Genome assembly strategies

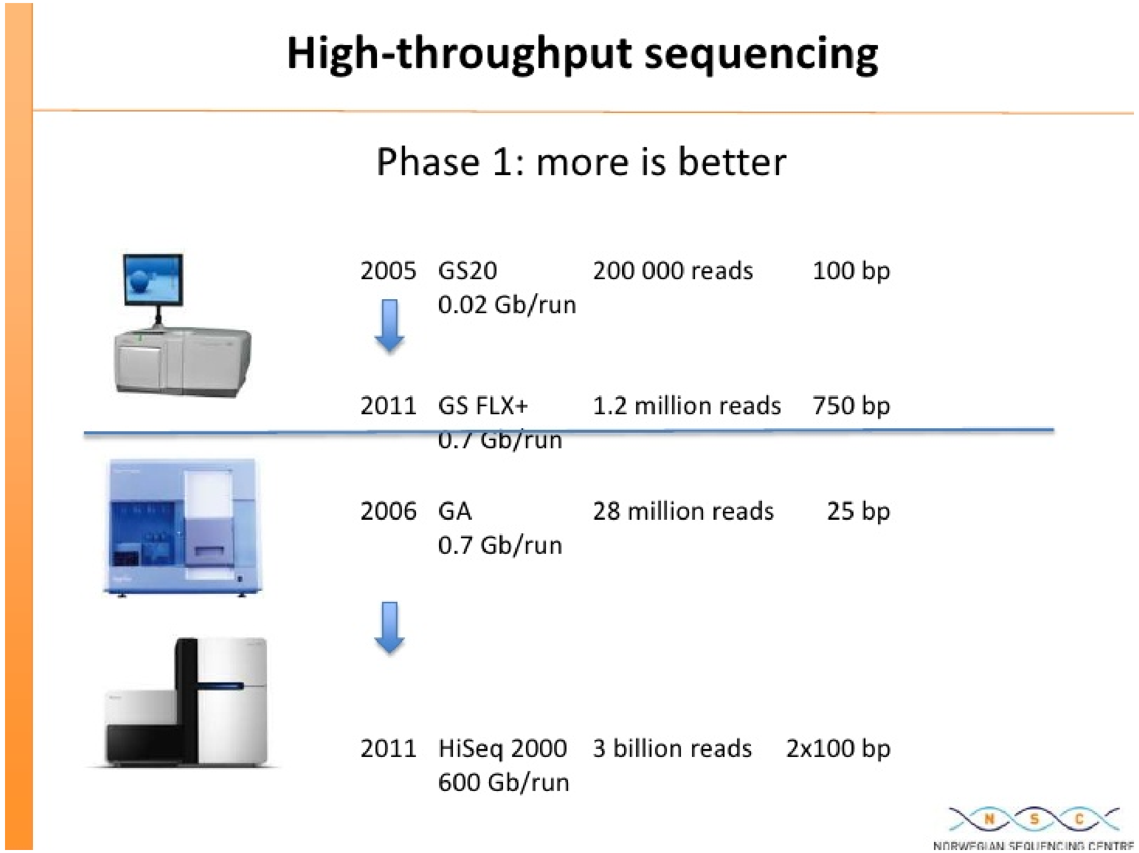
Arturo Vera Ponce de Leon

May 2019

veraponcedeleon.1@osu.edu

History of NGS and Quality control

High throughput sequencing or NGS

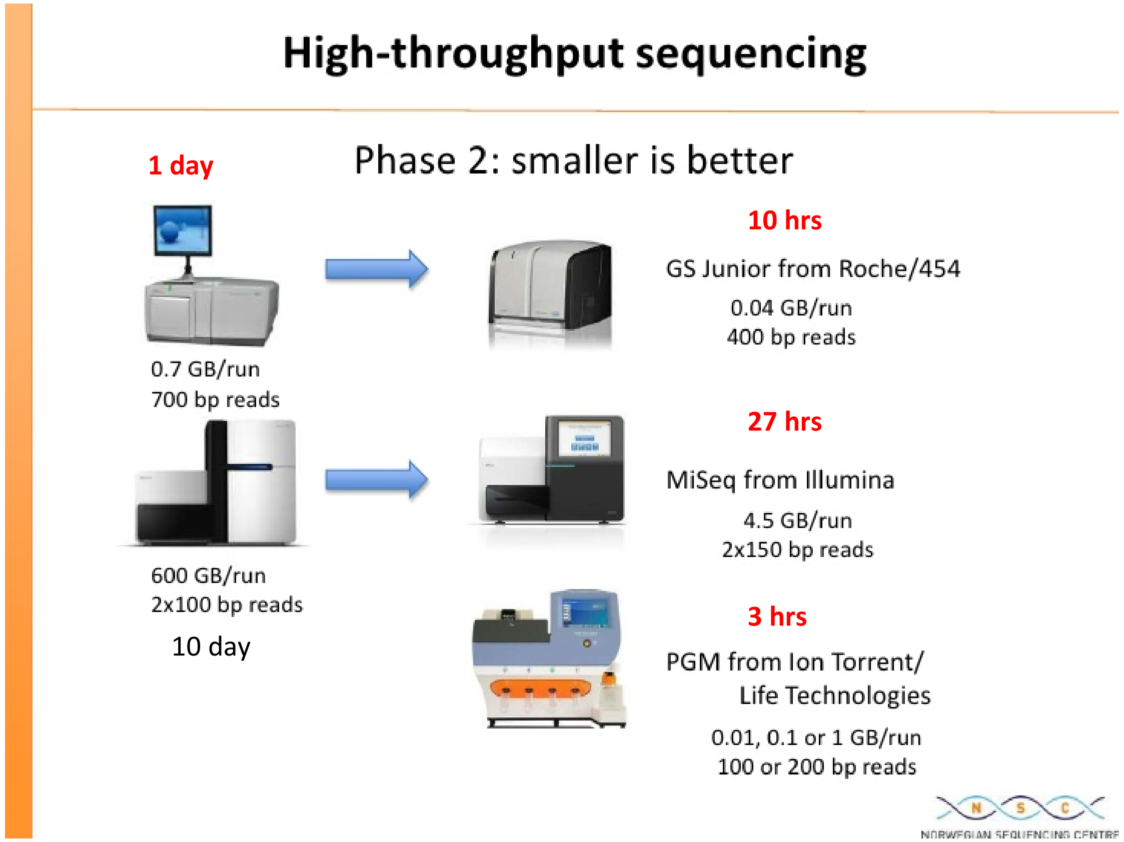


High-throughput sequencing

Phase 1: more is better

2005	GS20 0.02 Gb/run	200 000 reads	100 bp
2011	GS FLX+ 0.7 Gb/run	1.2 million reads	750 bp
2006	GA 0.7 Gb/run	28 million reads	25 bp
2011	HiSeq 2000 600 Gb/run	3 billion reads	2x100 bp

NORWEGIAN SEQUENCING CENTRE



High-throughput sequencing

Phase 2: smaller is better

1 day

0.7 GB/run
700 bp reads

600 GB/run
2x100 bp reads

10 day

10 hrs

GS Junior from Roche/454

0.04 GB/run
400 bp reads

27 hrs

MiSeq from Illumina

4.5 GB/run
2x150 bp reads

3 hrs

PGM from Ion Torrent/
Life Technologies

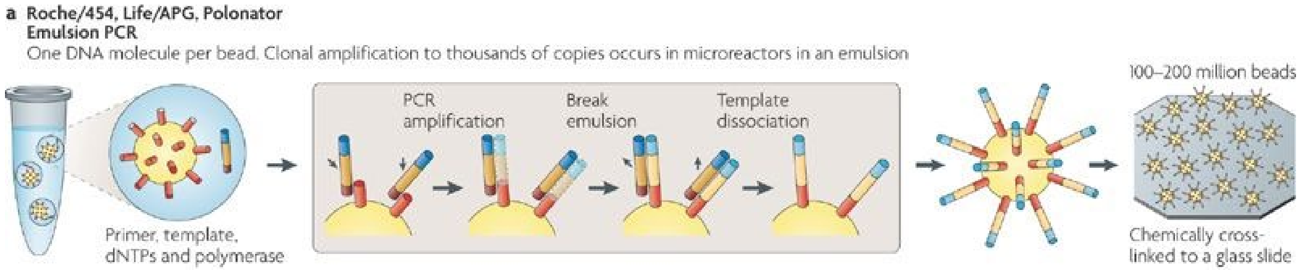
0.01, 0.1 or 1 GB/run
100 or 200 bp reads

NORWEGIAN SEQUENCING CENTRE

Sequencing-by-synthesis categories. SBS is a term used to describe numerous DNA-polymerase-dependent methods

454 and IonTorrent sequencing

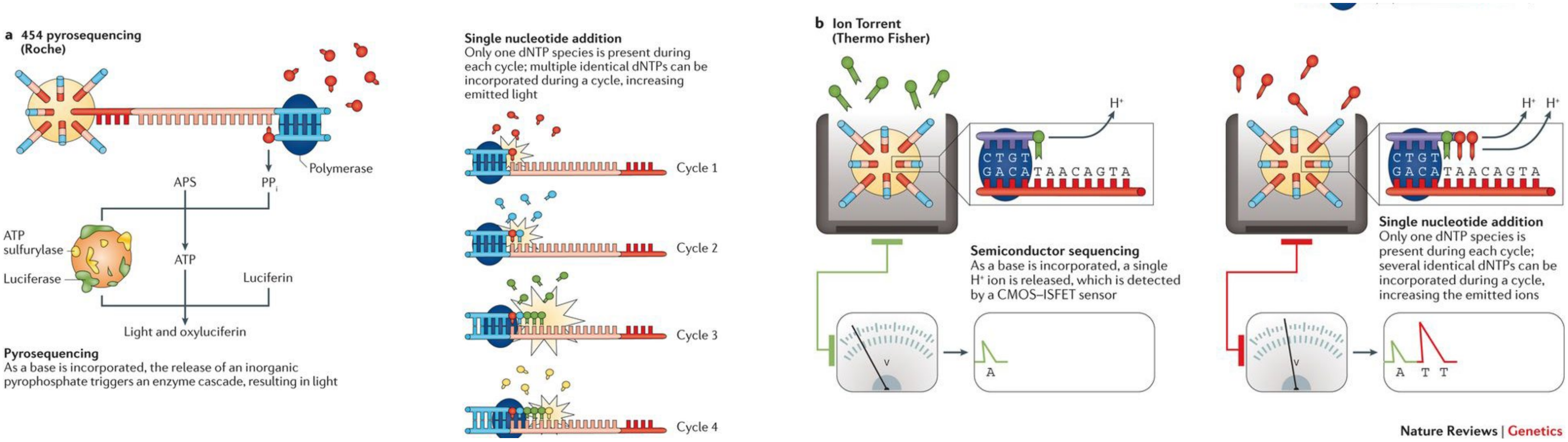
Template immobilization strategies.



a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion

Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide

Sequencing by synthesis: single-nucleotide addition approaches.



a 454 pyrosequencing (Roche)

Pyrosequencing
As a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light

Single nucleotide addition
Only one dNTP species is present during each cycle; multiple identical dNTPs can be incorporated during a cycle, increasing emitted light

b Ion Torrent (Thermo Fisher)

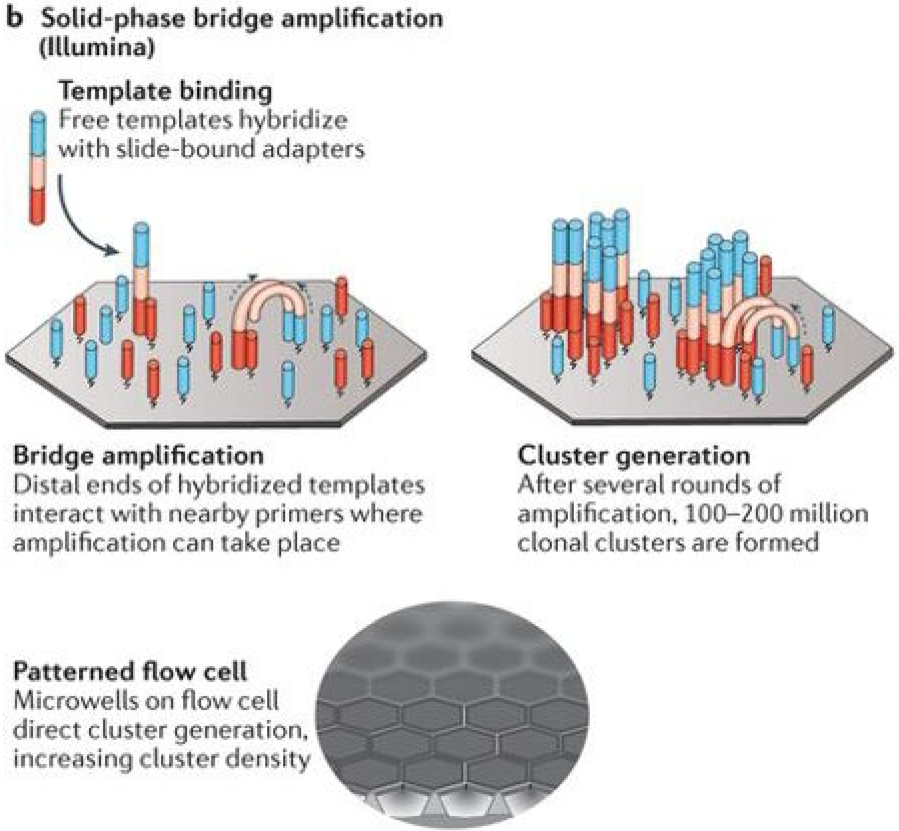
Semiconductor sequencing
As a base is incorporated, a single H$^+$ ion is released, which is detected by a CMOS–ISFET sensor

Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions

Nature Reviews Genetics **volume17**, 333–351 (2016)

Illumina technology

Template immobilization strategies.



b Solid-phase bridge amplification (Illumina)

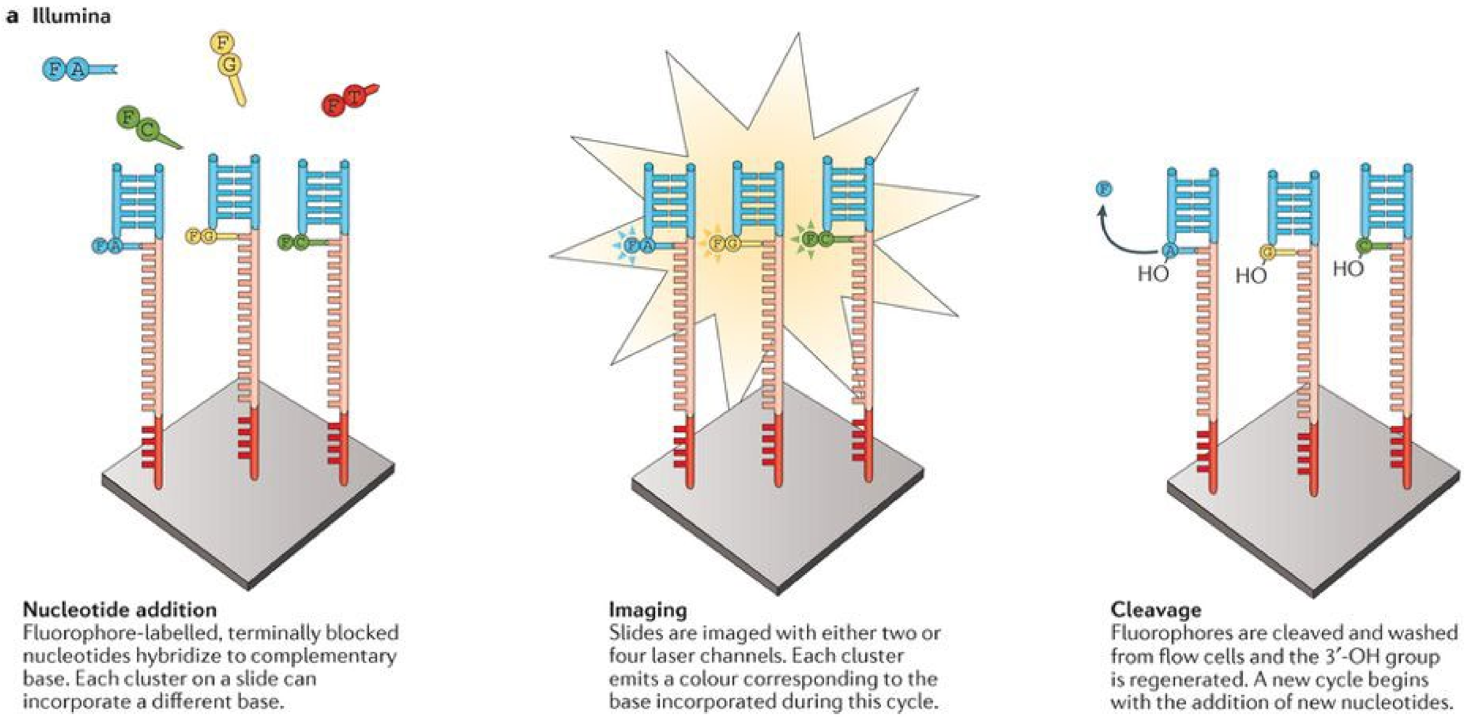
Template binding
Free templates hybridize with slide-bound adapters

Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

Sequencing by synthesis: cyclic reversible termination approaches.

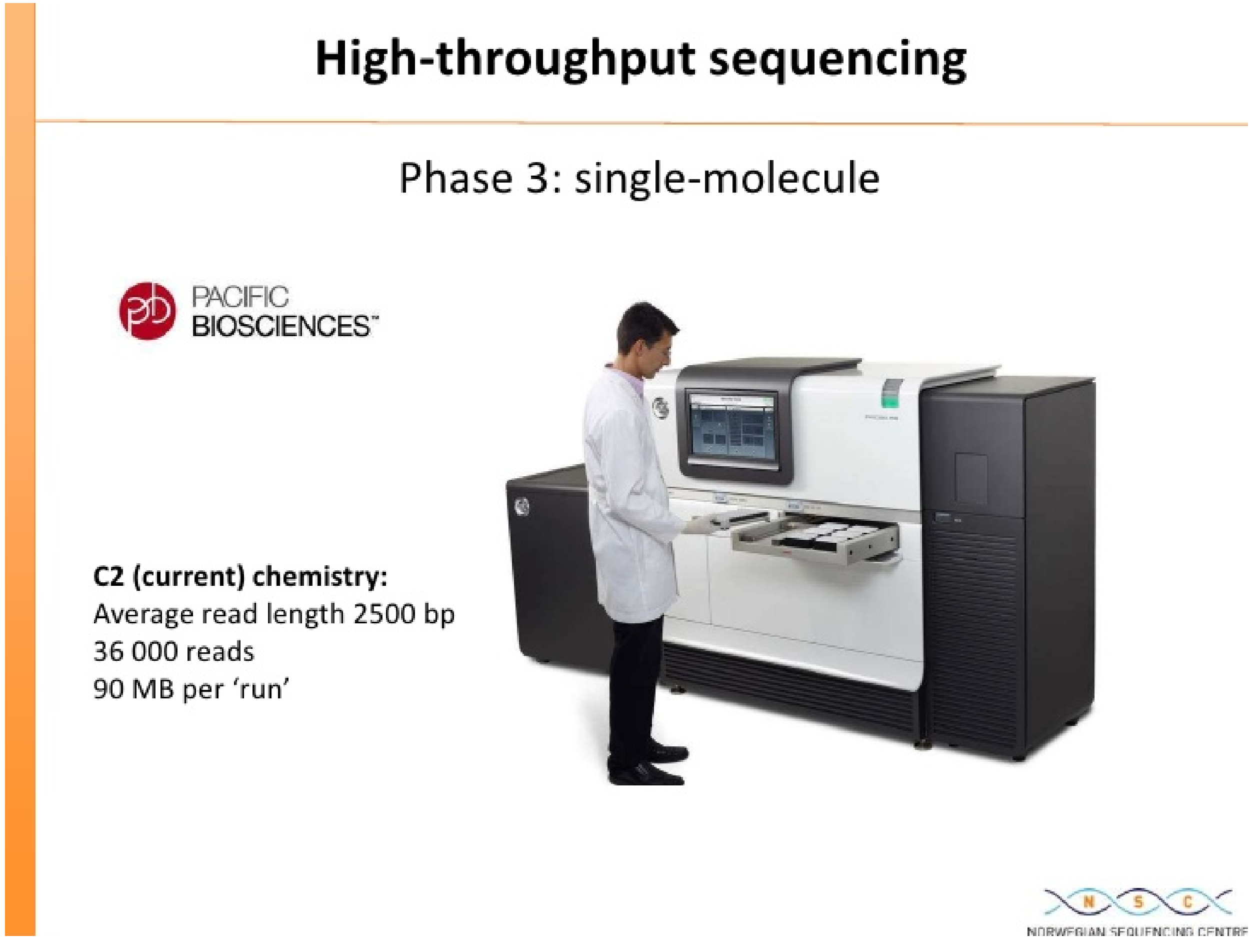


Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Nature Reviews Genetics **volume17**, 333–351 (2016)



High-throughput sequencing

Phase 3: single-molecule

PACIFIC BIOSCIENCES™

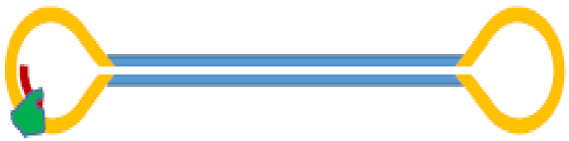
C2 (current) chemistry:
Average read length 2500 bp
36 000 reads
90 MB per 'run'

NORWEGIAN SEQUENCING CENTRE

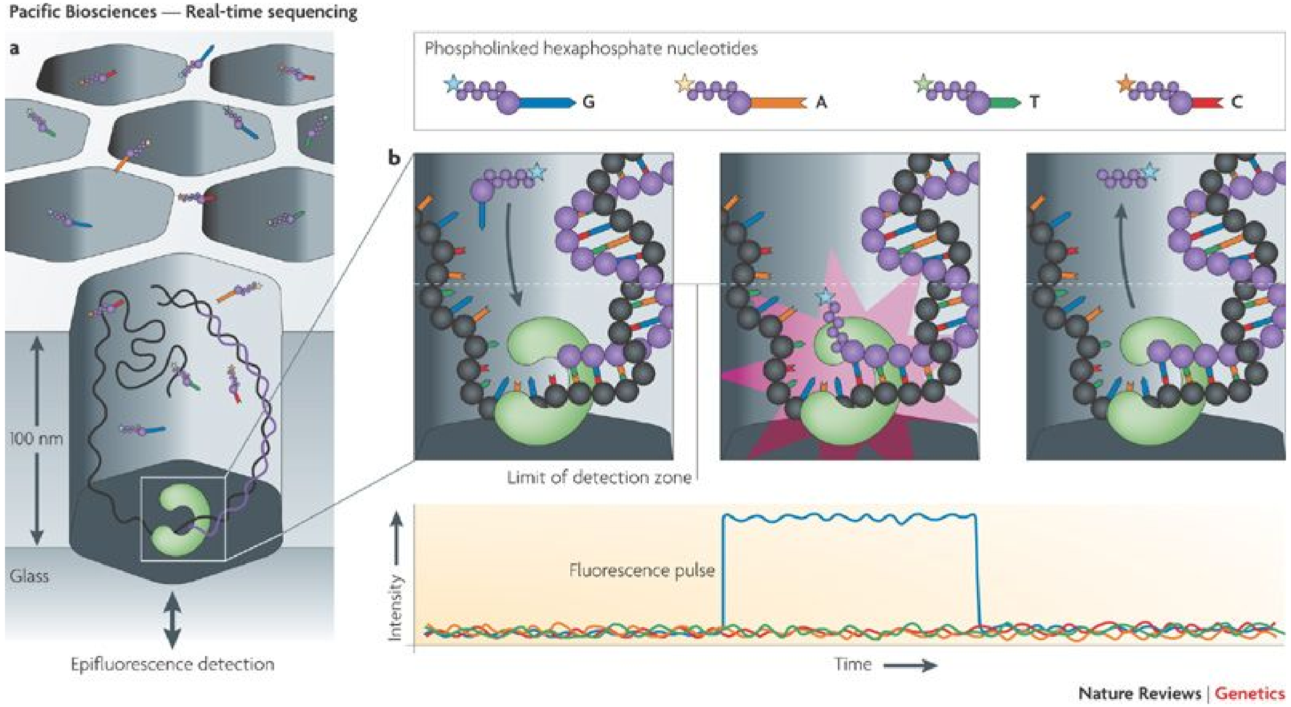
Oxford NANOPORE Technologies

Real-time sequencing.

(A)

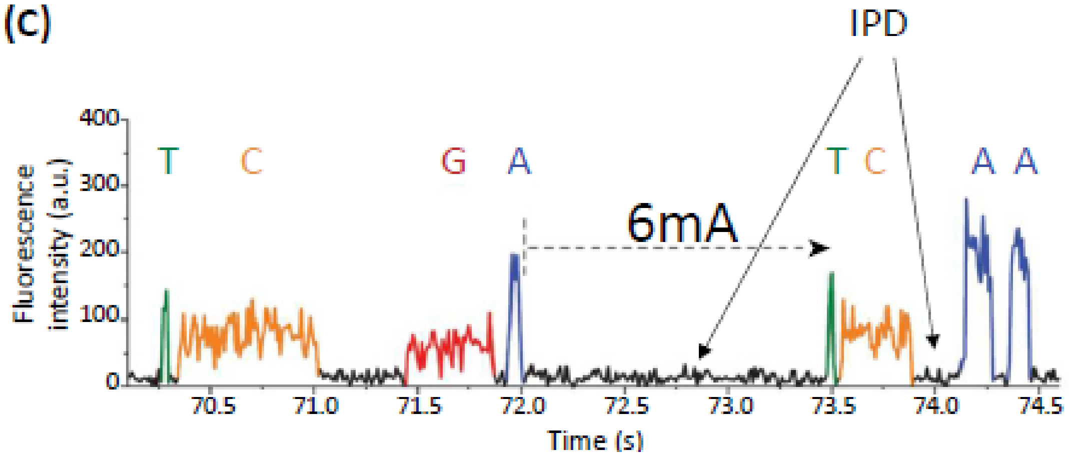


Library preparation comprises the ligation of hairpin adapters (yellow) to double-stranded DNA molecules (blue), thereby creating circular molecules called ‘SMRTbells’.


Pacific Biosciences — Real-time sequencing
a
b
Phospholinked hexaphosphate nucleotides
G
A
T
C
100 nm
Glass
Epifluorescence detection
Limit of detection zone
Fluorescence pulse
Intensity
Time
Nature Reviews | Genetics

Nature Reviews Genetics **volume17**, 333–351 (2016)

the interpulse duration (IPD)



The presence of an epigenetic modification, such as 6-methyladenosine (6 mA), results in a delayed IPD

Trends in Genetics (2018) Vol. 34, No. 9 666-681

Platform Features

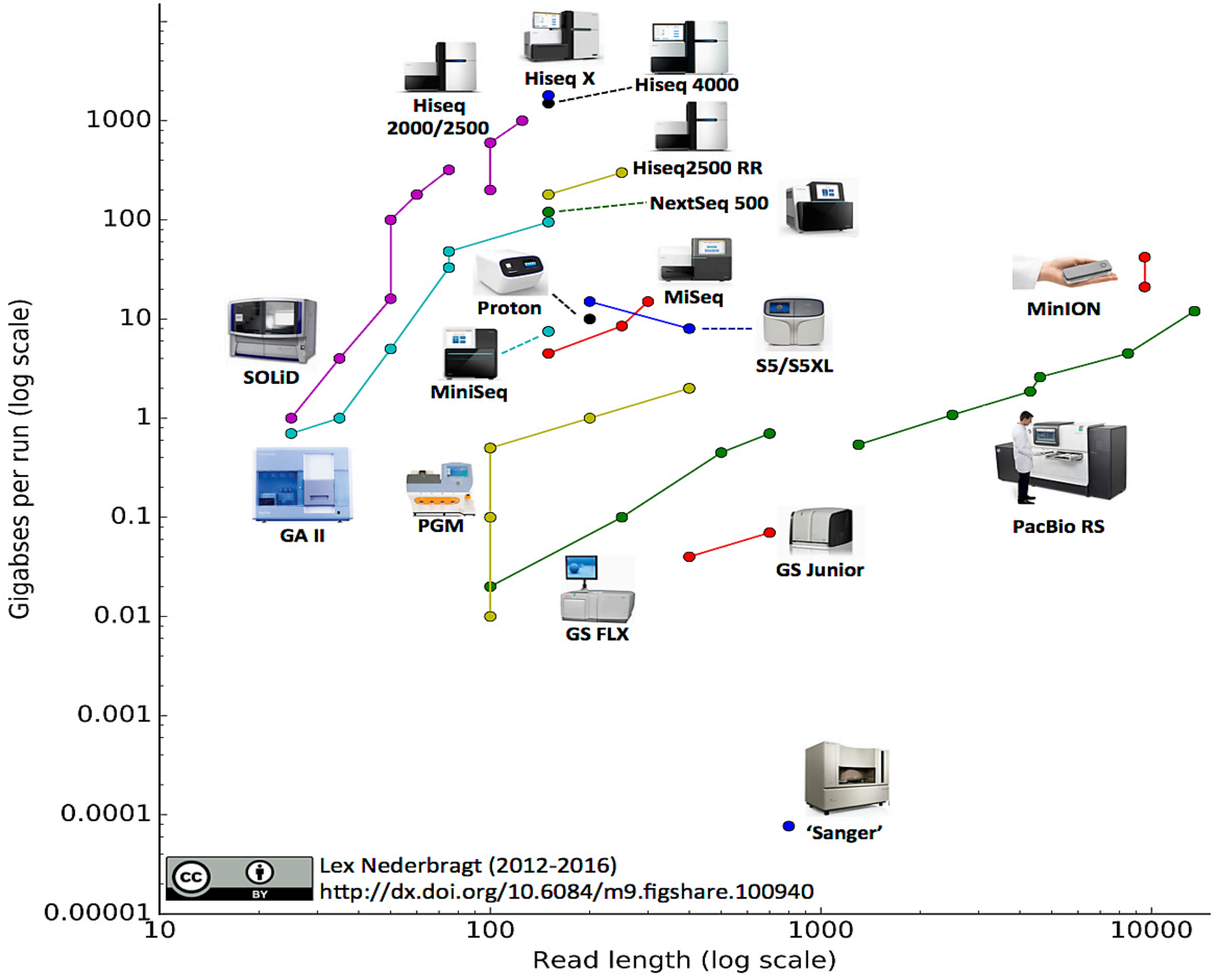
Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	$59 (PE100)	$53 (PE150)	$108 (PE300)	$697

Applications

Platform	454	Illumina HiSeq	Illumina MiSeq*	Ion Torrent	PacBio
resequencing	-	+++	++	-	+
de novo	+++	+	+	+++	+++
metagenomics	+++	++	+	+++	+/-
mRNA	++	+++	++	++	++
miRNA	-	+++	+++	-	-
ChIP	-	+++	++	-	-
DNA meth	-	+++	+	-	!!!
SNP validation	+	-	-	-	++

NORWEGIAN SEQUENCING CENTRE

Multiple technologies diverse features



Which one is the good one?

Yields

A Genome of 1Mb (1 x 10^6 bases):

- By Sanger:

  C = nl/L

  10 = n(500)/1,000,000

  n = 1,000,000*10/500

  20,000 reads

  Cost per read~ 1-2 USD

  20,000 USD (~360,000 MX pesos)
- A 454 run ~700Mb (700X)
  - Cost arpox de 20,000 USD
- Un SMRT cell de PacBio (P6-C4) ~150,000 reads (1Gb)
  - Cost 800 USD (~14,400 pesos)
- An Illumina lane ~300 millions of reads (HiSeq2000)
  - An average length of 100 bp = 30 Gb = 30,000 X
    - Cost per lane 2,000 USD (~36,000 pesos)

Coverage:
C = nl/L

C=Coverage

N=Number of reads

l=Read length

L=Genome size (length) in bases

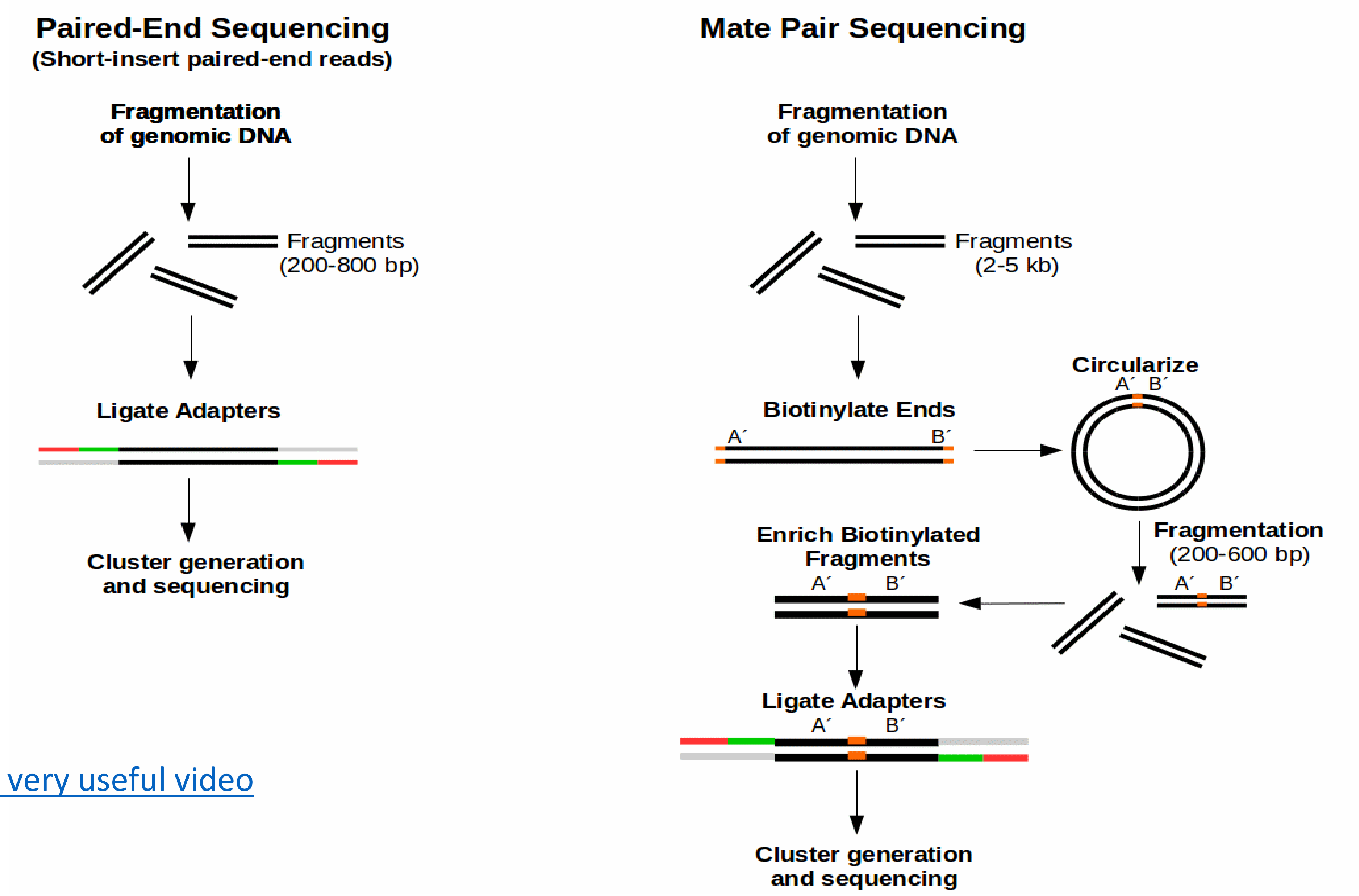
30, 000 X coverage ~ $ 36, 000 Mxpesos Illumina

10x coverage ~ $ 360, 000 MX pesos Sanger

Minimal coverage for SNPs, annotation and completeness assessments

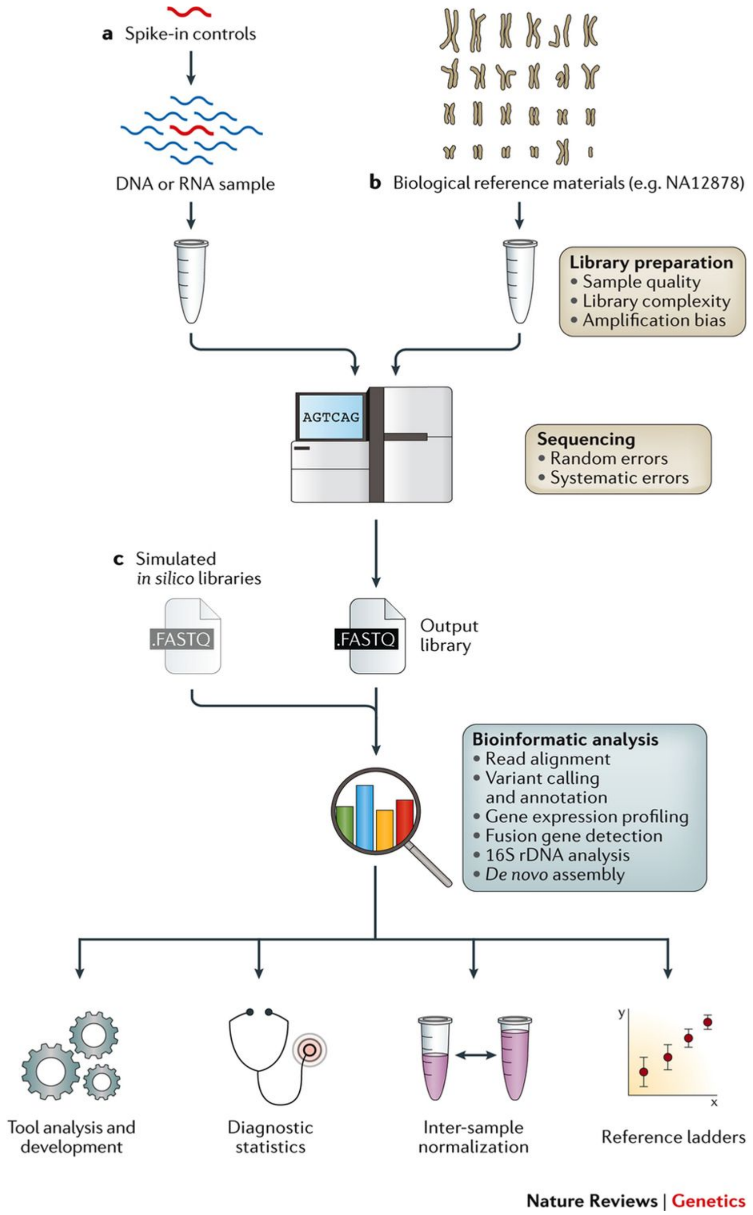
> 50 x

Pair end vs Mate Pair


Paired-End Sequencing
(Short-insert paired-end reads)
Fragmentation of genomic DNA
Fragments (200-800 bp)
Ligate Adapters
Cluster generation and sequencing
Mate Pair Sequencing
Fragmentation of genomic DNA
Fragments (2-5 kb)
Biotinylate Ends
A´
B´
Circularize
A´ B´
Fragmentation (200-600 bp)
A´ B´
Enrich Biotinylated Fragments
A´ B´
Ligate Adapters
A´ B´
Cluster generation and sequencing

Lets watch a very useful video

HTS general analysis flow chart



‘Wet-lab’ experimental design

Bioinformatics hard work

Nature Reviews Genetics volume 18, pages 473–484 (2017)

HTS general analysis time flow chart

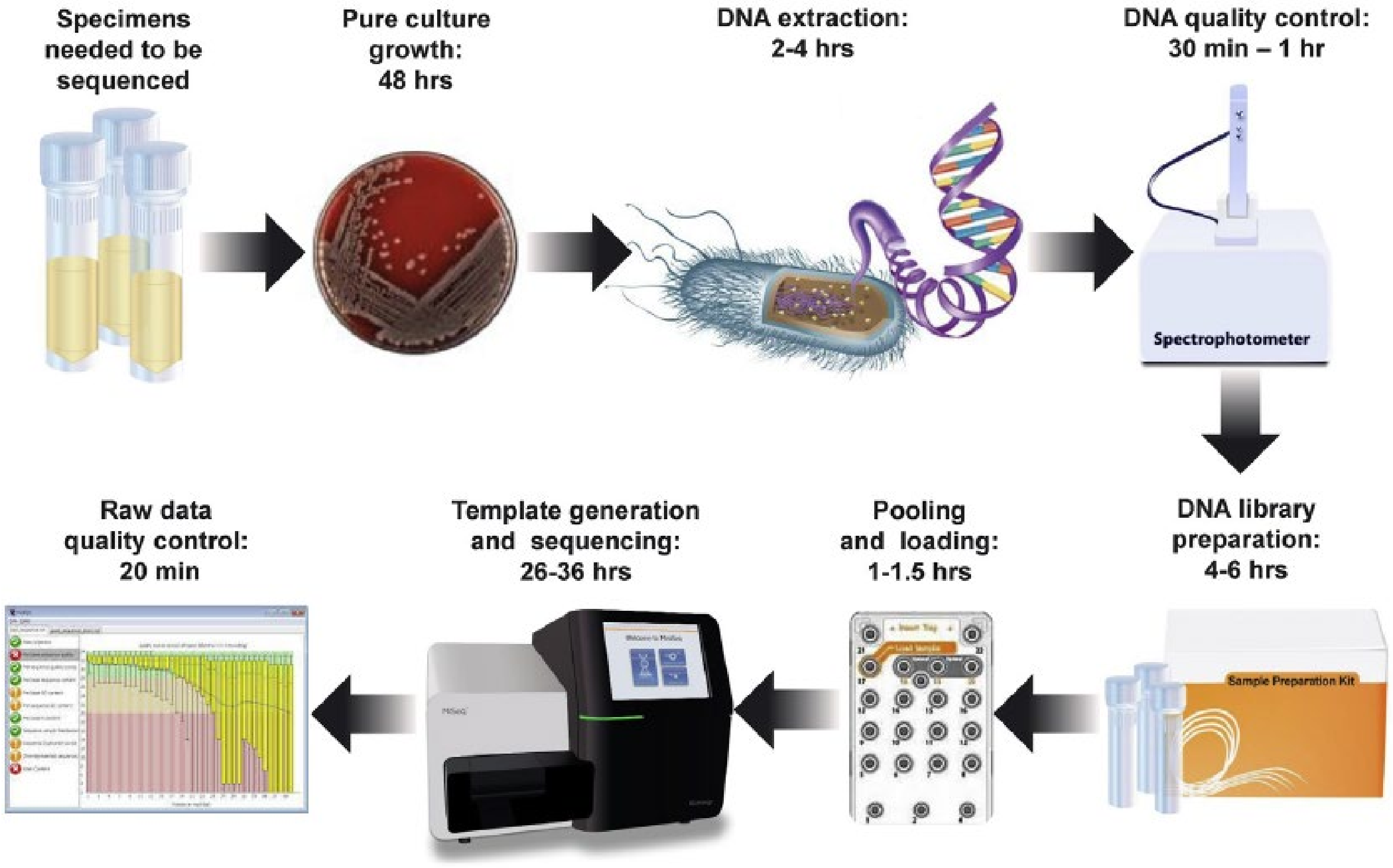
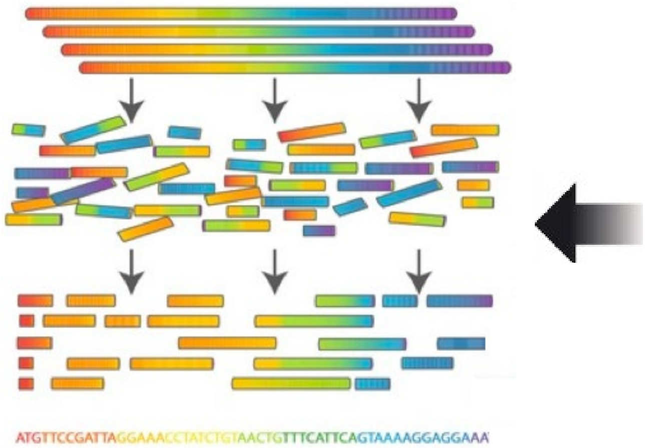
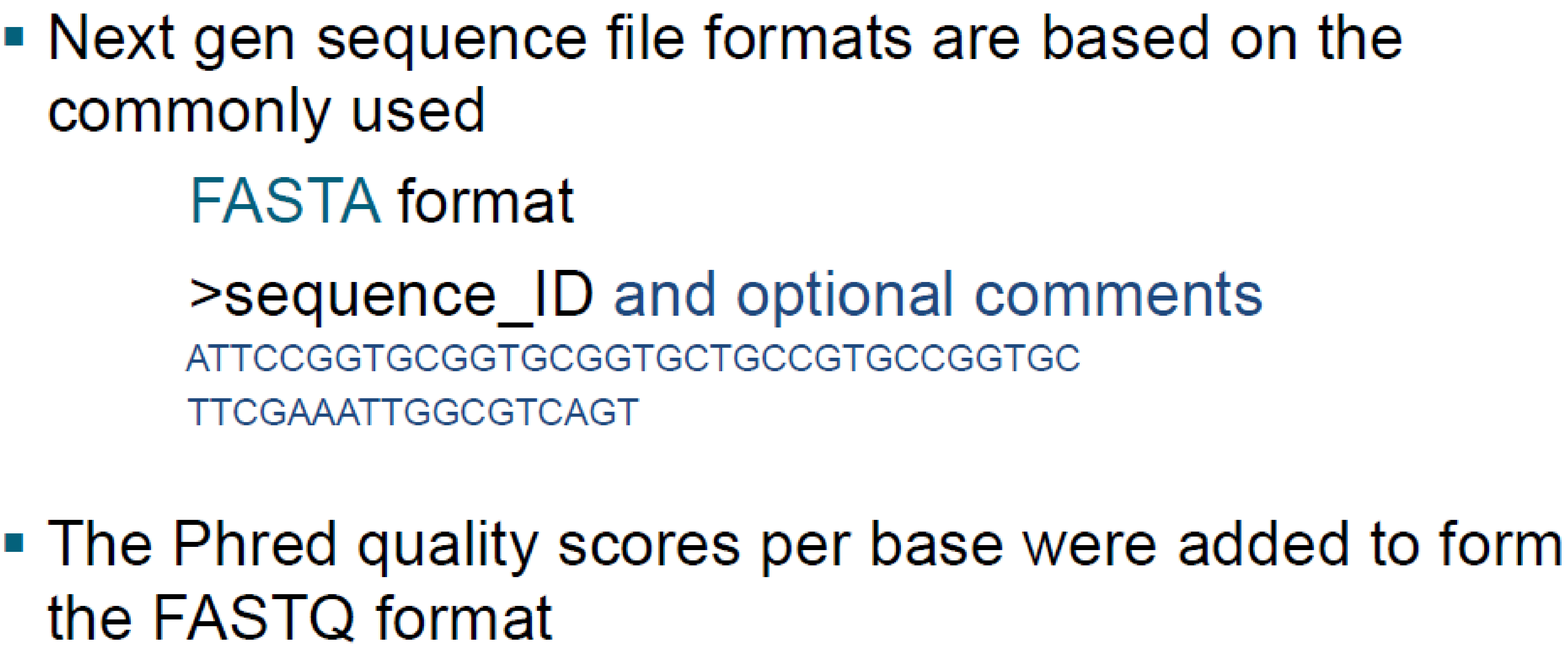

Specimens needed to be sequenced
Pure culture growth: 48 hrs
DNA extraction: 2-4 hrs
DNA quality control: 30 min – 1 hr
Spectrophotometer
DNA library preparation: 4-6 hrs
Sample Preparation Kit
Pooling and loading: 1-1.5 hrs
Template generation and sequencing: 26-36 hrs
Raw data quality control: 20 min

Fig. 2. Typical whole genome sequencing workflow in a clinical or public health laboratory.



Genome
assembly

Sequence file formats



- Next gen sequence file formats are based on the commonly used

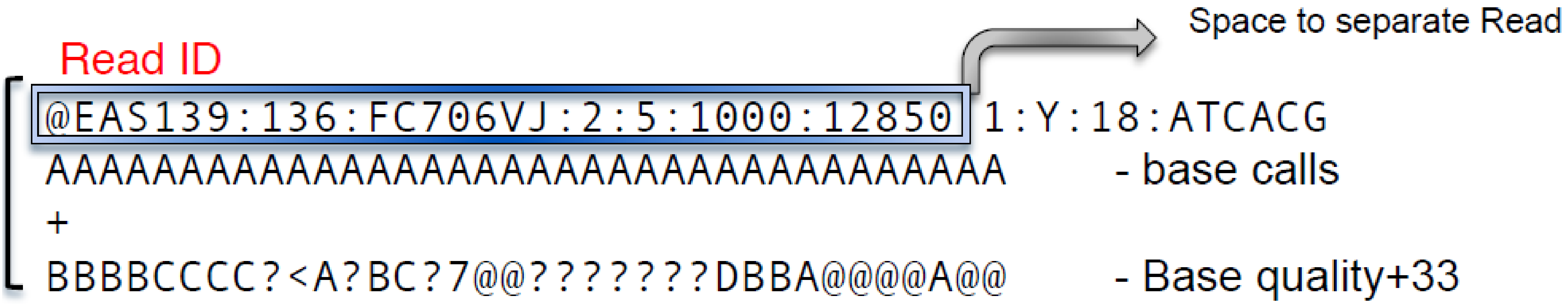
  FASTA format

  >sequence_ID and optional comments
  ATTCCGGTGCGGTGCGGTGCTGCCGTGCCGGTGC
  TTCGAAATTGGCGTCAGT

- The Phred quality scores per base were added to form the FASTQ format

Sequence file formats

- Illumina Fastq format (fasta format with **Q**uality values for each base)



Read ID

Space to separate Read

```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
BBBBCCCC?<A?BC?7@@????????DBBA@@@@A@@
```

- base calls

- Base quality+33

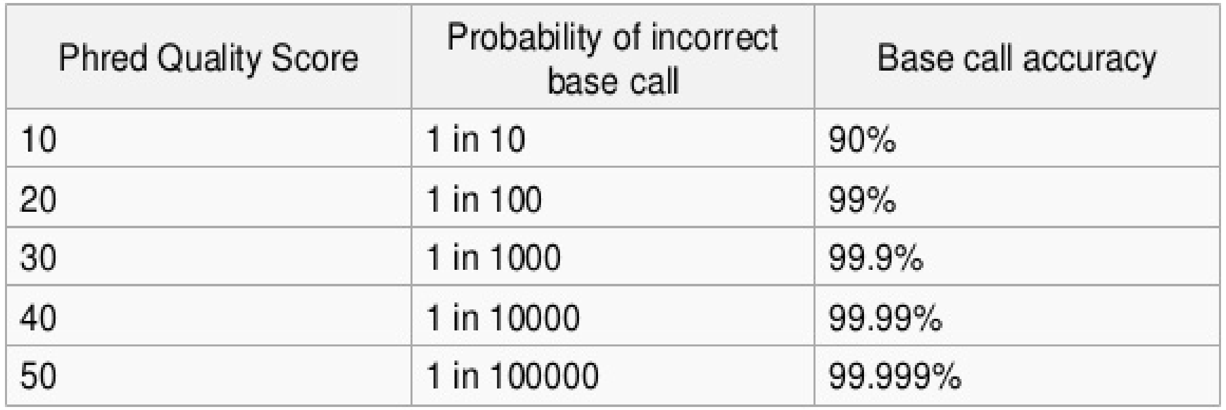
Full read header description

@ <instrument-name>:<run ID>:<flowcell ID>:<lane-number>:<tile-number>: <x-pos>: <y-pos>
<read number>:<is filtered>:<control number>:<barcode sequence>

The phred quality score

Quality score interpretation

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

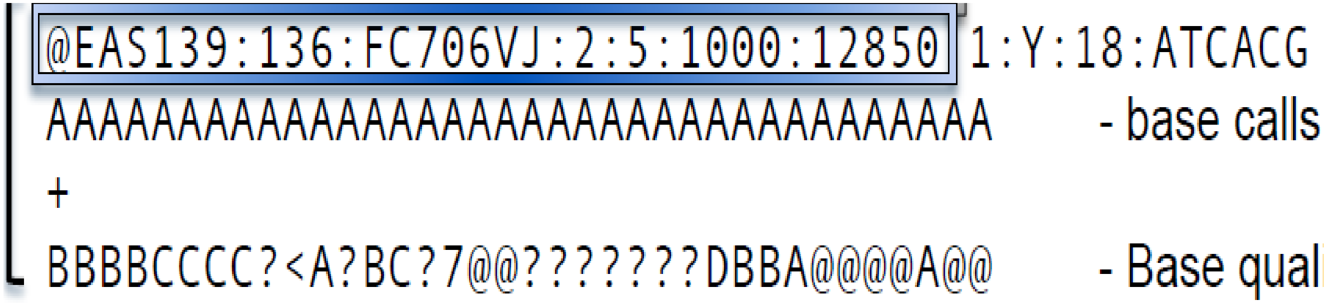


Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

If base quality = 35

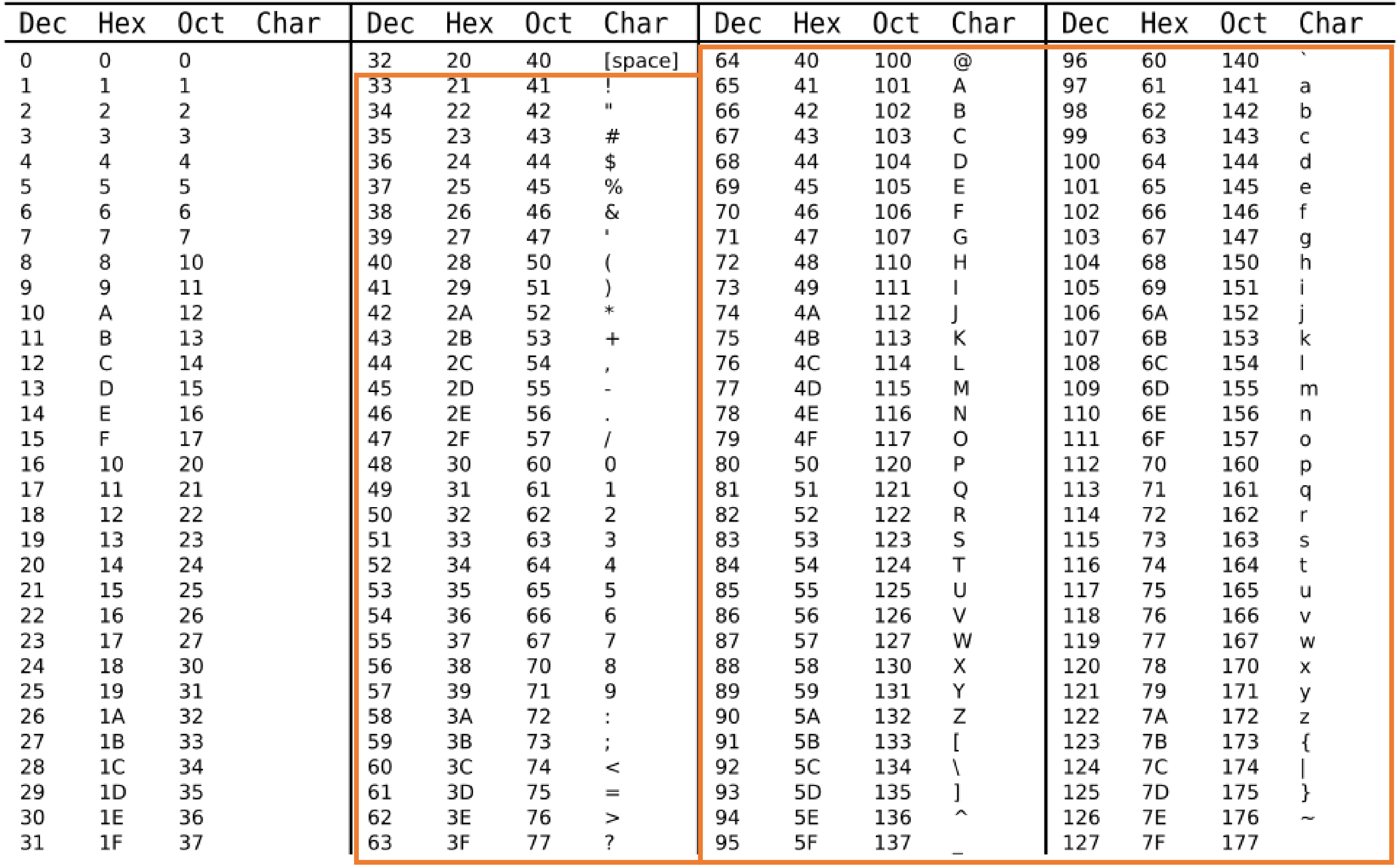
$P=10^{-35/10} = 0.00032$

or 1/3200 incorrect



```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA       - base calls
+
BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@       - Base quality+33
```

ASCII Table



Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char
0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	`
1	1	1		33	21	41	!	65	41	101	A	97	61	141	a
2	2	2		34	22	42	"	66	42	102	B	98	62	142	b
3	3	3		35	23	43	#	67	43	103	C	99	63	143	c
4	4	4		36	24	44	$	68	44	104	D	100	64	144	d
5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
7	7	7		39	27	47	'	71	47	107	G	103	67	147	g
8	8	10		40	28	50	(	72	48	110	H	104	68	150	h
9	9	11		41	29	51	)	73	49	111	I	105	69	151	i
10	A	12		42	2A	52	*	74	4A	112	J	106	6A	152	j
11	B	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	l
13	D	15		45	2D	55	-	77	4D	115	M	109	6D	155	m
14	E	16		46	2E	56	.	78	4E	116	N	110	6E	156	n
15	F	17		47	2F	57	/	79	4F	117	O	111	6F	157	o
16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
19	13	23		51	33	63	3	83	53	123	S	115	73	163	s
20	14	24		52	34	64	4	84	54	124	T	116	74	164	t
21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
23	17	27		55	37	67	7	87	57	127	W	119	77	167	w
24	18	30		56	38	70	8	88	58	130	X	120	78	170	x
25	19	31		57	39	71	9	89	59	131	Y	121	79	171	y
26	1A	32		58	3A	72	:	90	5A	132	Z	122	7A	172	z
27	1B	33		59	3B	73	;	91	5B	133	[	123	7B	173	{
28	1C	34		60	3C	74	<	92	5C	134	\	124	7C	174	|
29	1D	35		61	3D	75	=	93	5D	135	]	125	7D	175	}
30	1E	36		62	3E	76	>	94	5E	136	^	126	7E	176	~
31	1F	37		63	3F	77	?	95	5F	137	_	127	7F	177	

AAAAA

+

BBBBC

ASCII val 66 67

-33 -33

Q value 33 34

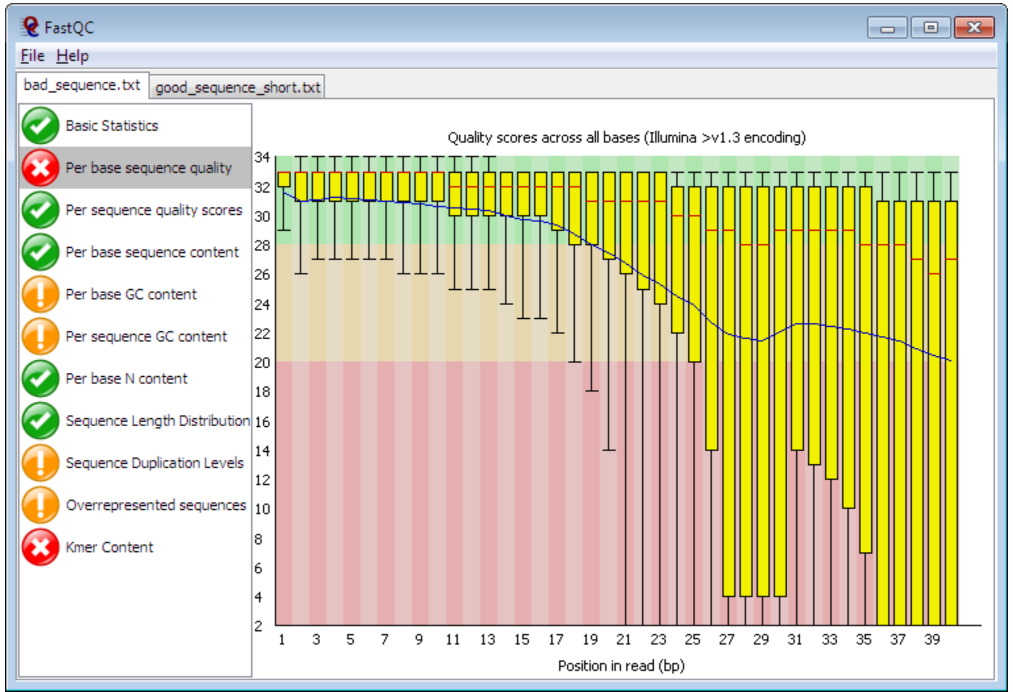
$Q = -10 \ \log_{10} P \longrightarrow P = 10^{\frac{-Q}{10}}$

```
> 10^(-33/10)
[1] 0.0005011872 = 1/5000
> 10^(-34/10)
[1] 0.0003981072=1/39000
```

Let’s play with FasQC to quality control visualization

Open FastQC program

Open in browser:
fastqc_report.html



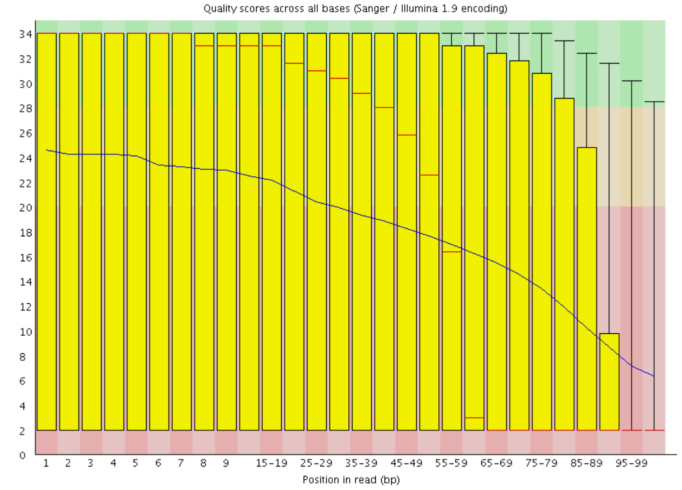
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content

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

Quality filter trim galore

Babraham Bioinformatics

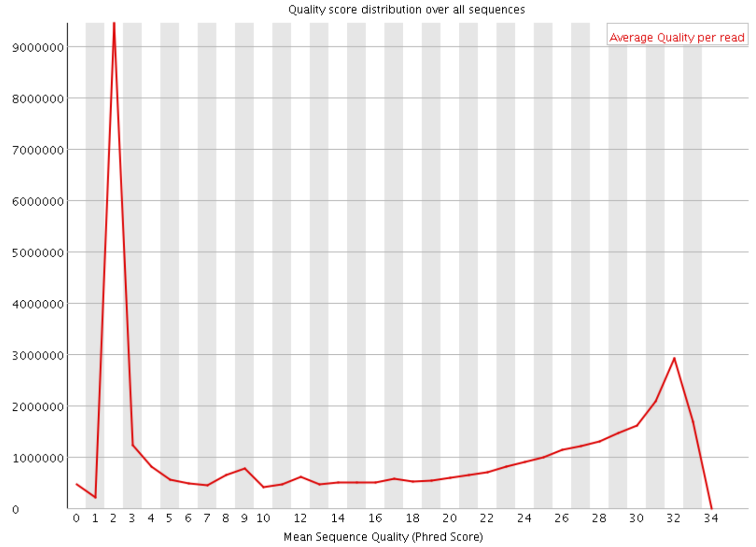
Before trimGalore



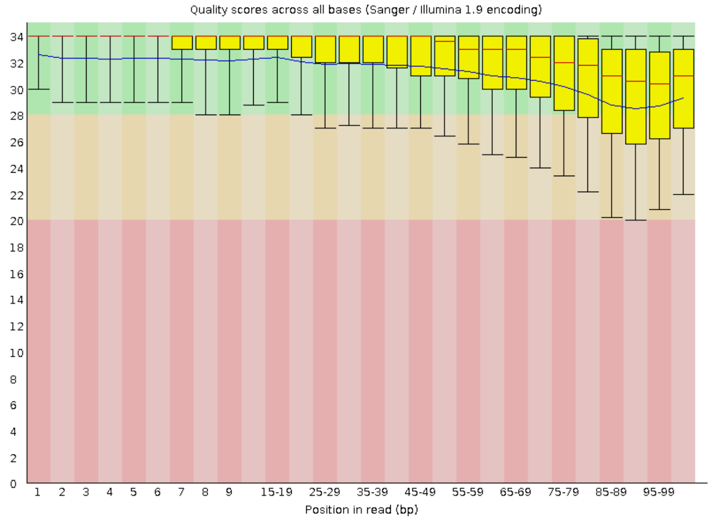
Good

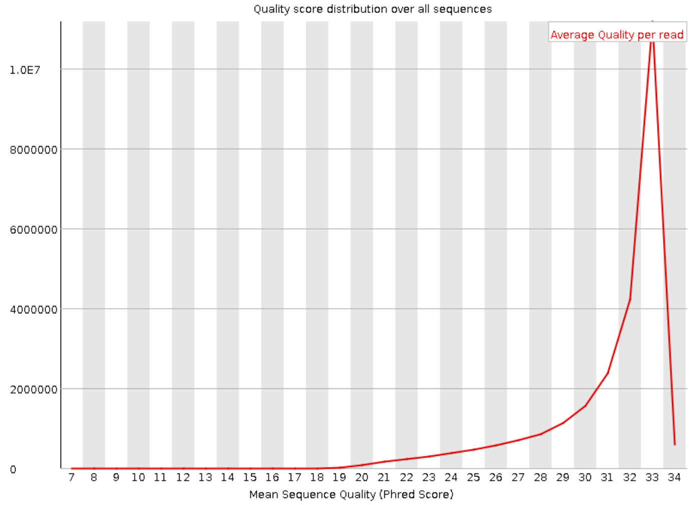
Reasonable

bad



After trimGalore


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


Quality score distribution over all sequences
Average Quality per read
1.0E7
Mean Sequence Quality (Phred Score)