WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

World Health Organization
Philippines

KDCA
Korea Disease Control and Prevention Agency

NGS DATA QUALITY CONTROL

FRANCIS A. TABLIZO
Core Facility for Bioinformatics
Philippine Genome Center
University of the Philippines System

April 15, 2024

CENTRE FOR PATHOGEN GENOMICS

Doherty Institute

THE UNIVERSITY OF MELBOURNE

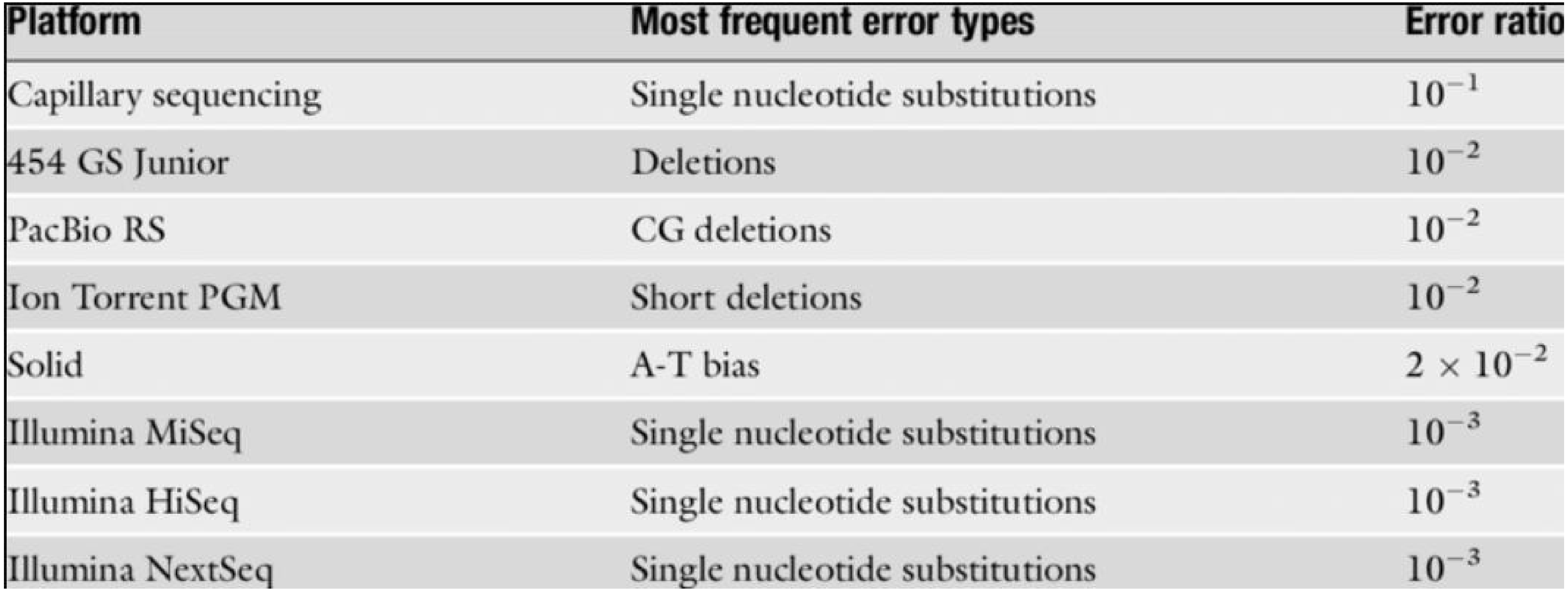
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Objectives

- Learn how to assess the quality of NGS data
- Implement quality control procedures to minimize the incorporation of sequencing errors in downstream analyses

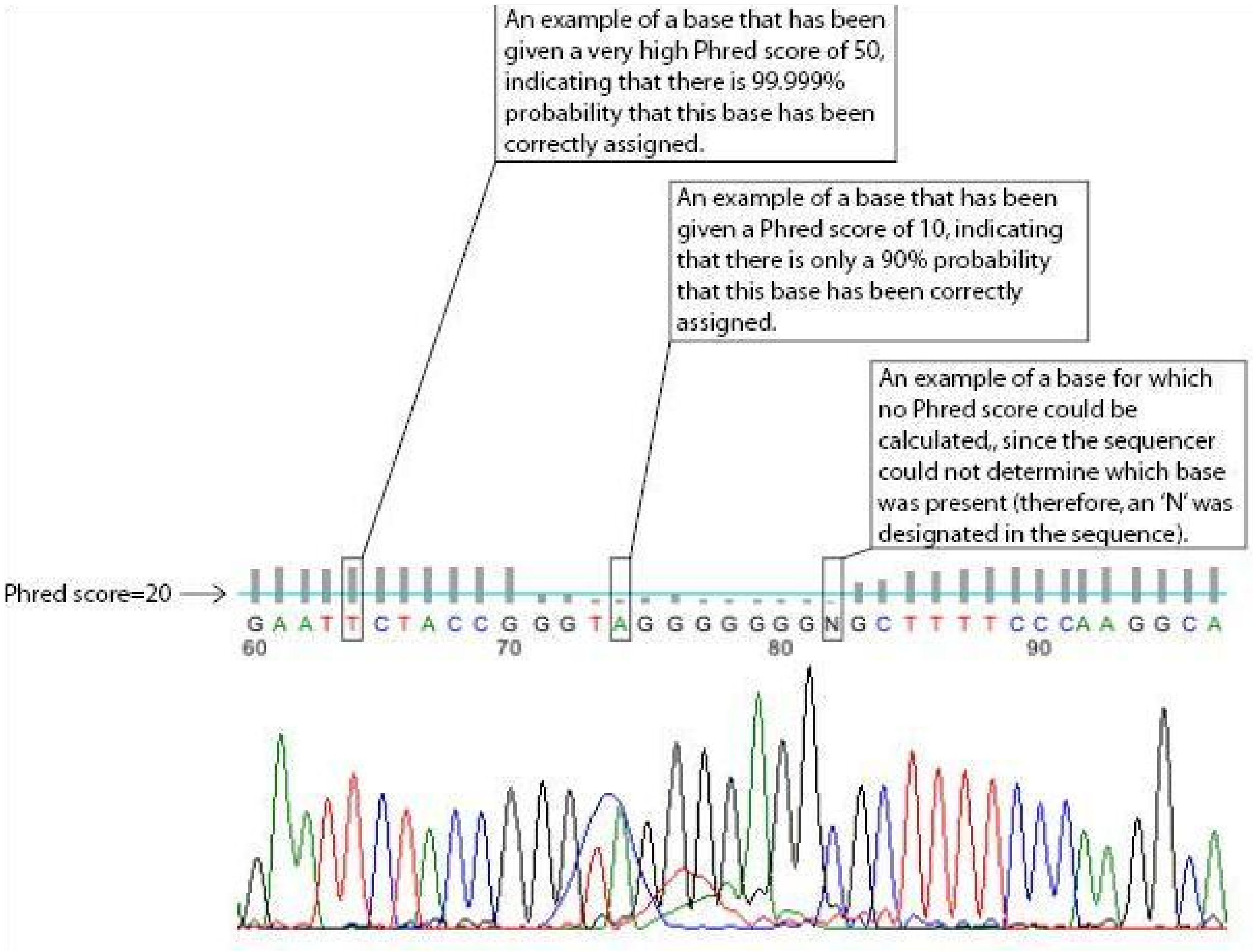
Sequencing Error Rates



Platform	Most frequent error types	Error ratio
Capillary sequencing	Single nucleotide substitutions	10^{-1}
454 GS Junior	Deletions	10^{-2}
PacBio RS	CG deletions	10^{-2}
Ion Torrent PGM	Short deletions	10^{-2}
Solid	A-T bias	2×10^{-2}
Illumina MiSeq	Single nucleotide substitutions	10^{-3}
Illumina HiSeq	Single nucleotide substitutions	10^{-3}
Illumina NextSeq	Single nucleotide substitutions	10^{-3}

https://www.researchgate.net/profile/Shifu_Chen/publication/323733012/figure/tbl1/AS:688075940302849@1541061496360/A-comparison-of-sequencing-error-ratios-of-different-sequencing-platforms.png

Chromatogram



An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

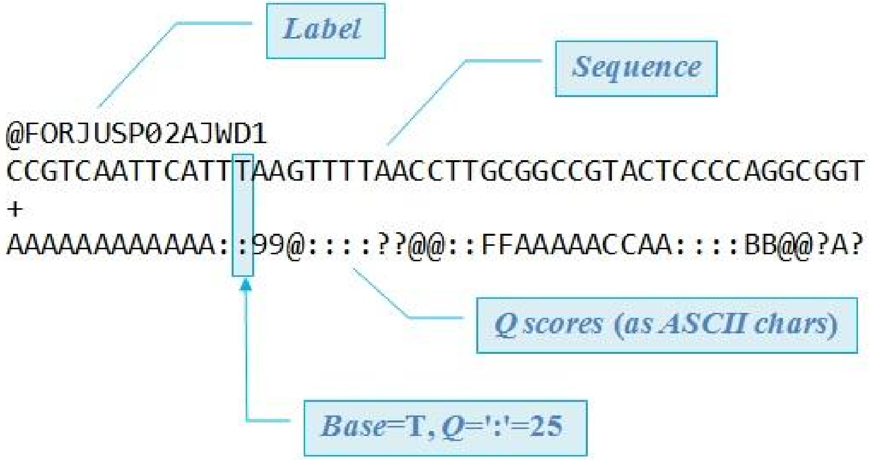
An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

Phred score=20
G A A T T C T A C C G G G T A G G G G G G N G C T T T T C C C A A G G C A
60
70
80
90

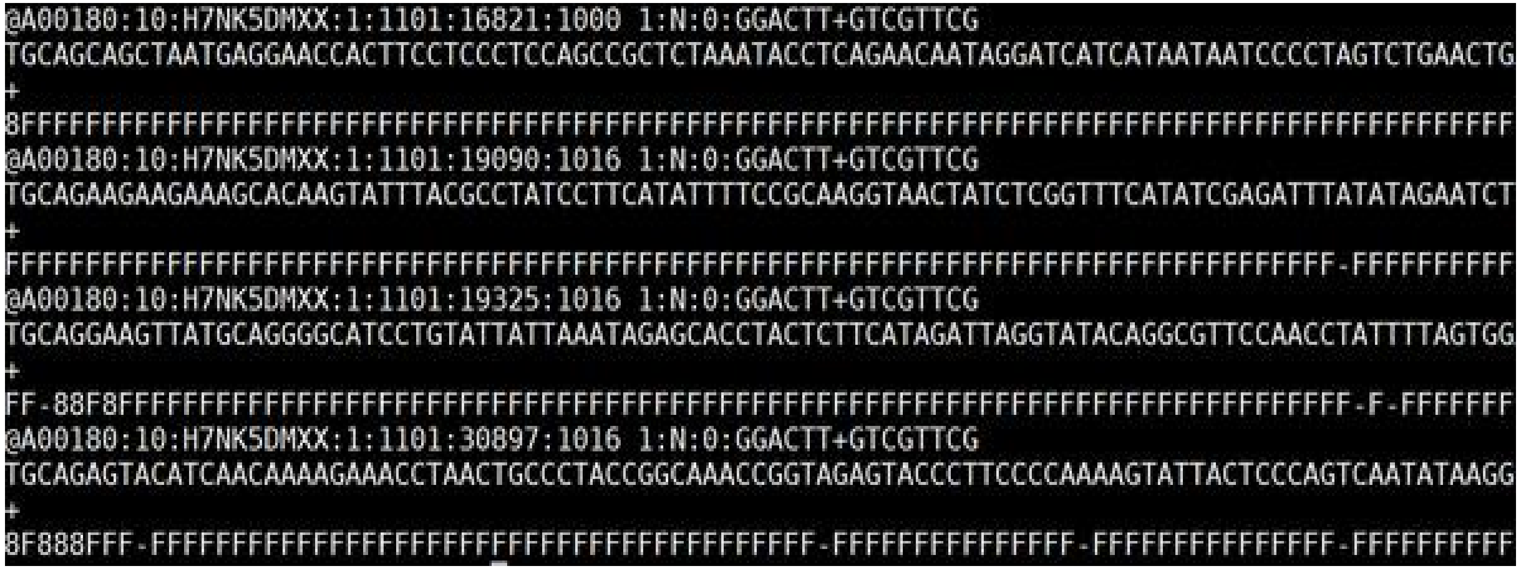
http://upload.wikimedia.org/wikipedia/en/e/e7/Phred_Figure_1.jpg

FASTQ File



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

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

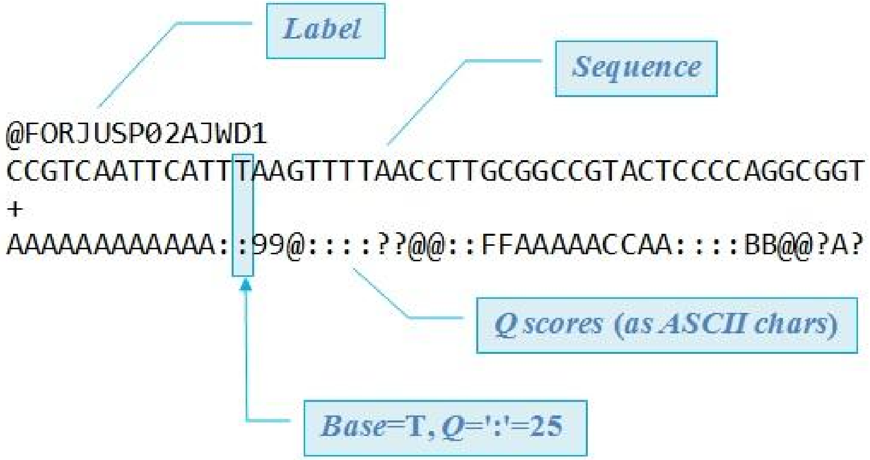


```
@A00180:10:H7NK5DMXX:1:1101:16821:1000 1:N:0:GGACTT+GTCGTTCG
TGCAGCAGCTAATGAGGAACCACTTCCTCCCTCCAGCCGCTCTAAATACCTCAGAACAATAGGATCATCATAATAATCCCCTAGTCTGAACTG
+
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19090:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTTTCATATCGAGATTTATATAGAATCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19325:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATACAGGCGTTCCAACCTATTTTAGTGG
+
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-F-FFFFFFF
@A00180:10:H7NK5DMXX:1:1101:30897:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAAAGTATTACTCCCAGTCAATATAAGG
+
8F888FFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFF
```

https://www.reneshbedre.com/blog/fqqualfmt.html

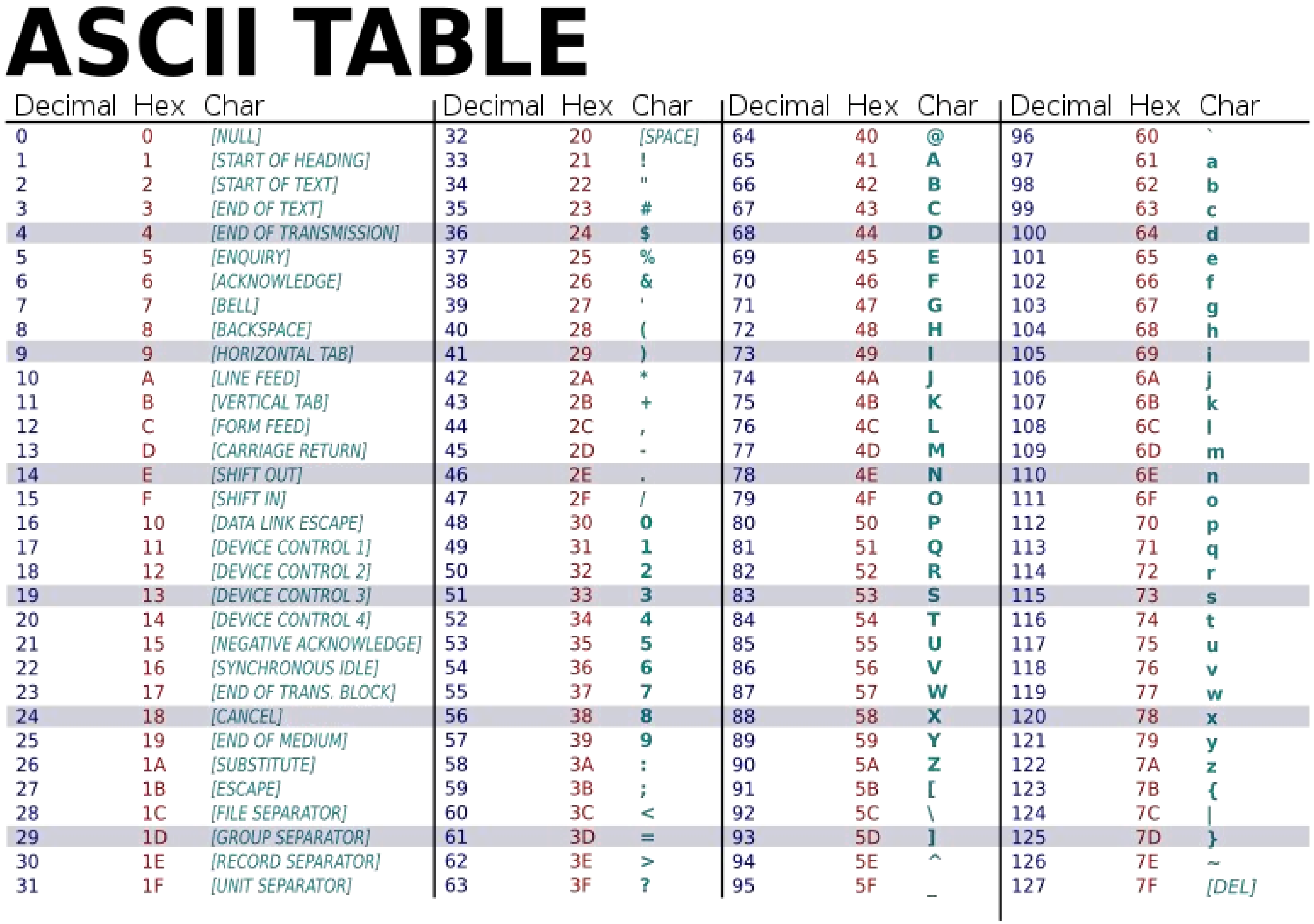
FASTQ File

A joint venture between The University of Melbourne and The Royal Melbourne Hospital



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

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

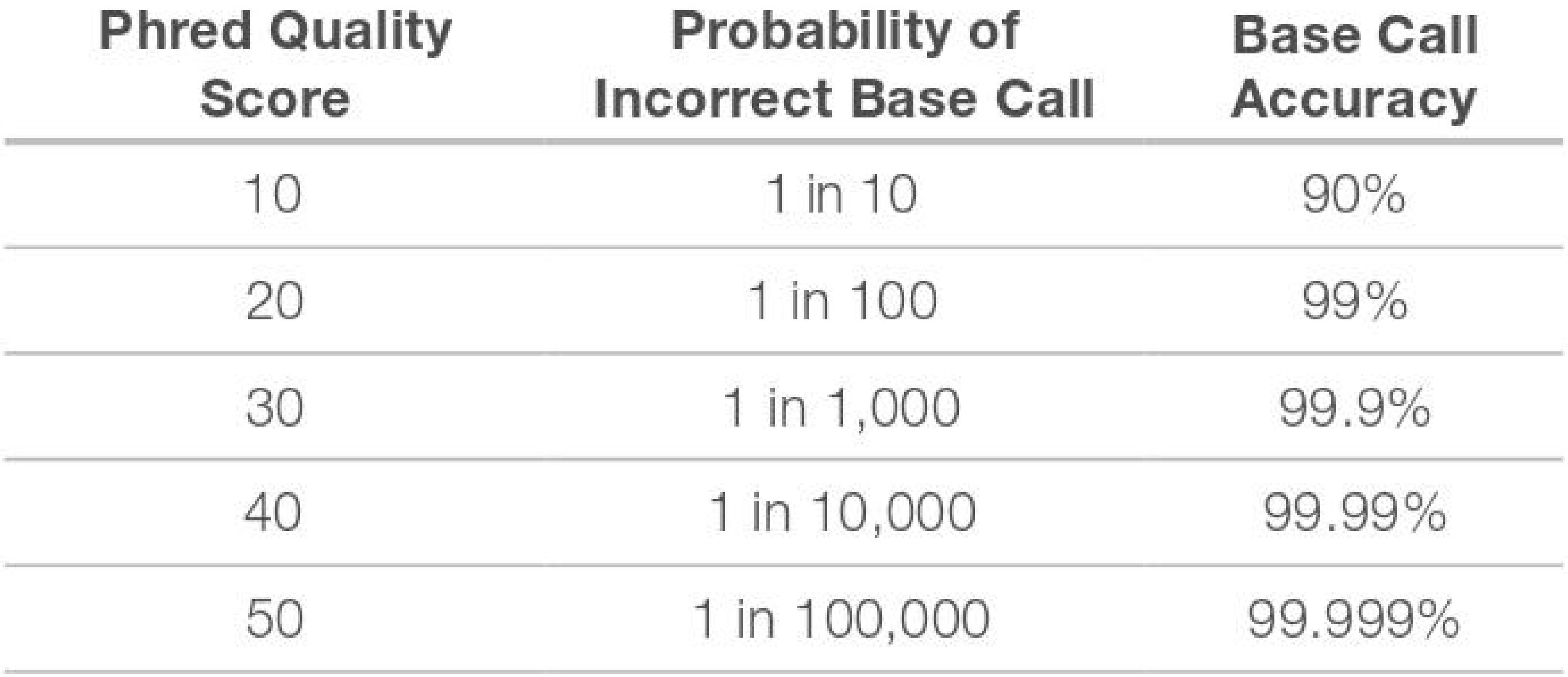


ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[END OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

https://simple.m.wikipedia.org/wiki/File:ASCII-Table-wide.svg

Base Call Quality



Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Sequence Quality Control

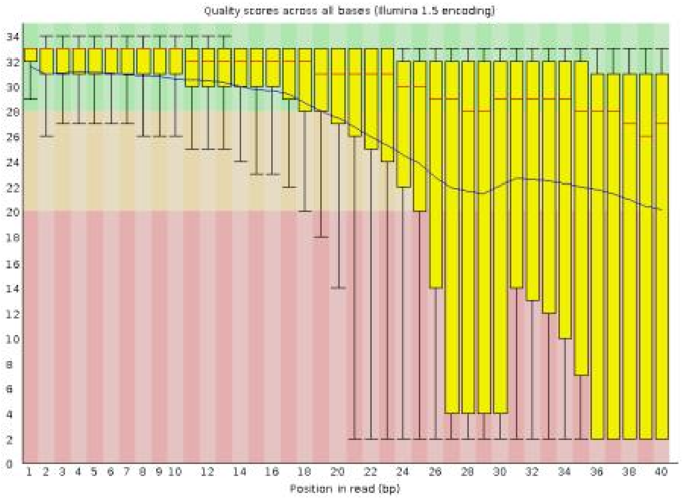
General Steps

1. Initial sequence quality assessment
2. Adapter clipping
3. Trimming of low quality sequence ends
4. Read quality filtering
5. Pairing of reads (for paired-end reads)
6. Final sequence quality assessment

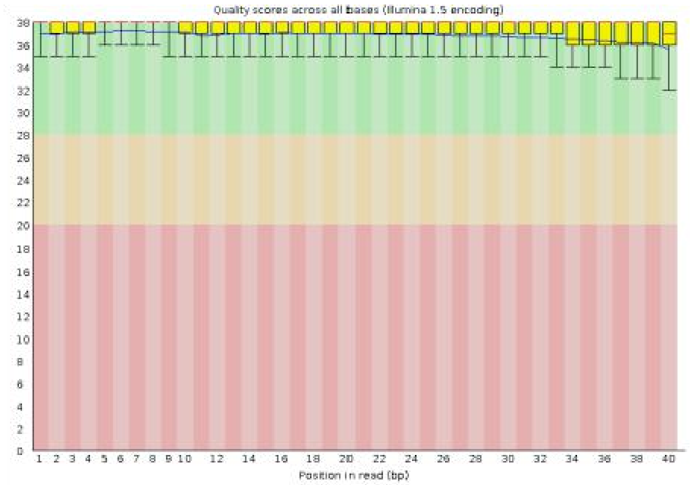
Sequence Quality Control

Initial Quality Assessment

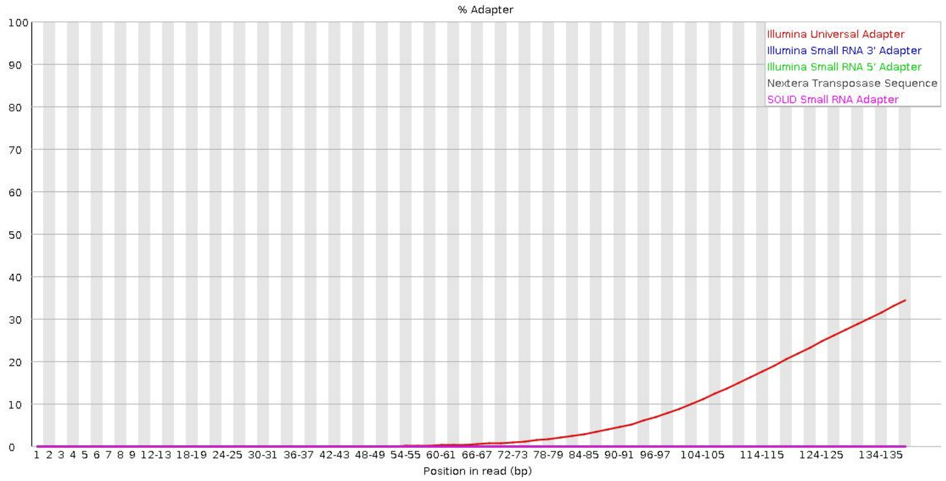
👎



👍



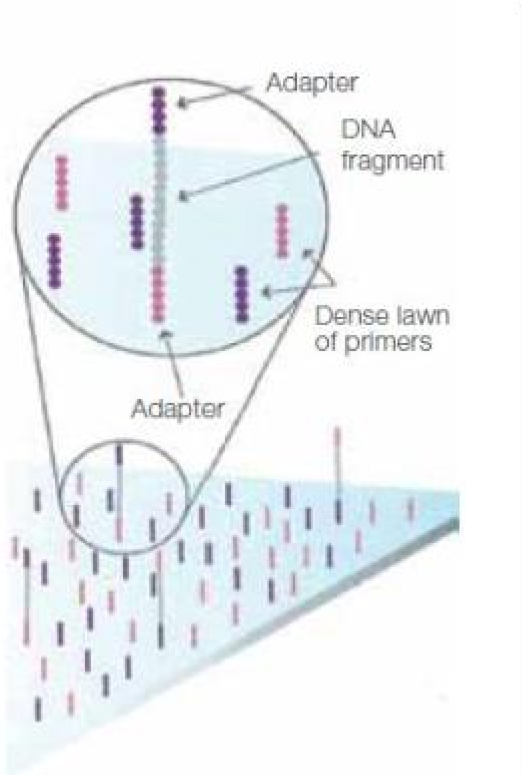
✖ **Adapter Content**

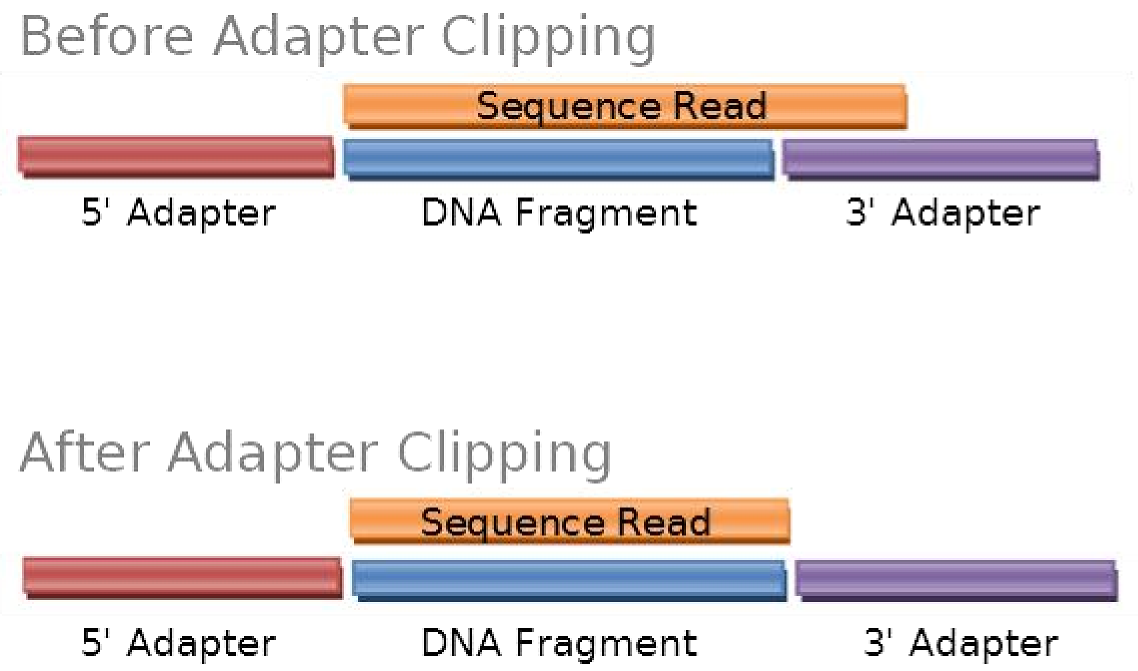


http://www.bioinformatics.babraham.ac.uk/projects/fastqc
https://www.biostars.org/p/463741/

Sequence Quality Control

Adapter Clipping

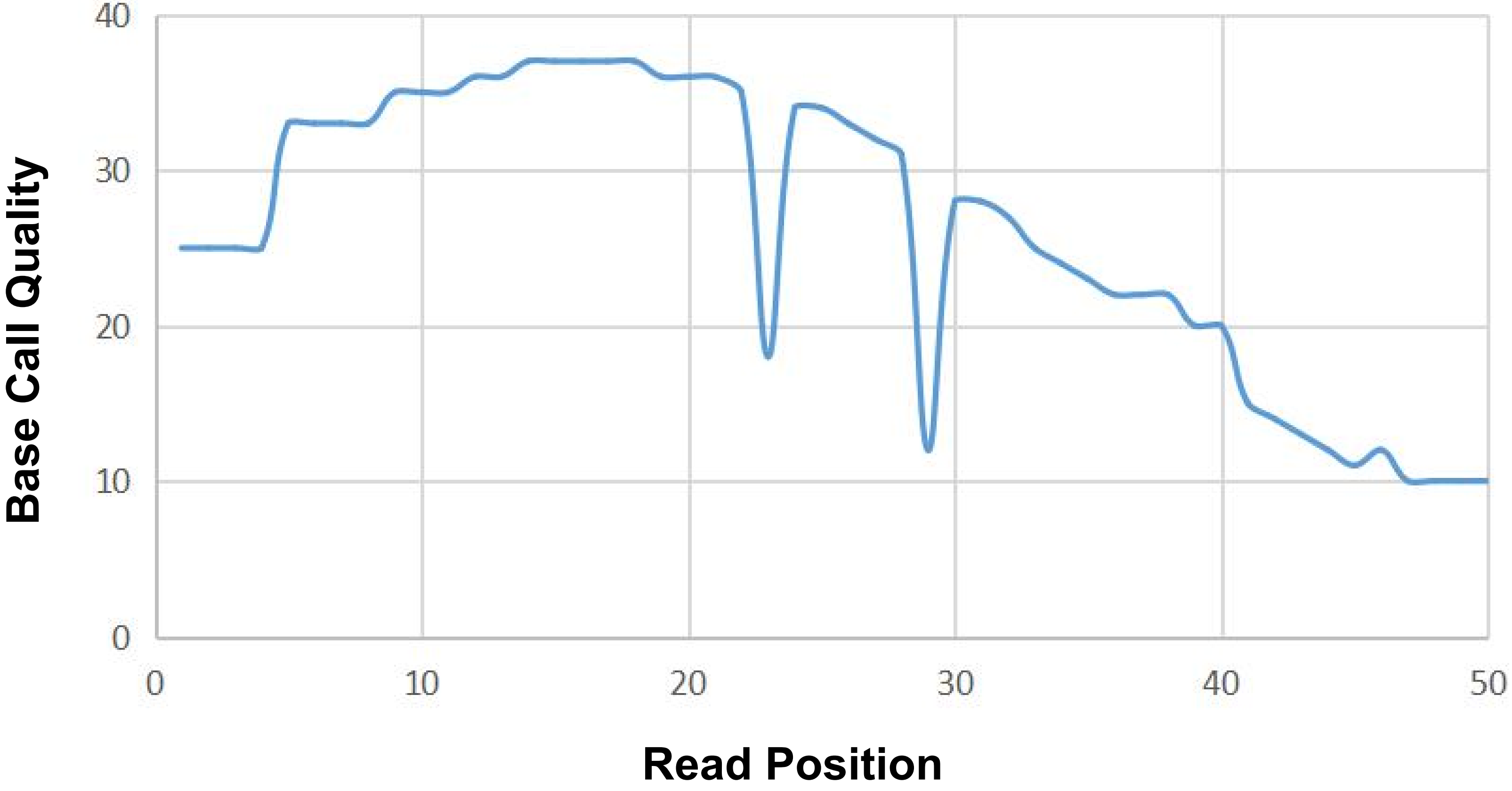




http://www.eisenlab.org/FunFly/wp-content/uploads/2012/06/illumina-attach-dna.jpg;
http://srna-workbench.cmp.uea.ac.uk/wp-content/uploads/2012/01/AR3.png

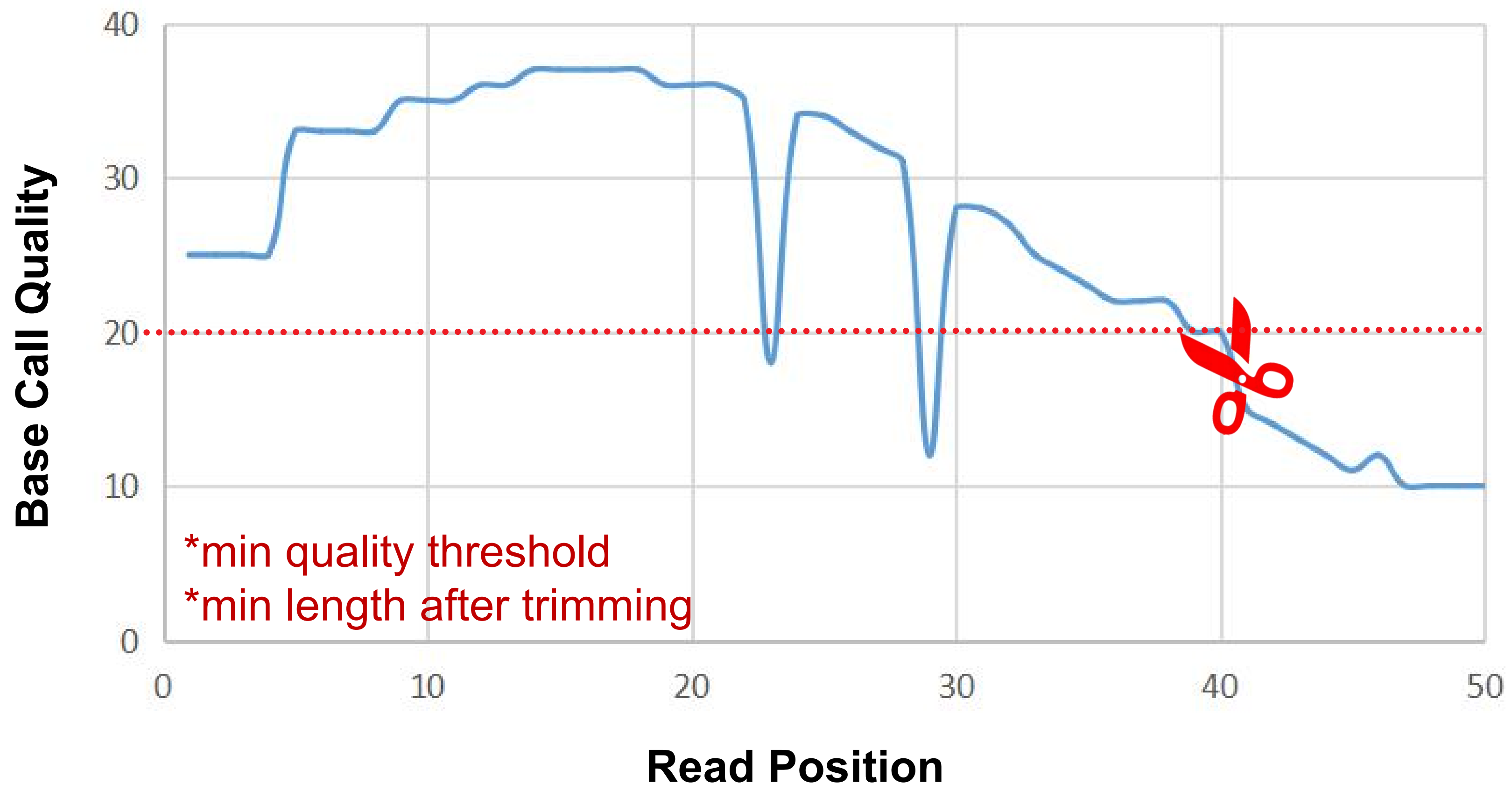
Sequence Quality Control

Trimming of Low Quality Sequence Ends



Sequence Quality Control

Trimming of Low Quality Sequence Ends



*min quality threshold
*min length after trimming

Sequence Quality Control

CENTRE FOR PATHOGEN GENOMICS

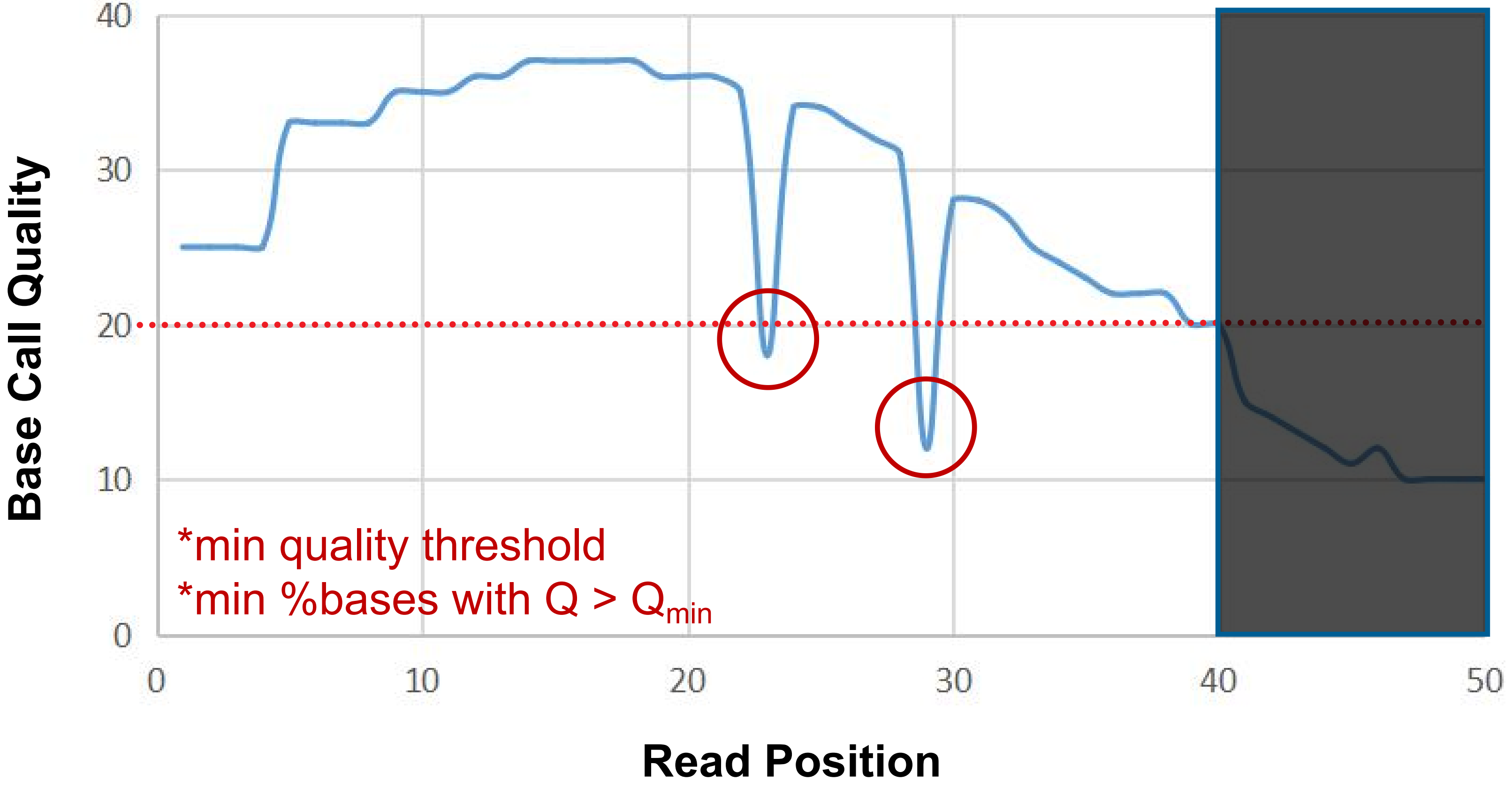
Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Quality Filtering

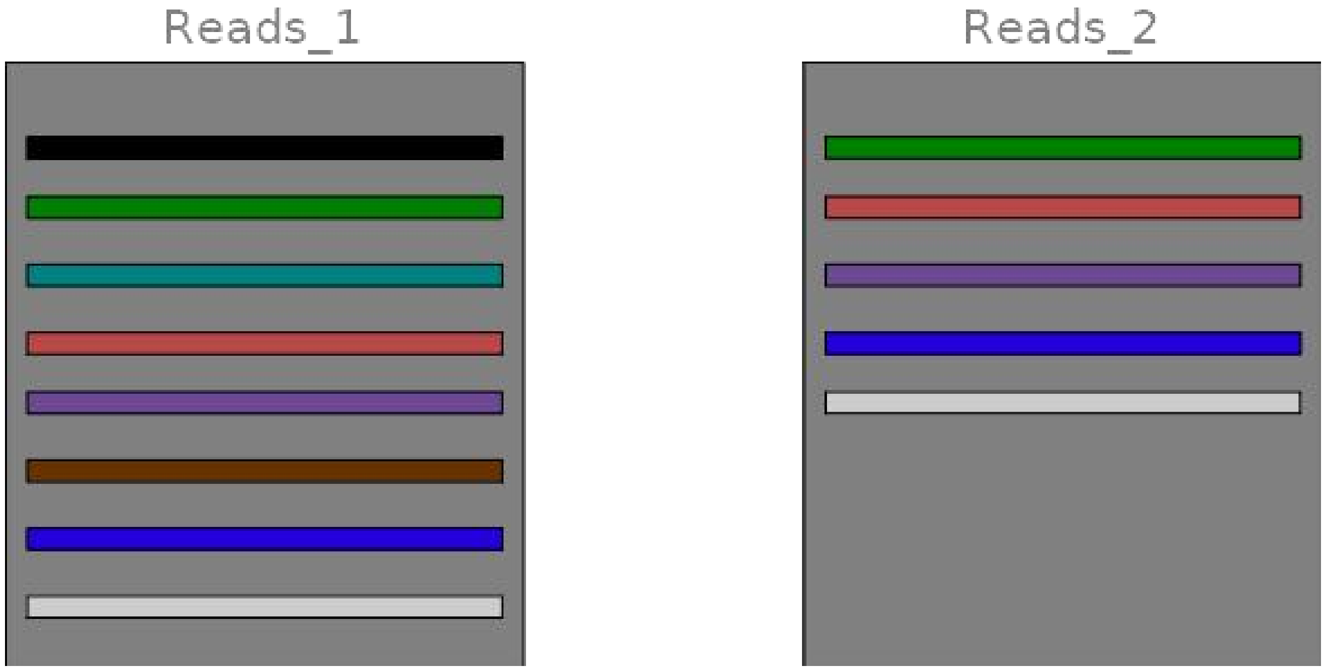


*min quality threshold

*min %bases with Q > Q_{min}

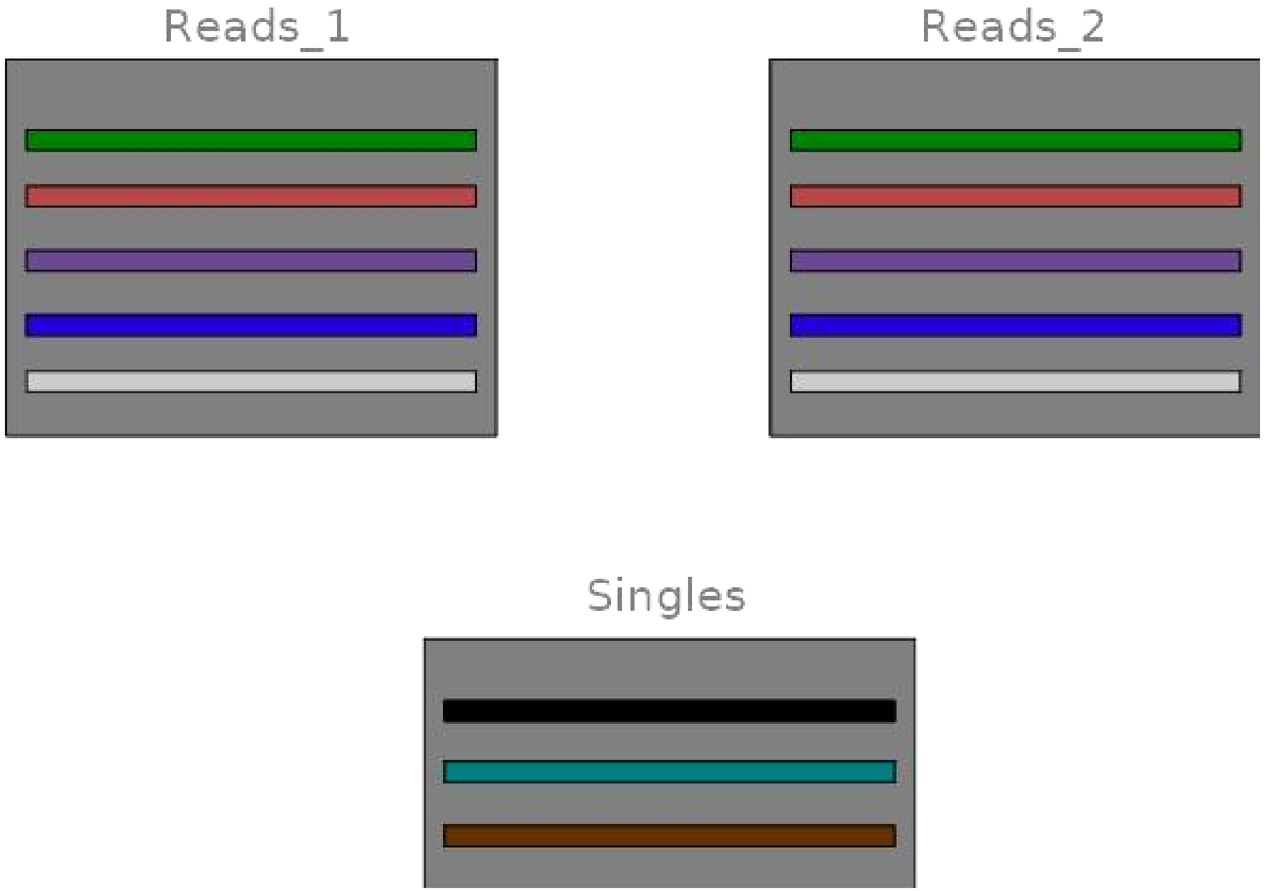
Sequence Quality Control

Read Pairing (For Paired-End Data)



Sequence Quality Control

Read Pairing (For Paired-End Data)



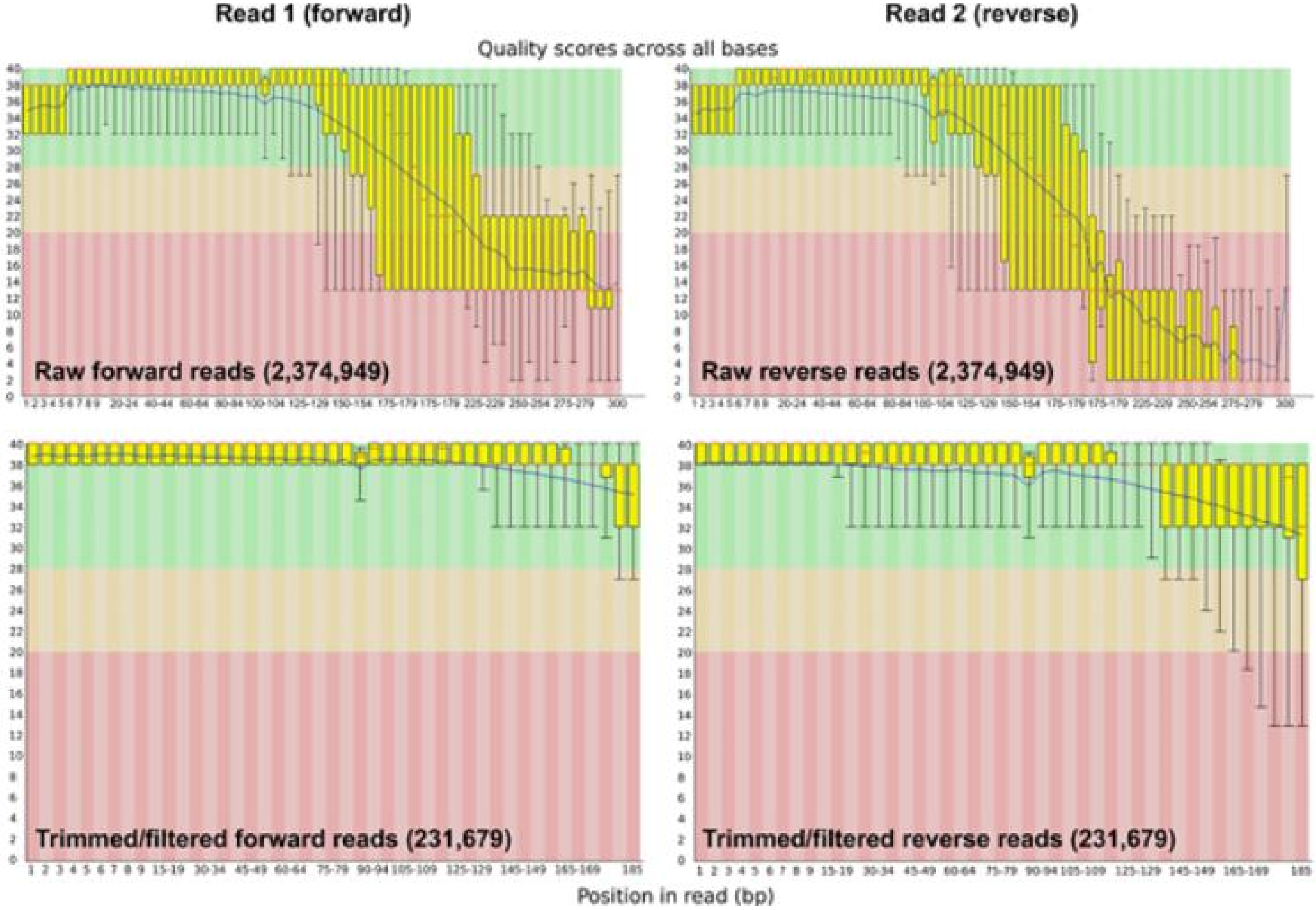
Reads_1

Reads_2

Singles

Sequence Quality Control

Final Quality Assessment



https://www.researchgate.net/profile/Richard-Tennant/publication/312355161/figure/fig2/AS:450870568591361@1484507328098/Sequence-Quality-Per-base-Before-and-After-Trimming-and-Adapter-Removal-The-per-base.png

World Health Organization
Philippines

CENTRE FOR PATHOGEN GENOMICS

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

QUESTIONS?

fatablizo@up.edu.ph

bioinformatics@pgc.up.edu.ph

KDCA
Korea Disease Control and Prevention Agency

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital