Genetic Diversity Centre Zurich

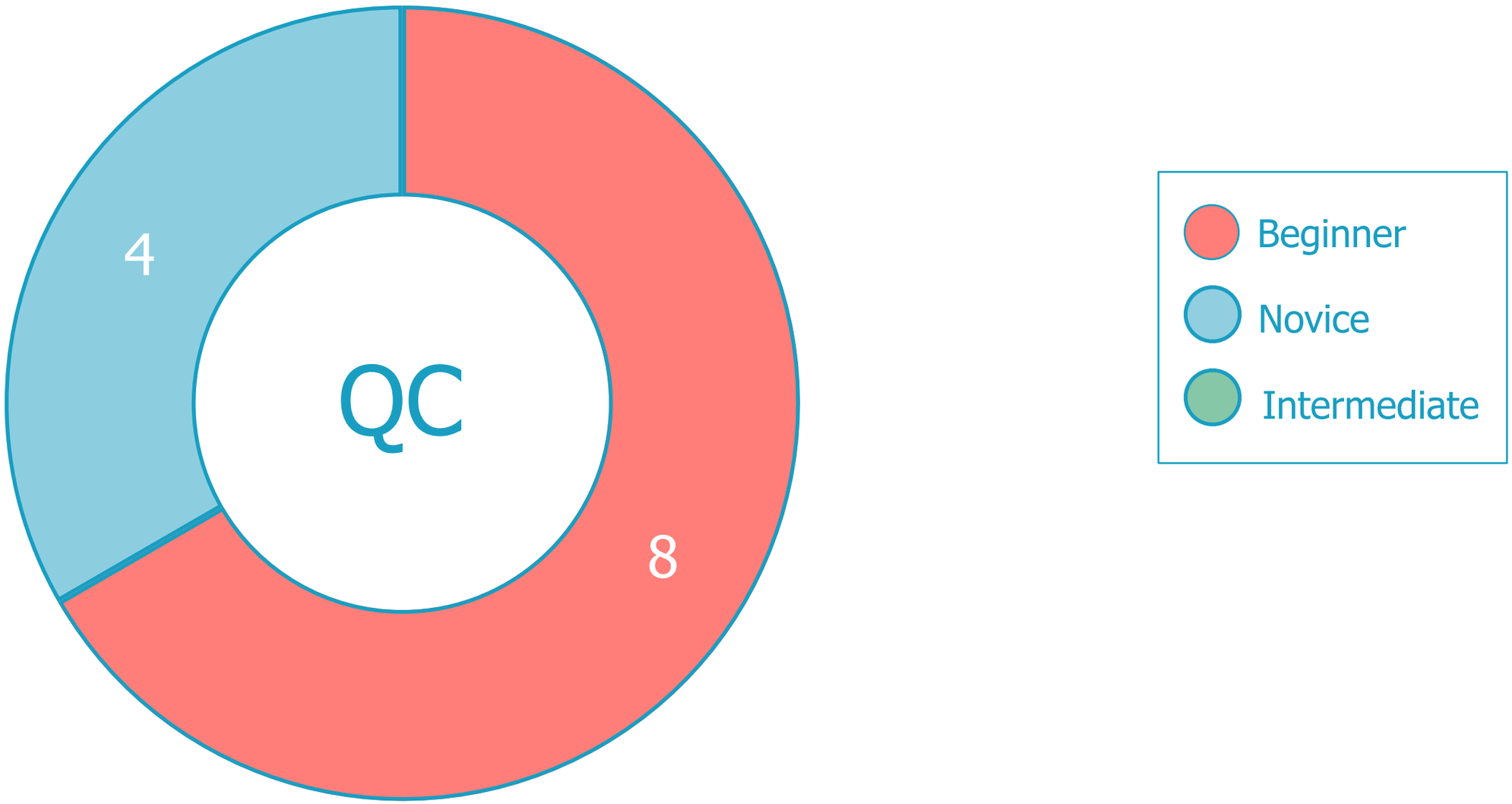
Genetic Diversity: Analysis

Sequence Quality Control

Friday, June 21, 2024

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

MPS ▷ Quality Control



Genetic Diversity Centre Zurich

1 **Sample** Quality Control

2 **Library** Quality Control

3 **Run** Quality Control

4 **Sequencing** Quality Control

5 **Data** Quality Control (e.g. bias, outlier)

Genetic Diversity Centre Zurich

1 **Sample** Quality Control

2 **Library** Quality Control

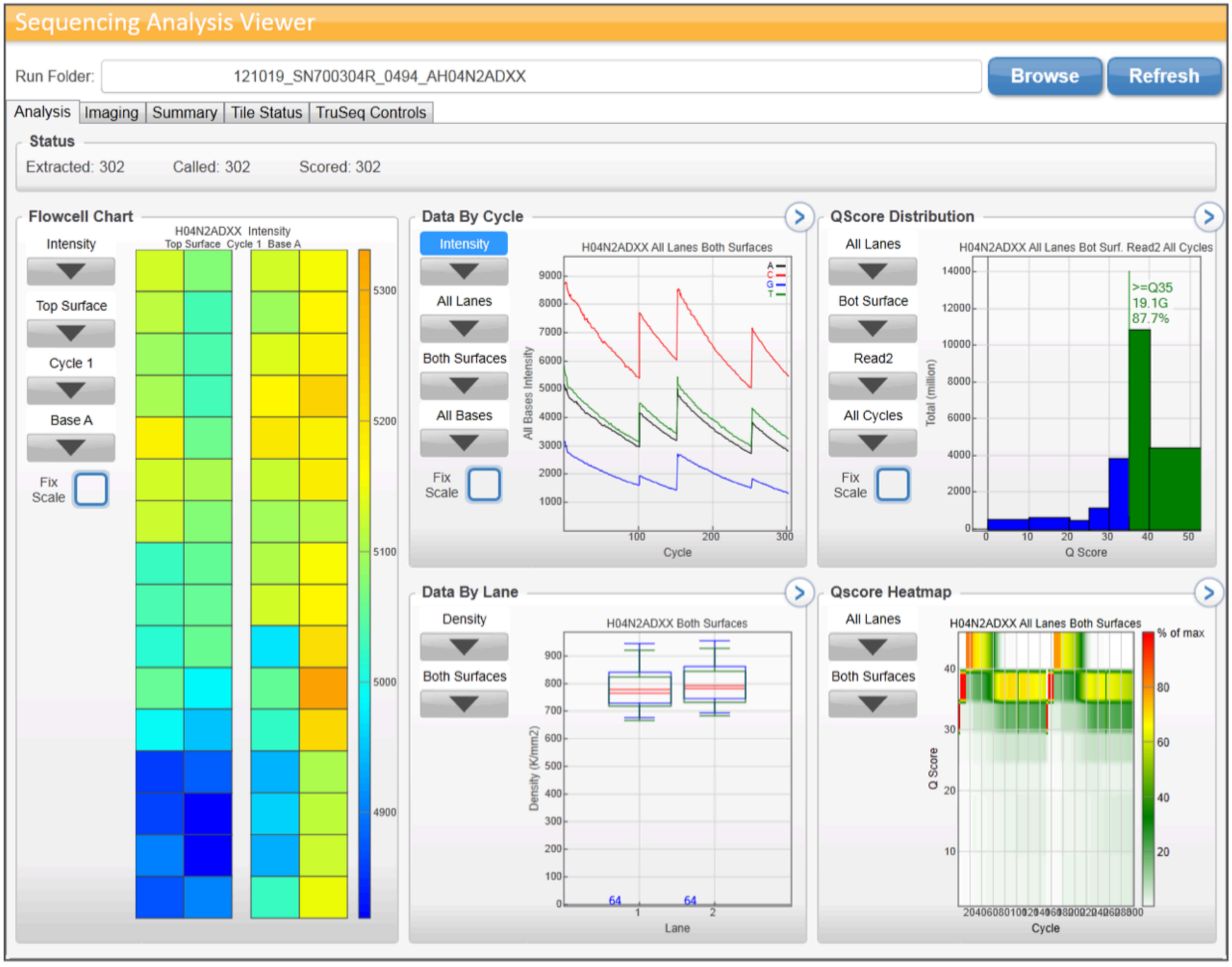
3 **Run** Quality Control

4 **Sequencing** Quality Control

5 **Data** Quality Control

Run QC

MPS ▷ Quality Control



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

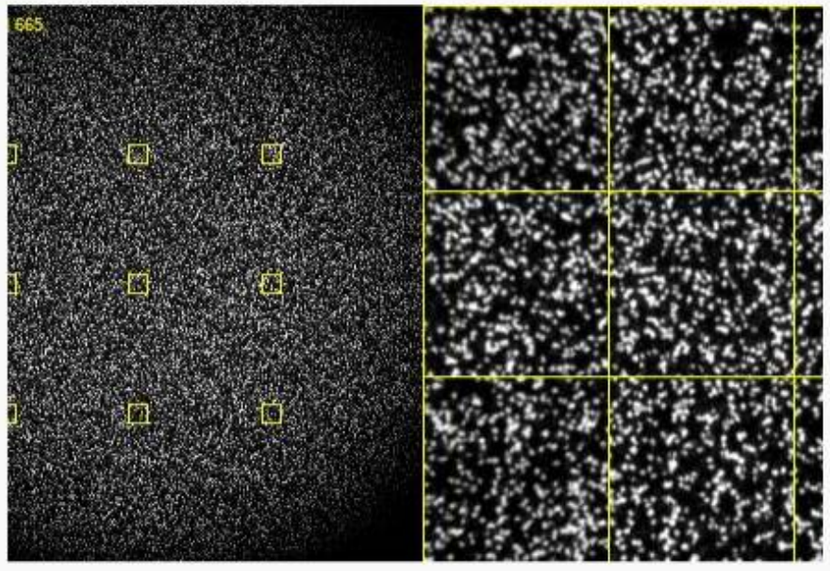
Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

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 865-965 k/mm^2)
Reads Total: 27.69 M (goal 30 M)
Reads PF: 21.60 M (78%)
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 mm^2) 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.

Sequence QC

Data Format(s)

Genetic Diversity Centre Zurich

- Fasta
- Fastq (Fasta with Quality)
- Bam (PacBio)
- POD5/FAST5 (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…
. . .
```

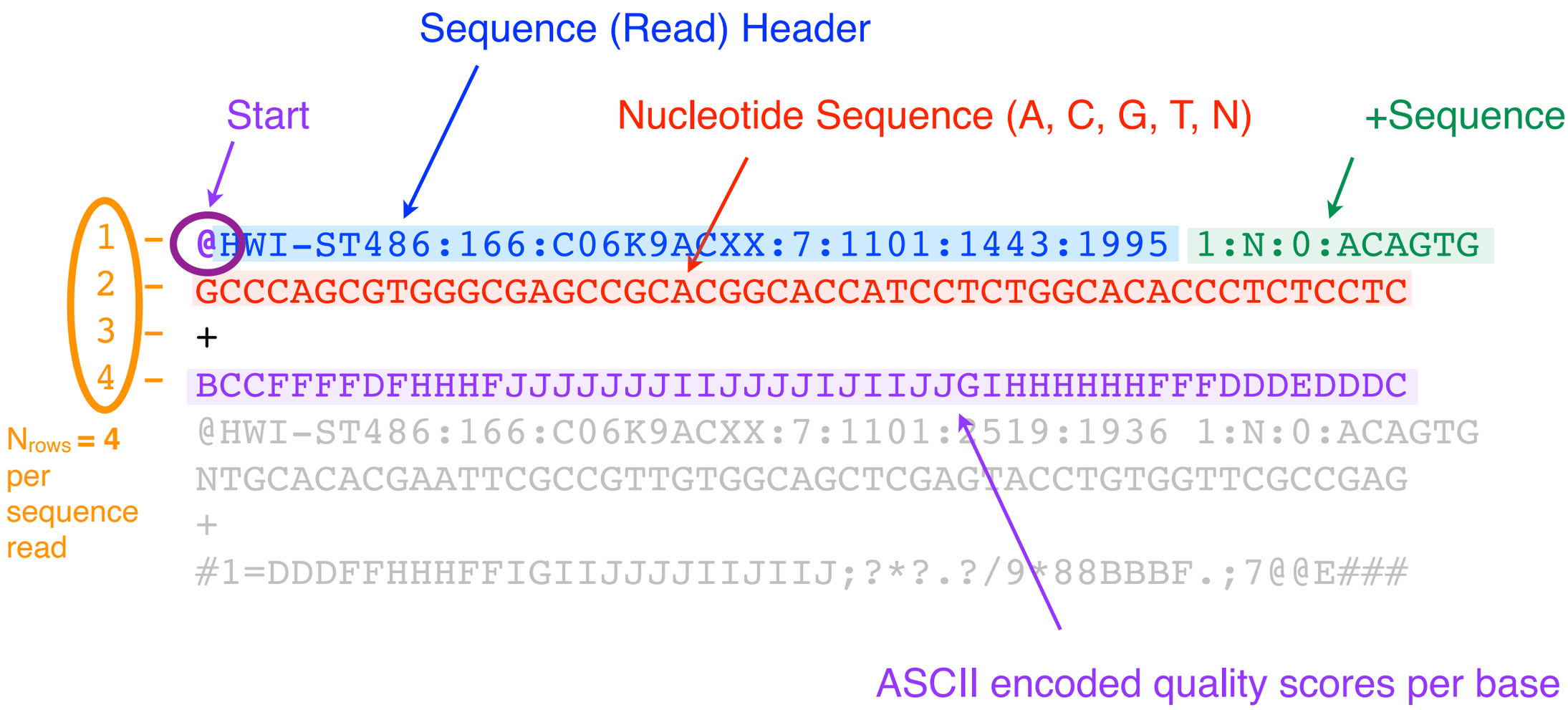
$N_{rows} \geqq 2$ per sequences read

Sequence (nucleotide or protein)

File extension: sequence(s).**fa**, sequence(s).**fasta**

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

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



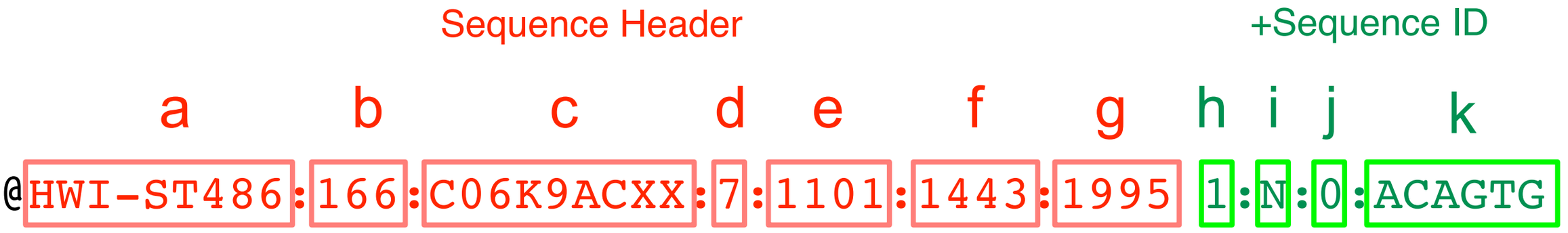
File extension: reads.**fq**, reads.**fastq**

Special cases: read_**R**[12].fq (> paired reads)

read_**I**[12].fq (> index reads)

GDC Genetic Diversity Centre Zurich

Illumina Fastq Header Format (version > 1.8)



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

Older Illumina 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)*

Genetic Diversity Centre Zurich

PacBio (Revio) Fastq Header Format

BAM → pbtk::bam2fastx → FASTQ

```
@m84151_240325_161603_s2/97453283/ccs bc=24,25 bl=GGTAGCACACGCACTGAGATACAGGAAACAGCTATGAC bq=100
bt=GTCATAGCTGTTTCCTGTACTCTCAGAGACACACTACCAT bx=38,40 cx=12 qe=3280
ql=IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII qs=38 qt=IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
TATCGCTCTTAAACGAGGAATCTCTAGTAAGTGCATTTCATGAAAATGCGCTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATT
GAATGATGTGATGAAAATCTCAGAGATTGTATACGCCAATAGACTGGGCAACTGGTCTATGTATACTGGTCAAAGAGATTTGAATCTTATTATTTAGAGG
AAGGAGAAGTCGTAACAAGGTATCTGTAGGTGAACCTGCAGATGGATCATTCGTTACTAAACCATTTTTTTTATTTTAAACAACCCAATTGTTAATAGTC
TTTTTATAAACCAAACCATTCAATTATTTTTTATTTCAA…
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII…
```

PacBio (Revio) Fastq Header Format

```
@m84151_240325_161603_s2/97453283/ccs - QNAME (query template name)
bc=24,25 - Barcode Calls
bl=GGTAGCACACGCACTGAGATACAGGAAACAGCTATGAC - Barcode sequence clipped from leading end
bq=100 - Barcode Quality [0-100]
bt=GTCATAGCTGTTTCCTGTACTCTCAGAGACACACTACCAT - Barcode sequence clipped from trailing end
bx=38,40 - Pair of clipped barcode sequence lengths
cx=12 - subread tag
qe=3280 - For CCS reads, the 0-based end of the query in its original CCS read.
ql=IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII - Qualities of bc bases clipped from leading end
qs=38 - For CCS reads, the 0-based start of the query in its original CCS read.
qt=IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII - Qualities of bc bases clipped from trailing end
```

The `QNAME` for by-strand CCS reads includes a suffix **`fwd`** or **`rev`** to indicate strand relative to the other by-strand read for the ZMW. Strand assignment by CCS is arbitrary and does not imply the strand that may be assigned during mapping.

Source: https://pacbiofileformats.readthedocs.io/en/13.0/index.html

ONT Fastq Header Format

```
@c66cf9d5-8ceb-49a2-87ba-573bdf744476 runid=da30afcc85dd427c946e129369e698f701722998 read=19 ch=21
start_time=2023-12-13T17:44:43.197441+01:00 flow_cell_id=FAY08440
protocol_group_id=20231213_LakeGeneva_Sasi sample_id=LG1_Stepeffluent
parent_read_id=c66cf9d5-8ceb-49a2-87ba-573bdf744476
basecall_model_version_id=dna_r10.4.1_e8.2_400bps_sup@v4.2.0
ATCATATTTACTTCGTTCAGTTACGTATTGCTGTTTCAATCTTGGTGAAATTAATTCTGTTTGATTATTGTTTCCATCTTGTTTACTTTCACCATAATTA
AAATATAAAATTTCTTTATATAATGTTTCTTTTGAATAGTTTTCTATTAAAAAGCTTTTGTCATCAATATTATTATTTTTATATTCTTCAAATAATTTTA
AAAATAAATTGTAATCAGTTTTATTAAATTGATAGCAATACATTA
+
#$$#$$%'('*/001:<;?
<<;920000==:::;;QHIHHKLFKSSNSKNJMF?;78888??;BILGKJIKRMHHHSGSJKKSHLPSISLSSMKNLHJIPHSHHPSPMSHNSSMQSSPM
NKJHGJSSSNSHIPLJHIKSOHKLJHHGJSGOLILSMMSNIL9999;ISNKMS99999SKSHSSISJSIHIMKSHJLSSNMSKHLSSMKKKIIHKSPMJG
KSSJSSJSJSIKSIGIJNJEFJ755)
```

The headers in FASTQ files generated by Oxford Nanopore Technologies (ONT) basecallers are defined by the specific basecaller used (such as Dorado, Guppy, Albacore, Bonito, etc.) and have changed many times over the years.

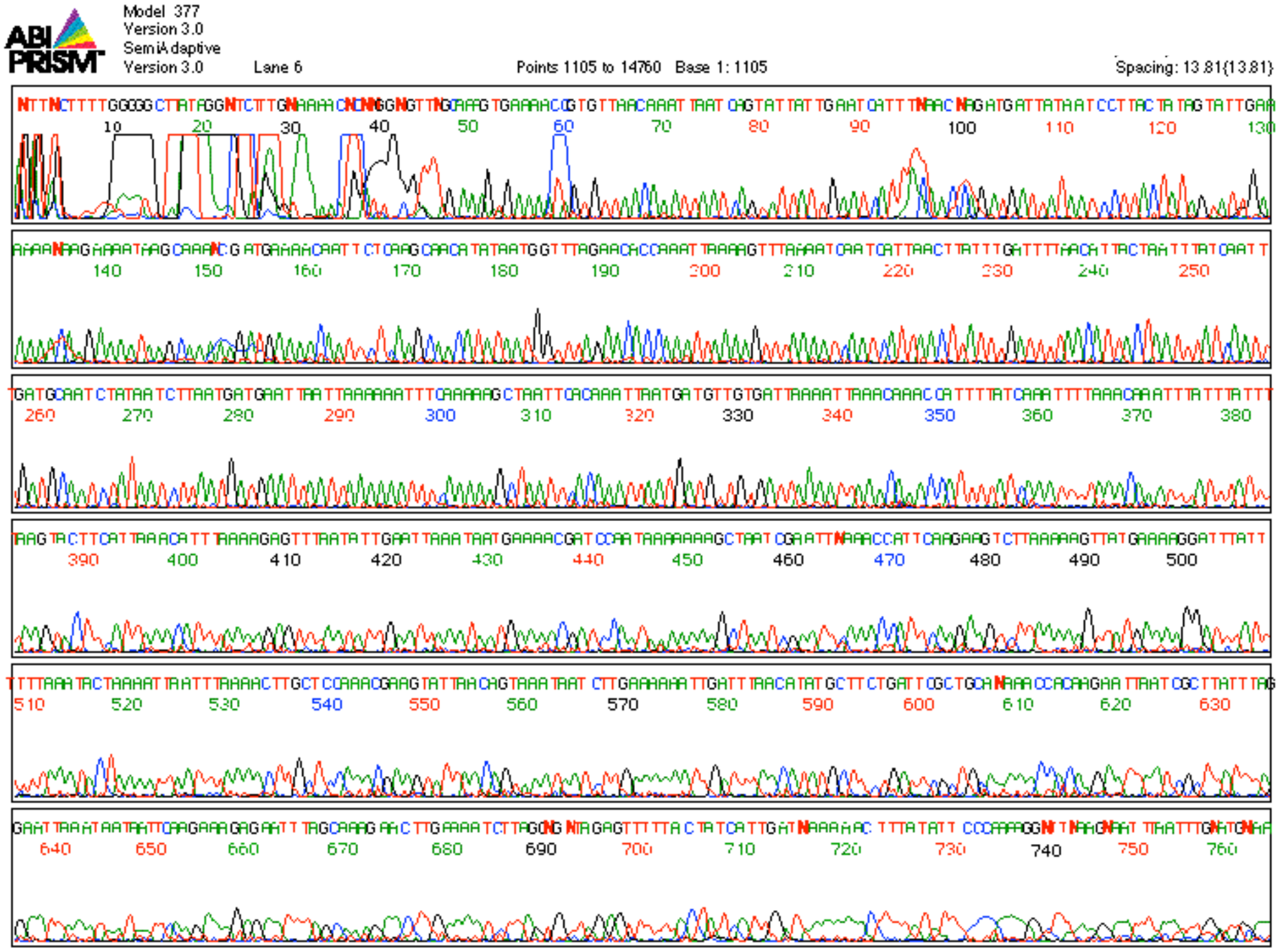
Genetic Diversity Centre Zurich

ONT Fastq Header Format

```
@c66cf9d5-8ceb-49a2-87ba-573bdf744476
runid=da30afcc85dd427c946e129369e698f701722998
read=19
ch=21
start_time=2023-12-13T17:44:43.197441+01:00
flow_cell_id=FAY08440
protocol_group_id=p1023_run240512_Meta
sample_id=Shower-123sw
parent_read_id=c66cf9d5-8ceb-49a2-87ba-573bdf744476
basecall_model_version_id=dna_r10.4.1_e8.2_400bps_sup@v4.2.0
```

The FASTQ headers generated by Dorado, like those from other ONT basecallers, are designed to provide comprehensive metadata about each read. Understanding the specific format used by the version of Dorado you are working with is critical to effective data management and analysis.

Sequence QC


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

Fastq Sequence Read

```
1 - @HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG
2 - ACTGAGCGTGGGCGAGCCGCACGGCACCATCCTCTGGCACACCCTCTCCTC
3 - +
4 - 5576FFFDFHHHFJJJJJJJIIJJJJIJIIJJGIHHHHHHFFFDDDEDDDC
```

⇧

ASCII **encoded** quality scores per base

Genetic Diversity Centre Zurich

Sequencing (phred) **quality scores (Q)** measures the (error) **probability (P)** that a base is called (in)correctly.

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

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

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

Base-Calling Error Probability

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

Phred Quality Score

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

Sequencing (phred) **quality scores (Q)** measures the (error) **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	...

$$P = 10^{\frac{-Q}{10}} = 10^{-2} = 0.01 \rightarrow 1\% \ (\rightarrow 99\%)$$

Sequencing (phred) **quality scores (Q)** measure the (error) **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 (1-P)	0.99	0.99	0.994	0.992	...

One character encoding!

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

1	2	3	4		1	2	3	4		1	2	3	4
A	C	G	T	or	A	C	G	T	or	A	C	G	T
a	a	c	b		+	+	*	"		😏	😏	😜	😀
20→a					20→+					20→😏			
21→b					21→"					21→😀			
22→c					22→*					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]

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	_

$$Q + 33 = ASCII$$

$$20 + 33 = 53 \rightarrow 5$$

position	1	2	3	4	...
nucleotide	A	C	G	T	...
quality score Q	20	20	22	21	...
Ascii	53	53	55	54	
char endcoding	5	5	7	6	...

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
 .............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ...................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ...................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..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)
```

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")
```

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

Phred Quality Score

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

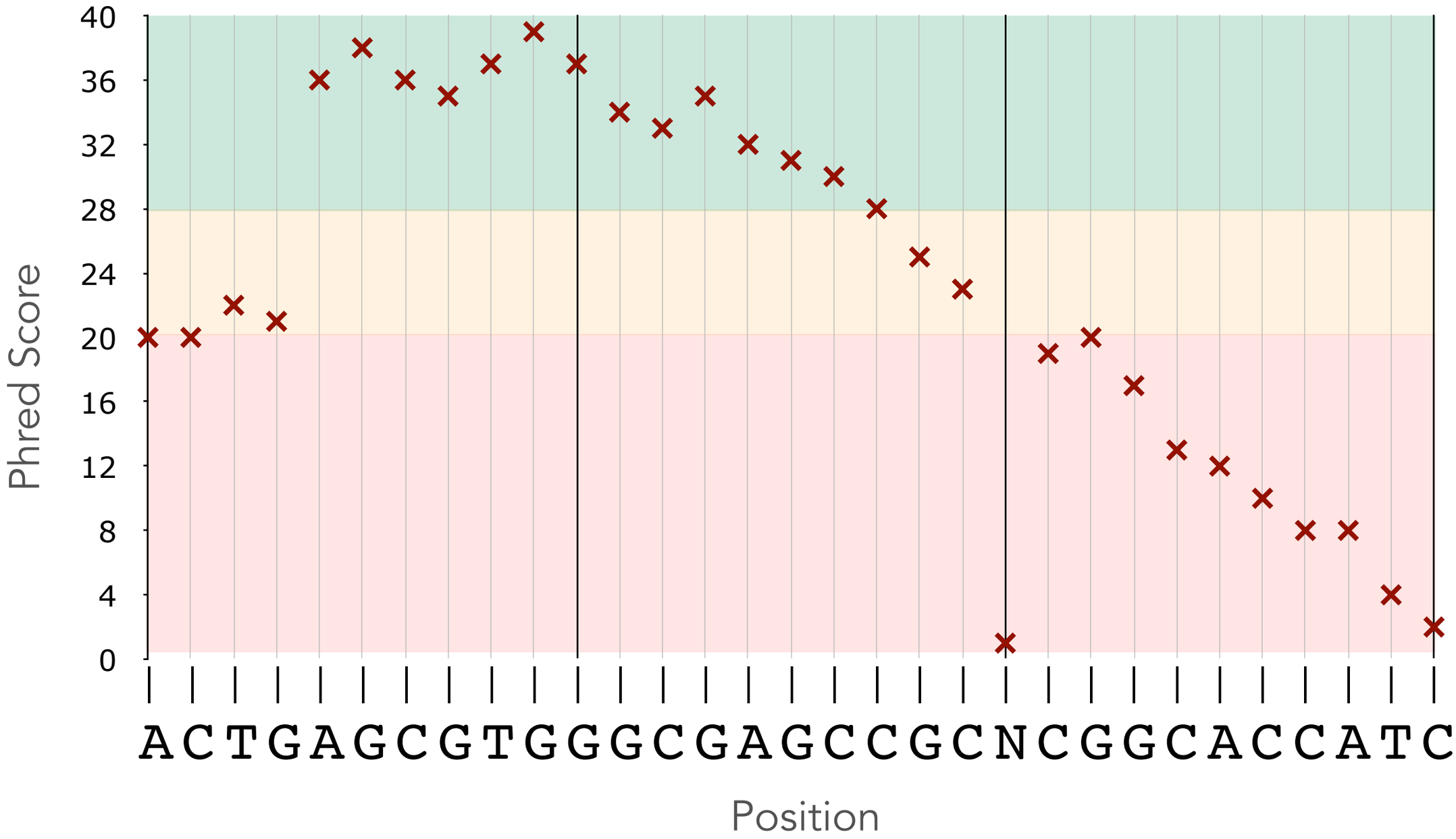
Base-Calling Error Probability

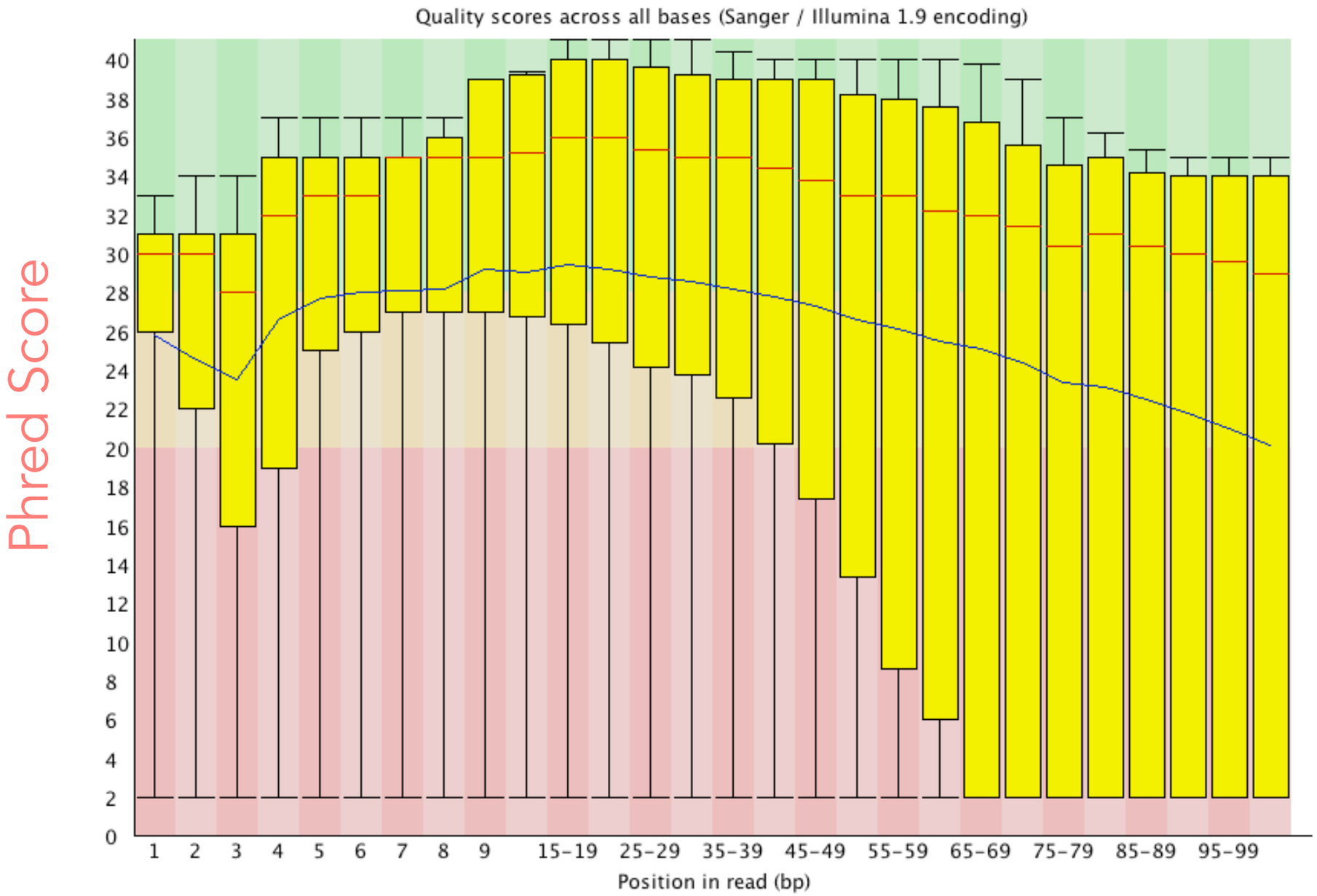
$$P = 10^{\frac{-Q}{10}}$$

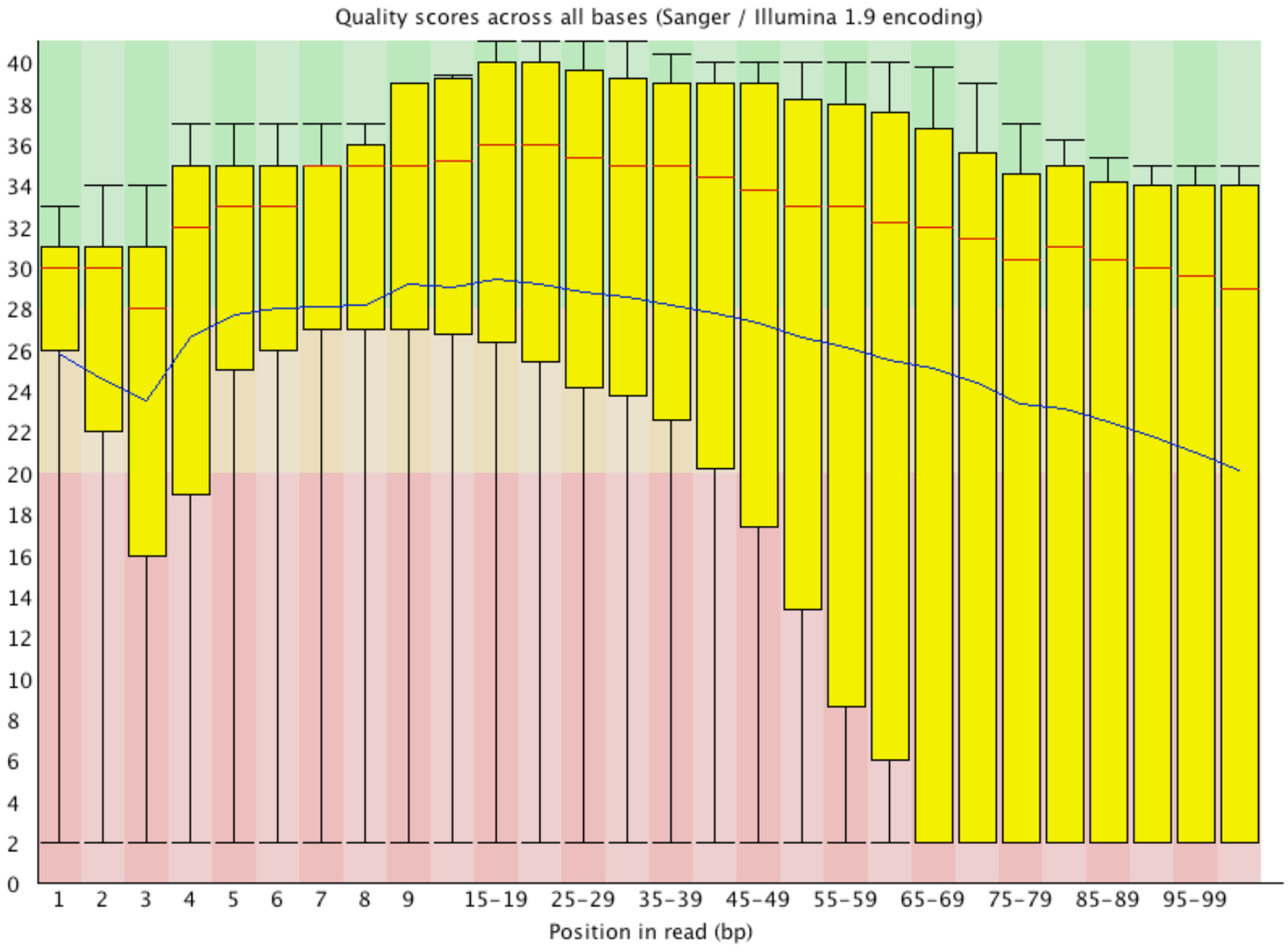
Q	P	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

Genetic Diversity Centre Zurich

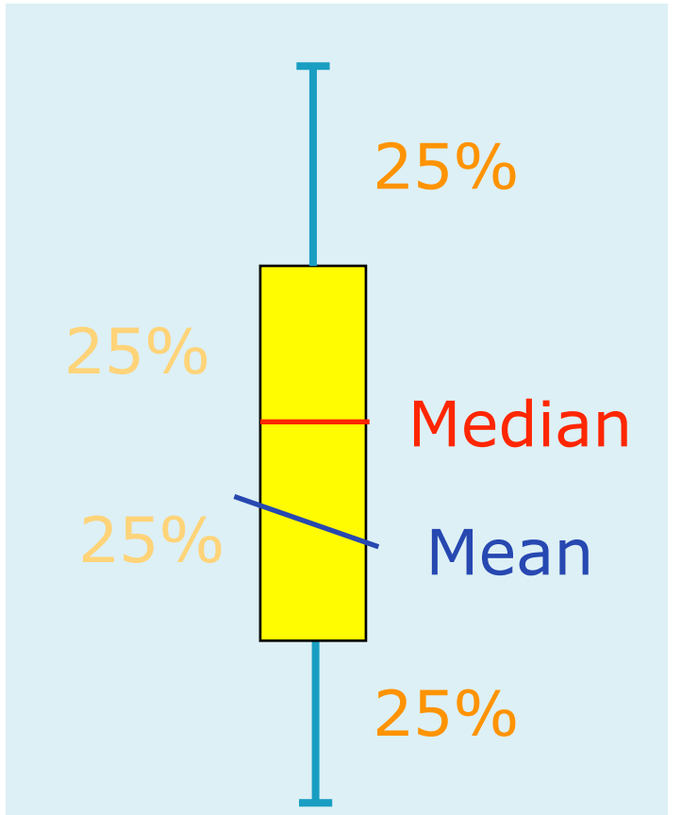
Phred Scores per Base




Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Phred Score
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)


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

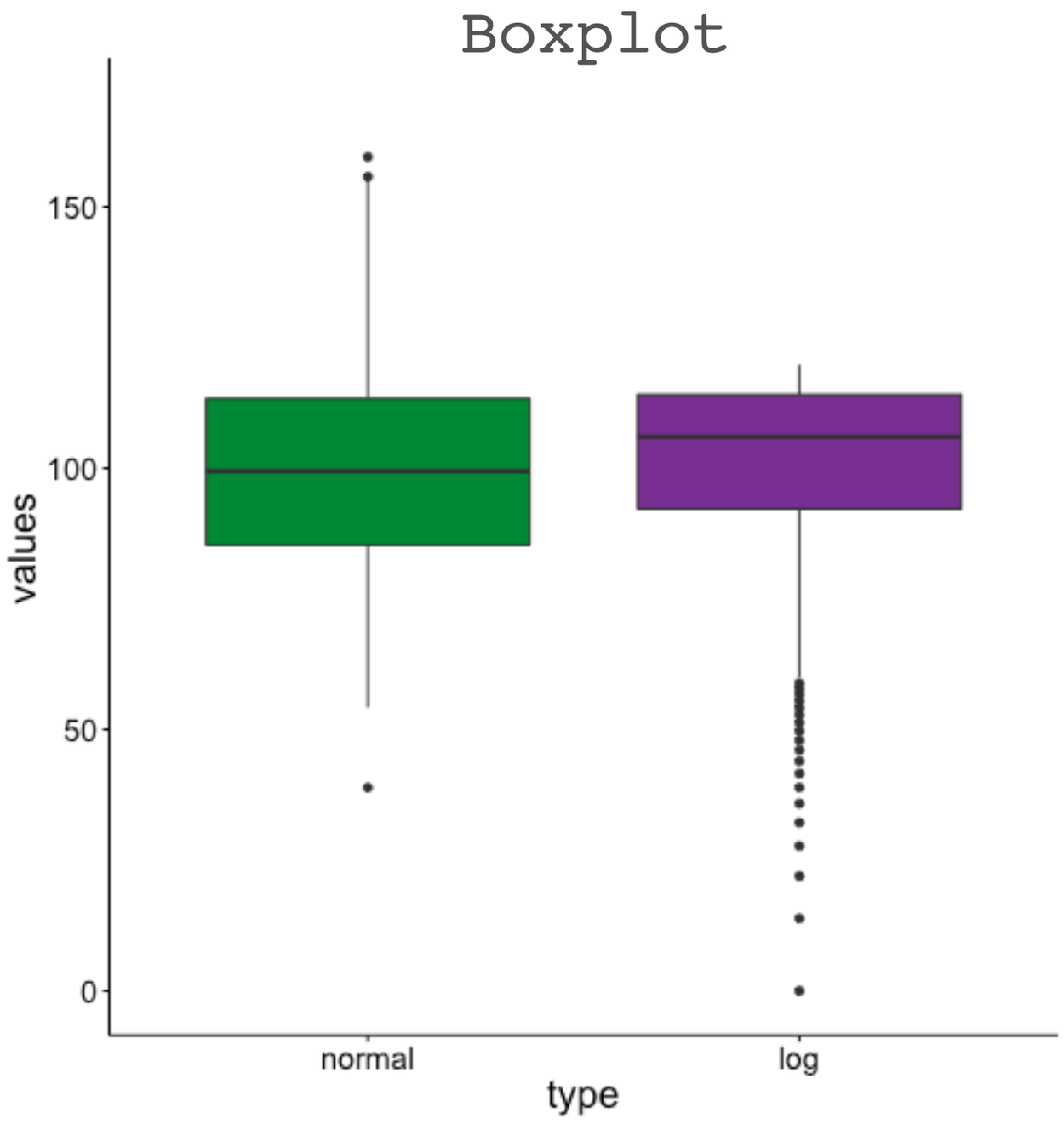

25%
25%
Median
Mean
25%

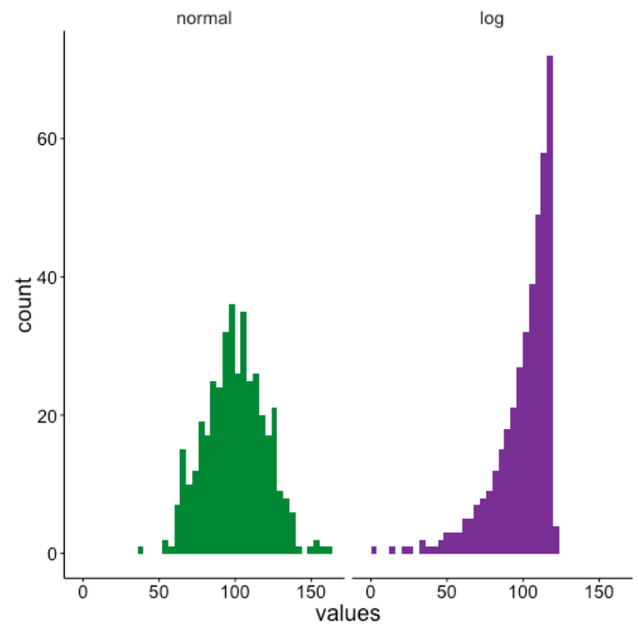
Why do we use a box plot and not a bar plot?

Genetic Diversity Centre Zurich

Why do we use a box plot and not a bar plot?

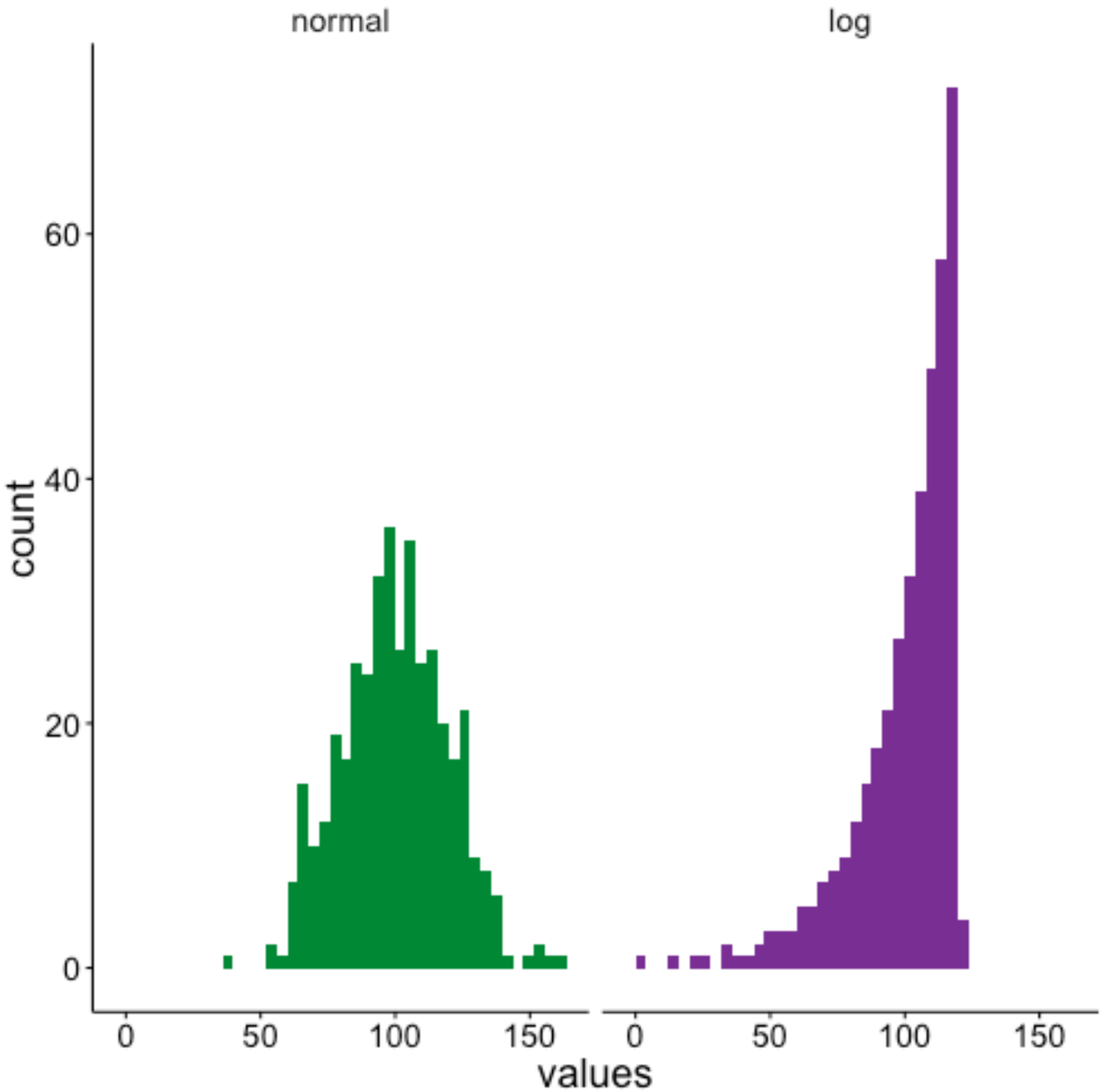
Box plots are generally more informative than bar charts, especially when you need to understand the distribution and variability of your data.





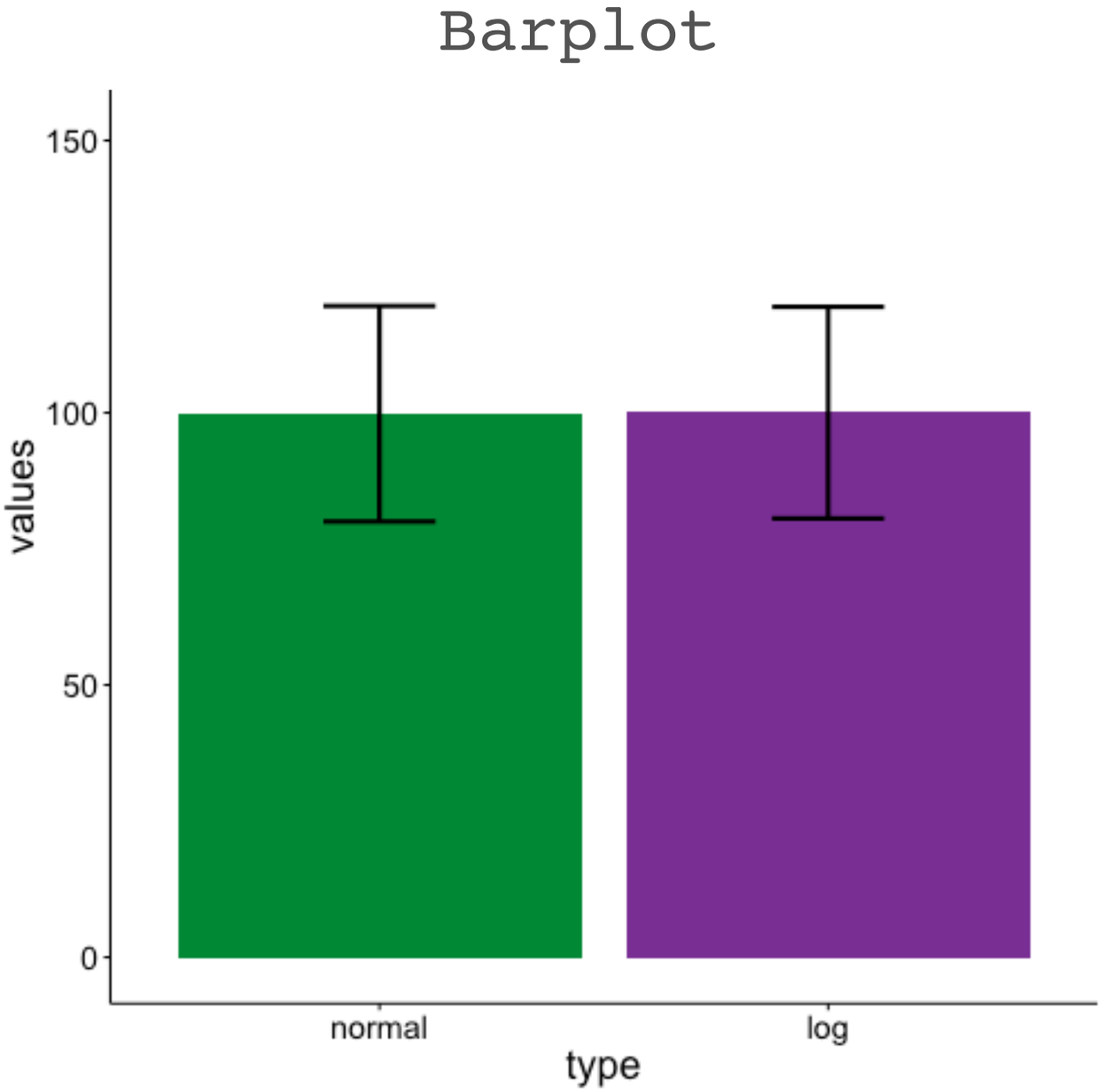
https://pagepiccinini.com/2016/02/23/boxplots-vs-barplots/

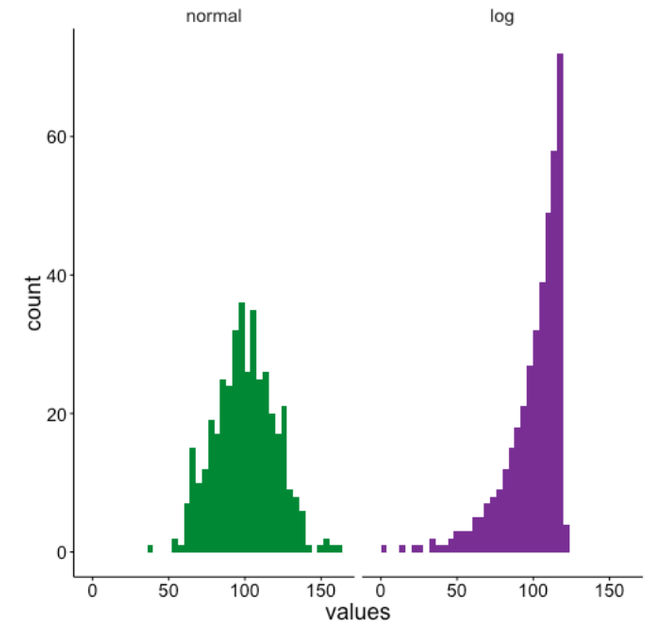
Histogram

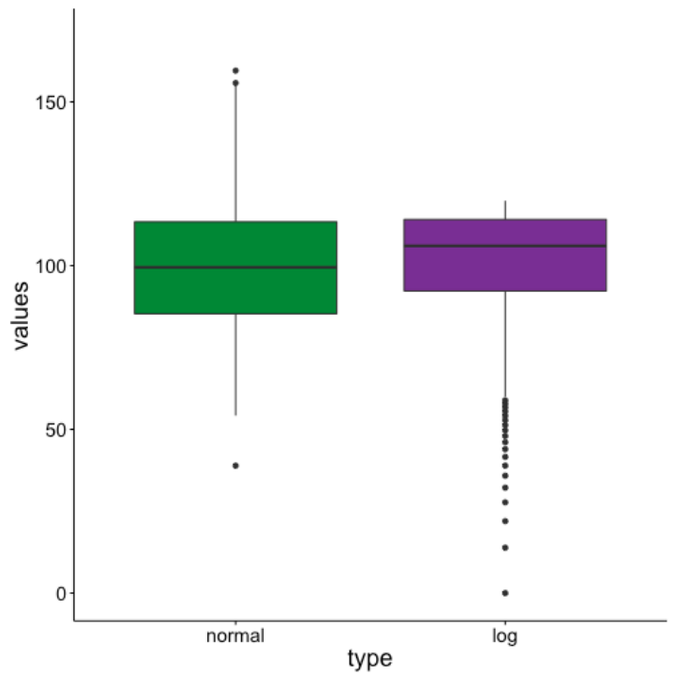


mean:100
sd: 20

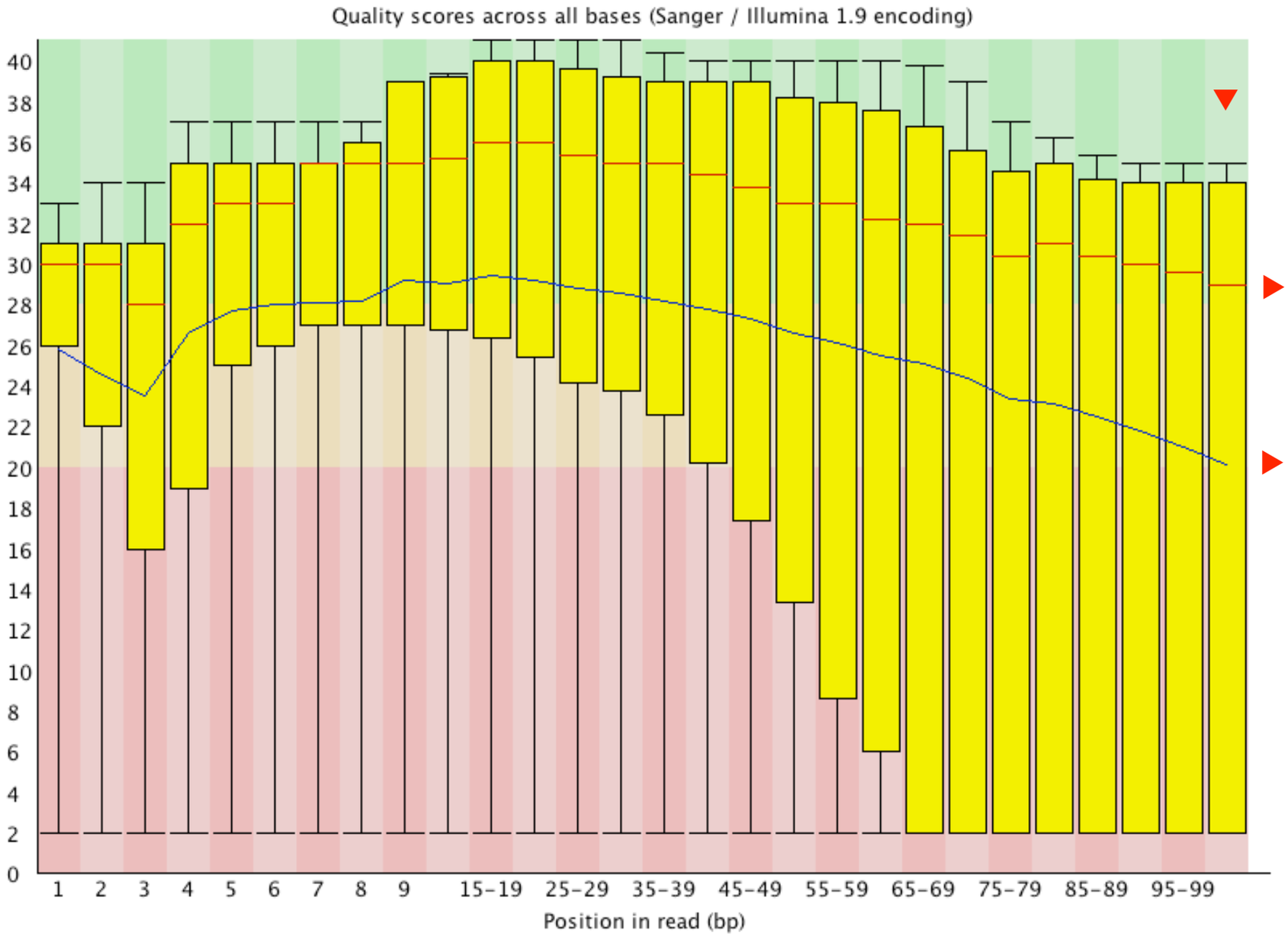
https://pagepiccinini.com/2016/02/23/boxplots-vs-barplots/







https://pagepiccinini.com/2016/02/23/boxplots-vs-barplots/

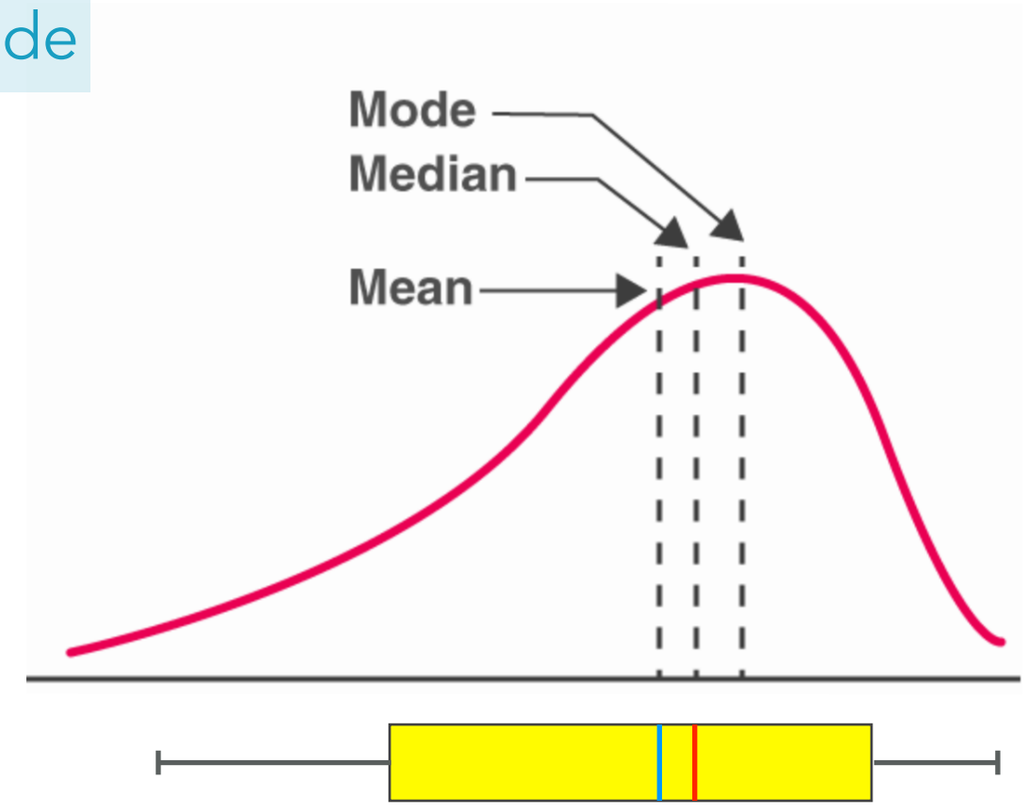


- $Q_{(median)}$ = 30 → 99.9%
- $Q_{(mean)}$ = 20 → 99%

GDC Genetic Diversity Centre Zurich

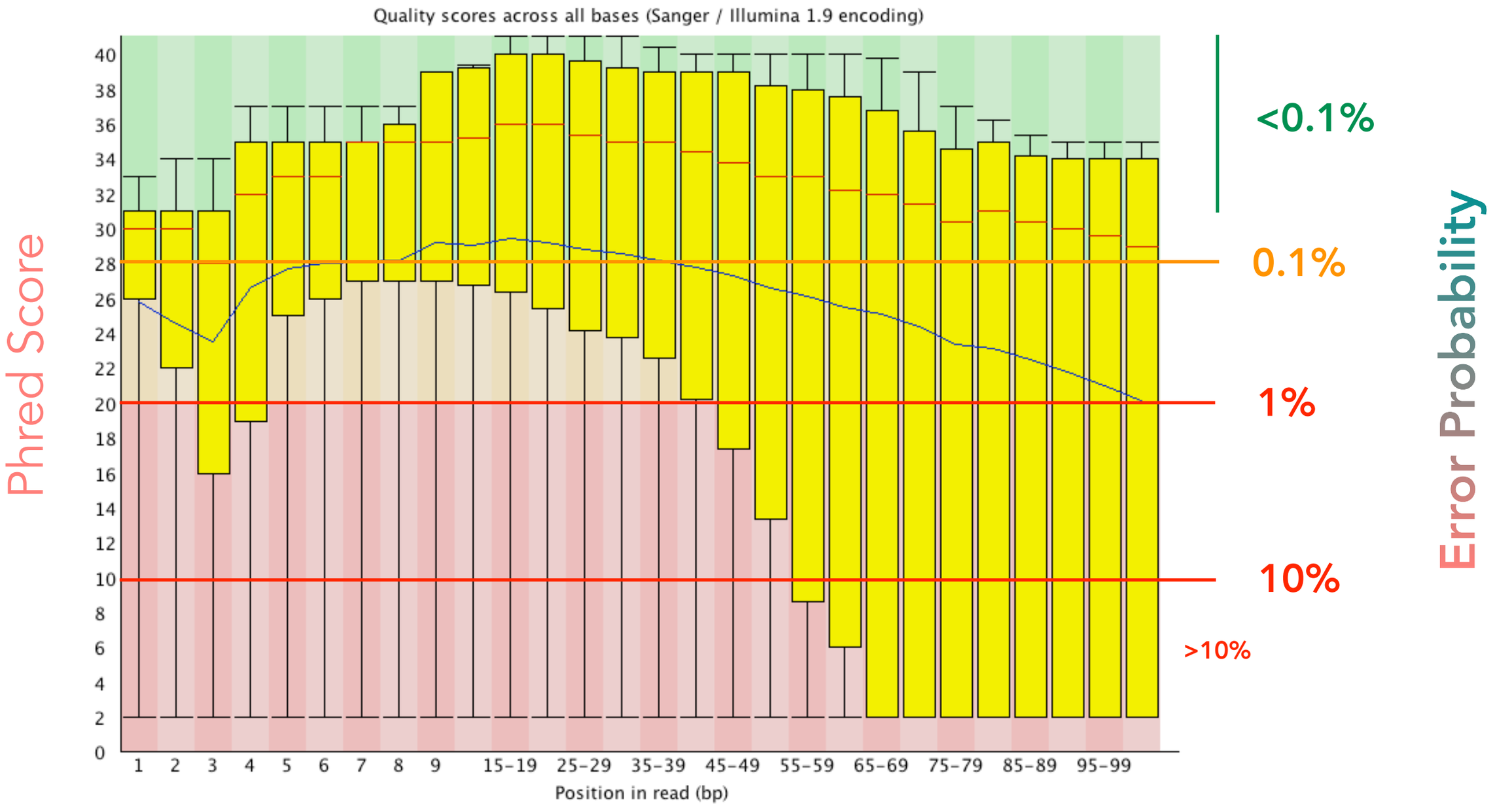
Symmetrical Distribution

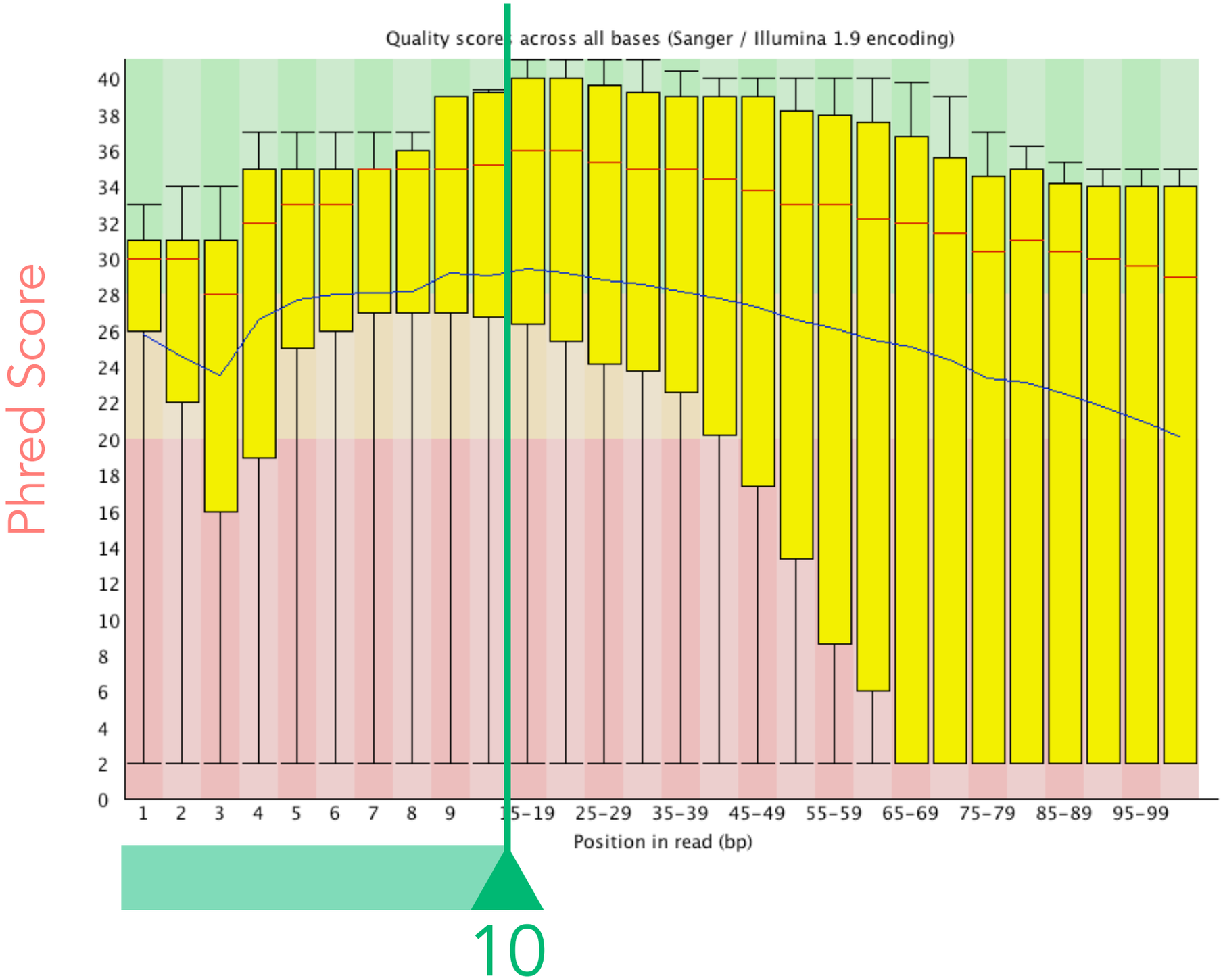
mean = median = mode

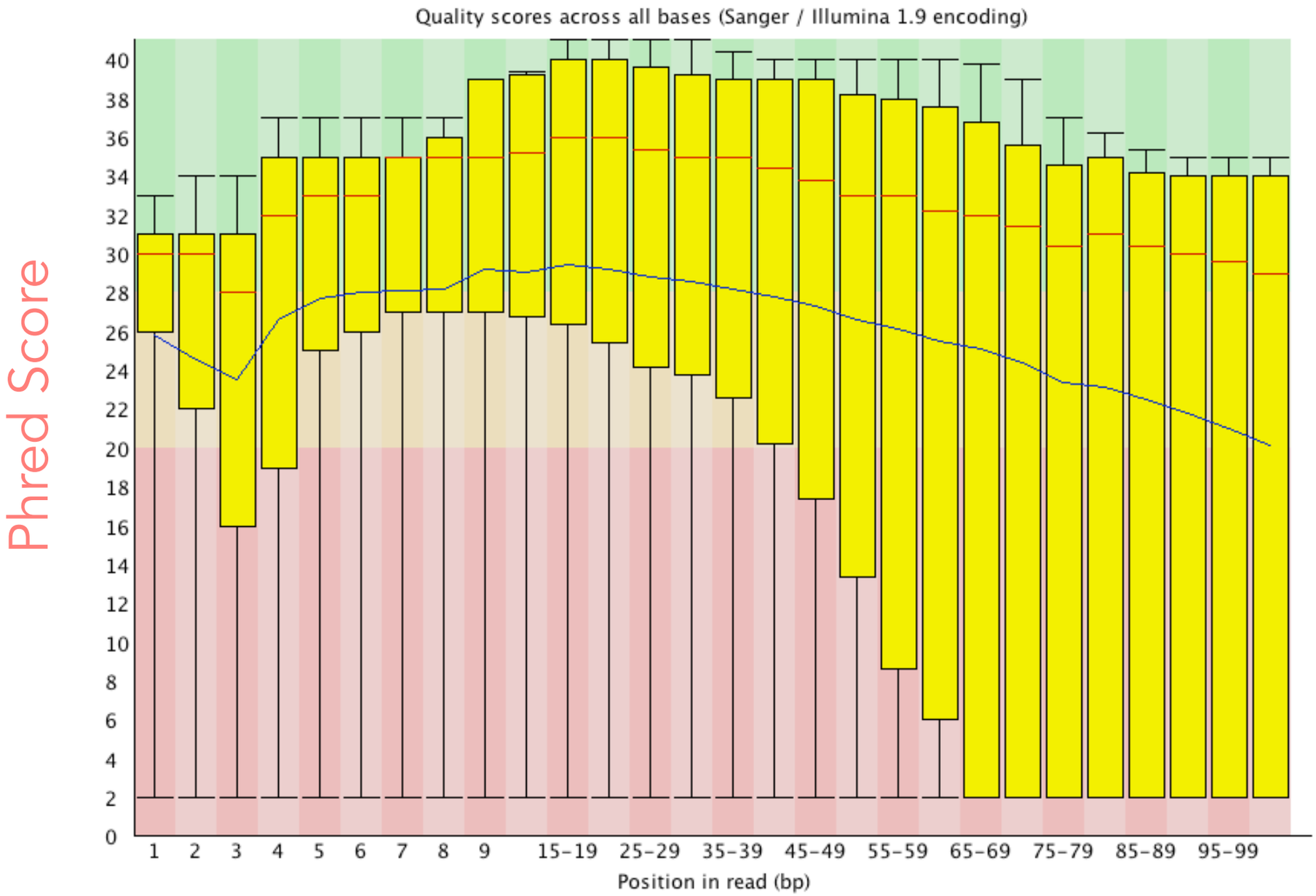


mean < median

skewed to the left


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Phred Score
Error Probability
<0.1%
0.1%
1%
10%
>10%
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)


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Phred Score
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
10-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
Position in read (bp)
10



Note: Error rate increases along the length of the read.

$Q_{mean\ phred\ score}$: 30

150nt

% 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).

$Q_{mean} = 30$

Genetic Diversity Centre Zurich

Q_{mean} =30

Q_{mean} =30

Genetic Diversity Centre Zurich

Q_{mean} =30

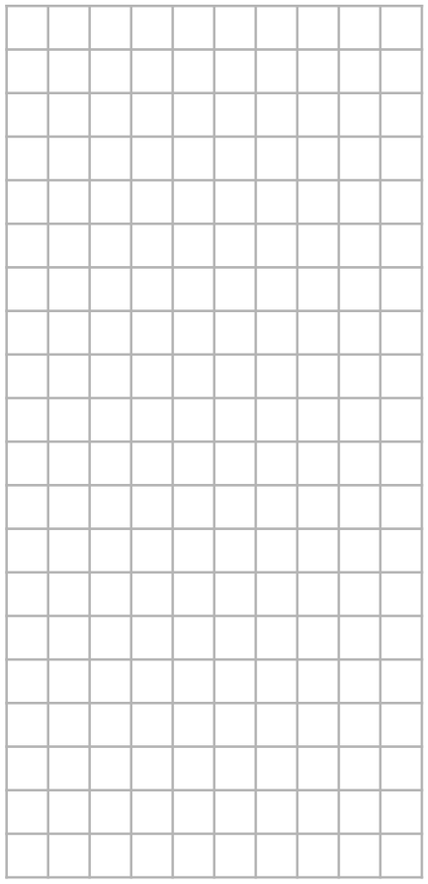
Q_{mean} =30

Q_{mean} =30

Genetic Diversity Centre Zurich

10 position

20 sequences



error rate:

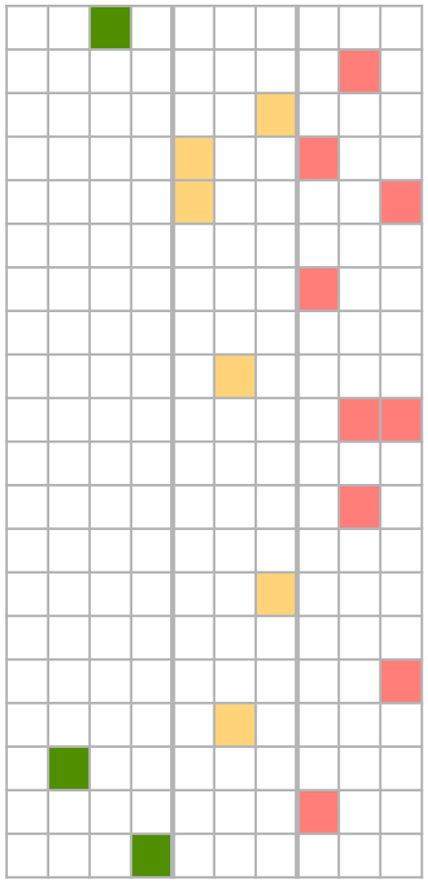
μ: 0.05 0.1 0.3

sequence #1

error at position #3

error rate:

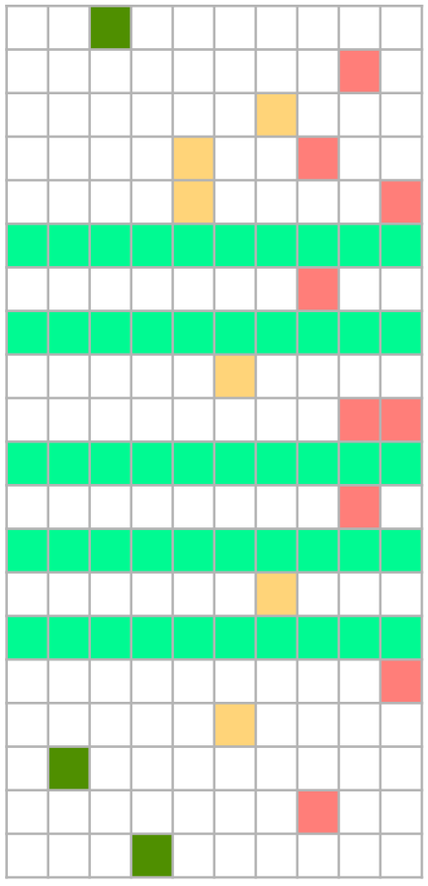
μ: 0.05 0.1 0.3



Genetic Diversity Centre Zurich

error rate:

μ: 0.05 0.1 0.3



5 (25%) error free reads

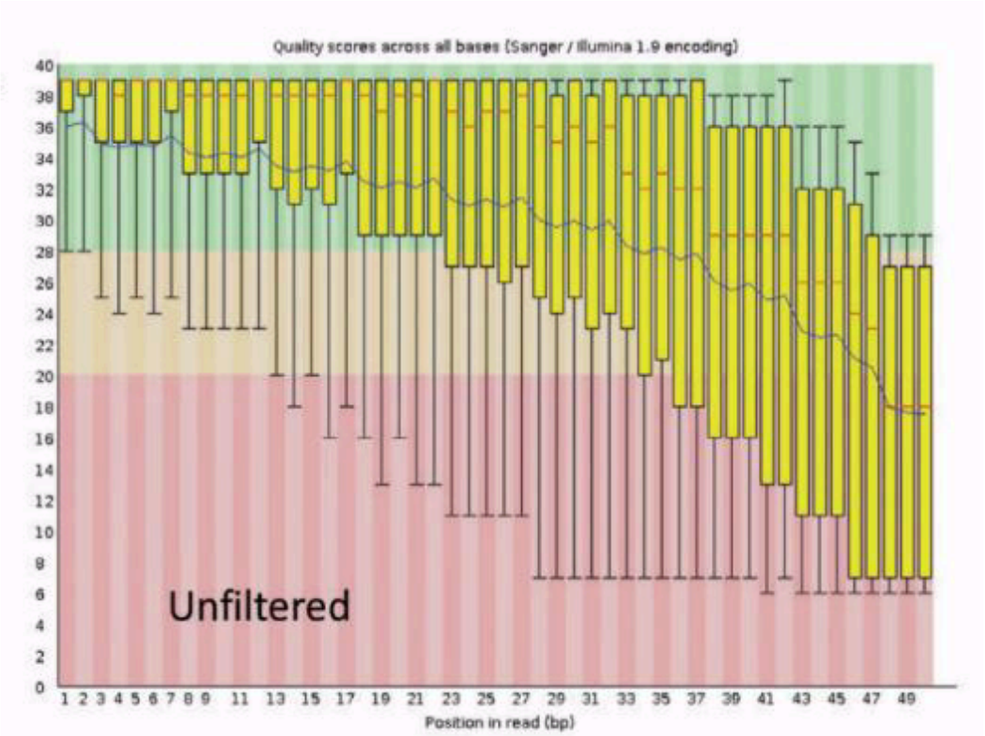
Genetic Diversity Centre Zurich

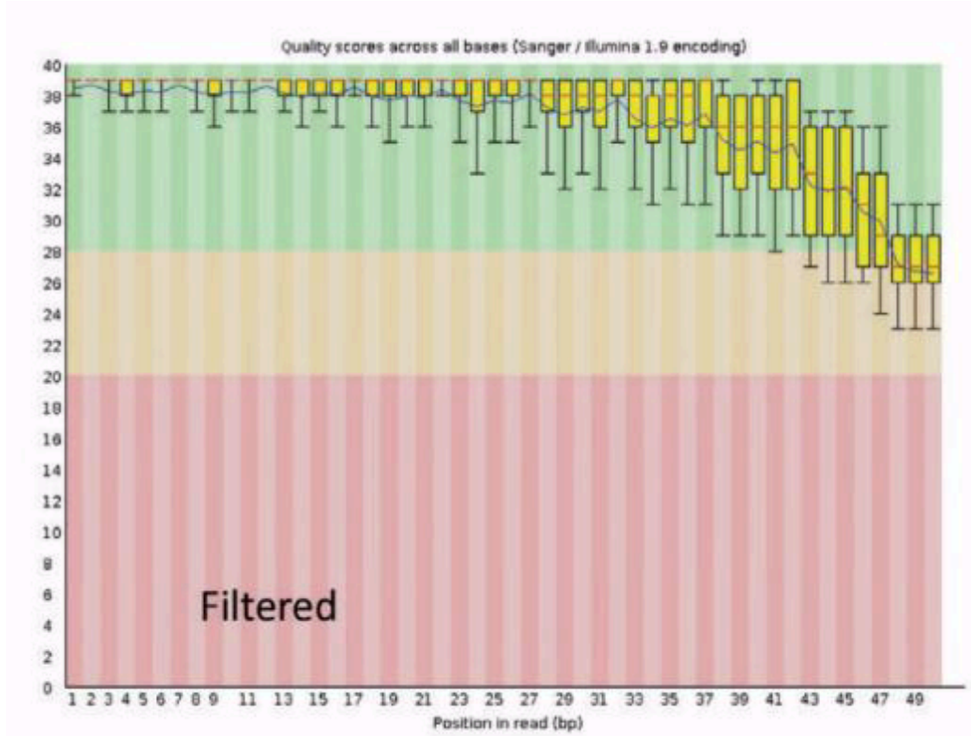
Q20 - 150nt

99% Accuracy / 1% Error Rate

$0.99^{150} \rightarrow$ **22% Error Free Reads**

For Better or Worse





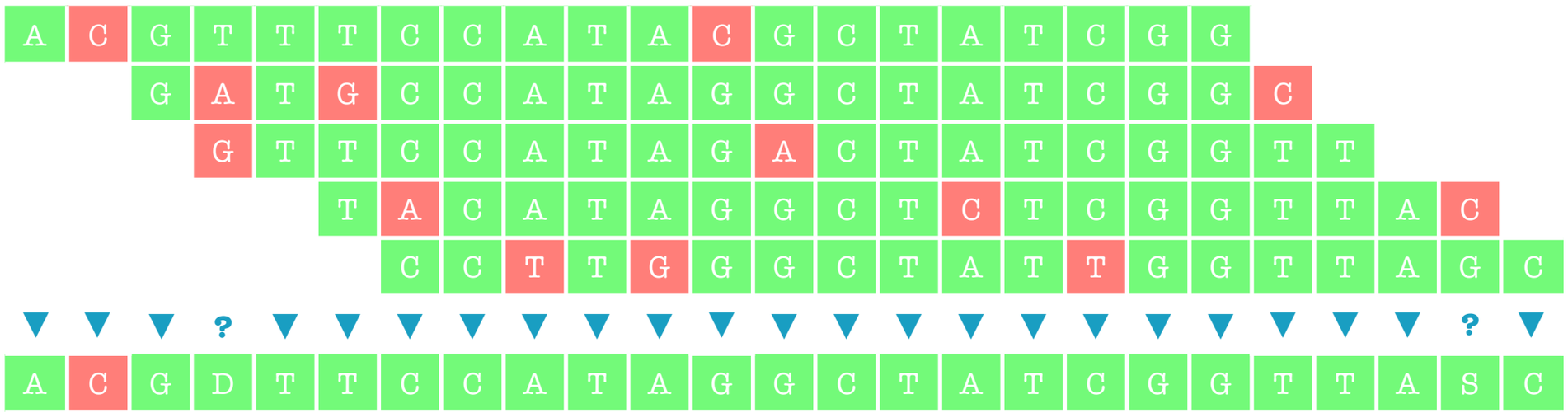
N_{reads} = 6Mio

N_{reads} = 2.5Mio (-60%)

Sequencing Error Rate

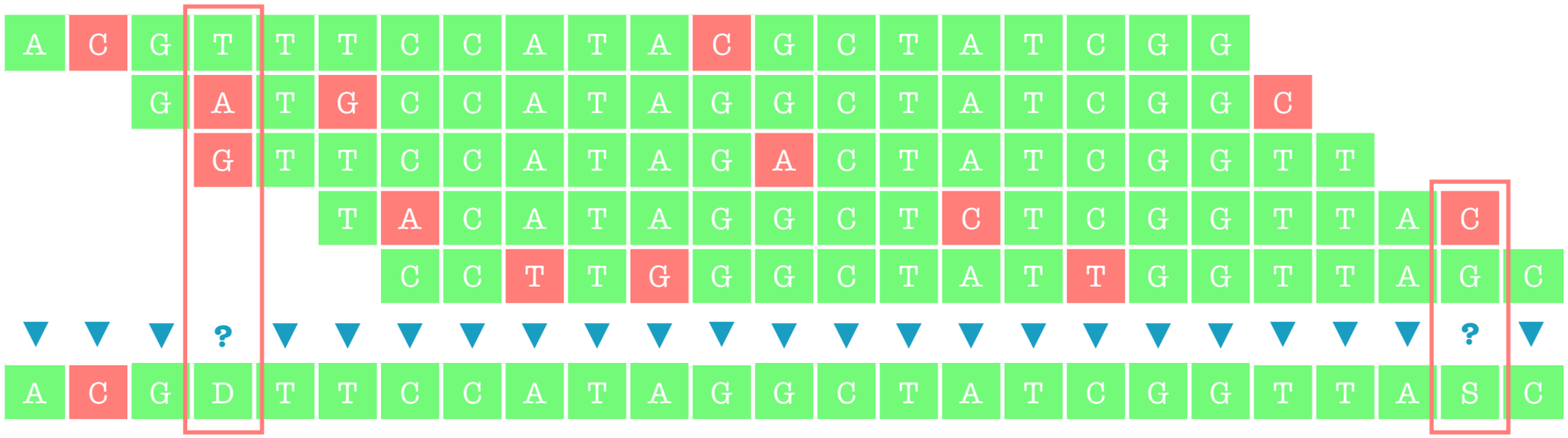
Zurich Genetic Diversity Centre GDC

Error Correction



- Read quality
- Number of reads (coverage)

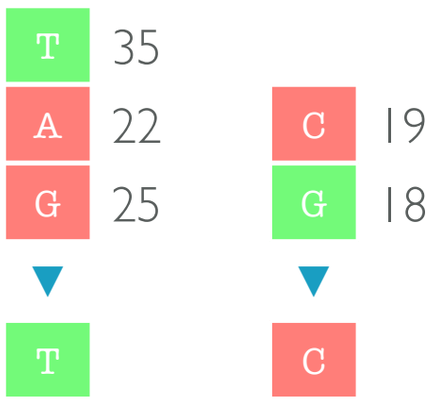
Error Correction



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.

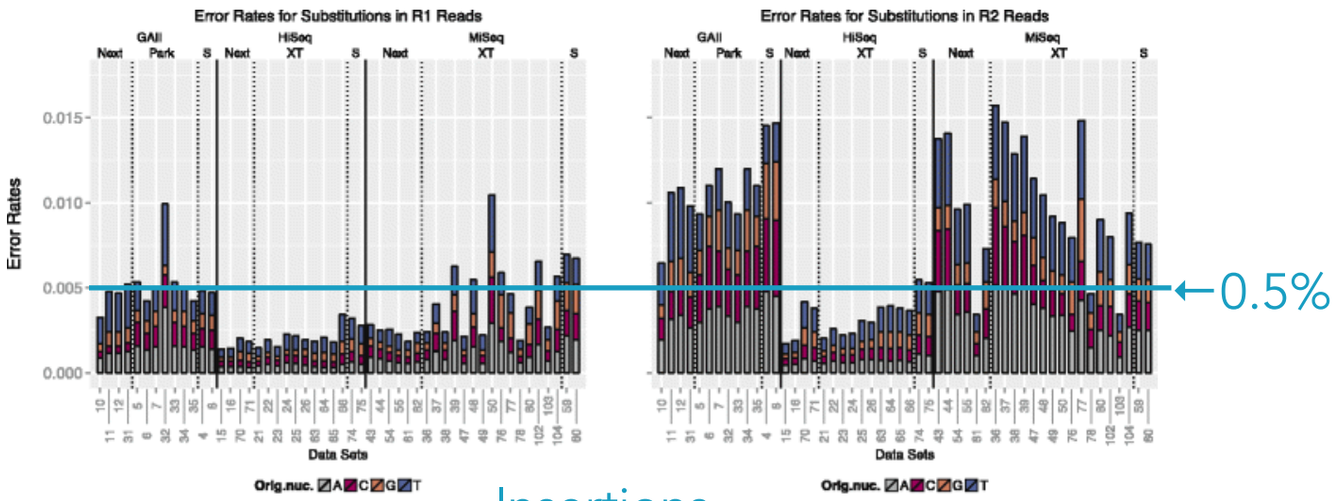
Results: We conducted a systematic investigation of errors and biases in Illumina data based on the largest collection of in vitro metagenomic data sets to date. We evaluated the Genome Analyzer II, HiSeq and MiSeq and tested state-of-the-art low input library preparation methods. Analysing in vitro 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**.

MPS ▷ Quality Control

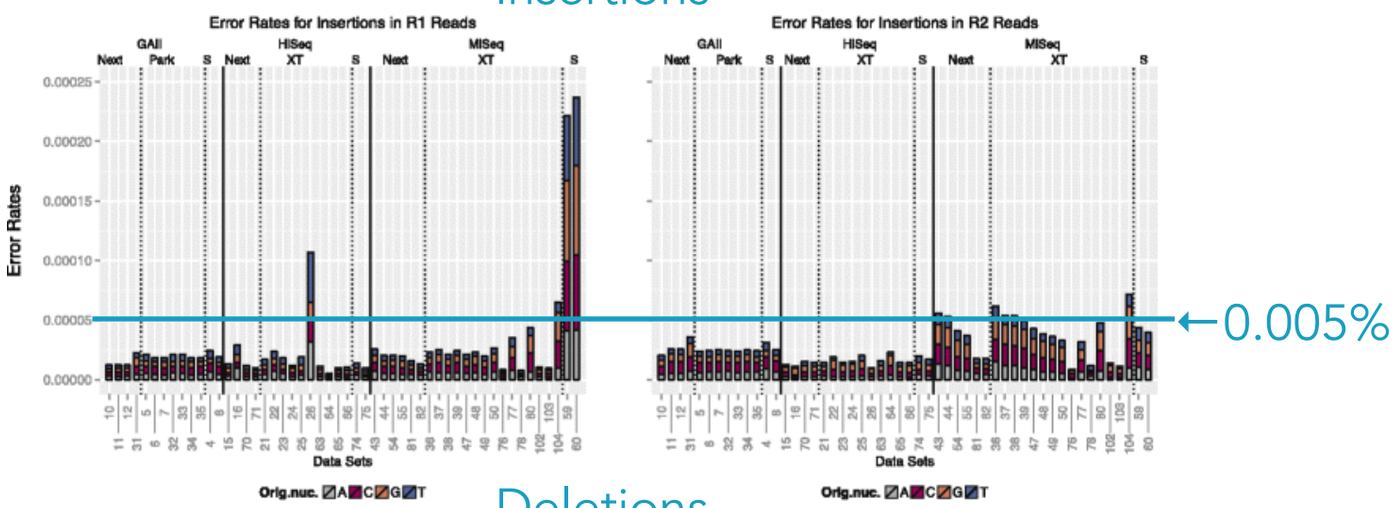
Zurich Genetic Diversity Centre

Substitutions



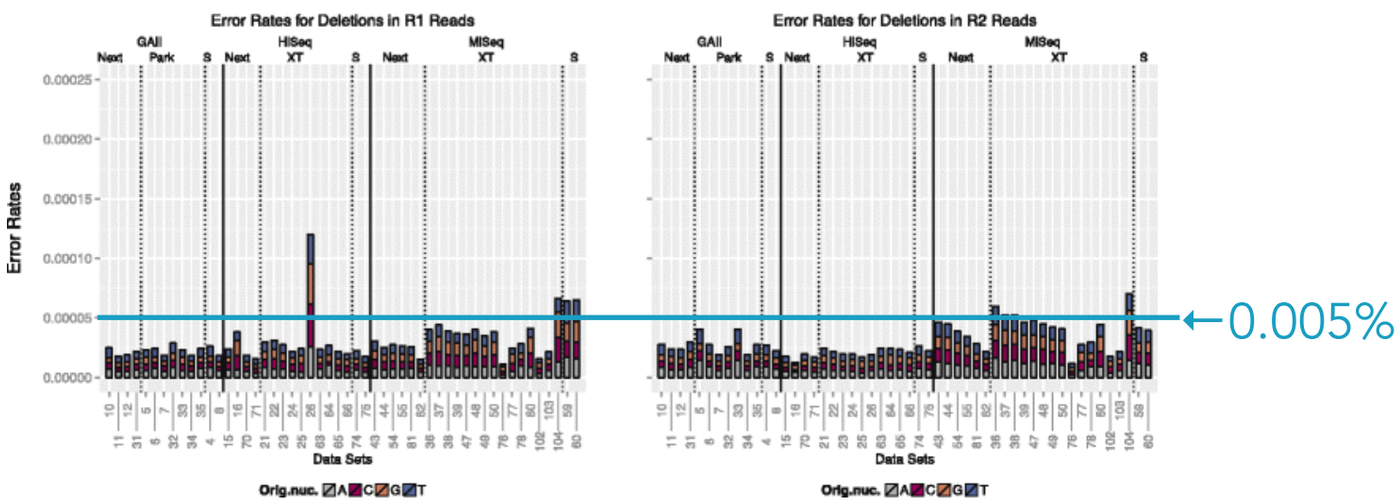
←0.5%

Insertions



←0.005%

Deletions



←0.005%

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

EE(R1) < EE(R2)

EE(Sub) > EE(Ins) ≈ EE(Del)

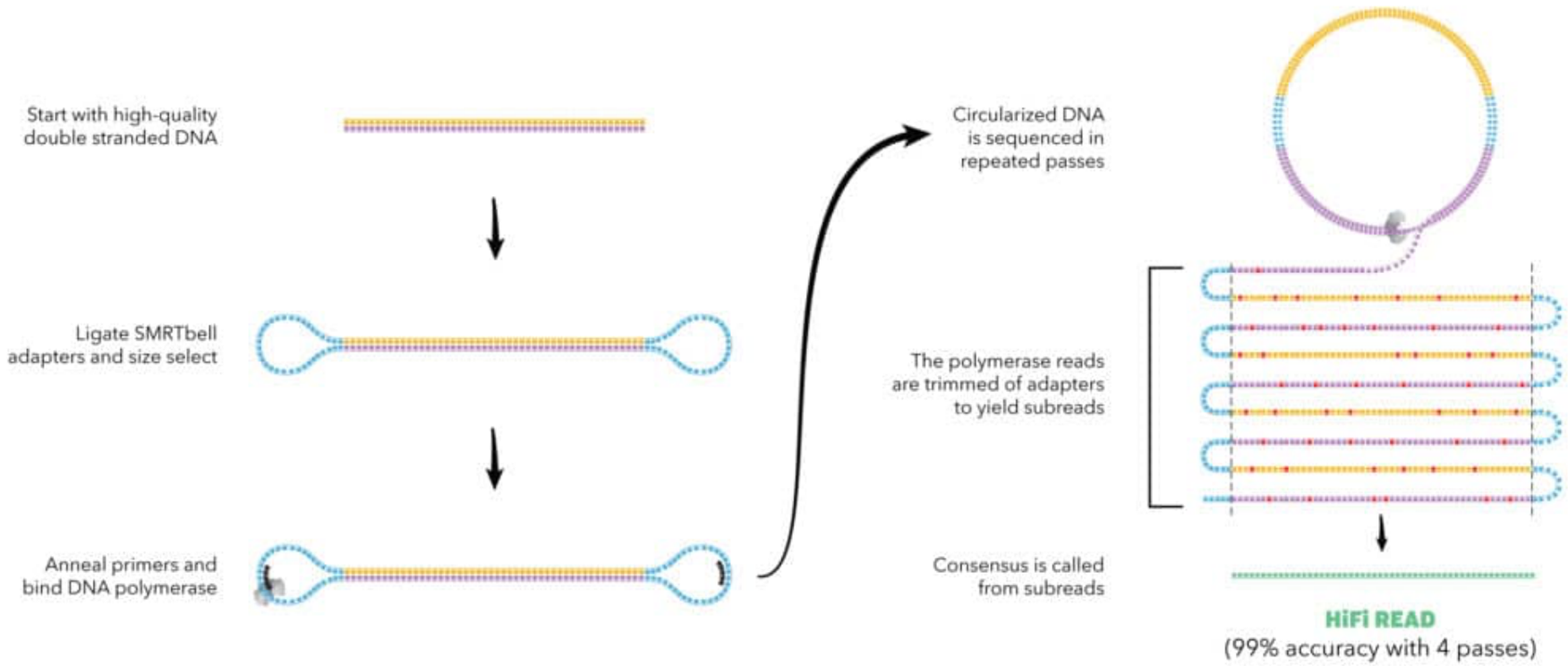
EE(HiSeq) < EE(MiSeq) ≈ EE(GAII)

Error Rate 8-12%

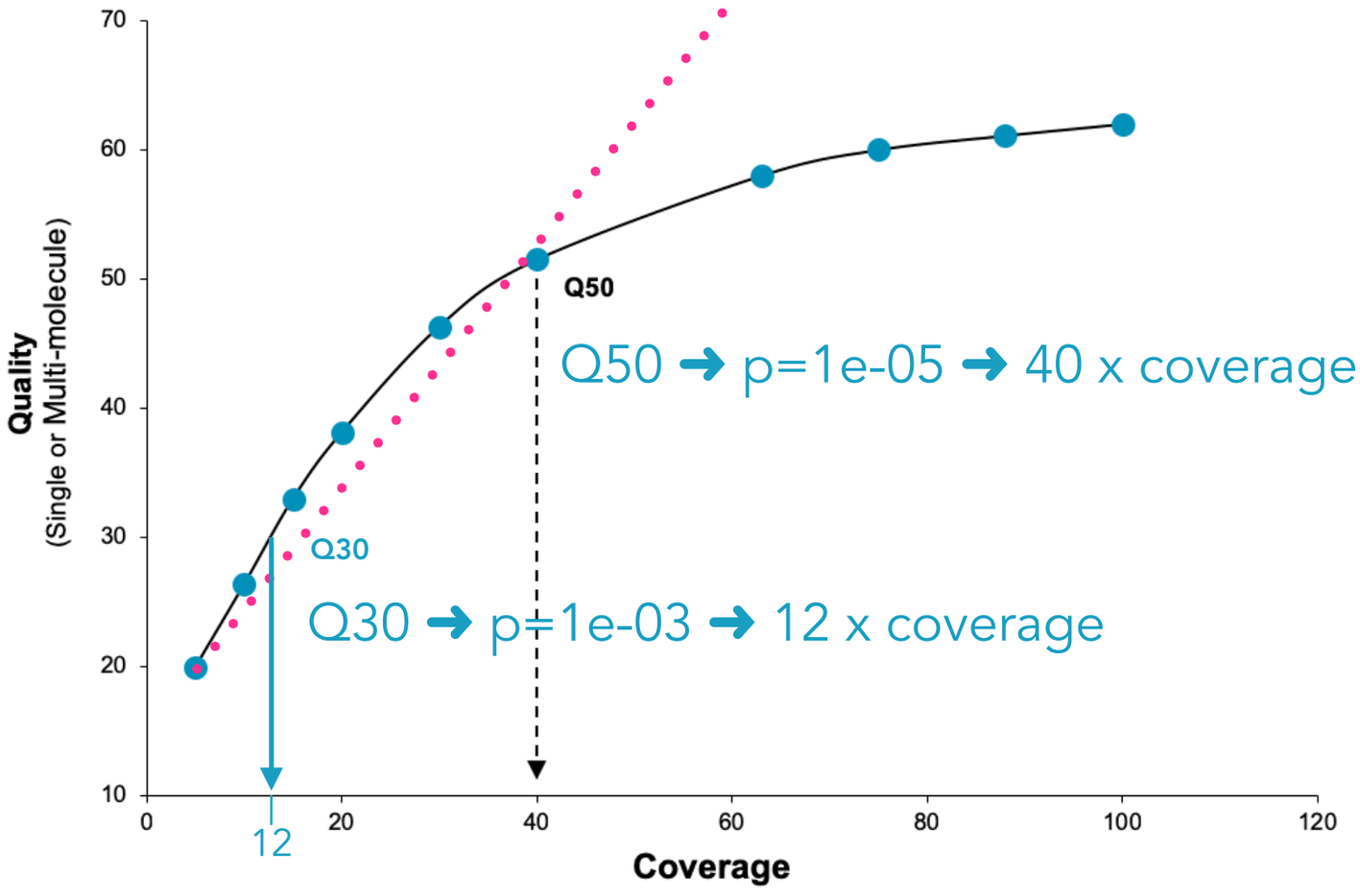
BAM ➜ FASTQ

BAM ➜ CCS.FASTX

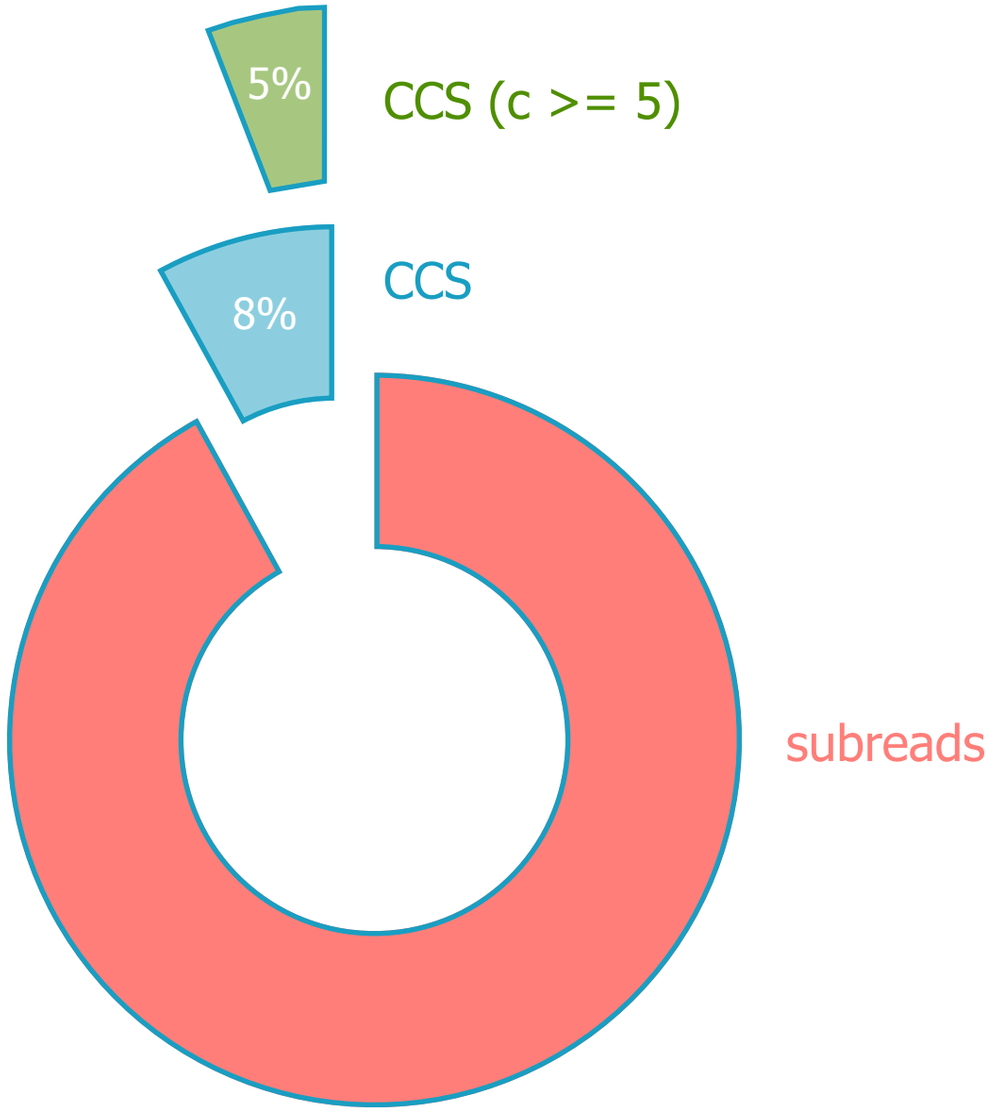
Circular Consensus Sequences (CCS)



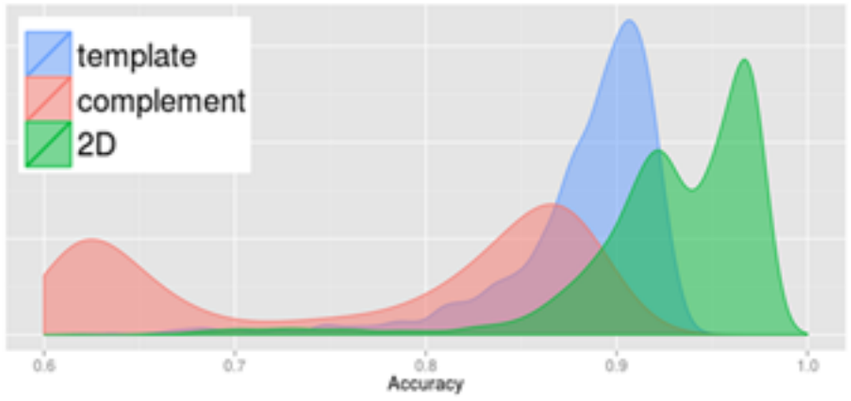
Why does it not improve anymore?

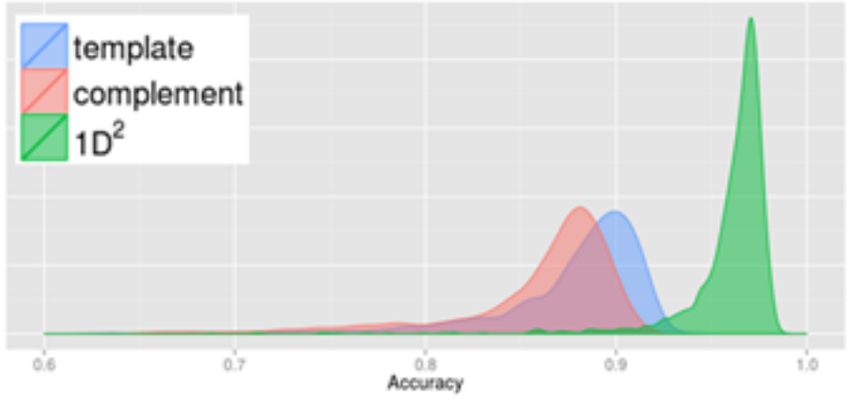


p: base-calling error probability


5%
CCS (c >= 5)
8%
CCS
subreads

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

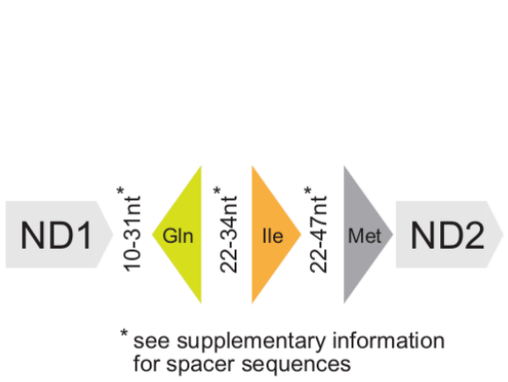
ARTICLE

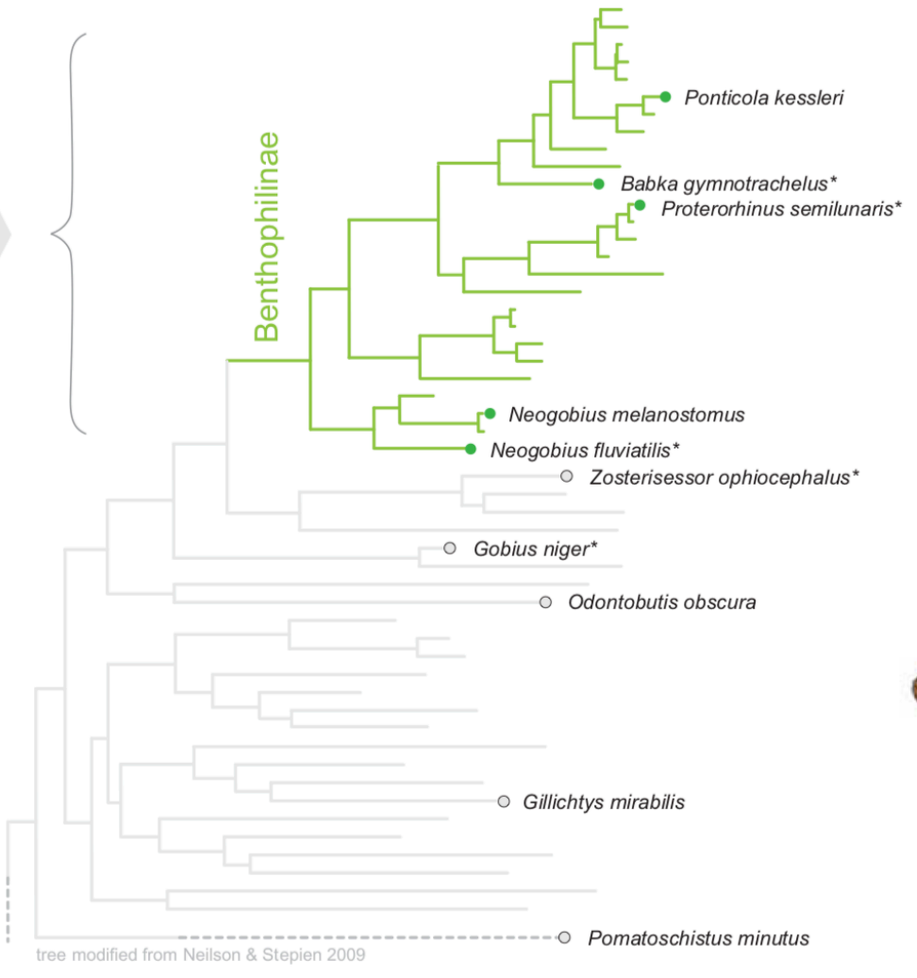
OPEN ACCESS Check for updates

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

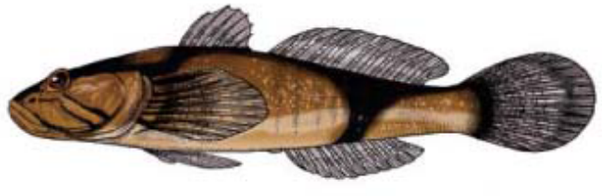
Silvia Gutnik[a], Jean-Claude Walser[b] and Irene Adrian-Kalchhauser[c]

[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.



```
Illumina MiSeq R1 -> EE mean  0.1
Illumina MiSeq R2 -> EE mean  0.2
PacBio Sequel IIe -> EE mean  2.9
PacBio   Revio    -> EE mean  1.2
ONT MinION        -> EE mean 30.7
```

Extras

Genetic Diversity Centre Zurich

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

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

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

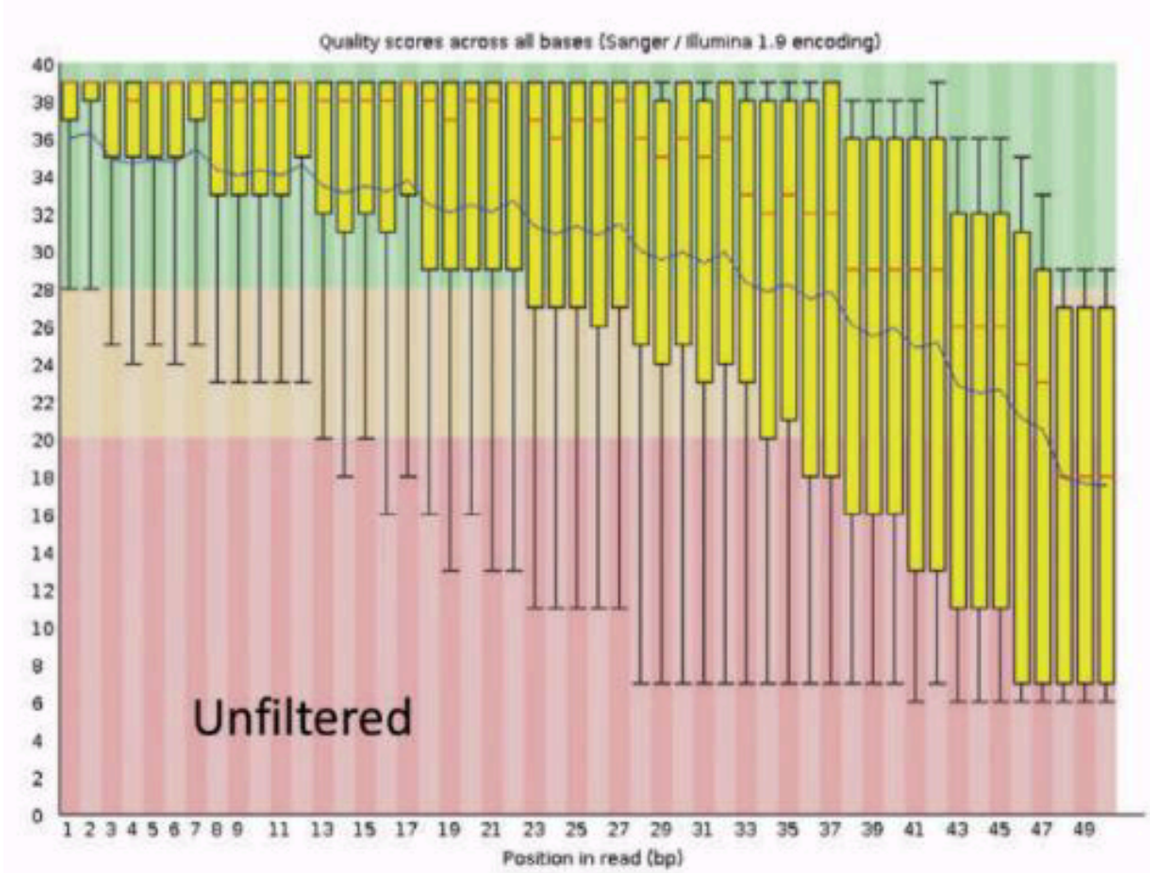
PRINSEQ
(http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi)

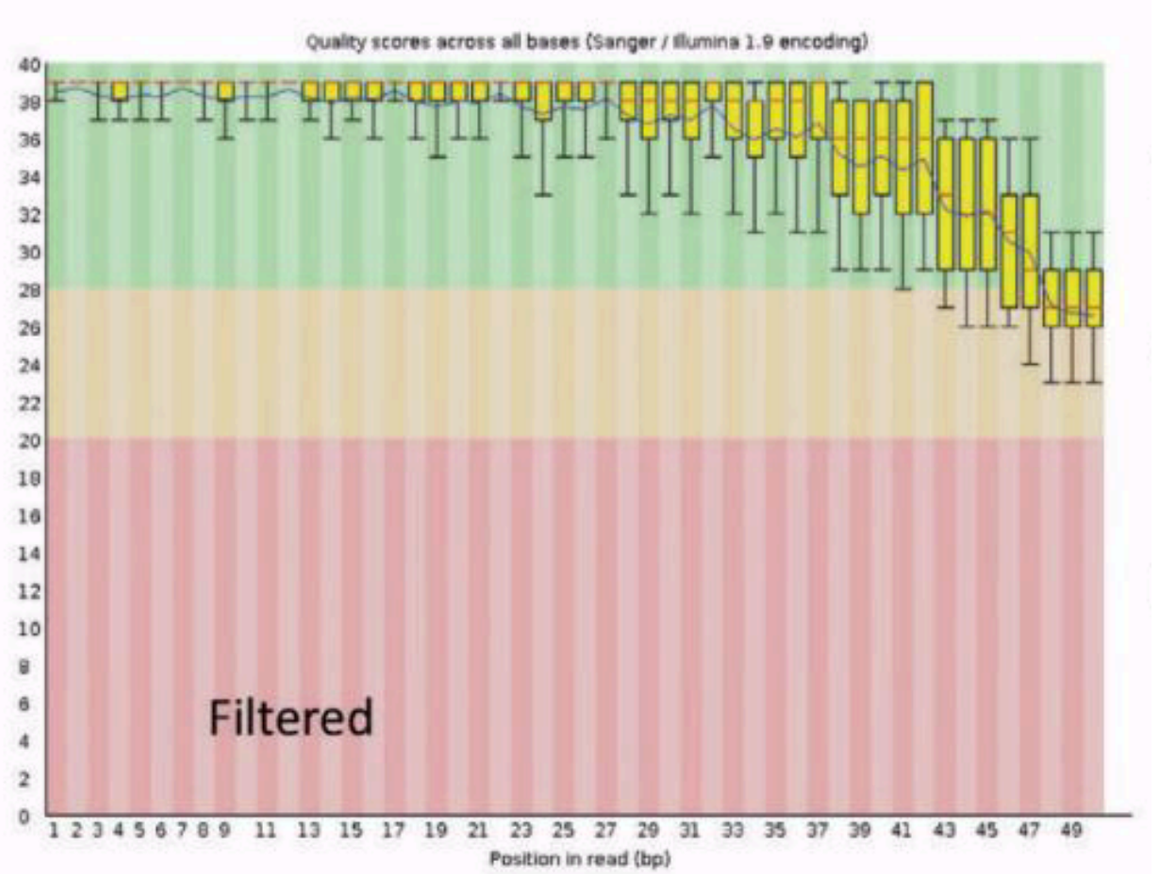
Galaxy
(http://galaxyproject.org)

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/)


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


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

Nucleic Acids Research Advance Access published December 16, 2009

Nucleic Acids Research, 2009, 1–5
doi:10.1093/nar/gkp1137

The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants

Peter J. A. Cock[1,*], Christopher J. Fields[2], Naohisa Goto[3], Michael L. Heuer[4] and Peter M. Rice[5]

ASCII encoded quality scores

```
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAAT
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfffffgggfafffcdfcfffbfdddeaegfgfgafaffW^a]
```

HiSeq score

39 33 38 33 38 38 23 30 33 29

Phred score

ASCII stands for American Standard Code for Information Interchange. Computers can only understand numbers, so an ASCII code is the numerical representation of a character or an action of some sort.

DEC	OCT	HEX	BIN	Symbol	HTML Number	HTML Name	Description
32	040	20	00100000		&#32;		Space
33	041	21	00100001	!	&#33;		Exclamation mark
34	042	22	00100010	"	&#34;	&quot;	Double quotes (or speech marks)
35	043	23	00100011	#	&#35;		Number
36	044	24	00100100	$	&#36;		Dollar
37	045	25	00100101	%	&#37;		Procenttecken
38	046	26	00100110	&	&#38;	&amp;	Ampersand
39	047	27	00100111	'	&#39;		Single quote
40	050	28	00101000	(	&#40;		Open parenthesis (or open bracket)
41	051	29	00101001	)	&#41;		Close parenthesis (or close bracket)
42	052	2A	00101010	*	&#42;		Asterisk
43	053	2B	00101011	+	&#43;		Plus
44	054	2C	00101100	,	&#44;		Comma
45	055	2D	00101101	-	&#45;		Hyphen
46	056	2E	00101110	.	&#46;		Period, dot or full stop
47	057	2F	00101111	/	&#47;		Slash or divide
48	060	30	00110000	0	&#48;		Zero
49	061	31	00110001	1	&#49;		One
50	062	32	00110010	2	&#50;		Two
51	063	33	00110011	3	&#51;		Three
52	064	34	00110100	4	&#52;		Four
53	065	35	00110101	5	&#53;		Five
54	066	36	00110110	6	&#54;		Six
55	067	37	00110111	7	&#55;		Seven

Nucleic Acids Research Advance Access published January 13, 2015

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gku1341

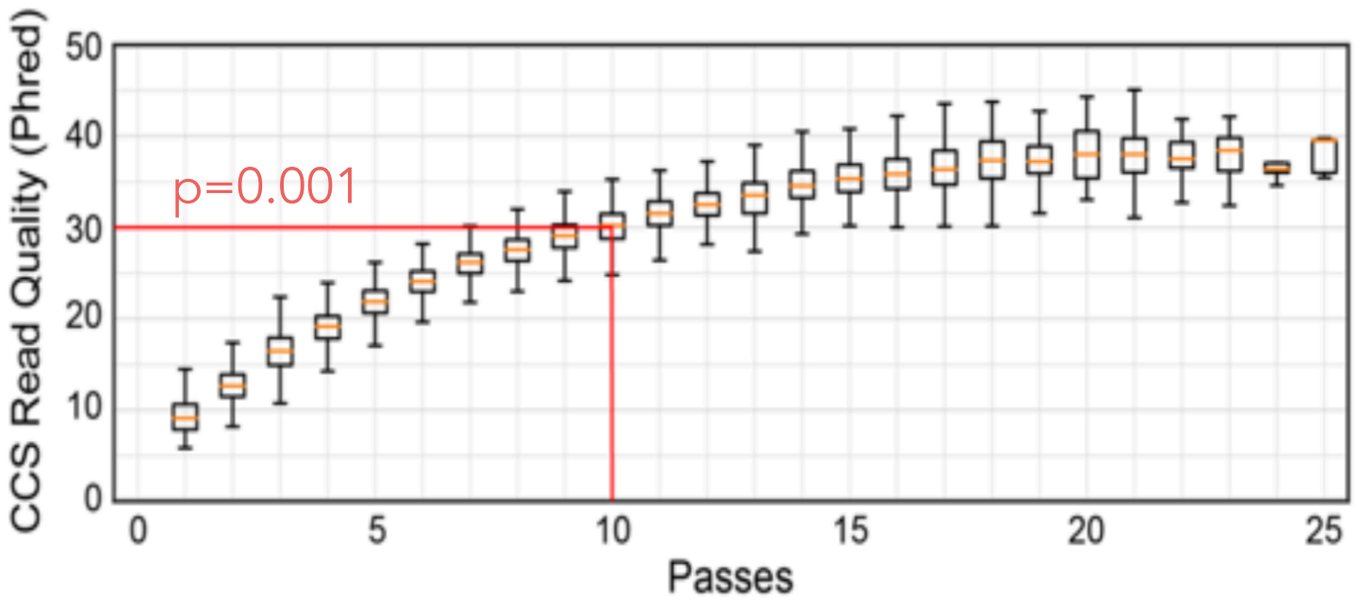
Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform

Melanie Schirmer[1,*], Umer Z. Ijaz[1], Rosalinda D'Amore[2], Neil Hall[2], William T. Sloan[1] and Christopher Quince[1]

[1]School of Engineering, University of Glasgow, Glasgow, UK and [2]Functional and Comparative Genomics, University of Liverpool, Liverpool, UK

With read lengths of currently up to 2 × 300 bp, high throughput and low sequencing costs Illumina's MiSeq is becoming one of the most utilized sequencing platforms worldwide. The platform is manageable and affordable even for smaller labs. This enables quick turnaround on a broad range of applications such as targeted gene sequencing, metagenomics, small genome sequencing and clinical molecular diagnostics. However, **Illumina error profiles are still poorly understood and programs are therefore not designed for the idiosyncrasies of Illumina data.** A better knowledge of the error patterns is essential for sequence analysis and vital if we are to draw valid conclusions. Studying true genetic variation in a population sample is fundamental for understanding diseases, evolution and origin. We conducted a large study on the error patterns for the MiSeq based on 16S rRNA amplicon sequencing data. We tested state-of-the-art library preparation methods for amplicon sequencing and showed that the library preparation method and the choice of primers are the most significant sources of bias and cause distinct error patterns. Furthermore we tested the efficiency of various error correction strategies and identified quality trimming (Sickle) combined with error correction (BayesHammer) followed by read overlapping (PANDAseq) as the most successful approach, reducing substitution error rates on average by 93%.

Wenger et al. (2019) Highly-accurate long-read sequencing improves variant detection and assembly of a human genome.



A SMRTbell library tightly distributed at 15 kb was chosen for circular consensus sequencing based on estimates of 150 kb polymerase read length and a requirement of 10 passes to achieve Q30 read accuracy. CCS reads with a predicted accuracy of at least Q20 (99%) were retained. The total CCS read yield was 89 Gb, an average of 2.3 Gb per SMRT Cell, with an average read length of 13.5 kb ± 1.2 kb. The predicted accuracy of the CCS reads has a median of Q30 (99.9%) and a mean of Q27 (99.8%).

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

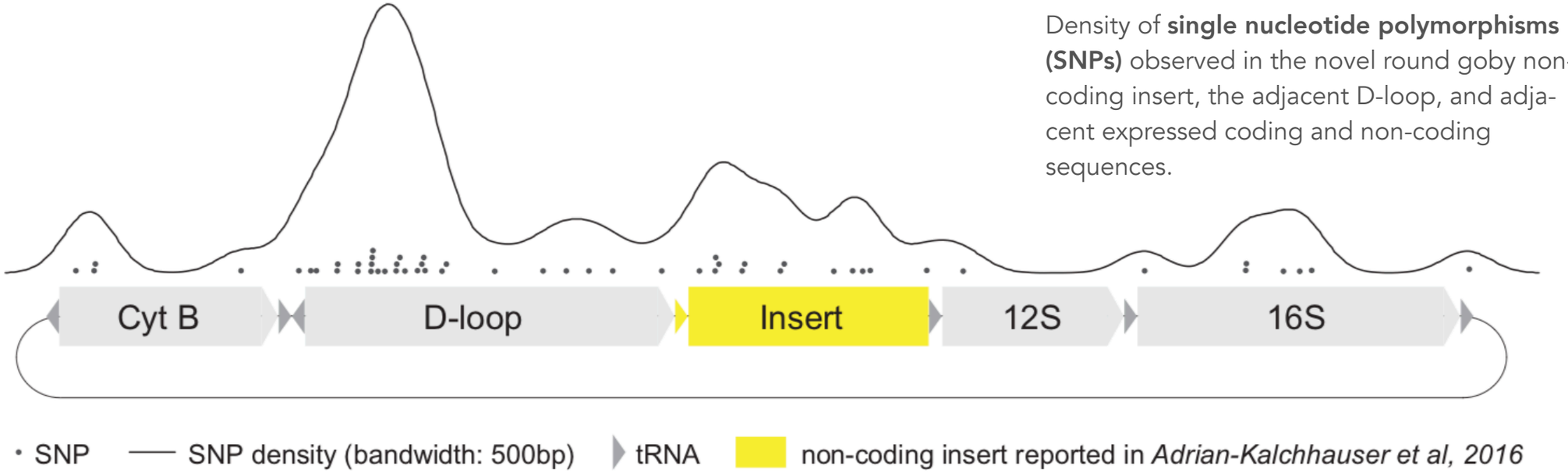
ARTICLE

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



Density of **single nucleotide polymorphisms (SNPs)** observed in the novel round goby non-coding insert, the adjacent D-loop, and adjacent expressed coding and non-coding sequences.