מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

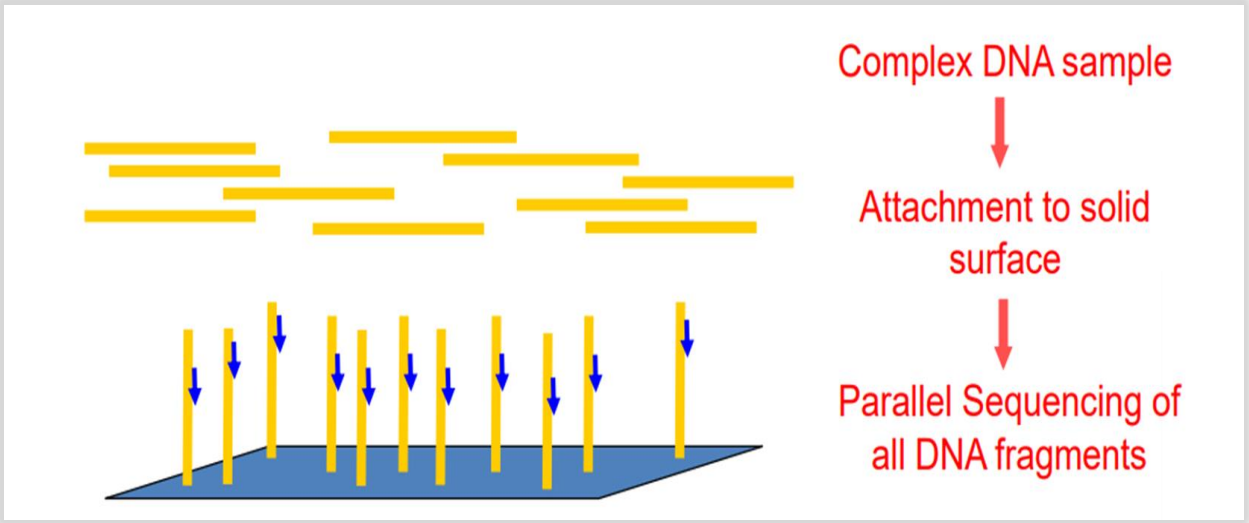
Illumina Primary Analysis Pipeline & Quality Control

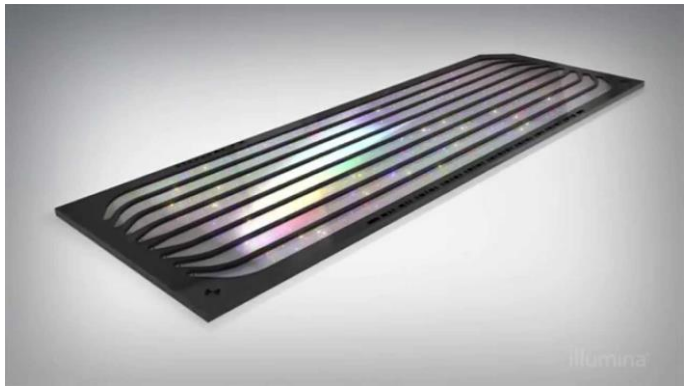
Noa Wigoda

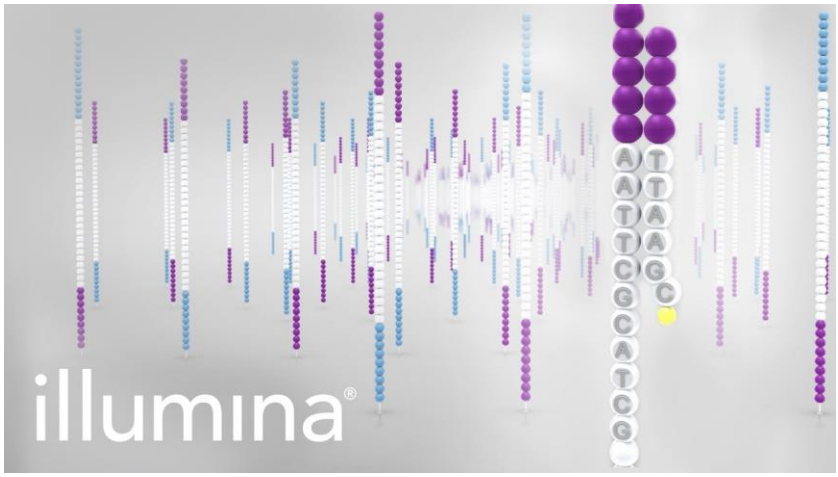
19.11.19

An Introduction to deep-sequencing analysis for biologists

NGS – What is the essence of this technology?



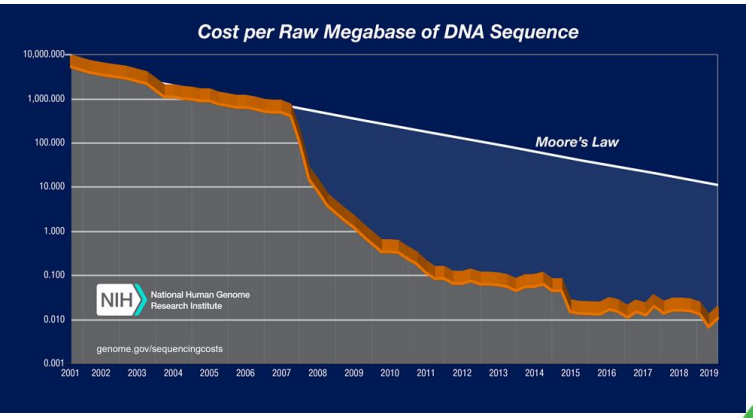




Why Illumina?

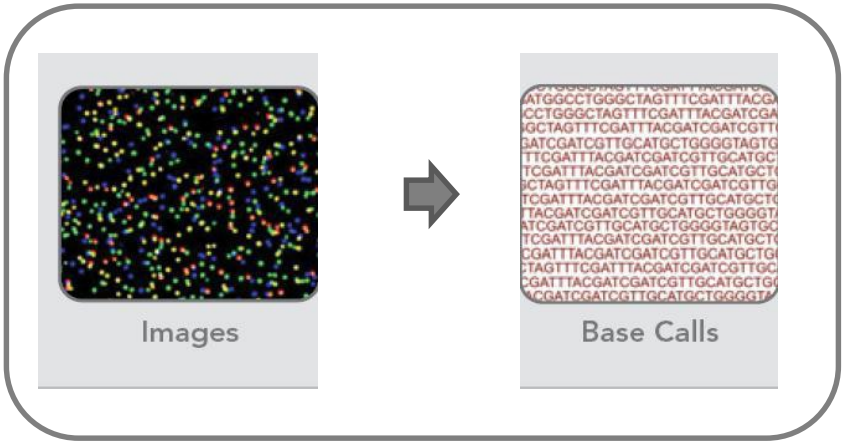
- Responsible for generating more than 90% of the world's sequencing data
- Several machines in the Weizmann Institute

  LSCF, G-INCPM and several labs



Topics to be discussed

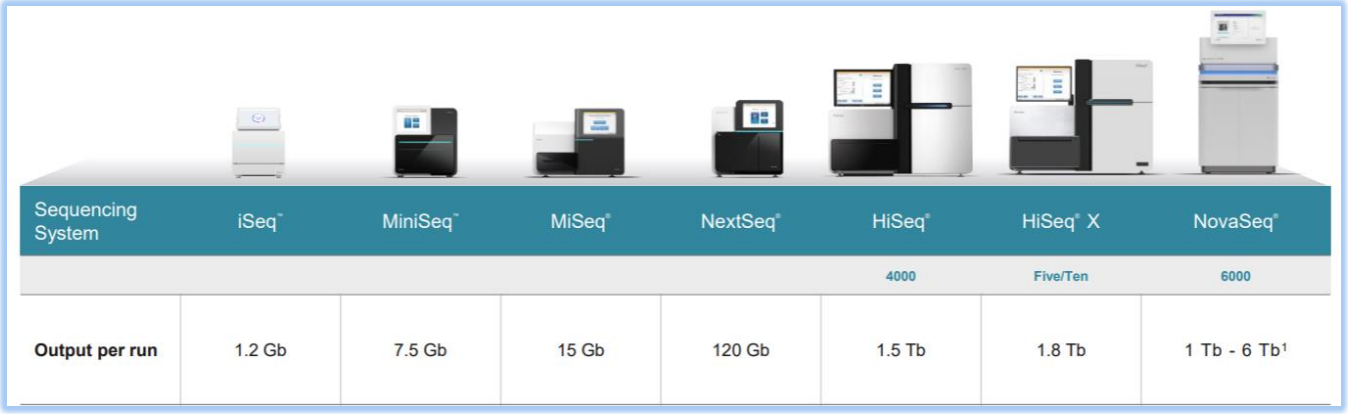
- Illumina sequencing technology advances
- Illumina primary analysis



- Sequencing quality control (QC)

Illumina sequencing technology advances

- Machines



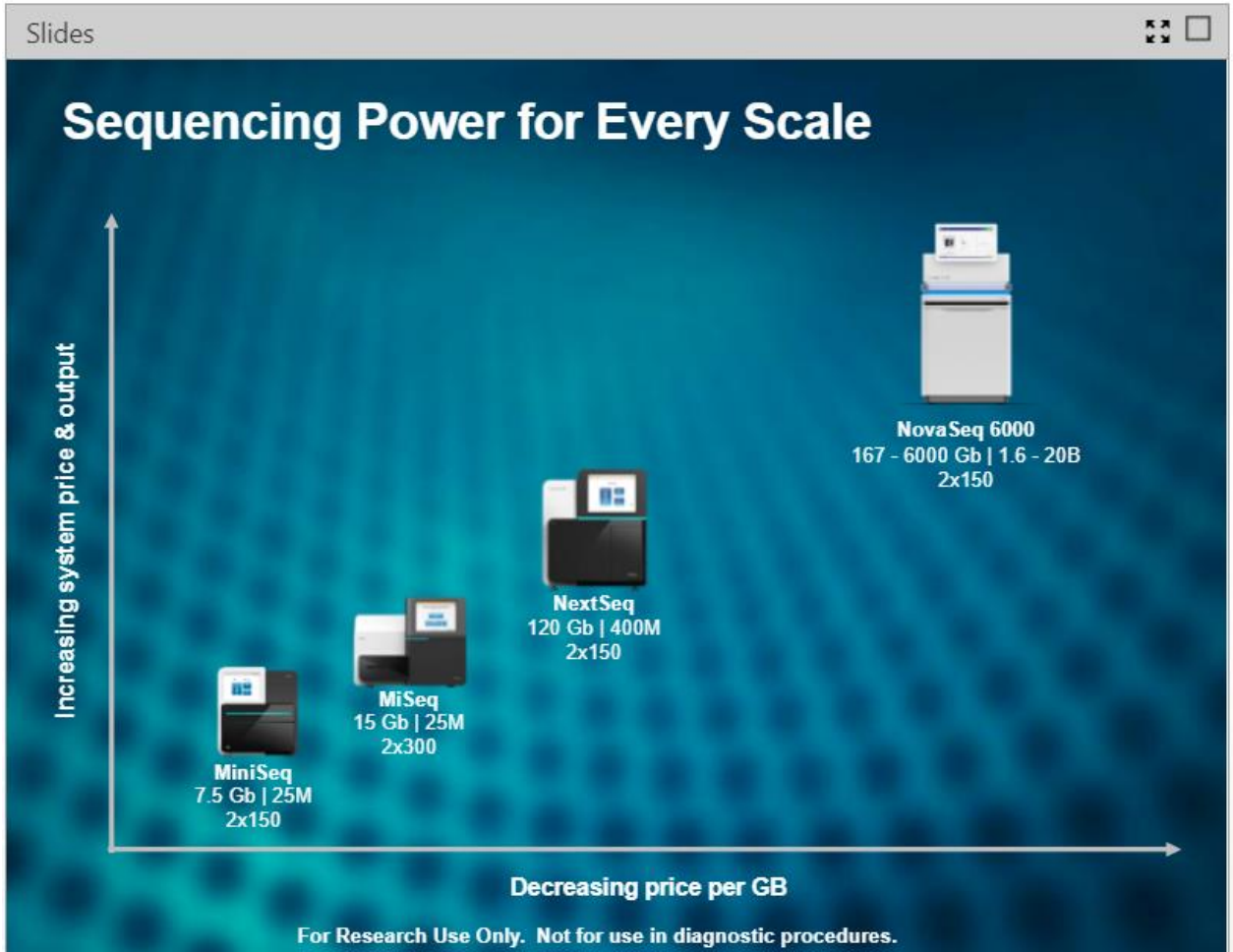
Sequencing System	iSeq™	MiniSeq™	MiSeq®	NextSeq®	HiSeq®	HiSeq® X	NovaSeq®
					4000	Five/Ten	6000
Output per run	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb[1]

- Flow cells
- Chemistry

Progress: a remarkable increase in data and decrease in price

Illumina Webinar

Next-generation solutions for the next era of sequencing



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

Illumina machines in Weizmann



Product	MiSeq	NextSeq	NovaSeq
Cartridge for reagents (greater laboratory efficiency) 	No	Yes	Yes
Color SBS chemistry	4-colour	2-colour	2-colour
Two flow cells can be run independently	No	No	Yes

SBS - sequencing by synthesis

NovaSeq

Able to:

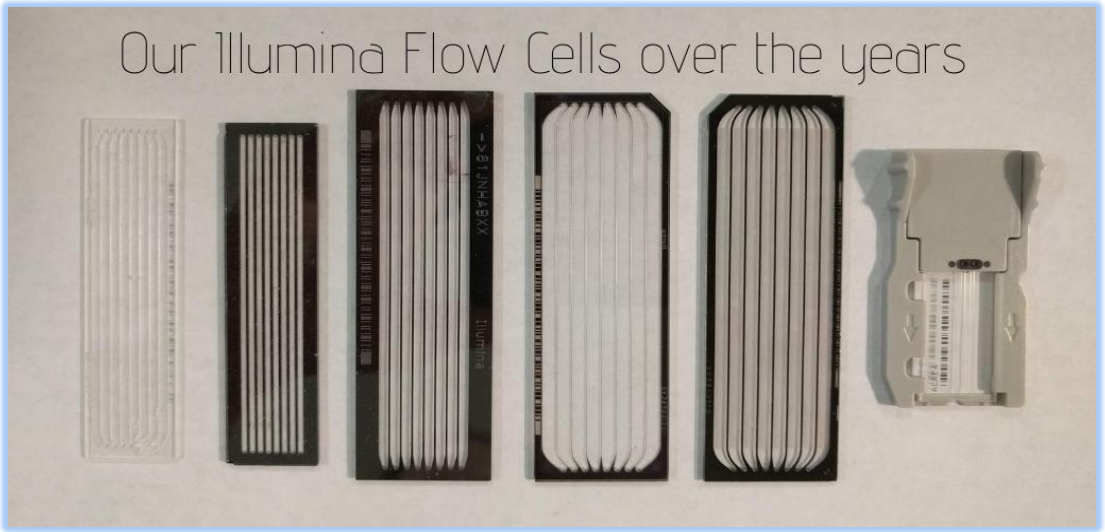
- Sequence 48 human genomes within 2 days.
- Run 500 RNA-seq samples on one @illumina S4 NovaSeq flow cell, this would bring sequencing down to just $5-10 per sample!

Need to multiplex sequences with unique-at-both-ends dual-indexed adapters.

NovaSeq will be available in Weizmann in a few weeks

Illumina sequencing technology advances

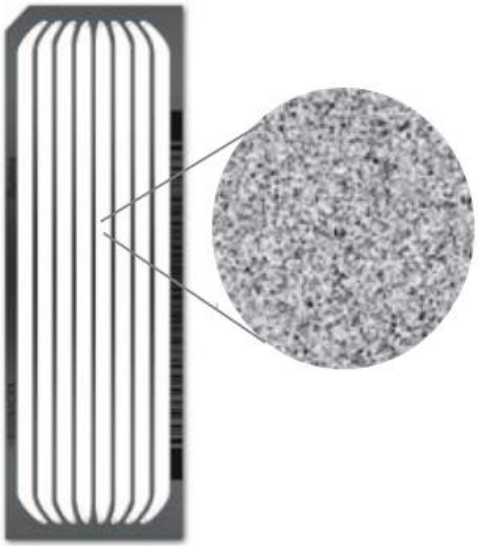
- Machines
- Flow cells
- Chemistry



Flow cell architecture

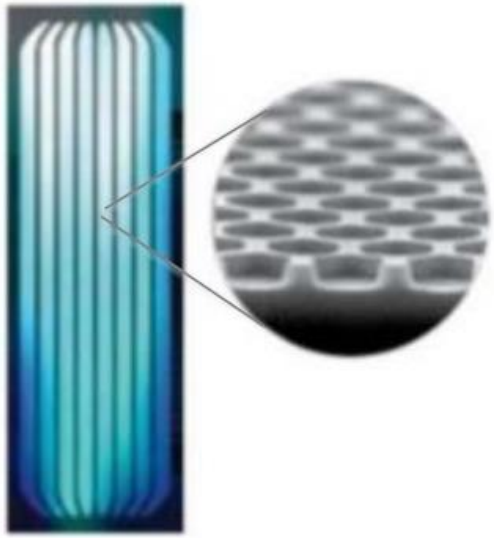
Random Flow Cell

- HiSeq 2500, MiSeq, NextSeq, MiniSeq
- Randomly spaced clusters
- Variable Insert Sizes
- Lower Duplication Rates



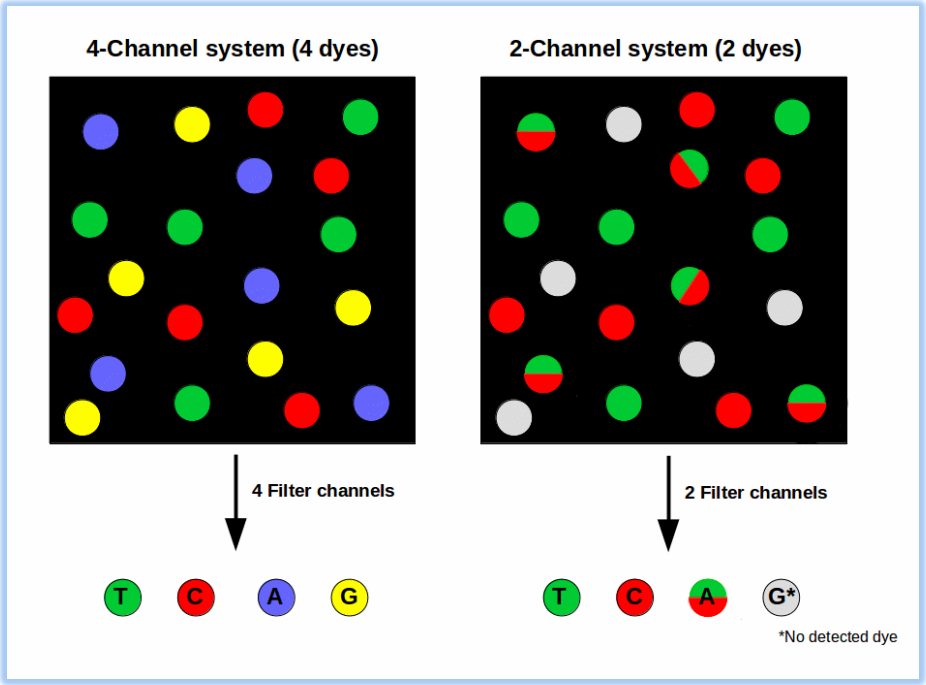
Patterned Flow Cell

- HiSeq 3K/4K/X, NovaSeq, iSeq 100
- Defined size and spacing
- Increased Cluster density
- Simplified imaging

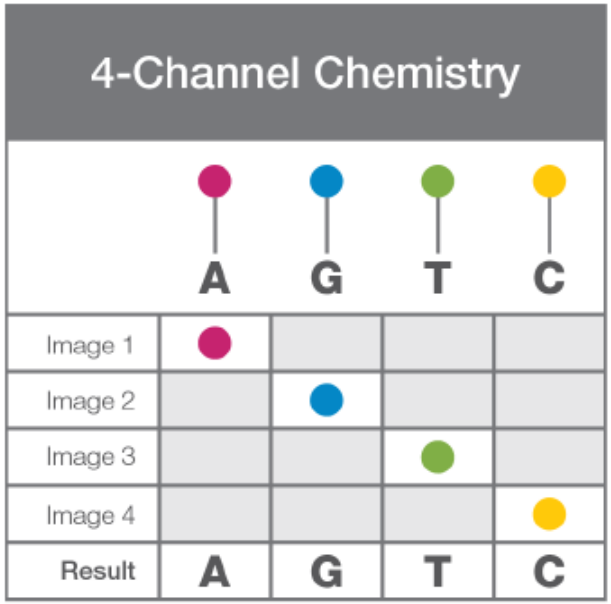


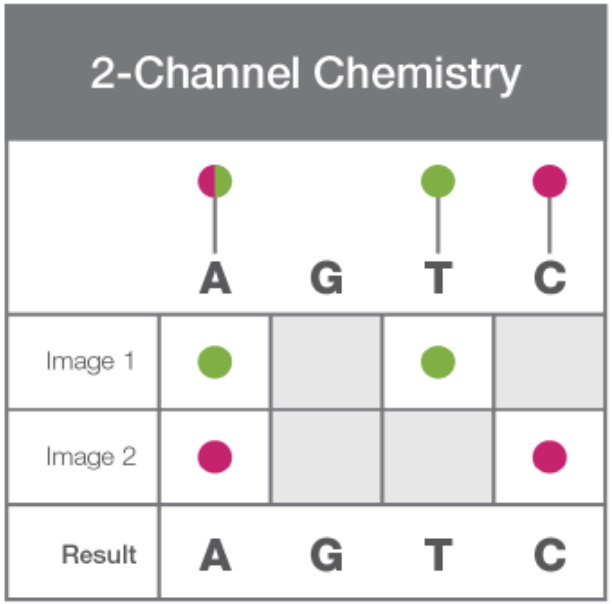
Illumina sequencing technology advances

- Machines
- Flow cells
- Chemistry



Illumina chemistry comparison





Four-channel SBS

- Bases are identified using four different fluorescent dyes, one for each base and four images per sequencing cycle

Two-channel SBS

- Simplified nucleotide detection by using two fluorescent dyes and two images to determine all four base calls

Four Channels SBS:

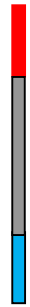
- MiSeq

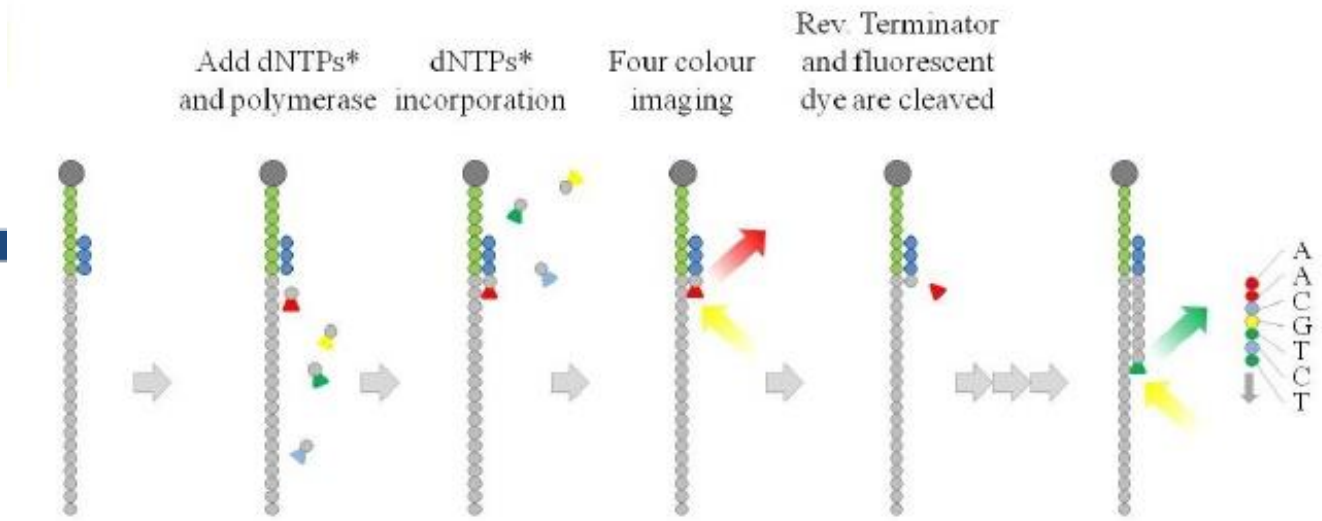
Two Channels SBS:

- MiniSeq, NextSeq, NovaSeq
- Accelerates sequencing and data processing times

Illumina SBS workflow

cDNA
(0.05-1.0 µg)





synthesis utilizes reversible terminator-based method in order to detect single bases as they are incorporated into the DNA.

The base incorporated is detected upon laser excitation of the base fluorophore.
Images are collected to record the emission.

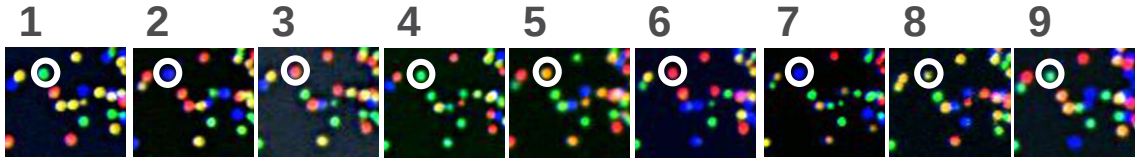
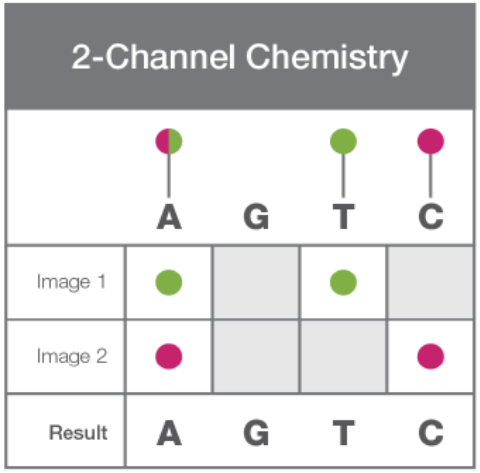
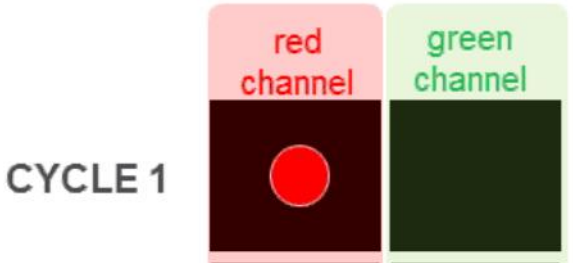


Image Acquisition

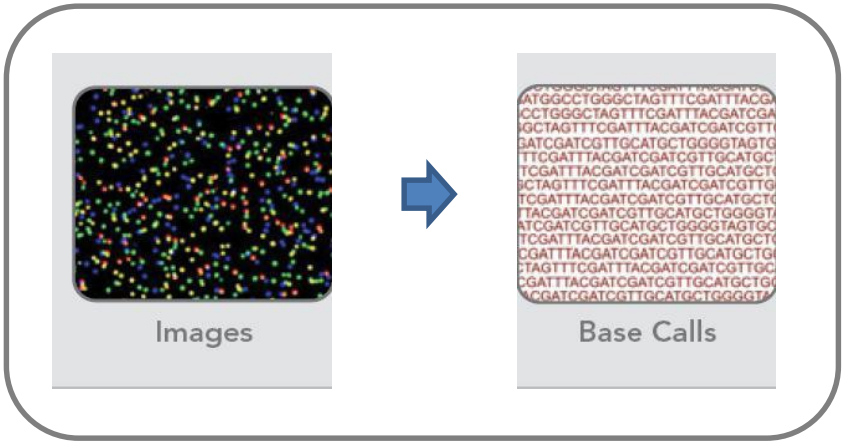
Using 2 dyes: 2-channel method





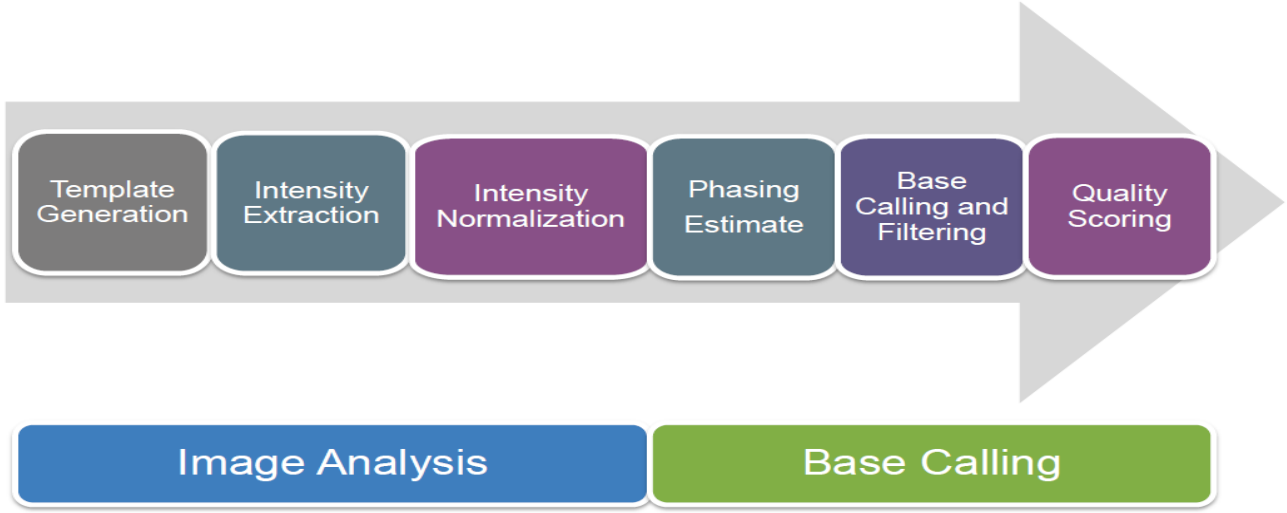
Topics to be discussed

- Illumina sequencing technology advances
- **Illumina primary analysis**



- Sequencing quality control (QC)

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.

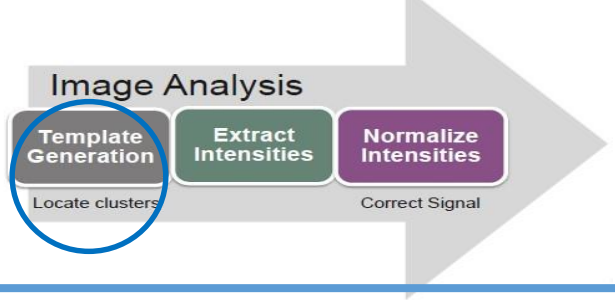
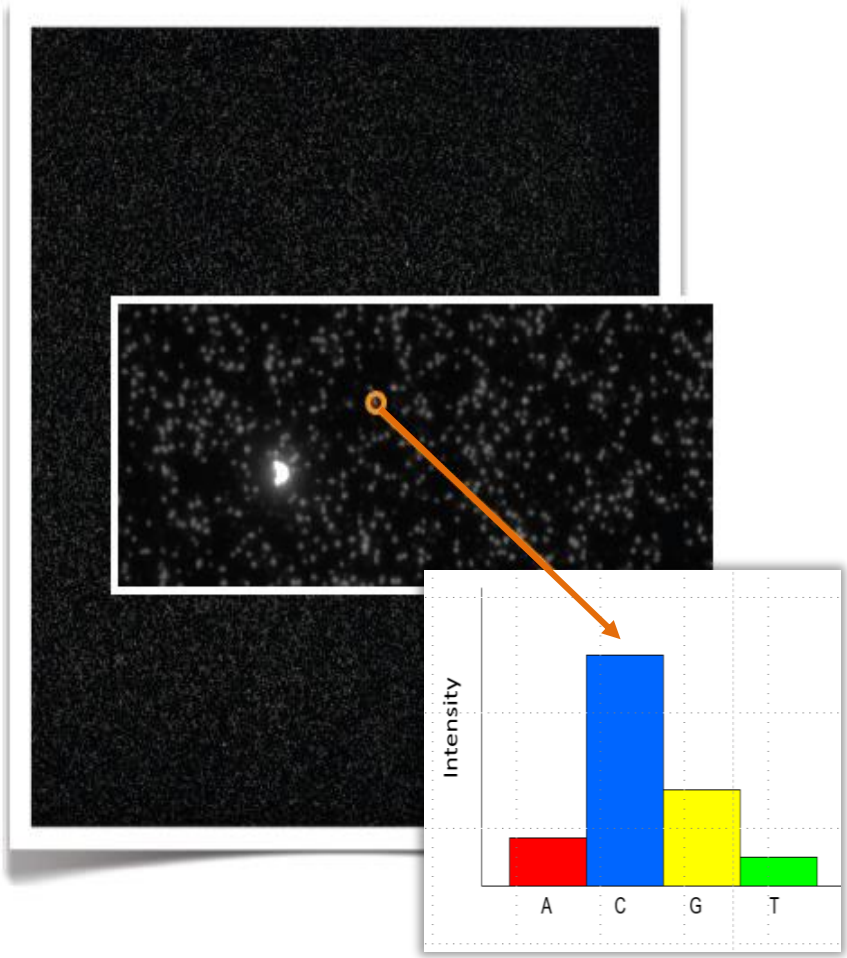
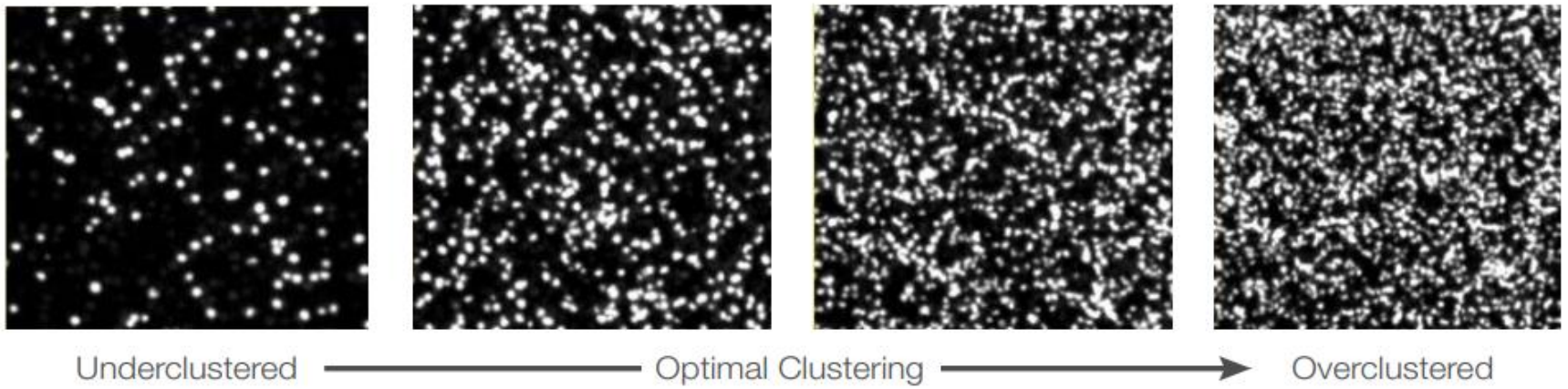


Image analysis steps

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



Cluster density influences sequence quality



Under clustering:

- Maintains high data quality.
- Results in lower data output.

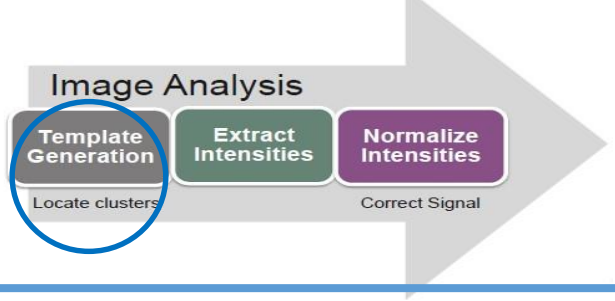
Over clustering can lead to:

- Poor run performance
- Lower Q30 scores
- Possible introduction of sequencing artifacts

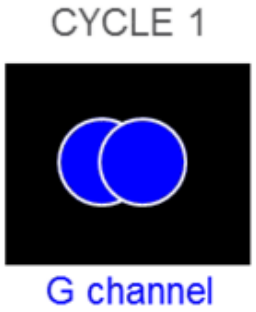
Counterintuitively:

lower total data output.

https://*emea.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf*



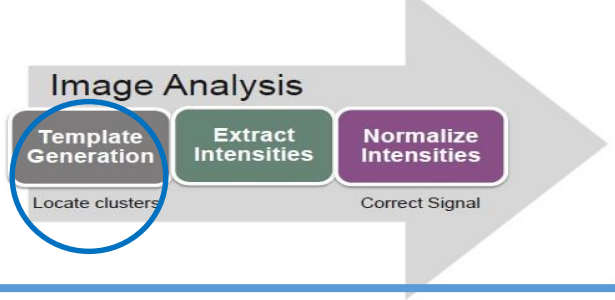
Detection of cluster position



What if we have two adjacent clusters which are both G in cycle 1

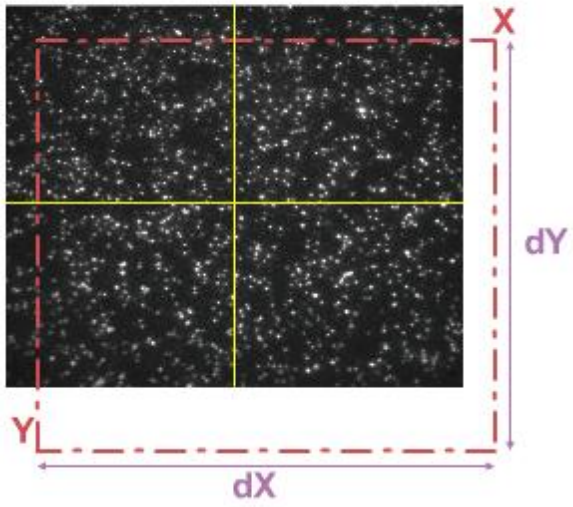
They will appear as a single cluster

Can you suggest a solution?



Aligning the Images

Default Offsets



Offsets/offsets.txt

X	Y	dx	dy	
0.00	0.00	0.00000	0.00000	A
0.32	1.41	0.00069	0.00068	T
-0.01	1.82	-0.00123	-0.00125	C
0.14	1.59	-0.00097	-0.00092	G

- Images of the various flourophores can be slightly offset from each other due to their different optical path

- Offset matrix is computed to allow shift, shrink and stretch

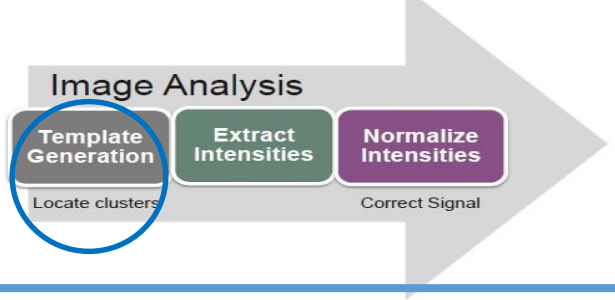
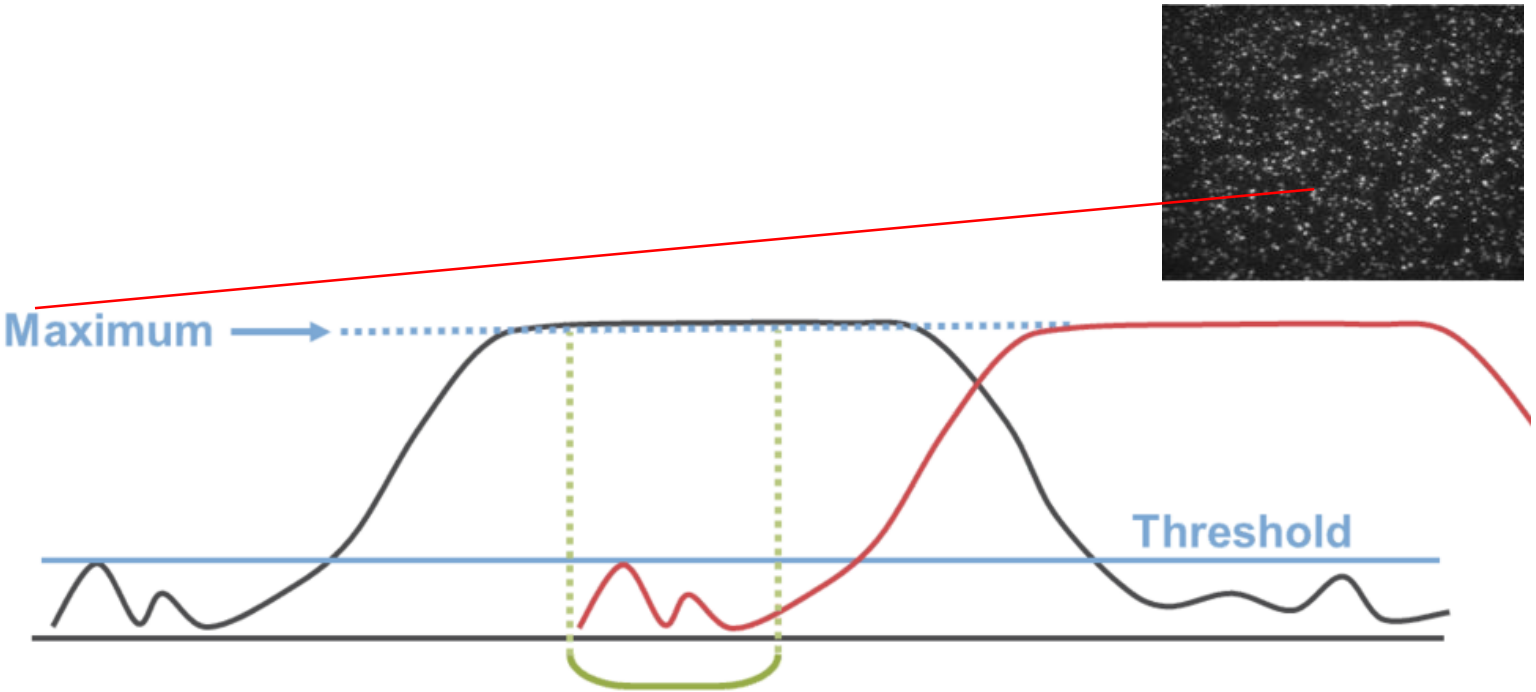
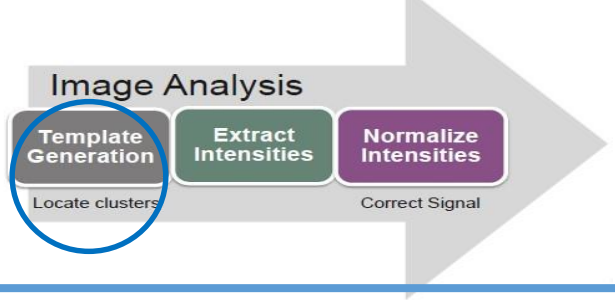


Image analysis algorithm



Defining cluster location within the image:

- Looking at the first cycle, the program scans the images to detect regions of intensity above background
- Only a slice of this region is used for cluster definition, in order to prevent overlap with a near by cluster



Template generation

Template Generation

HiSeq 2500, NextSeq, MiniSeq, MiSeq

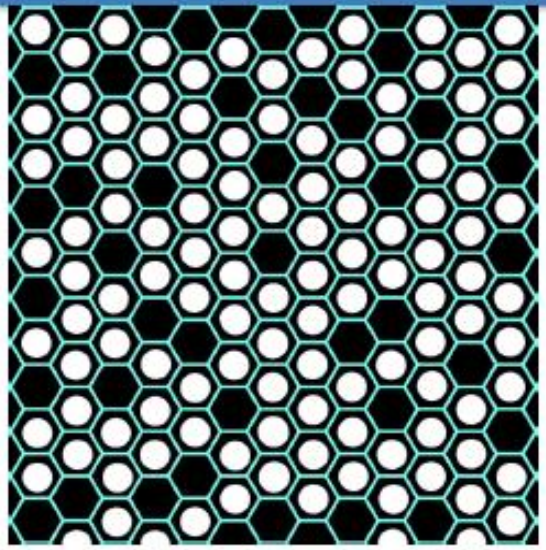
Identify the location of every cluster and create a map

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

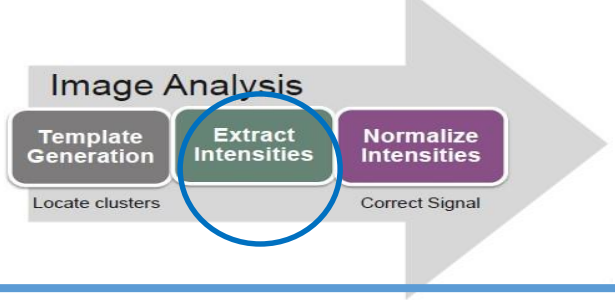


/black-light-studio.deviantart.com/art/Free-Hexagon-pattern-02-371945610

Identifying cluster location is based on the first 5 cycles.

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

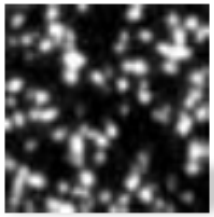
illumina



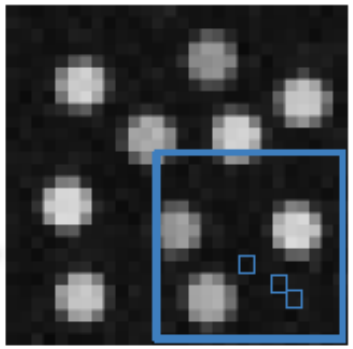
Intensity extraction

Background Subtraction

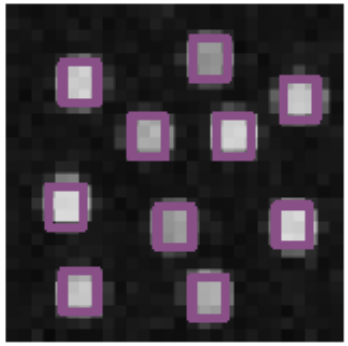
Compute background → **Compute signal for each cluster** → **Subtract background from each cluster**



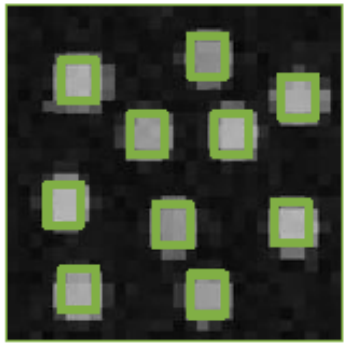
Clusters are not perfect circles. RTA sees them as shown



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

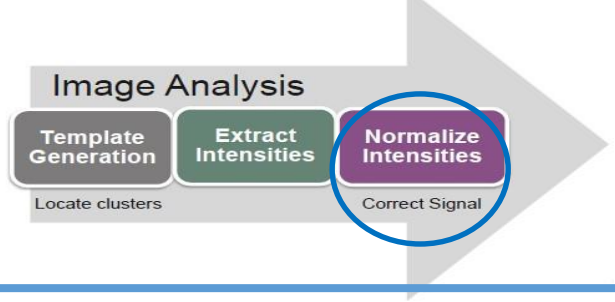


Cluster intensity is extracted from the brightest part of each cluster



Background is subtracted from signal of each cluster

illumina®



Intensity correction

1. **Cross talk correction:** emission spectra of the four dyes overlaps

2. Normalization: scaling factor to make intensities equivalent

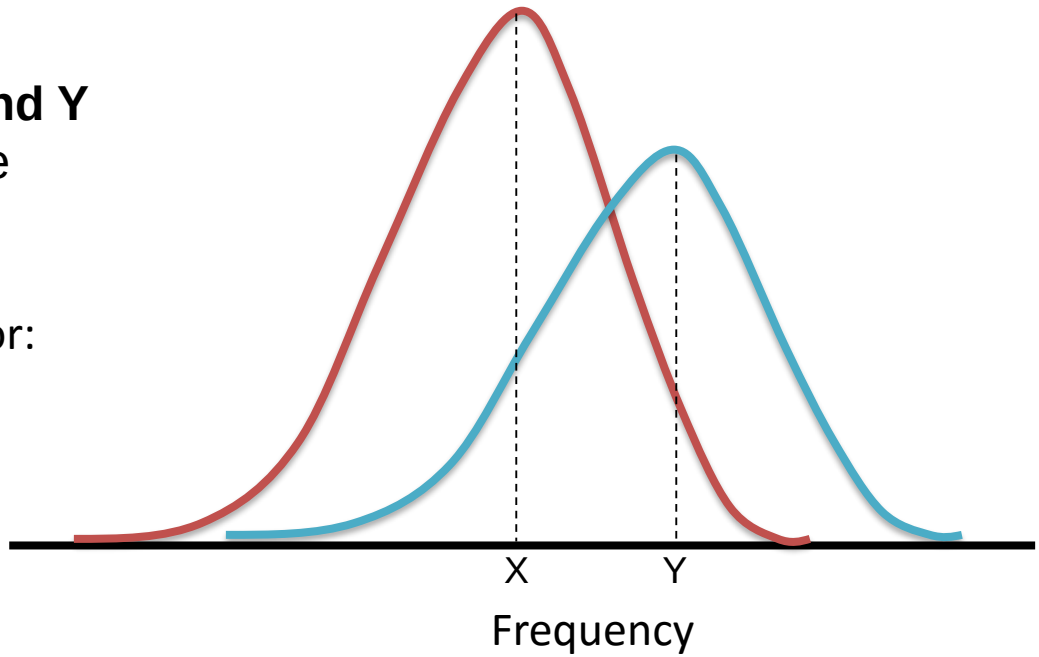
Emission spectra plot of dyes X and Y

As seen here – upon excitation in the frequency of dye X, there is some emission from dye Y

A strong correlation of the intensities for:

- A and C
- G and T

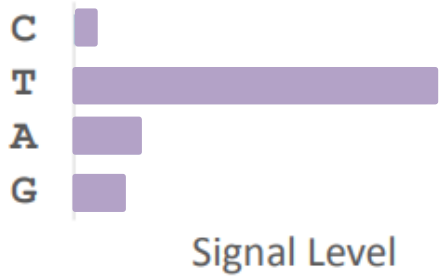
Due to similar emission spectra of the fluorophores



A correction matrix is calculated by data of cycles 1-4

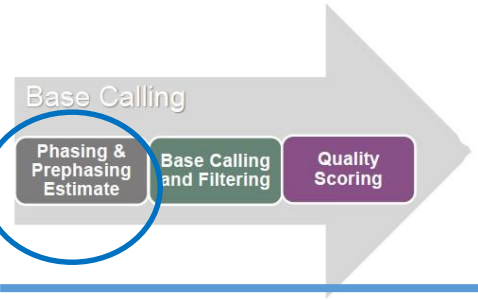
https://support.illumina.com/content/dam/illumina-support/courses/MiSeq_Imaging_and_Base_Calling/story_html5.html

Goal (simplified): identify base



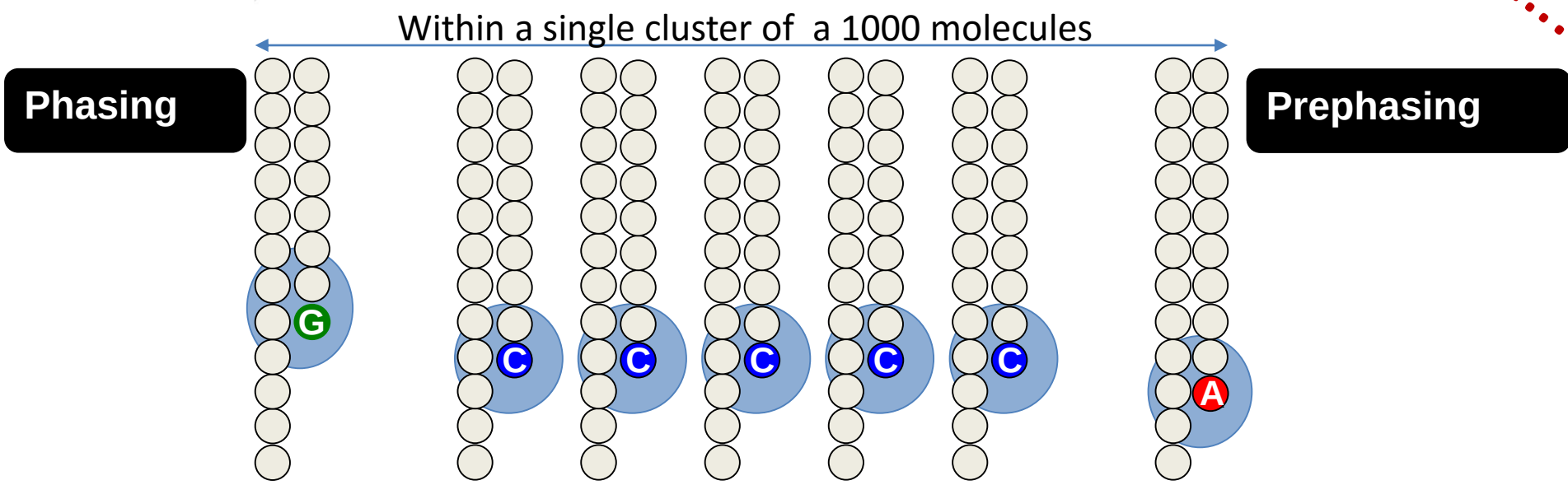
Call = T

Confidence = High

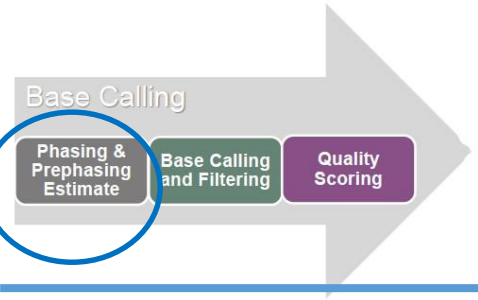


Base calling
Phasing/ prephasing correction

Advanced…



- The rate of the prephasing and phasing is calculated after 12 cycles (lower than 0.1% per cycle is OK)
- Requires a sample with a random, balanced base composition and therefore is usually done on a known predefined sequence – our control
- Phasing is biased due to a lower removal rate of T fluorophores
- The phasing is the reason for the decay in fluorescent signal intensity with each cycle



Base calling
Phasing/ prephasing correction

within a single cluster of thousands of strands

Phasing **Prephasing**

T A A A A A C

illumina

Empirical Phasing Correction

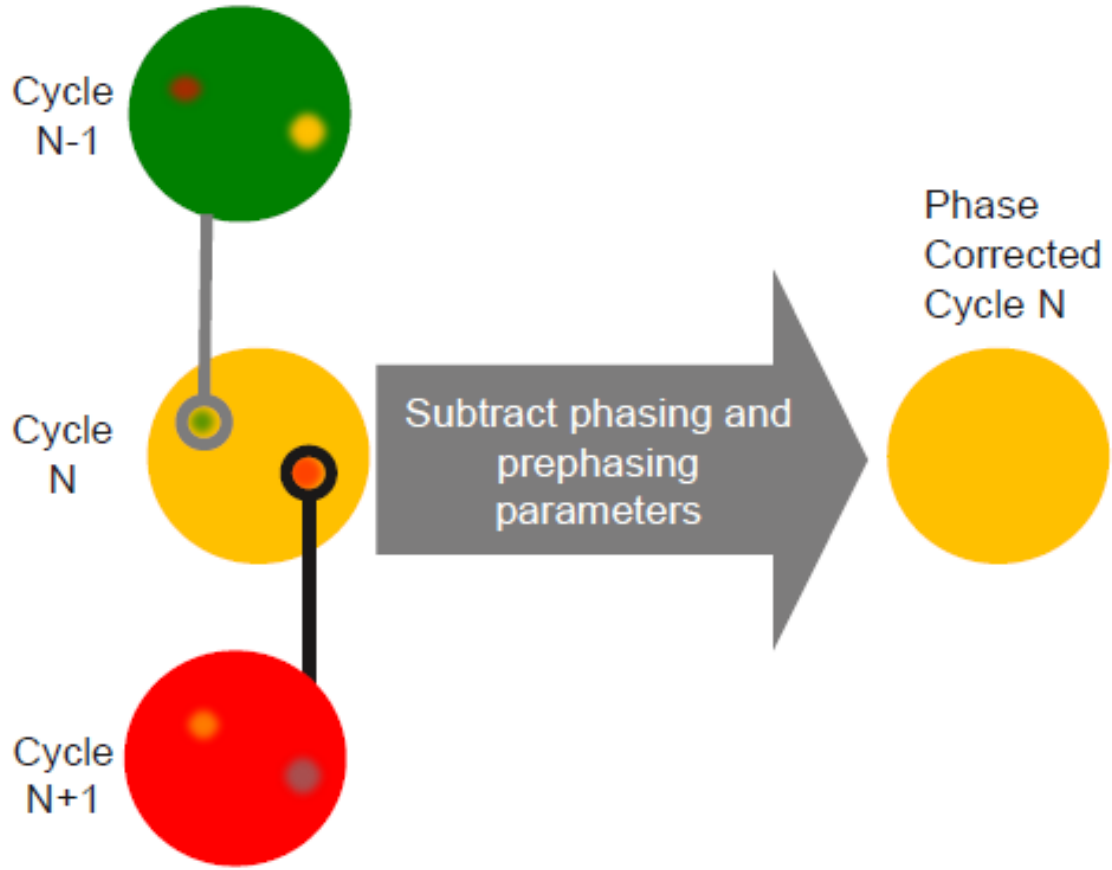
Advanced…

Phasing Correction Parameter

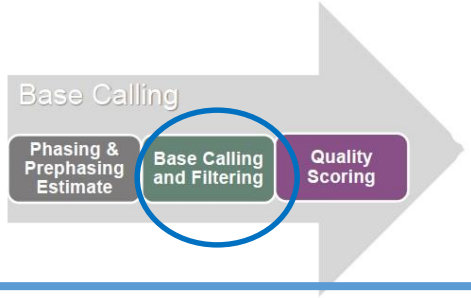
- How much signal of the previous base is present?

Prephasing Correction Parameter

- How much signal of the next base is present?



illumina

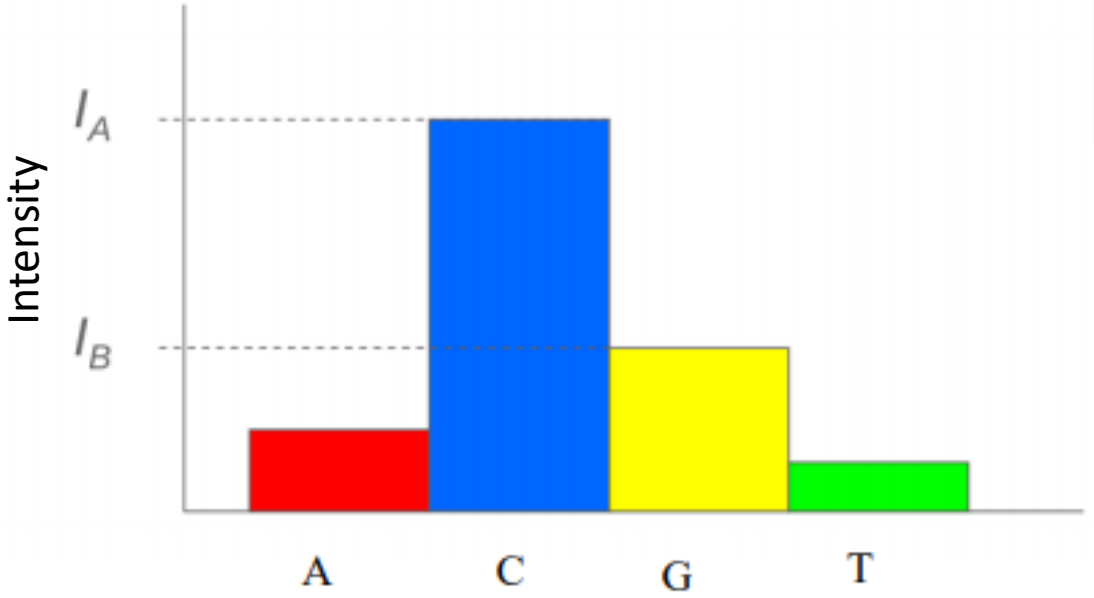


4 color base calling

Chastity is defined as the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities.

CHASTITY formula:

$$C = \frac{I_A}{I_A + I_B}$$



$$Chastity = \frac{Intensity\ (C)}{Intensity\ (C) + Intensity\ (G)}$$

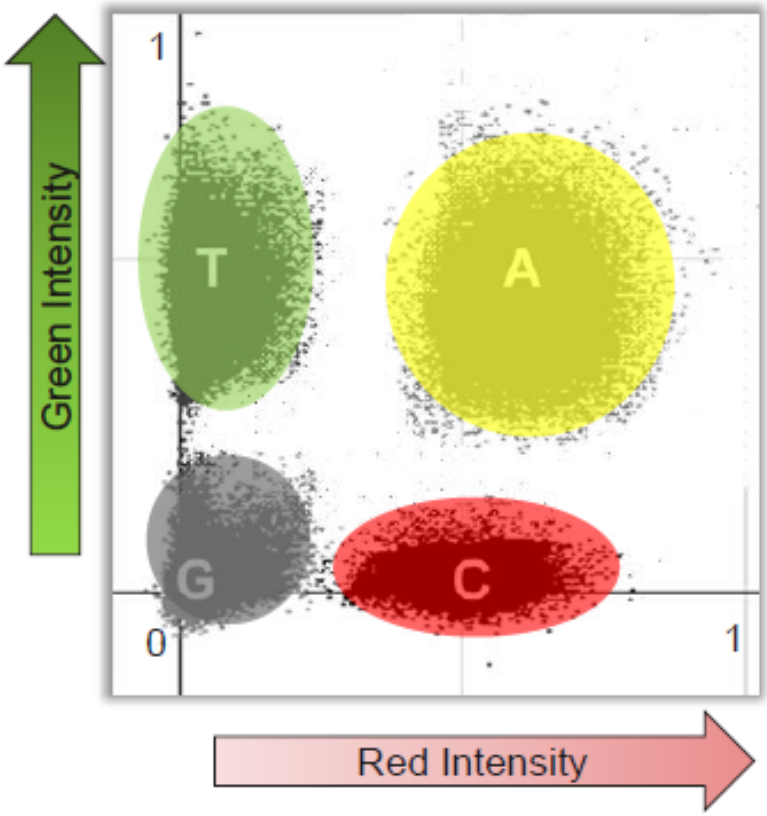
If the chastity is below 0.6 the base called will be N

2 Color Population-Based Base Calling

MiniSeq, NextSeq, NovaSeq

Advanced…

- **Scatterplot of 4 distinct populations (nucleotides) is created from extracting intensities from one image versus the other image**
- **Base calls are made according to which channel is on (1) or off (0) for each cluster according to (x, y):**
  - (1, 0) → C
  - (0, 1) → T
  - (1, 1) → A
  - (0, 0) → G



illumina

2-Color Calculating Clusters Passing Filter

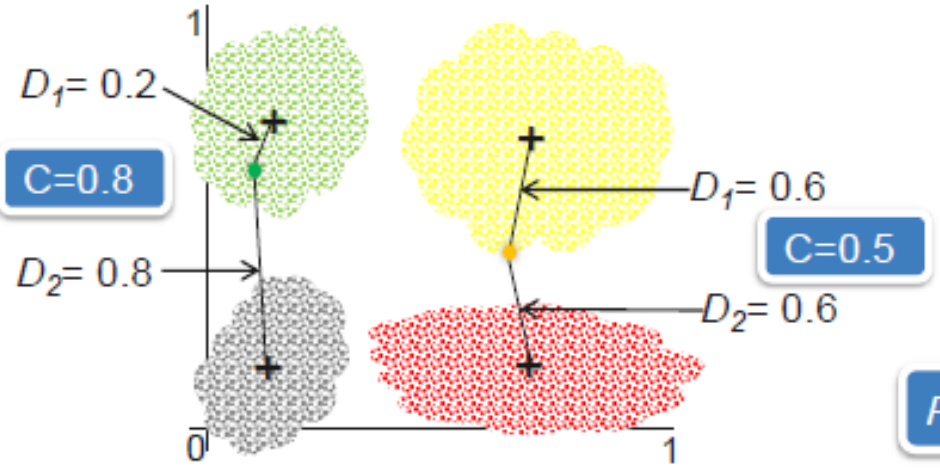
MiniSeq, NextSeq, NovaSeq

Advanced…

Pass filter is:

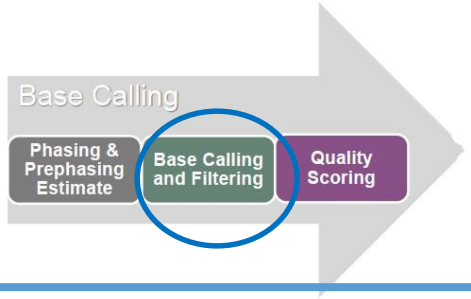
$$C = 1 - \frac{D_1}{D_1 + D_2}$$

- The ratio of the sum of the most prominent and second most prominent population intensities
- Calculated for each cluster over the first 25 bases of the sequence
- Filters cluster by signal purity
  - Removes overlapping and low-intensity clusters



Passing value: ≥ 0.63

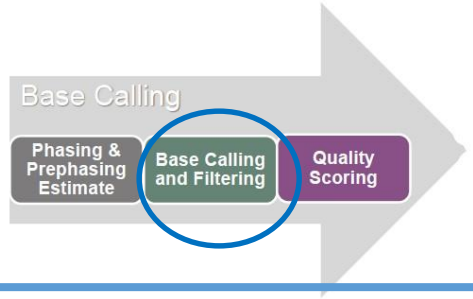
illumina



Random flow cells – calculation of %PF clusters passing filter

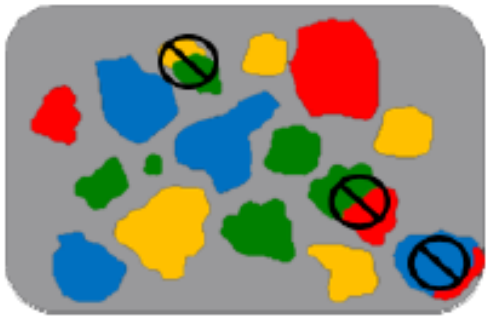
- Clusters "**P**ass **F**ilter" (PF) if no more than one base call has a chastity value below 0.6 in the first 25 cycles.

- This filtration process removes the least reliable clusters from the image analysis results.



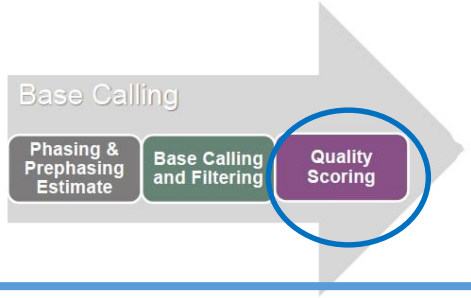
Calculation of %PF clusters passing filter

	Random flow cell	Patterned flow cell
Total Clusters	16	25 (Fixed)
Polyclonal Clusters	-3	-2
Empty Wells	NA	-5
Clusters Passing Filter	13 (81% PF)	18 (72% PF)



Pass filter is lower but output is higher

illumina



Quality scores

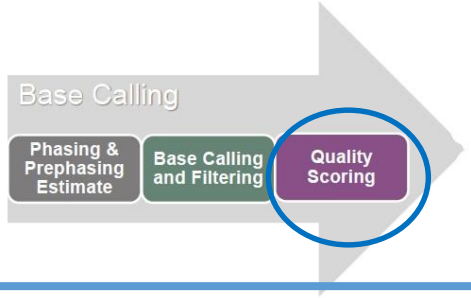
- Quality scores are used to measure base accuracy
- Q score (Phred) is the probability that the base call is wrong.

Phred quality scores are logarithmically linked to error probabilities

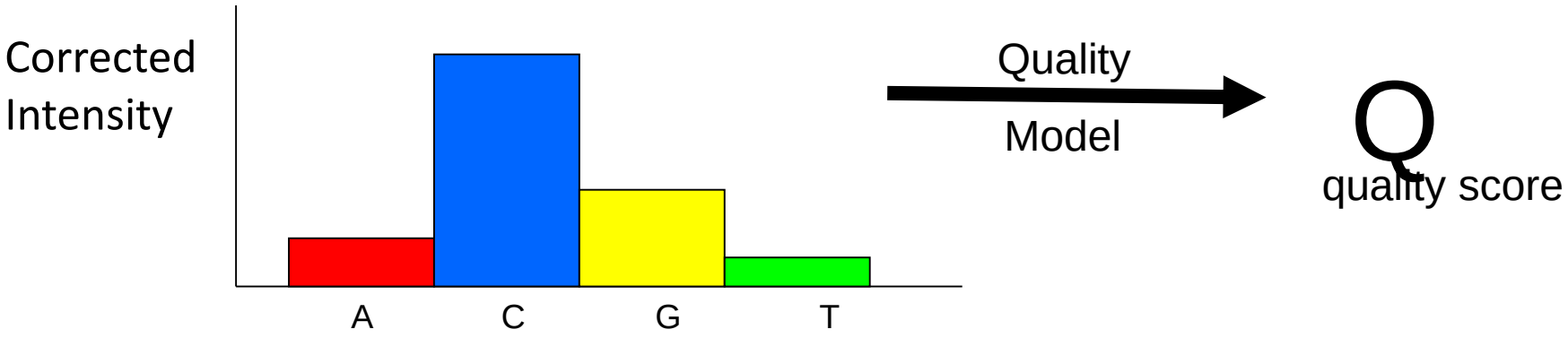
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

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

p = Probability call is incorrect



Assigning quality scores



- Quality predictor values - observable properties:
  - Intensity
  - Signal-to-noise ratios

- Quality model –relates quality predictor values to quality scores:
  - Based on empirical data of various well-characterized human and non-human samples sequenced on a number of instruments.
  - Updated when characteristics of the sequencing platform change.

https://emea.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

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

Line 1: Unique ID for a sequencing read

`@HISEQ:126:H14YJADXX:1:1101:1118:2101 1:N:0:ATCACG`

HISEQ	the unique instrument name
126	the run id
H14YJADXX	the flowcell id
1	flowcell lane
1101	tile number within the flowcell lane
1118	'x'-coordinate of the cluster within the tile
2101	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
N	Y if the read is filtered, N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Quality score representation

Quality scores are represented as ASCII characters in order to save space, so that there is one ASCII character per base.

ASCII code = Q score + 33 Q = ASCII_CODE - ASCII_BASE

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			

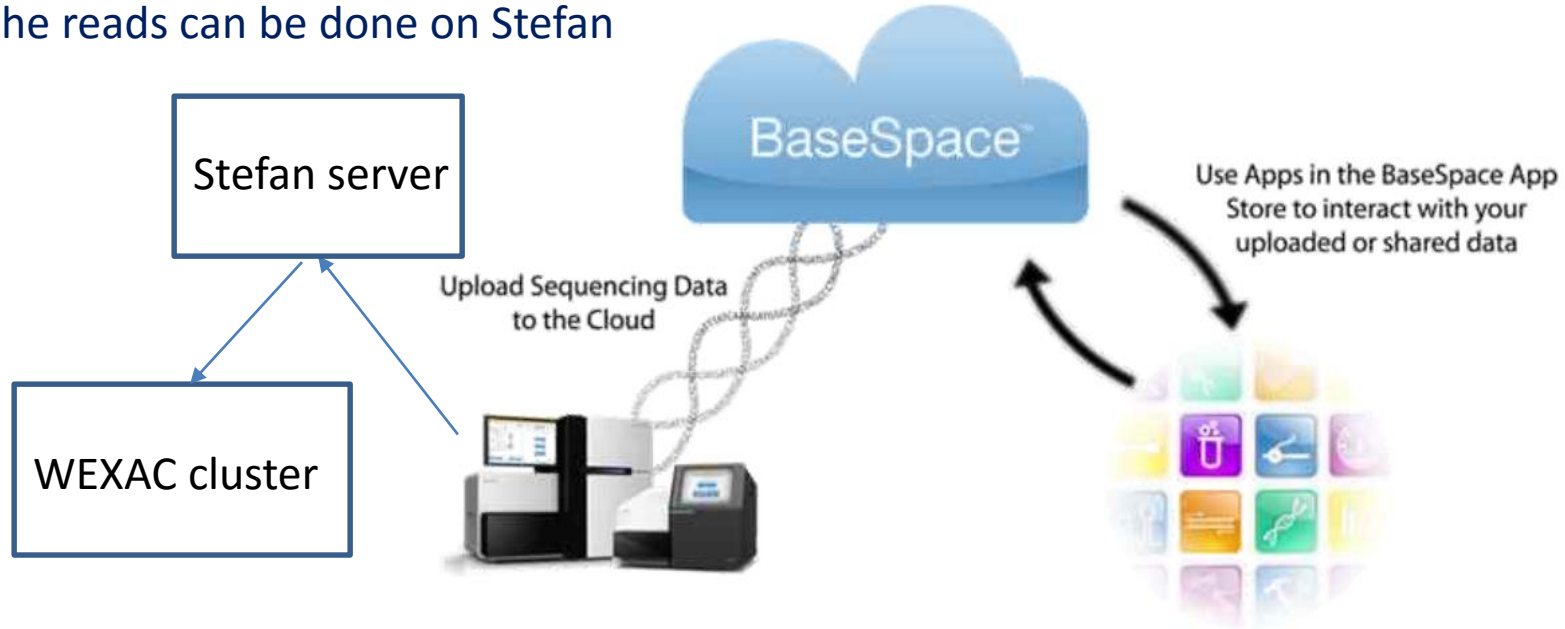
ASCII stands for American Standard Code for Information Interchange.
ASCII code is the numerical representation of a character such as 'a' or '@'

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

Storing sequence information

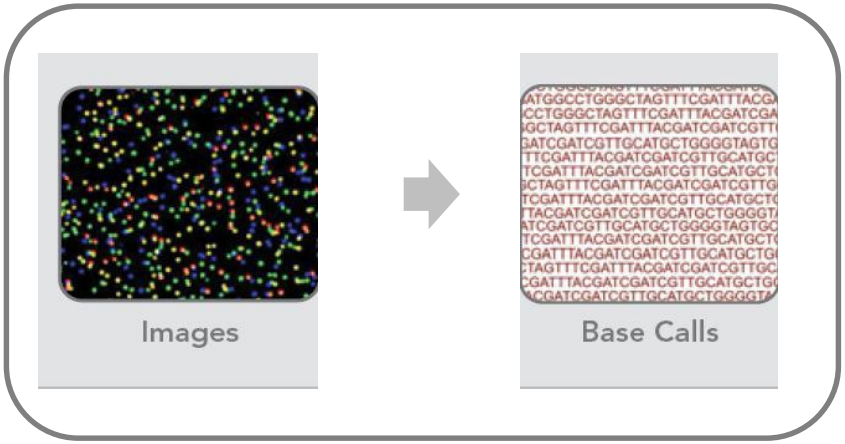
- A website where biologists and informaticians can easily store, analyse, and share genetic data from Illumina instruments
- How do you get there? → https://basespace.illumina.com

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

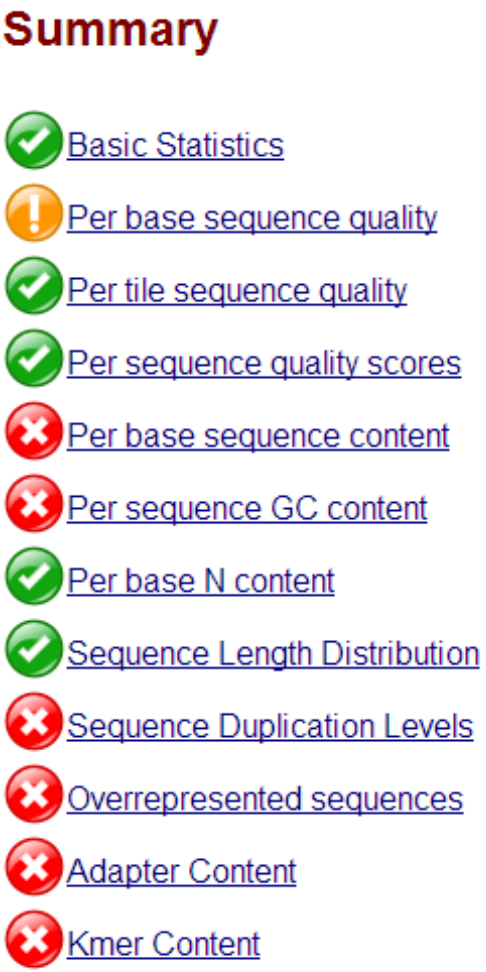
Developed by Simon Andrews at the Babraham Institute in Cambridge
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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.



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

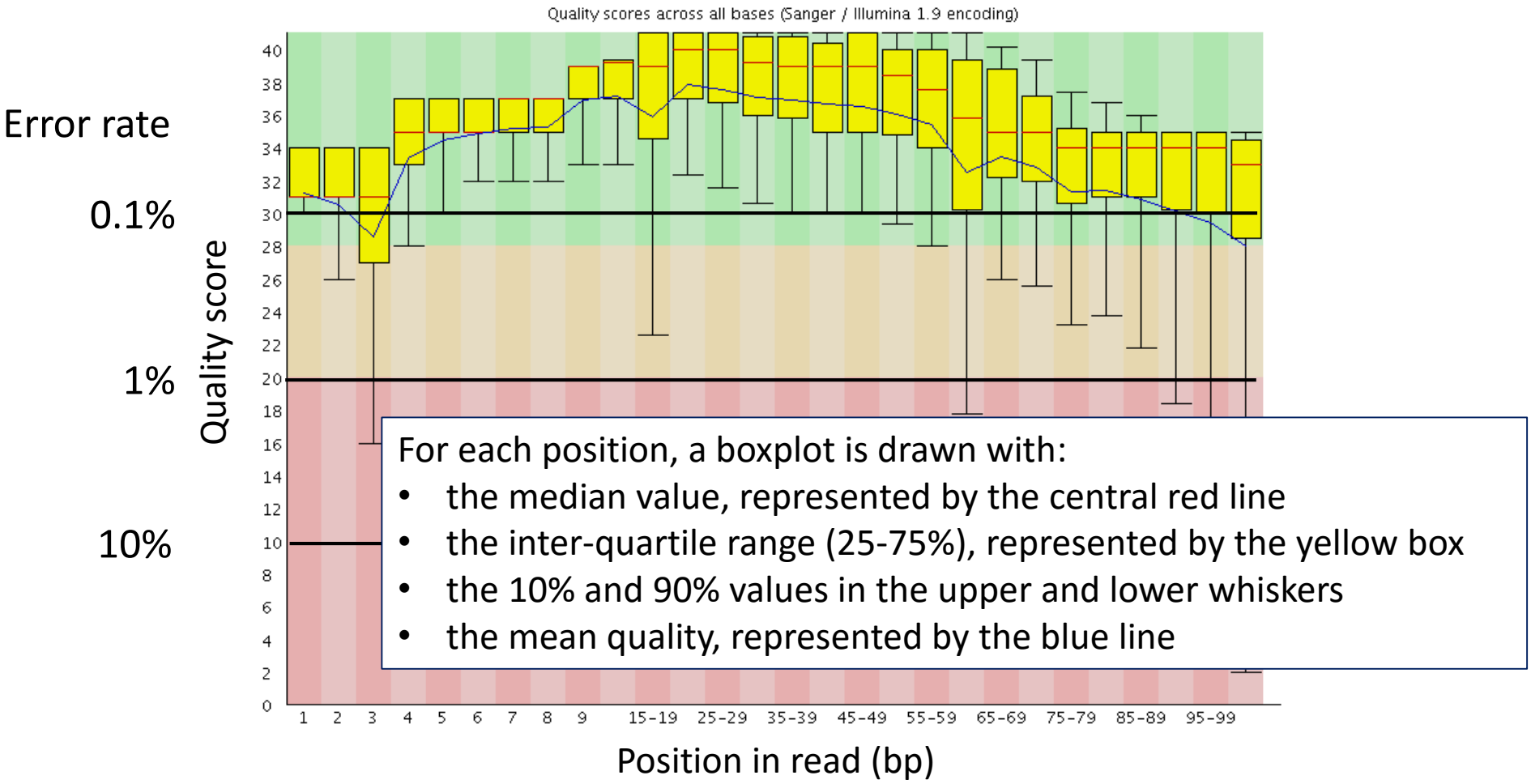
- # of sequences
- Base call qualities
- Base composition
- Potential contaminants
- Expected duplication rate

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

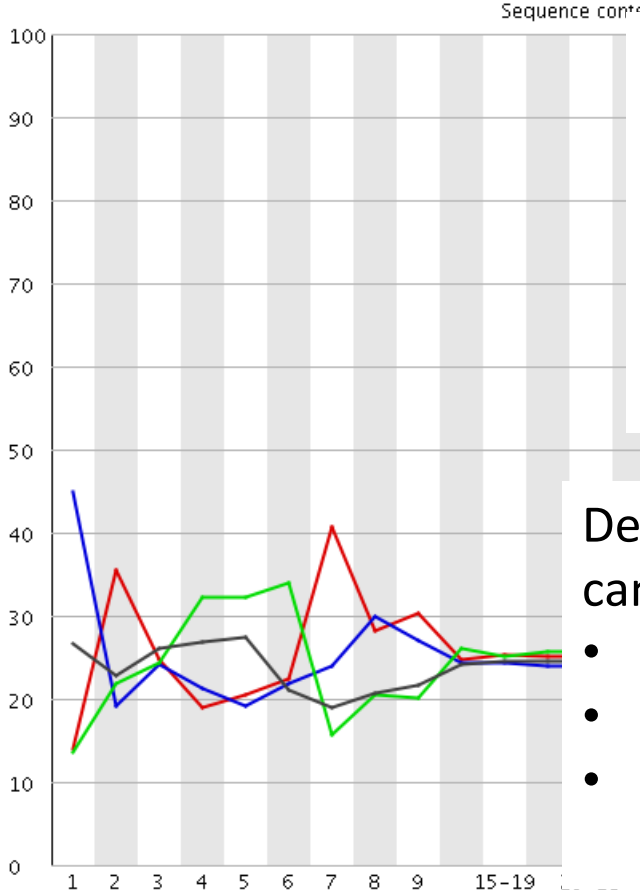
This plots the Q-score of the raw sequence reads as a box-plot for each cycle.
Higher is always better, and the characteristic decay of quality is seen in most runs.

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

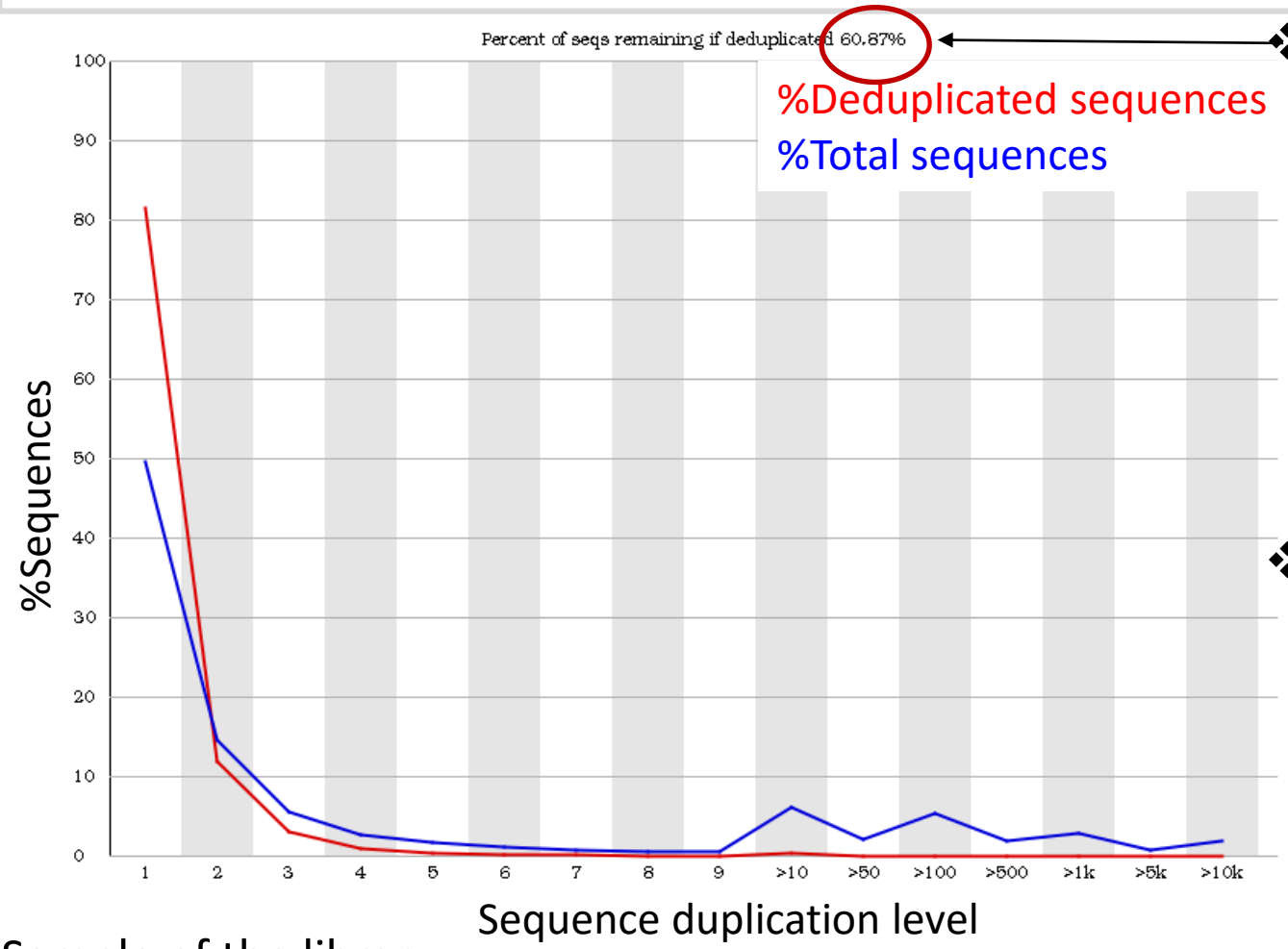
Deviation from normal base content (equally represented) can be:

- 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)

FASTQC – Sequence Duplication Level

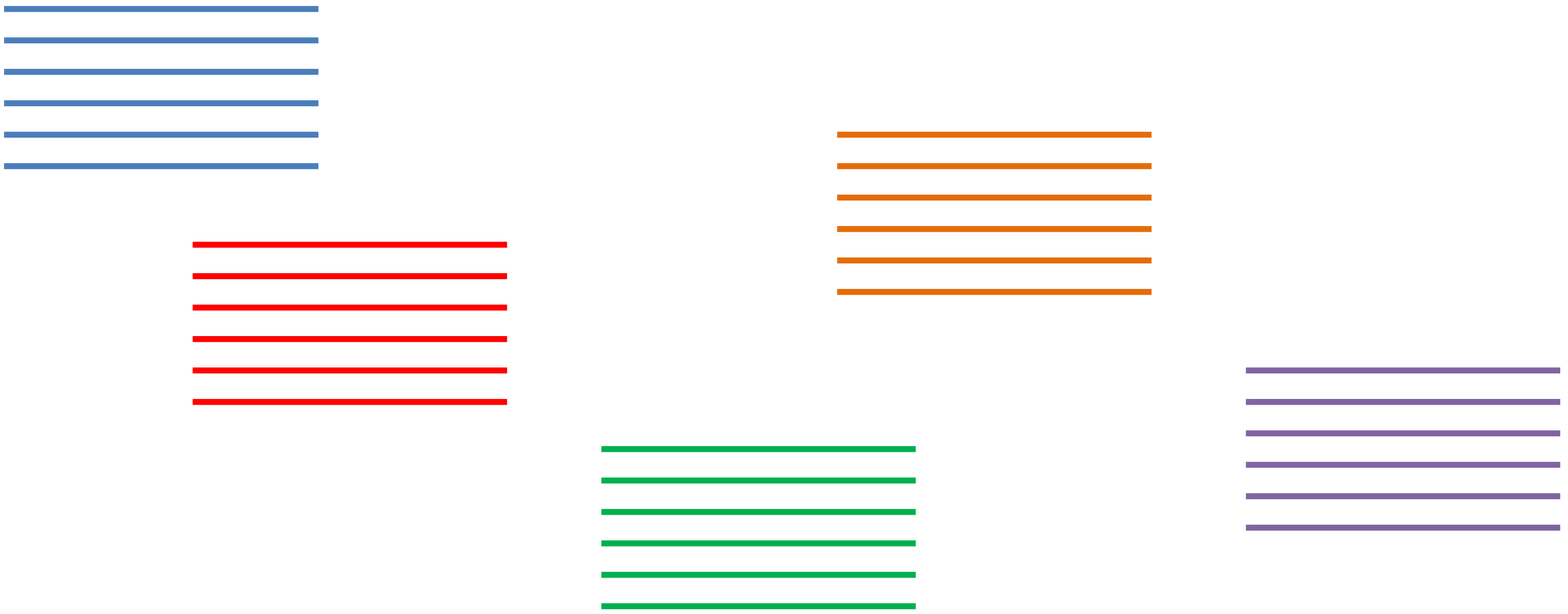
Duplicated read = copy of **exactly the same** sequence



- 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.

Example
Explaining Deduplicated Calculation



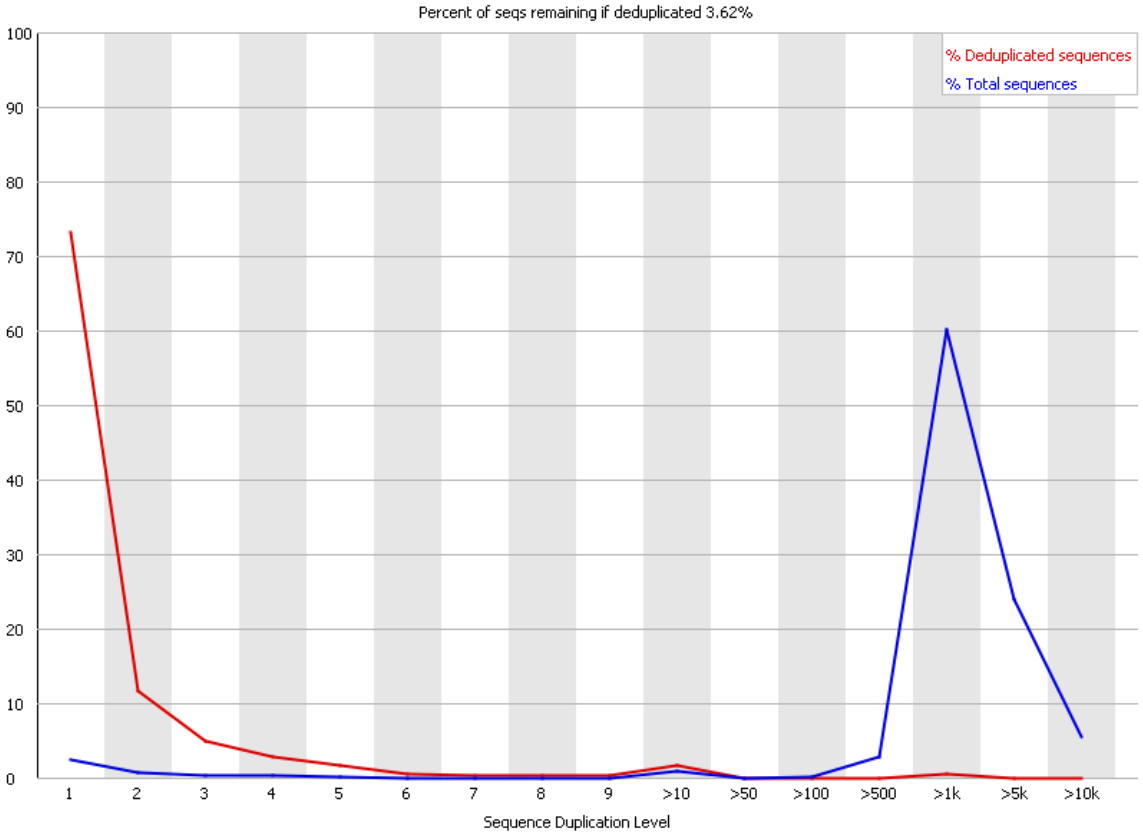
Look at the bin of sequences duplicated 6 times.

We have 5 different sequences each duplicated 6 times.

How many reads do we have in this bin in total

and after deduplication?

Examples
Sequence Duplication Plots

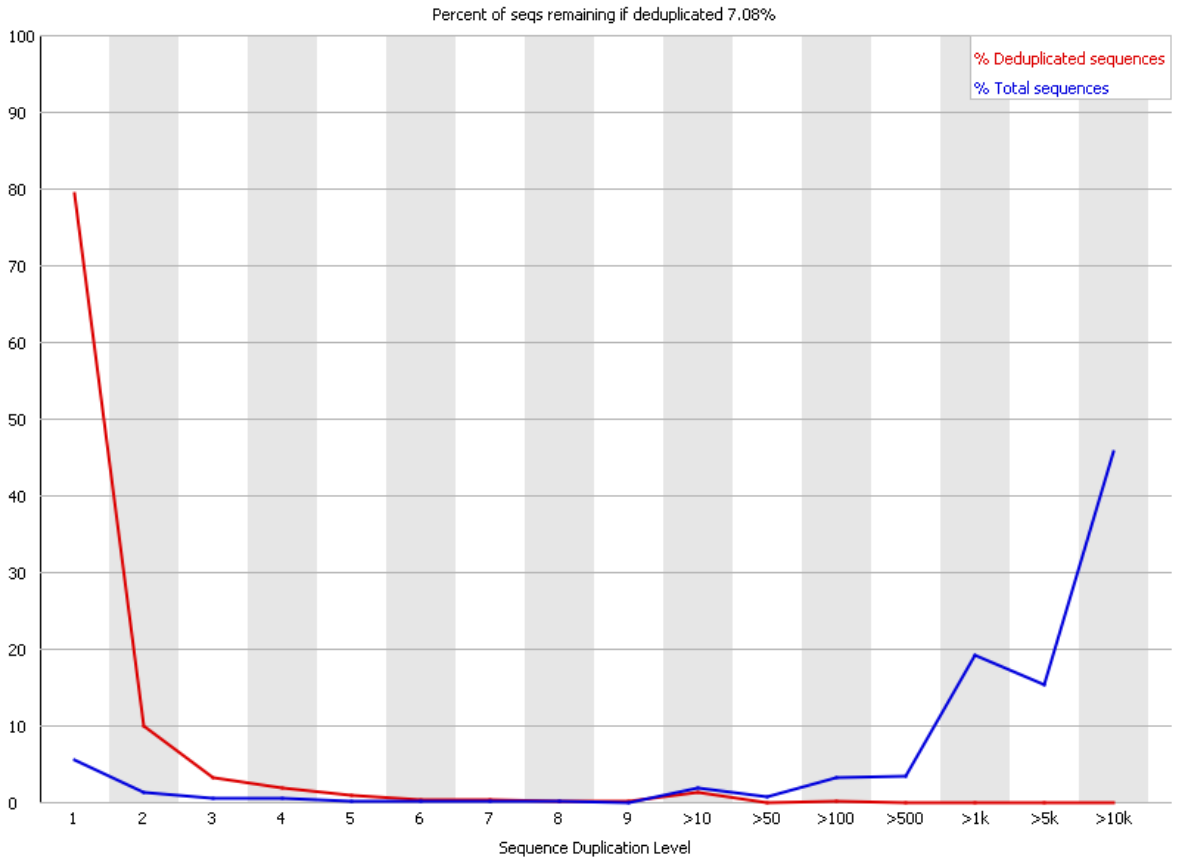


PhiX is a bacteriophage with a well-defined genome sequence of 5386 nucleotides.
PhiX is commonly used as a control for Illumina sequencing runs.

- 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

QC reports for NextSeq runs at LSCF

General QC for run AHNCYJBGX7

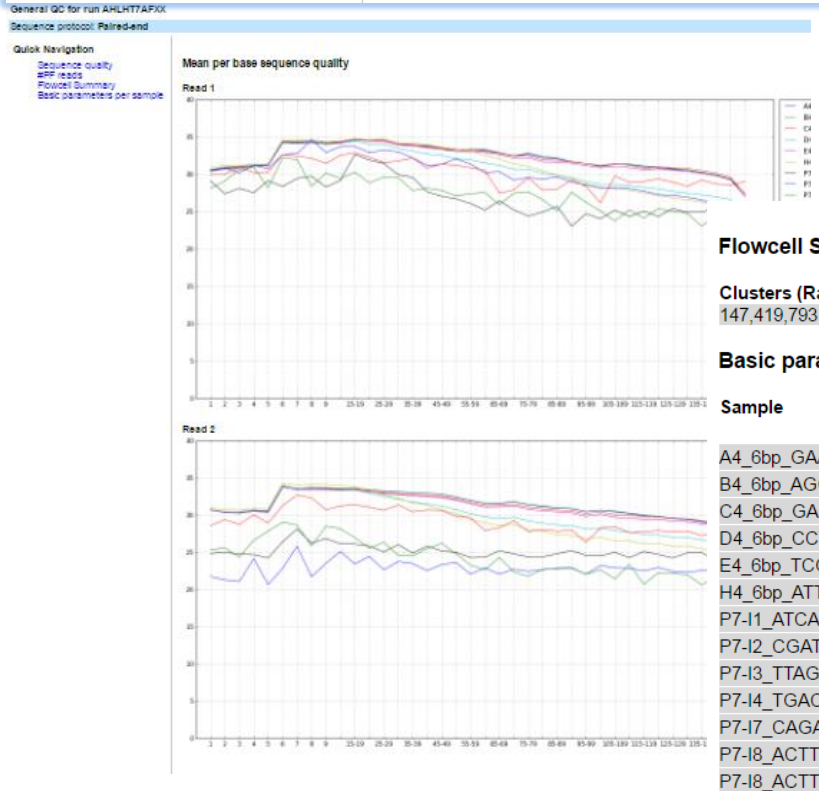
Sequence protocol: **Paired-end**

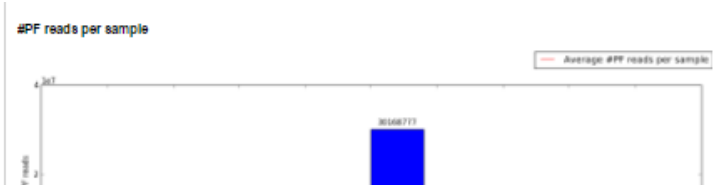
Quick Navigation

- 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/

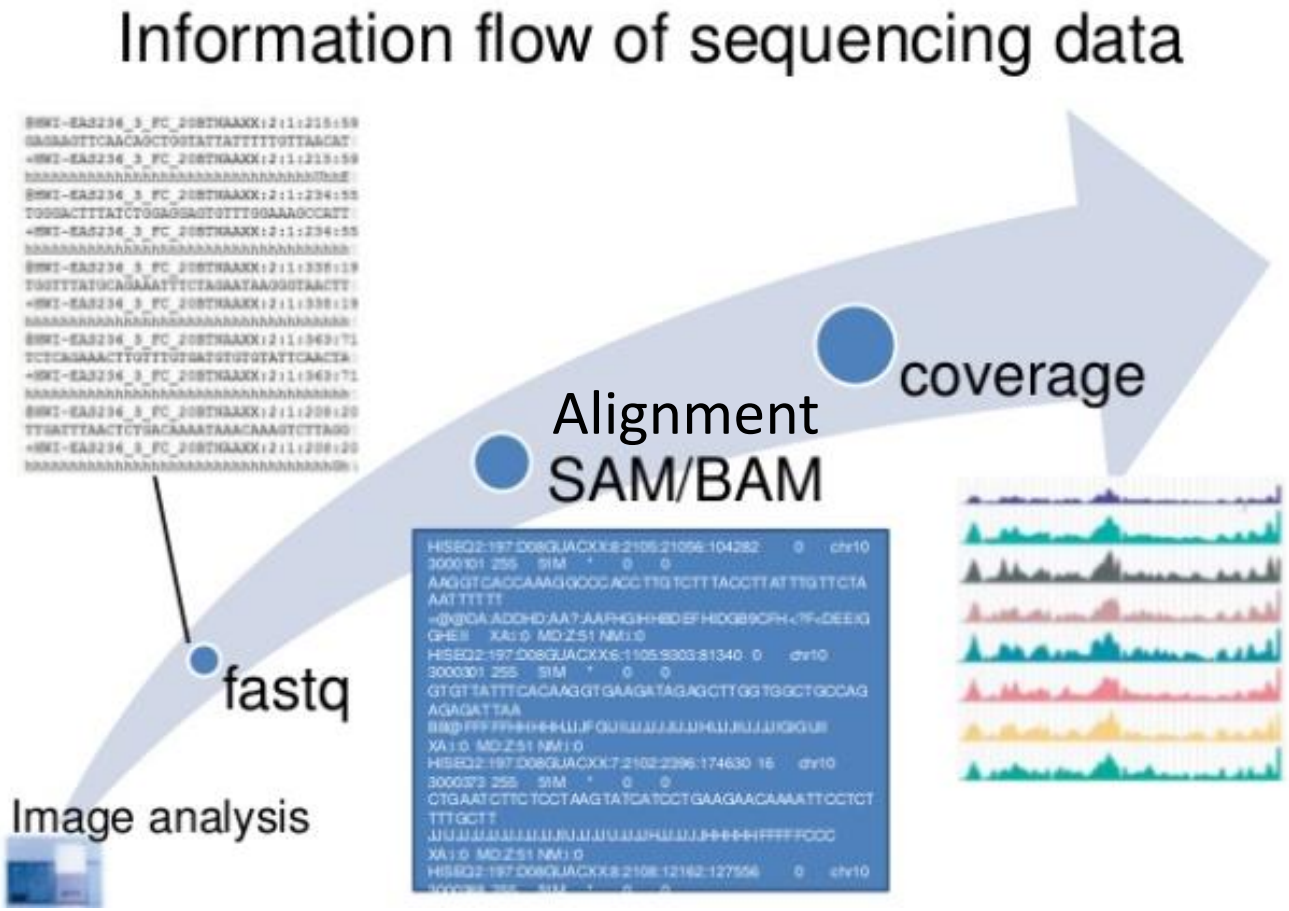
Summary

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