Quality Control for Raw Data

Matías Marín Falco

Mar 6th 2017

GDA

International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Contents

- Data formats
  - Sequence capture
  - Fasta and fastq formats
  - Sequence quality encoding
- Quality Control
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artifacts & filtering

Sequence capture

illumina

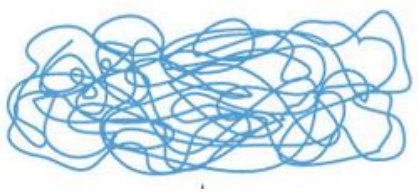
Roche

applied biosystems

RAW data
Propietary format

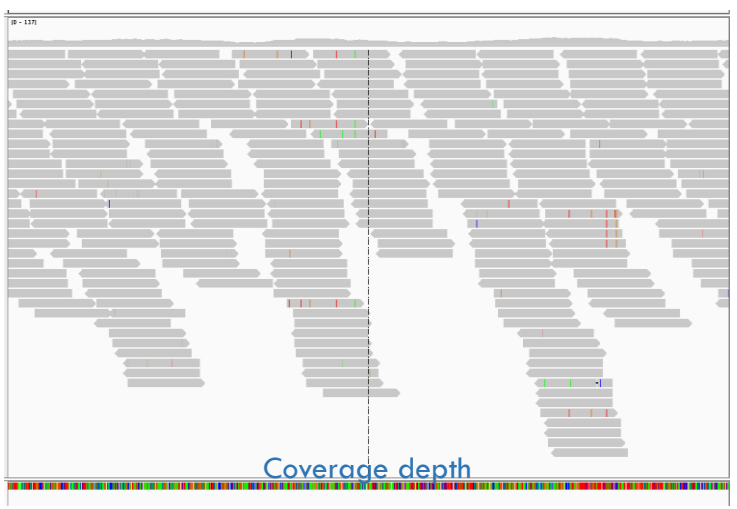
FastQ

Genome Sequencing

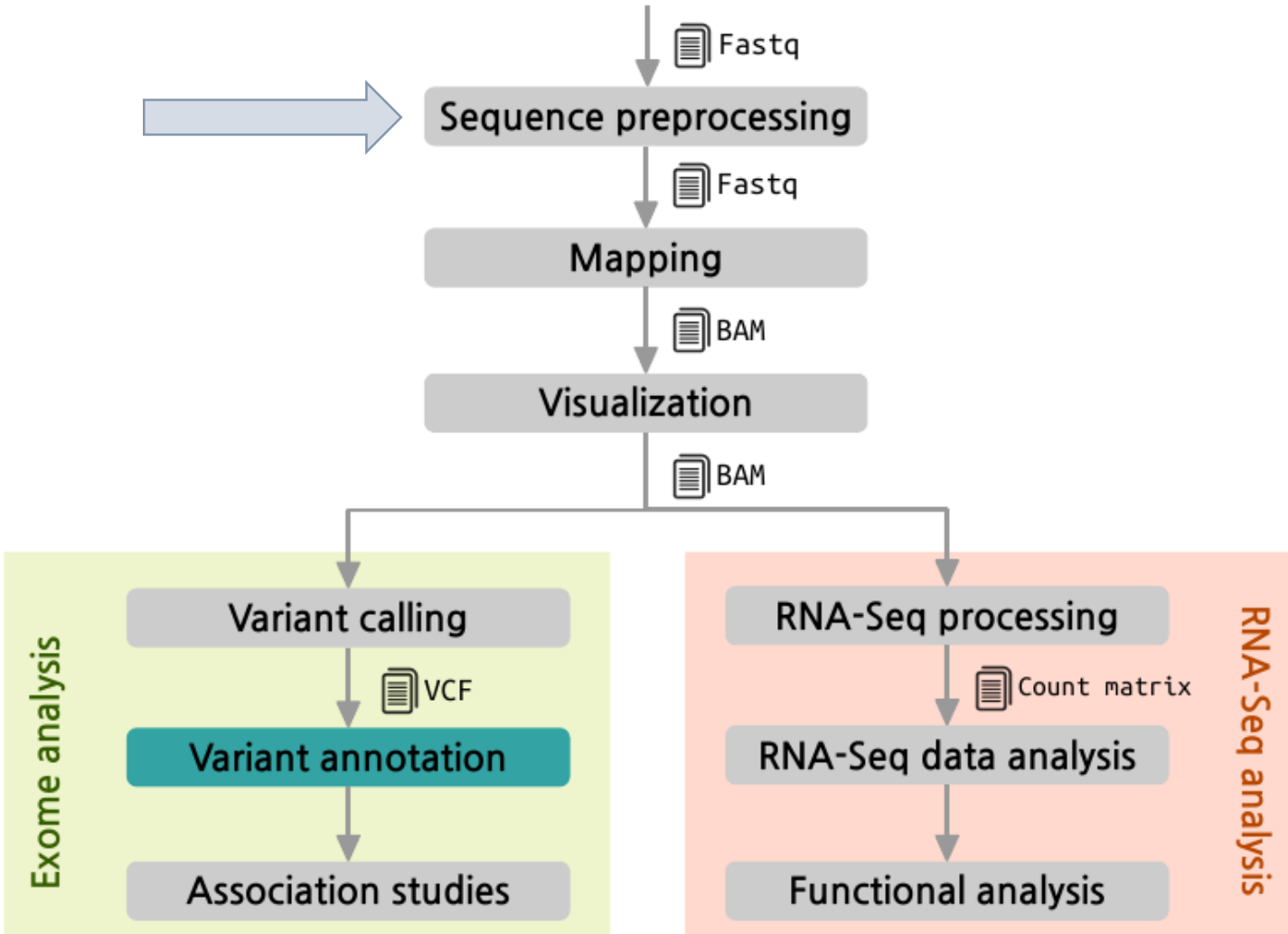


Reads

Reference genome



Course pipeline



From sequencers to digital data

- **What structure does the data have?**
  - Text-based formats (easy to use!)
  - If not compressed, it can be huge

- **Data formats:**
  - Different sequencers output different files (sff, fasta, csfasta, qual file, fastq...)
  - There are some data formats widely accepted (e.g. FastQ format)

Fasta format

- Two lines per sequence:
  - 1. Header lines starts with “>” followed by a sequence ID
  - 2. Sequence (string of nt or peptides)

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

```
>SEQ_ID
tgcaccaaacatgtctaaagctggaaccaaaattactttctttgaagacaaaaactttca
Aggccgccactatgacagcgattgcgactgtgcagatttccacatgtacctgagccgctg
>SEQ_ID2
caactccatcagagtggaaggaggcacctgggctgtgtatgaaaggcccaattttgctgg
gtacatgtacatcctaccccggggcgagtatcctgagtaccagcactggatgggcctcaa
```

- Typical file extensions (.fasta, .fa, .fna, .fnn, .faa, …)

Fastq format

- We could say "it is a fasta with **qualities**":
  - 1. Header (like the fasta but starting with "@")
  - 2. Sequence (string of nt)
  - 3. "+" and sequence ID (optional)
  - 4. Encoded quality of the sequence

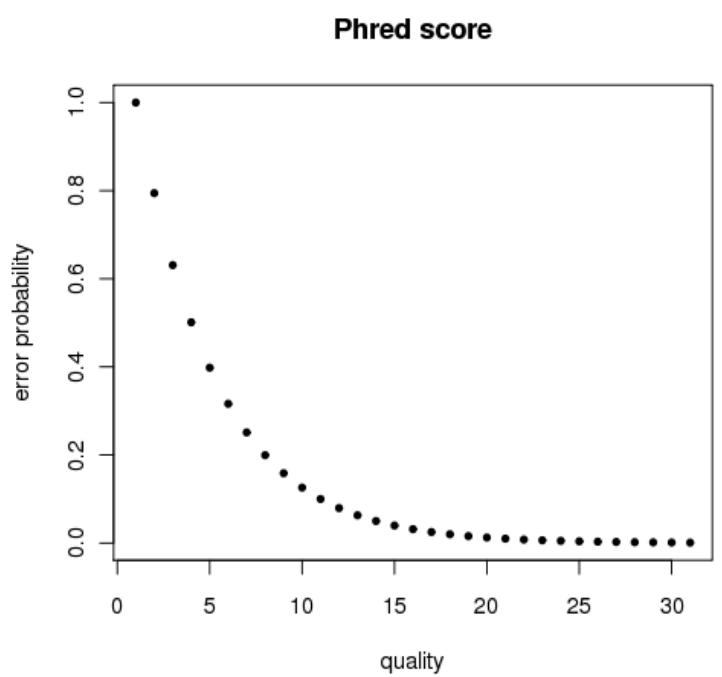
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality codification

- Phred scores

$$Q = -10 \log_{10} P \quad \Longleftrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%



Quality codification

- Phred quality score
- Error probability
- ASCII encoded

- Phred +33
  - Sanger [0,40]
  - Illumina 1.8 [0,41]
  - Illumina 1.9 [0,41]

- Phred +64
  - Illumina 1.3 [0,40]
  - Illumina 1.5 [3,40]

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			

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 `			

Sequence quality evaluation

- If we evaluate our sequence in depth …

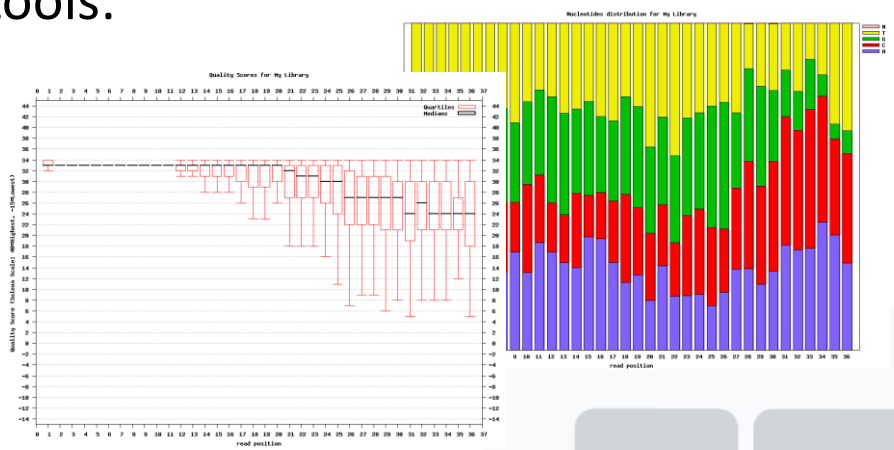
… we will know how reliable our results are

- **Problem:**
- **Huge files** → Need of a tool to do it

Sequence quality evaluation

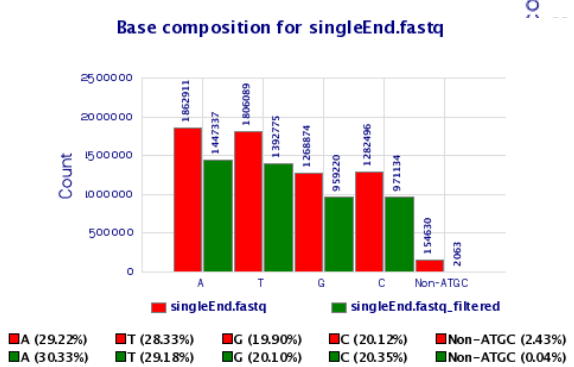
- Quality control tools:

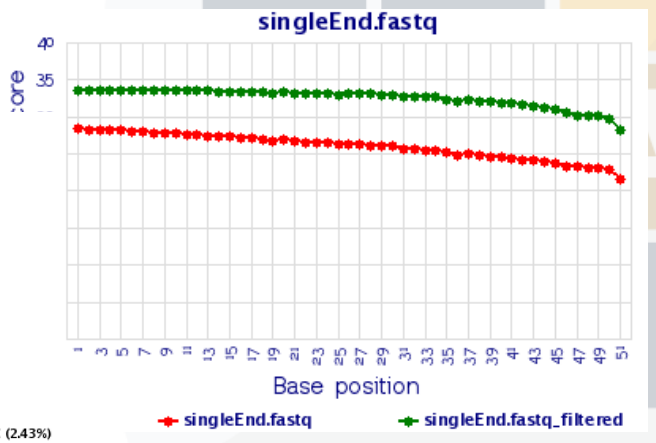
- **Fastx-toolkit**



http://hannonlab.cshl.edu/fastx_toolkit/download.html

- **NGS QC Toolkit**



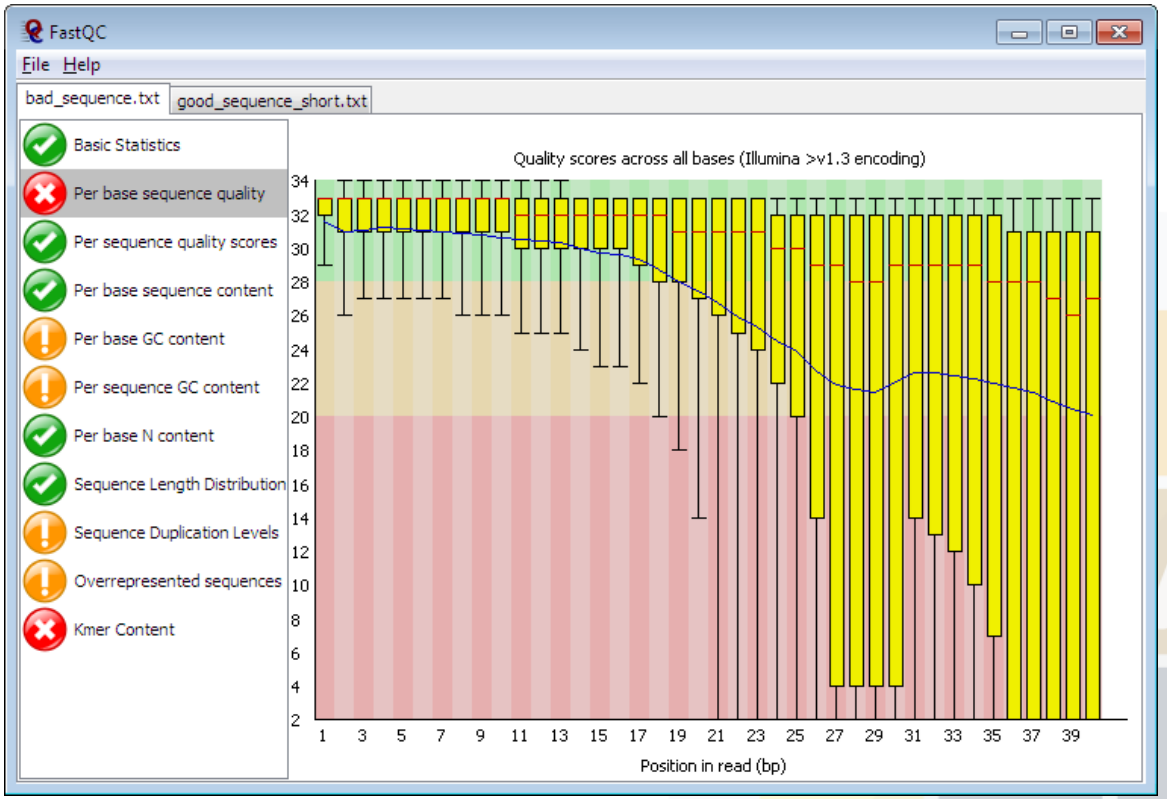


http://www.nipgr.res.in/ngsqctoolkit.html

Sequence quality evaluation

- Other quality control tool: **FastQC**

Analysis modules



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

Quality analysis examples

GOOD quality

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

POOR quality

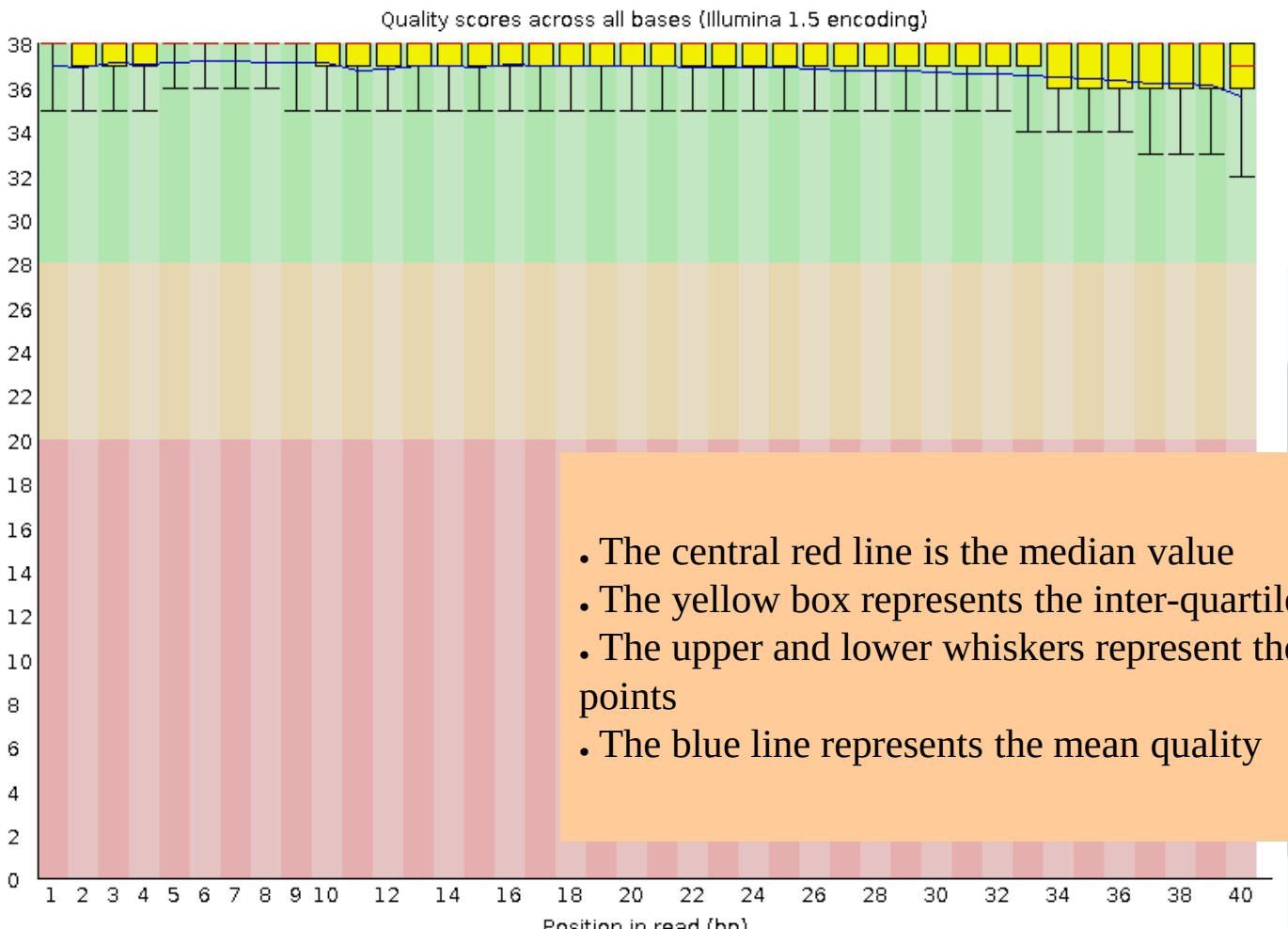
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Basic statistics

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

Sequence quality per base position

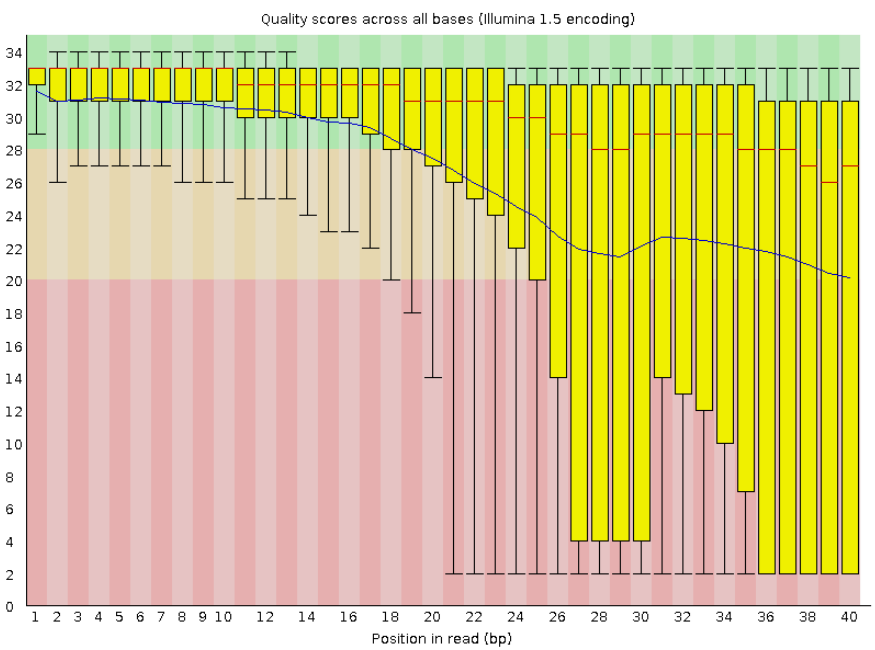


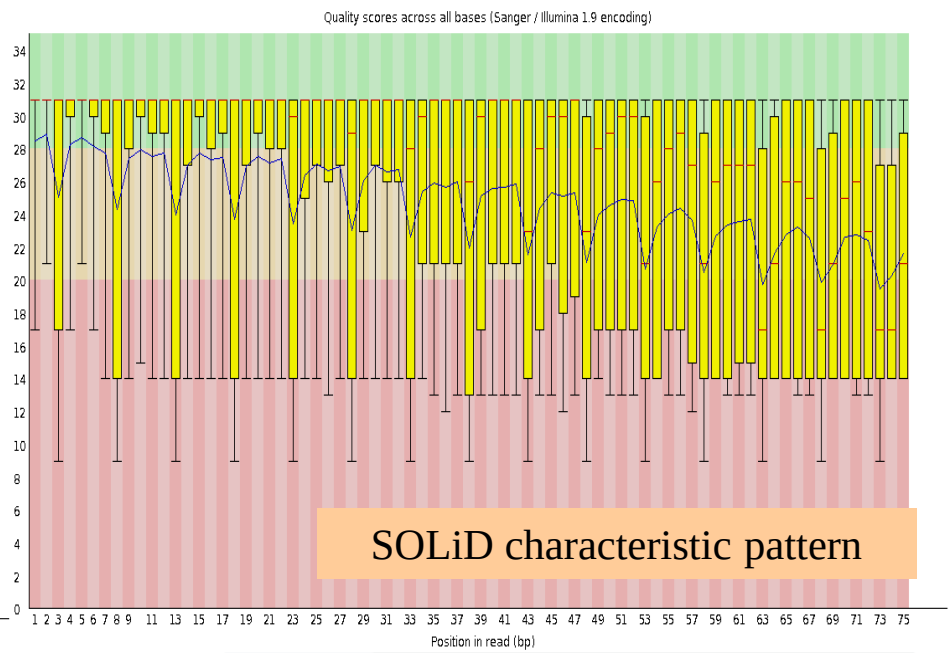
- Good data
- Consistent
- High quality along the read

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

Sequence quality per base position

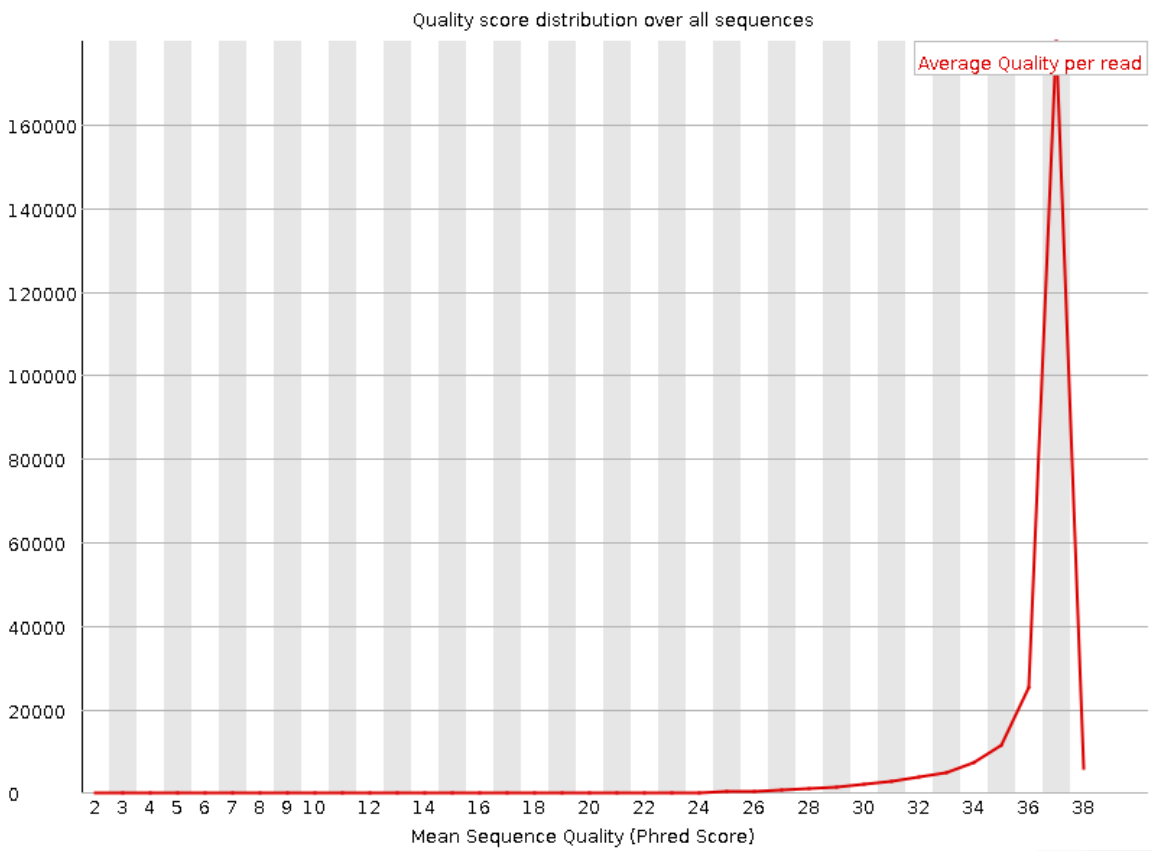
- Bad data
  - High variance
  - Quality decrease with length





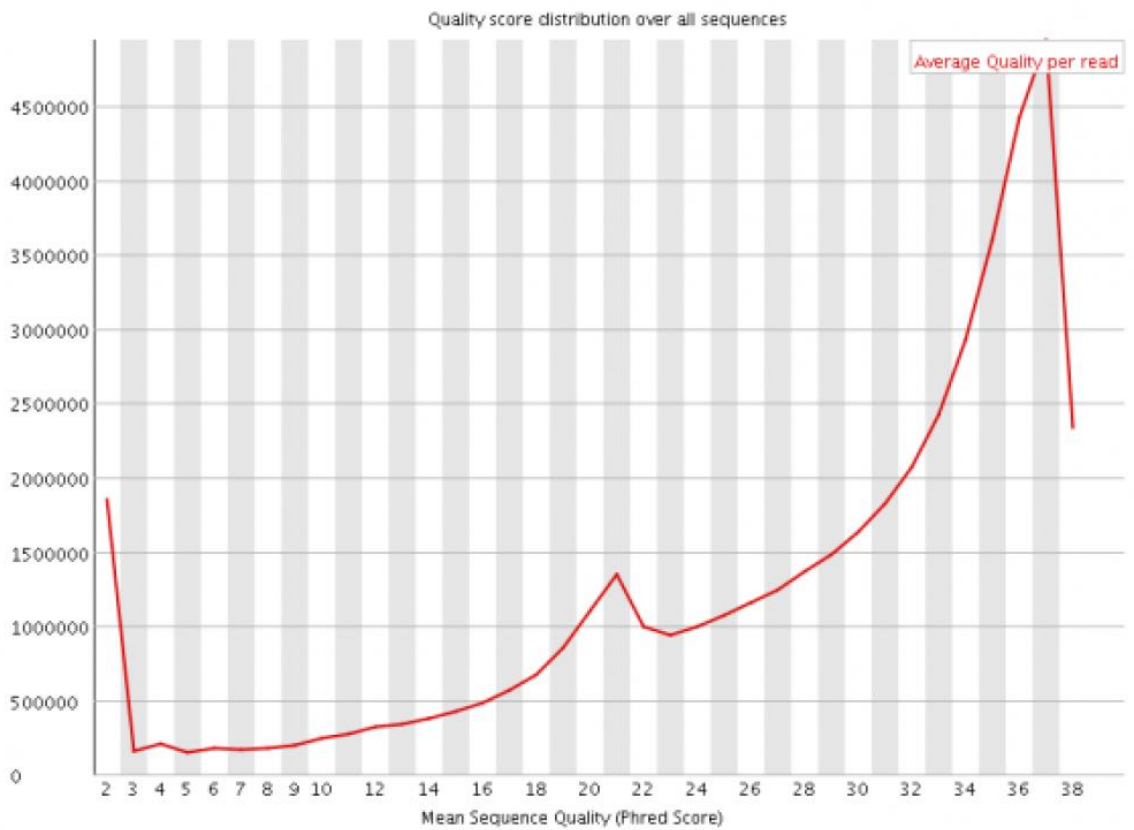
SOLiD characteristic pattern

Per sequence quality distribution



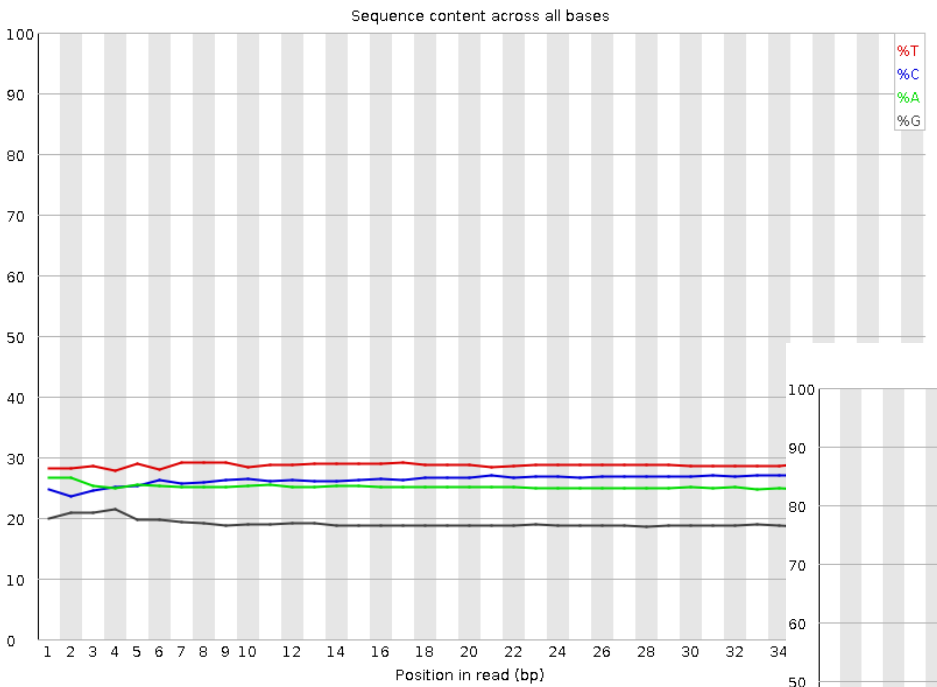
- Good data
- Most are high-quality sequences

Per sequence quality distribution

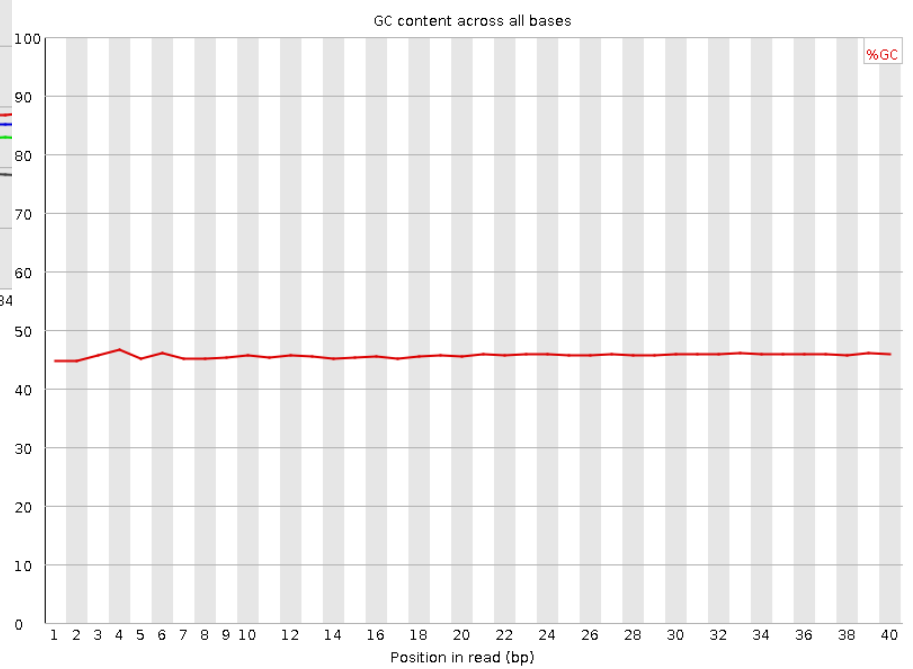


- Bad data
- Non-uniform distribution

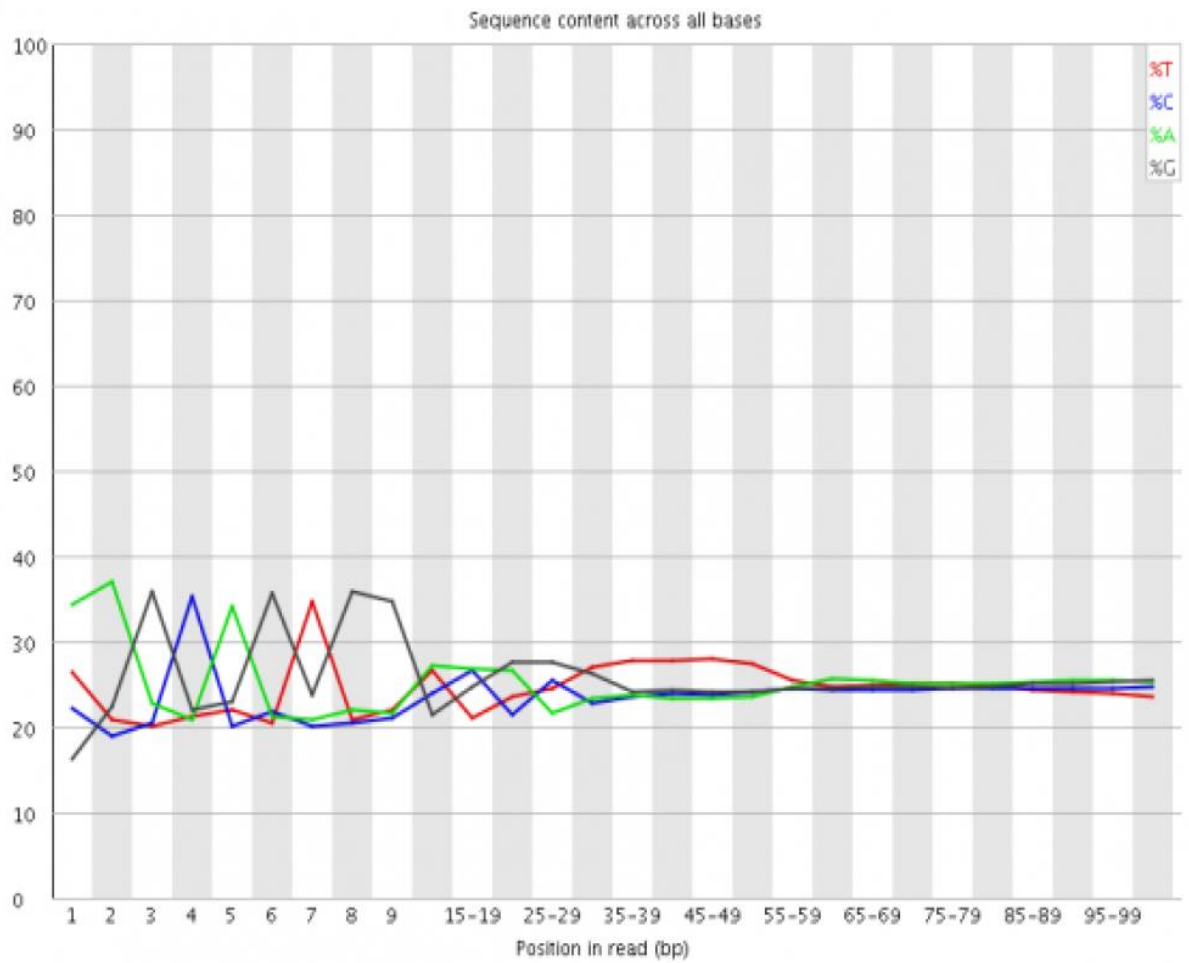
Per base sequence content



- Good data
  - Smooth over length
  - Organism dependent (GC)



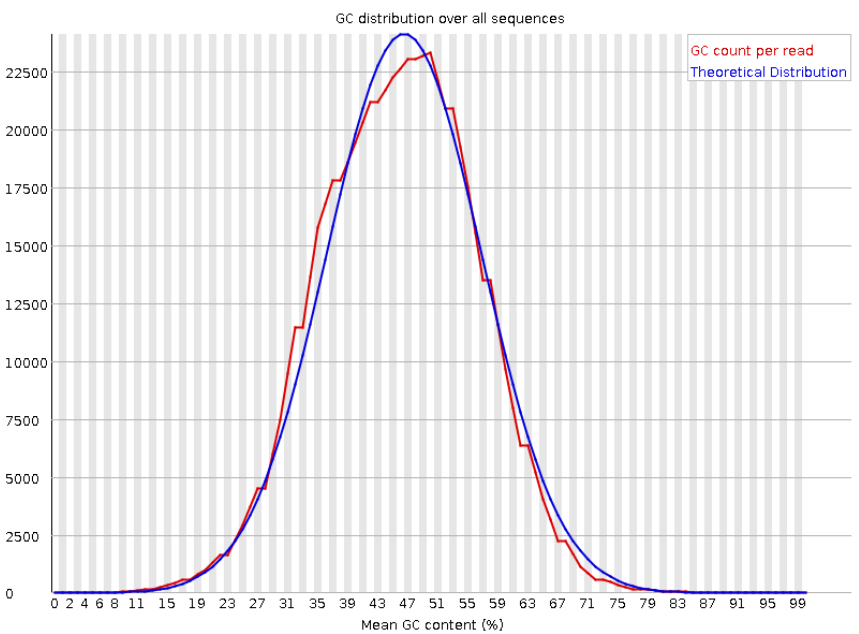
Per base sequence content



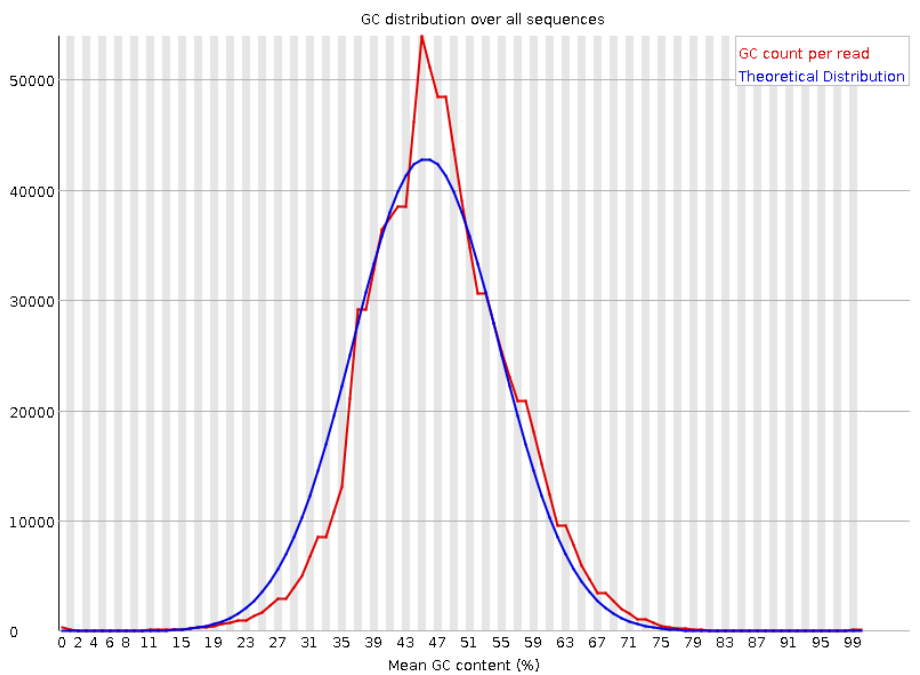
- Bad data
- Sequence position bias

Per sequence GC content

- Good data
  - Fits with expected
  - Organism dependent

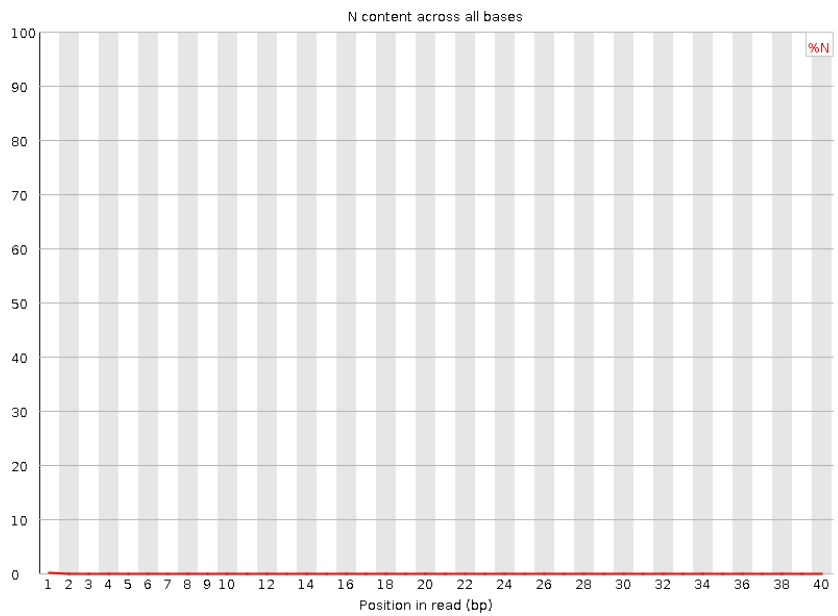


- Bad data
  - Does not fit with expected
  - Library contamination

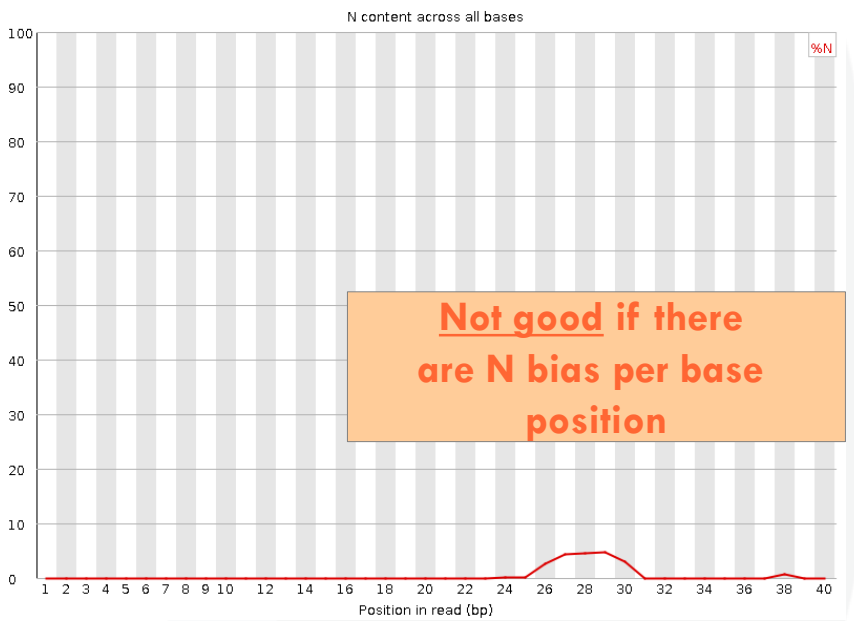


Per base N content

- Good data



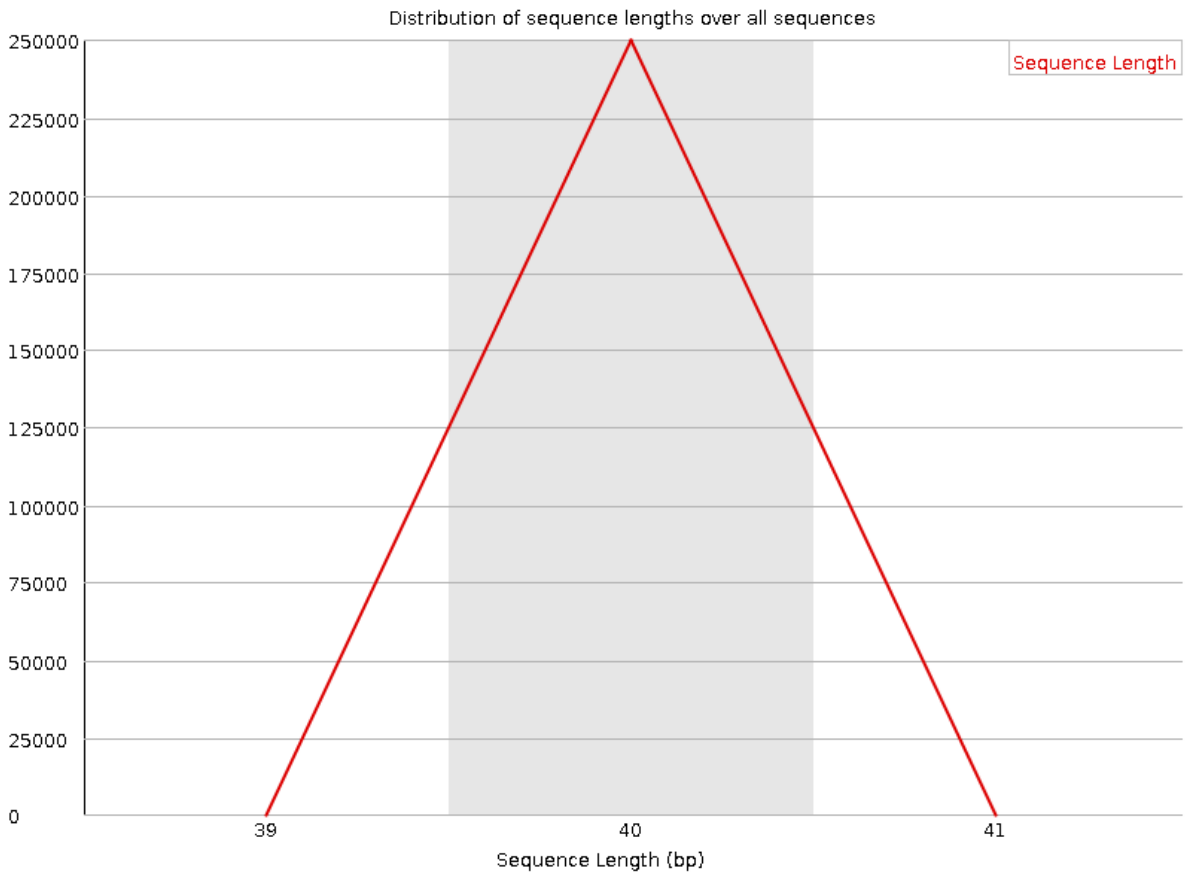
- Bad data



Not good if there are N bias per base position

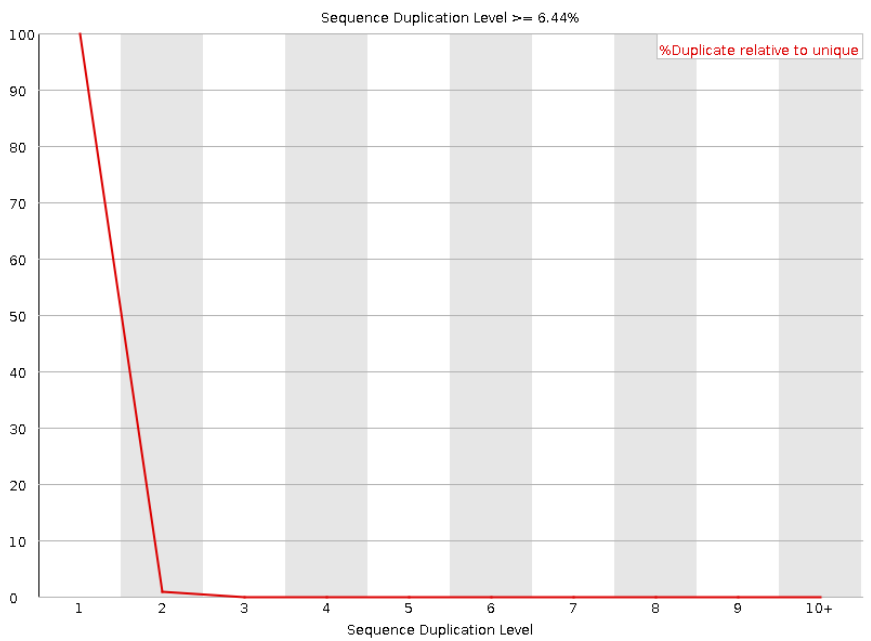
Sequence length distribution

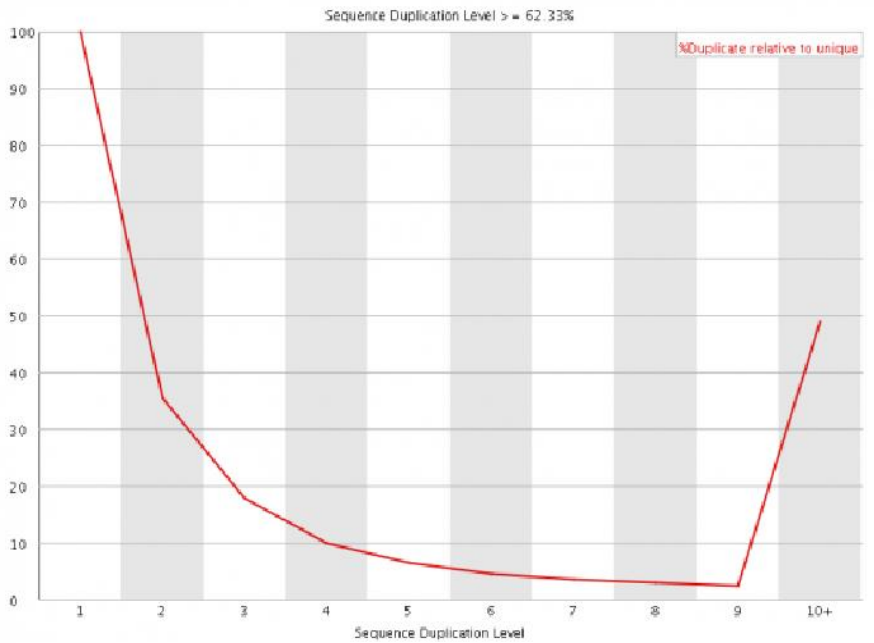
- Just descriptive:
  - Some sequencers output sequences of different length (e.g. 454)



Sequence duplication levels

- In **transcriptomics**, you expect higher number of duplicated sequences.
- In **genomics** you should be worried if this happens → PCR artifact





Overrepresented sequences & Kmer content

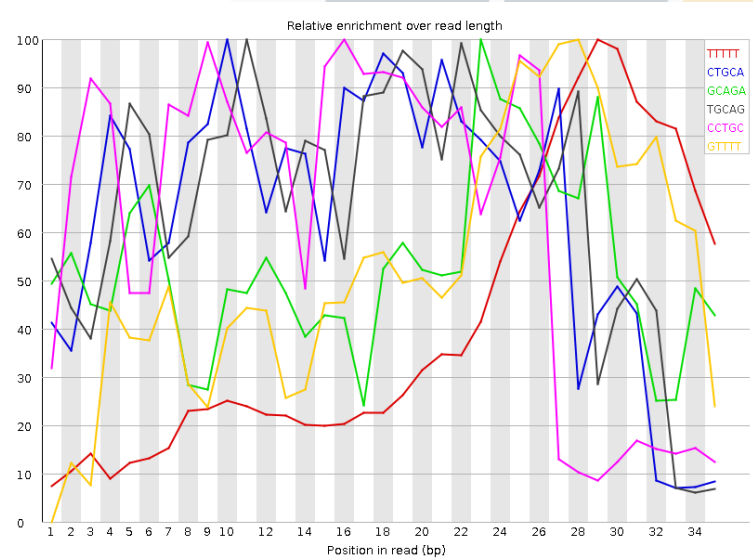
- Question:
  - If we obtain the exact same sequences too many times

→ **Do we have a problem?**

Sequence	Count	Percentage
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754

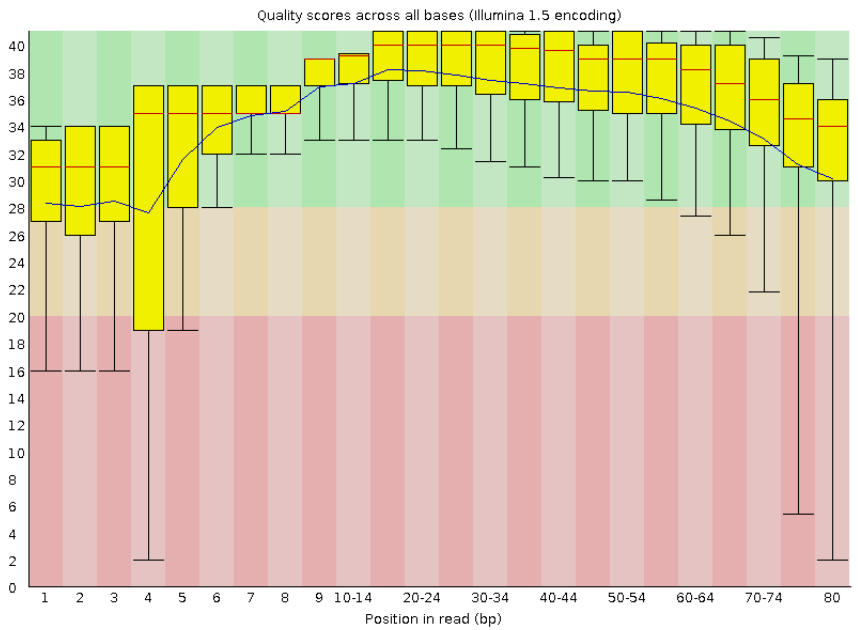
- Answer:
  - **Sometimes !**

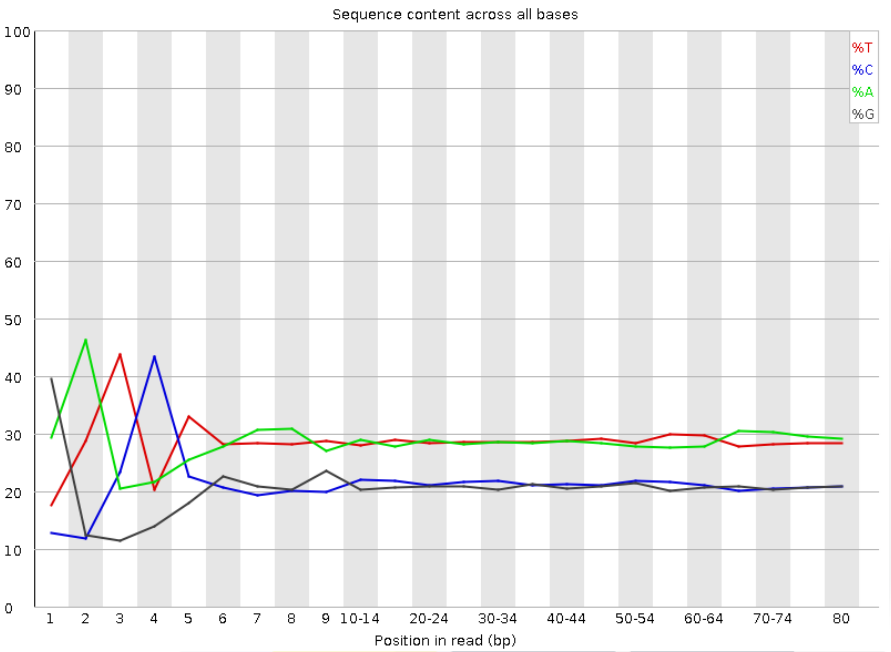
- Examples:
  - PCR primers, adapters …



Typical artefacts

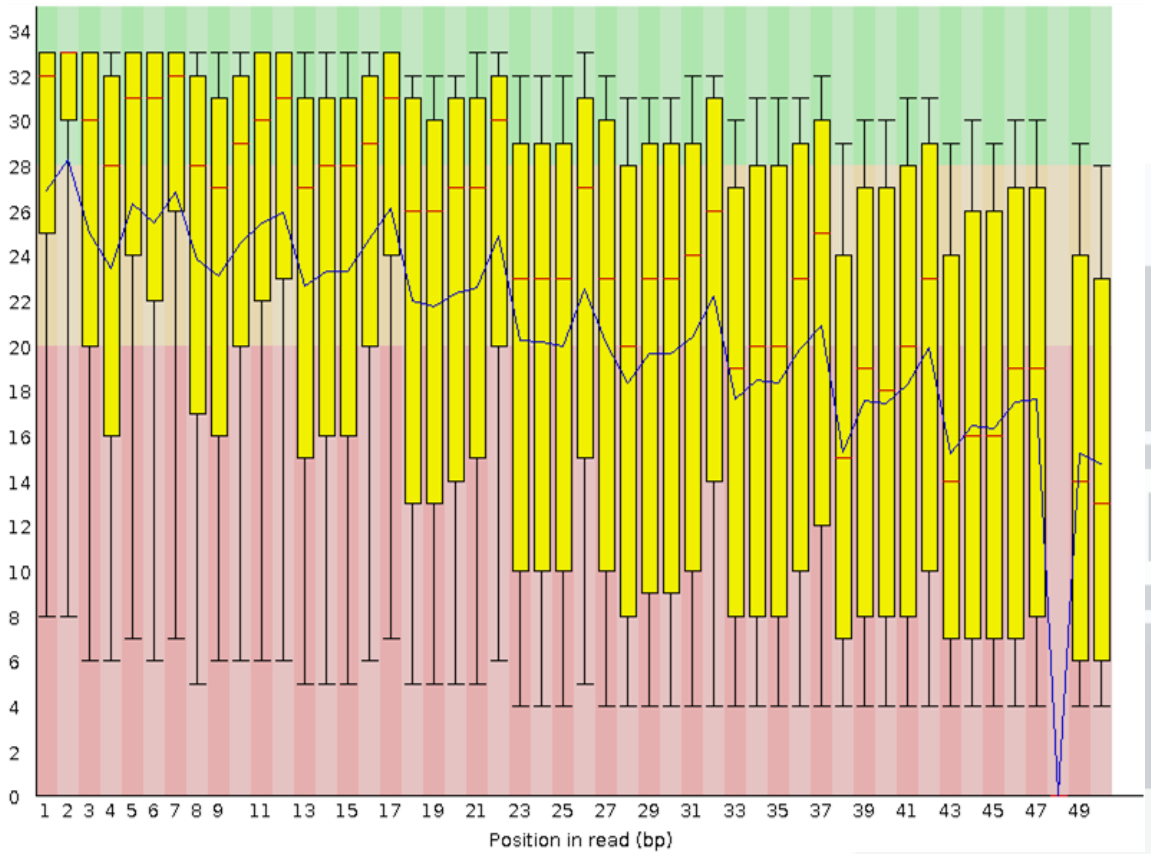
- Sequence adapters



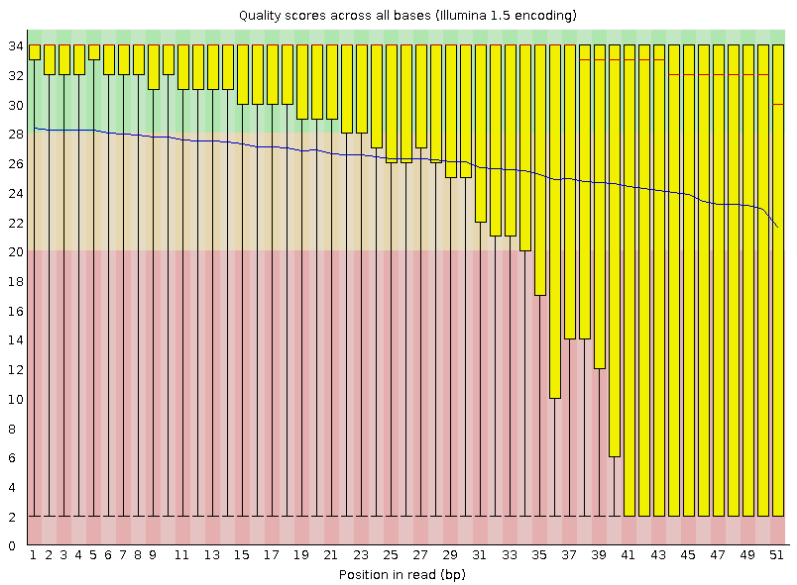


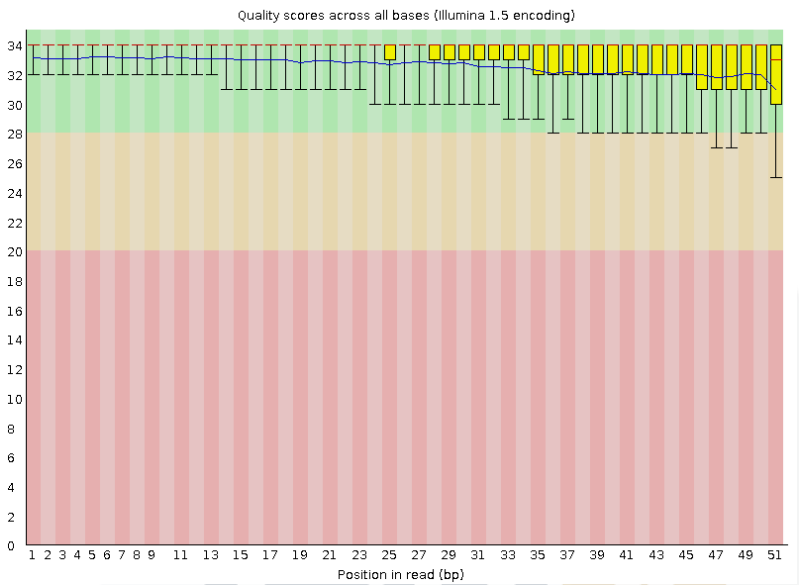
Typical artefacts

- Platform dependent



Filtering & trimming

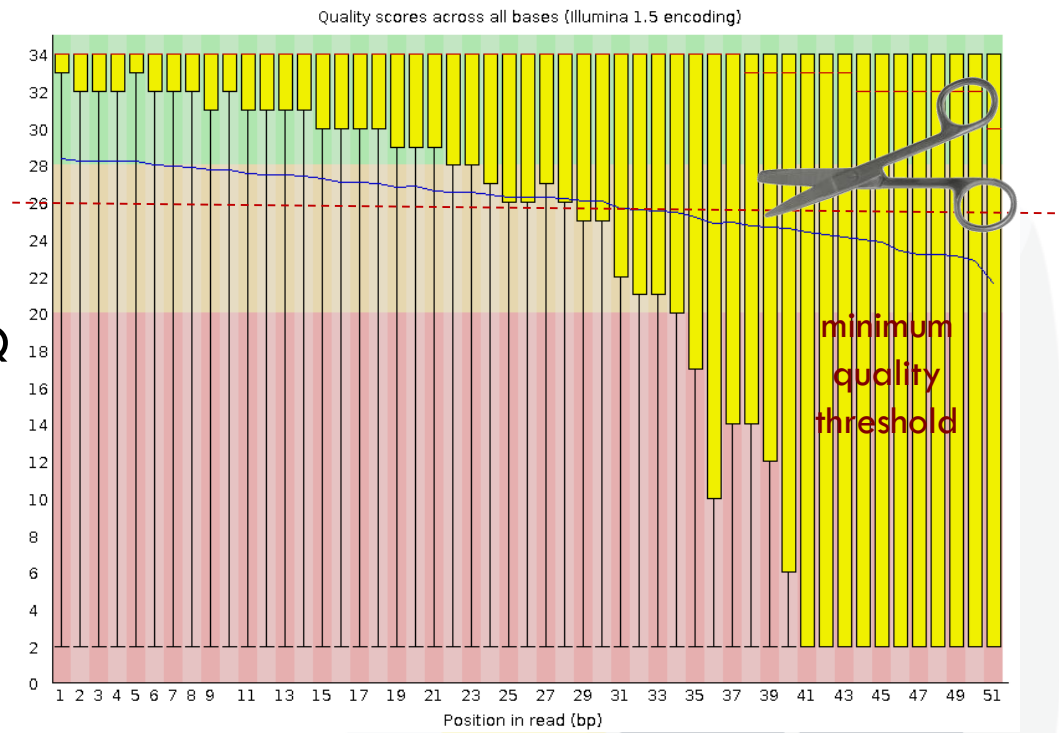




- Removing bad quality data will improve our confidence on downstream analysis

Filtering & trimming

- Sequence filtering
  - Mean quality
  - Read length
  - Read length after trimming
  - Percentage of bases above Q
  - Adapter trimming
  - Adapter reads



Improving sequence quality

- Sequence filtering tools
  - Fastx-toolkit
  - Galaxy (https://main.g2.bx.psu.edu/)
  - SeqTK (https://github.com/lh3/seqtk)
  - Cutadapt (http://code.google.com/p/cutadapt/)
  - And more....

Some practice …

- Download example file and evaluate quality
  - fastqc & + Open file
  - fastqc example_file.fastq
- Cut the last (or the first) bases of the reads
  - cutadapt -l 60 -o example_file_trimmed.fastq example_file.fastq
  - cutadapt -u -40 -o example_file_trimmed.fastq example_file.fastq
  - cutadapt -u 5 -o example_file_trimmed.fastq example_file.fastq
- Quality trimming
  - cutadapt -q 28 -o example_file_quality.fastq example_file.fastq
- Remove adapters
  - cutadapt -a adapter1=ACTG example_file.fastq