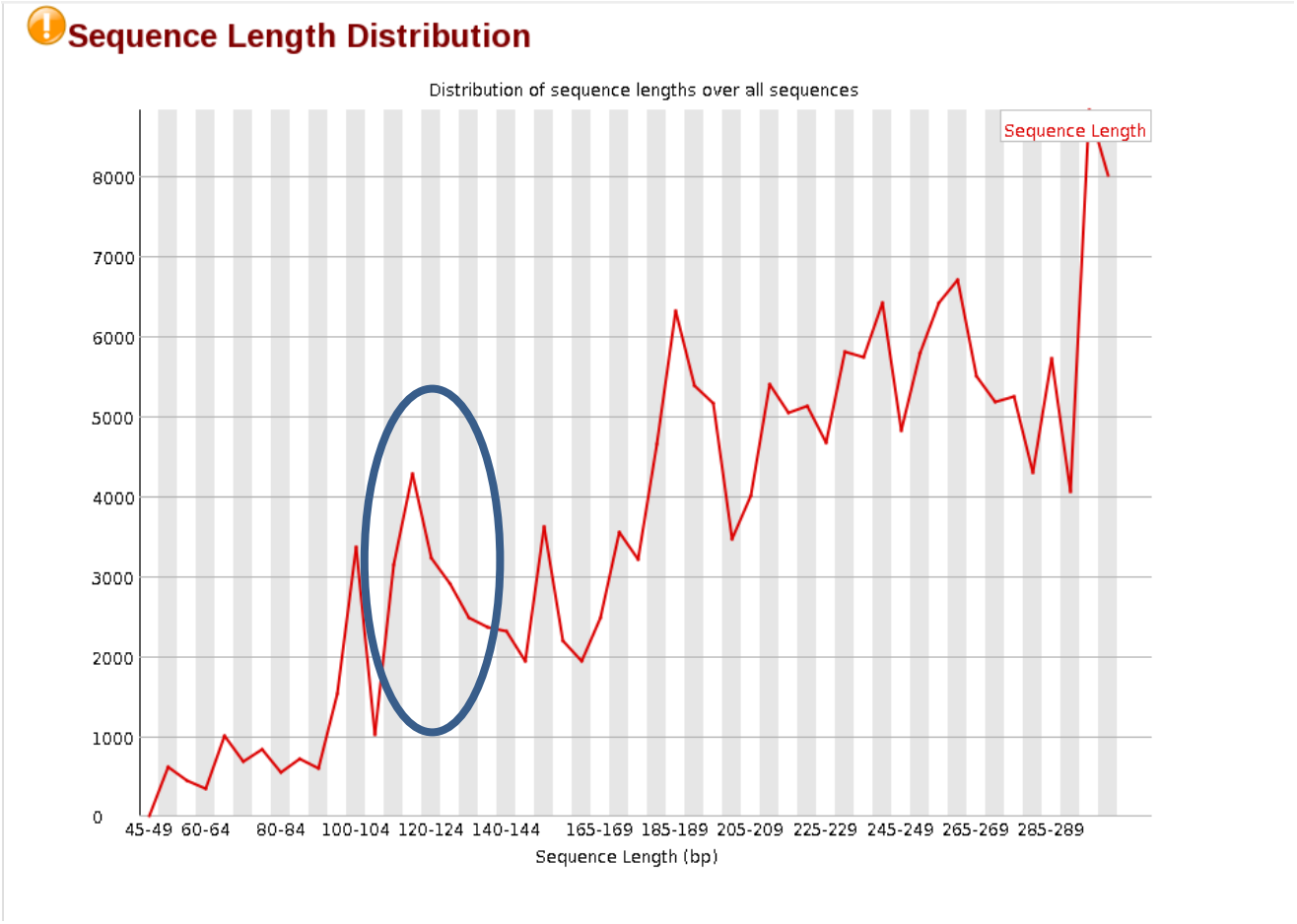
Quality filtering in metagenomic samples



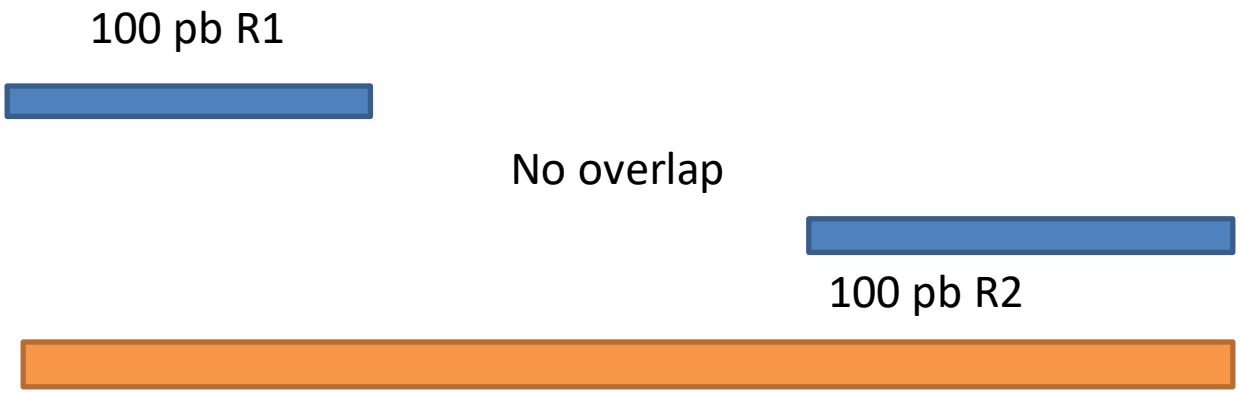
Quality filtering in metagenomic samples



Quality filtering in metagenomic samples



Quality filtering in metagenomic samples



Questions?