



# Quality control of Illumina data

Céline Keime  
keime@igbmc.fr



# Quality control of Illumina data

---

- Primary analysis
- Quality control
- Data pre-processing

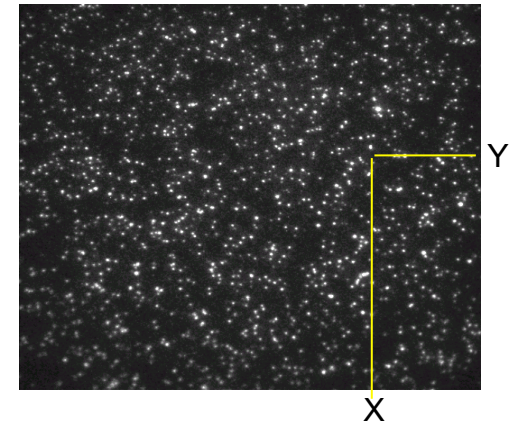
# Quality control of Illumina data

---

- Primary analysis
- Quality control
- Data pre-processing

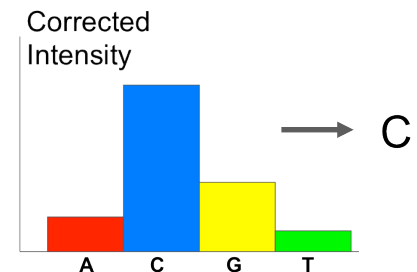
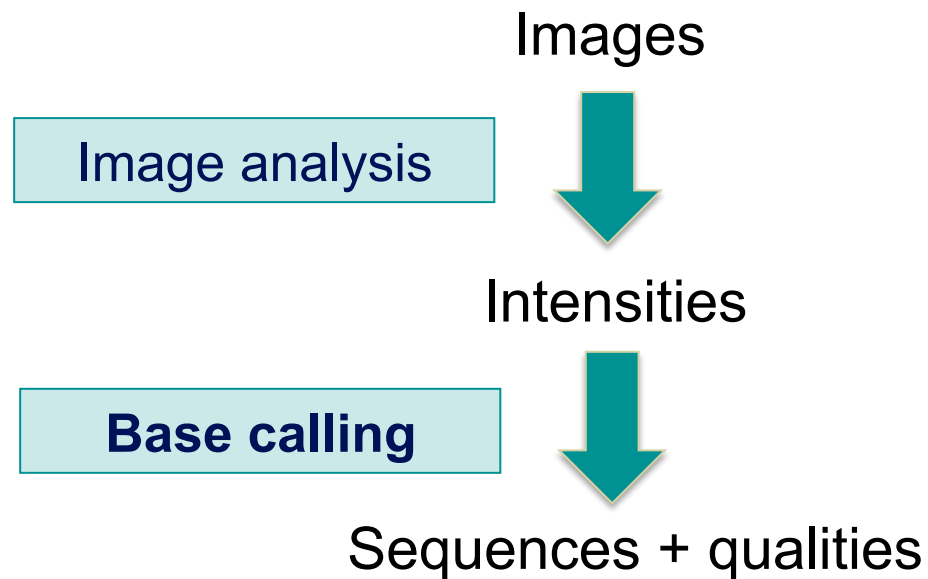
# Primary analysis

---



- Determination of cluster position (only for non-patterned flow cells)
- Extraction of intensities for each cluster

# Primary analysis



- Intensity correction
  - Take into account  $\neq$  intensities per molecule for the 4 bases
- Call the base with the maximum intensity
- Determine “Passing filter” clusters
  - Remove clusters that have “too much” intensity corresponding to bases other than the called base

# Phred quality scores

---

- Prediction of the probability of error in base calling

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

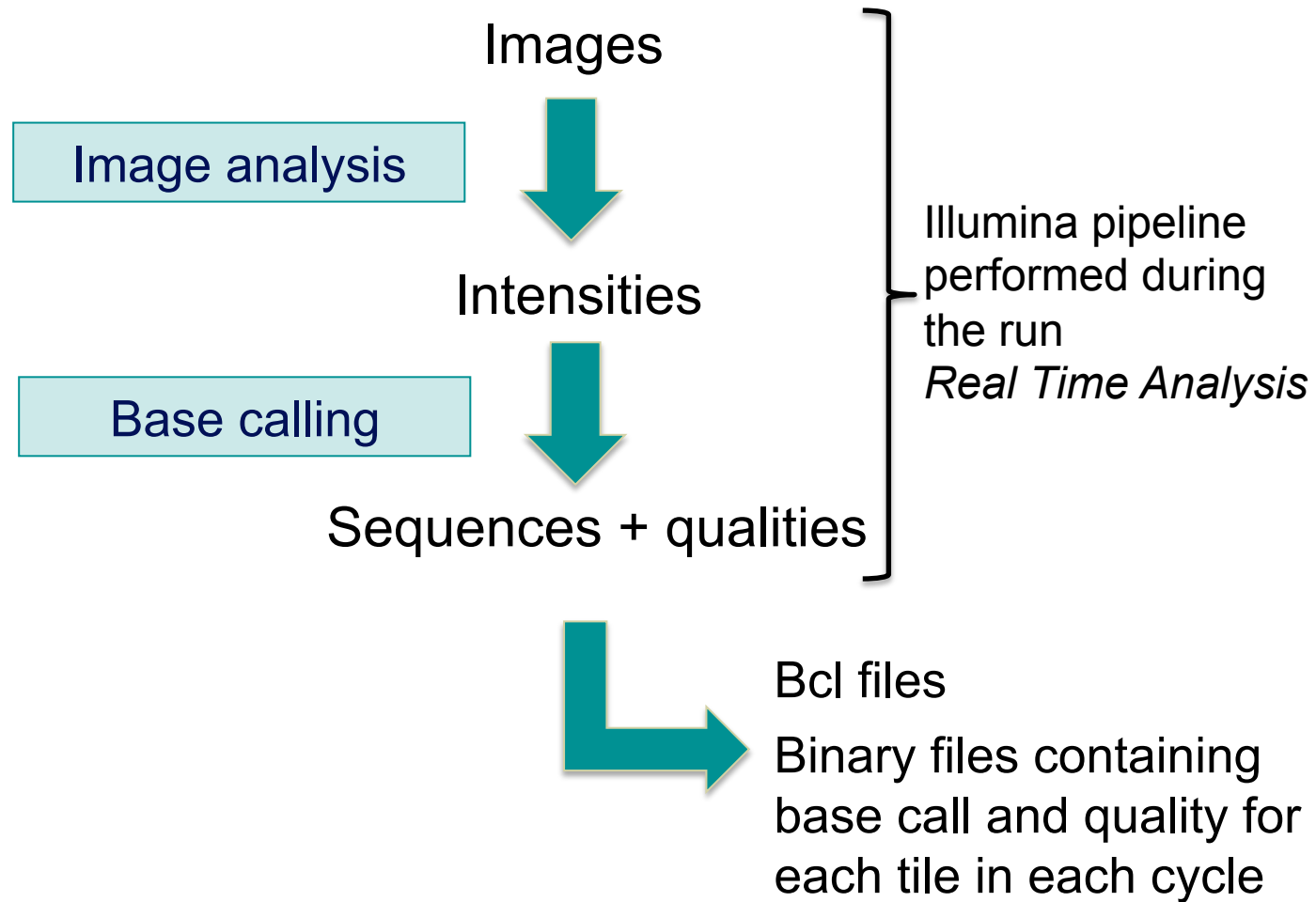
$Q$  : *quality score*

$P$  : *error probability*

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

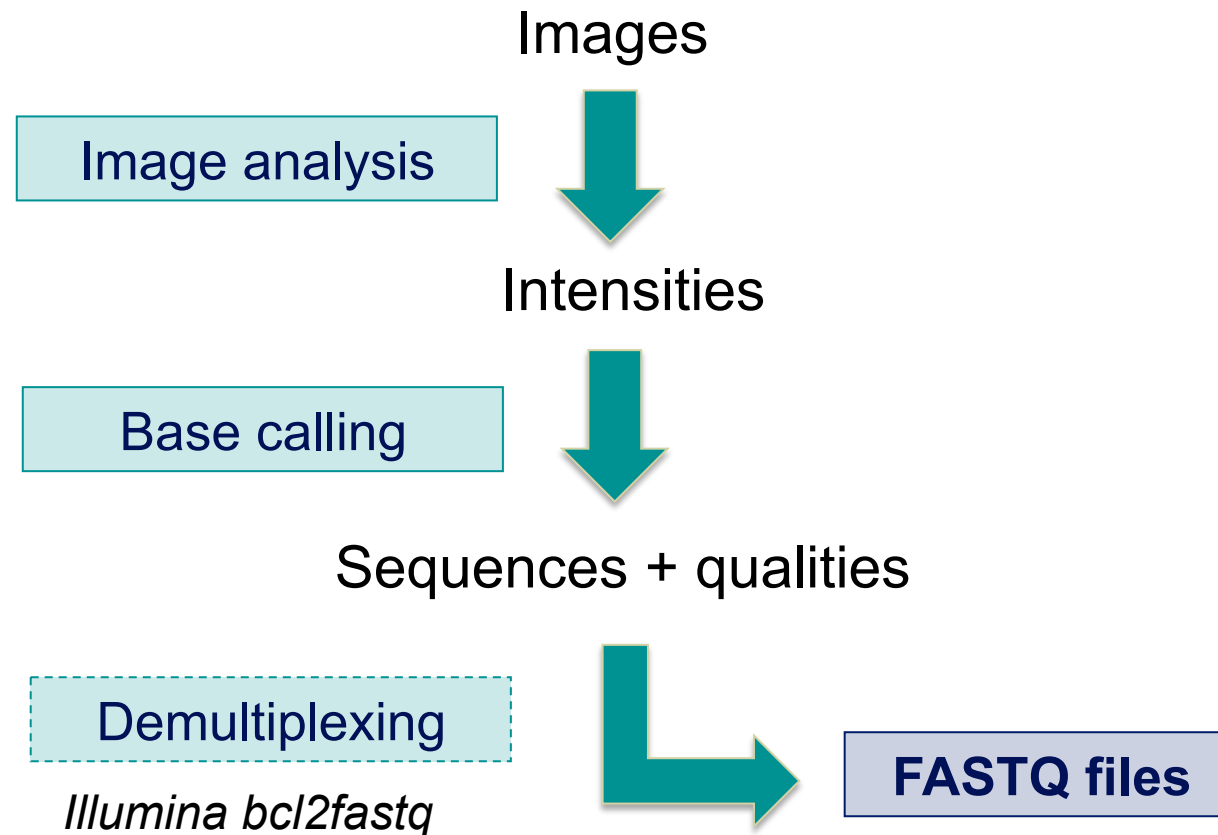
# Primary analysis

---



# Primary analysis

---





# FASTQ file

- Text file containing

- Sequences

- Qualities

Probability that the corresponding base call is incorrect

## 4 lines per sequence :

```
@HWI-ST1136:97:HS041:7:1101:1681:2104 1:N:0:ACAGTG → 1. @Identifier
CTTTTTTATTGAATTCTATGATTCTTGTTAGATTCATAATGGCTGCTTA → 2. Sequence
+ → 3.+ optionally followed by same identifier as 1.
@@@DBDDDDFF8:D?EBAEAH,CF:AF9F+2**9?B?1C<<?9*8D?)9*? → 4. Quality
@HWI-ST1136:97:HS041:7:1101:1521:2119 1:N:0:ACAGTG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
@@@?DDDDFAFHIIHFDAB@B6@B@BBBBBBBBBBBBBB359BBBB8BBB
@HWI-ST1136:97:HS041:7:1101:1669:2145 1:N:0:ACAGTG
CTGCTGTTTTCAAATGTCCGATGTGTGCTATGACTGACAACACTTTTC
+
@@<1?DDDFDBDFE>+<CCF>FAG++2+<<F**?:?1:C?:8B:9BBBD4
```

(Cock et al. NAR 2009; 38(1): 1767-1771)

# Beginning of siLuc3\_S12040.fastq file

The screenshot displays the Galaxy web interface. At the top, the navigation bar includes 'Galaxy / Galaxeast', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 43%' indicator is visible in the top right corner. On the left, a 'Tools' sidebar lists various categories such as 'Get Data', 'Send Data', 'Text Manipulation', 'Convert Formats', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Statistics', and 'Graph/Display Data'. The main content area features a yellow warning box stating 'This dataset is large and only the first megabyte is shown below.' Below this, the beginning of a FASTQ file is displayed, showing header lines (e.g., '@HWI-ST1136:52:HS008:4:1101:2560:2035') and sequence lines. On the right, a 'History' panel shows a search for 'RNA-seq data analysis' with one result shown: '1: siLuc3\_S12040.fastq'. This entry is highlighted in green, and a red circle highlights the eye icon next to it, indicating that the dataset is visible.

Galaxy / Galaxeast Analyze Data Workflow Shared Data Visualization Help User Using 43%

Tools

search tools

Get Data  
Send Data  
Text Manipulation  
Convert Formats  
Filter and Sort  
Join, Subtract and Group  
Extract Features  
Fetch Sequences  
Statistics  
Graph/Display Data

NGS TOOLBOX BETA  
NGS: QC and manipulation  
NGS: SAM Tools

This dataset is large and only the first megabyte is shown below.  
[Show all](#) | [Save](#)

```
@HWI-ST1136:52:HS008:4:1101:2560:2035 1:N:0:GCGAAT
GCCGGTGGGGTCGATGCCATGTTTCATCACTGATCAACTCCCAGAACTTGG
+
?@;BBD<?)@<@@:):1?GFD?:?GF<9*9BG9B99?*0?CCBBF9@F
@HWI-ST1136:52:HS008:4:1101:2669:2093 1:N:0:GCCAAT
GCTGTTTGCTTTTGTCTCCCTCTGTCTTAGGAAAAGCCATCTTTAATAT
+
??7DD;=D+?CDD<EEEIIEEECFFFCFD<F<AEEE@DIDIIIIIIEIAD
@HWI-ST1136:52:HS008:4:1101:2690:2156 1:N:0:GCCAAT
TTTGCAATTACGCCTGTAATGTATTCATTCTTAATTTATGTAAGGTTT
+
???DDDDHFDHF<FHIGEHIII9?HBFFF<CHH@FFHCGHIGDIICDGH
@HWI-ST1136:52:HS008:4:1101:2663:2212 1:N:0:GCCAAT
CAAATAGACTACATAATATACGTGGGCAAAAAGGCAATTAAGTGAATCTC
+
?8?DD?A:CCCFF??ECFH@,CAFHFGGIIHIGCGGE??<FDRGGEGGIE
```

History

search datasets

RNA-seq data analysis  
1 shown  
7.23 GB

1: siLuc3\_S12040.fastq

# Sequence identifier in FASTQ files

- Begins with @ followed by sequence ID and an optional description
- Illumina sequence identifiers :

Instrument Name    Run number    Flowcell ID    Lane    Tile    X\_pos    Y\_pos    Read    Is Filtered    Control Number    Index Sequence

@HWI-ST1136:97:HS041:7:1101:1681:2104 1:N:0:ACAGTG

- Read :  
The member of a pair = 1 or 2 (for paired-end or mate-pair reads)
- Is filtered  
Y if the read is bad (the cluster do not pass filter), N otherwise  
Recent versions of Illumina pipeline only supply passing filter reads

# Quality in FASTQ files

- Phred quality score (Sanger format)
- Encoded in ASCII characters to save space
- 1 ASCII symbol = 1 quality value
- Phred quality scores from 0 to 93 are encoded using ASCII 33 to 126 :

032 sp	048 ò	064 @	080 P	096 `	112 p
033 !	049 1	065 A	081 Q	097 a	113 q
034 "	050 2	066 B	082 R	098 b	114 r
035 #	051 3	067 C	083 S	099 c	115 s
036 \$	052 4	068 D	084 T	100 d	116 t
037 %	053 5	069 E	085 U	101 e	117 u
038 &	054 6	070 F	086 V	102 f	118 v
039 '	055 7	071 G	087 W	103 g	119 w
040 (	056 8	072 H	088 X	104 h	120 x
041 )	057 9	073 I	089 Y	105 i	121 y
042 *	058 :	074 J	090 Z	106 j	122 z
043 +	059 ;	075 K	091 [	107 k	123 {
044 ,	060 <	076 L	092 \	108 l	124
045 -	061 =	077 M	093 ]	109 m	125 }
046 .	062 >	078 N	094 ^	110 n	126 ~
047 /	063 ?	079 O	095 _	111 o	127 ò

- Binned in order to save space in the last version of Illumina software, e.g.
  - $2 < \text{real Q-score} < 9 \rightarrow \text{binned Q-score} = 6$
  - $10 < \text{real Q-score} < 19 \rightarrow \text{binned Q-score} = 15$
  - ...
  - $\text{real Q-score} \geq 40 \rightarrow \text{binned Q-score} = 40$

# Paired-end FASTQ files

- 2 FASTQ files per sample



XXXX.R1.fastq.gz

XXXX.R2.fastq.gz

```
@HWI-ST1136:163:HS087:7:2310:17264:70630 1 N:0:ATCACG
GTTAGAGCCAAGGTACAGTGGCCTGTCTTTGTAAATGTGCC TTTATGT
+
CCCCFFFFHHHHHJFHIIJHIJIIJIIJJJJIIHHIJIIGIJJJJJJII
@HWI-ST1136:163:HS087:7:2310:17415:70636 1:N:0:ATCACG
TGGAGCCTTGGTAACTTTTTGTAGTTTGTATGCGTTTTTGTGGTCTC
+
BCCFFFFHHHHHJJJJJJJJHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@HWI-ST1136:163:HS087:7:2310:17337:70637 1:N:0:ATCACG
CTGTTACCCCTCCATTGAGGGTATGAAGAAGGGCTTACCTGTAGTTC
+
@CCFFFFHHHHHJJJJJJJJBFHIIJIIJJIIIIJJIIJJHGHIIJJ
```

```
@HWI-ST1136:163:HS087:7:2310:17264:70630 2 N:0:ATCACG
TAATTTTTGCATCCTGAAAAGTGTGGAAGTTGGGTTTTTCATAGTCAA
+
CCCCFFFFHHHHHJJJJJJJJICHGIIJJIIJJIIJFHIIJHIJJJJJJ
@HWI-ST1136:163:HS087:7:2310:17415:70636 2:N:0:ATCACG
TGTTCATATGTATGAGATAGATTTGAAAAATCTACTAATTTTTAAAATC
+
CCCCFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@HWI-ST1136:163:HS087:7:2310:17337:70637 2:N:0:ATCACG
TCCTGACATCAAGCACACTGCTTCTGCATCTATGTGGCACCTAAAACAA
+
CCCCFFFFHHHHHJJIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

# Quality control of Illumina data

---

- Primary analysis
- Quality control
- Data pre-processing

# Quality control

---

## ■ Why ?

- Are the data consistent to what is expected ?
- Are the data suited to answer my biological questions ?  
With what limitations ?
- Identify any problems of which you should be aware before doing any further analysis

## ■ What to look for ?

- Number of reads
- Base qualities and N calls
- Base composition relative to reference genome
- Sequence duplication
- Presence of adapters
- Contaminations


# Quality control tools


---

- FastQC
  - <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>
- SolexaQA
  - <http://solexaqa.sourceforge.net/>
- NGS QC Toolkit
  - <http://www.nipgr.res.in/ngsqctoolkit.html>
- Picard
  - <http://picard.sourceforge.net/>
- RSeQC – quality controls specific to RNAseq data
  - <http://rseqc.sourceforge.net/>
- FastQ Screen – to verify the composition of a library and search for possible contaminations
  - [http://www.bioinformatics.babraham.ac.uk/projects/fastq\\_screen/](http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/)
- ...



# FastQ Screen on GalaxEast

Tools 

**fastq screen** 

NGS: QC and manipulation

[fastq screen](#) Screen for contamination

**fastq\_screen** Screen for contamination (Galaxy Version 0.4.2) Options

**Job narrative (included in output names as a reminder)**

fastq\_screen

Only letters, numbers and underscores \_ will be retained in this field

**Sample this number of reads. Set to 0 or less to use all**




500000

Time/precision trade off - fewer reads takes a little less time trading off precision of the estimates.

**Single ended or mate-pair ended reads in this library?**


Single-end

**RNA-Seq FASTQ file**

   1: siLuc3\_S12040.fastq


Nucleotide-space: Must have Sanger-sealed quality values with ASCII offset 33

**Installed organism reference sequences to check for alignment to your fastq**

1: Installed organism reference sequences to check for alignment to your fastq 


**Bowtie2 reference genome**

hg38

2: Installed organism reference sequences to check for alignment to your fastq 


**Bowtie2 reference genome**

mm10

3: Installed organism reference sequences to check for alignment to your fastq 

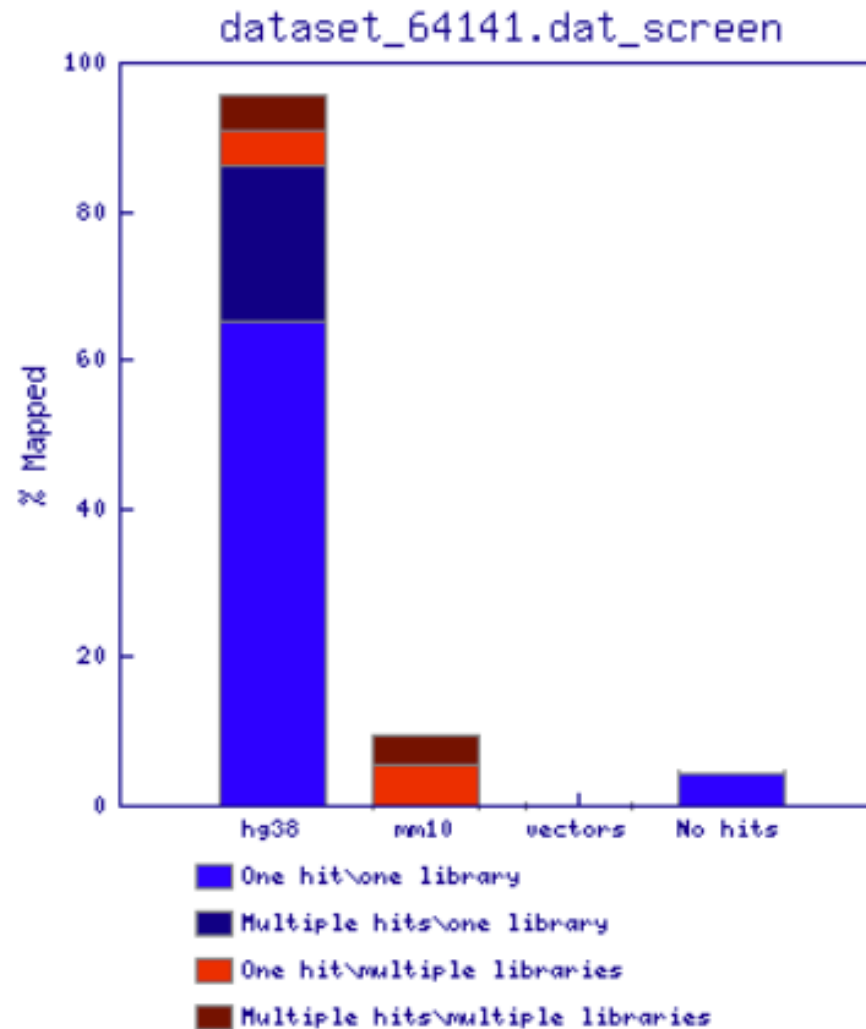
**Bowtie2 reference genome**

vectors

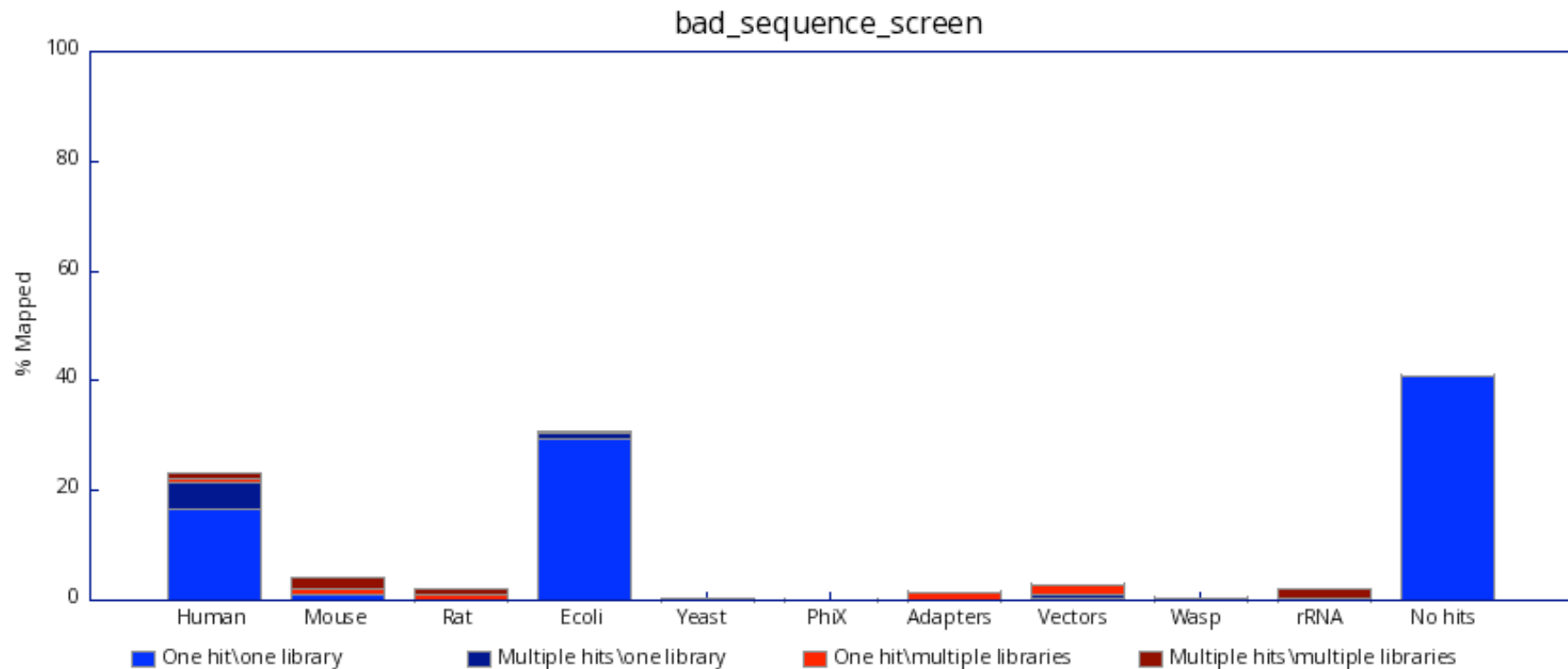
 Insert Installed organism reference sequences to check for alignment to your fastq

For checking cell culture sequence for contamination, Mycoplasma Genitalium might be a good choice eg

# FastQ Screen result on siLuc3\_S12040.fastq



# FastQ Screen result on a bad sample



# FastQC

---

- Allows quality control of NGS data
  - FASTQ, gzip compressed FASTQ (base or colorspace)
  - SAM, BAM alignment files
- Can be used *via* a graphical interface, in command-line or in Galaxy
- Generates graphs and tables with several quality control analyses
  - ➔ Allows a global quality assessment of NGS data and rapid identification of possible problems

# Exercise : quality analysis

---

- Analyse the quality of siLuc3\_S12040.fastq file
  - How many reads have been sequenced in this sample ?
  - What do you think about the quality of this sample ?
  - Do you identify bias in these data ?

# FastQC results

## FastQC Report

mar. 3 janv. 2017  
siLuc3\_S12040.fastq

History ↻ ⚙ ☰

search datasets ✕

RNA-seq data analysis  
3 shown  
7.23 GB ☑ 🗨 💬

3: FastQC on data 1:  
RawData 👁 ✎ ✕

2: FastQC on data 1:  
Webpage 👁 ✎ ✕

View data

1: siLuc3\_S12040.fastq 👁 ✎ ✕

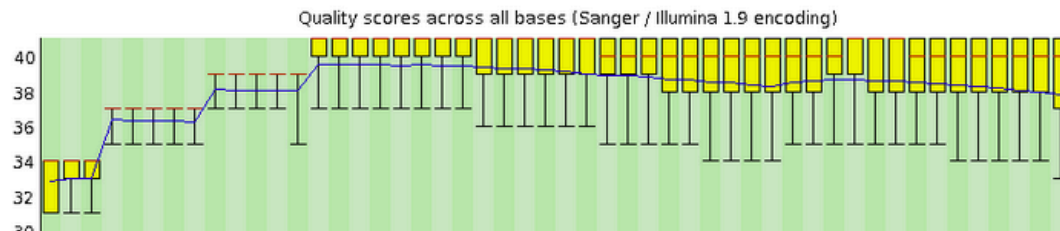
### Summary

- ✔ Basic Statistics
- ✔ Per base sequence quality
- ✔ Per tile sequence quality
- ✔ Per sequence quality scores
- ✘ Per base sequence content
- ✔ Per sequence GC content
- ✔ Per base N content
- ✔ Sequence Length Distribution
- ✘ Sequence Duplication Levels
- ✔ Overrepresented sequences
- ✔ Adapter Content
- ✘ Kmer Content

### ✔ Basic Statistics

Measure	Value
Filename	siLuc3_S12040.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50079515
Sequences flagged as poor quality	0
Sequence length	50
%GC	49

### ✔ Per base sequence quality

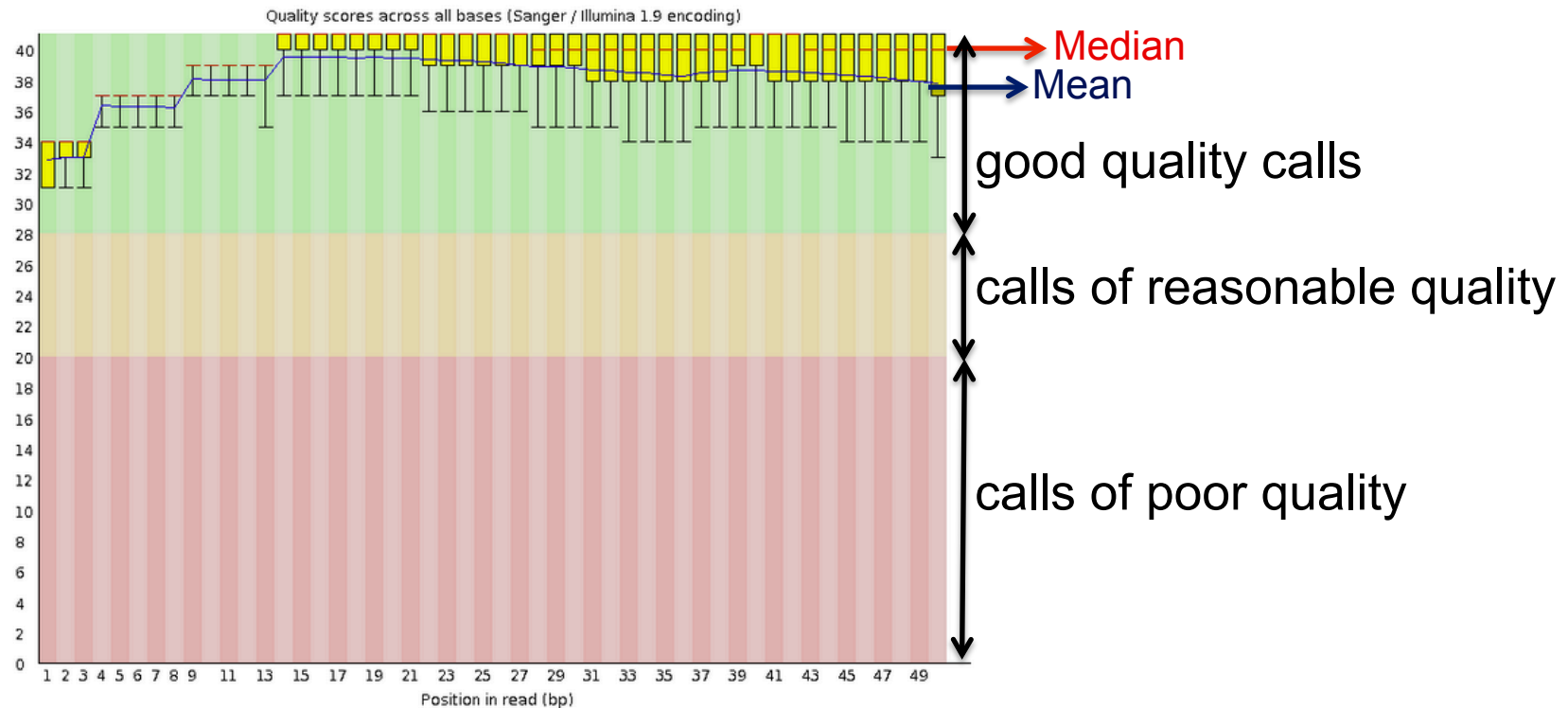


# Basic Statistics

Measure	Value
Filename	siLuc3_S12040.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50079515
Sequences flagged as poor quality	0
Sequence length	50
%GC	49

- **File type** : Base calls or colorspace data
- **Encoding** : Which ASCII encoding of quality values was found in this file
- **Total Sequences**: A count of the total number of sequences in the file
- **Filtered Sequences** : Sequences flagged to be filtered will be removed from all analyses  
The number of such sequences removed will be reported here  
The total sequences count above will not include these filtered sequences
- **Sequence length**: Length of the shortest and longest sequence  
If all sequences have the same length only one value is reported
- **%GC**: The overall %GC of all bases in all sequences

# Per base sequence quality

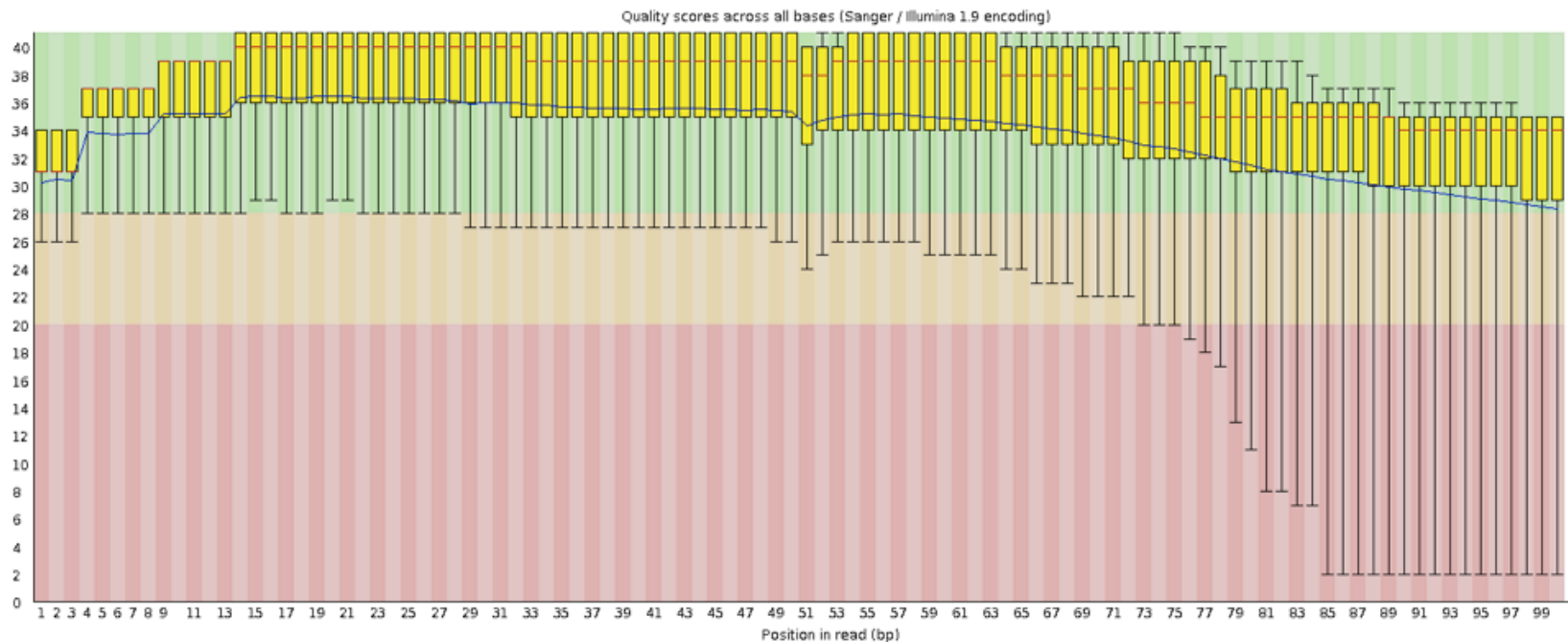


- Yellow boxes : inter-quartile range (25-75%)
- Upper and lower whiskers : 10% and 90%
- ➔ **Sample of good quality**



# Per base sequence quality on another sample

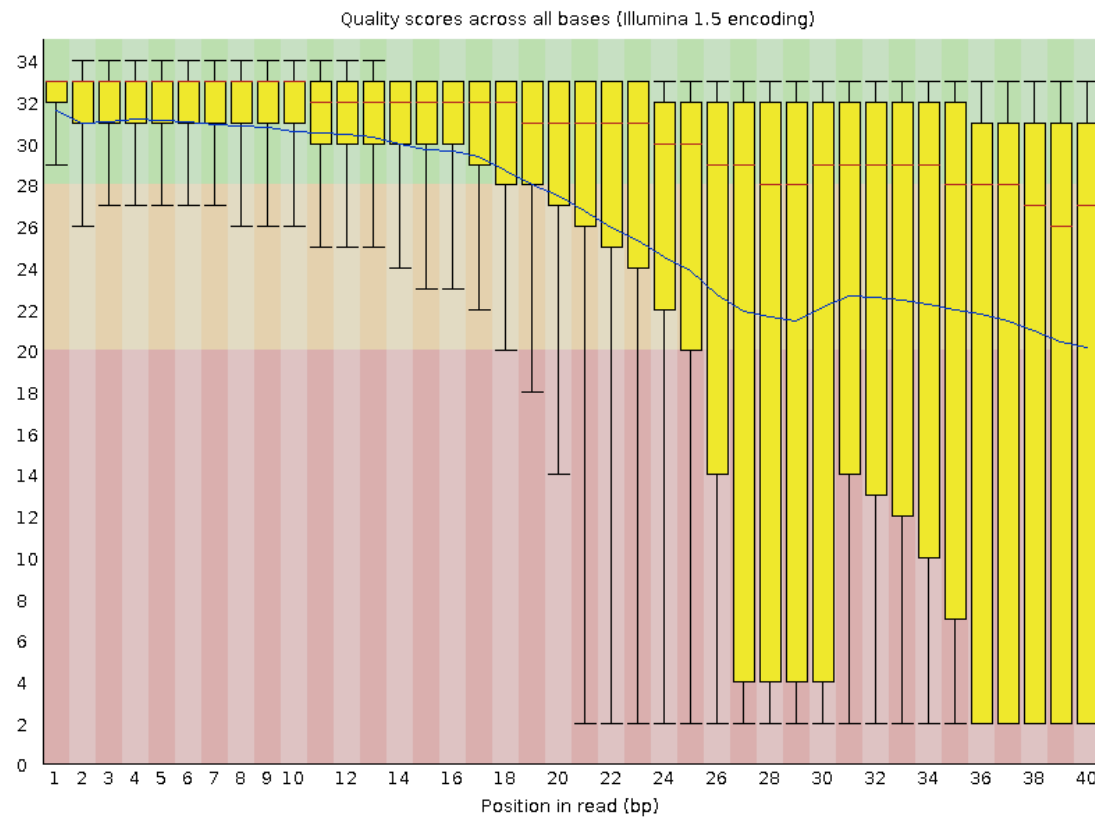
- The quality of calls decreases as the run progresses
  - ➔ common to see base calls decreasing towards the end of a read
- e.g. 2<sup>nd</sup> read of a 2x100 run :



➔ In such cases reads can be trimmed

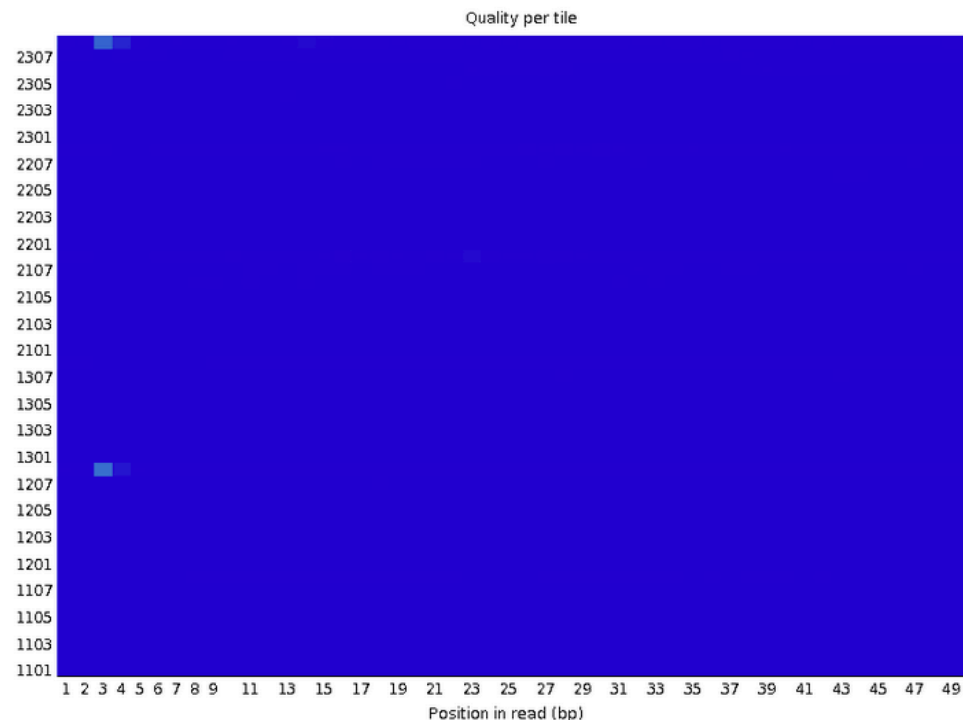
# Per base sequence quality on another sample

- Example of a bad quality sample



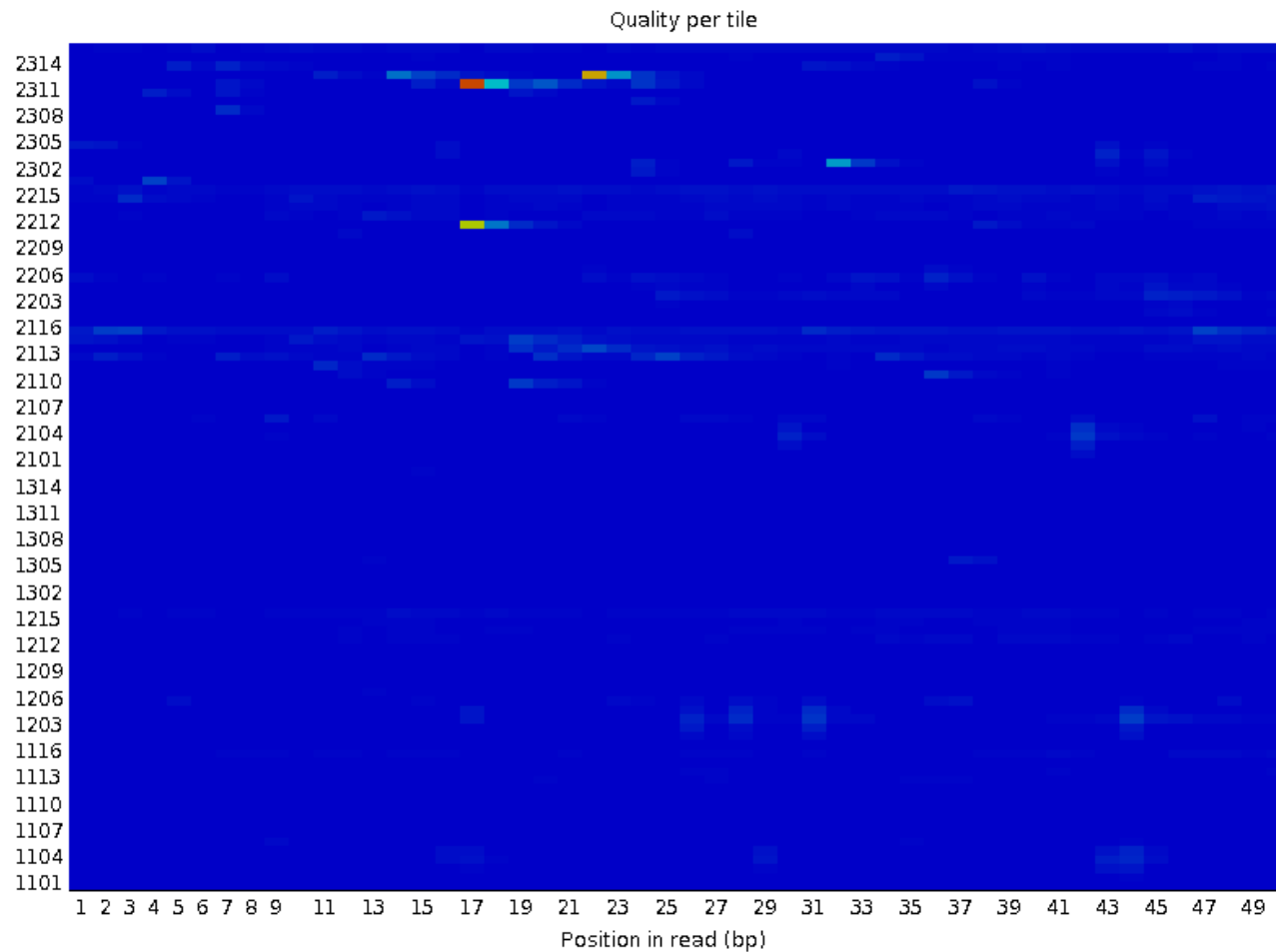
# Per tile sequence quality

Quality scores from each tile across all bases :  
show the deviation from the average quality for each tile

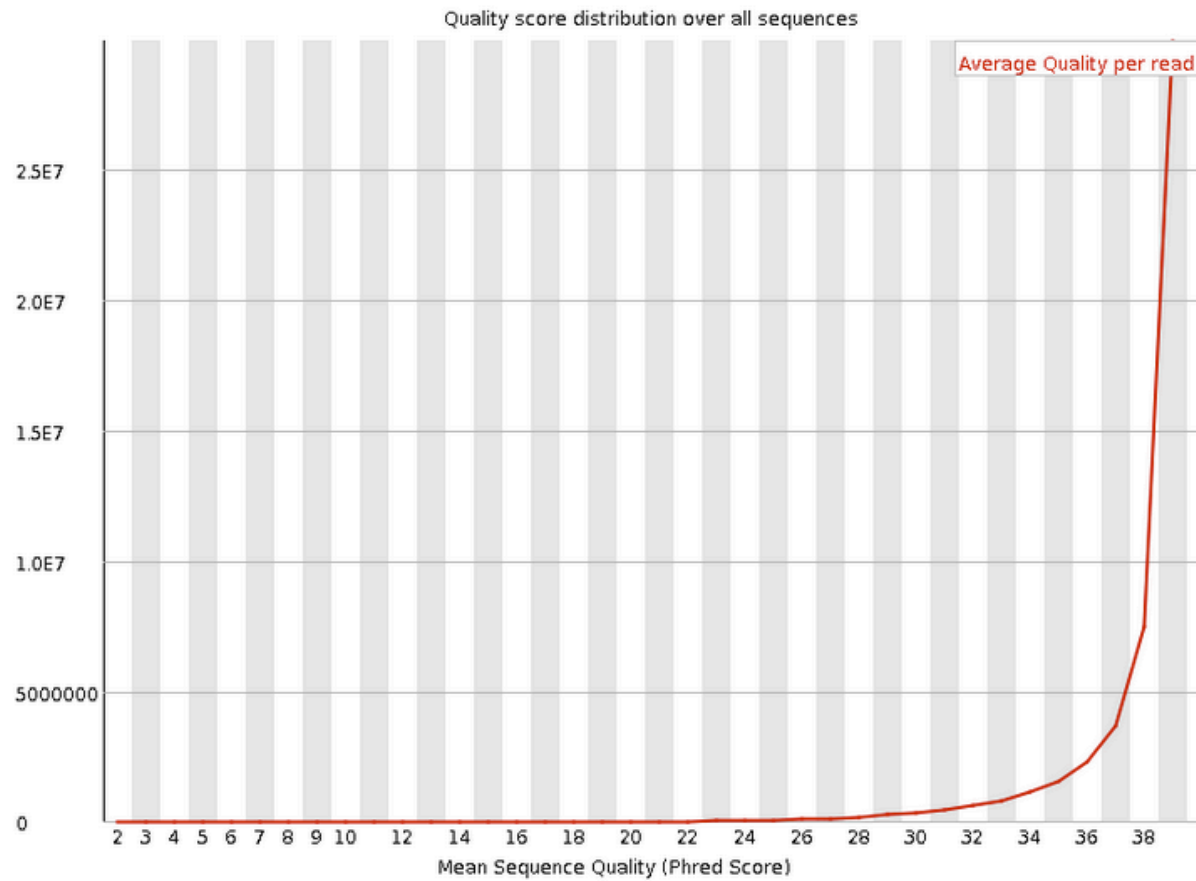


- To see if there was a loss in quality associated with only one part of the flowcell
- No poor quality tile for this sample

# Per tile sequence quality on another sample



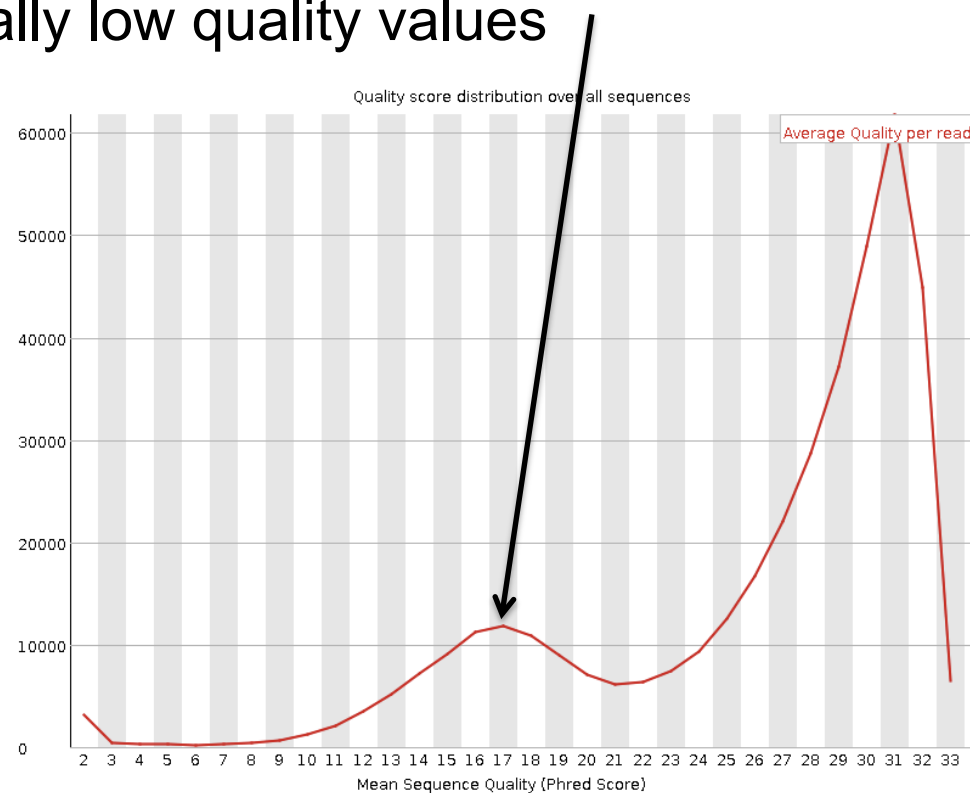
# Per sequence quality scores



→ Good quality of all sequences

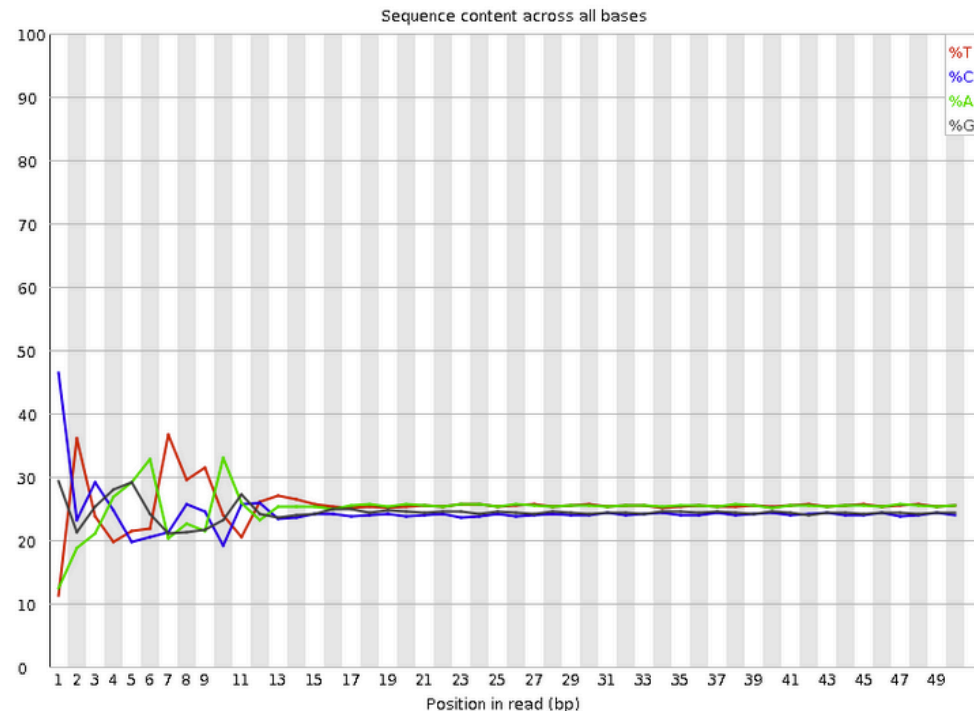
# Per sequence quality score on another sample

- Allows you to see if a subset of your sequences have universally low quality values



→ these should represent only a small percentage of the total sequences

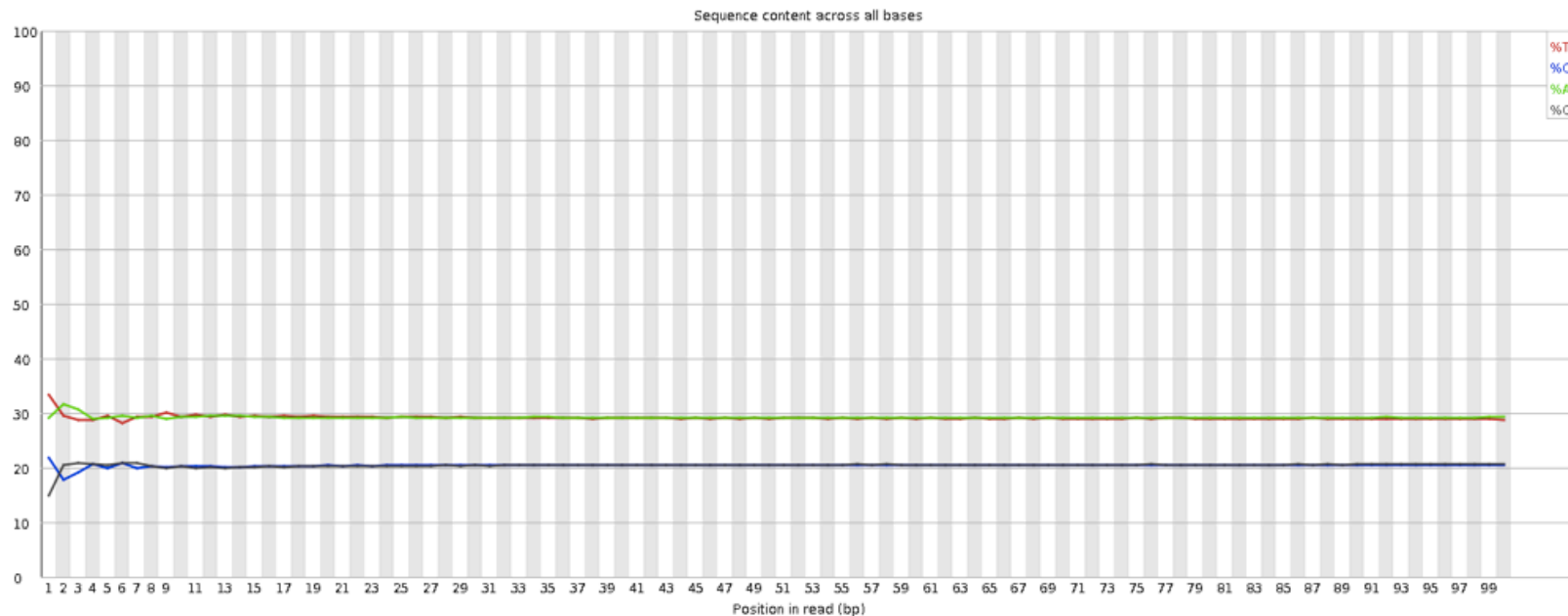
# Per base sequence content



- **Known bias in the repartition of the first nt in RNA-seq libraries**
  - Because random primers used during RT are “not so random”
  - “Reproducible bias” → Comparative analyses OK
  - c.f. Hansen et al. 2010;38(12):e131.  
Li et al. Genome Biology 2010;11(5):R50.

# Per base sequence content on other samples

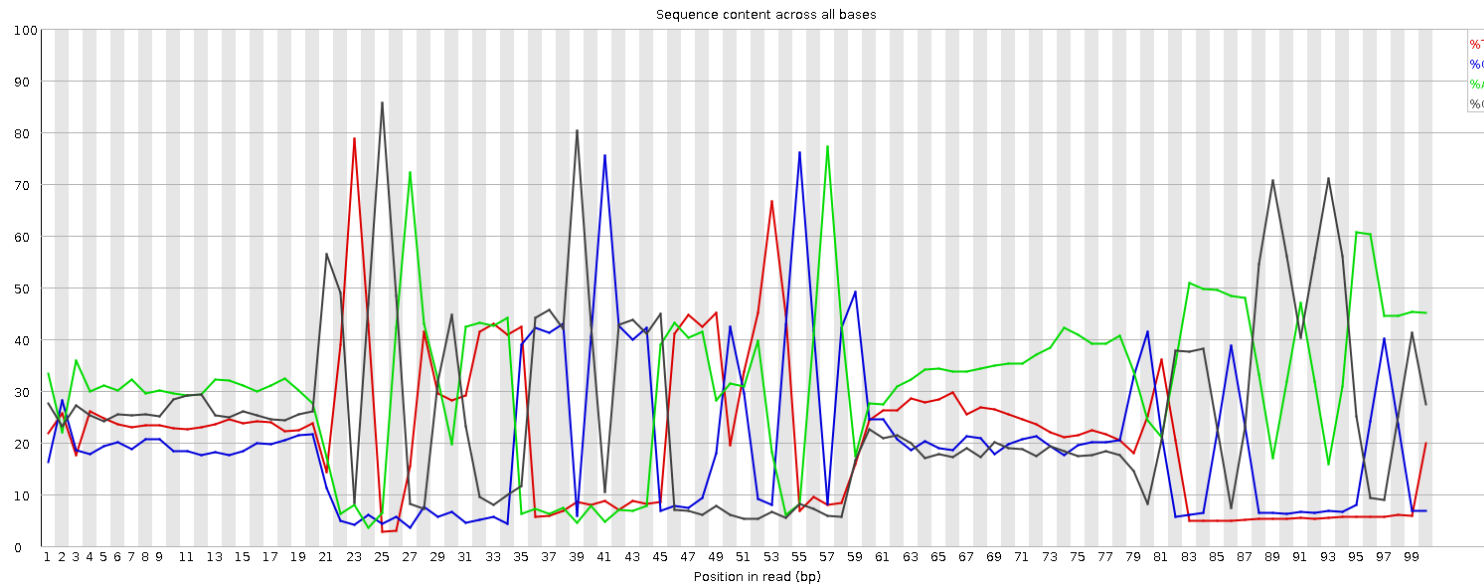
- The lines in this plot should run parallel with each other
- The relative amount of each base should reflect the overall amount of these bases in your genome
- Example for a DNaseq sample :





# Per base sequence content on other samples

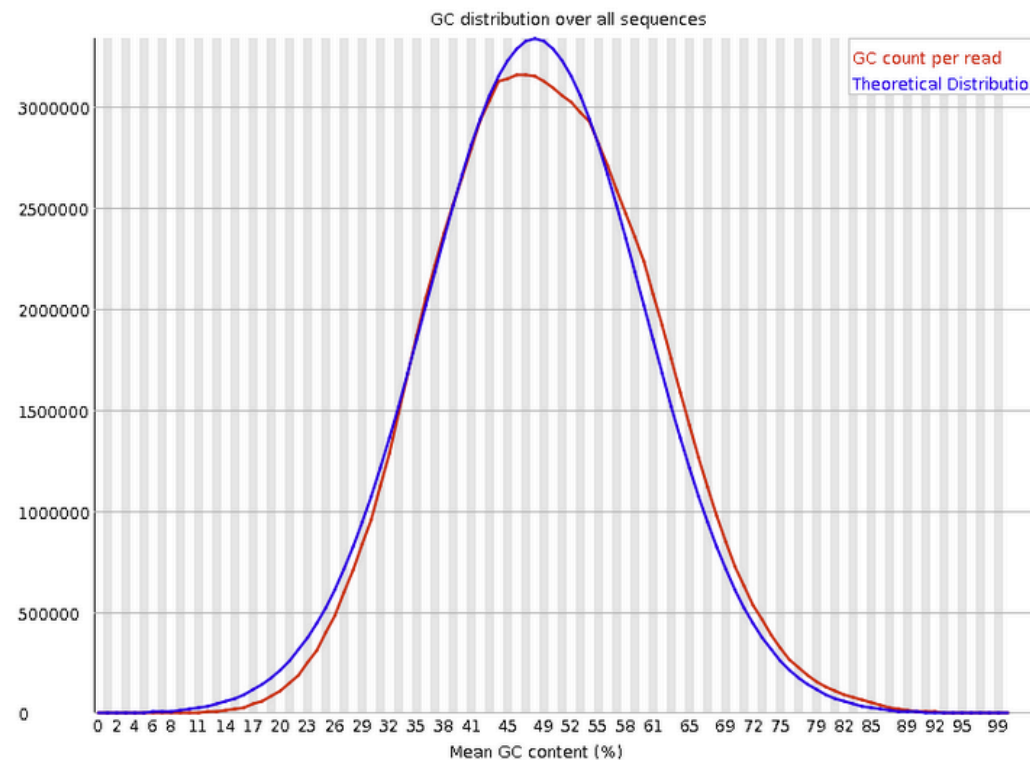
- Strong biases which change in different bases
  - Usually indicates an overrepresented sequence, e.g. adapters :



- Bias which is consistent across all bases
  - indicates that the original library was sequence biased
  - or that there was a systematic problem during sequencing

# Per sequence GC content

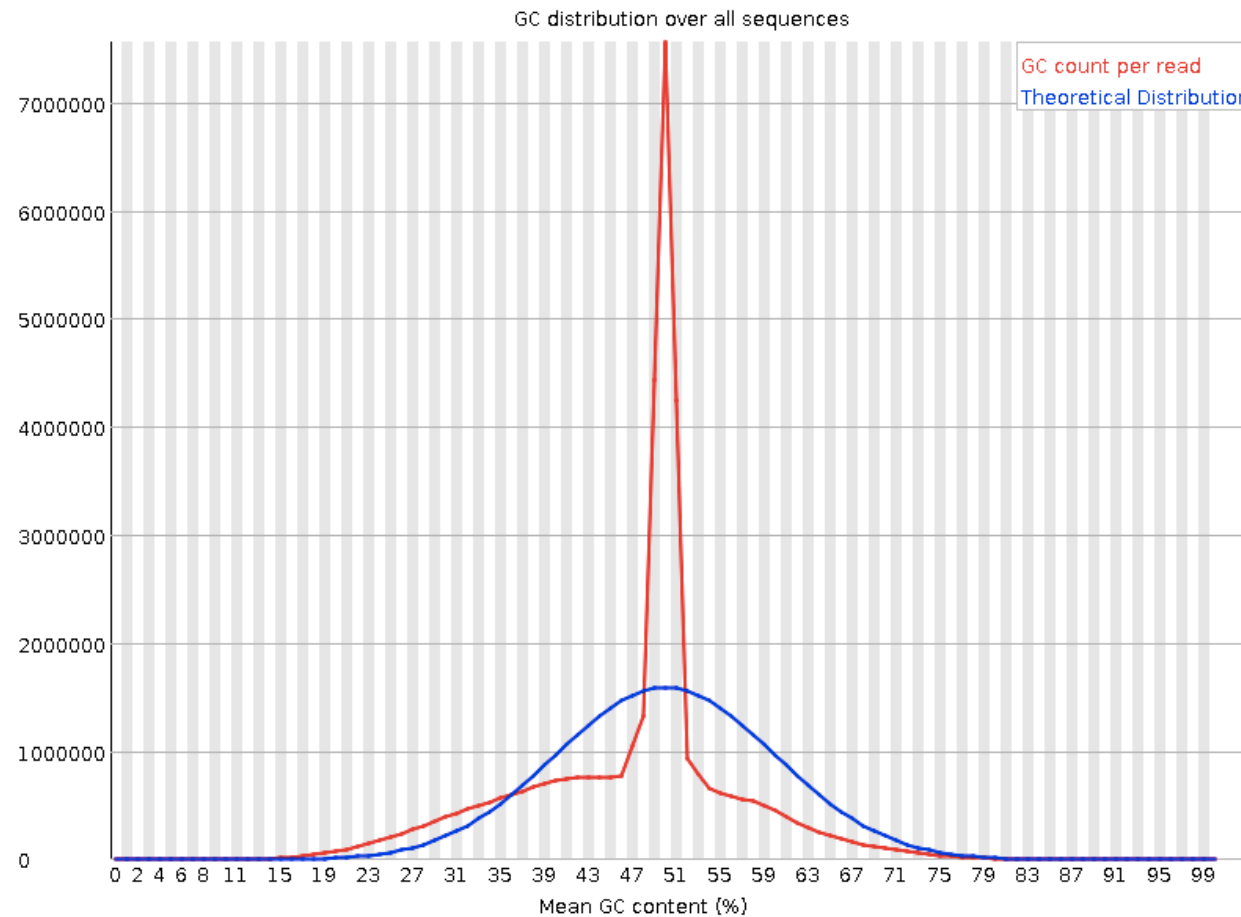
- Compares GC content of all sequences to a modelled normal distribution of GC content (mode calculated from the data and used to build the reference distribution)



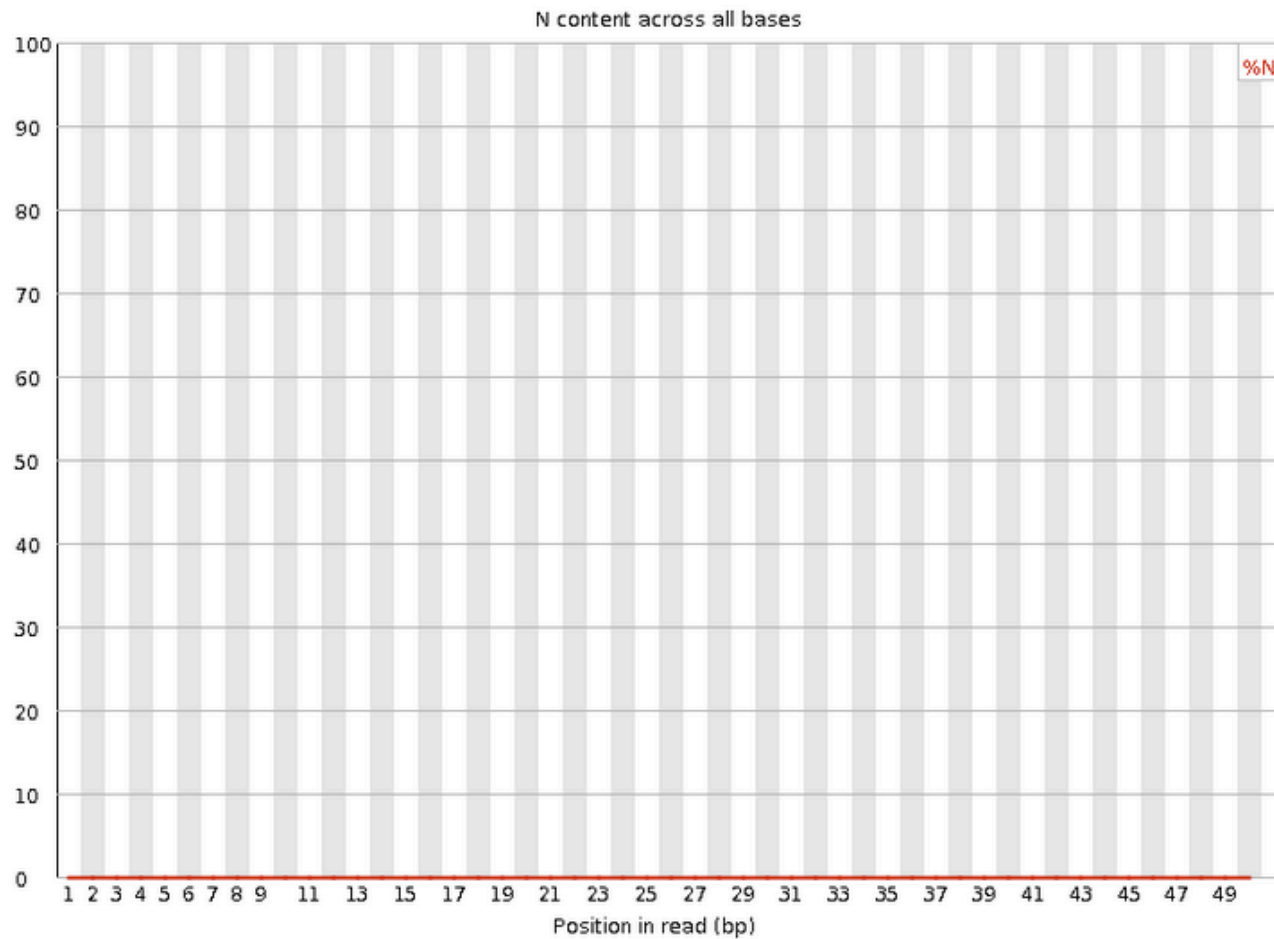
→ Observed GC distribution similar to the theoretical one

# Per sequence GC content on another sample

- Observed GC distribution very different to expected :



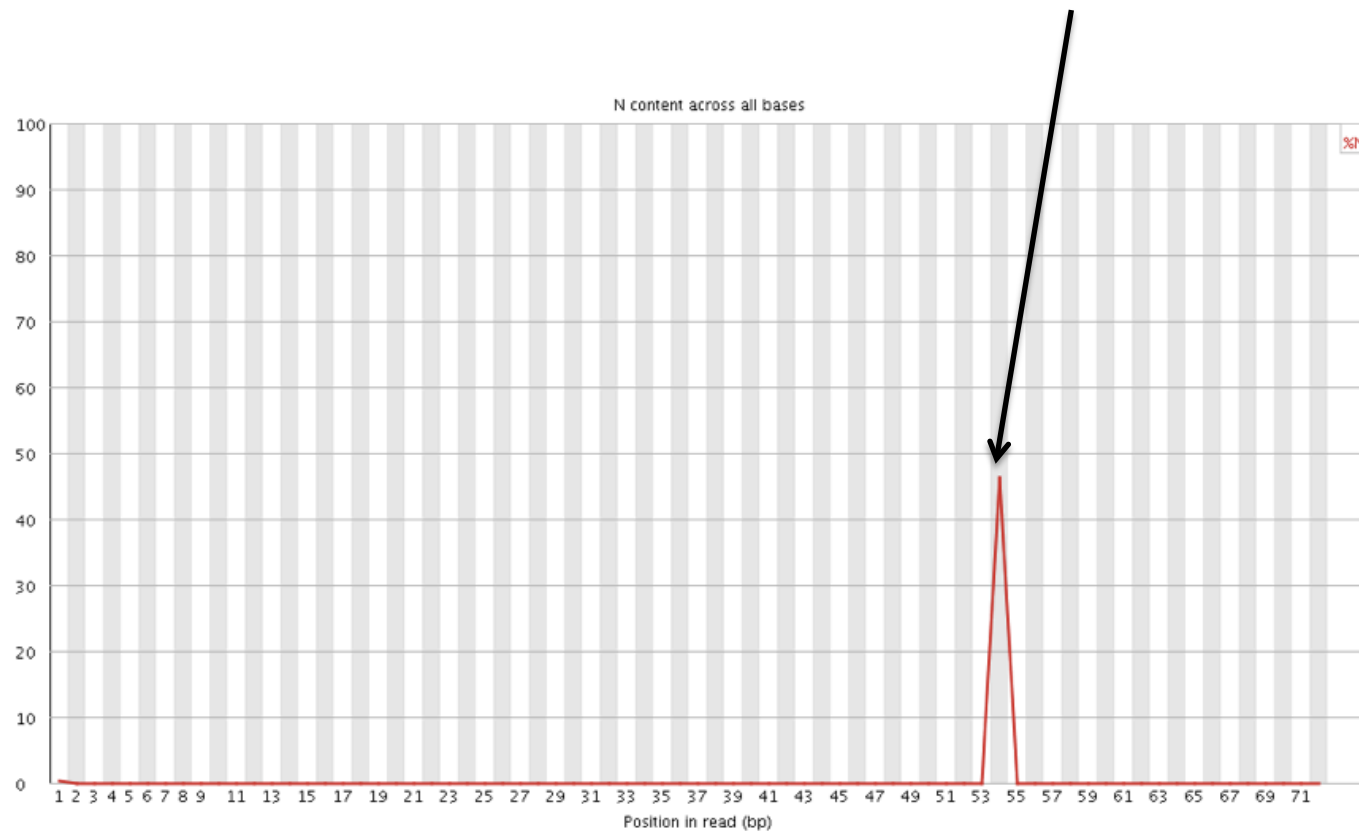
# Per base N content



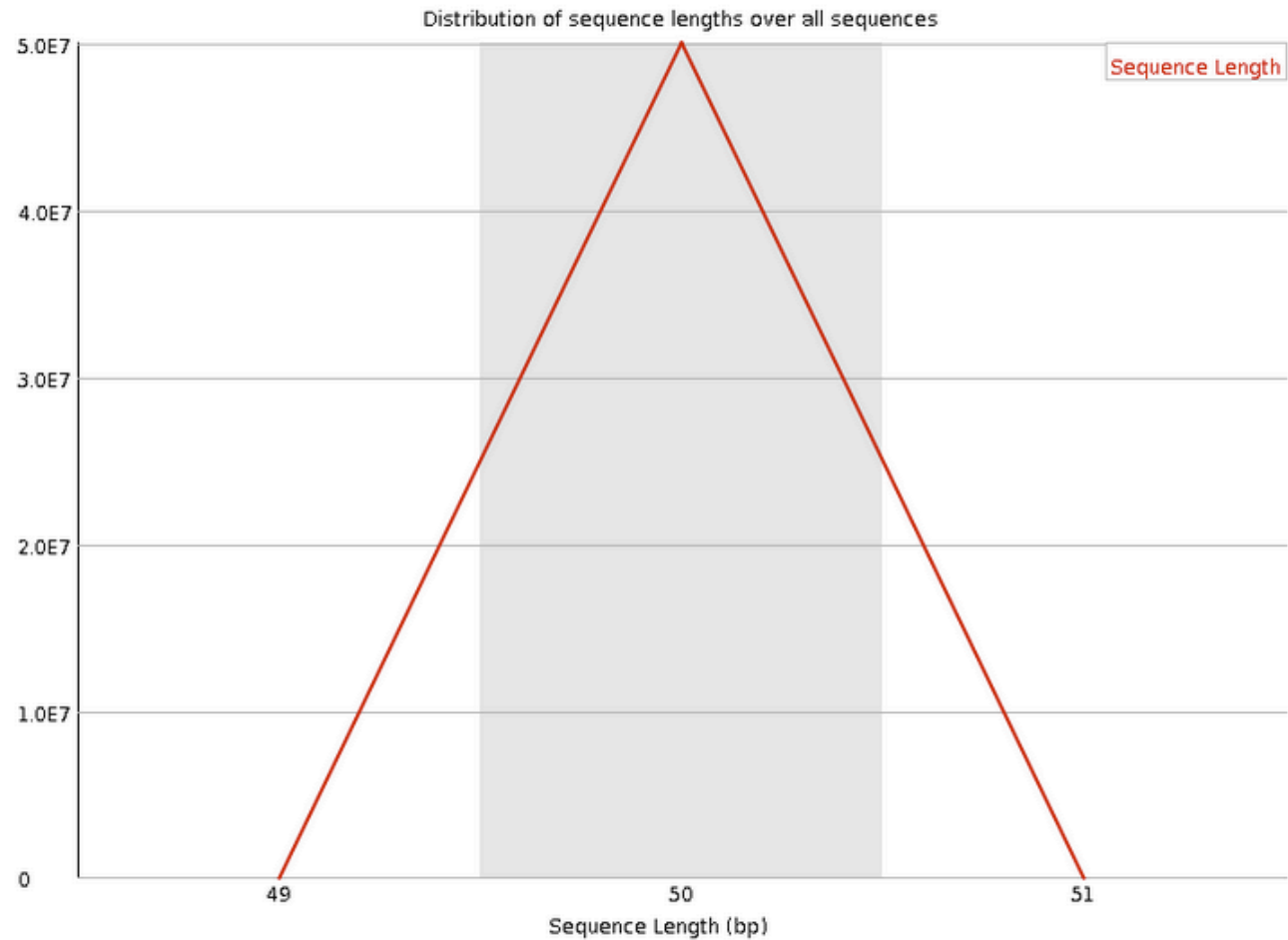
→ Very low N content

# Per base N content on another sample

- Can be used to detect bubbles (“Bottom Middle Swath”)



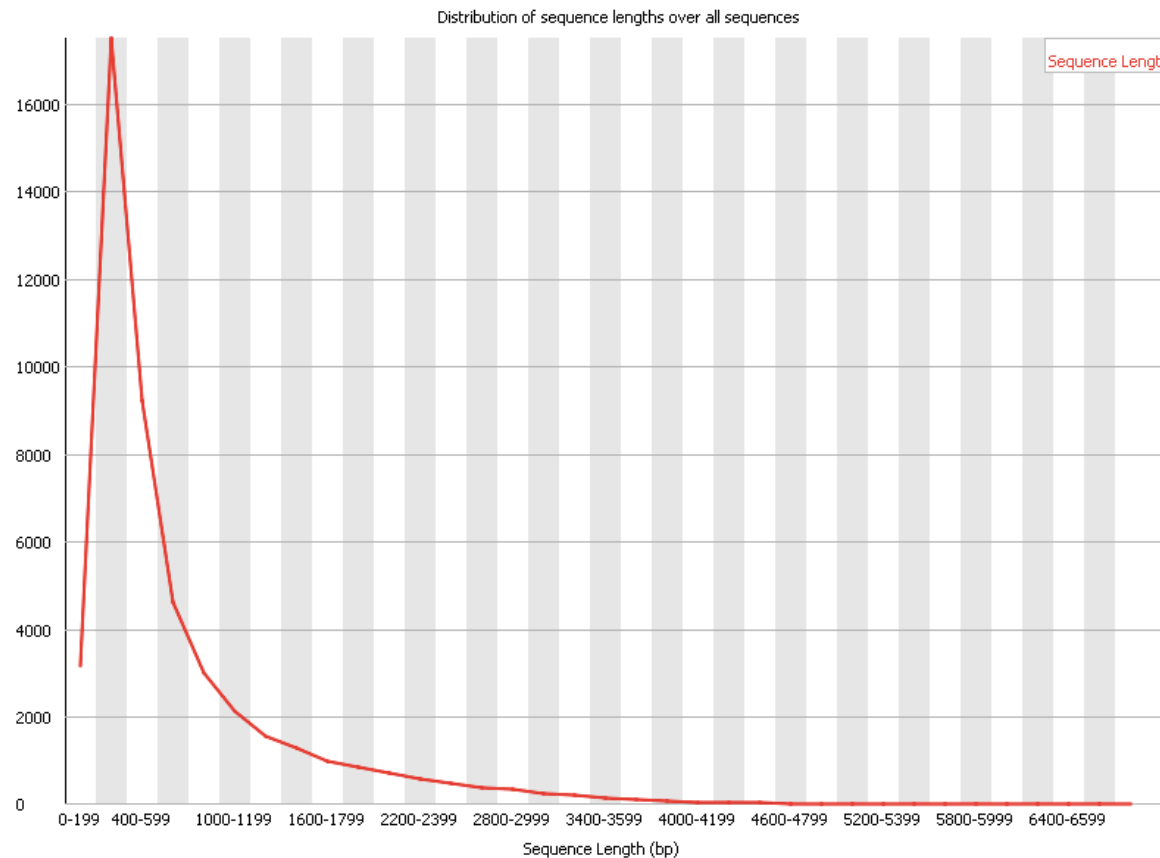
# Sequence length distribution



→ All sequences = 50bp reads

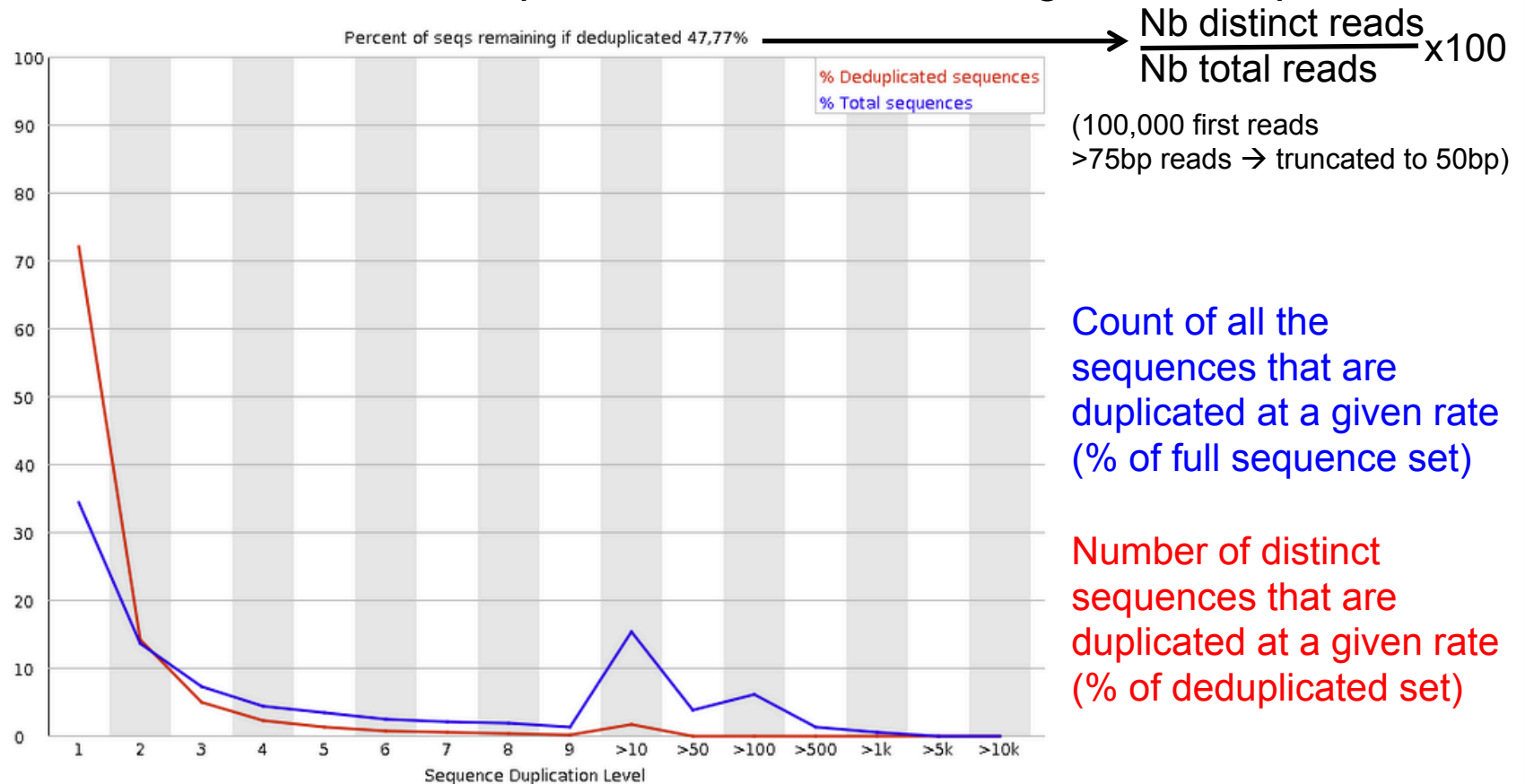
# Sequence length distribution on another sample

- Useful when different sequence lengths in the file



# Sequence duplication levels

- Relative number of sequences with different degrees of duplication



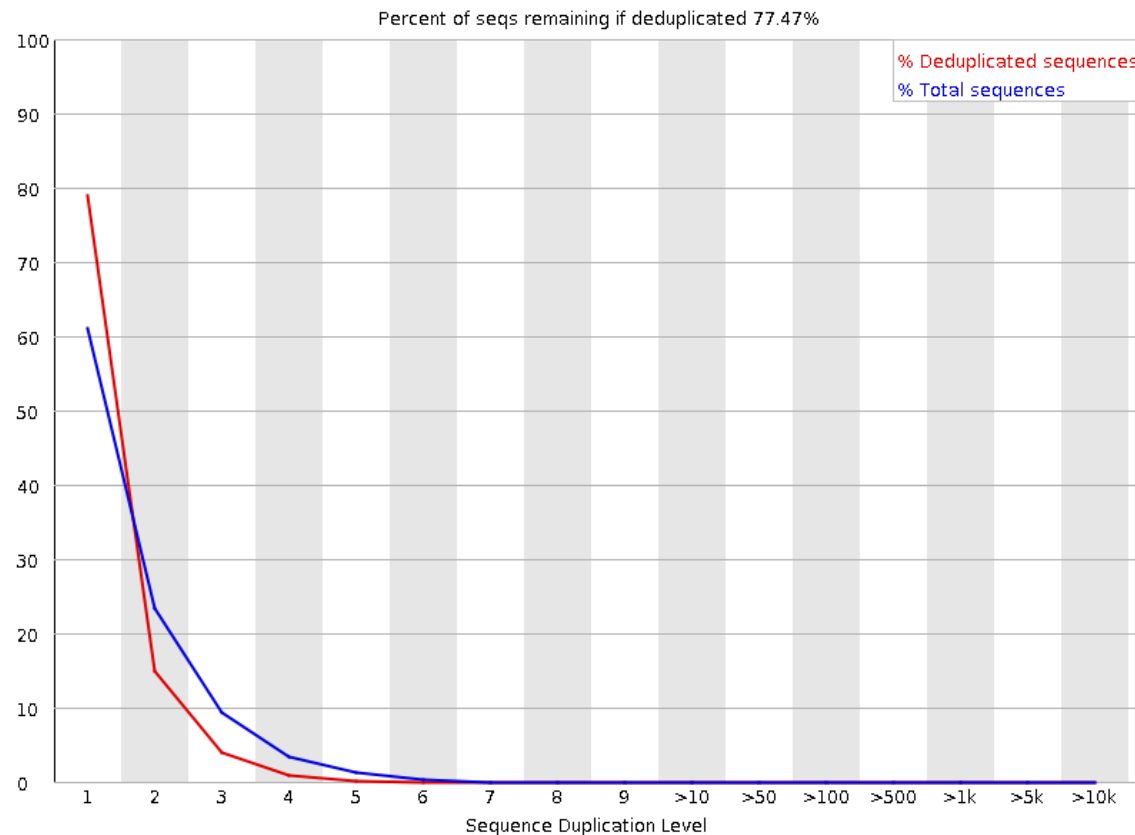
➔ OK for an RNA-seq sample :

Abundant mRNAs could lead to duplicated sequences



# Sequence duplication levels on other samples

## ■ Example for a DNA-seq sample



- A high level of duplication may indicate an enrichment bias, e.g. PCR over amplification

# Overrepresented sequences

---

- Lists all sequences representing more than 0.1% of the total

No overrepresented sequences

**→ No sequences representing > 0.1% of the total**

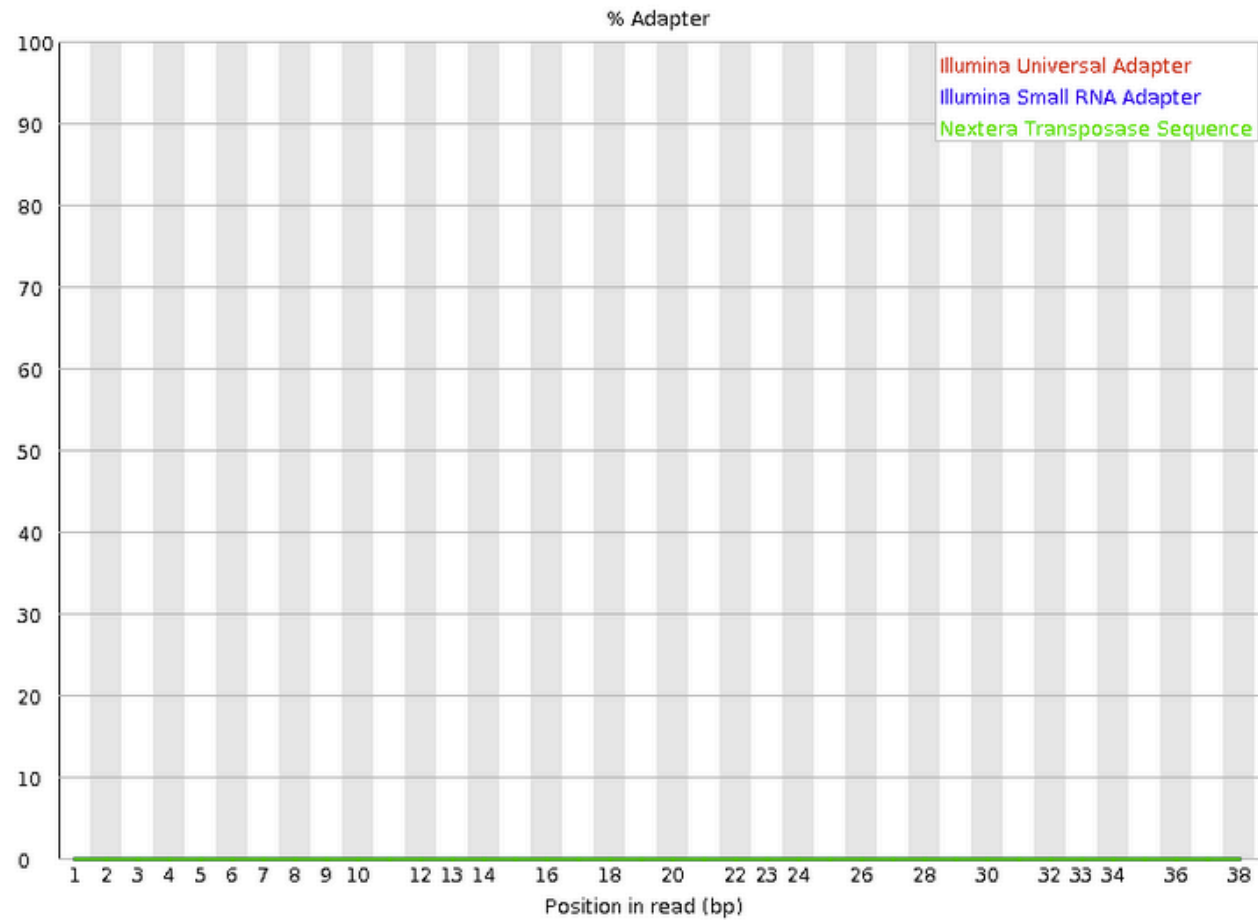
# Overrepresented sequences on another sample

---

- For each overrepresented sequence, FastQC will look for matches in a database of common contaminants
  - ➔ report the best hit, e.g. :

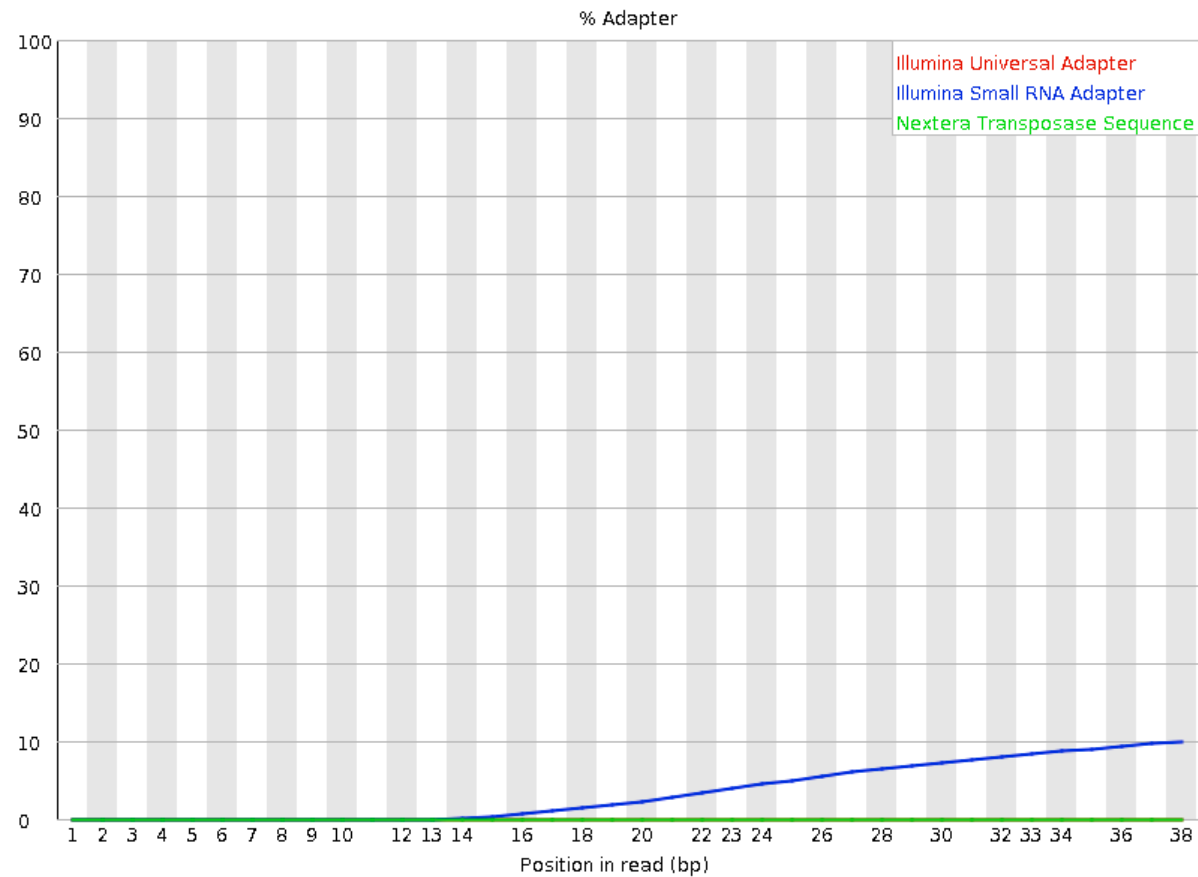
Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACTTCTGAACTCCAGTCACCGATGTATCTCGTATG	113163	0.614990735439532	TruSeq Adapter, Index 2 (97% over 49bp)
AGATCGGAAGAGCACACGTCTGAACTCAAGTCACCGATGTATCTCGTATG	41889	0.22764814397662272	TruSeq Adapter, Index 2 (97% over 49bp)
AGATCGGAAGAGCACACCTCTGAACTCCAGTCACCGATGTATCTCGTATG	39078	0.21237160520228368	TruSeq Adapter, Index 2 (97% over 49bp)

# Adapter content



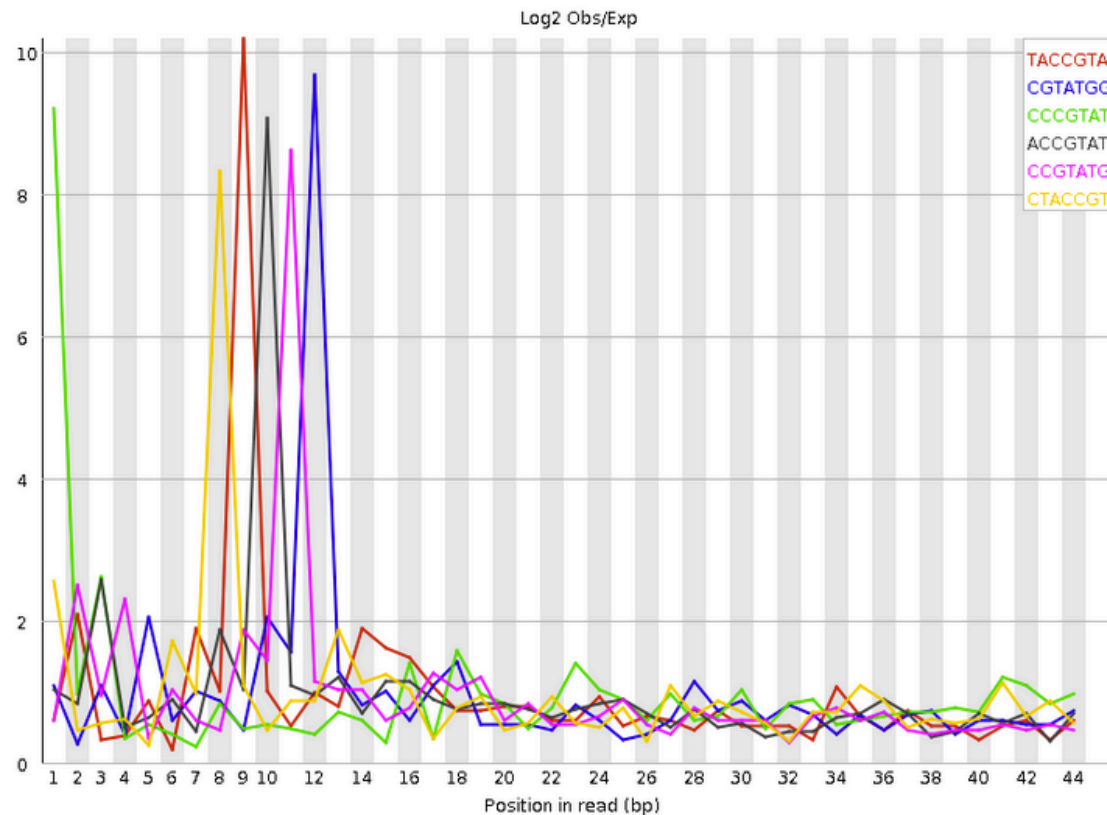
➔ No adapters

# Adapter content on another sample



➔ Reads have to be trimmed before analysis

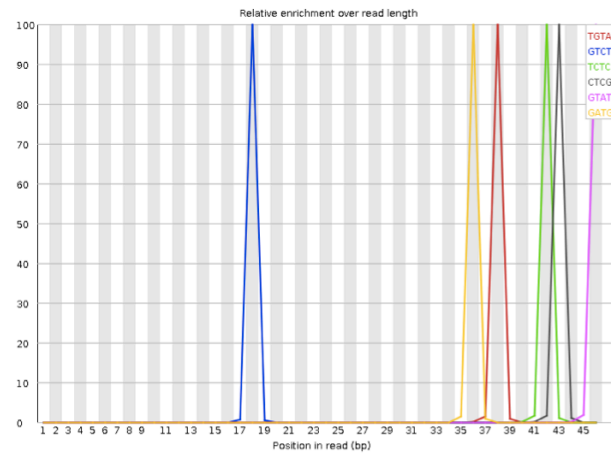
# K-mer content



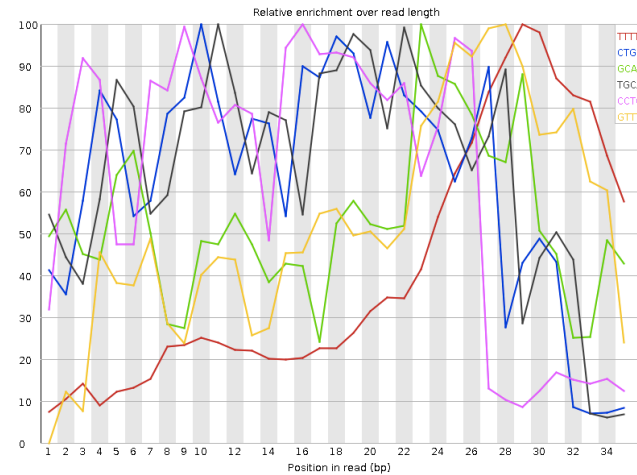
→ Bias in the repartition of the first nucleotides in RNA-seq libraries (as in the “per base sequence content” graph)

# K-mer content on other samples

- Presence of overrepresented sequences, e.g. adapters



- Bad quality sequence



# Quality control of Illumina data

---

- Primary analysis
- Quality control
- Data pre-processing



# Data pre-processing

---

## ■ Why ?

- Remove bad quality/contaminant data
- Improve confidence of downstream analysis

## ■ Needed ?

- Depend on what type of data and what type of analysis you want to perform on your data
  - e.g. smallRNA-seq : adapters removal required
  - e.g. assembly : cleaned data required
  - e.g. variant calling : has to be performed only on good quality reads / part of reads

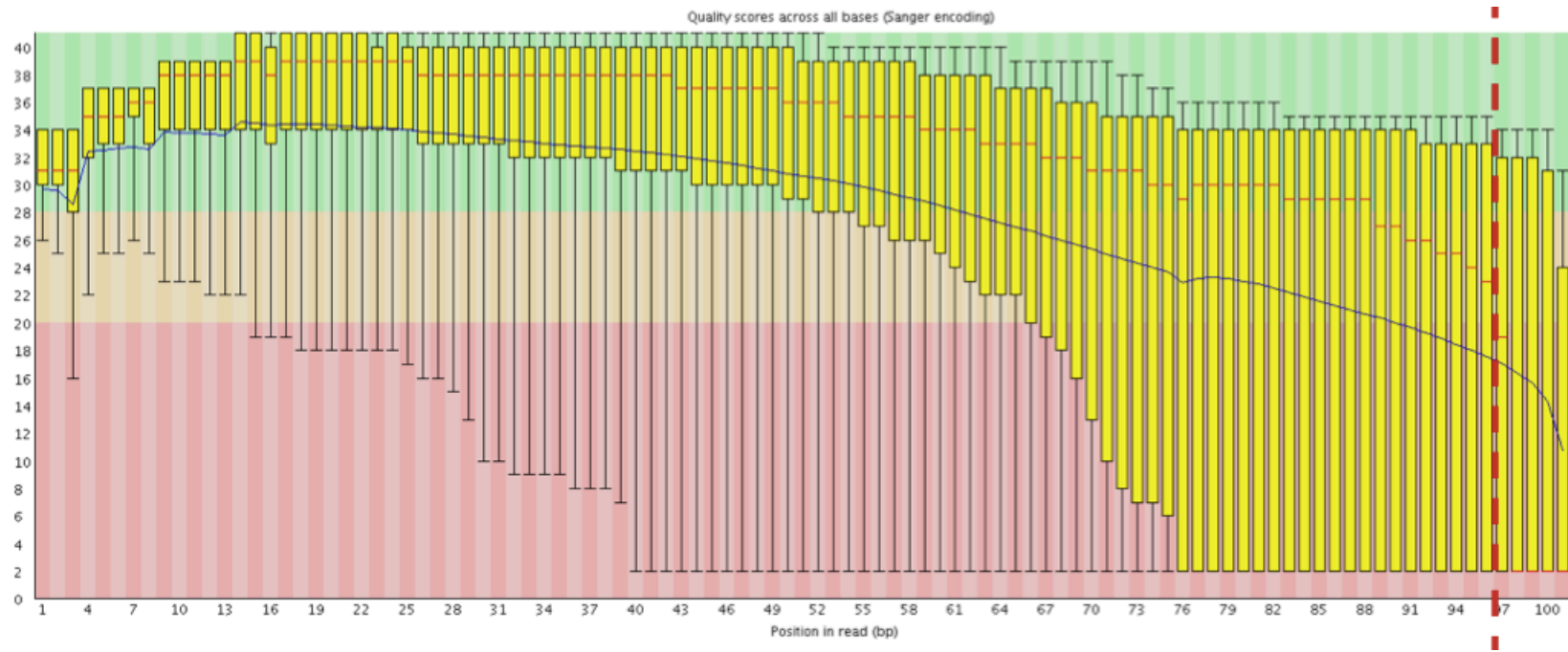
# Preprocessing tools

---

- Galaxy
  - e.g. <http://www.galaxeast.fr/>
- DeconSeq
  - <http://deconseq.sourceforge.net/>
- FASTX-Toolkit
  - [http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)
- Cutadapt
  - <https://code.google.com/p/cutadapt/>
- Trimmomatic
  - <http://www.usadellab.org/cms/?page=trimmomatic>
- Picard
  - <http://picard.sourceforge.net/>
- SolexaQA
  - <http://solexaqa.sourceforge.net/>
- ...

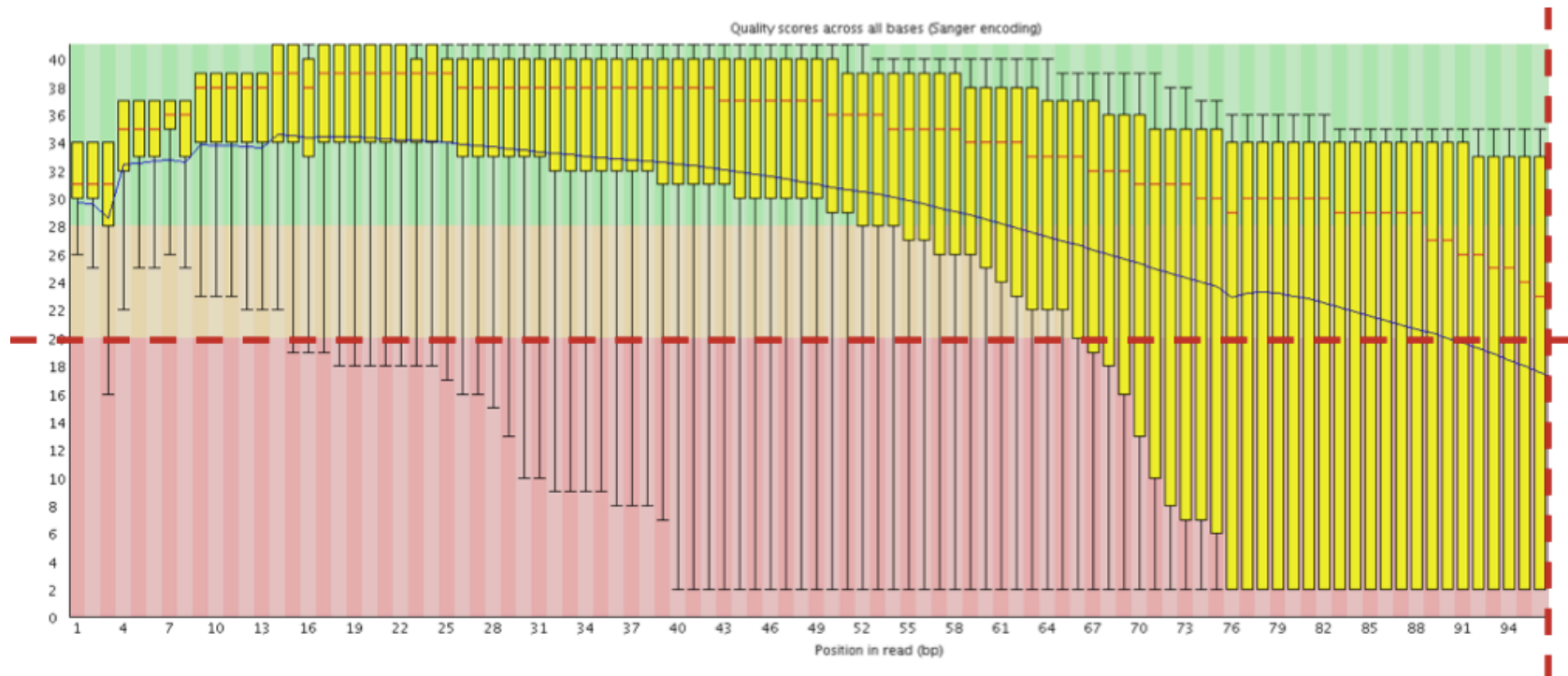
# Trimming

- Remove low quality bases from the sequence end
- e.g. trim reads when the median base quality falls below 20



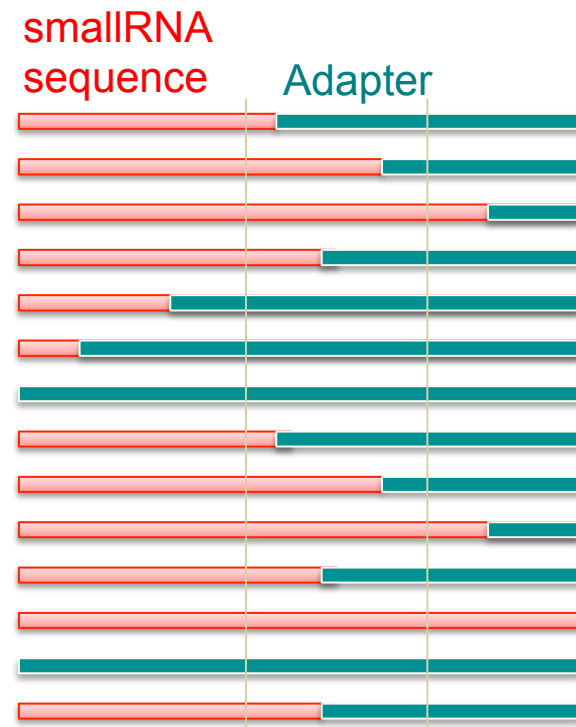
# Filtering low quality reads

- Keep only reads with a sufficient quality
- e.g. retain only reads with an average base quality score  $\geq 20$



# Removing/clipping adapter sequences

- e.g. small RNA-seq library
  - Remove adapter sequences
  - Remove too-short sequences
  - Remove too-long sequences
  - Clip adapters



# Removing contaminants

---

Possibly :

- Sequences used during library preparation
  - e.g. Spikes
- Sequences from other organisms
  - e.g. Xenografts
- rRNA sequences