מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

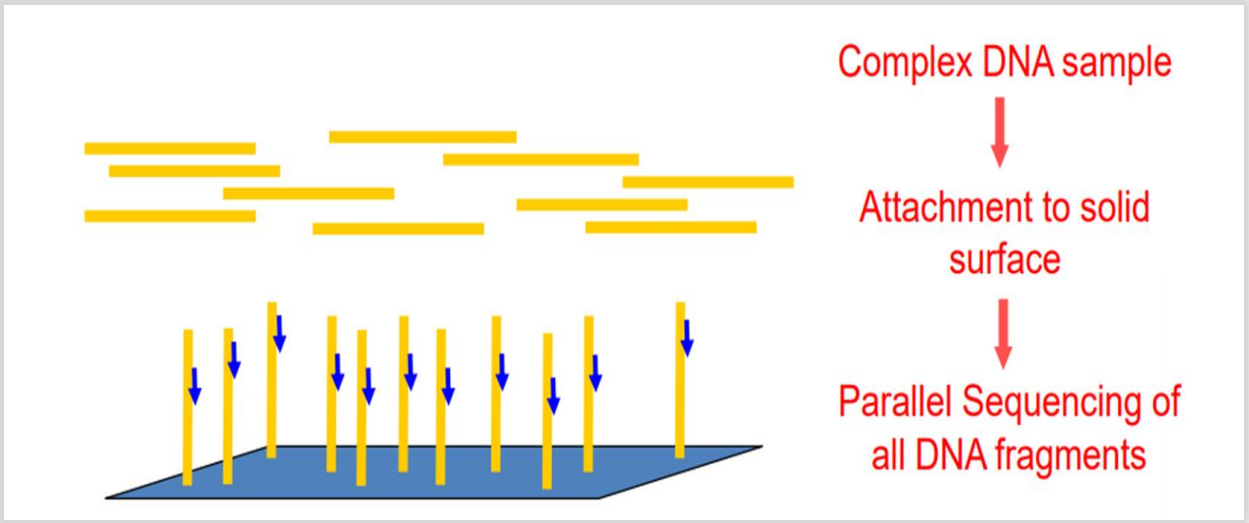
Illumina Primary Analysis Pipeline & Quality Control

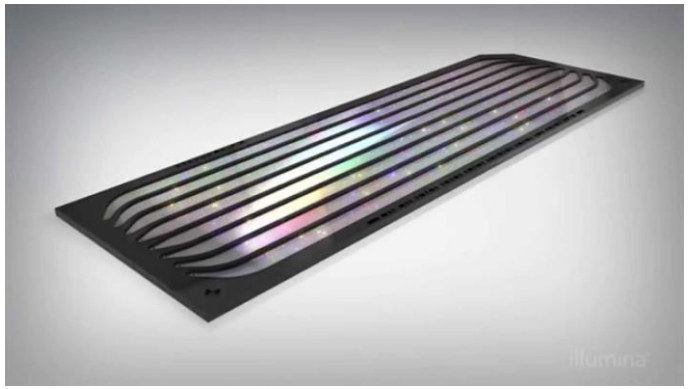
Noa Wigoda

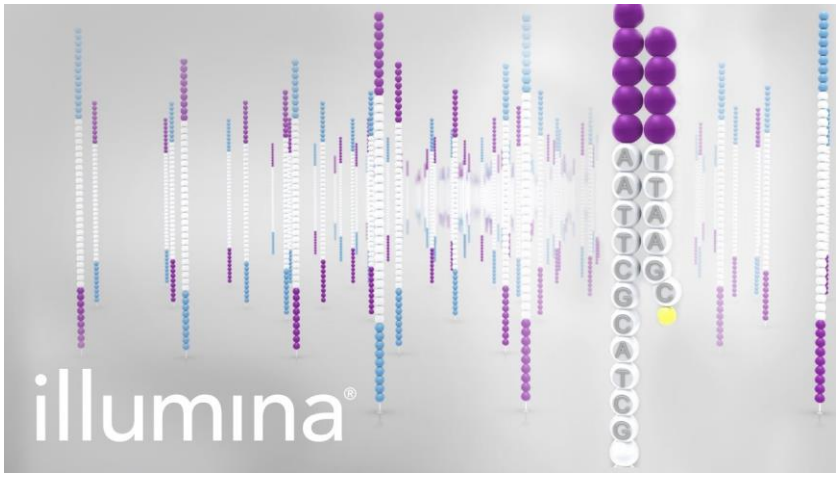
19.11.19

An Introduction to deep-sequencing analysis for biologists

NGS – What is the essence of this technology?







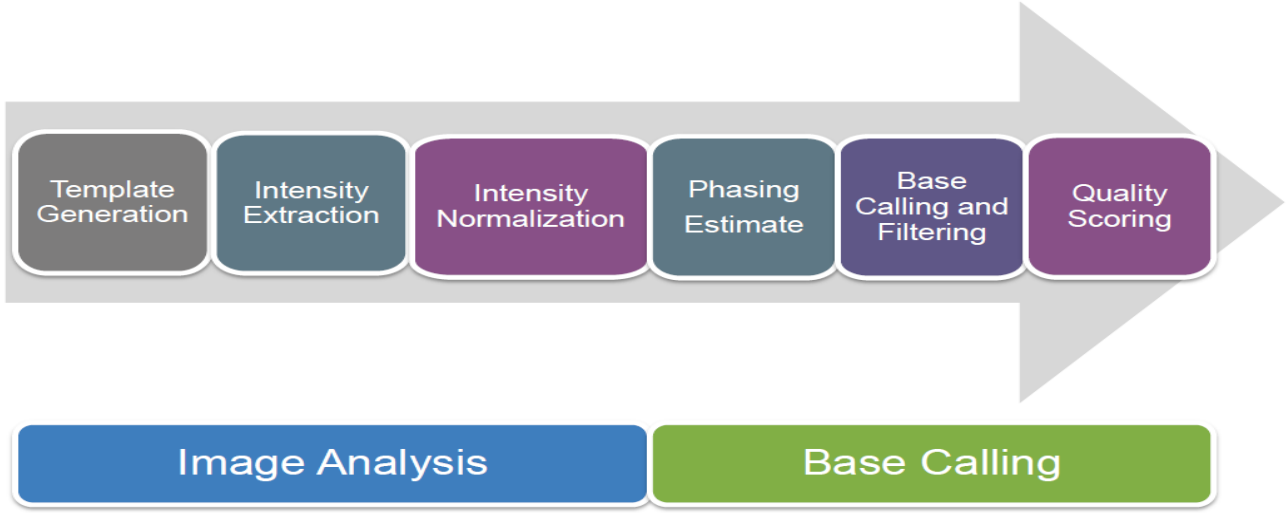
Illumina machines in Weizmann

Product	MiSeq	NextSeq	NovaSeq
Run Time	4–55 hours	12–30 hours	~13 - 38 hours, FC: dual SP ~13–25 hours, FC: dual S1 ~16–36 hours, FC: dual S2 ~44 hours, FC: dual S4
Maximum Reads Per Run	25 million	400 million	20 billion
Maximum Read Length	2 × 300 bp	2 × 150 bp	2 x 250

https://emea.illumina.com/systems/sequencing-platforms.html#

FC: Flow Cell

Primary data analysis workflow



illumina

- RTA – Real time analysis of images
  - Images generation per color and per cycle (of all the flow cell)
  - From the images we perform base calling & quality scoring
- Images are stored only for the first cycles in order to identify clusters (for random flow cells only).
- Once intensities per cluster are extracted the images are deleted.

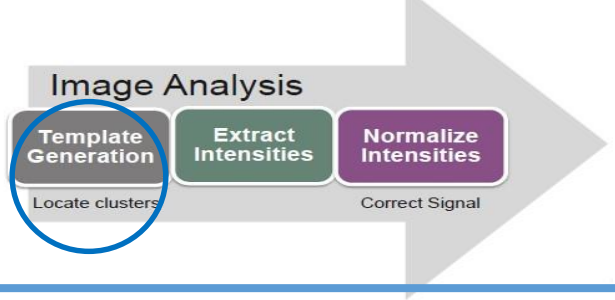
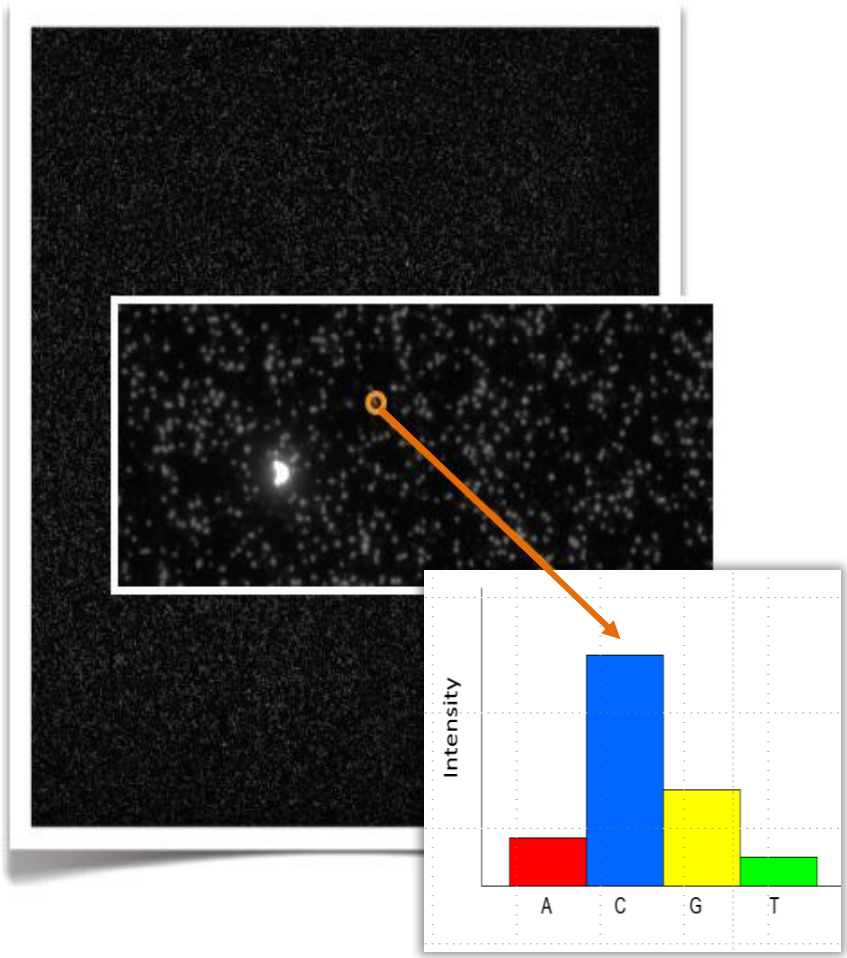
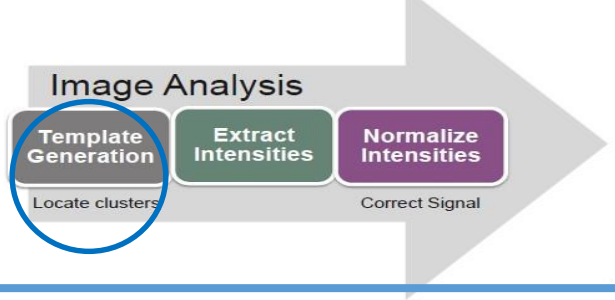


Image analysis steps

- Each single cluster is identified and quantified across all images of a cycle
- In this case the highest base quantified is C





Template generation

Template Generation

HiSeq 2500, NextSeq, MiniSeq, MiSeq

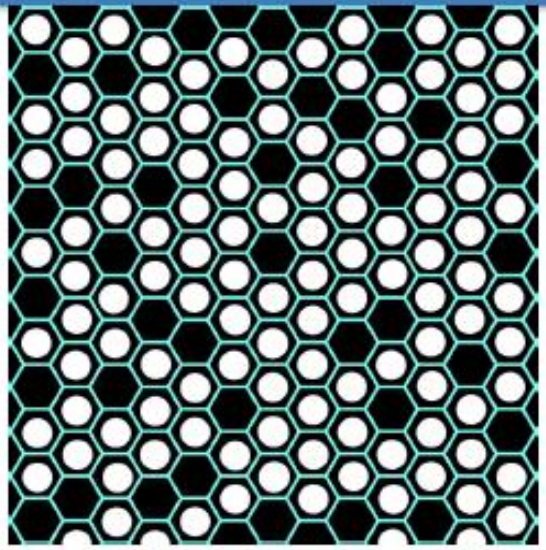
Identify the location of every cluster and create a map

www.lookandlearn.com/blog/5793/constellations-mapping-stars/

Rigid Registration

HiSeq 3000, HiSeq 4000, HiSeq X, NovaSeq

Start with a map of where clusters could be and overlay on images

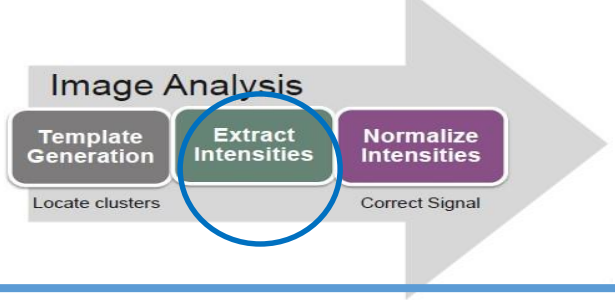


/black-light-studio.deviantart.com/art/Free-Hexagon-pattern-02-371945610

Identifying cluster location is based on the first 5 cycles.

Fixed cluster locations on patterned flow cells eliminates the need for template generation.

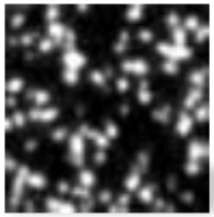
illumina



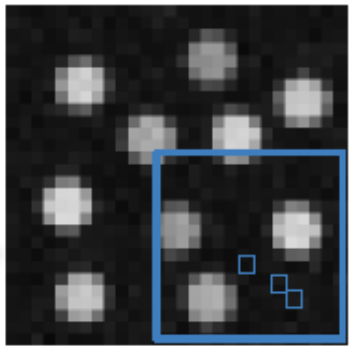
Intensity extraction

Background Subtraction

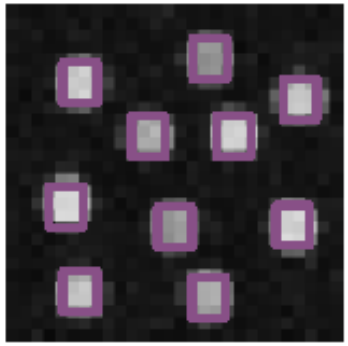
Compute background → **Compute signal for each cluster** → **Subtract background from each cluster**



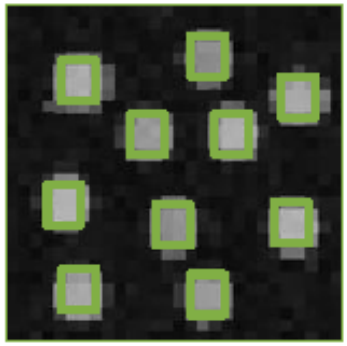
Clusters are not perfect circles. RTA sees them as shown



Background is calculated by averaging the intensity of the dimmest pixels in a region



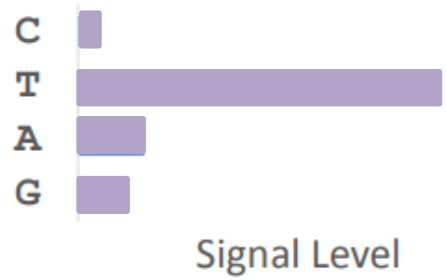
Cluster intensity is extracted from the brightest part of each cluster



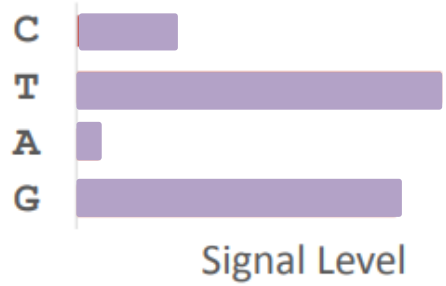
Background is subtracted from signal of each cluster

illumina®

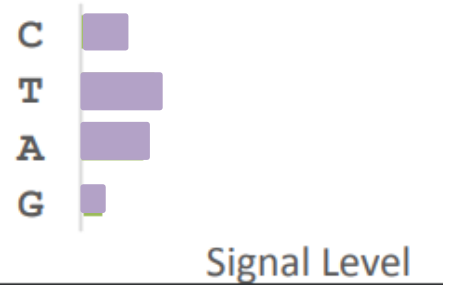
Goal (simplified): identify base



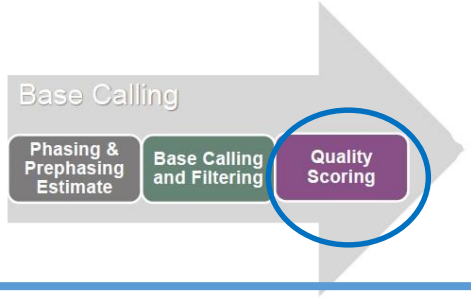
Call = T
Confidence = High



Call = T
Confidence = Low



Call = T / N ?
Confidence = Low



Quality scores

A quality score is a prediction of the probability of an error in base calling.

Phred quality scores are logarithmically linked to error probabilities

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

$$\text{Phred} = -10 \log_{10} p$$

p = Probability call is incorrect

P	Log10(P)
0.1	-1
0.01	-2
0.05	-1.30103
0.001	-3

Sequence output format

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. (From Wikipedia)

Line 1: Unique ID for a sequencing read
Line 2: Sequences
Line 3:+
Line 4: Base calling quality score (Analogous to Phred scores but in ASCII value)

Example:

```
@HISEQ:126:H14YJADXX:1:1101:1118:2101 1:N:0:ATCACG
CTCCATAGTCAGAAACTTCAGCATGACAGTACCTCATGCTGCATCAGGTGATCATGAAAAGATTAC
+
@@?ADDDD?ADHDIIIIIIIEIIIGEFHC<?FH4C9E9BGAFIGH<DG9BD?@DGGEGHHG<DCBB
```

https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Quality score representation

There is an ASCII character associated with each nucleotide, representing its Phred quality score, the probability of an incorrect base call.

Illumina (L; 1.8+) Phred Score encoding

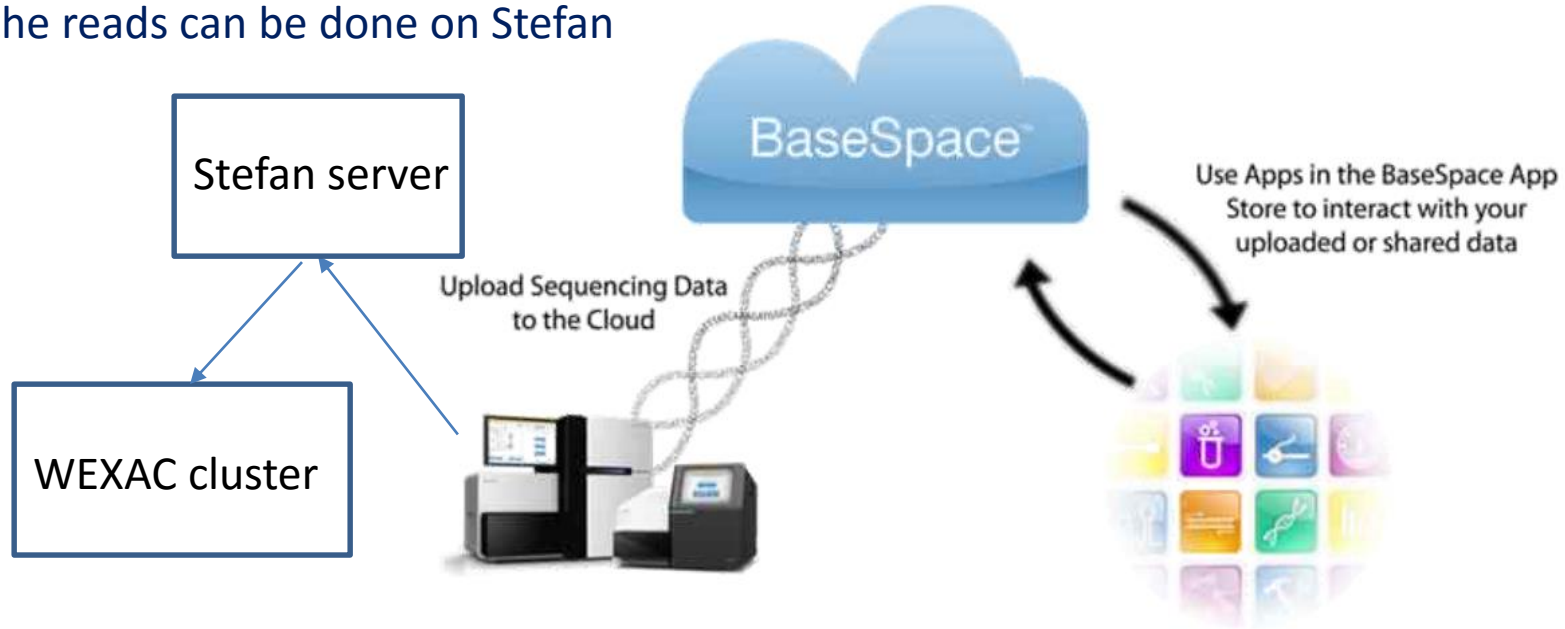
Symbol	Phred Score	Symbol	Phred Score	Symbol	Phred Score	Symbol	Phred Score
!	0	,	11	7	22	B	33
"	1	-	12	8	23	C	34
#	2	.	13	9	24	D	35
$	3	/	14	:	25	E	36
%	4	0	15	;	26	F	37
&	5	1	16	<	27	G	38
'	6	2	17	=	28	H	39
(	7	3	18	>	29	I	40
)	8	4	19	?	30	J	41
*	9	5	20	@	31		
+	10	6	21	A	32		

http://drive5.com/usearch/manual/quality_score.html

Storing sequence information

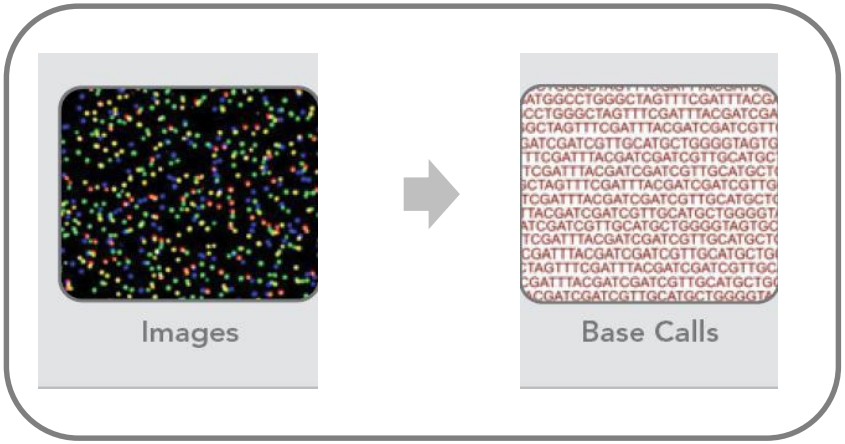
- A website where biologists and informaticians can easily store, analyse, and share genetic data from Illumina instruments
- How do you get there? → https://basespace.illumina.com

The process of creating Fastq files & demultiplexing the reads can be done on Stefan



Topics to be discussed

- Illumina sequencing technology advances
- Illumina primary analysis



- Sequencing quality control (QC)

Quality Control

There are 3 main areas where QC should be applied:

- Starting at the nucleic acids
- After Library preparation
- Post-Sequencing

Sequencing QC

Common sequence artifacts in NGS data:

- Read errors (base calling errors and small insertions/deletions)
- Poor quality reads
- Primer/adaptor contamination

FASTQC gives a quick impression of whether your data has any problems of which you should be aware **before** doing any further analysis

Developed by Simon Andrews at the Babraham Institute in Cambridge
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

QC reports for NextSeq runs at LSCF

General QC for run AHNCYJBGX7

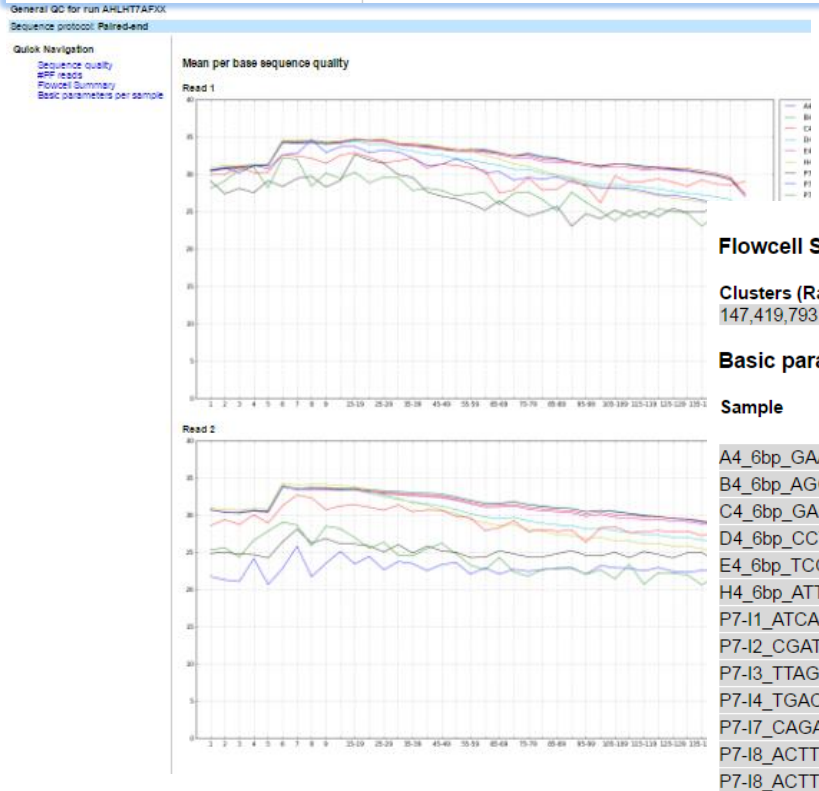
Sequence protocol: **Paired-end**

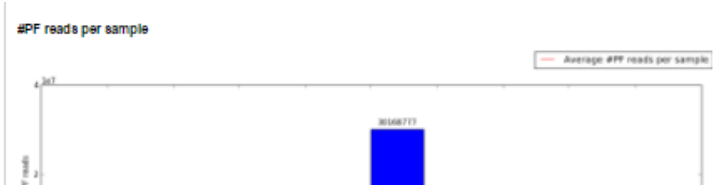
Quick Navigation

- Sequence quality
- #PF reads
- Flowcell Summary
- Basic parameters per sample

Ewels, Philip, et al. **"MultiQC: summarize analysis results for multiple tools and samples in a single report."** Bioinformatics 32.19 (2016): 3047-3048.

See here a more comprehensive report of MultiQC software





Flowcell Summary

Clusters (Raw)	Clusters (PF)	Yield (MBases)
147,419,793	115,248,805	34,575

Basic parameters per sample

Sample	Index	# PF Clusters	% Clusters per sample	Yield (MBases)	%≥Q30	FastQC Analysis	
A4_6bp_GAAGAA	GAAGAA	7,784,351	672.75	2,335	81.71 (R1) 72.38 (R2)	R1	R2
B4_6bp_AGGATC	AGGATC	5,678,462	490.50	1,703	82.04 (R1) 72.68 (R2)	R1	R2
C4_6bp_GACAGT	GACAGT	4,848,140	419.25	1,455	82.48 (R1) 71.31 (R2)	R1	R2
D4_6bp_CCTATG	CCTATG	11,618,248	1006.25	3,485	64.39 (R1) 54.91 (R2)	R1	R2
E4_6bp_TCGCCT	TCGCCT	4,487,566	387.50	1,346	80.13 (R1) 68.38 (R2)	R1	R2
H4_6bp_ATTCTA	ATTCTA	30,168,777	2621.00	9,050	61.09 (R1) 44.91 (R2)	R1	R2
P7-I1_ATCACG_Benny	ATCACG	130	0.00	0	23.08 (R1) 22.31 (R2)	R1	R2
P7-I2_CGATGT_Benny	CGATGT	98	0.00	0	57.14 (R1) 21.43 (R2)	R1	R2
P7-I3_TTAGGC_Benny	TTAGGC	6	0.00	0	(R1) (R2)	R1	R2
P7-I4_TGACCA_Benny	TGACCA	11	0.00	0	(R1) (R2)	R1	R2
P7-I7_CAGATC_Benny	CAGATC	29	0.00	0	20.69 (R1) 6.90 (R2)	R1	R2
P7-I8_ACTTGA_Benny	ACTTGA	103	0.00	0	50.49 (R1) 43.69 (R2)	R1	R2
P7-I8_ACTTGA_Benny	ACTTGA	103	0.00	0	50.49 (R1) 43.69 (R2)	R1	R2
Undetermined Indices	Undetermined	50,662,884	4402.25	15,200			

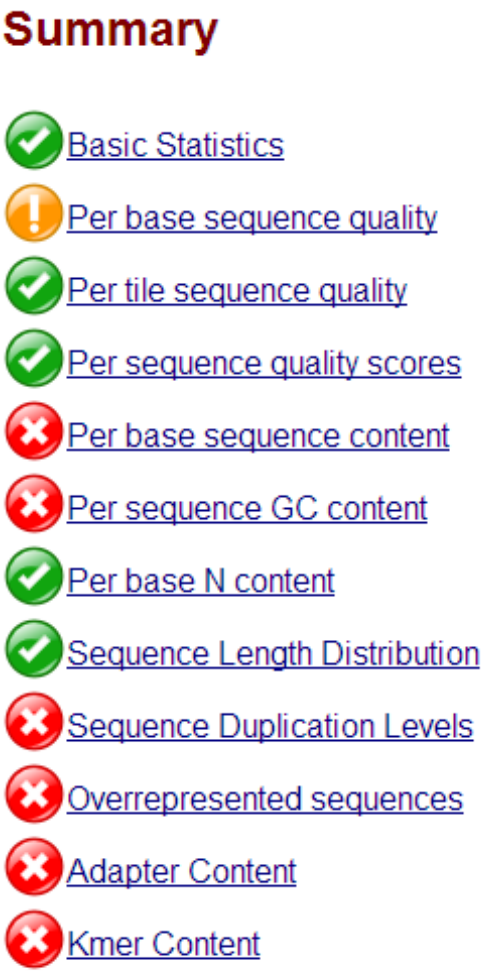
Example : http://stefan.weizmann.ac.il/fqc/181018_NB501465_0390_AHNCYJBGX7/

Report of Quality: FASTQC

Traffic light warning system:

- normal (green)
- abnormal (orange)
- bad (red)

Aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.



https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

Initial QC -
What does QC tell you about your library?

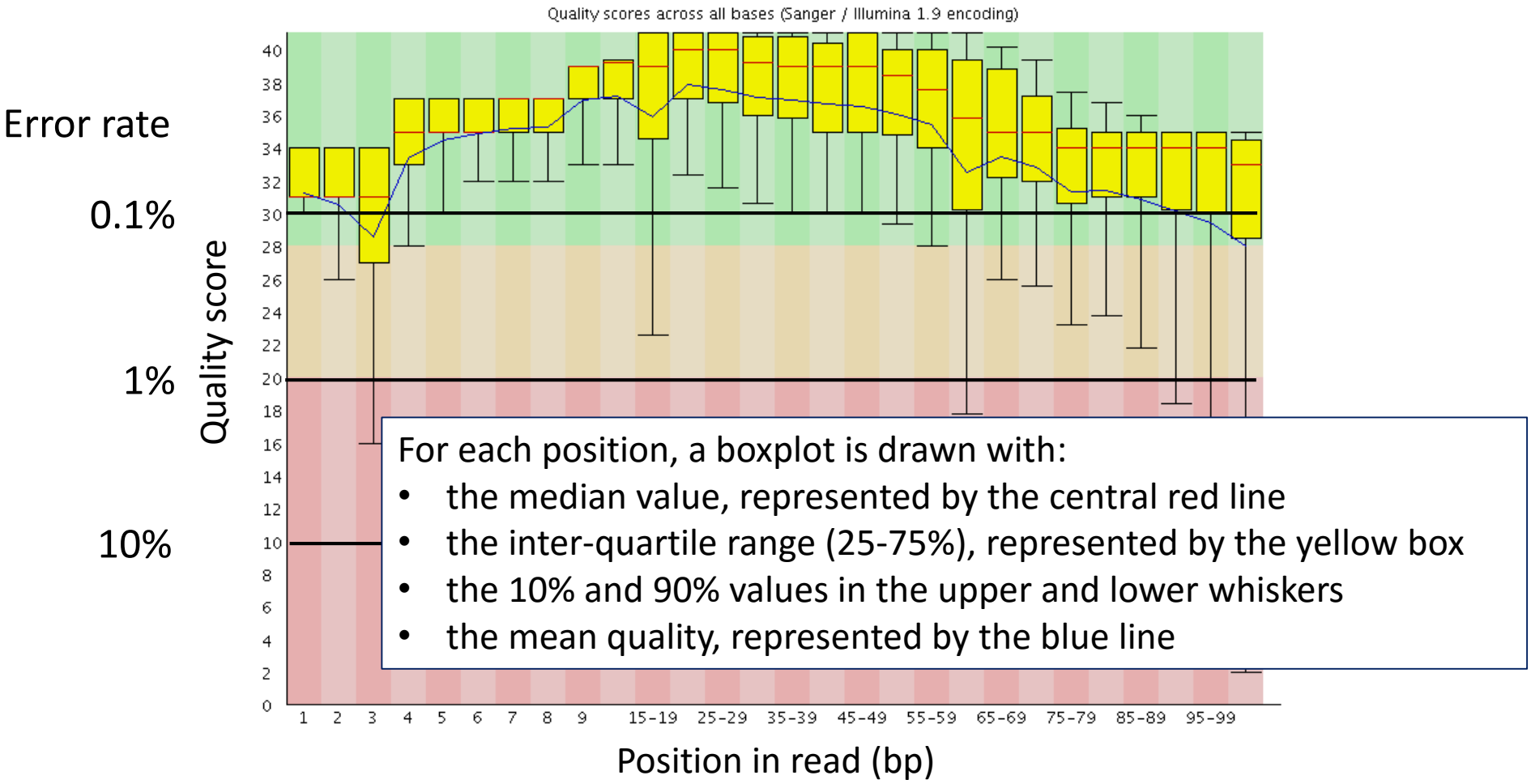
- # of sequences
- Base call qualities
- Base composition
- Potential contaminants
- Expected duplication rate

Basic Statistics

Measure	Value
Filename	s_4_1_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	35290120
Sequence length	40
%GC	46

Babraham Bioinformatics

FASTQC
Per base distribution of sequence quality



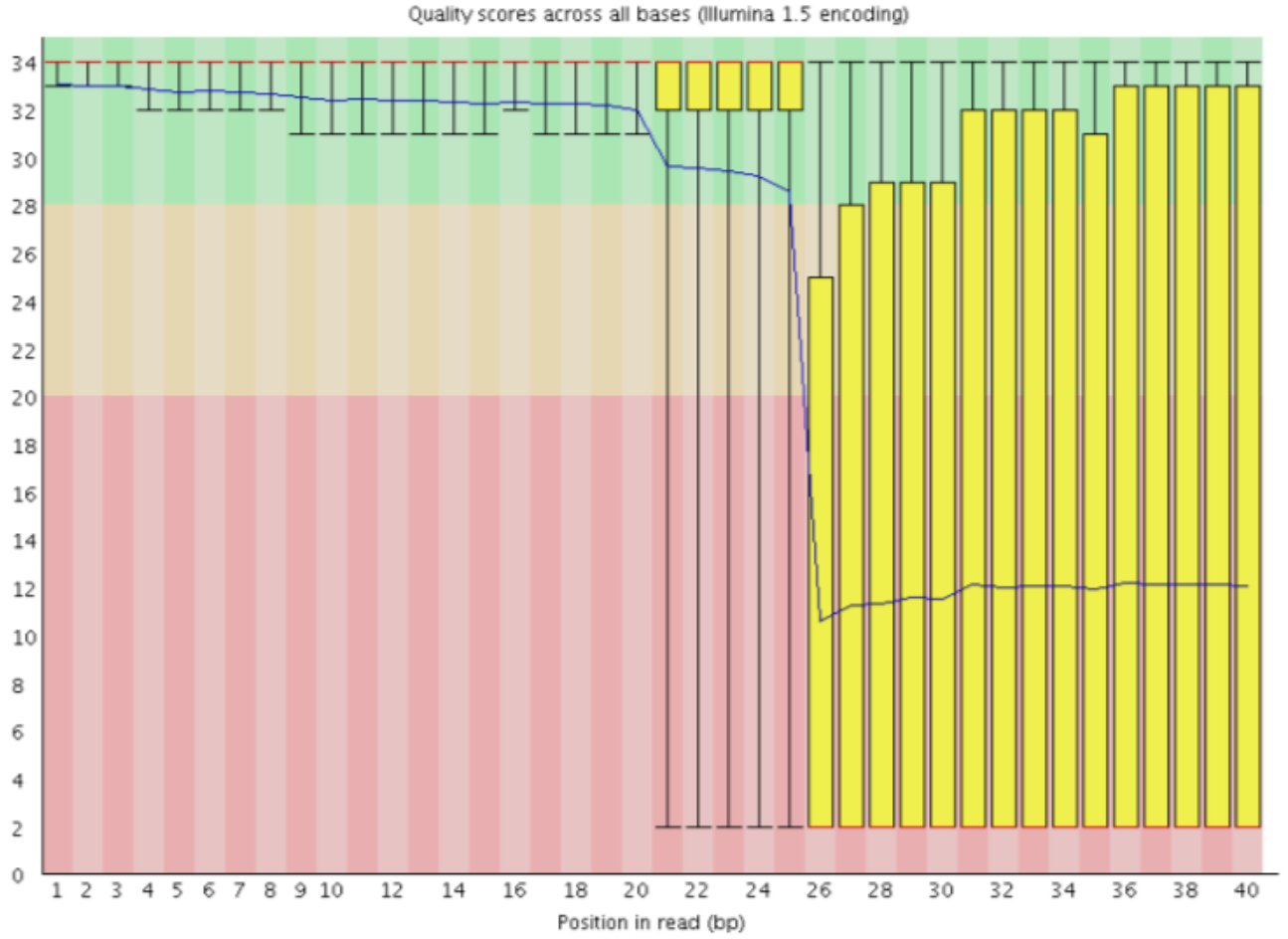
For each position, a boxplot is drawn with:

- the median value, represented by the central red line
- the inter-quartile range (25-75%), represented by the yellow box
- the 10% and 90% values in the upper and lower whiskers
- the mean quality, represented by the blue line

This plots the Q-score of the raw sequence reads as a box-plot for each cycle.
Higher is always better, and the characteristic decay of quality is seen in most runs.

Over clustering of the flow cells:

- Results in small distances between clusters and an overlap in the signals.
- Two clusters can be interpreted as a single cluster with mixed fluorescent signals being detected, decreasing signal purity.
- It generates lower quality scores across the entire read.



- Instrumentation breakdown

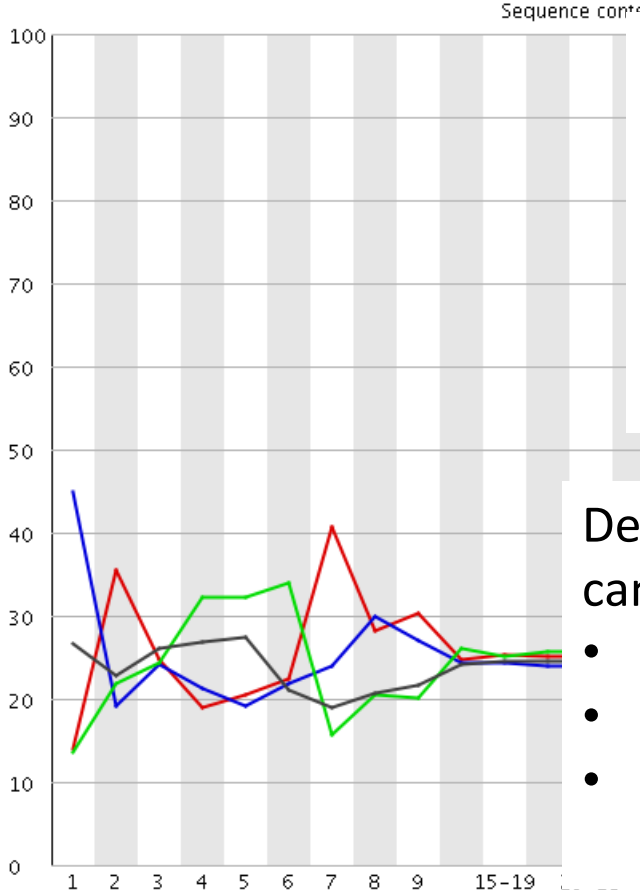
Some issues can occasionally happen with the sequencing instruments during a run. Any sudden drop in quality or a large percentage of low quality reads across the read could indicate a problem at the facility.

FASTQC

Per base sequence content

- Is this plot problematic?

This plots the proportion of each base at each cycle.



Published online 14 April 2010

Nucleic Acids Research, 2010, Vol. 38, No. 12 ***e131***
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358, [2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and [3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010

Deviation from normal base content (equally represented) can be:

- Library quality issues
- Genomes that are very GC biased
- NGS applications that introduce a strong GC bias (such as Bisulfite-seq).

Position in read (bp)

Bisulfite-sequencing (measures DNA methylation) introduce a strong GC bias

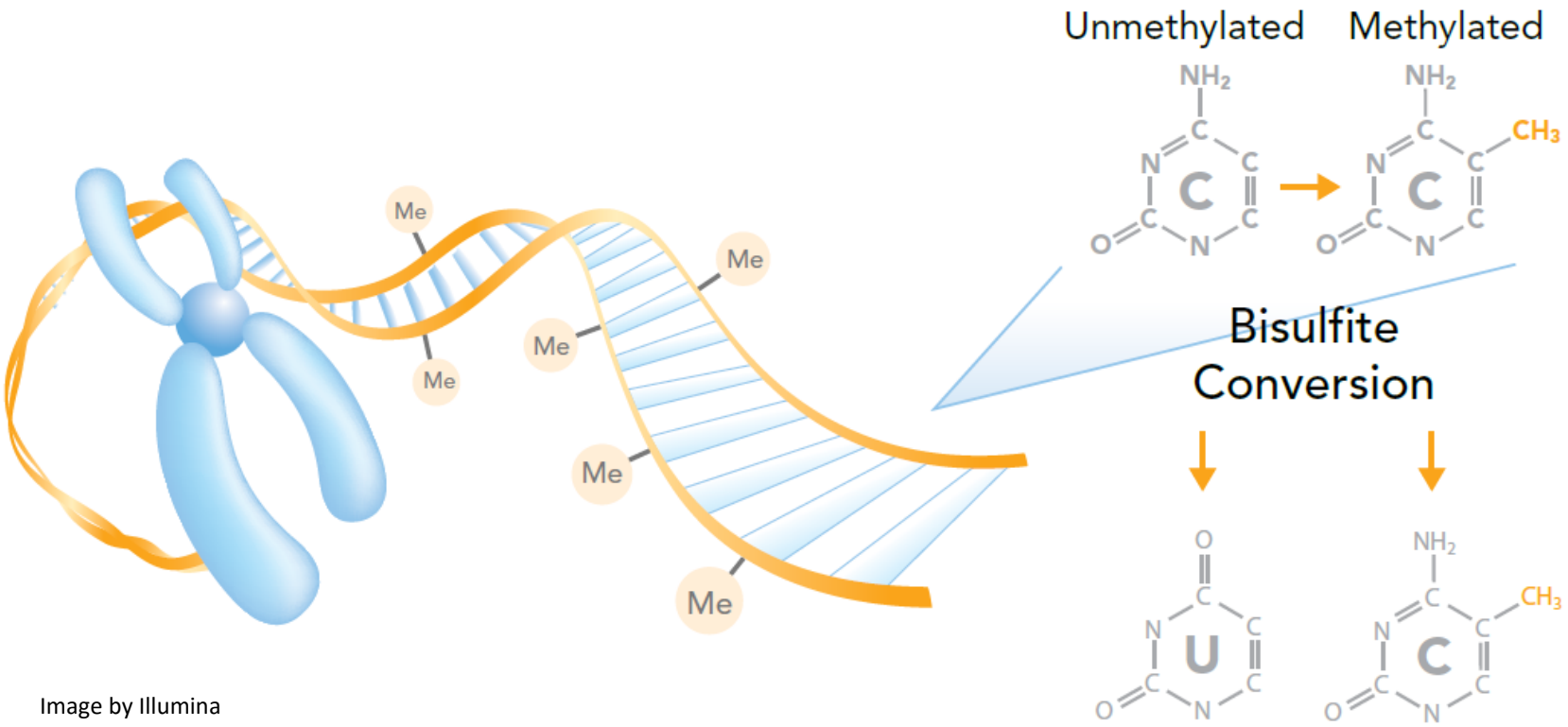
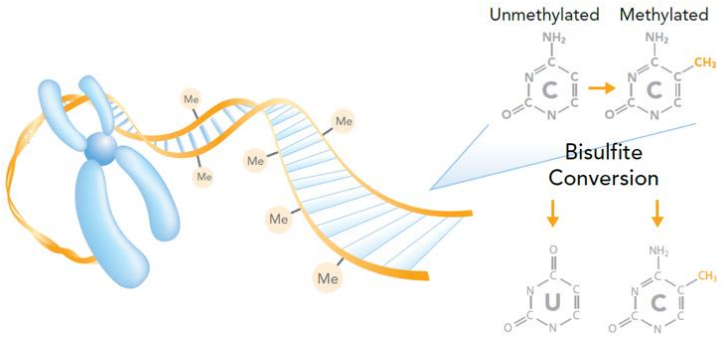


Image by Illumina

Babraham Bioinformatics

Bisulfite Informatics



me me

CCAGTCGCTATAGCGCGATATCGTA

Convert

TTAGTTGCTATAGTGCGATATTGTA

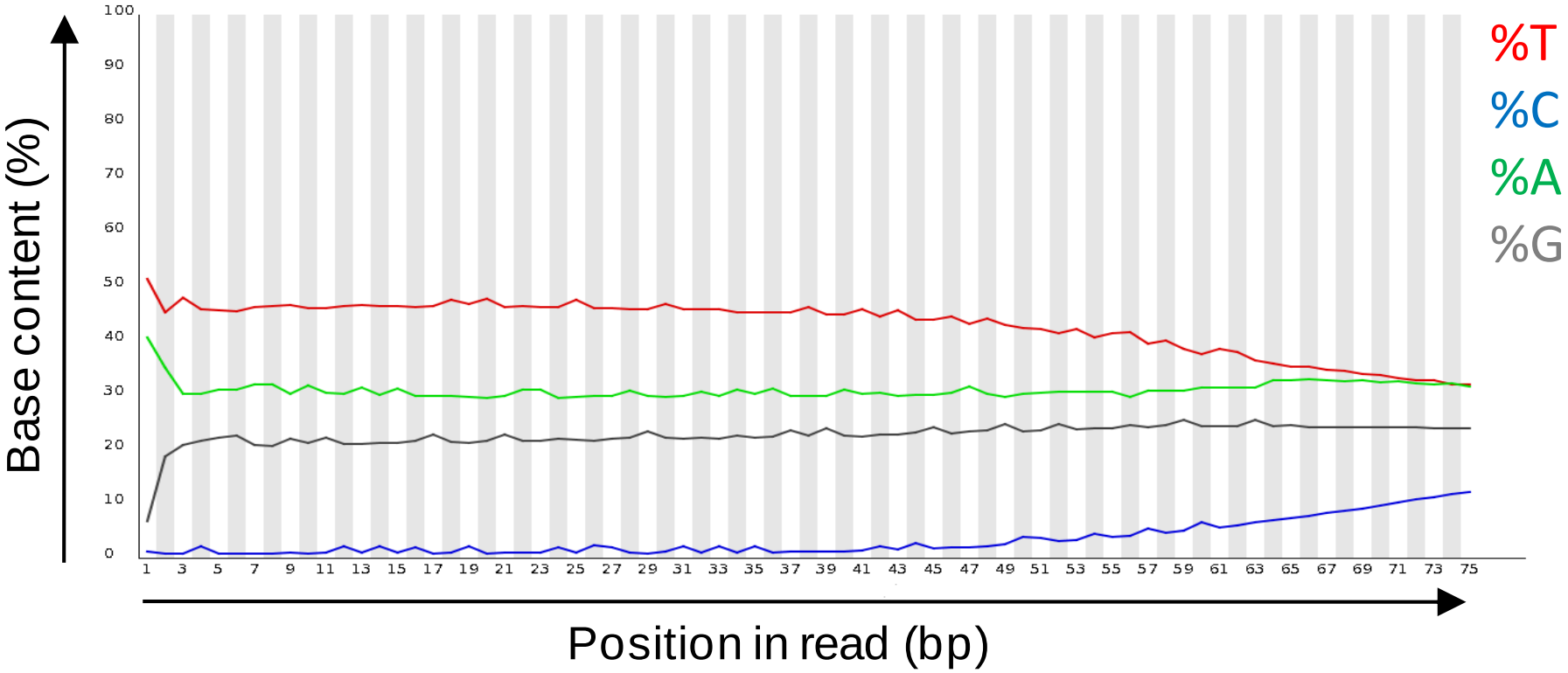
Map

TTAGTTGCTATAGTGCGATATTGTA

. . . CCAGTCGCTATAGCGCGATATCGTA . . .

Babraham Bioinformatics

Common result in Bisulfite -Sequencing



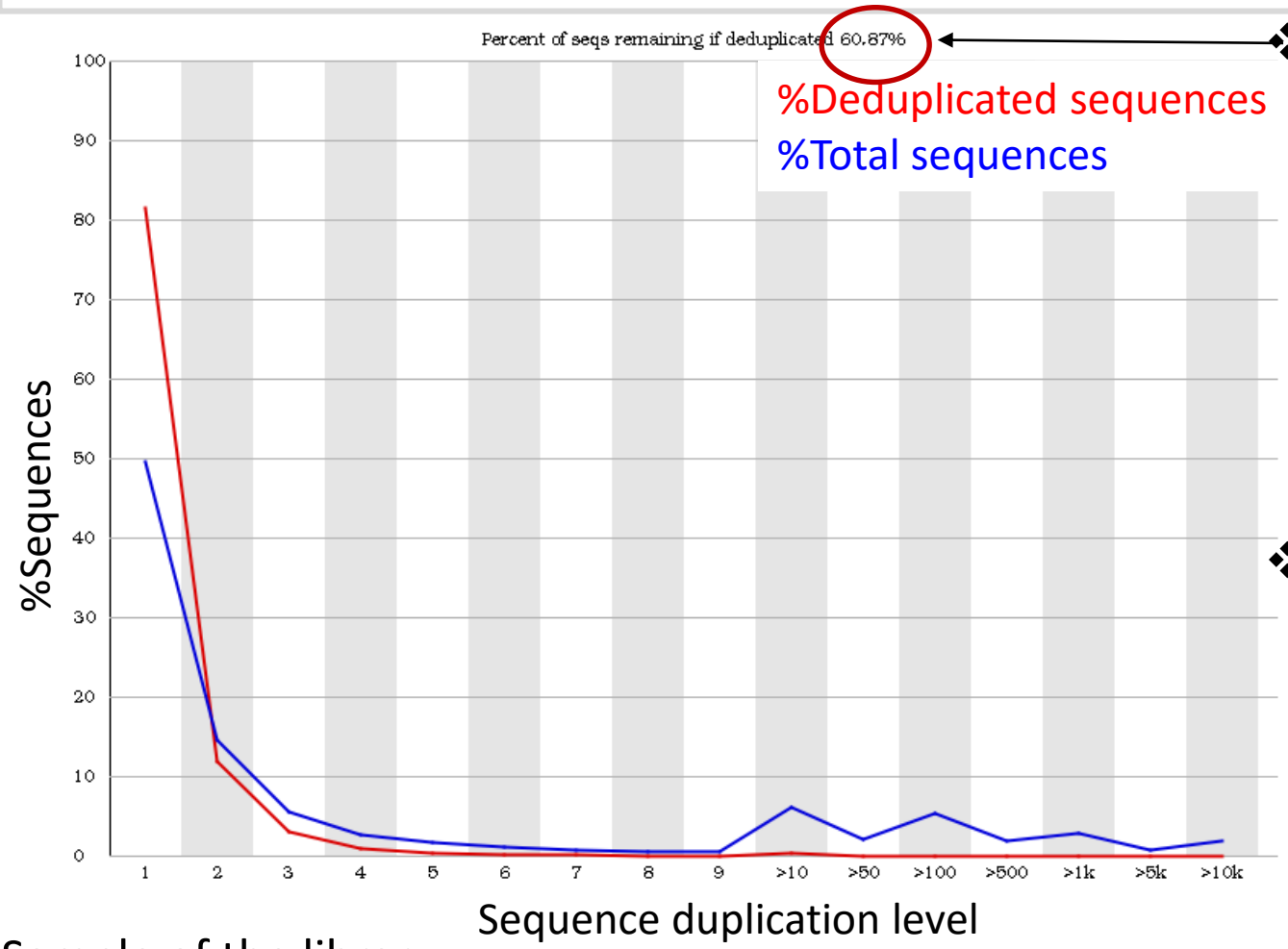
Not observed in 'normal' libraries, e.g. ChIP or RNA-Seq

Developed– Bismark - QC application specific for bisulfite methylation

Babraham Bioinformatics

FASTQC – Sequence Duplication Level

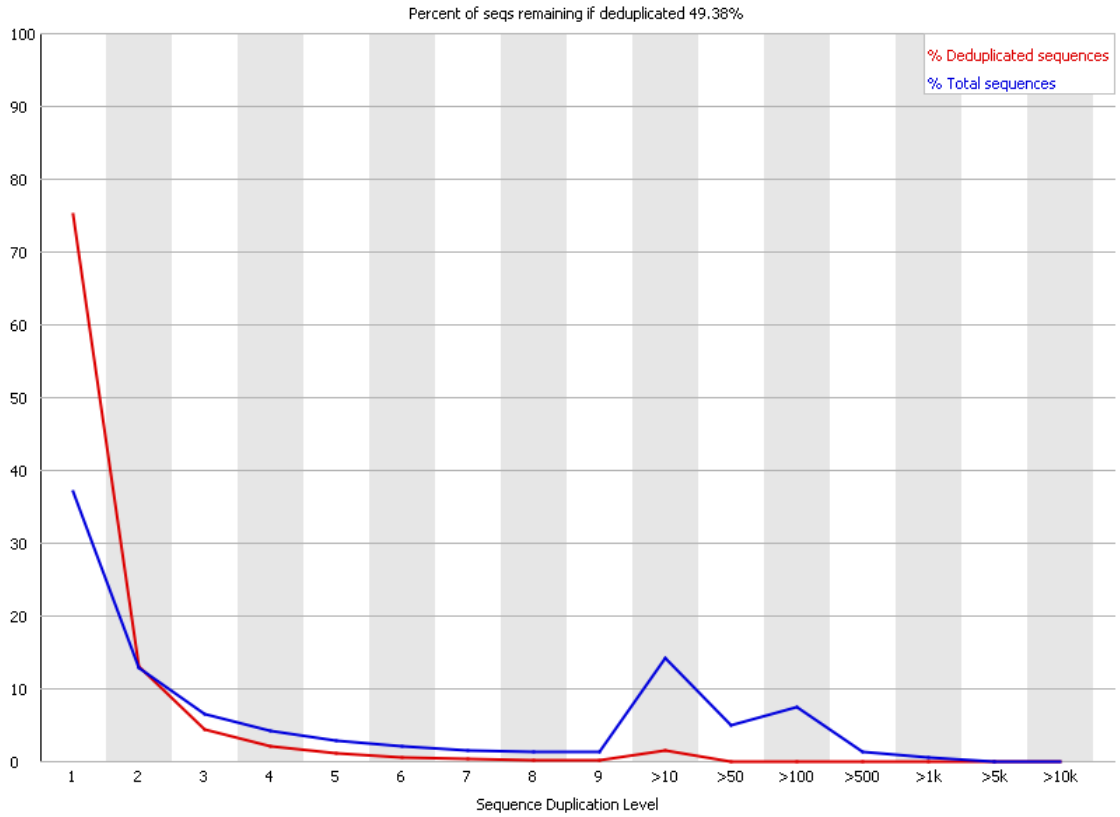
Duplicated read = copy of **exactly the same** sequence



- Sample of the library
- Only look at the first 50bp

- What percentage of the library would remain if you deduplicated it to keep only one copy of every different sequence.
- The two traces show the proportion of the library which comes from sequences with different levels of duplication.

This is a good RNA-Seq library

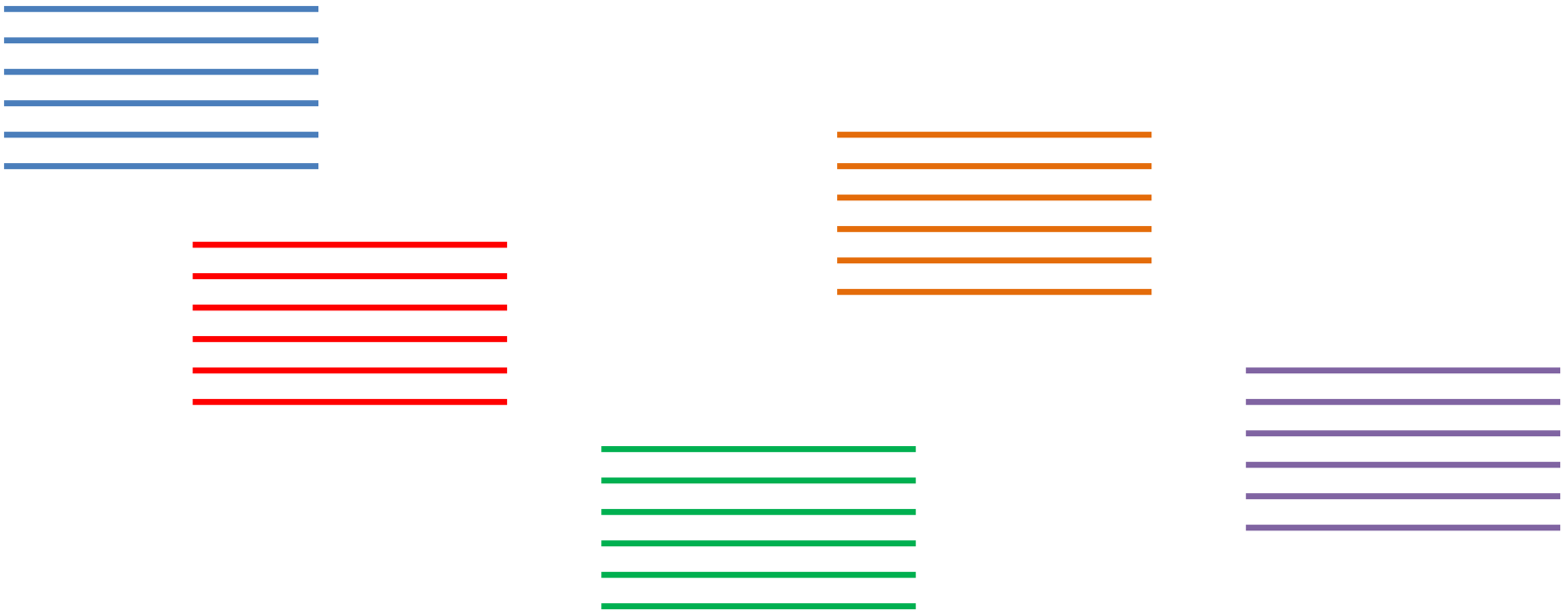


There are very few highly duplicated sequences which might be more indicative of a technical problem with the library.

low duplication levels (2-9 copies) which probably cover most of the ‘normal’ genes

Moderate duplication (10 – 100 copies) which may cover things like rRNA and some repeat sequences

Example
Explaining Deduplicated Calculation



Look at the bin of sequences duplicated 6 times.

We have 5 different sequences each duplicated 6 times.

How many reads do we have in this bin in total

and after deduplication?

Two sources of duplicate reads can be found:

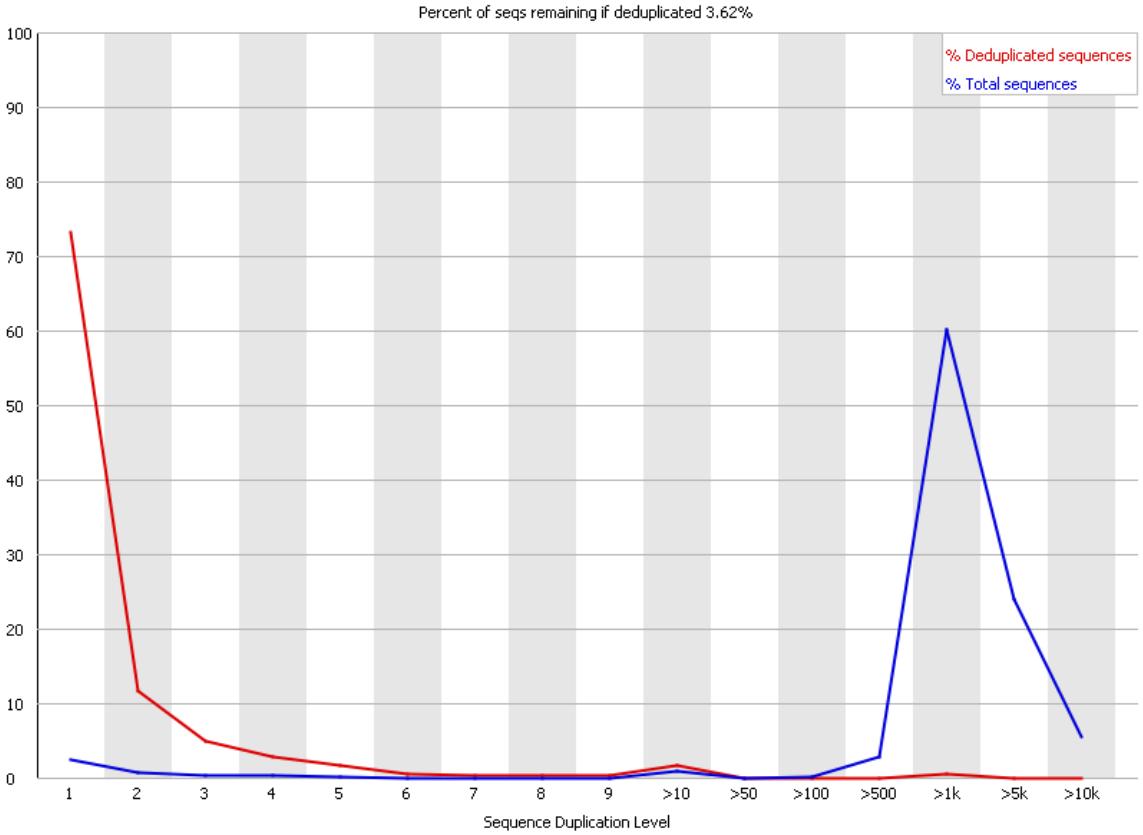
- PCR duplication in which library fragments have been over-represented due to biased PCR enrichment

  It is a concern because PCR duplicates misrepresent the true proportion of sequences in the input.

- Truly over-represented sequences such as very abundant transcripts in an RNA-Seq library

  It is an expected case and not of concern because it does faithfully represent the input.

Examples
Sequence Duplication Plots

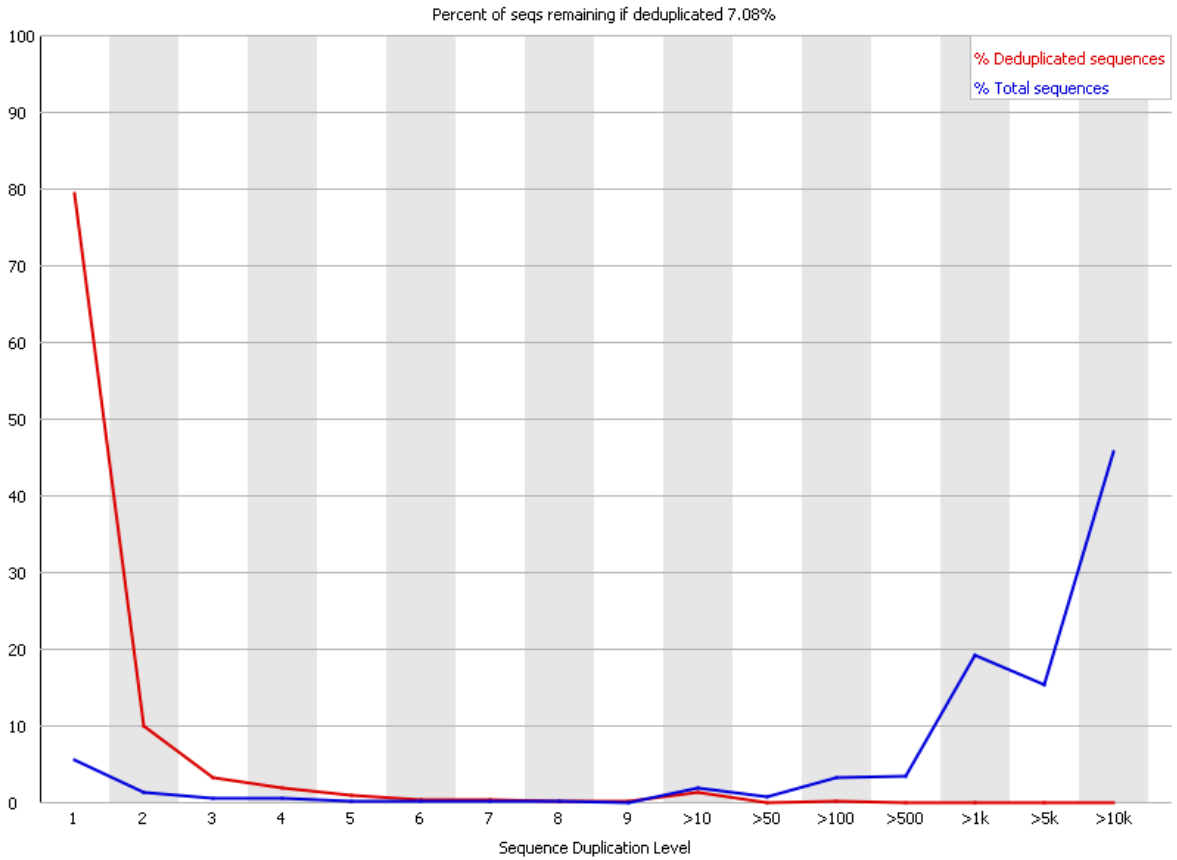


PhiX is a bacteriophage with a well-defined genome sequence of 5386 nucleotides.
PhiX is commonly used as a control for Illumina sequencing runs.

- Many PhiX sequences are present thousands of times due to high coverage.
- The red line shows that when we deduplicate the library the vast majority of sequences come from reads which were present only once in the original library.
- Why do we have sequences present only once?

Examples
Sequence Duplication Plots

What do you think on the quality of this RNA-Seq library?



http://proteo.me.uk/2013/09/a-new-way-to-look-at-duplication-in-fastqc-v0-11/

FASTQC
Over represented sequences

Finding that a single sequence is very overrepresented in the set means:

- It is highly biologically significant
- Indicates that the library is contaminated
- Library is not as diverse as you expected

Sequence	Count	Percentage	Possible Source
AGCCTTTCATCCCTTCTCAACATGAGTAAGAGAAATACGGGTAGGAAATC	6399	0.8001210372189448	No Hit
AGCCTCTCCGAGCGCGTTTCCTAAAAAGGGGGAGTCCTCATTTAAAAAAA	3452	0.43163272706357203	No Hit
ATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCGCGAAATCTCGTATG	2061	0.25770424405504694	TruSeq Adapter, Index 6 (97% over 35bp)
ATGACGCTCTTCTTGAGCGTCTTTTGTCTGCCGCCTCTGTGCGGCTTTTT	1277	0.1596740997856841	No Hit
ATGACGCCTCTCTTTTCGGCGCTGTTTTGGAGCTTCAAAAAATGGCTGGG	1030	0.1287896028028619	No Hit
ATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCGCGAAAACTCGTATG	998	0.12478837242452054	TruSeq Adapter, Index 6 (97% over 35bp)
GCCCCCTTAACATTTTCTTAACAATTTCTTAACAATCCCTACATAGTTAT	804	0.10053091325582617	No Hit

For each overrepresented sequence the program will look for matches in a database of common contaminants

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/9%20Overrepresented%20Sequences.html

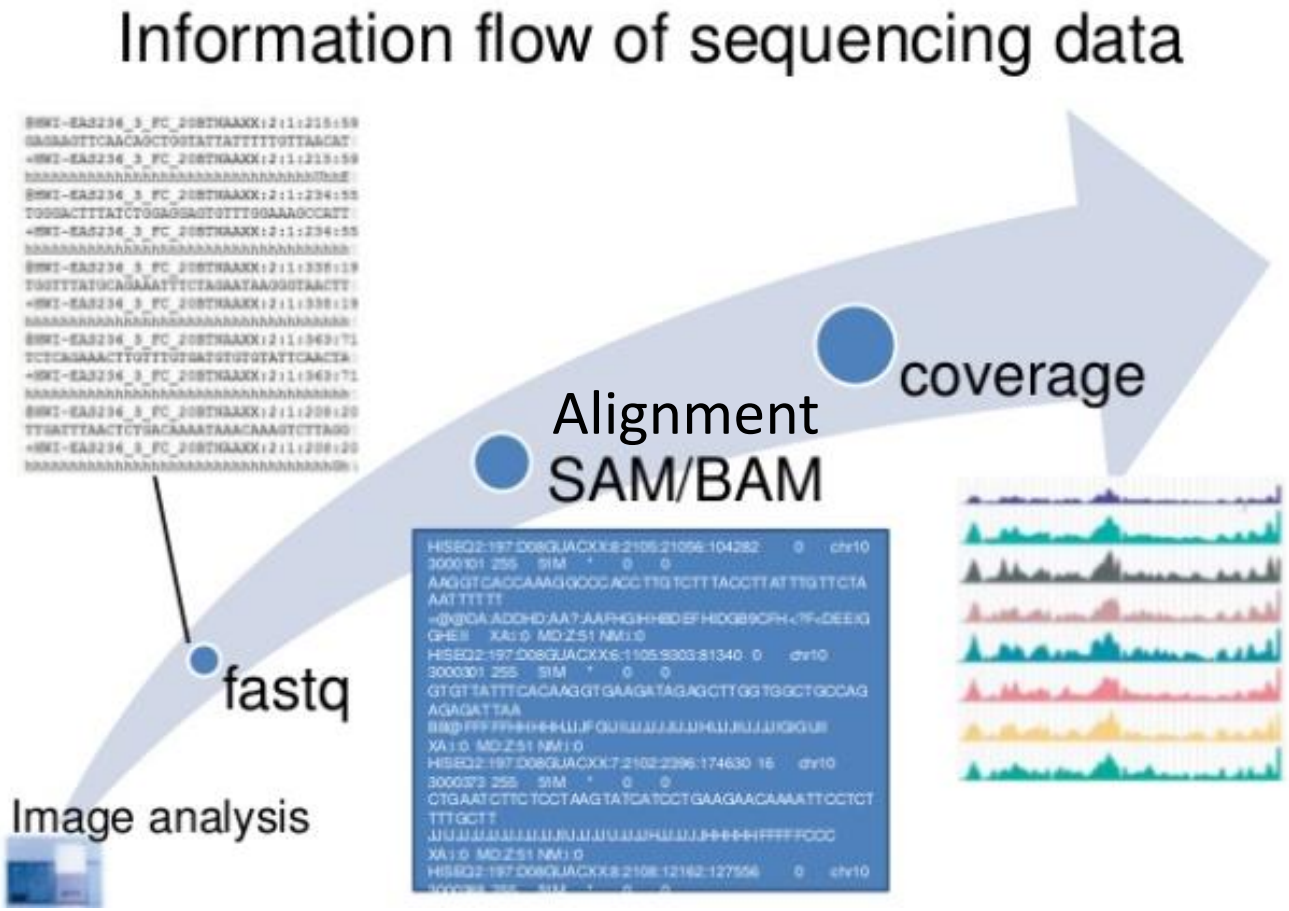
Summary

The more time and effort you spend on QC

the better quality

your results and conclusion will be.

Questions?


Information flow of sequencing data
fastq
Alignment
SAM/BAM
coverage
Image analysis