Genetic Diversity Centre Zurich

701-1425-00L - Genetic Diversity: Analysis

NGS: Quality Control

Friday, June 19, 2019

Jean-Claude Walser

jean-claude.walser@env.ethz.ch

Genetic Diversity Centre Zurich

Sample Quality Control

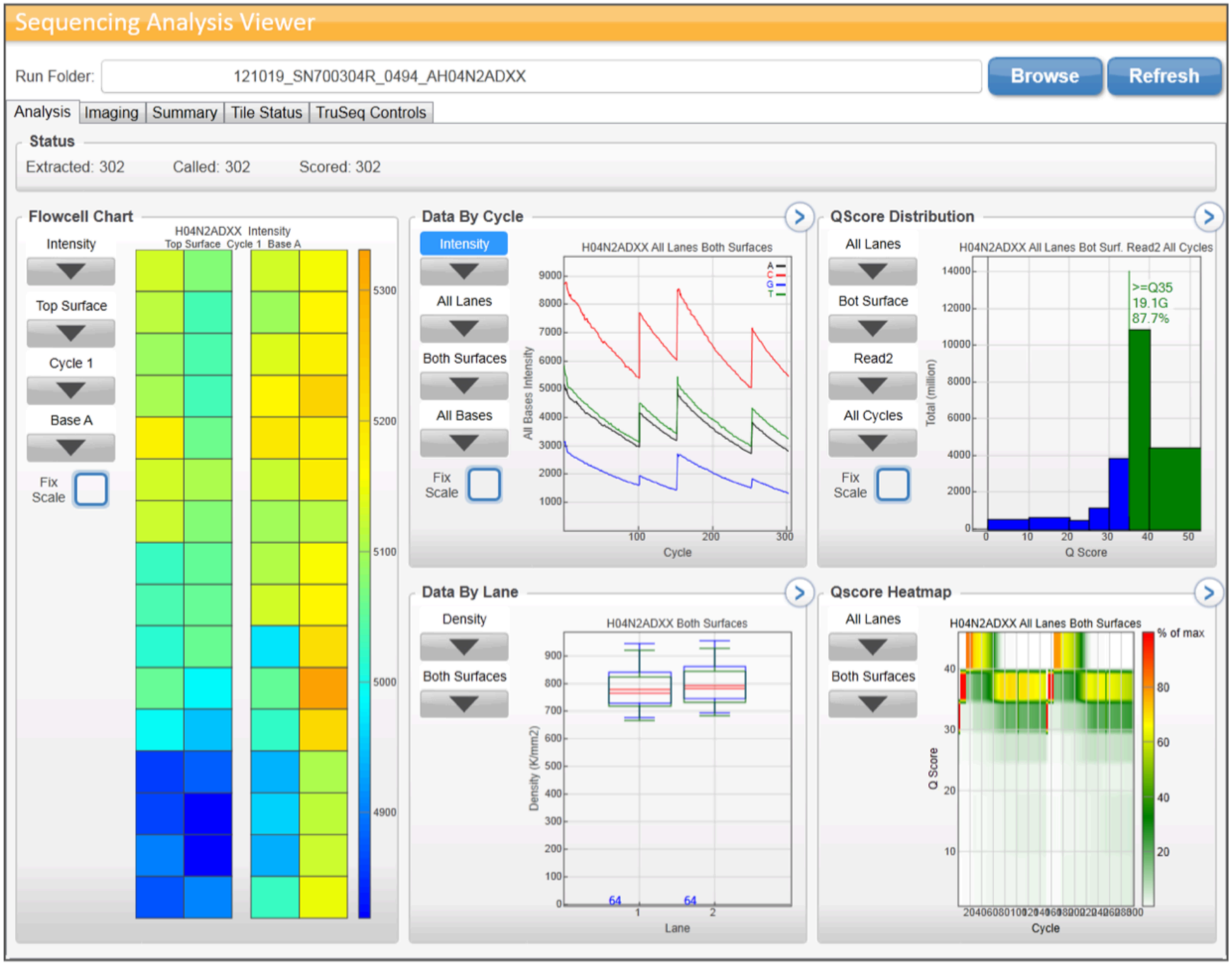
Library Quality Control

Run Quality Control

Sequencing Quality Control

Outlier Control

RUN QUALITY CONTROL


Sequencing Analysis Viewer
Run Folder: 121019_SN700304R_0494_AH04N2ADXX
Browse
Refresh
Analysis
Imaging
Summary
Tile Status
TruSeq Controls
Status
Extracted: 302
Called: 302
Scored: 302
Flowcell Chart
Intensity
Top Surface
Cycle 1
Base A
Fix Scale
H04N2ADXX Intensity
Top Surface Cycle 1 Base A
5300
5200
5100
5000
4900
Data By Cycle
Intensity
All Lanes
Both Surfaces
All Bases
Fix Scale
H04N2ADXX All Lanes Both Surfaces
All Bases Intensity
Cycle
QScore Distribution
All Lanes
Bot Surface
Read2
All Cycles
Fix Scale
H04N2ADXX All Lanes Bot Surf. Read2 All Cycles
Total (million)
>=Q35
19.1G
87.7%
Q Score
Data By Lane
Density
Both Surfaces
H04N2ADXX Both Surfaces
Density (K/mm2)
64
64
Lane
Qscore Heatmap
All Lanes
Both Surfaces
H04N2ADXX All Lanes Both Surfaces
% of max
Q Score
Cycle

Illumina Sequencing Analysis Viewer 1.8.37 - 151204_M03930_0010_000000000-AK6GL

Sequencing Analysis Viewer

Run Folder: D:\Illumina\MiSeqAnalysis\151204_M03930_0010_000000000-AK6GL

Browse | Refresh

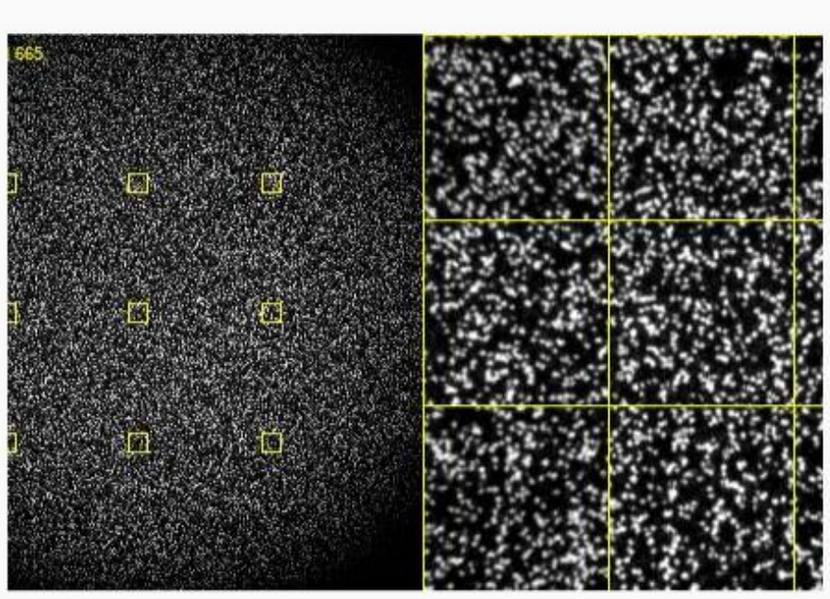
Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

Cycle All | Lane All | Surface Both | Swath All

A C G T

Index	Lane	Tile	Section	Cycle	Surface	Swath	Time	P90 A
1	1	1101	1	1	Top	1	12/04/201...	363
2	1	1101	1	2	Top	1	12/04/201...	369
3	1	1101	1	3	Top	1	12/04/201...	378
4	1	1101	1	4	Top	1	12/04/201...	382
5	1	1101	1	5	Top	1	12/04/201...	375
6	1	1101	1	6	Top	1	12/04/201...	381
7	1	1101	1	7	Top	1	12/04/201...	392
8	1	1101	1	8	Top	1	12/04/201...	381
9	1	1101	1	9	Top	1	12/04/201...	406
10	1	1101	1	10	Top	1	12/04/201...	396
11	1	1101	1	11	Top	1	12/04/201...	404
12	1	1101	1	12	Top	1	12/04/201...	399
13	1	1101	1	13	Top	1	12/04/201...	400
14	1	1101	1	14	Top	1	12/04/201...	402
15	1	1101	1	15	Top	1	12/04/201...	412
16	1	1101	1	16	Top	1	12/04/201...	426
17	1	1101	1	17	Top	1	12/04/201...	423

Rows=14504 | Disp=14504 | Sel=1 | Filter



Cluster Density: 1017 K/mm2 (Optimal 1200-1400 k/mm2)
Reads Total: 27.69 M (goal 30 M)
Reads PF: 21.60 M
PhiX Conc: 2.03 % (loaded 2%)
%>=Q30: Total 63.06% (should be at least 70%)

- The **density** of clusters for each tile (in thousands per mm2) and the number of **clusters** for each tile (in millions).
- Total **yield** is the number of bases generated in the run.
- The calculated **error rate**, as determined by a spiked in PhiX control sample if available and it refers to the percentage of bases called incorrectly at any one cycle.
- The total fraction of passing filter reads (**PF**) assigned to an index.
- **% Q-score** >= Q30 (percentage of bases that have a Q-score above or equal to 30; Q30 is a probability of incorrect base calling of 1 in 1000).
- The **signal to noise ratio** is calculated as mean called intensity divided by standard deviation of non-called intensities. Not calculated for NextSeq two-channel sequencing or HiSeq X.
- The percentage of molecules in a cluster for which sequencing falls behind (**phasing**) or jumps ahead (**prephasing**) the current cycle within a read.

NGS DATA FORMAT[S]

Genetic Diversity Centre Zurich

- Fasta
- Fastq (Fasta with Quality - Illumina)
- Bam (PacBio)
- Fast5 (HDF5 - ONT)

GDC Genetic Diversity Centre Zurich

Sequence Data Format: **Fasta** (**>**)

Start

Unique Sequence Header

```
1 - >BY999847.1 BY999847 Moon Jellyfish cDNA library Aurelia aurita cDNA
    clone Aa_plW_142145_H14, mRNA sequence
2 - AAAATACCGCATGATTGTTCGTTTCACAAACAAAGATATAGCTTGCCAGATAGCGTATGCCAGATTGCAA
3 - GGAGATGTGATCATTTGTGCAGCTTATGCTCATGAACTCCCAAGATATGGTGTCAAGGTCGGGTTGACCA
4 - ACTATGCAGCTGCTTATTGCACTGGCCTCTTGCTCGCAAGAAGGCTCCTTTCAAAATTGAAATTGGCTGA
5 - CACTTACAAAGGTTGTGAAGAAGTGAATGGTGATGAATACCTTGTGGAAGGAGAGGAGGGACAGCCTGGA
6 - CCTTTCCGTTGTTACCTTGATATTGGCCTTGCCAGAACCTCAACTGGTGCCAAGATCTTTGGTGCATTGA
7 - AAGGTGCAGTTGATGGTGGACTTGACATCCCACACAGCAACACGAGATTCCCTGGTTATGACAATGAAGC
8 - AAAGGAATTTGACCCAGAGGTGCACAGACAACACA…
...
```

Sequence (nucleotide or protein)

File Suffix: sequence(s).fa, sequence(s).fasta

Special cases: sequences.mfa (multiple - aligned - sequences)
sequences.afa (aligned sequences)

Sequence Data Format: **Fastq** (@)

Sequence (Read) Header

Start

Nucleotide Sequence (A, C, G, T, N)

+Sequence ID

```
1 @HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG
2 GCCCAGCGTGGGCGAGCCGCACGGCACCATCCTCTGGCACACCCTCTCCTC
3 +
4 BCCFFFFDFHHHFJJJJJJJJIIJJJJJIJIIJJGIHHHHHHFFFDDDEDDDC
@HWI-ST486:166:C06K9ACXX:7:1101:2519:1936 1:N:0:ACAGTG
NTGCACACGAATTCGCCGTTGTGGCAGCTCGAGTACCTGTGGTTCGCCGAG
+
#1=DDDFFHHHFFIGIIJJJJIIJIIJ;?*?.?/9*88BBBF.;7@@E###
```

$N_{rows} = 4$

ASCII encoded quality scores per base

File Suffix: reads.fq, reads.fastq

Special cases: read_R[12].fq (> paired reads)

read_I[12].fq (> index)

GDC Genetic Diversity Centre Zurich

Current Fastq Header Format (**version > 1.8**)

Sequence Header +Sequence ID

a b c d e f g h i j k

@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG

a. **unique instrument name**
b. run id
c. flowcell id
d. flowcell lane
e. tile number within the flowcell lane
f. x-coordinate of the cluster within the tile
g. y-coordinate of the cluster within the tile
h. **the member of a pair, 1 or 2 (paired-end or mate-pair reads only)**
i. Y if the read fails filter (read is bad), N otherwise (read passed filter)
j. 0 when no control bits are on
k. index sequence

GDC Genetic Diversity Centre Zurich

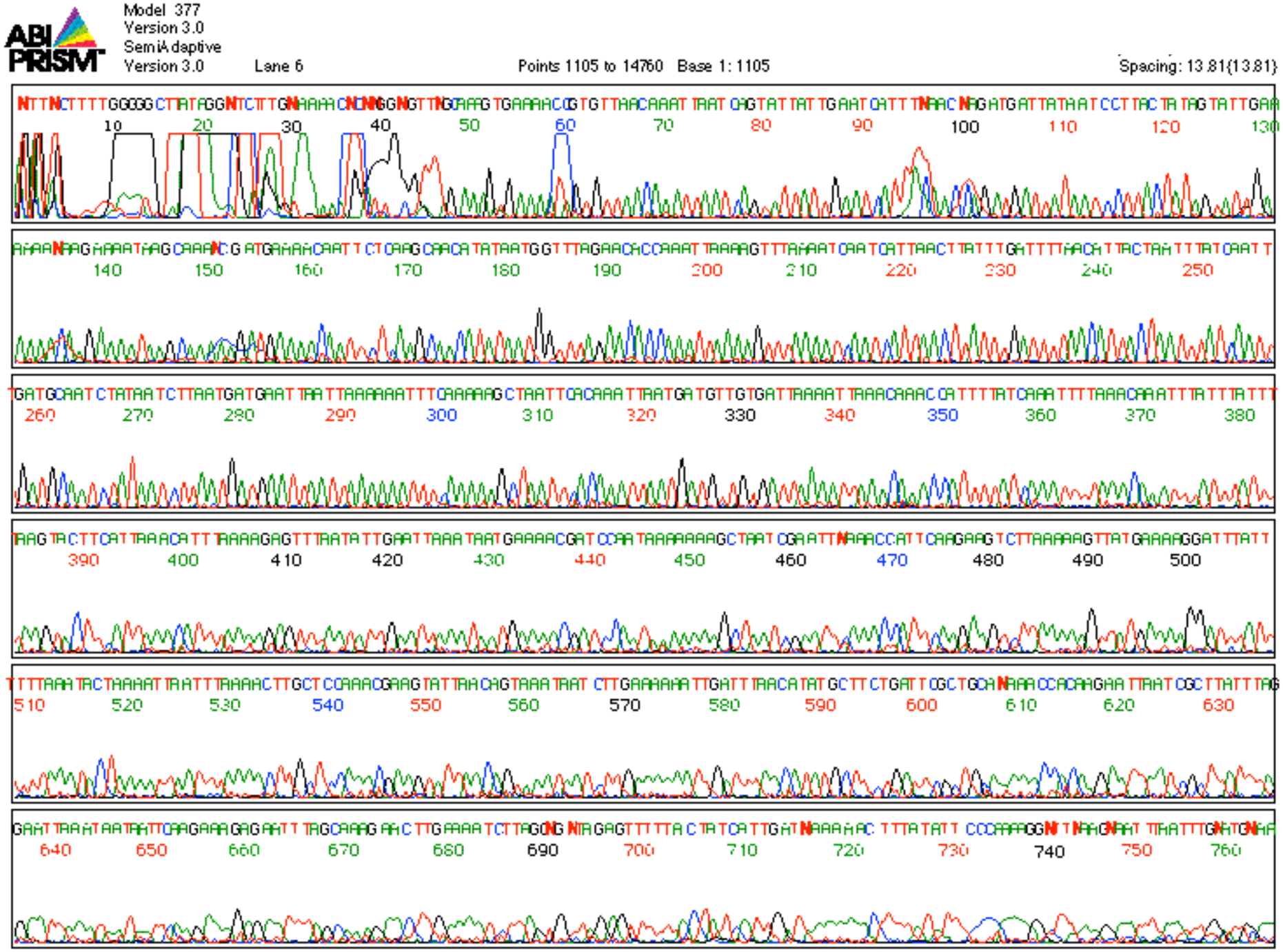
Older Fastq Header Format (**version < 1.8**)

a b c d e f g

@HWUSI-EAS100R:6:73:941:1973#0/1

a. unique instrument name
b. flowcell lane
c. tile number within the flowcell lane
d. x-coordinate of the cluster within the tile
e. y-coordinate of the cluster within the tile
f. index number for a multiplexed sample (0 for no indexing)
g. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

SEQUENCE QUALITY CONROL


Model 377
Version 3.0
SemiAdaptive
Version 3.0
Lane 6
Points 1105 to 14760 Base 1: 1105
Spacing: 13.81(13.81)

Sequencing (phred) **quality scores (Q)** measure the **probability (P)** that a base is called incorrectly.

position	1	2	3	4	...
nucleotide	A	C	G	T	...
quality score (Q)	20	20	22	21	...

https://www.phrap.com/phred/

Genetic Diversity Centre Zurich

Sequencing **quality scores (Q)** measure the **probability (P)** that a base is called incorrectly.

position	1	2	3	4	...
nucleotide	A	C	G	T	...
quality score (Q)	20	20	22	21	...

$$P = 10^{\frac{-Q}{10}} = 10^{-2} = 0.01$$

Sequencing **quality scores (Q)** measure the **probability (P)** that a base is called incorrectly.

position	1	2	3	4	...
nucleotide	A	C	G	T	...
quality score (Q)	20	20	22	21	...
probability (P)	0.01	0.01	0.006	0.008	...
accuracy	0.99	0.99	0.994	0.992	...

Sequencing **quality scores (Q)** measure the **probability (P)** that a base is called incorrectly.

Base-Calling Error Probability

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

Phred Quality Score

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

Genetic Diversity Centre Zurich

Sequencing **quality scores (Q)** measure the **probability (P)** that a base is called incorrectly.

position	1	2	3	4	…
nucleotide	A	C	G	T	…
quality score (Q)	20	20	22	21	…

One character encoding!

1	2	3	4
A	C	G	T
20	20	22	21

1	2	3	4
A	C	G	T
a	a	c	b

20→a
21→b
22→c

or

1	2	3	4
A	C	G	T
+	+	*	"

20→+
21→"
22→*

or

1	2	3	4
A	C	G	T
😉	😉	😜	😀

20→😉
21→😀
22→😜

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	[ENG 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]

GDC Genetic Diversity Centre Zurich

$$Q*2 = ASCII$$

Decimal	Hex	Char
32	20	[SPACE]
33	21	!
34	22	"
35	23	#
36	24	$
37	25	%
38	26	&
39	27	'
40	28	(
41	29	)
42	2A	*
43	2B	+
44	2C	,
45	2D	-
46	2E	.

position	1	2	3	4	...
nucleotide	A	C	G	T	...
quality score Q	20	20	22	21	...
Ascii	40	40	44	42	
char endcoding	(	(	,	*	...

$$(\rightarrow \frac{ASCII}{2} = Q \rightarrow P = 10^{\frac{-Q}{10}} = 10^{-2} = 0.01$$

Illumina Quality Encoding (**version > 1.8**)

Decimal	Hex	Char	Decimal	Hex	Char
32	20	[SPACE]	64	40	@
33	21	!	65	41	A
34	22	"	66	42	B
35	23	#	67	43	C
36	24	$	68	44	D
37	25	%	69	45	E
38	26	&	70	46	F
39	27	'	71	47	G
40	28	(	72	48	H
41	29	)	73	49	I
42	2A	*	74	4A	J
43	2B	+	75	4B	K
44	2C	,	76	4C	L
45	2D	-	77	4D	M
46	2E	.	78	4E	N
47	2F	/	79	4F	O
48	30	0	80	50	P
49	31	1	81	51	Q
50	32	2	82	52	R
51	33	3	83	53	S
52	34	4	84	54	T
53	35	5	85	55	U
54	36	6	86	56	V
55	37	7	87	57	W
56	38	8	88	58	X
57	39	9	89	59	Y
58	3A	:	90	5A	Z
59	3B	;	91	5B	[
60	3C	<	92	5C	\
61	3D	=	93	5D	]
62	3E	>	94	5E	^
63	3F	?	95	5F	_

$$Q + 33 = ASCII$$

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
 .............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

 0........................26...31.......40
                         -5....0........9.............................40
                              0........9.............................40
                                 3.....9.............................40
 0.2......................26...31.......41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Encoding	ASCII	Q	P
!	33	0	1.00000
"	34	1	0.79433
#	35	2	0.63096
$	36	3	0.50119
%	37	4	0.39811
&	38	5	0.31623
'	39	6	0.25119
(	40	7	0.19953
)	41	8	0.15849
*	42	9	0.12589
+	43	10	0.10000
,	44	11	0.07943
-	45	12	0.06310
.	46	13	0.05012
/	47	14	0.03981
0	48	15	0.03162
1	49	16	0.02512
2	50	17	0.01995
3	51	18	0.01585
4	52	19	0.01259
5	53	20	0.01000
6	54	21	0.00794
7	55	22	0.00631
8	56	23	0.00501
9	57	24	0.00398
:	58	25	0.00316
;	59	26	0.00251
<	60	27	0.00200
=	61	28	0.00158
>	62	29	0.00126
?	63	30	0.00100
@	64	31	0.00079
A	65	32	0.00063
B	66	33	0.00050
C	67	34	0.00040
D	68	35	0.00032
E	69	36	0.00025
F	70	37	0.00020
G	71	38	0.00016
H	72	39	0.00013
I	73	40	0.00010
J	74	41	0.00008

Phred Quality Score

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

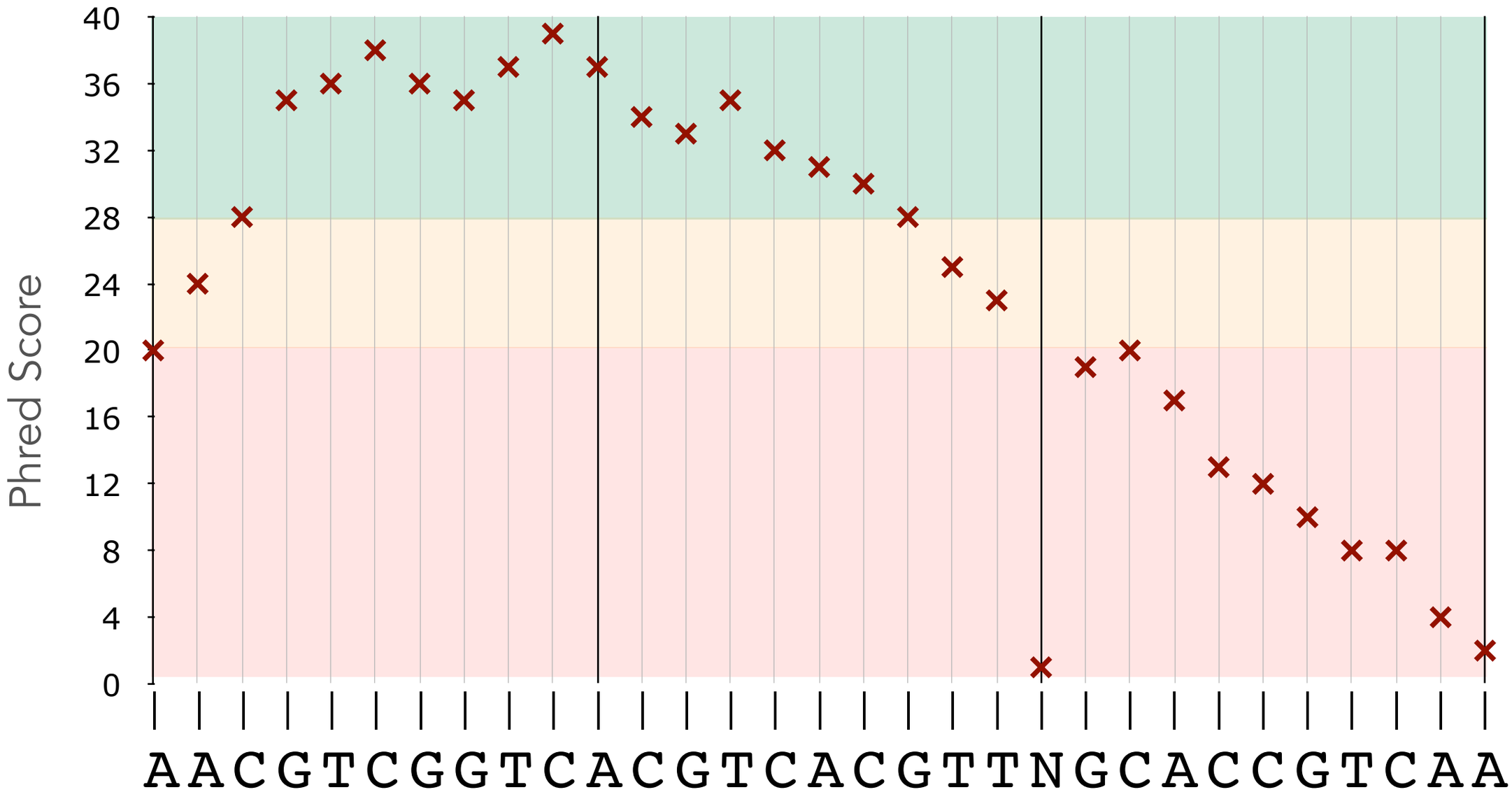
Base-Calling Error Probability

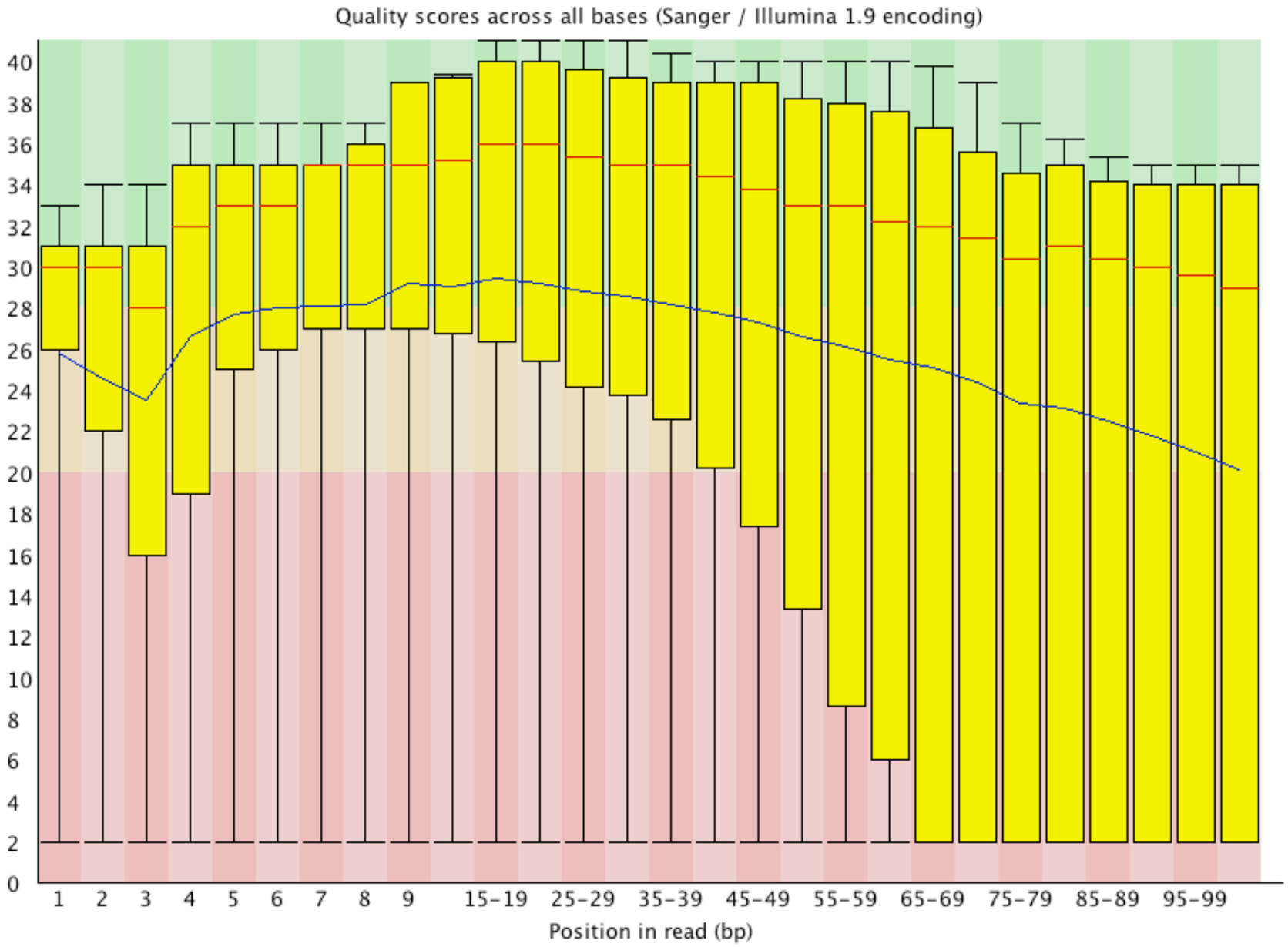
$$P = 10^{\frac{-Q}{10}}$$

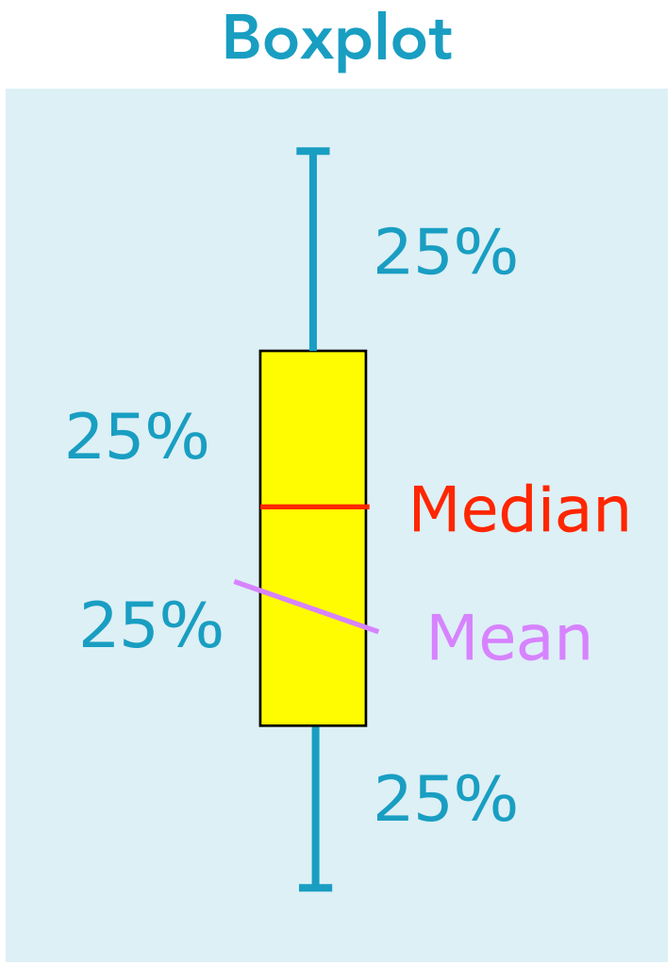
Q	P	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

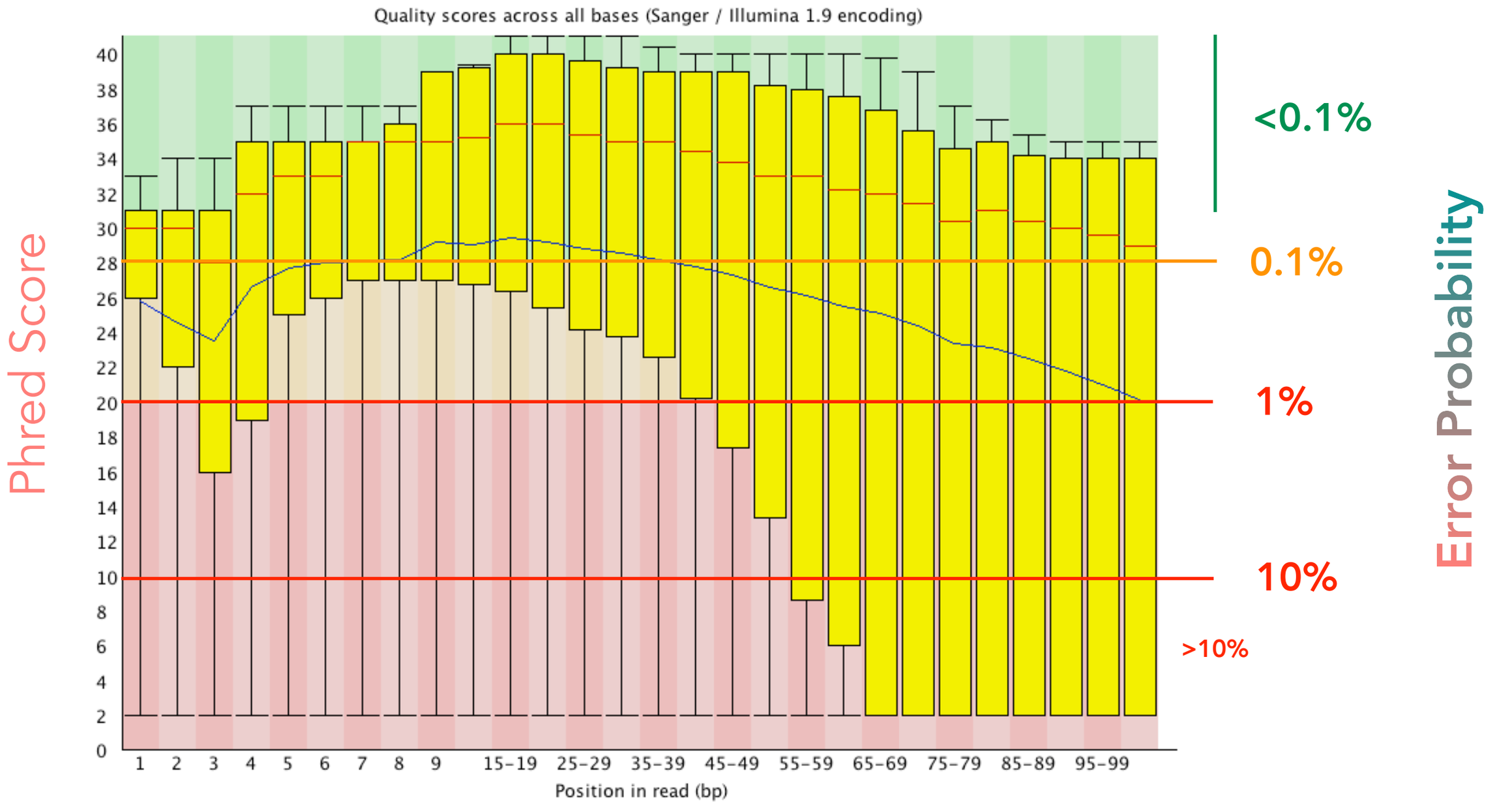
```
## R - Function

# ascii character > decimal value
asc <- function(x) {
          strtoi(charToRaw(x),16L)
          }
asc("!")

# decimal value > ascii character
chr <- function(n) {
          rawToChar(as.raw(n))
          }
chr("33")
```

Phred Scores per Base

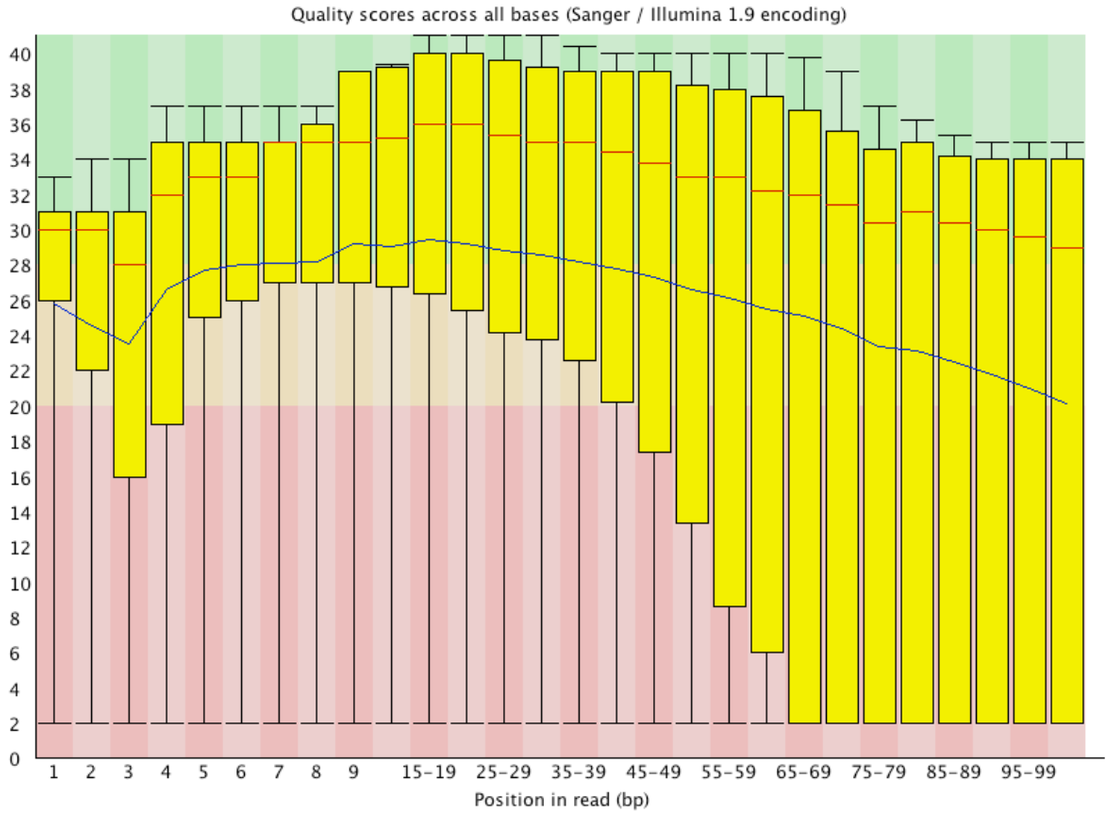



Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
Position in read (bp)


Boxplot
25%
25%
Median
25%
Mean
25%



Note: Color code is arbitrary!



Position #100: **Q = 30**

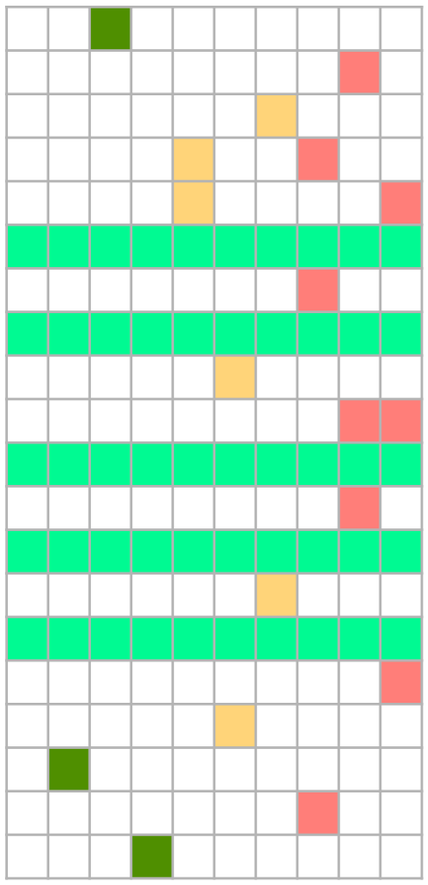
$$P = 10^{\frac{-30}{10}} = 0.001$$

Accuracy = 0.999

$$N_{(reads)} = 10^7 \rightarrow 10{,}000$$

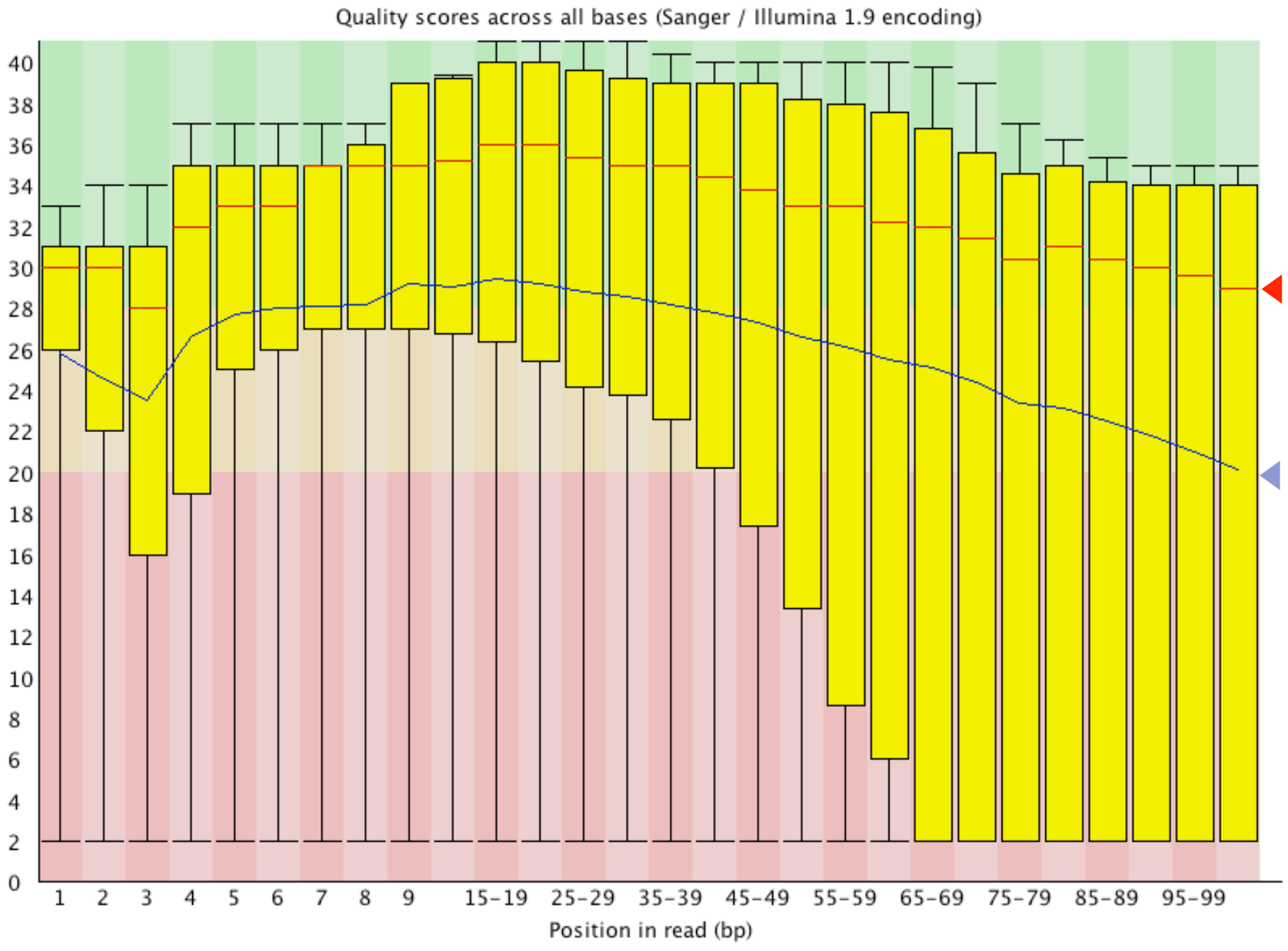
error rate:

0.05 0.1 0.3 = 0.135



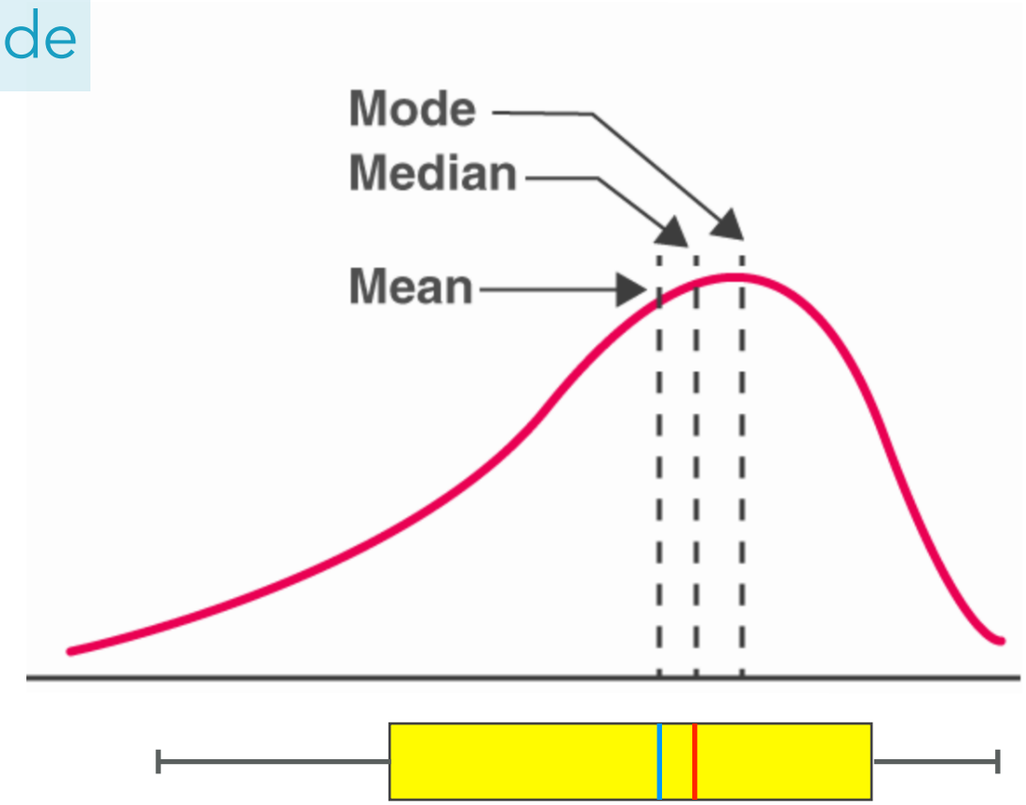
5 (25%) error free reads

Error rate increases along the length of the read.



Symmetrical Distribution

mean = median = mode



mean < median

skewed to the left

% Q-score >= Q30 (percentage of bases that have a Q-score above or equal to 30; Q30 is a probability of incorrect base calling of 1 in 1000).

Q30 = 30 (mean phred score)

Genetic Diversity Centre Zurich

Q30 =30

Q30 =30

Q30 = 30

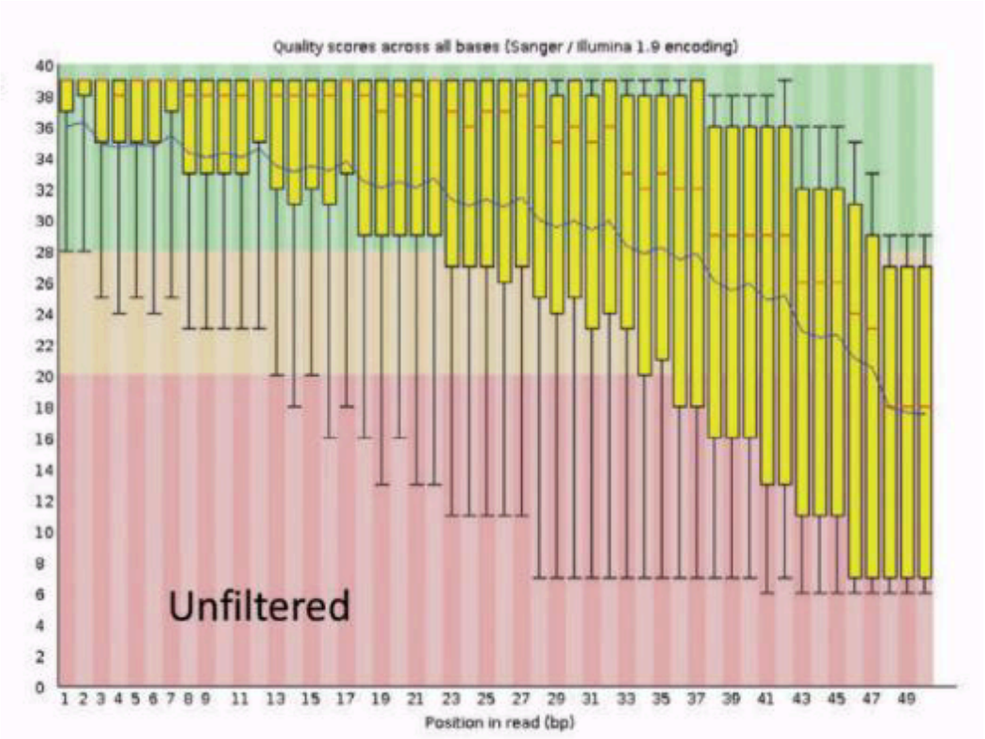
Q20 - 150nt

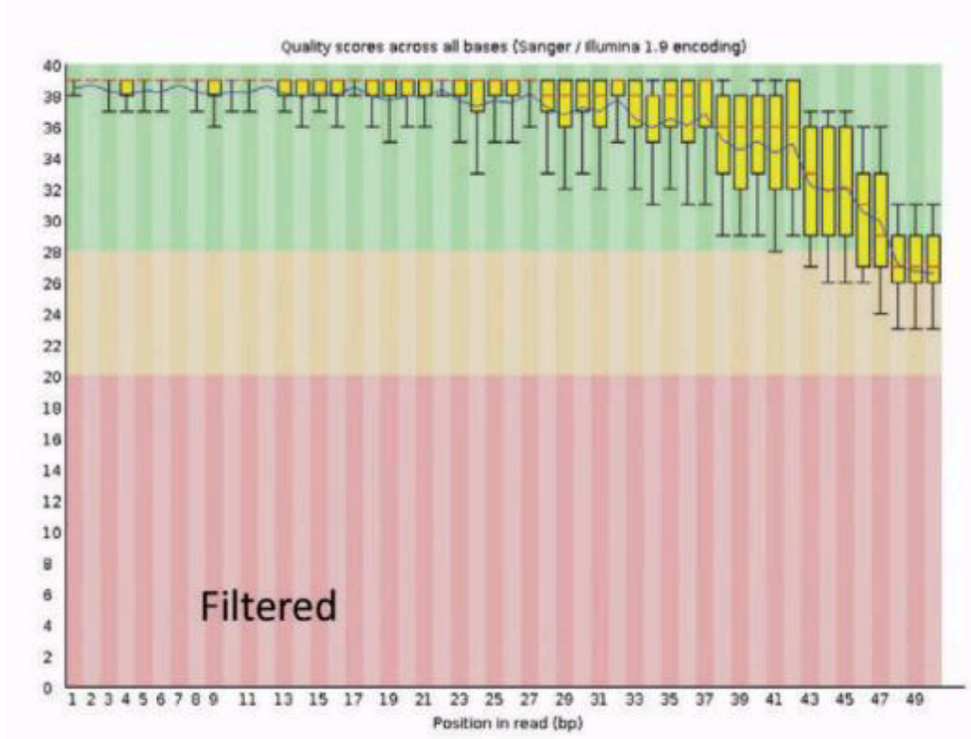
99% Accuracy / 1% Error Rate

15 Mio Reads - 1% → 15,000 Errors per Site

0.99^{150} → **22% Error Free Reads**

For Better or Worse





N_{reads} = 6Mio

N_{reads} = 2.5Mio

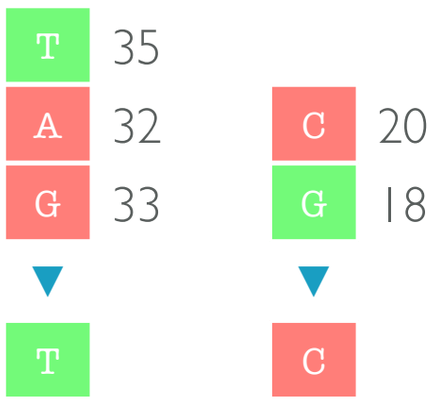
EXPECTED ERROR RATE

Error Correction

Read quality

Number of reads (coverage)

Phred score



Schirmer *et al. BMC Bioinformatics* (2016) 17:125
DOI 10.1186/s12859-016-0976-y

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data

Melanie Schirmer[1,2,4*], Rosalinda D'Amore[3], Umer Z. Ijaz[4], Neil Hall[3] and Christopher Quince[5]

Abstract

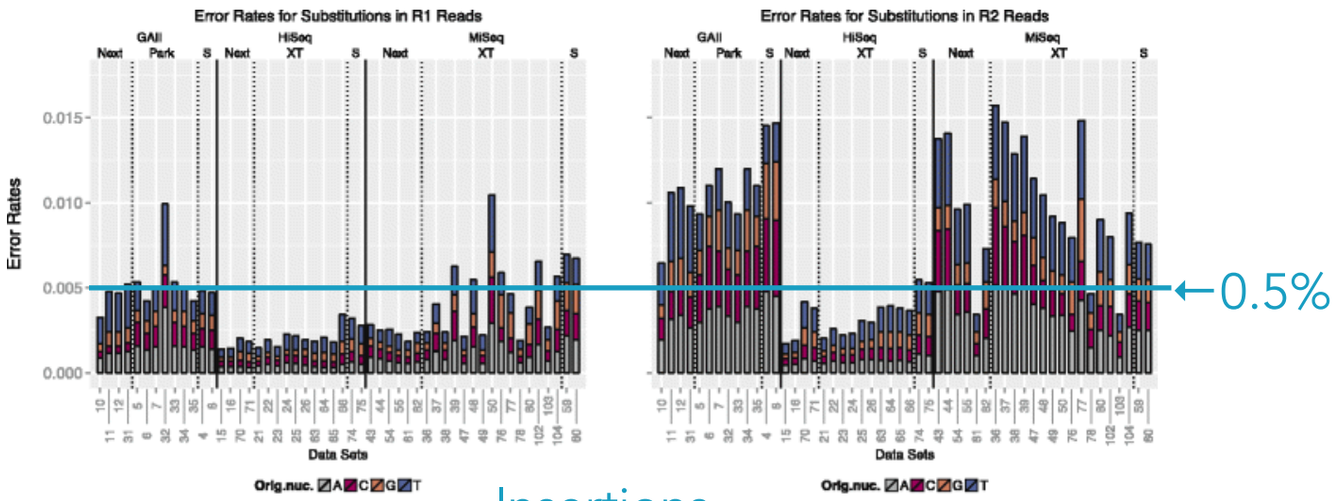
Background: Illumina's sequencing platforms are currently the most utilised sequencing systems worldwide. The technology has rapidly evolved over recent years and provides high throughput at low costs with increasing read-lengths and true paired-end reads. However, data from any sequencing technology contains noise and our understanding of the peculiarities and sequencing errors encountered in Illumina data has lagged behind this rapid development.

Results: We conducted a systematic investigation of errors and biases in Illumina data based on the largest collection of in vitro metagenomic data sets to date. We evaluated the Genome Analyzer II, HiSeq and MiSeq and tested state-of-the-art low input library preparation methods. Analysing in vitro metagenomic sequencing data allowed us to determine biases directly associated with the actual sequencing process. The position- and nucleotide-specific analysis revealed a substantial bias related to motifs (3mers preceding errors) ending in "GG". On average the top three motifs were linked to 16 % of all substitution errors. Furthermore, a preferential incorporation of ddGTPs was recorded. We hypothesise that all of these biases are related to the engineered polymerase and ddNTPs which are intrinsic to any sequencing-by-synthesis method. We show that quality-score-based error removal strategies can on average remove 69 % of the substitution errors - however, the motif-bias remains.

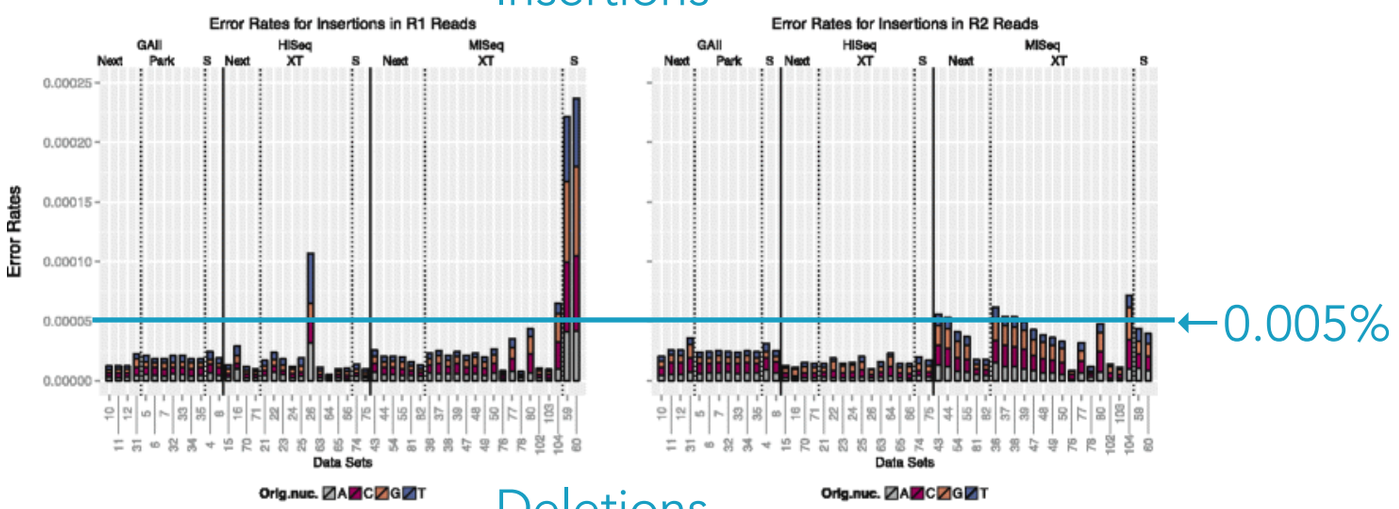
Conclusion: Single-nucleotide polymorphism changes in bacterial genomes can cause significant changes in phenotype, including antibiotic resistance and virulence, detecting them within metagenomes is therefore vital. Current error removal techniques are not designed to target the peculiarities encountered in Illumina sequencing data and other sequencing-by-synthesis methods, causing biases to persist and potentially affect any conclusions drawn from the data. In order to develop effective diagnostic and therapeutic approaches we need to be able to identify systematic sequencing errors and distinguish these errors from true genetic variation.

GDC Genetic Diversity Centre Zurich

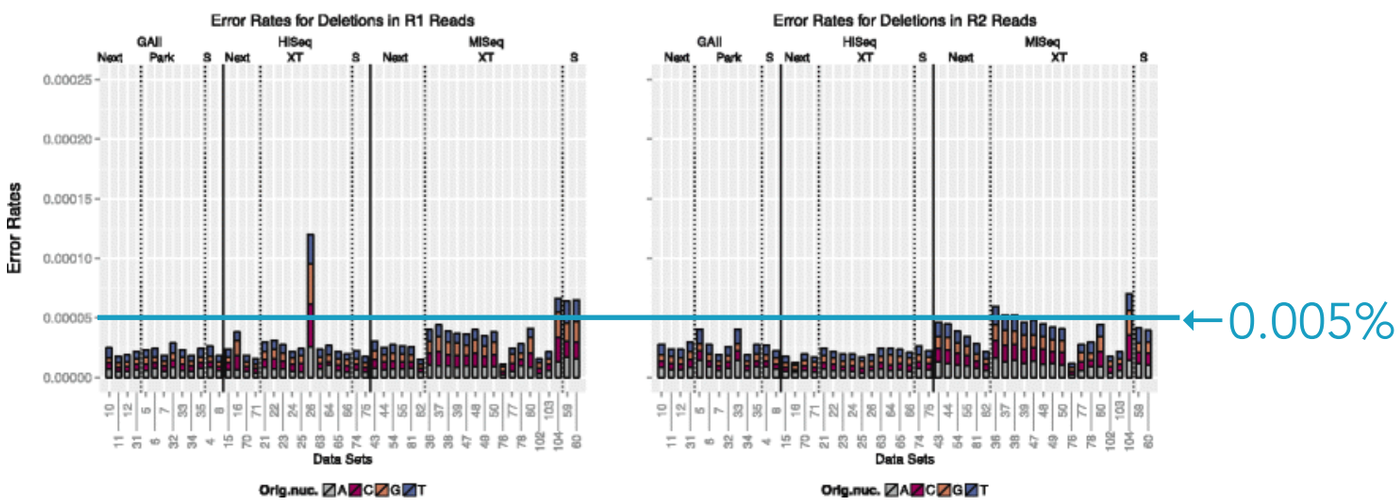
Substitutions



Insertions



Deletions



Illumina

Average substitution rates

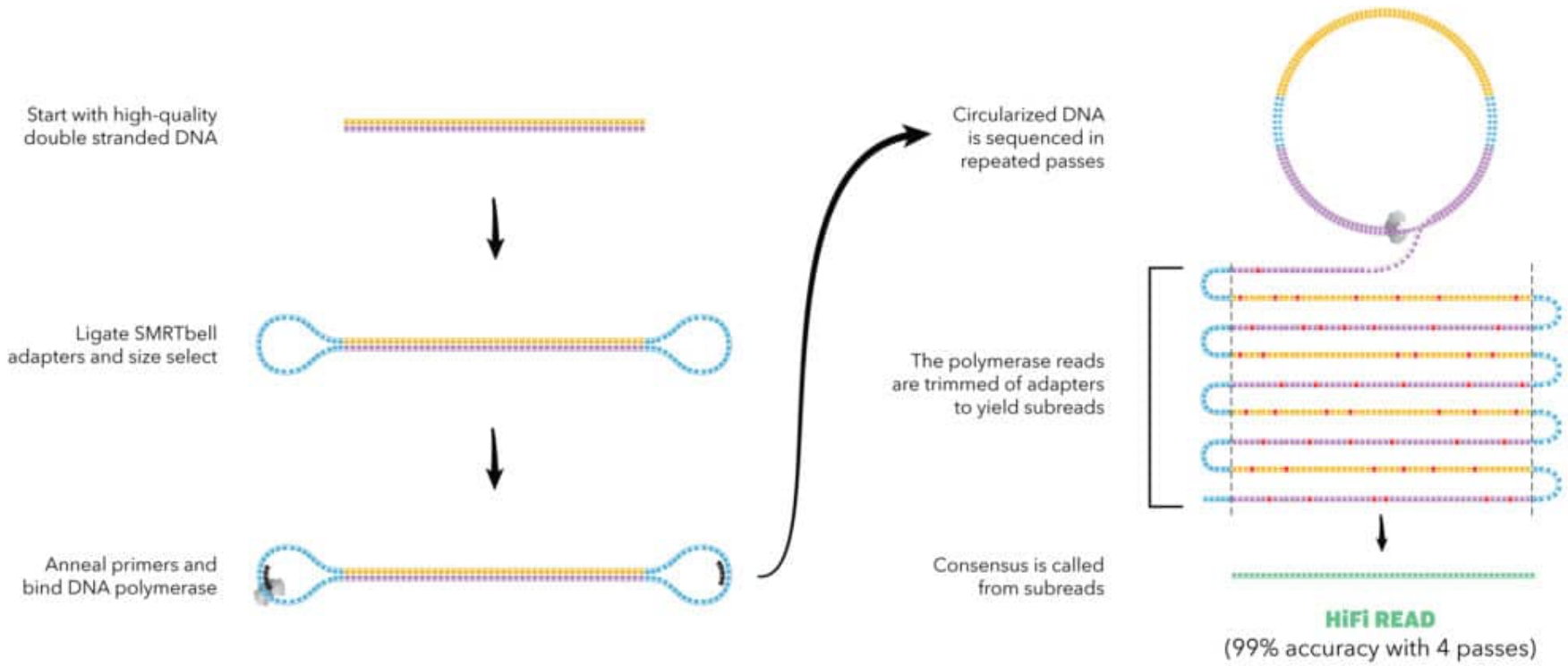
Platform	R1/R2	A	C	G	T
GAII	R1	0.0015	0.0010	0.0008	0.0018
GAII	R2	0.0035	0.0029	0.0019	0.0026
HiSeq	R1	0.0004	0.0004	0.0004	0.0008
HiSeq	R2	0.0007	0.0007	0.0007	0.0012
MiSeq	R1	0.0012	0.0009	0.0009	0.0012
MiSeq	R2	0.0033	0.0021	0.0015	0.0031

Error Rate 10-15%

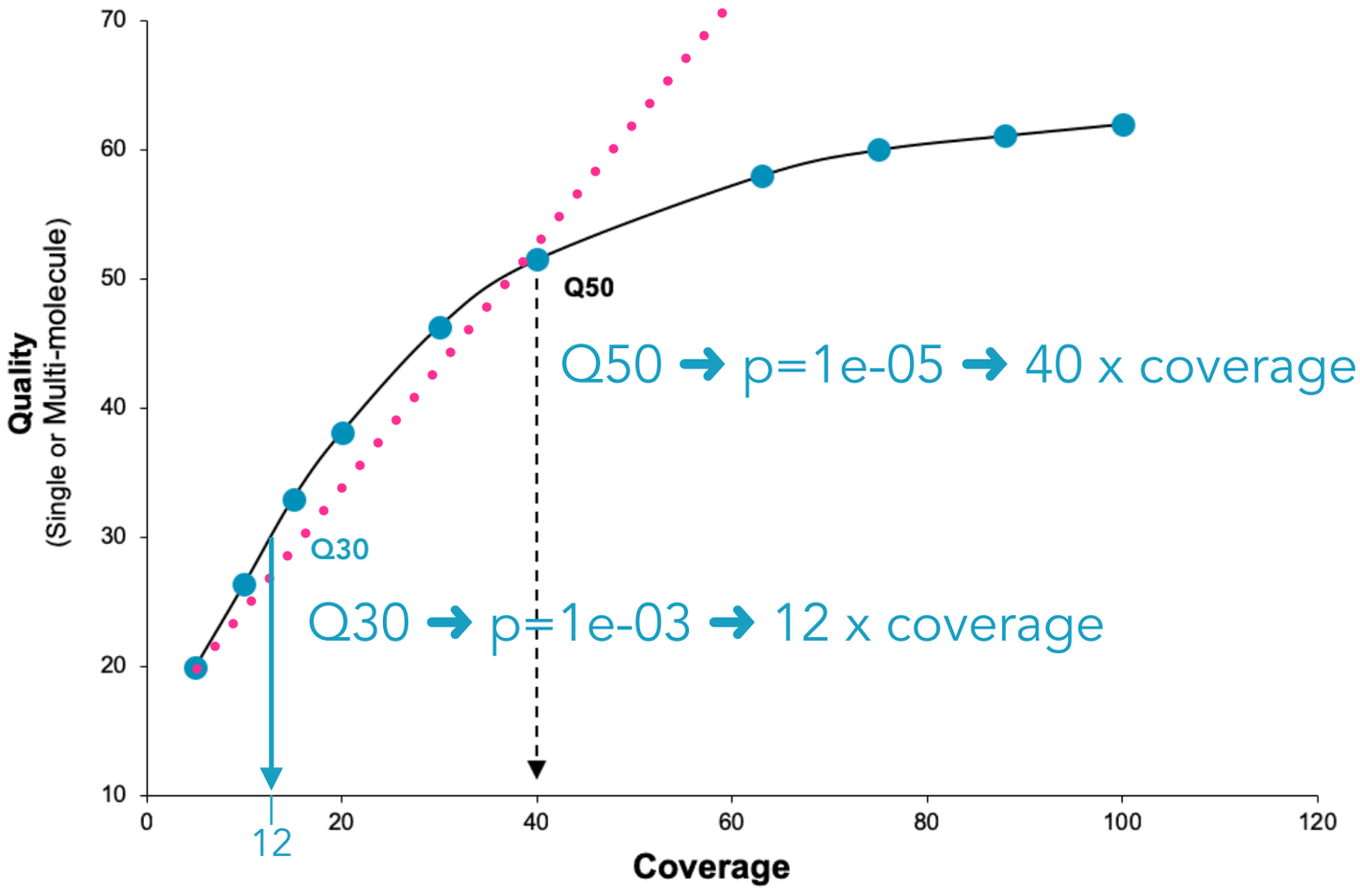
BAM ➜ FASTQ

BAM ➜ CCS.FASTX

Circular Consensus Sequences (CCS)

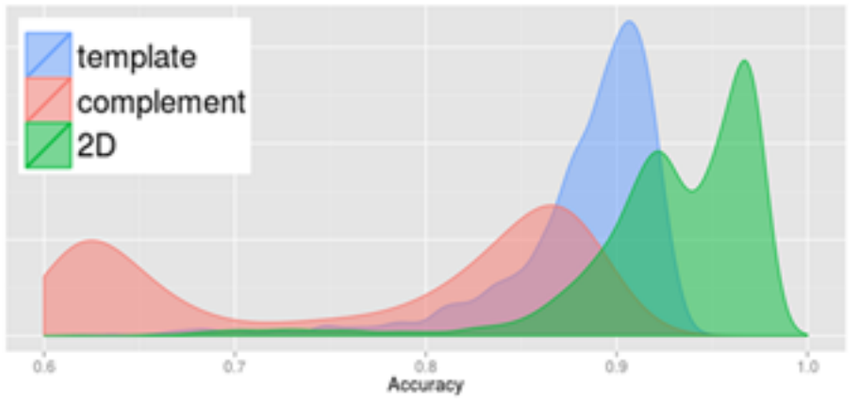


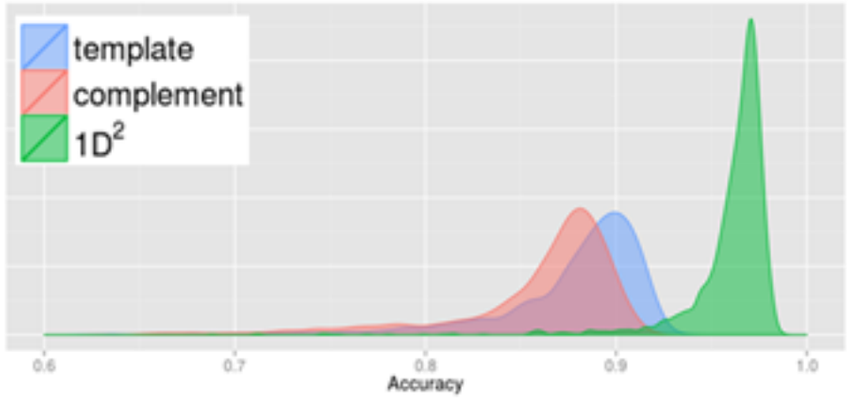
Why does it not improve anymore?



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

Oxford NANOPORE Technologies





Read type	Mappable length (bp)				Error rate (Proportion of overall error) (%)			
	Mean	Median	Standard deviation	Maximum	Overall	Insertion	Deletion	Mismatch
PacBio CCS	1772	1464	1132	8006	1.72	0.087 (5.06)	0.34 (19.48)	1.30 (**75.46**)
PacBio subread	1570	1299	1076	16040	14.20	5.92 (**41.71**)	3.01 (21.17)	5.27 (37.12)
ONT 2D	1861	1754	882	9126	13.40	3.12 (23.30)	4.79 (35.70)	5.50 (**40.99**)
ONT 1D	1695	1602	824	9345	20.19	2.93 (14.51)	7.52 (37.24)	9.74 (**48.25**)

MITOCHONDRIAL DNA PART B: RESOURCES
2019, VOL. 4, NO. 1, 408–409
https://doi.org/10.1080/23802359.2018.1547133

Taylor & Francis
Taylor & Francis Group

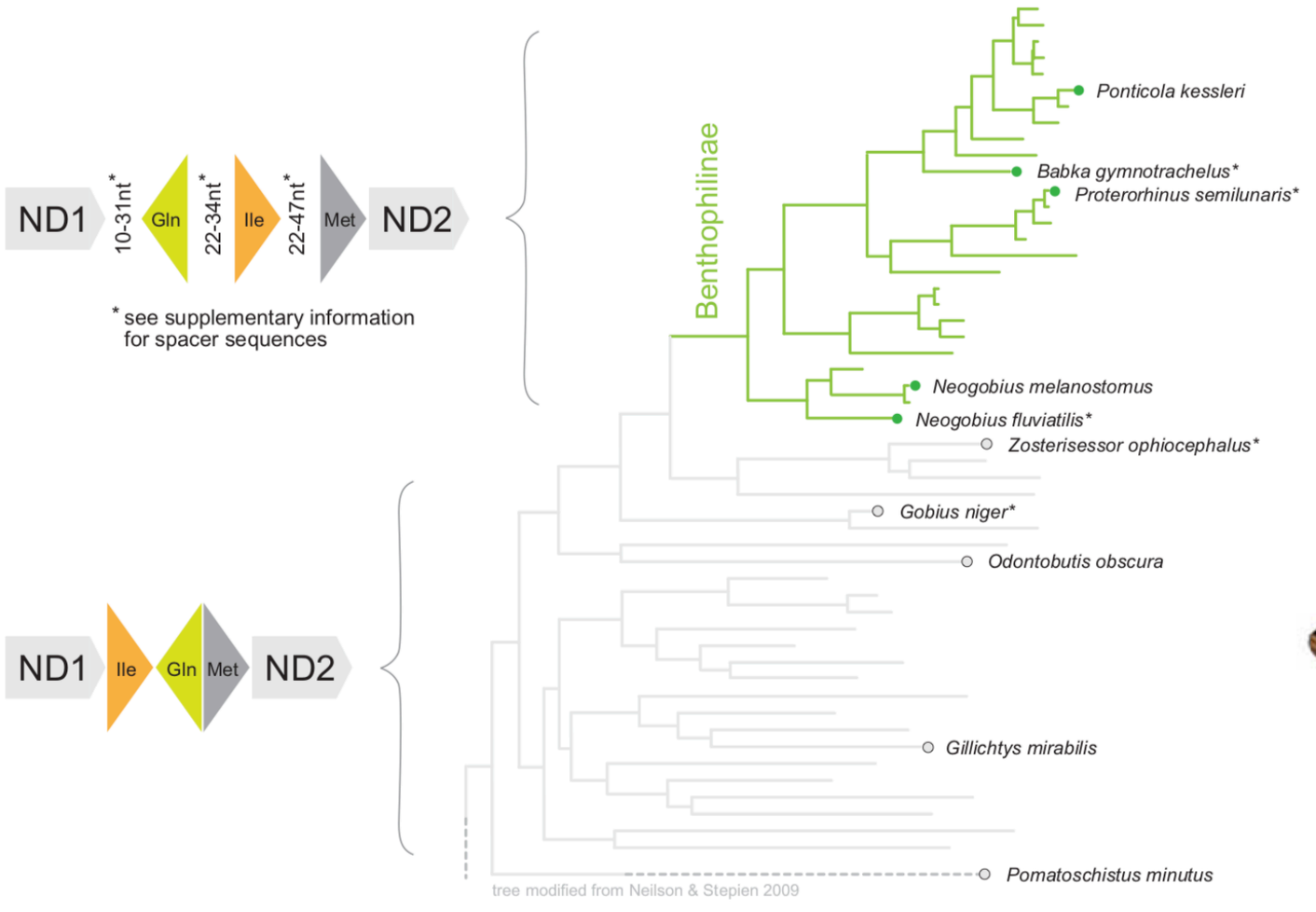
ARTICLE

OPEN ACCESS Check for updates

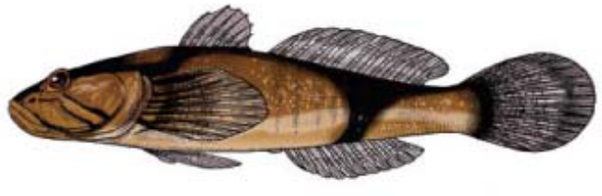
Long-read sequencing of benthophilinae mitochondrial genomes reveals the origins of round goby mitogenome re-arrangements

Silvia Gutnik[a], Jean-Claude Walser[b] and Irene Adrian-Kalchhauser[c]

[a]Biozentrum, Department Growth & Development, University of Basel, Basel, Switzerland; [b]Genetic Diversity Centre Zurich, ETH Zurich, Zurich, Switzerland;[c]Program Man-Society-Environment, Department of Environmental Sciences, University of Basel, Basel, Switzerland



Origin of the re-arranged **tRNA cluster** Gln, Ile, Met. Most Gobiidae carry the arrangement Ile, Gln, Met without spacers. Benthophilinae (subfamily of gobies) however carry the arrangement Gln, Ile, Met, and feature variable length spacers between the genes.



EXTRAS

FastQC
(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

FASTX-Toolkit
(http://hannonlab.cshl.edu/fastx_toolkit/)

USEARCH
(https://www.drive5.com/usearch/)

PRINSEQ
(http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi)

Galaxy
(http://galaxyproject.org)

Rqc
(https://bioconductor.org/packages/release/bioc/vignettes/Rqc/inst/doc/Rqc.html)

CLC Genomic Workbench
(http://www.clcbio.com/products/clc-genomics-workbench/)

Geneious
(http://www.geneious.com/)