Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Next **G**eneration **S**equencing

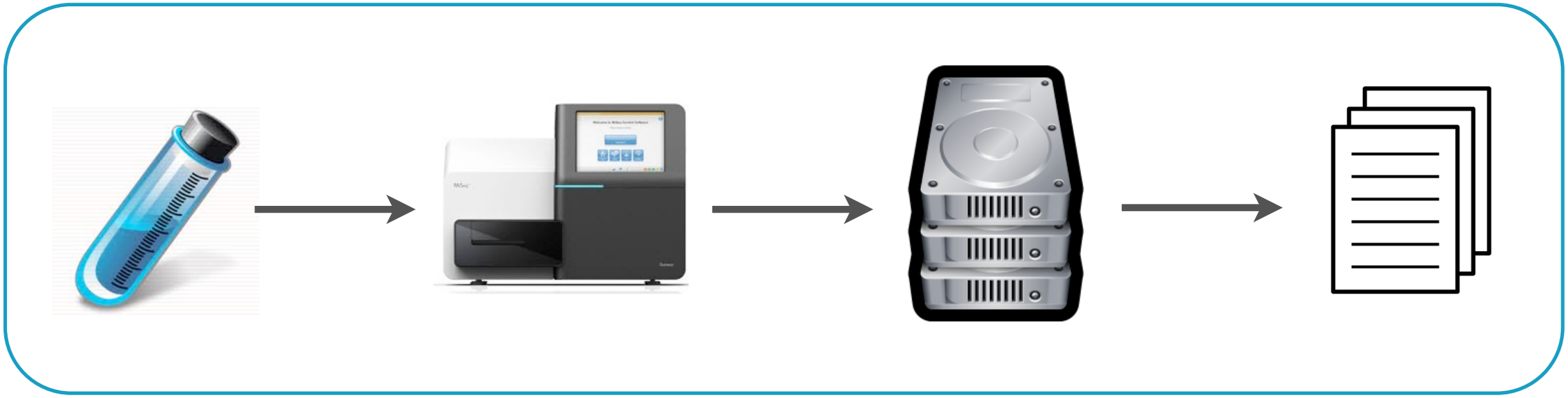
Jean-Claude Walser

jean-claude.walser@env.ethz.ch

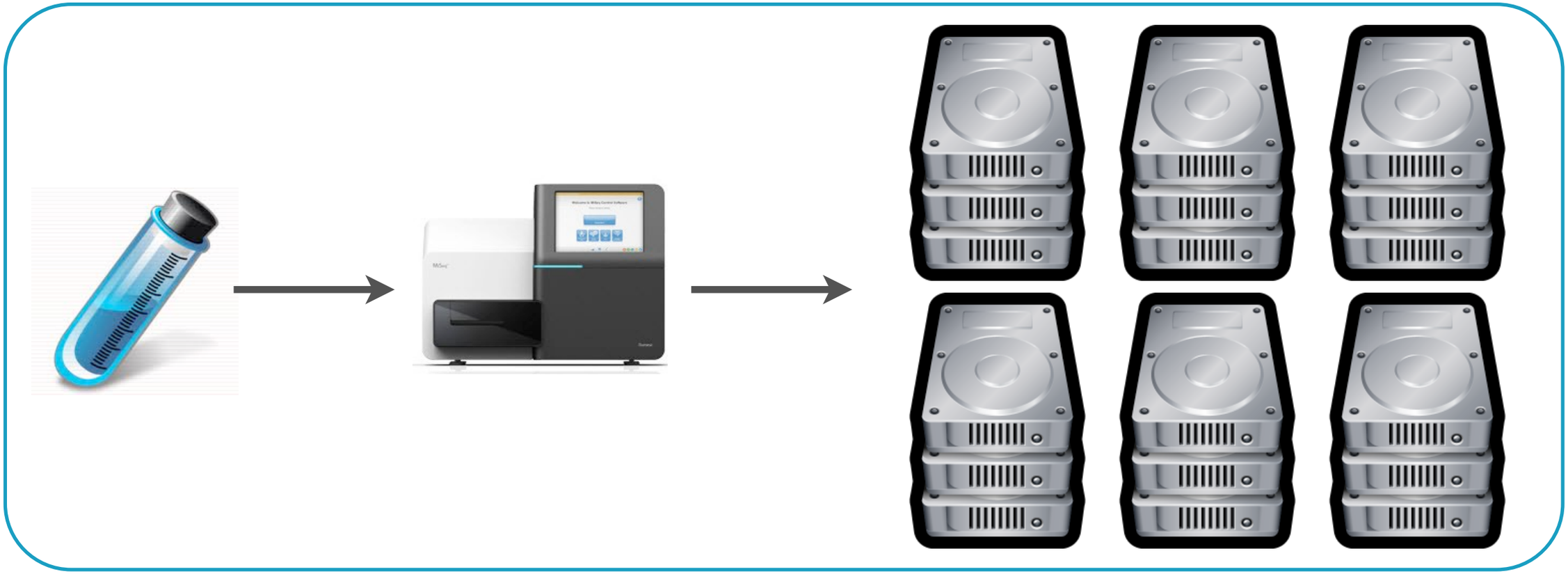
HS2020

Actually, that's the coffee machine...this is the next-gen sequencer.

Next (Next) Generation Sequencing **Hype**



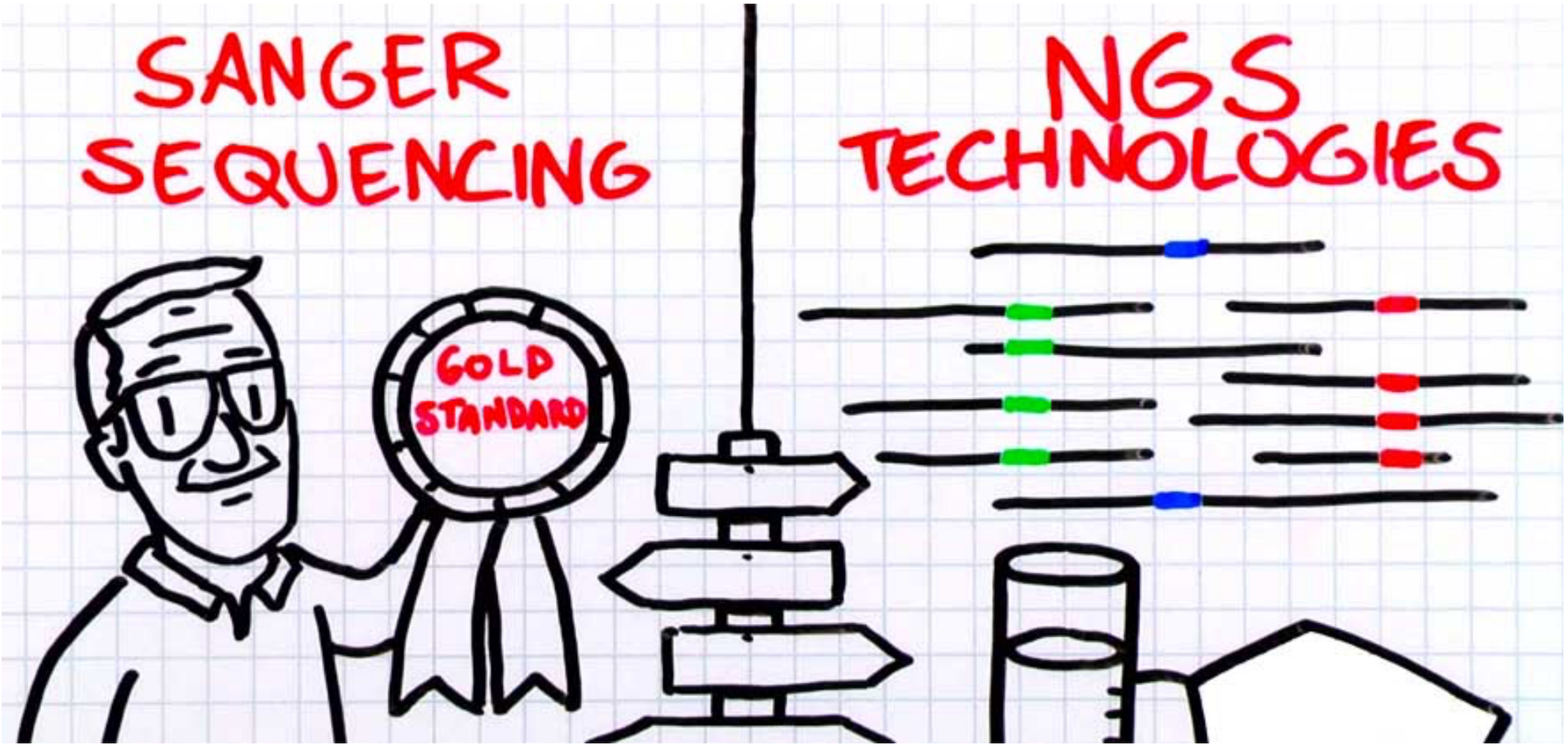
Next (Next) Generation Sequencing **Reality**

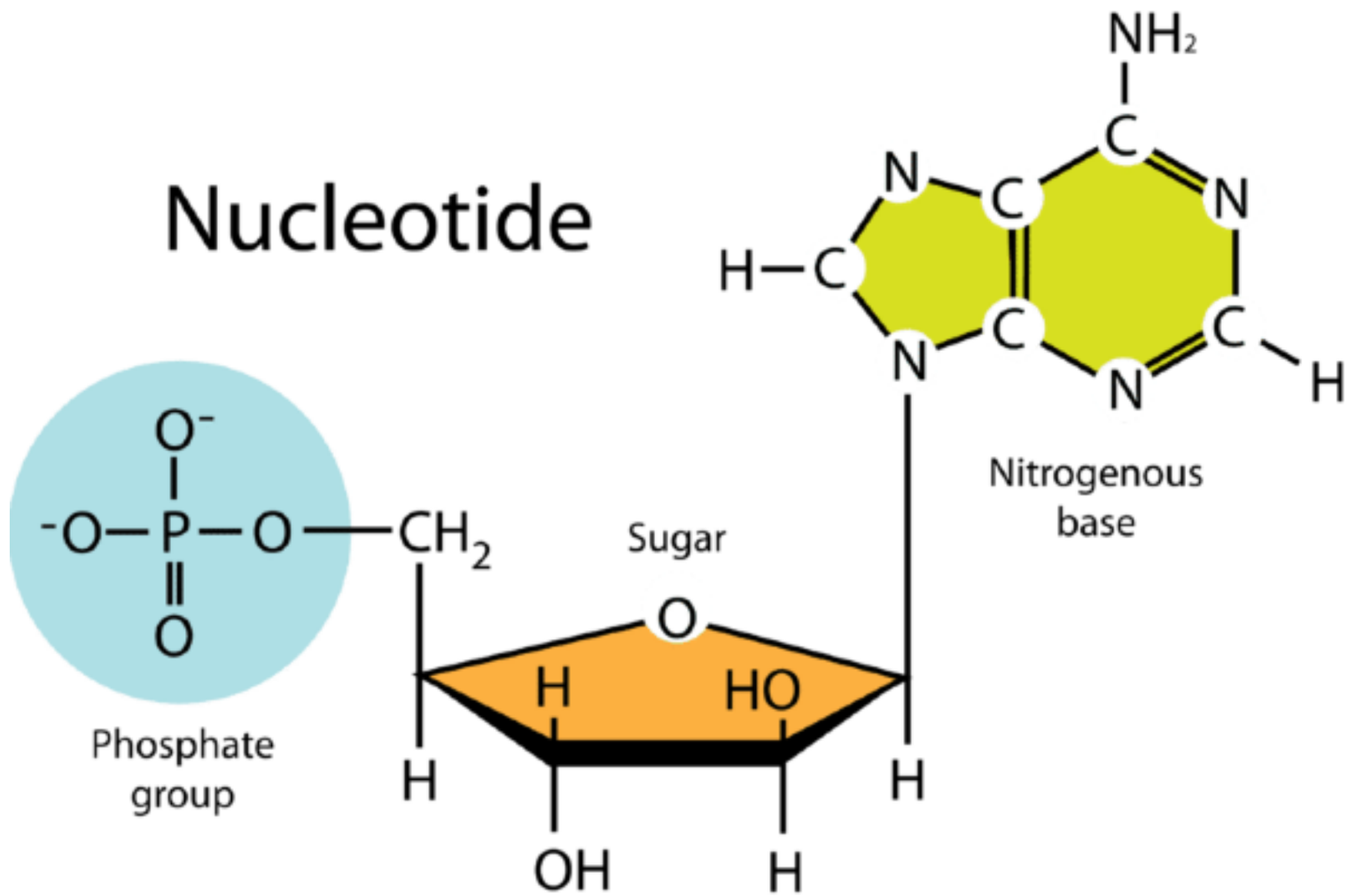


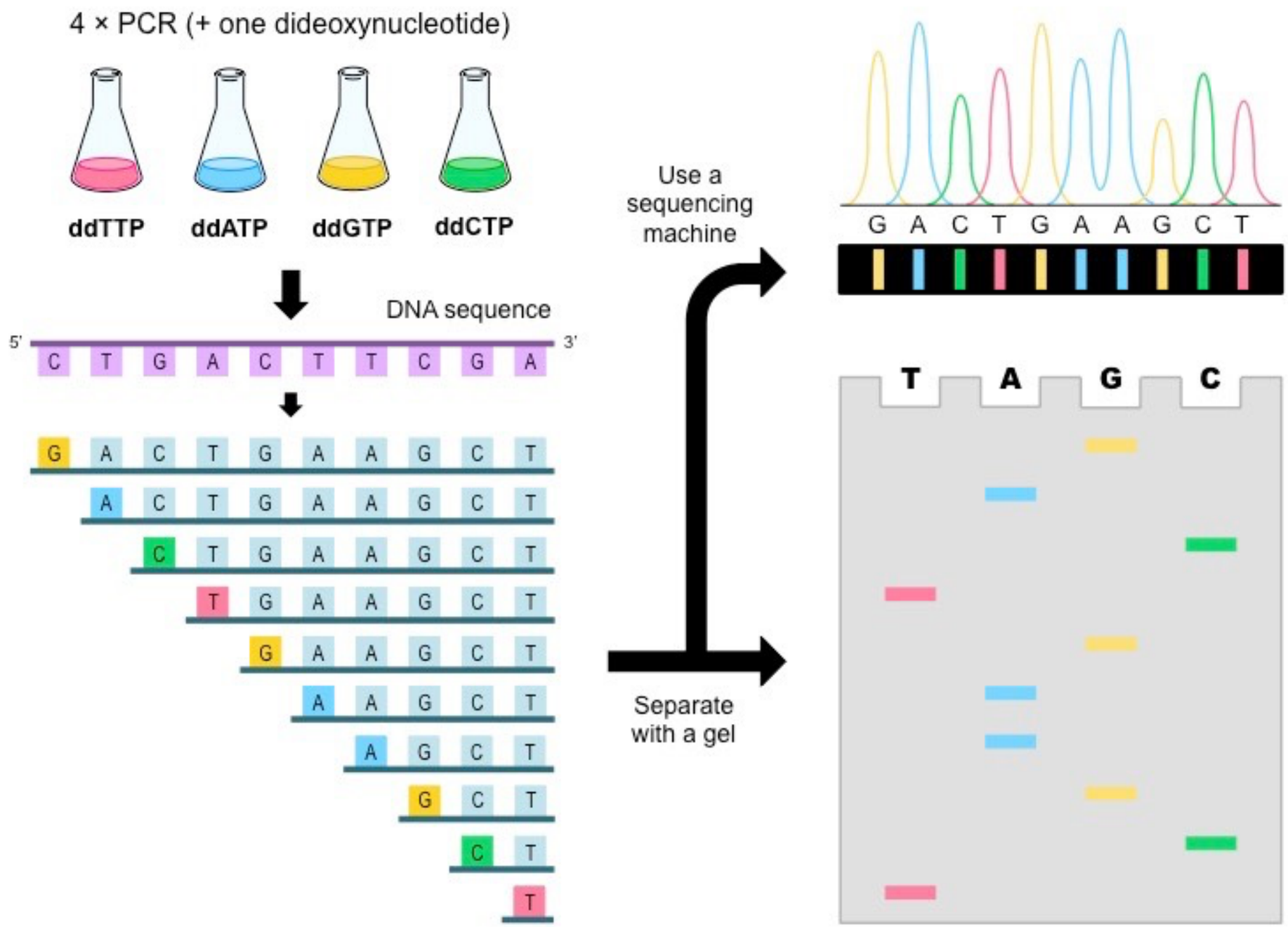
The **First Law of Technology** says we invariably **overestimate** the **short-term impact** of a truly transformational discovery, while **underestimating** its **longer-term effects**.

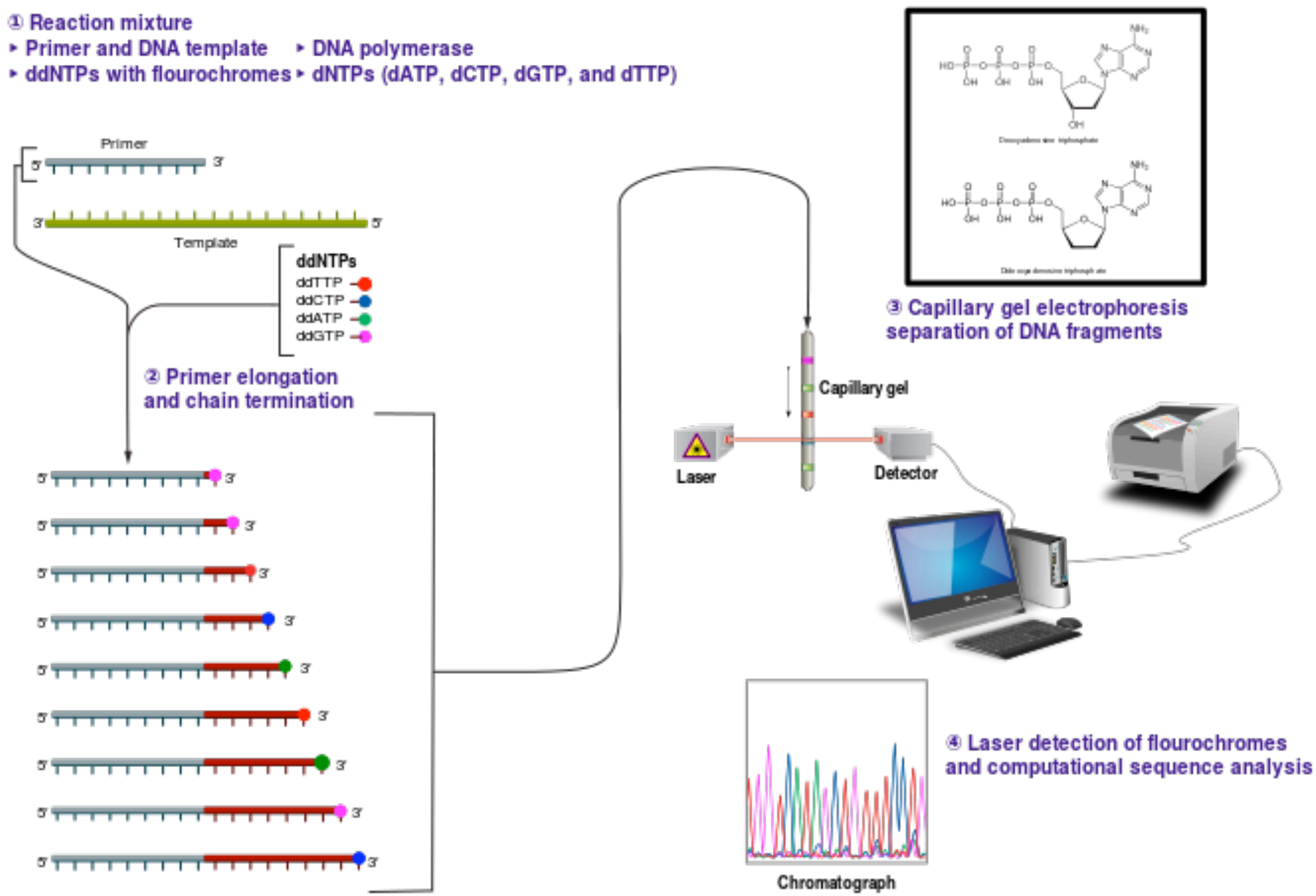
https://www.scientificamerican.com/

SEQUENCING TECHNOLOGIES


SANGER SEQUENCING
GOLD STANDARD
NGS TECHNOLOGIES


Nucleotide
NH2
C
N
C
N
H—C
C
C
N
N
H
Nitrogenous base
O⁻
-O—P—O—CH2
O
Sugar
O
H
HO
H
H
OH
H
Phosphate group


4 × PCR (+ one dideoxynucleotide)
ddTTP
ddATP
ddGTP
ddCTP
DNA sequence
5'
3'
C T G A C T T C G A
G A C T G A A G C T
A C T G A A G C T
C T G A A G C T
T G A A G C T
G A A G C T
A A G C T
A G C T
G C T
C T
T
Use a sequencing machine
Separate with a gel
G A C T G A A G C T
T A G C


① Reaction mixture
▸ Primer and DNA template ▸ DNA polymerase
▸ ddNTPs with flourochromes ▸ dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
5'
3'
3'
5'
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation
and chain termination
③ Capillary gel electrophoresis
separation of DNA fragments
Capillary gel
Laser
Detector
Chromatograph
④ Laser detection of flourochromes
and computational sequence analysis

Sanger (chain termination)

Roche 454 Pyrosequencing (pyrophosphate)

Ion Torrent (semiconductor technology)

Illumina Sequencing by Synthesis (fluorescent)

PacBio (fluorophore)

Nanopore (ionic current)

Helicos - SeqLL (fluorescent)

Bionano - Saphyr (third-generation optical mapping)

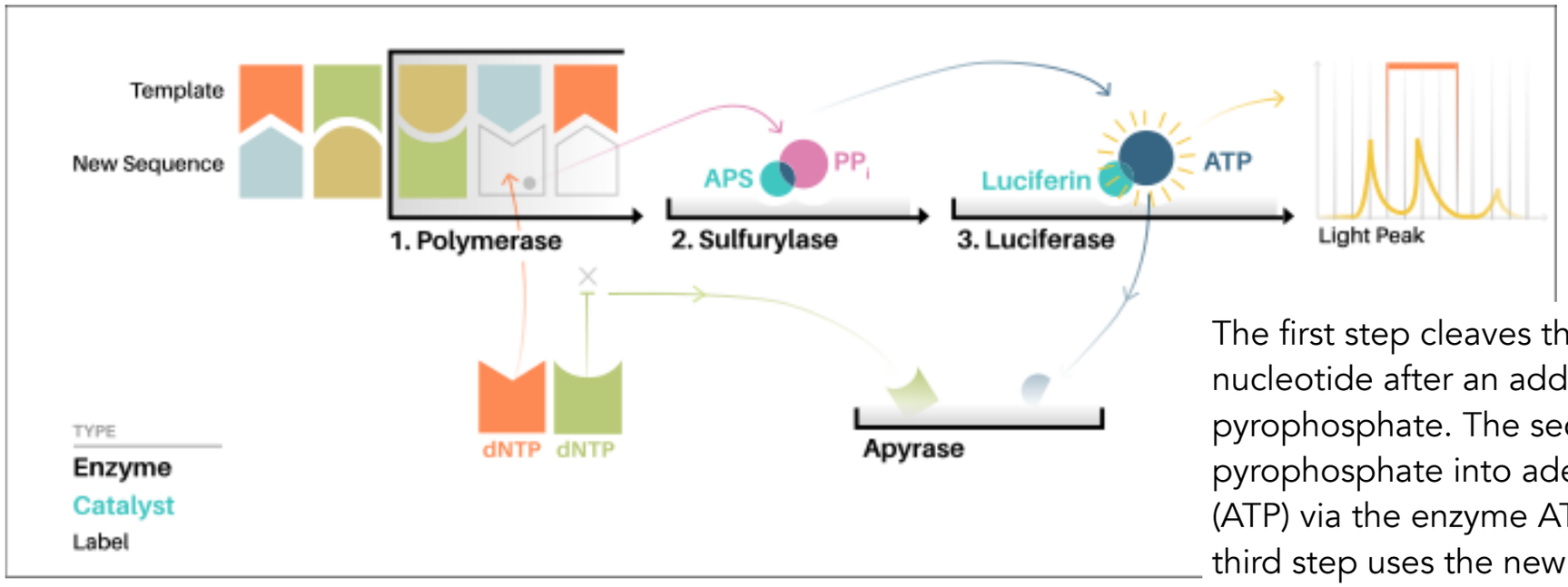
Pyrosequencing

GS Junior

Roche 454

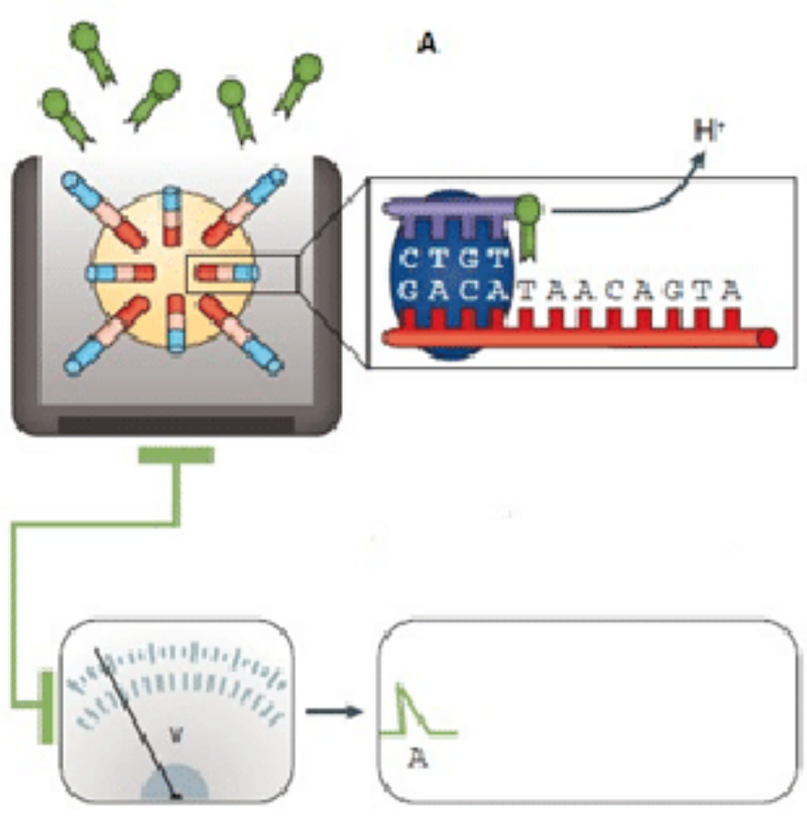
The **PyroMark** uses Pyrosequencing technology for real-time, sequence-based detection and quantification of sequence variants and epigenetic methylation. The PyroMark Q24 is highly suited for the analysis of CpG methylation, SNPs, insertion/deletions, STRs, variable gene copy number, as well as for microbial identification and resistance typing.

Pyrosequencing (pyrophosphate)



The first step cleaves the triphosphate nucleotide after an addition, releasing pyrophosphate. The second step converts pyrophosphate into adenosine triphosphate (ATP) via the enzyme ATP sulfurylase. The third step uses the newly synthesized ATP to catalyze the conversion of luciferin into oxyluciferin via the enzyme luciferase and this reaction generates a quanta of light that is captured from the picotiter plate by a charge- coupled camera.

Ion Torrent (semiconductor technology)

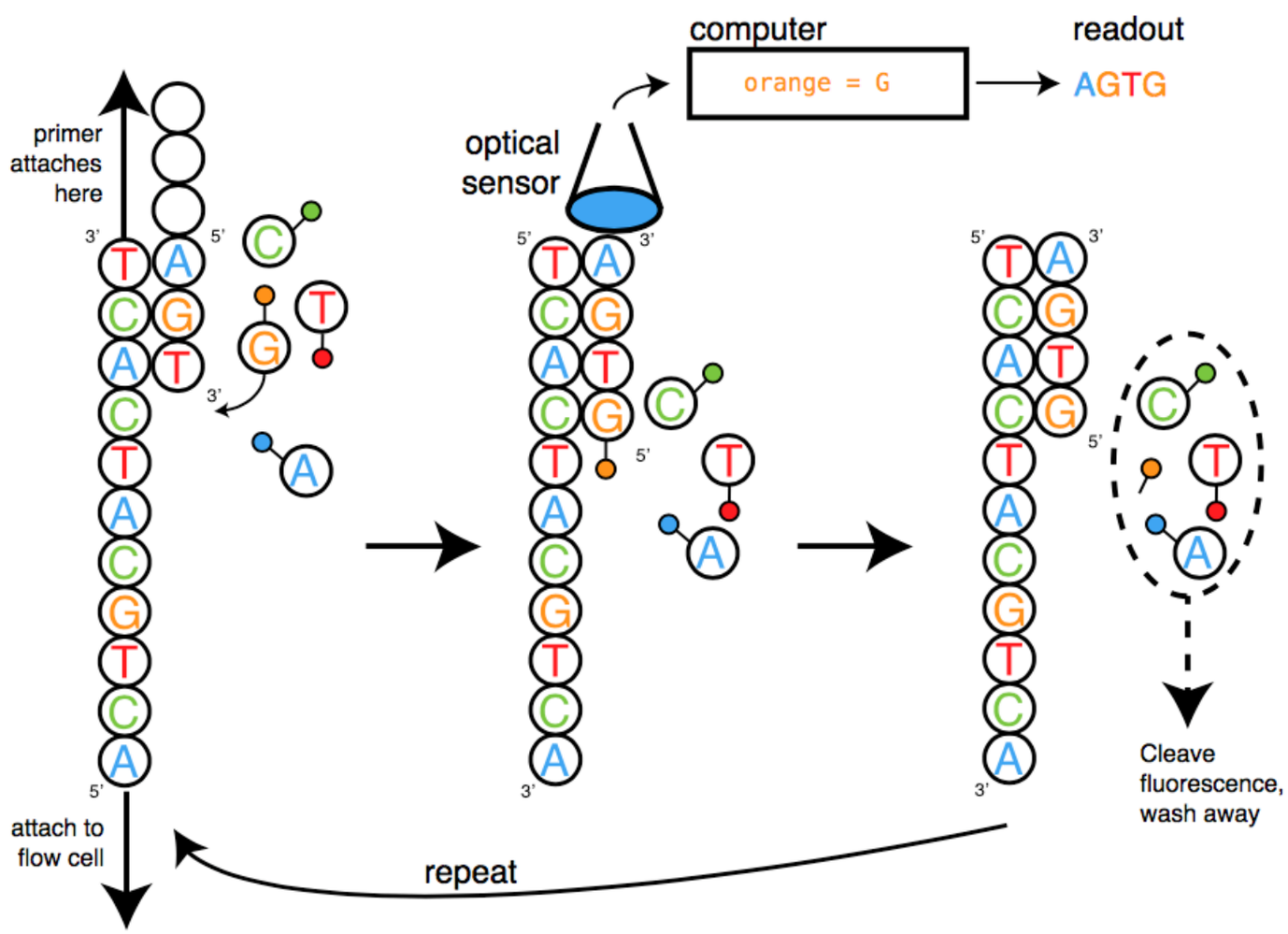


Illumina Sequencing by Synthesis (fluorescent)

MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series	NovaSeq Series
1.8-7.5 Gb	0.3-15 Gb	20-120 Gb	125-1500 Gb	900-1800 Gb	134-6000 Gb
8-25 million	1-25 million	130-400 million	2.5-5 billion	3-6 billion	Up to 20 billion
2 x 150 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp
50	384	96	12	16	48

http://www.illumina.com

Sequencing by Synthesis (fluorescent)



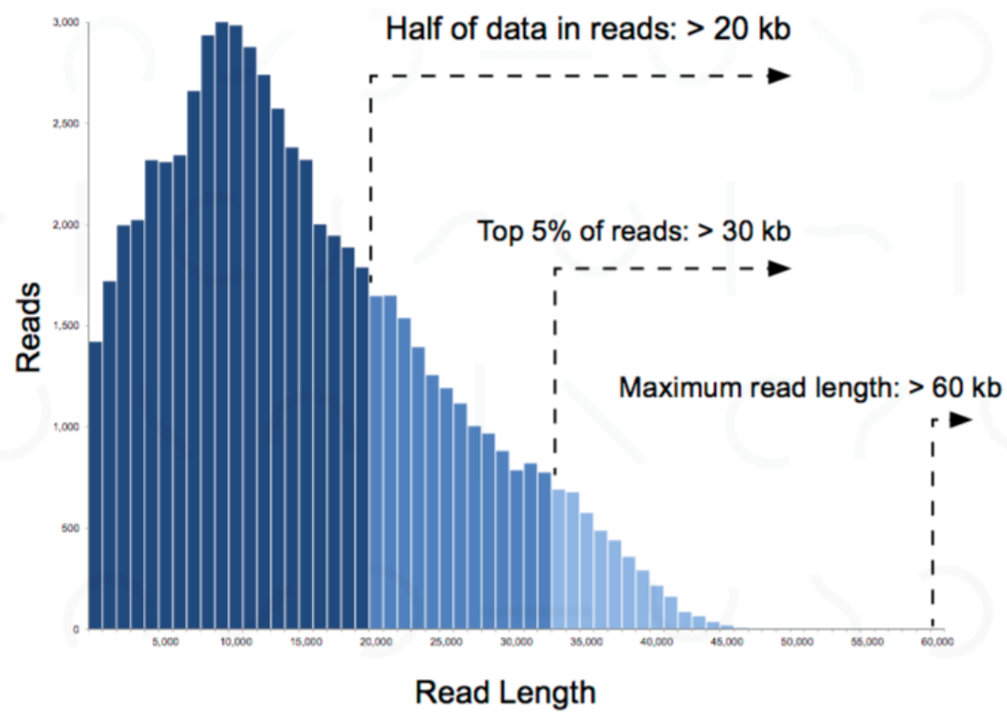
Sequencing by Synthesis. dNTP fluorescence is translated to a base call.

PacBio (fluorophore)

PacBio RS II

Long Read Lengths

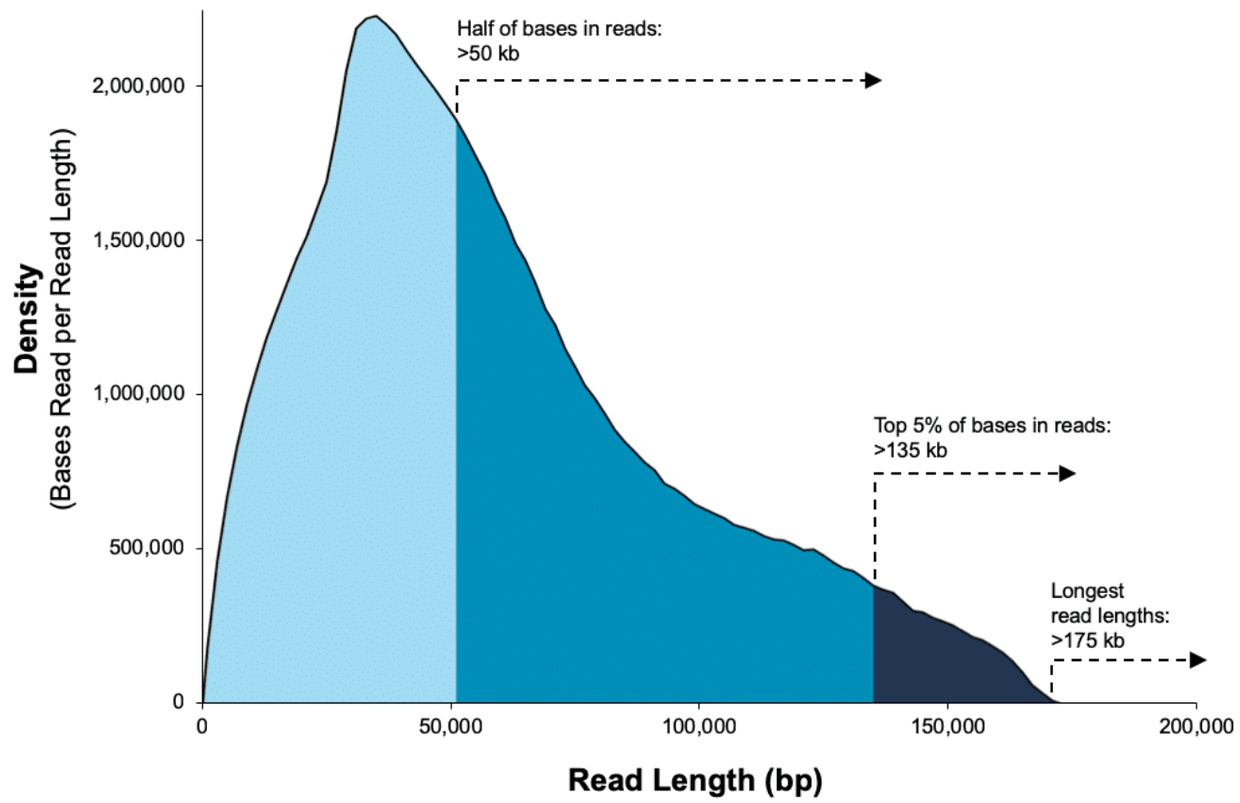
Read lengths > 20 kb
Data per SMRT Cell: 750 Mb – 1.25 Gb



Read-length data shown above is from a 20 kb size-selected human library run on a PacBio RS II (6-hour movie, P6-C4 chemistry). The PacBio RS II SMRT Cells generate ~55,000 reads. The Sequel System generates ~370,000 reads per SMRT Cell.

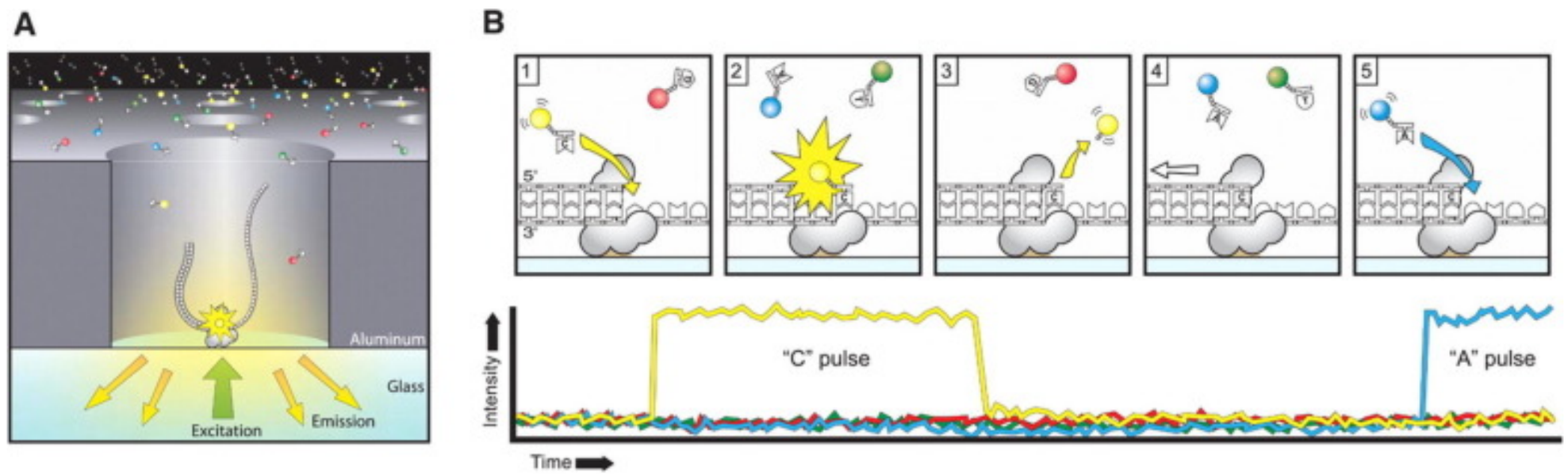
http://www.pacb.com

Sequel

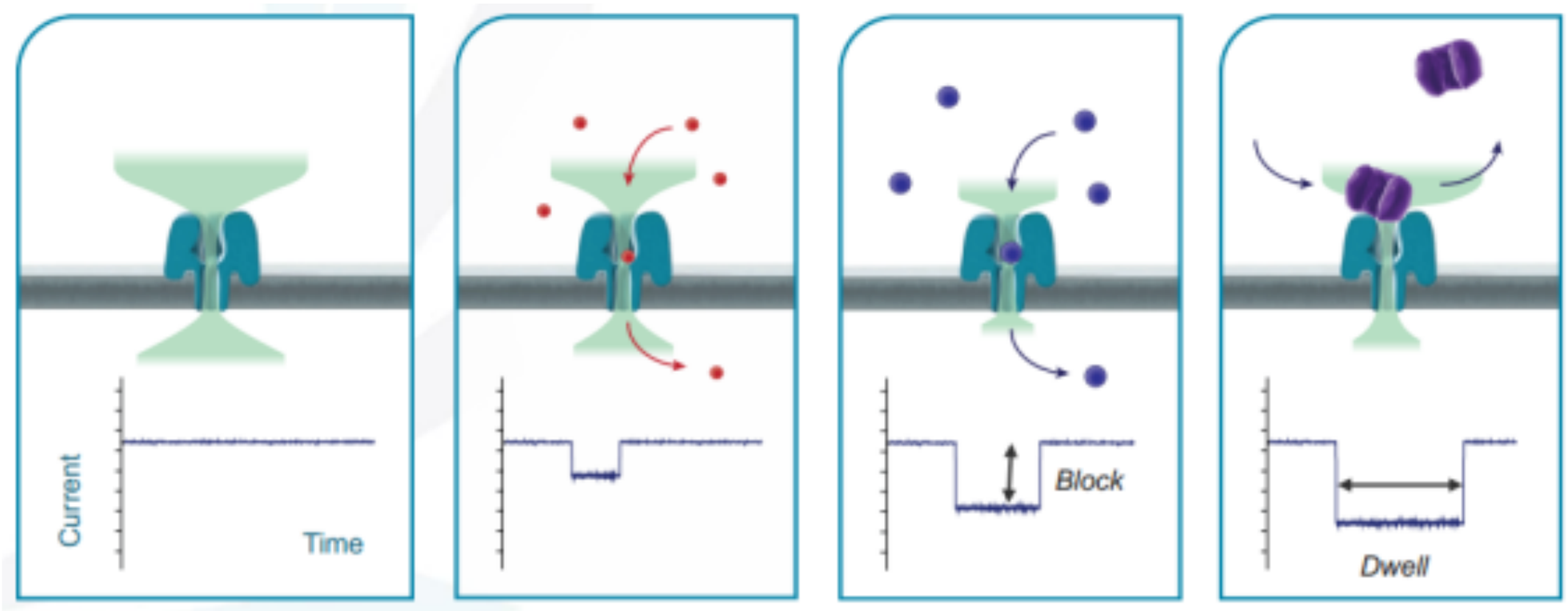


Data from a 35 kb size-selected *E. coli* library using the SMRTbell Express Template Prep Kit 2.0 on a Sequel II System (1.0 Chemistry, Sequel II System Software v7.0, 15-hour movie)*.

PacBio (fluorophore)

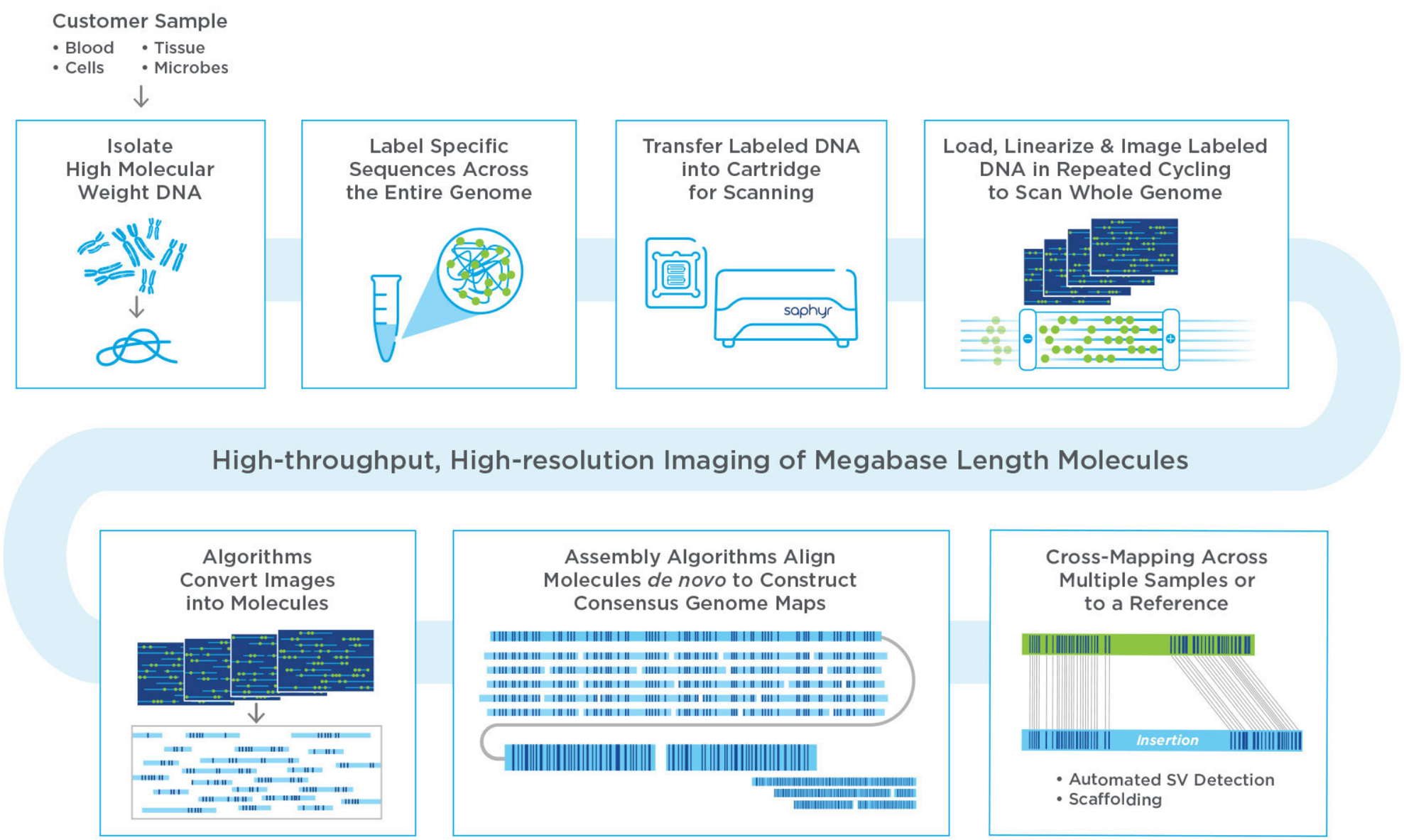



Oxford NANOPORE Technologies
GridION
PromethION
SmidgION
Flongle
MinION
GridION
PromethION
https://www.nanoporetech.com


Current
Time
Block
Dwell

Helicos
BioSciences Corporation


bionano
GENOMICS
saphyr


Customer Sample
• Blood
• Tissue
• Cells
• Microbes
Isolate High Molecular Weight DNA
Label Specific Sequences Across the Entire Genome
Transfer Labeled DNA into Cartridge for Scanning
saphyr
Load, Linearize & Image Labeled DNA in Repeated Cycling to Scan Whole Genome
High-throughput, High-resolution Imaging of Megabase Length Molecules
Algorithms Convert Images into Molecules
Assembly Algorithms Align Molecules de novo to Construct Consensus Genome Maps
Cross-Mapping Across Multiple Samples or to a Reference
Insertion
• Automated SV Detection
• Scaffolding

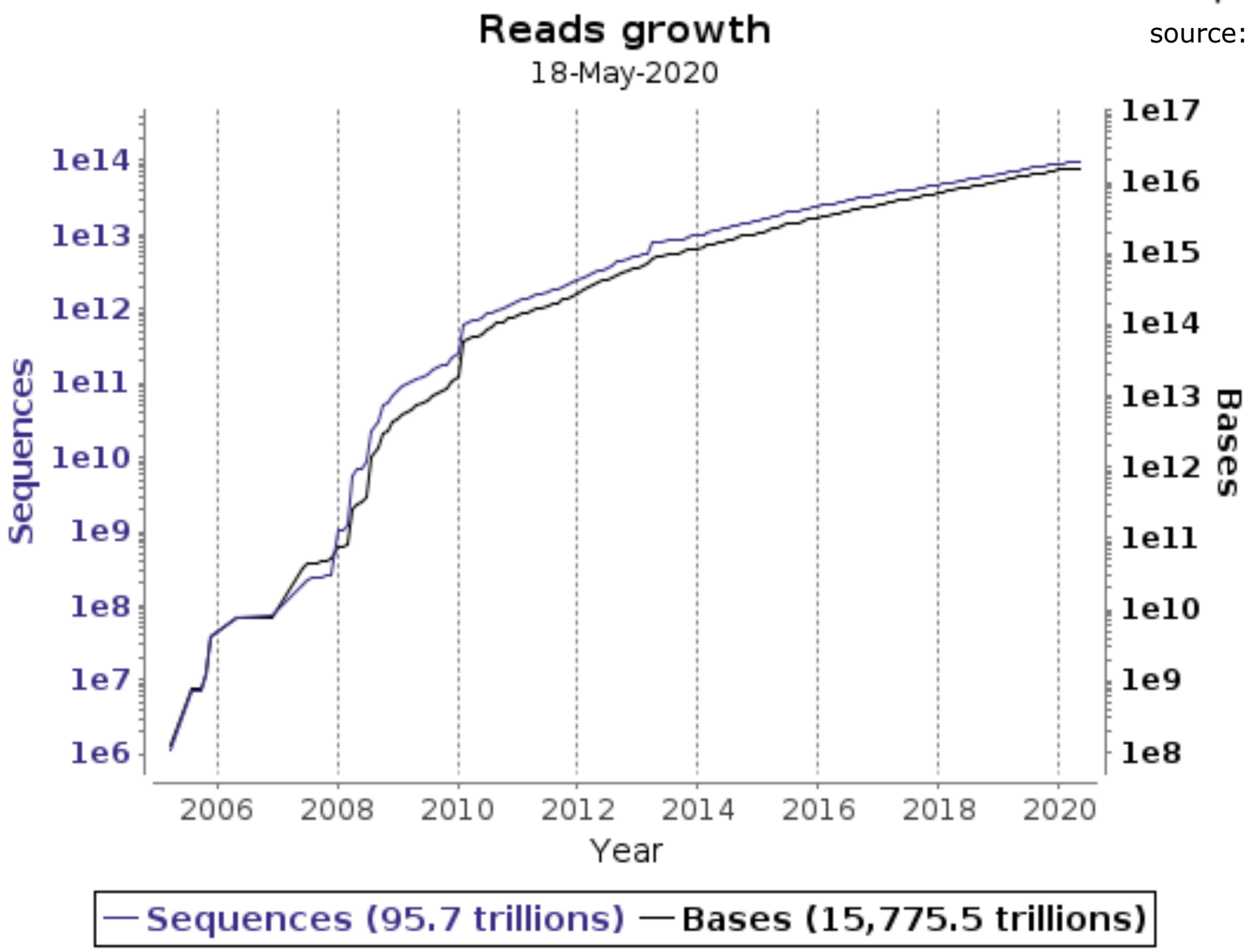

The Economist
Open-source medical devices
Dealing with nuclear waste
Burt Rutan, a maverick in flight
Technology Quarterly
December 1st 2012
The dream of the medical tricorder
the hand held devices inspiring medical add-ons for smartphones


Running
Stop Sequencing

MinION Mk1C

ENA
European Nucleotide Archive

source: http://www.ebi.ac.uk

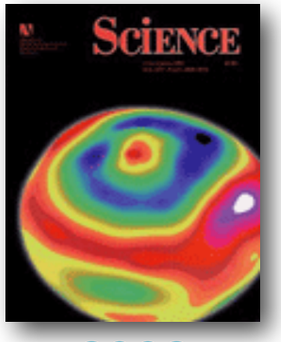


Bioinformatics ▶ NGS / NNGS

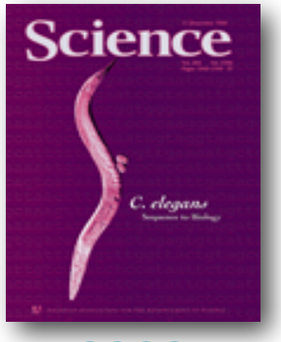
1995

1996

1997



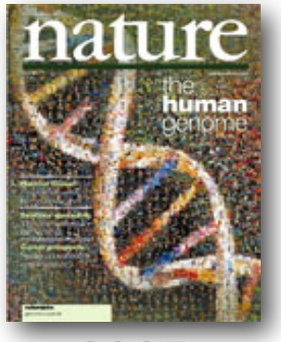
1998



2000

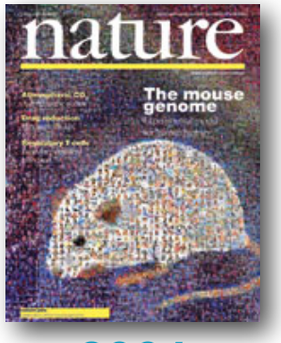
2000

2001



2002

2002



2002

2002

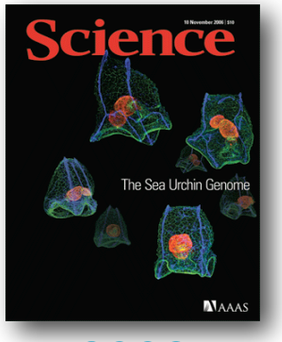
2004

2004

2005

2005

2006



2006

2007

2007

2008

2008

2009

2009

2010

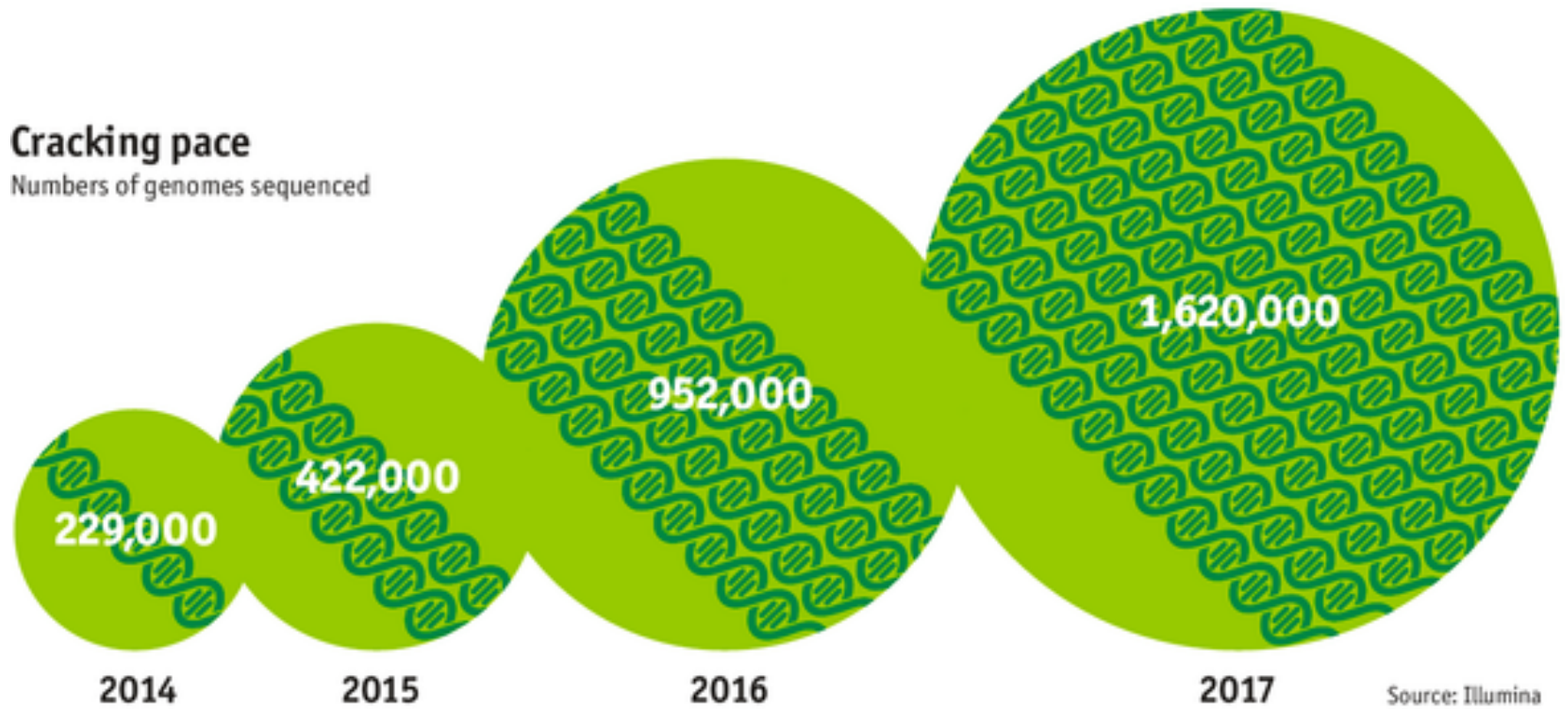
2011

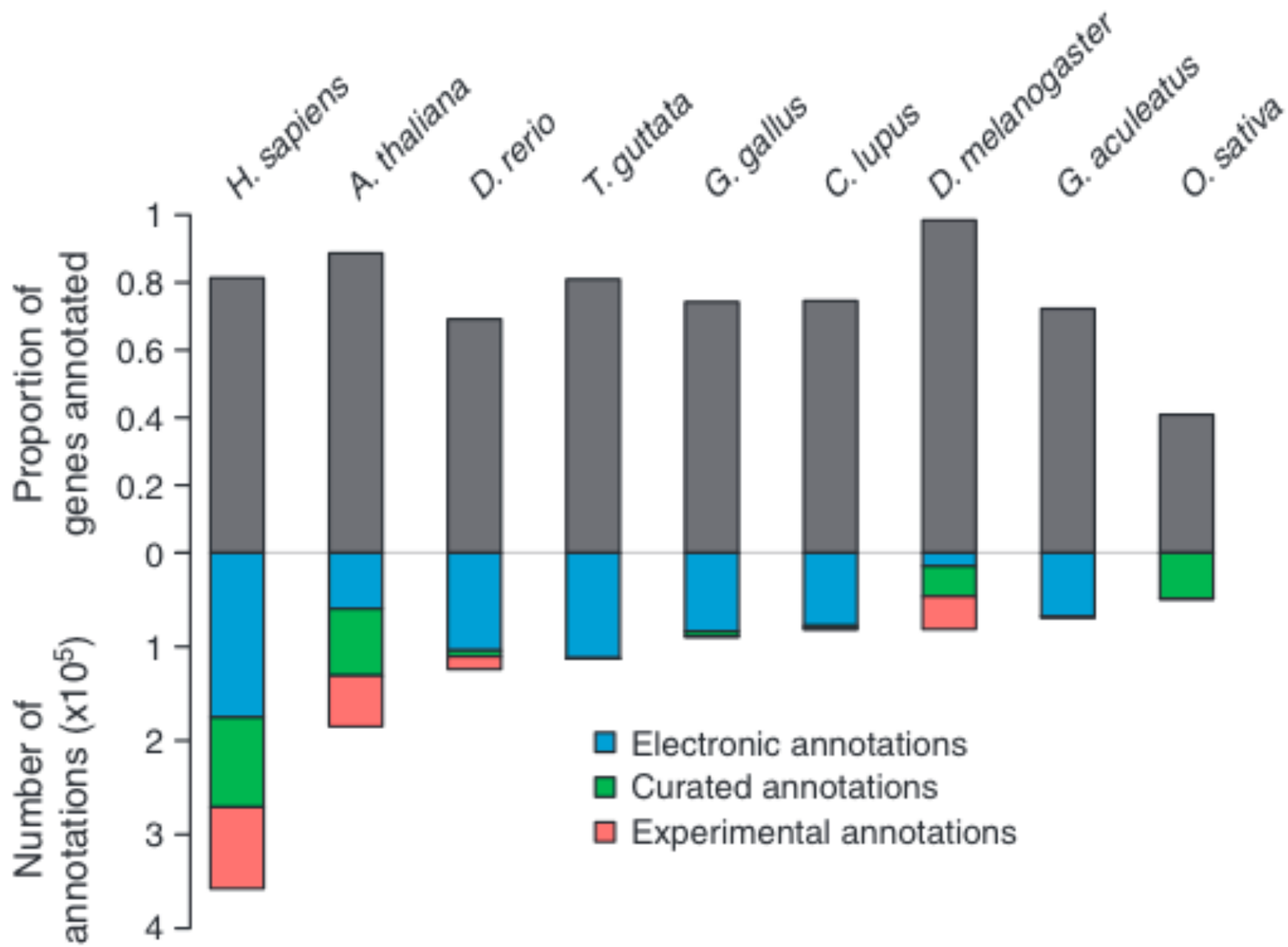
2011

2011

2012

...


Cracking pace
Numbers of genomes sequenced
229,000
422,000
952,000
1,620,000
2014
2015
2016
2017
Source: Illumina

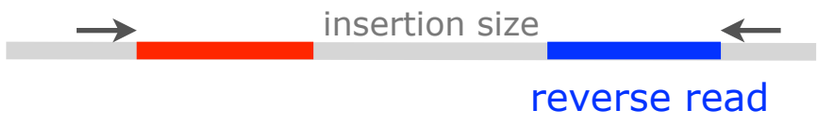


The proportion of annotated genes and their types of annotations for nine sequenced genomes (as of February 2013). Humans (*Homo sapiens*) and *Arabidopsis thaliana* have the highest number of annotations for animals and plants, respectively. They also have the most experimentally derived annotations. Most other species, except *Drosophila melanogaster*, are annotated mostly electronically.

Primmer et al. (2013) Mol Ecol

SEQUENCING DATA

Single Reads (SR)



Paired-End (PE) Reads

Overlapping Paired-End (PE) Reads

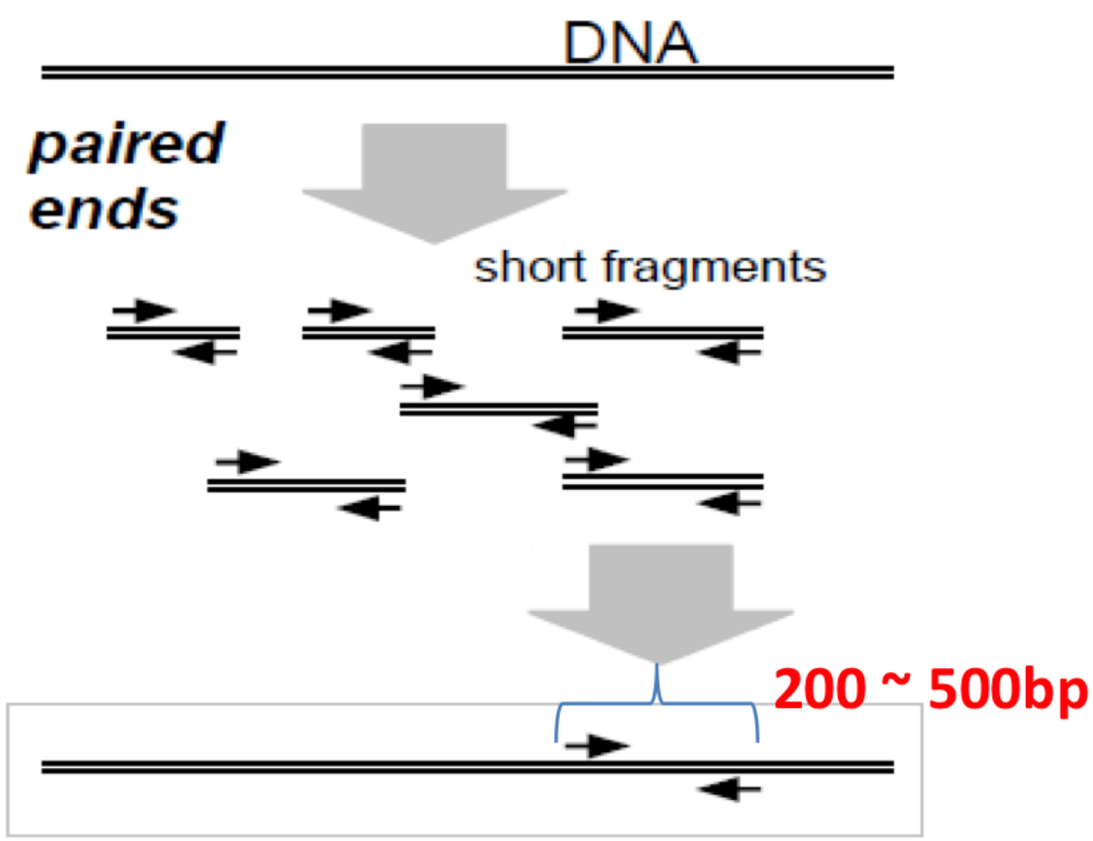
Single Reads (SR) with Index

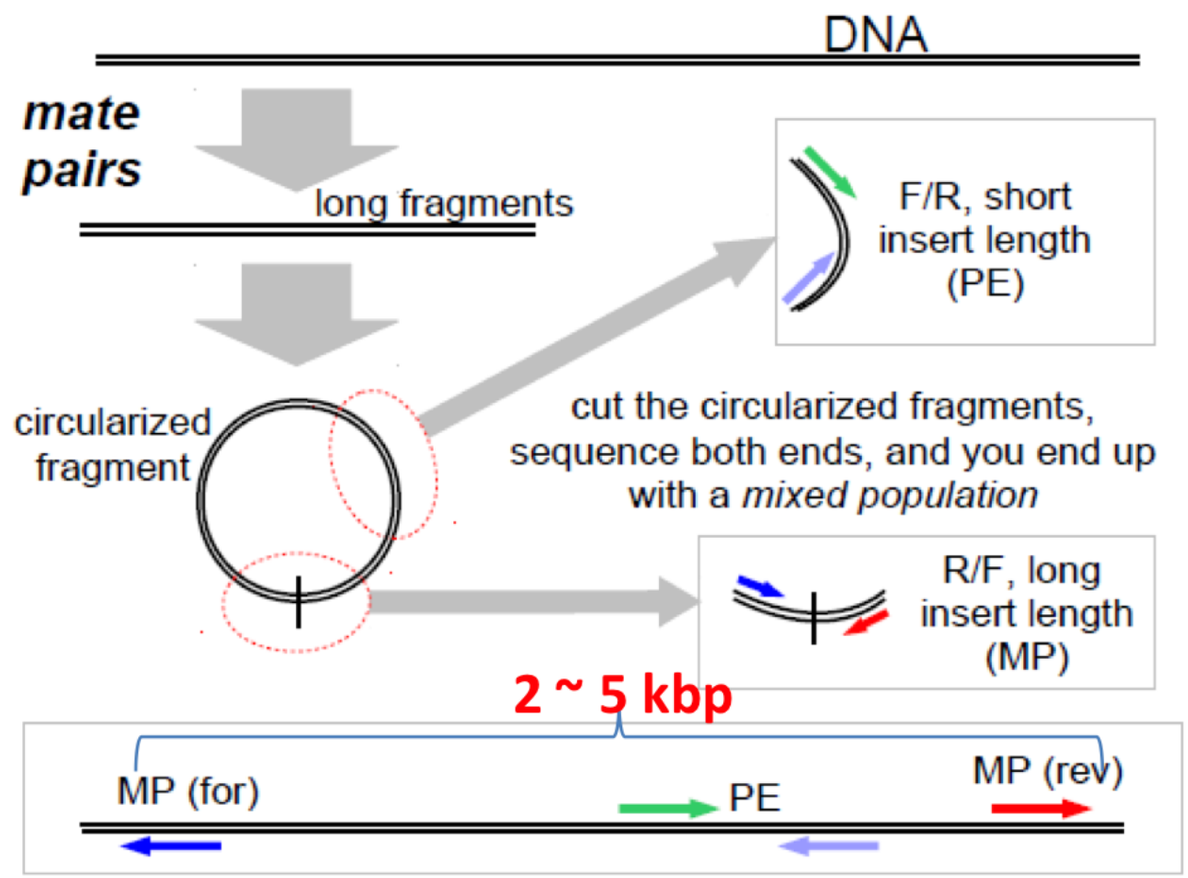
Paired-End (PE) Reads with Index

Paired-End (PE) Reads with Dual Indexing

Extended Single Reads (SR) with Index

paired-end (PE)	mate-pair (MP)





DATA FORMAT[S]

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

Fasta (**>**) Sequence Data Format

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 sequences)
sequences.afa (aligned sequences)

Fastq (@) Sequence Data Format

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 - BCCFFFFDFHHHFJJJJJJJJIIJJJJIJIIJJGIHHHHHHFFFDDDEDDDC
    @HWI-ST486:166:C06K9ACXX:7:1101:2519:1936 1:N:0:ACAGTG
    NTGCACACGAATTCGCCGTTGTGGCAGCTCGAGTACCTGTGGTTCGCCGAG
    +
    #1=DDDFFHHHFFIGIIJJJJIIJIIJ;?*?.?/9*88BBBF.;7@@E###
```

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)

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
j. 0 when no control bits are on
k. index sequence

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)*

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

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.........................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 ..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")
```

EXPECTED ERROR RATE

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]

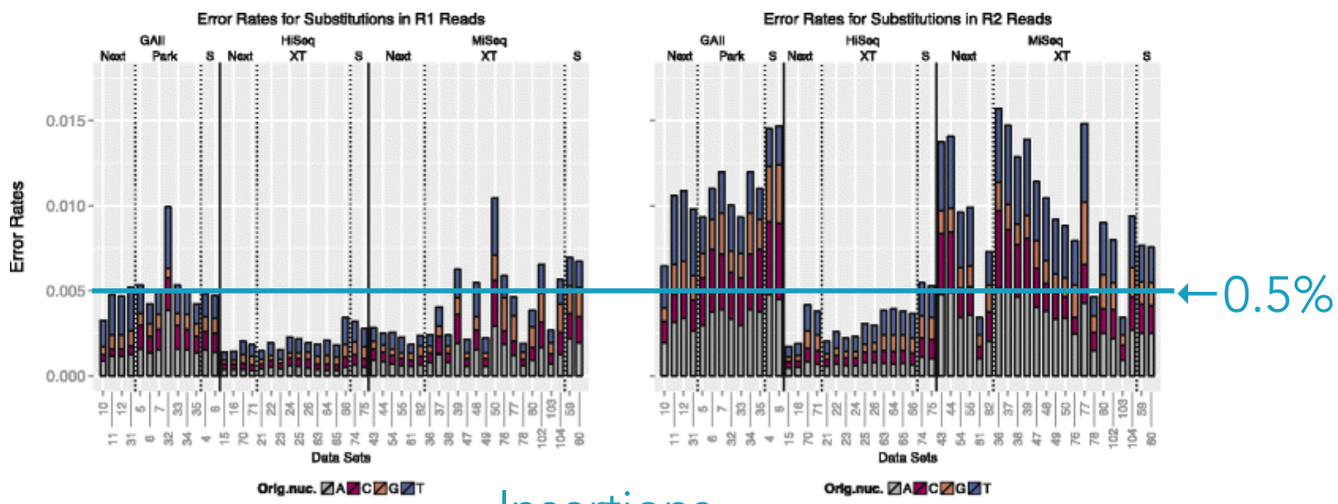
Abstract

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.

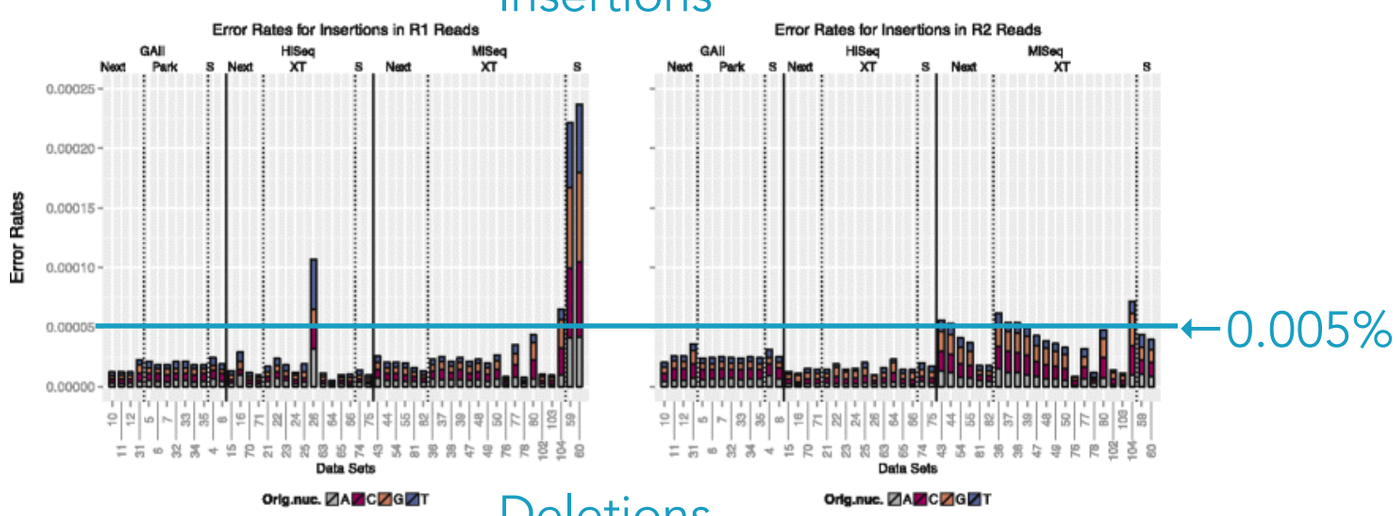
Results: Weconductedasystematicinvestigationoferrorsandbiasesin Illuminadatabasedonthelargestcollection of in vitro metagenomic data sets to date. We evaluated the Genome Analyzer II, HiSeq and MiSeq and tested state-of-the-art low input library preparation methods. Analysing in vitro metagenomic sequencing 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.

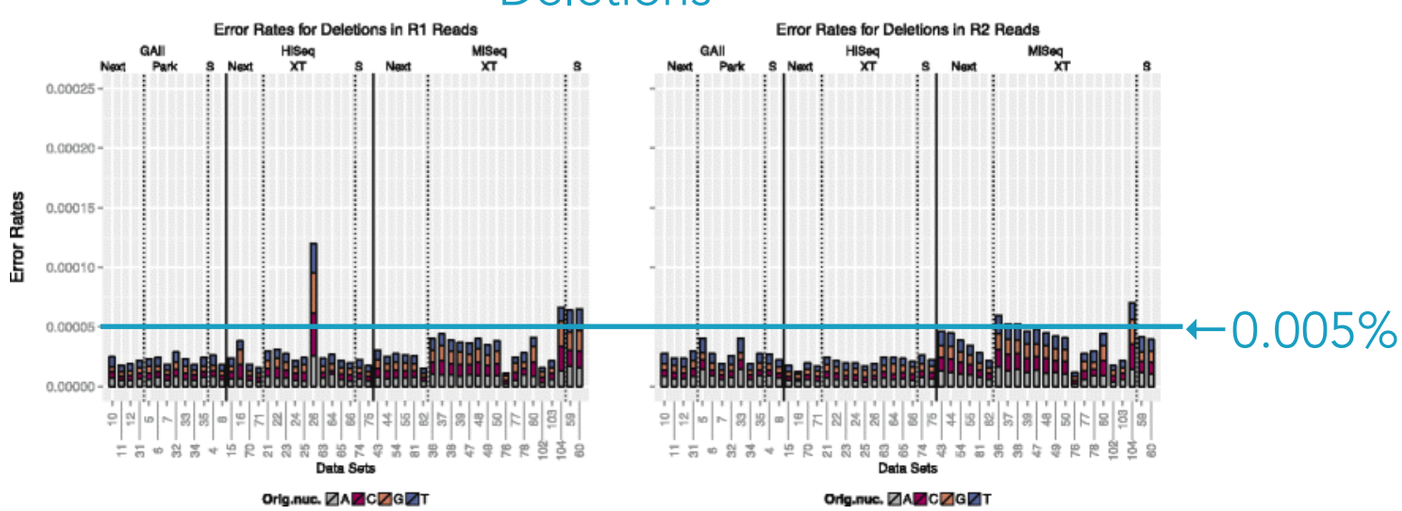
Substitutions



Insertions



Deletions

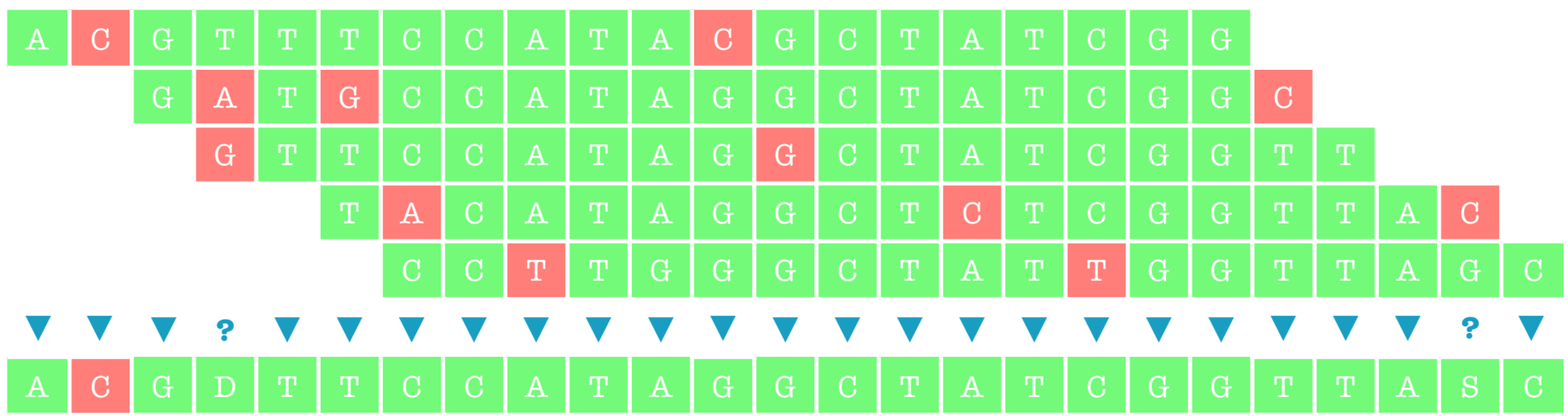


Illumina

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 Correction



Read quality

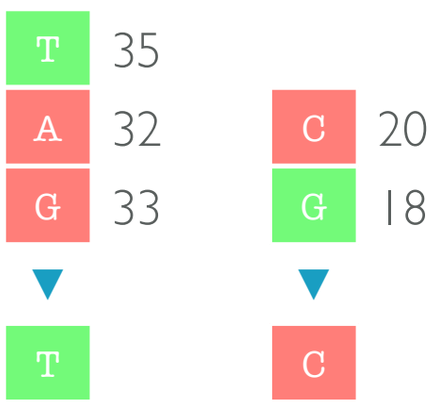
Number of reads (coverage)

Error Correction

Read quality

Number of reads (coverage)

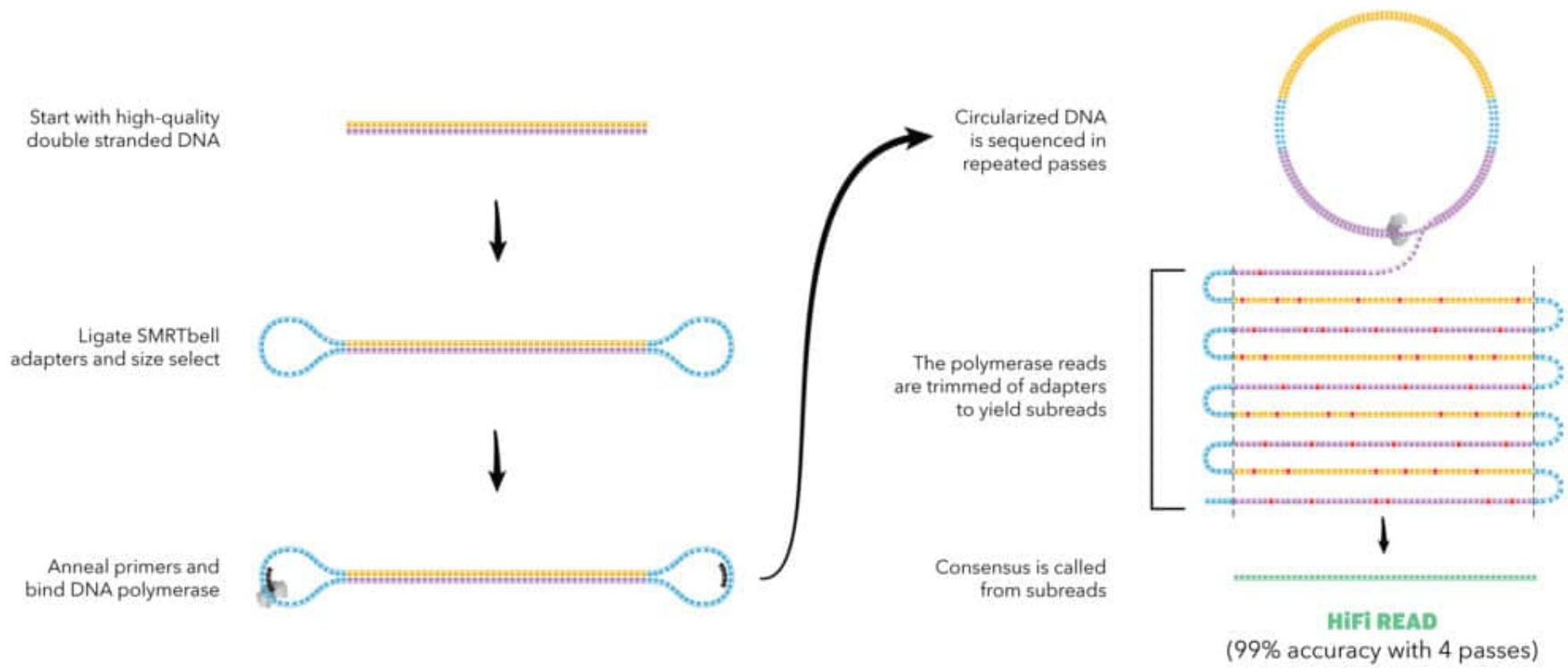
Phred score



Error Rate

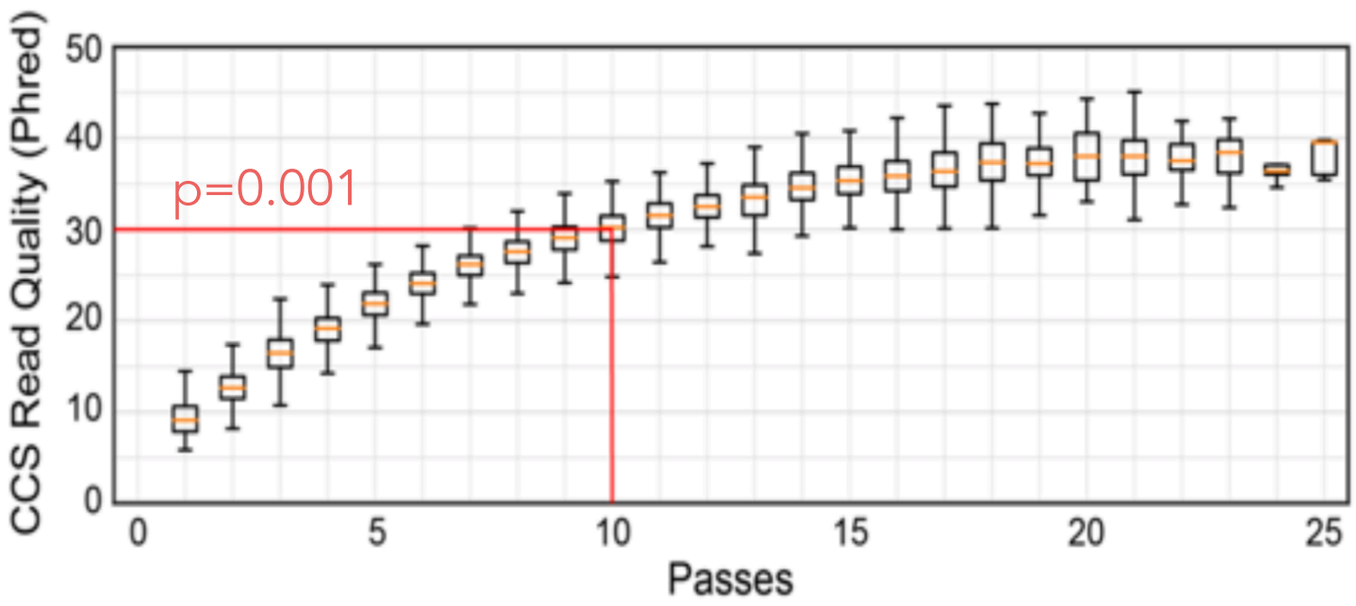
BAM ➜ FASTQ
BAM ➜ CCS.FASTX

Circular Consensus Sequences (CCS)



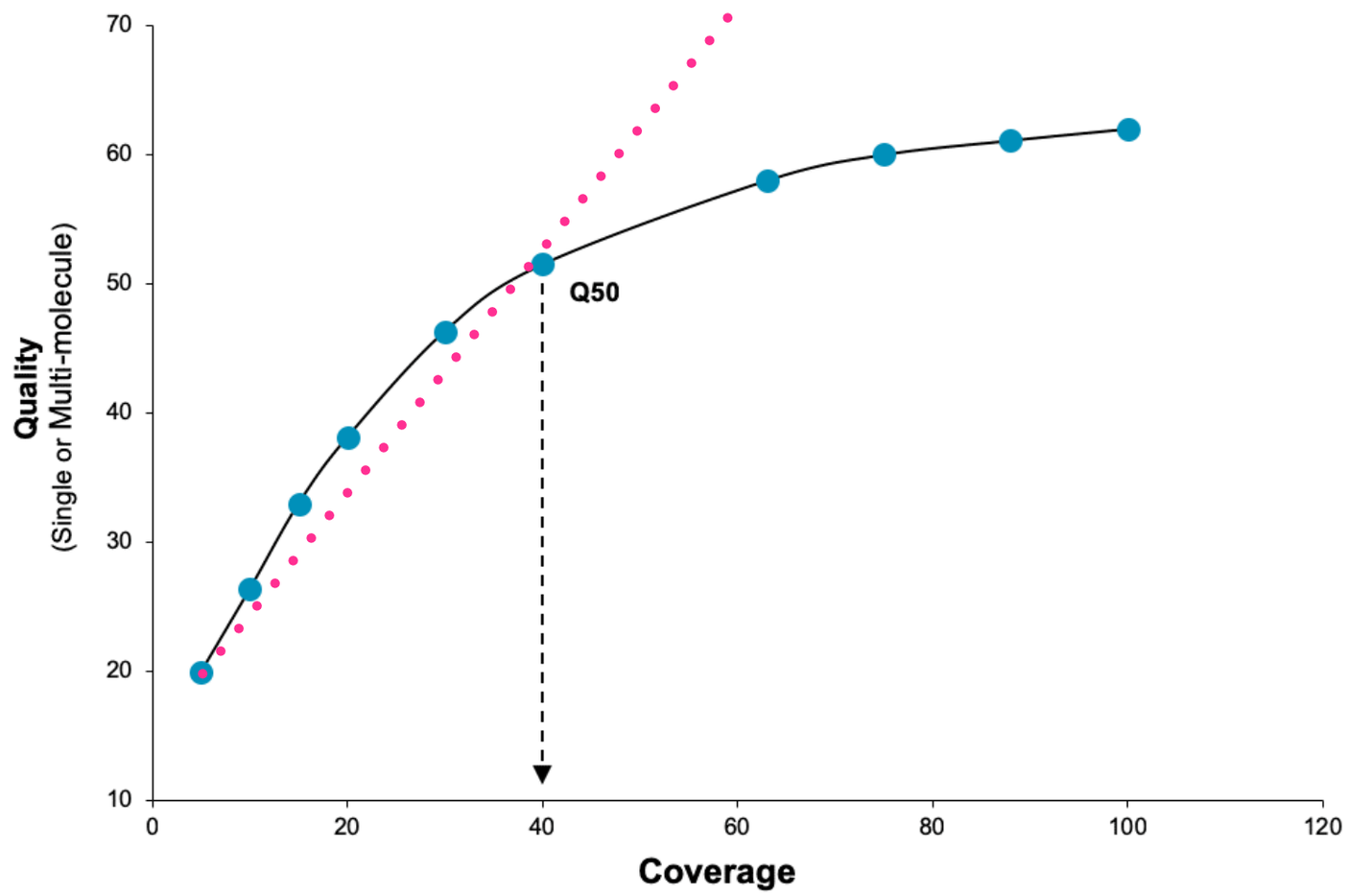
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

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

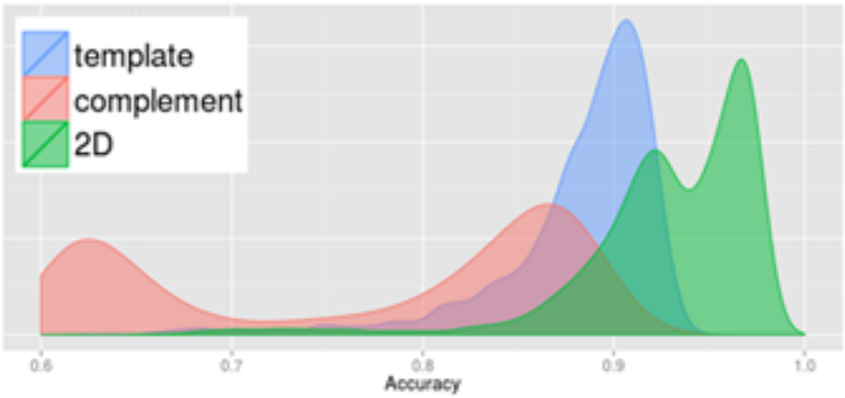


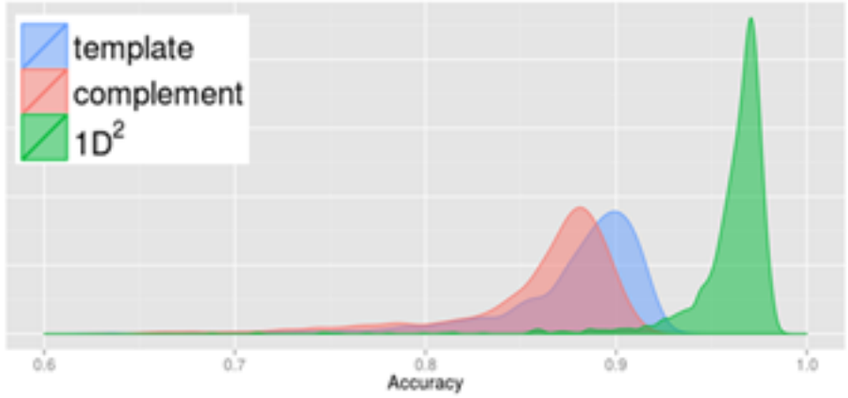
A SMRTbell library tightly distributed at 15 kb was chosen for circular consensus sequencing based on estimates of 150 kb polymerase read length and a requirement of 10 passes to achieve Q30 read accuracy. CCS reads with a predicted accuracy of at least Q20 (99%) were retained. The total CCS read yield 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%).

Why does it not improve anymore?



Q50 ➜ p=1e-05 ➜ 40 x coverage





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

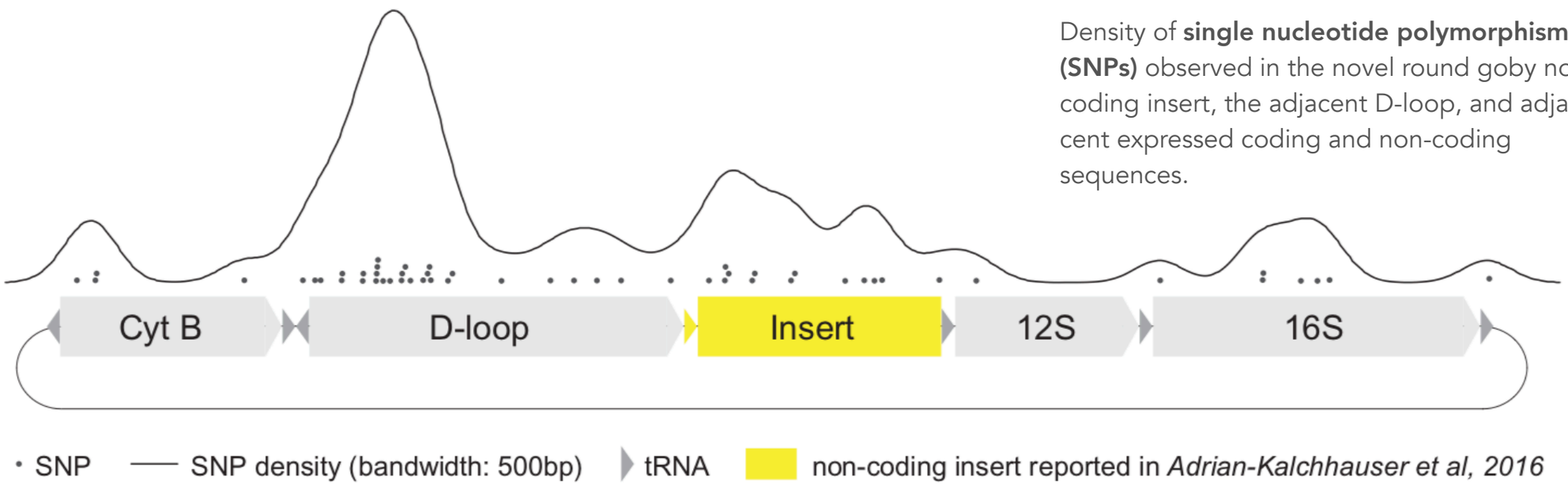
OPEN ACCESS Check for updates

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

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.



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

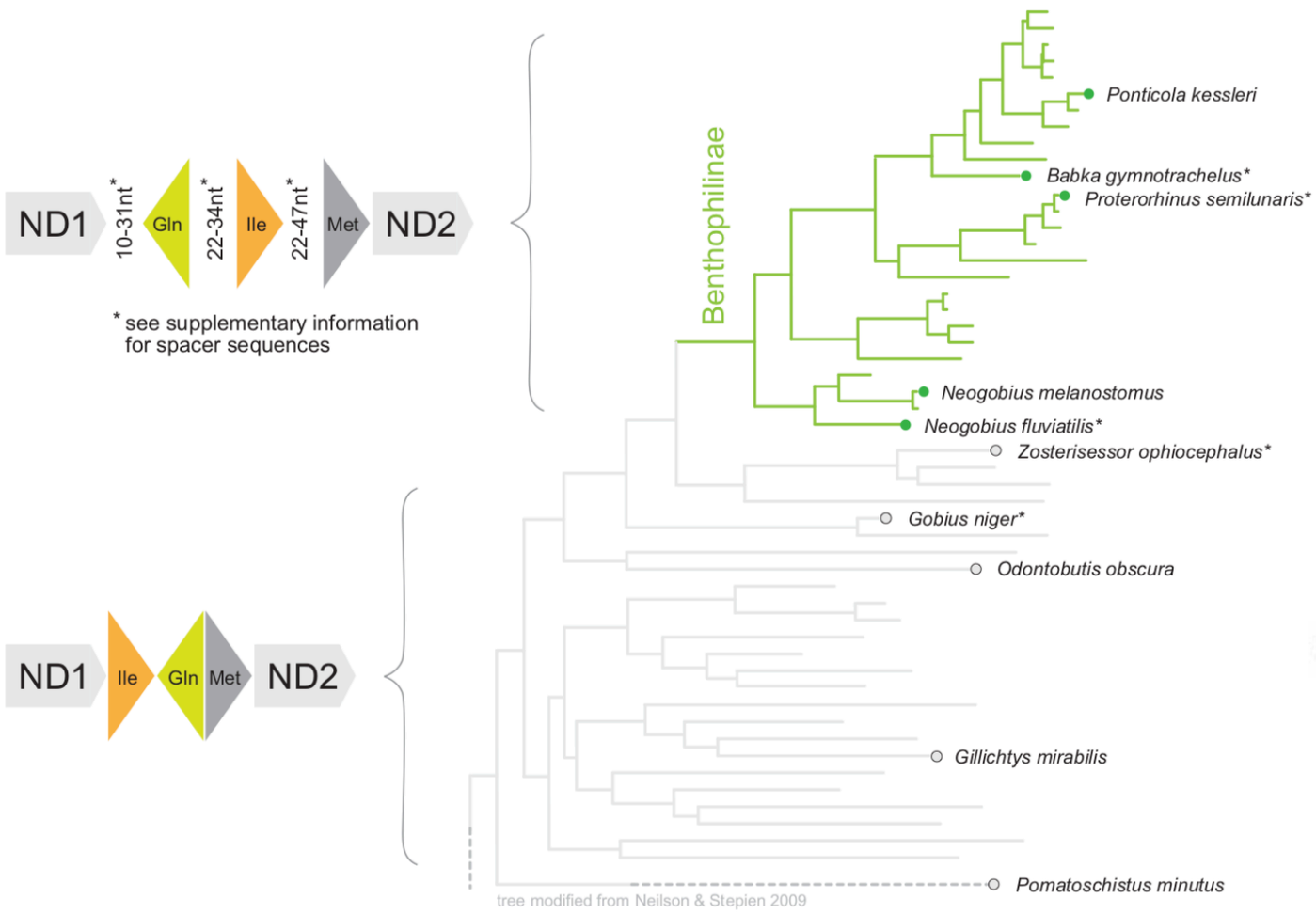
ARTICLE

OPEN ACCESS Check for updates

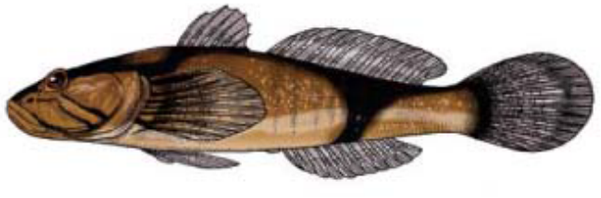
Long-read sequencing of benthophilinae mitochondrial genomes reveals the origins of round goby mitogenome re-arrangements

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

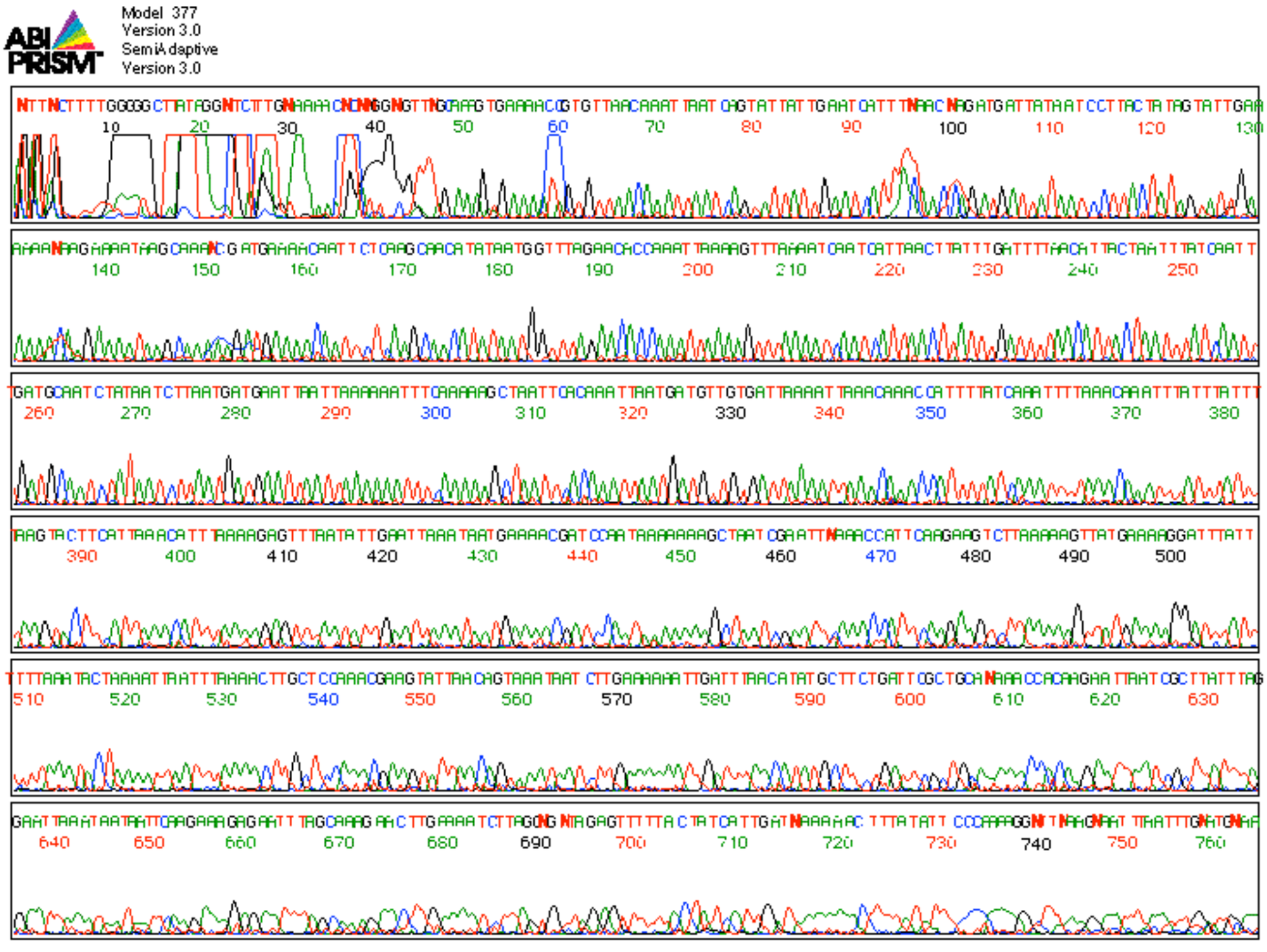
[a]Biozentrum, Department Growth & Development, University of Basel, Basel, Switzerland; [b]Genetic Diversity Centre Zurich, ETH Zurich, Zurich, Switzerland;[c]Program Man-Society-Environment, Department of Environmental Sciences, University of Basel, Basel, Switzerland

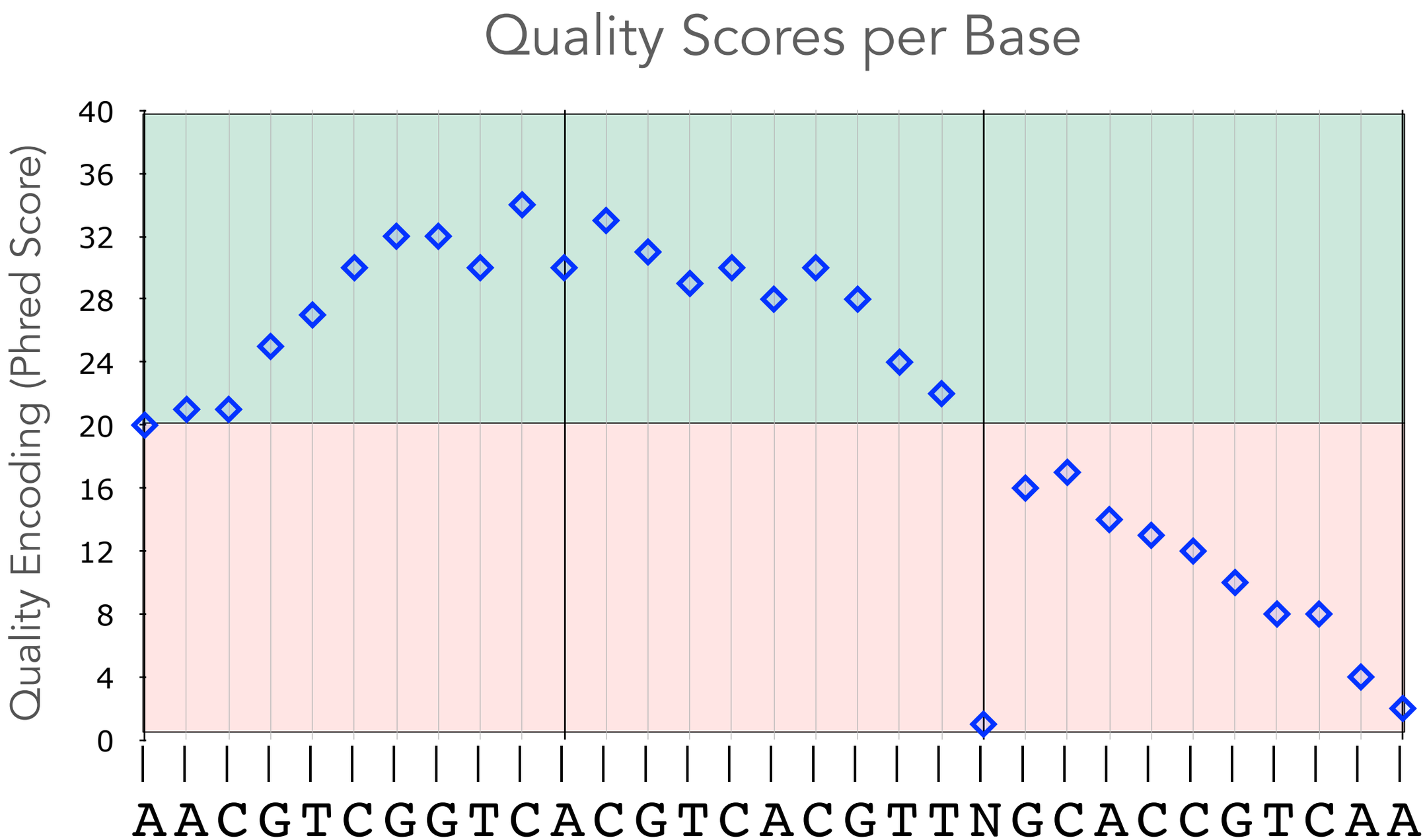


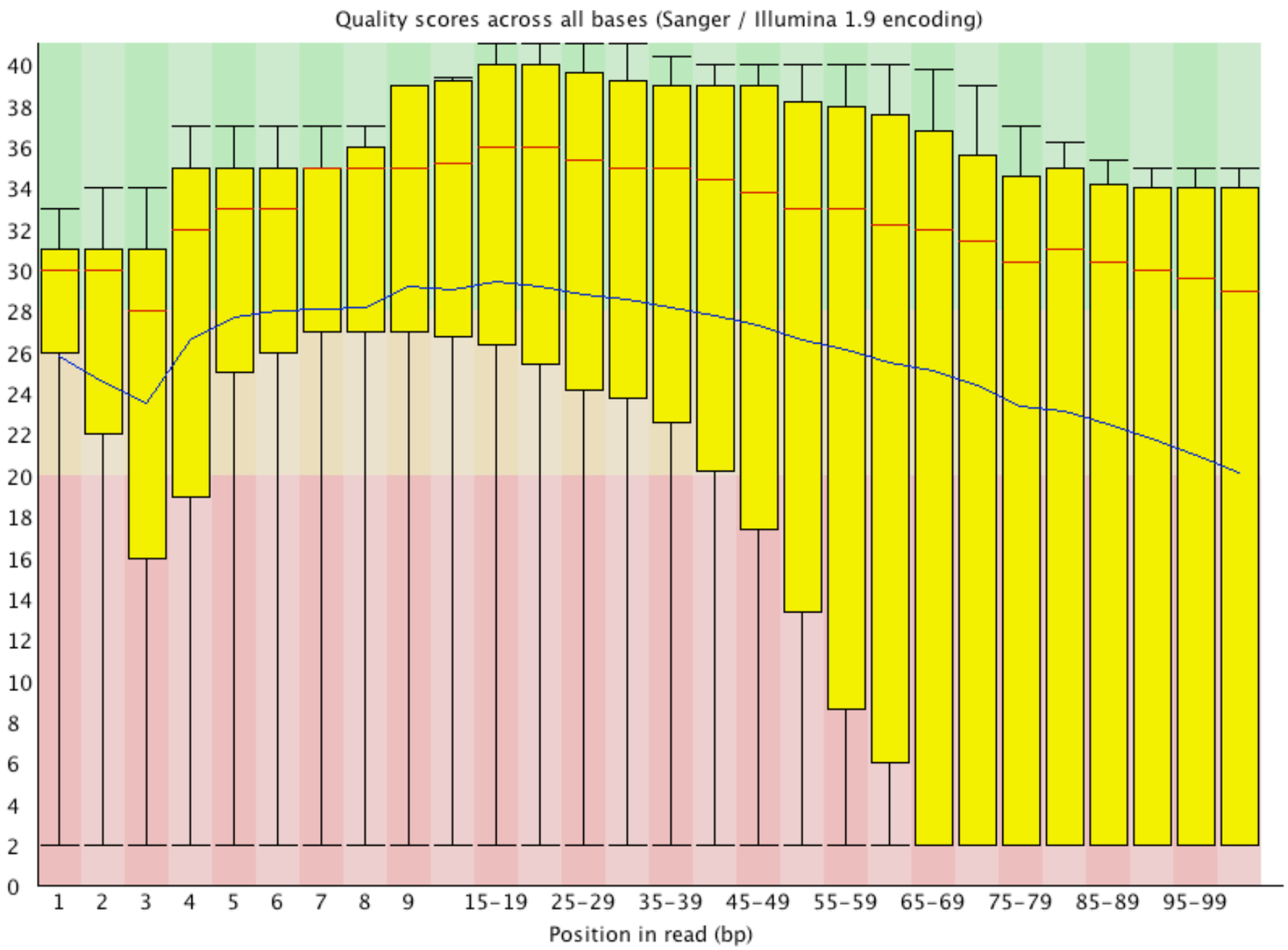
Origin of the re-arranged **tRNA cluster** Gln, Ile, Met. Most Gobiidae carry the arrangement Ile, Gln, Met without spacers. Benthophilinae (subfamily of gobies) however carry the arrangement Gln, Ile, Met, and feature variable length spacers between the genes.

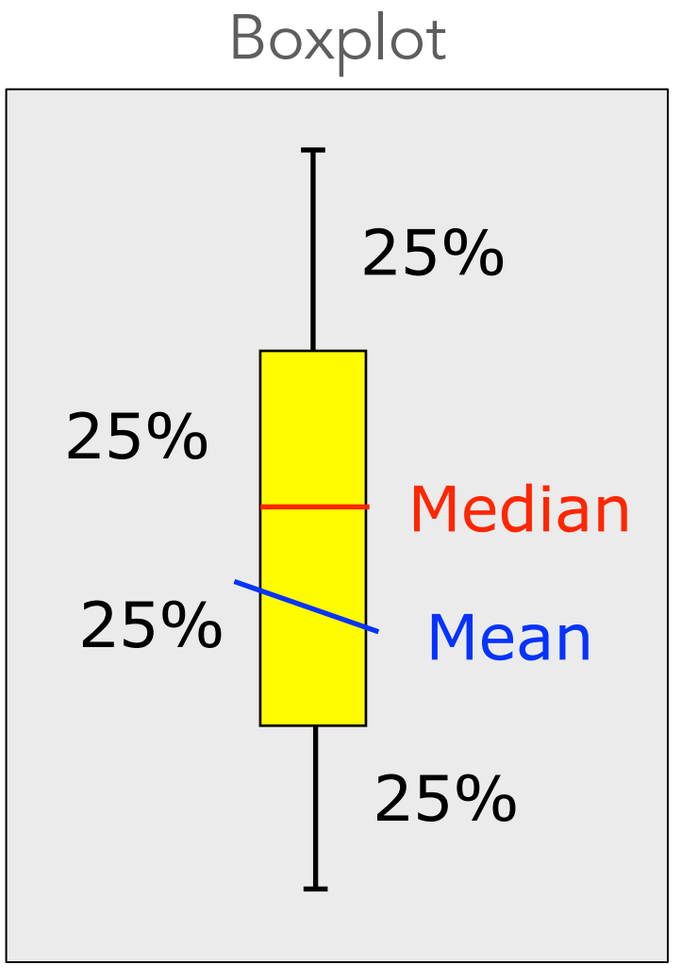


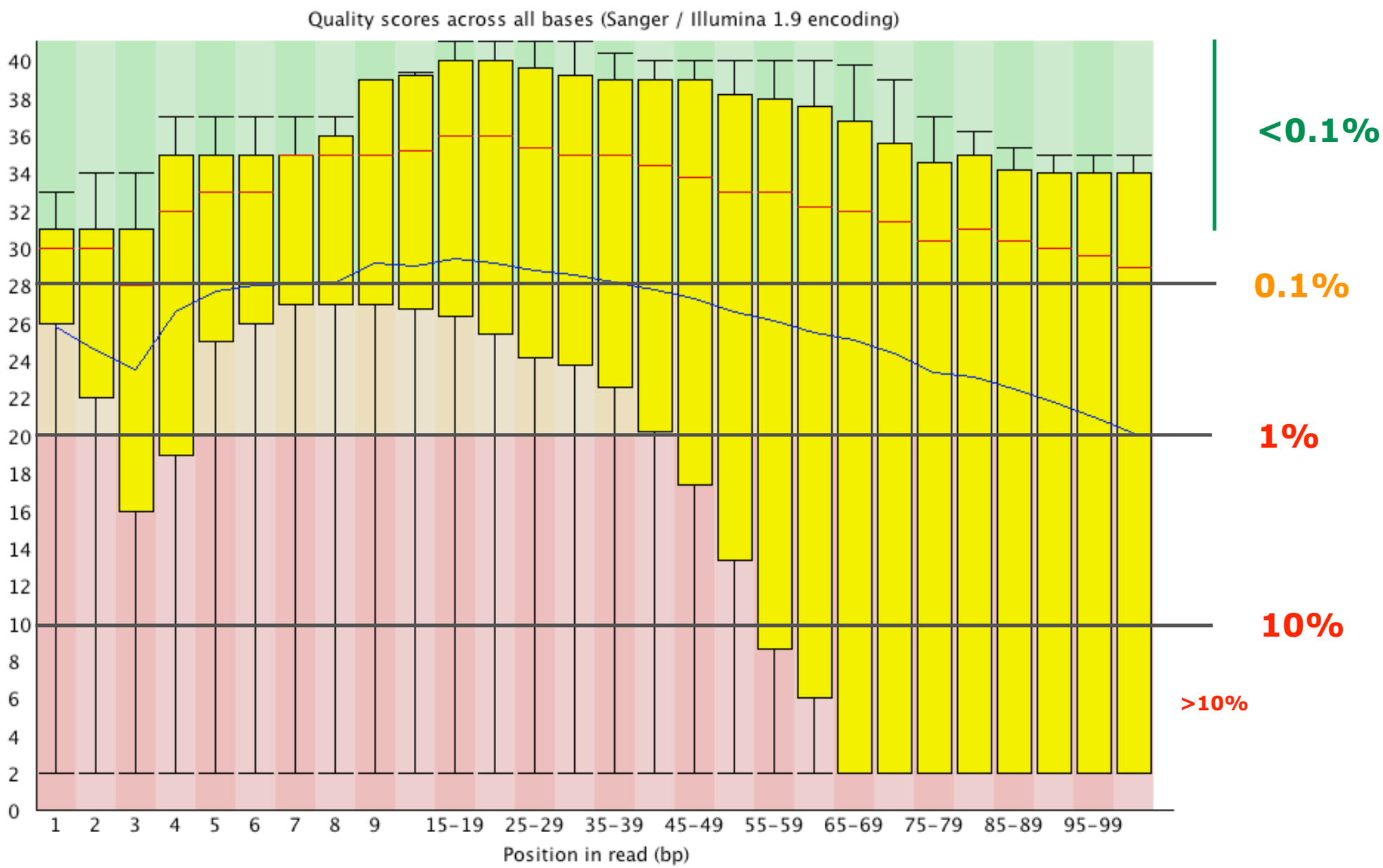
FASTQ QUALITY CONTROL


ABI PRISM
Model 377
Version 3.0
SemiAdaptive
Version 3.0


Quality Scores per Base
Quality Encoding (Phred Score)
40
36
32
28
24
20
16
12
8
4
0
AACGTCGGTCACGTCACGTTNGCACCGTCAA


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%


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)
<0.1%
0.1%
1%
10%
>10%
Error Rate

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/)

```
ssh -Y <student?>@gdcsrv2.ethz.ch
fastqc -v
fastqc
```

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

FastQC

File Help

1_S1_L001_R1_001.fastq.gz | 2_S2_L001_R2_001.fastq.gz

- 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

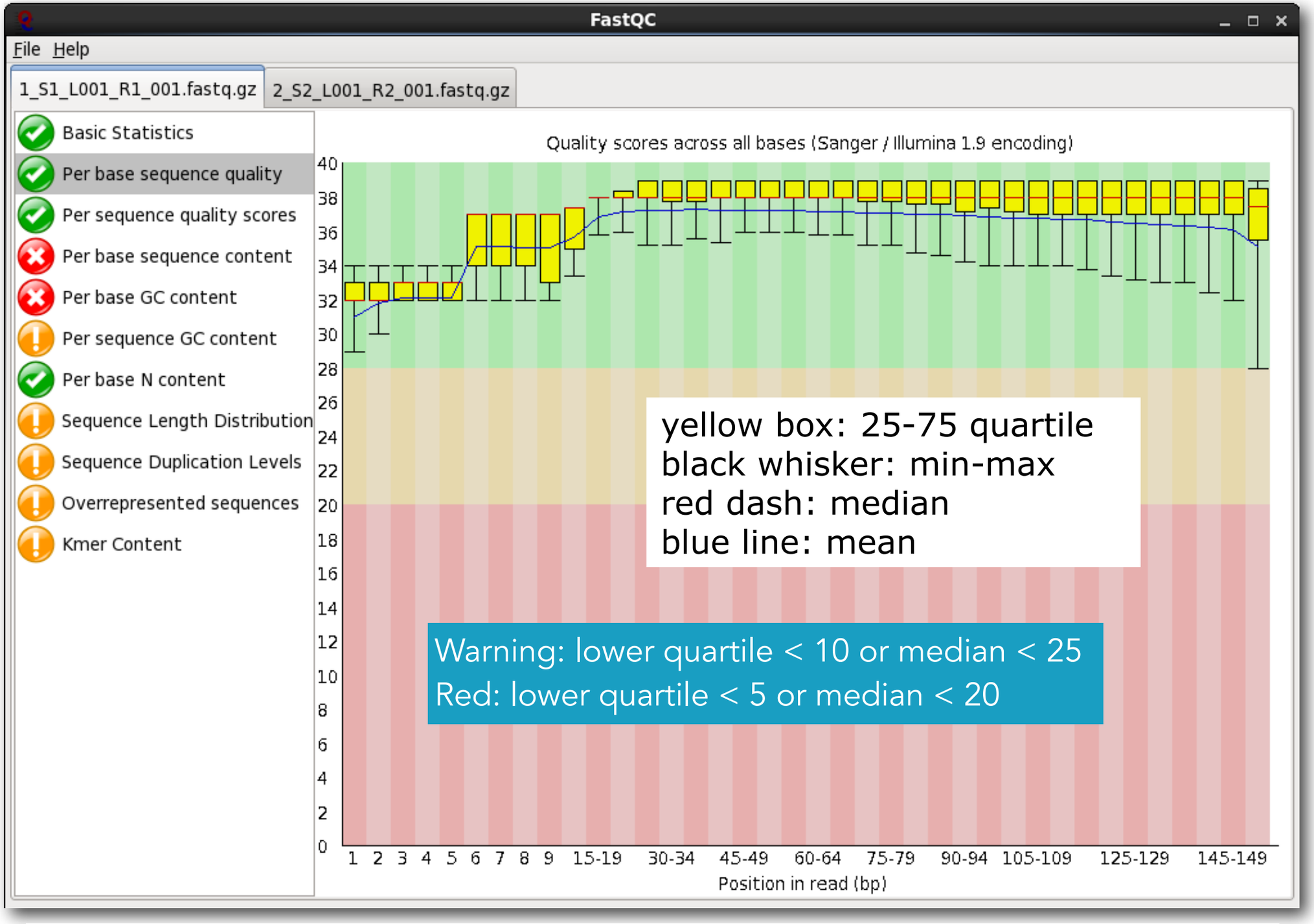
Basic sequence stats

Measure	Value
Filename	1_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	841152
Filtered Sequences	0
Sequence length	35-151
%GC	49

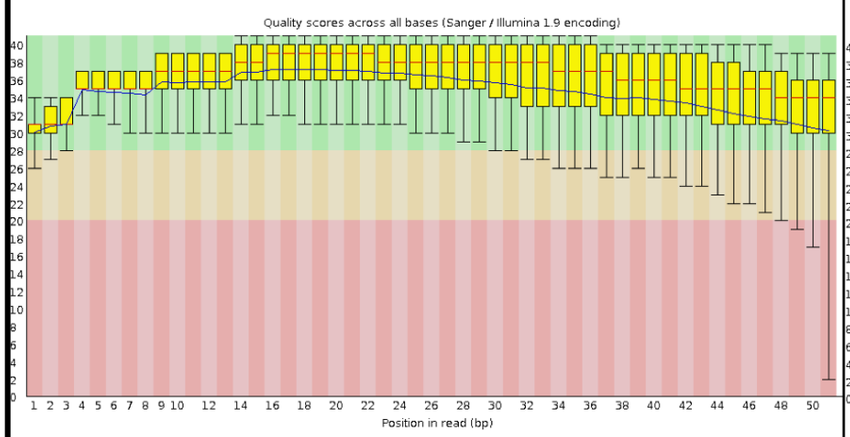
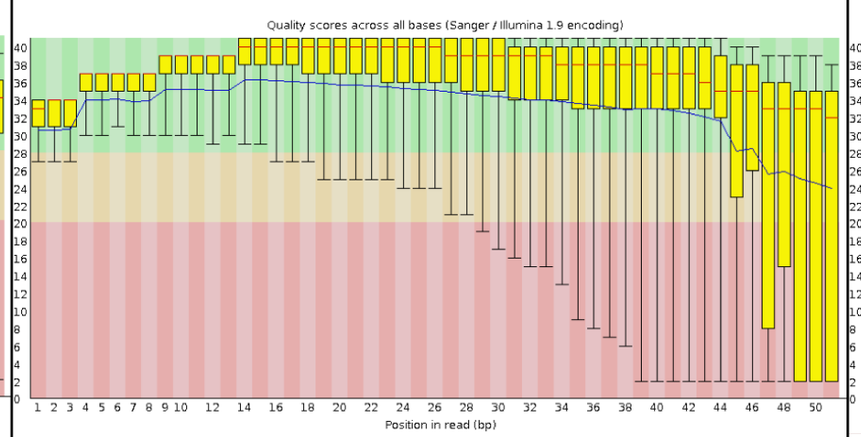
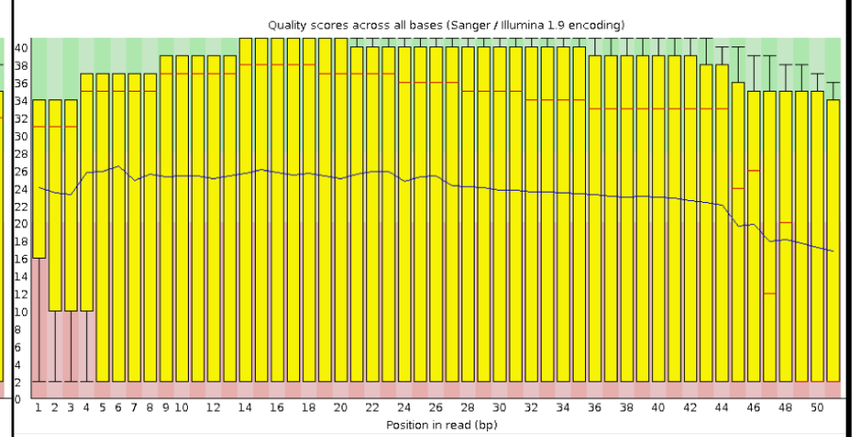
Basic Statisitcs never raises a warning nor an error.

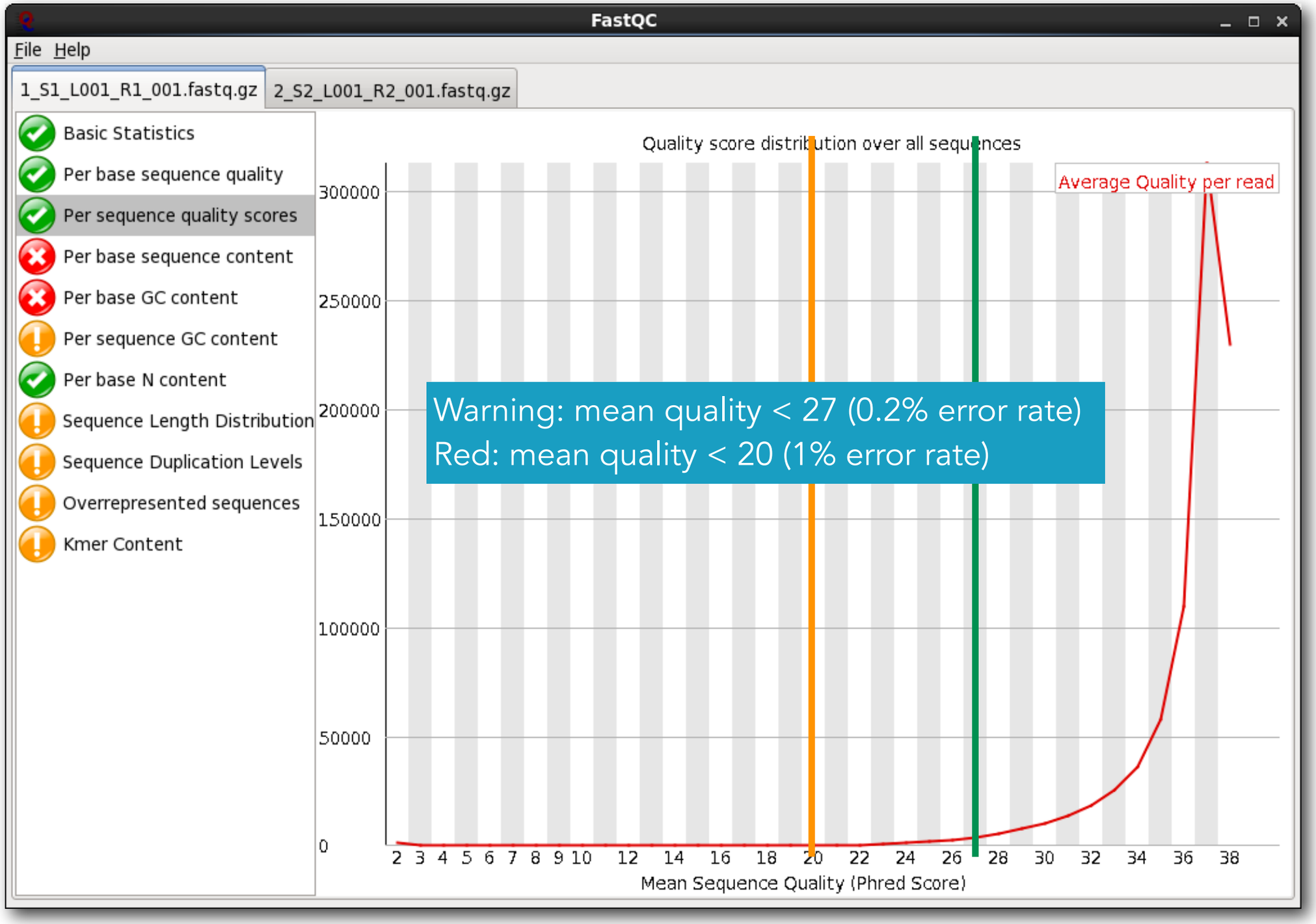
- Good
- Warning
- Error

The traffic lights are context dependend!

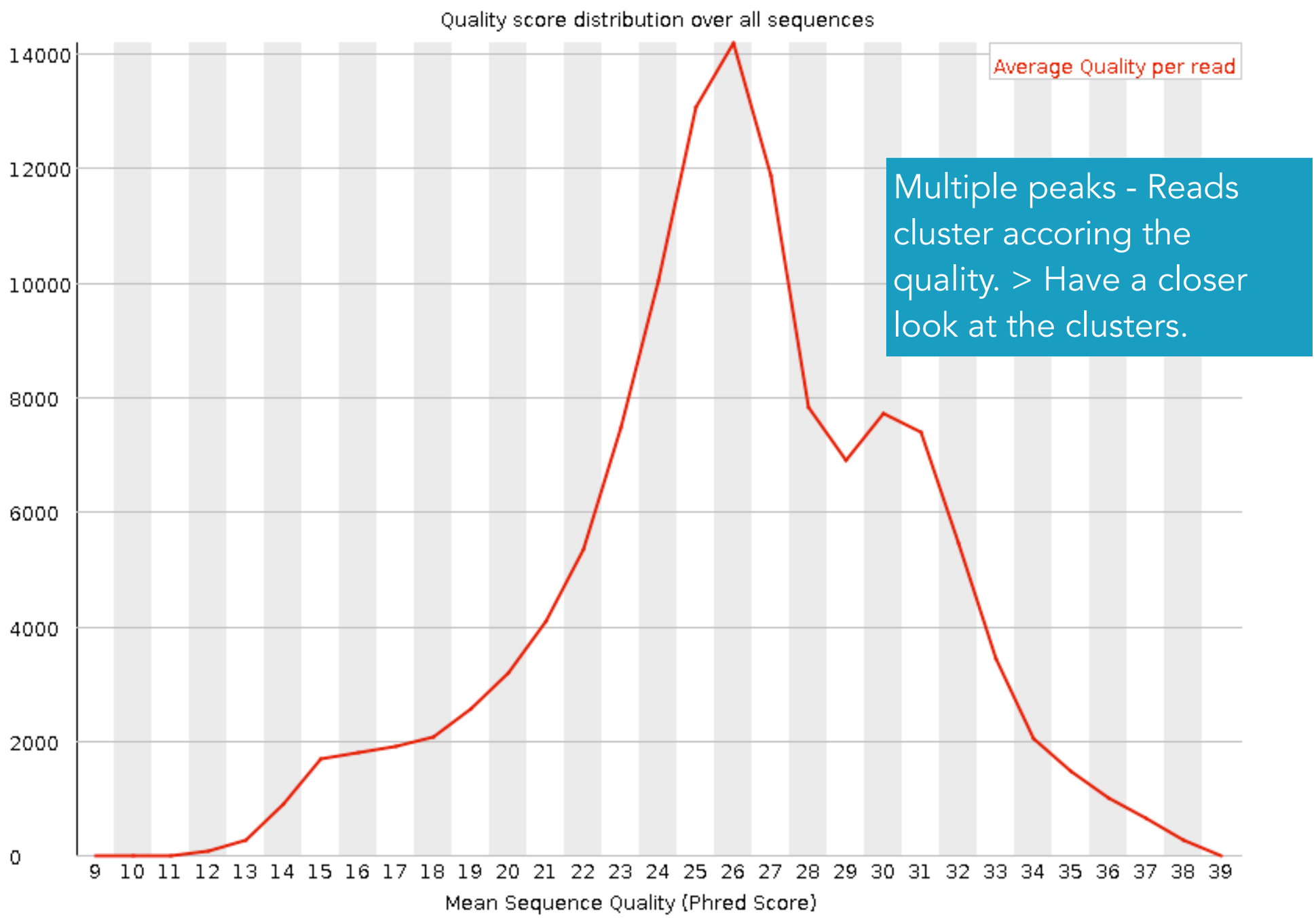


A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25.
This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

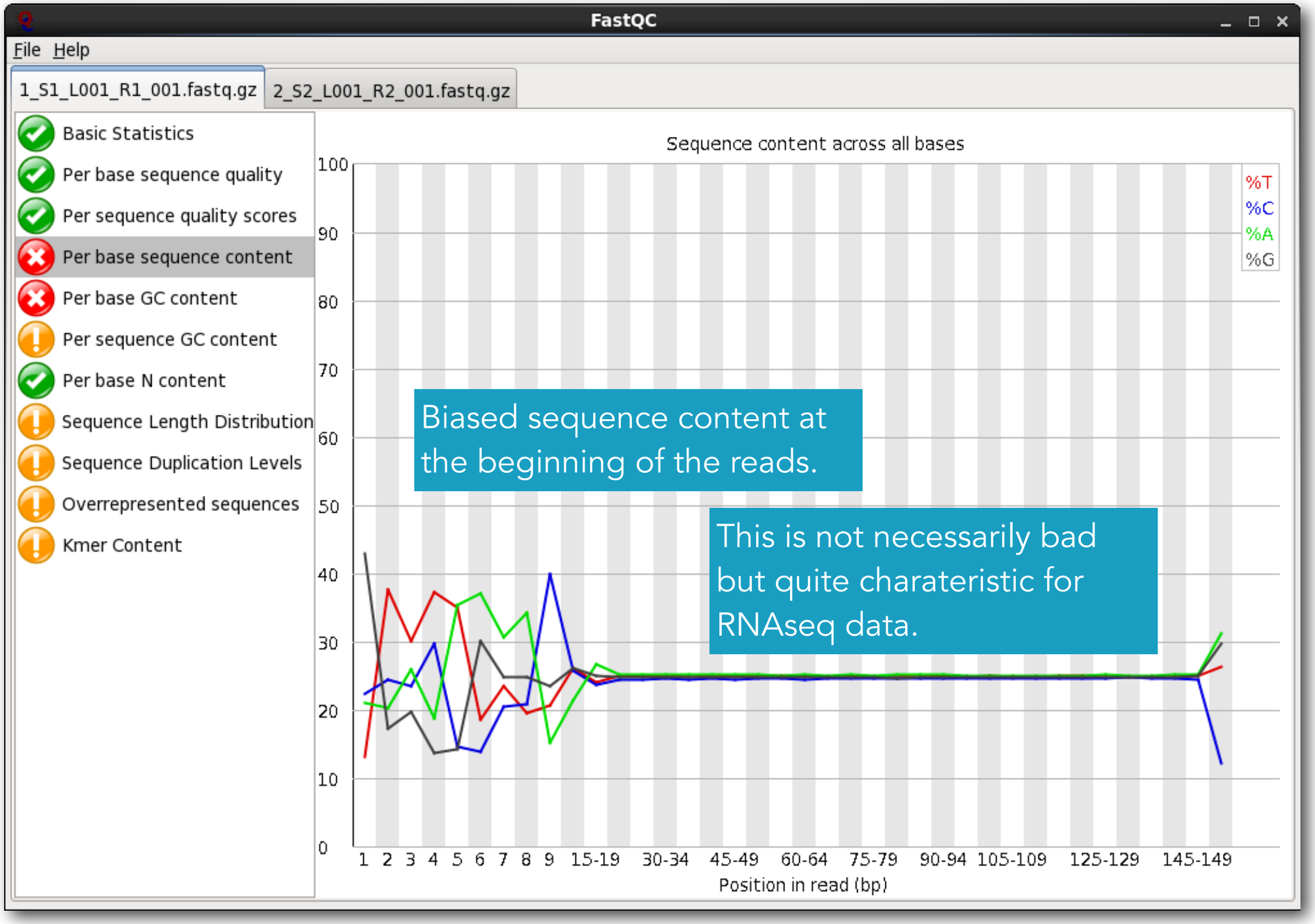
good	acceptable	bad
		



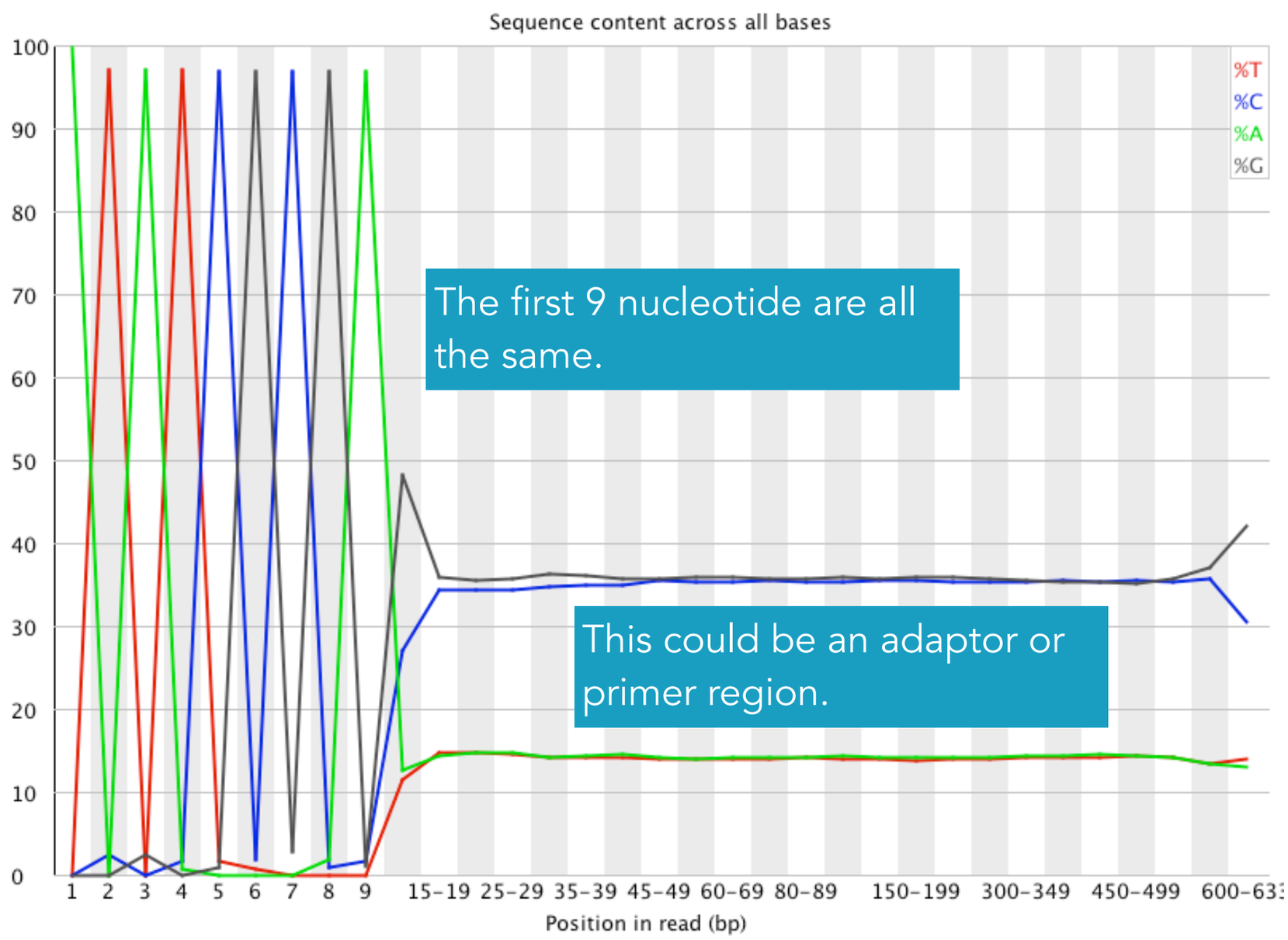
A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate.
An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

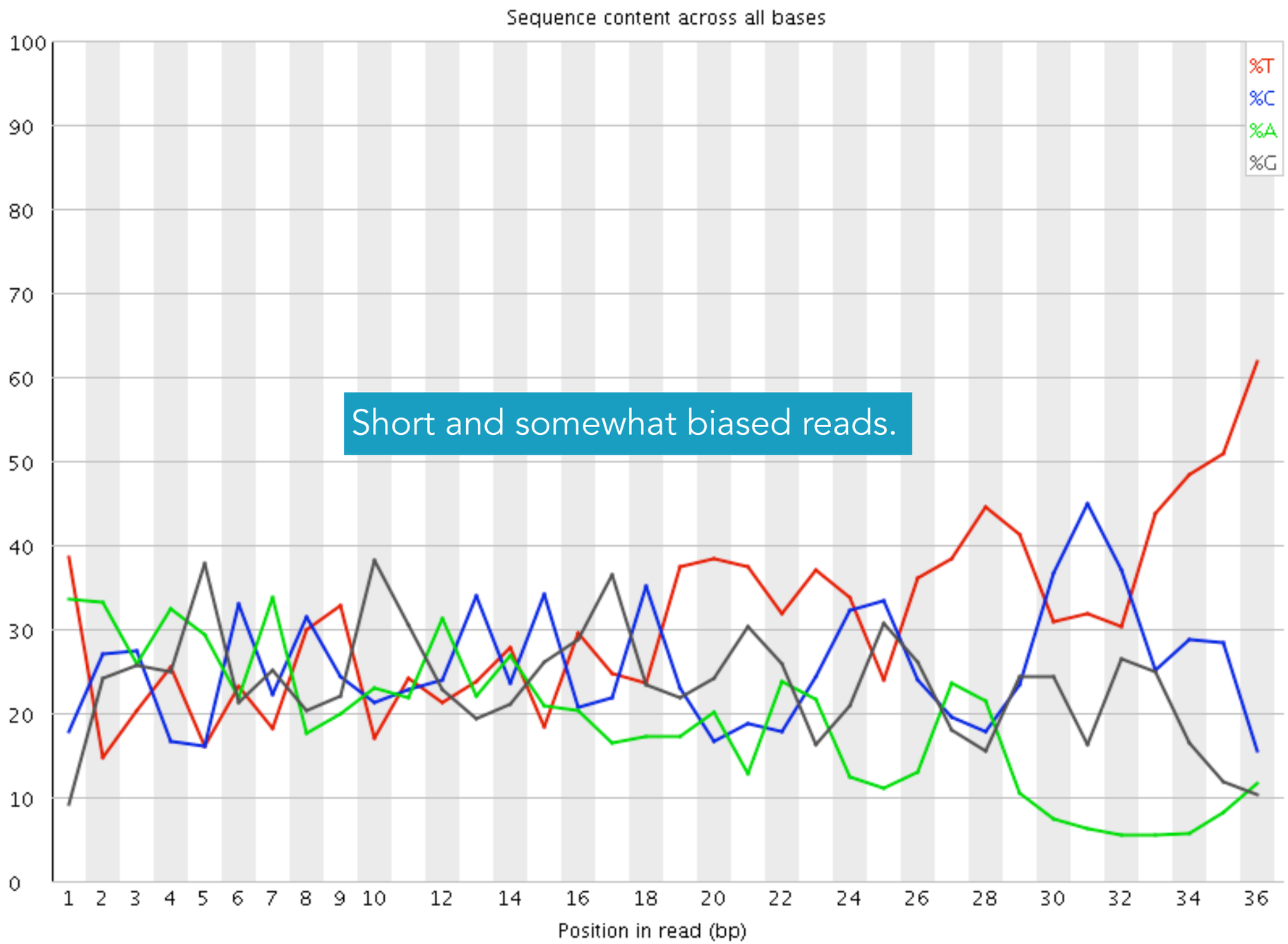


Multiple peaks - Reads cluster accoring the quality. > Have a closer look at the clusters.

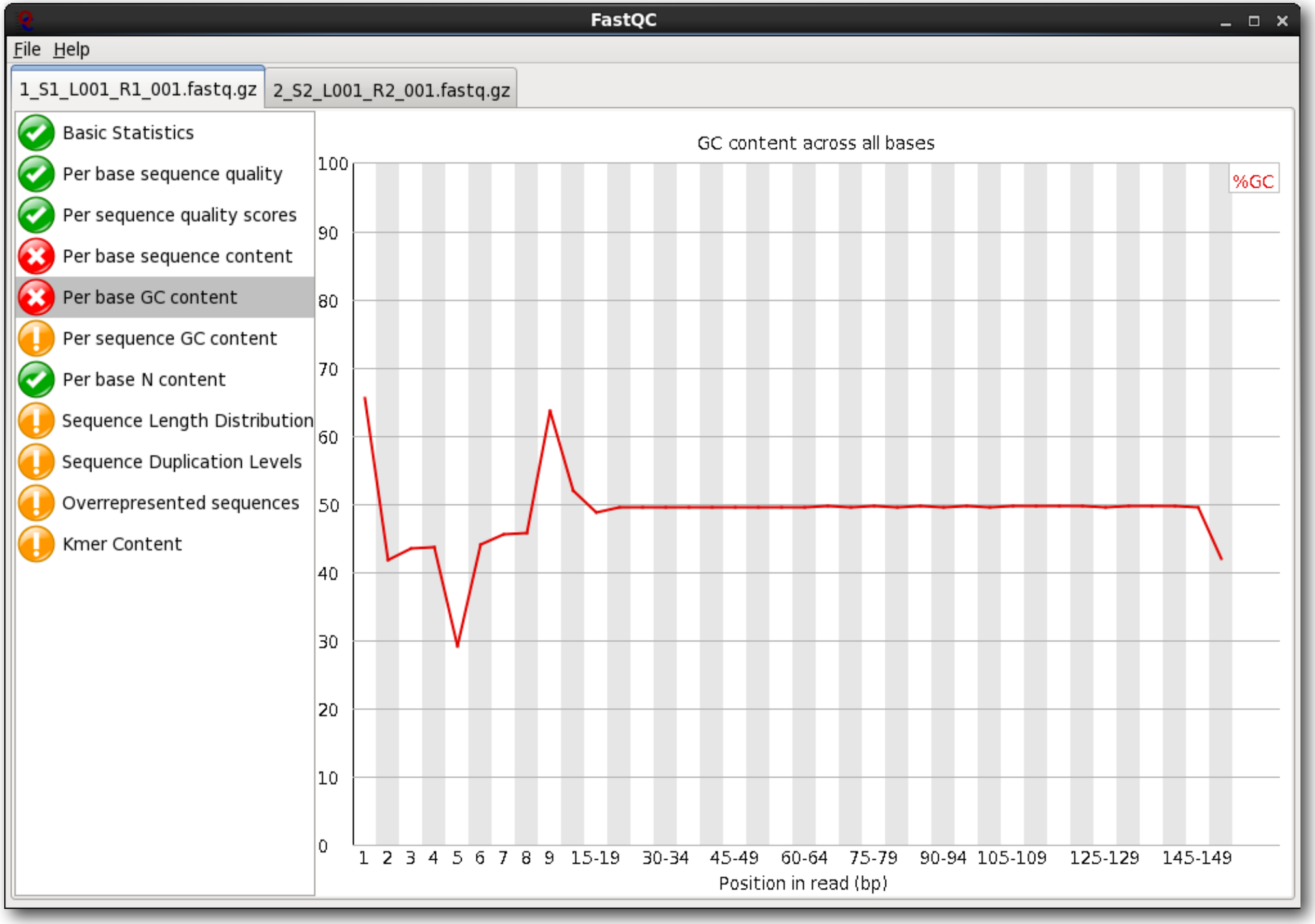


This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position.
This module will fail if the difference between A and T, or G and C is greater than 20% in any position.


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 15-19 25-29 35-39 45-49 60-69 80-89 150-199 300-349 450-499 600-633
Position in read (bp)
The first 9 nucleotide are all the same.
This could be an adaptor or primer region.

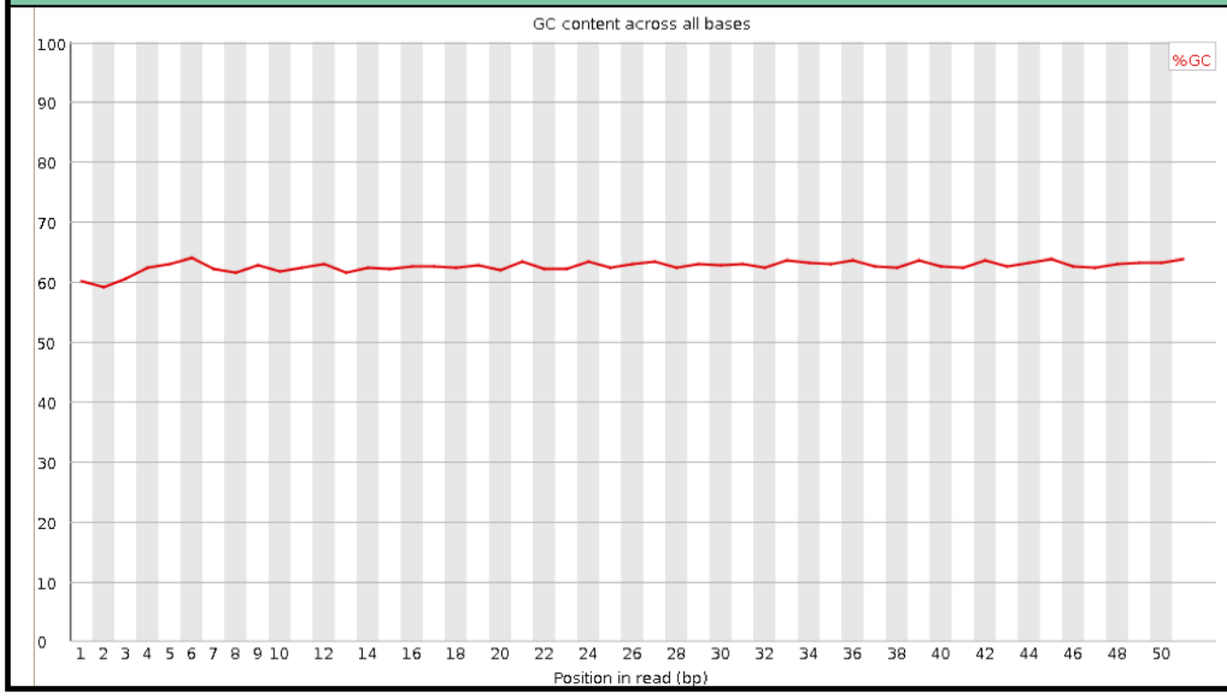
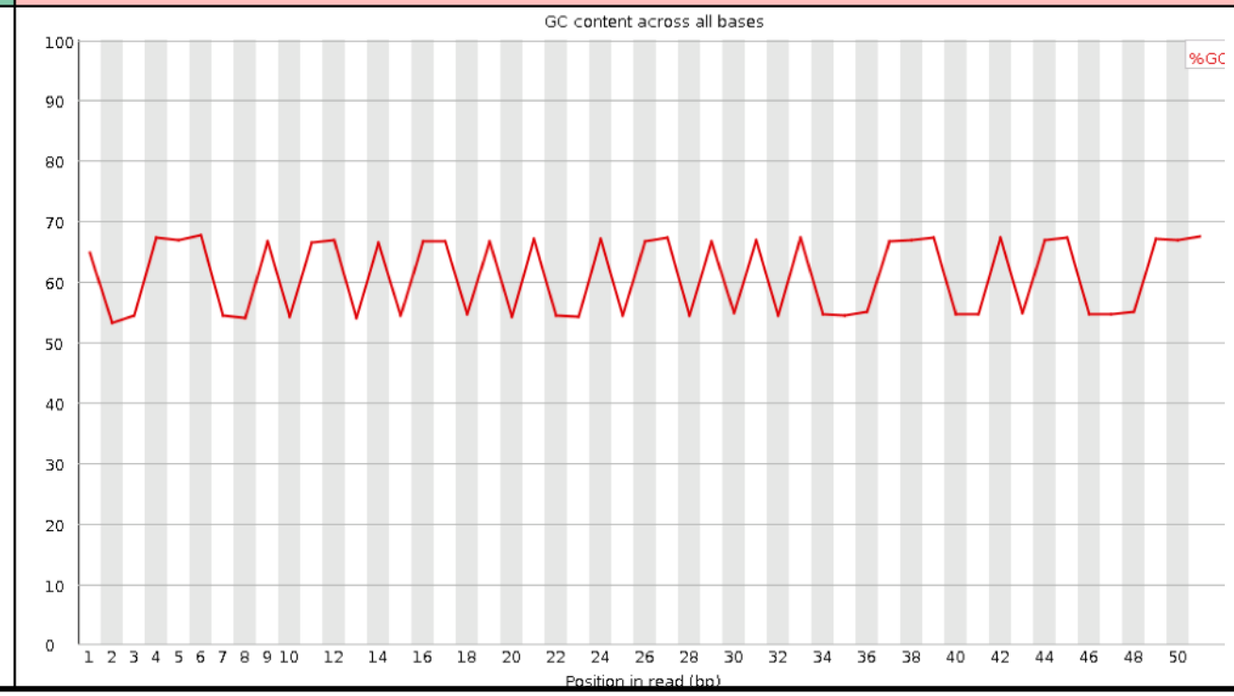


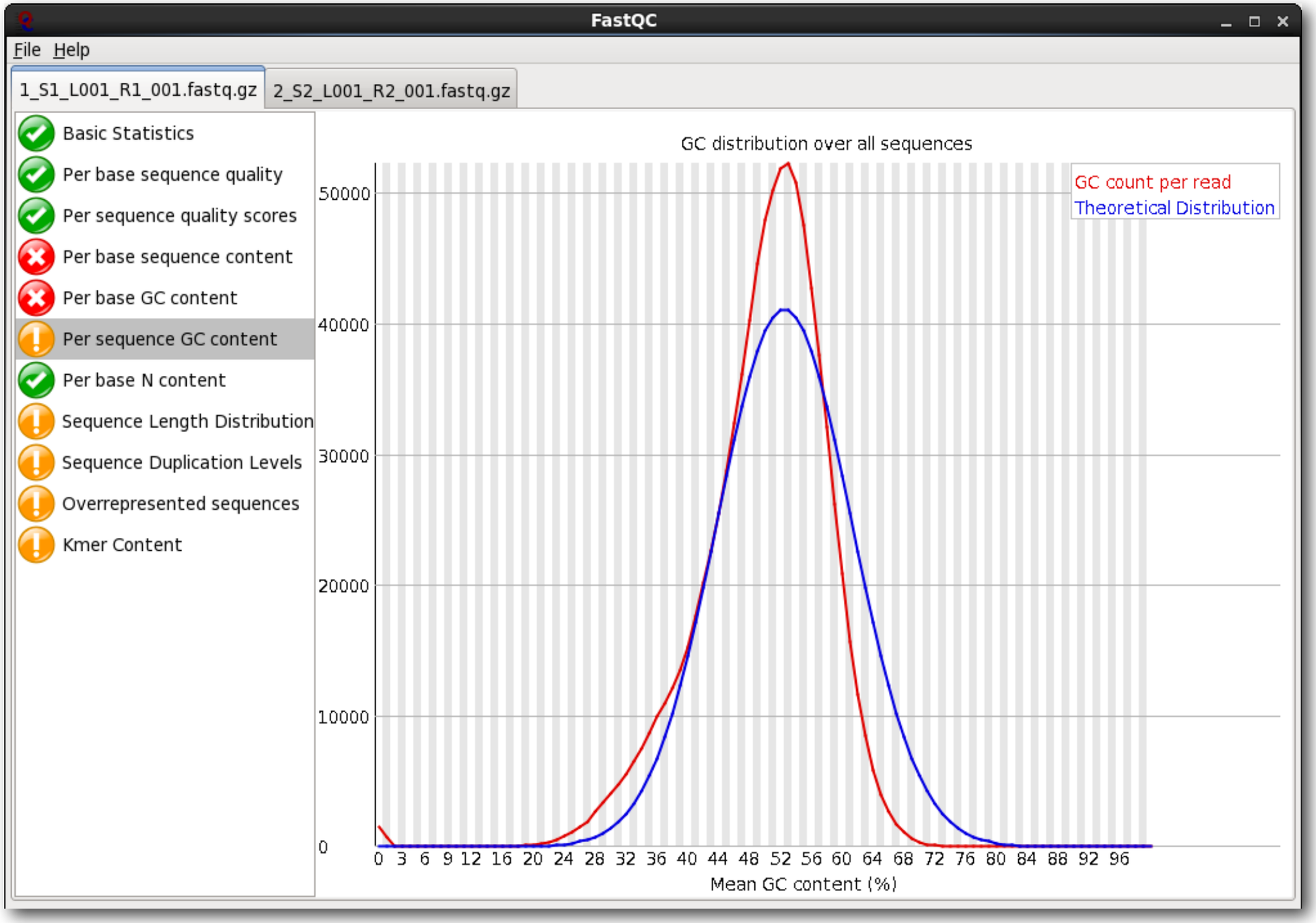
Short and somewhat biased reads.



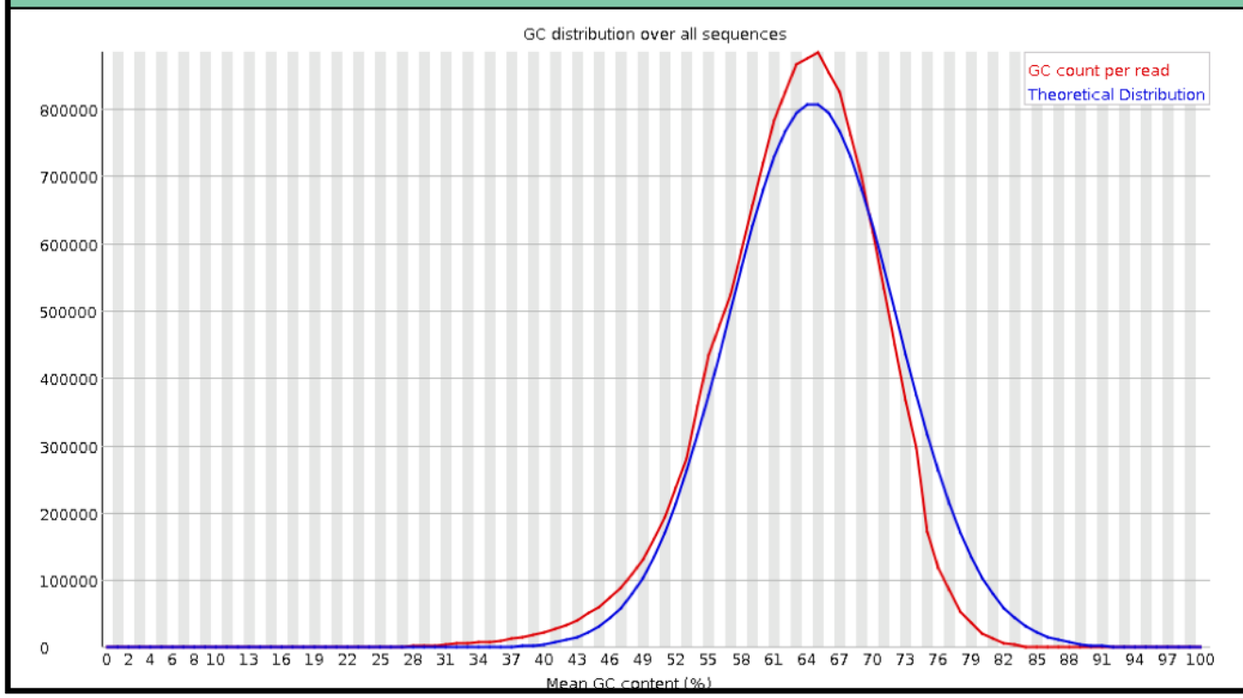
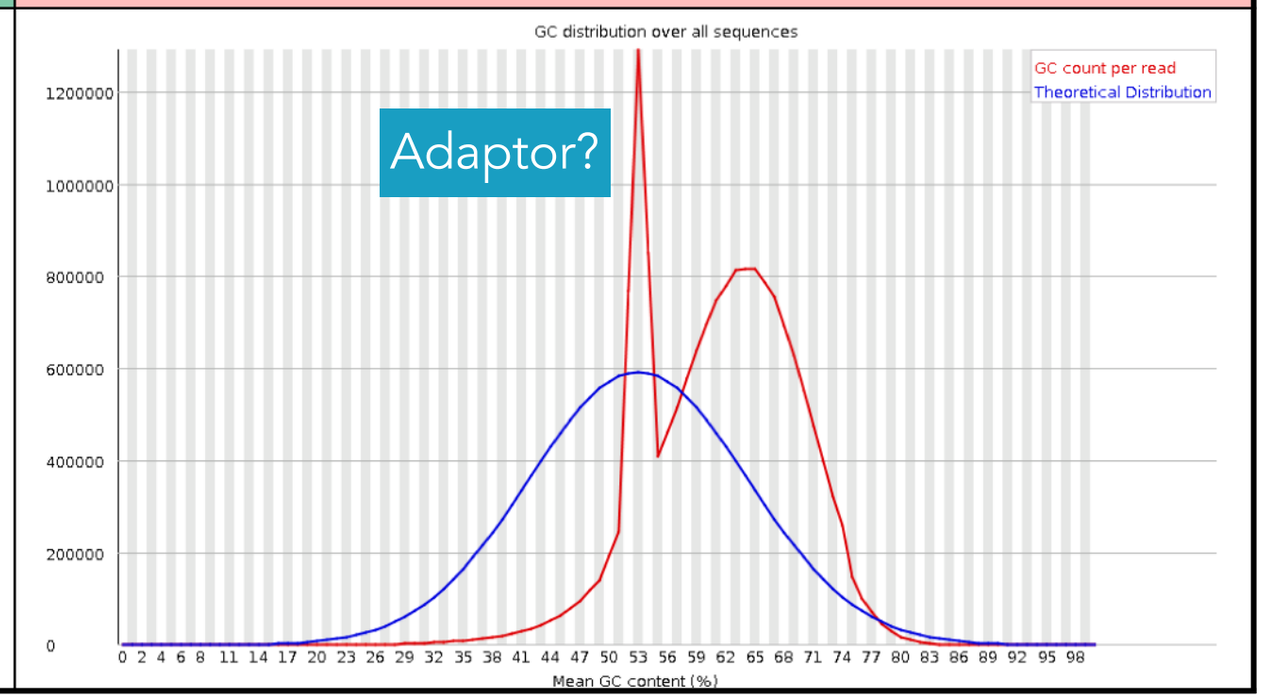
In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the line in this plot should run horizontally across the graph. The overall GC content should reflect the GC content of the underlying genome.

GC content per base:

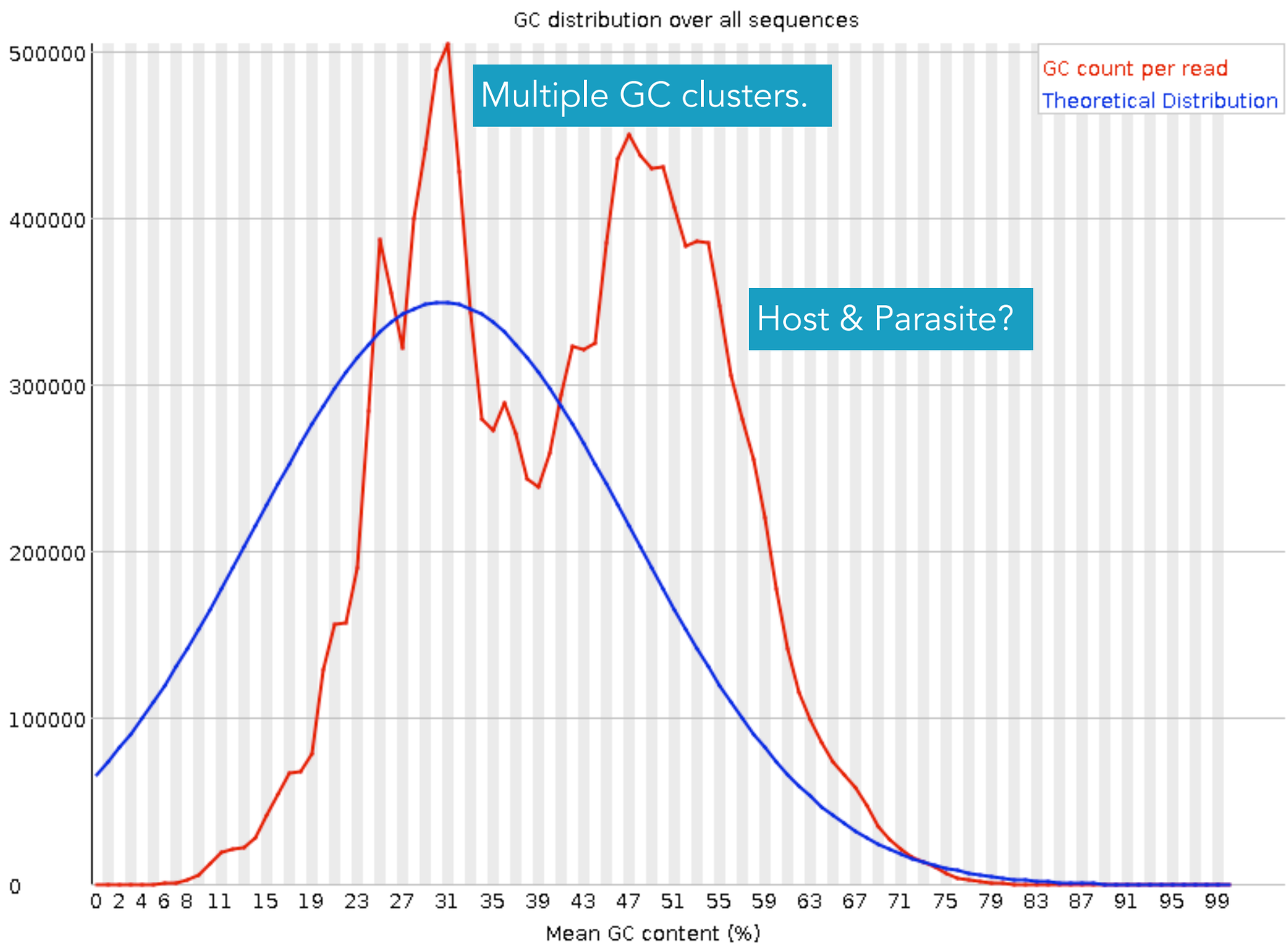
good	???
	



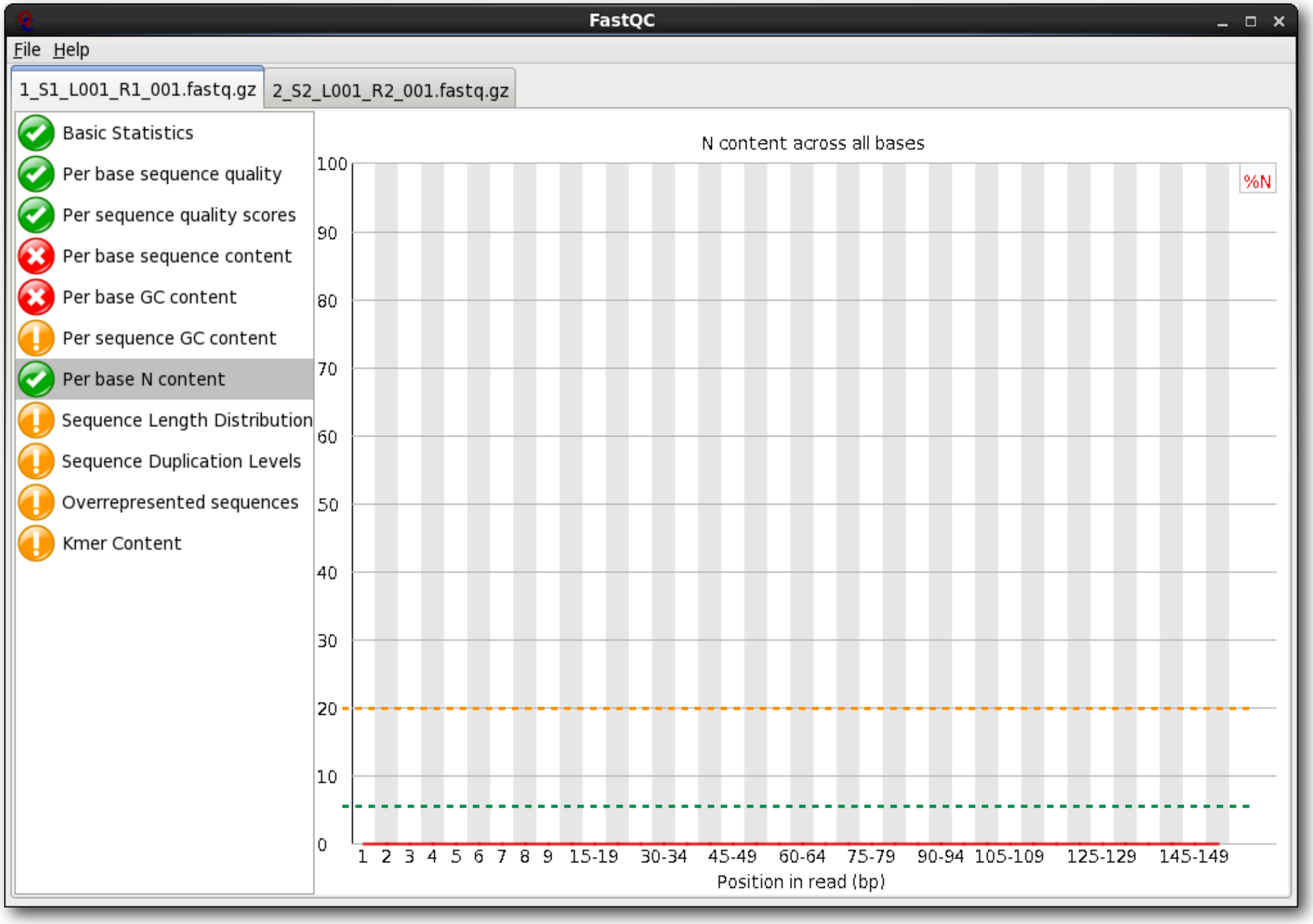
In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution.

good	bad
	

GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)
GC distribution over all sequences
GC count per read
Theoretical Distribution
Adaptor?
Mean GC content (%)

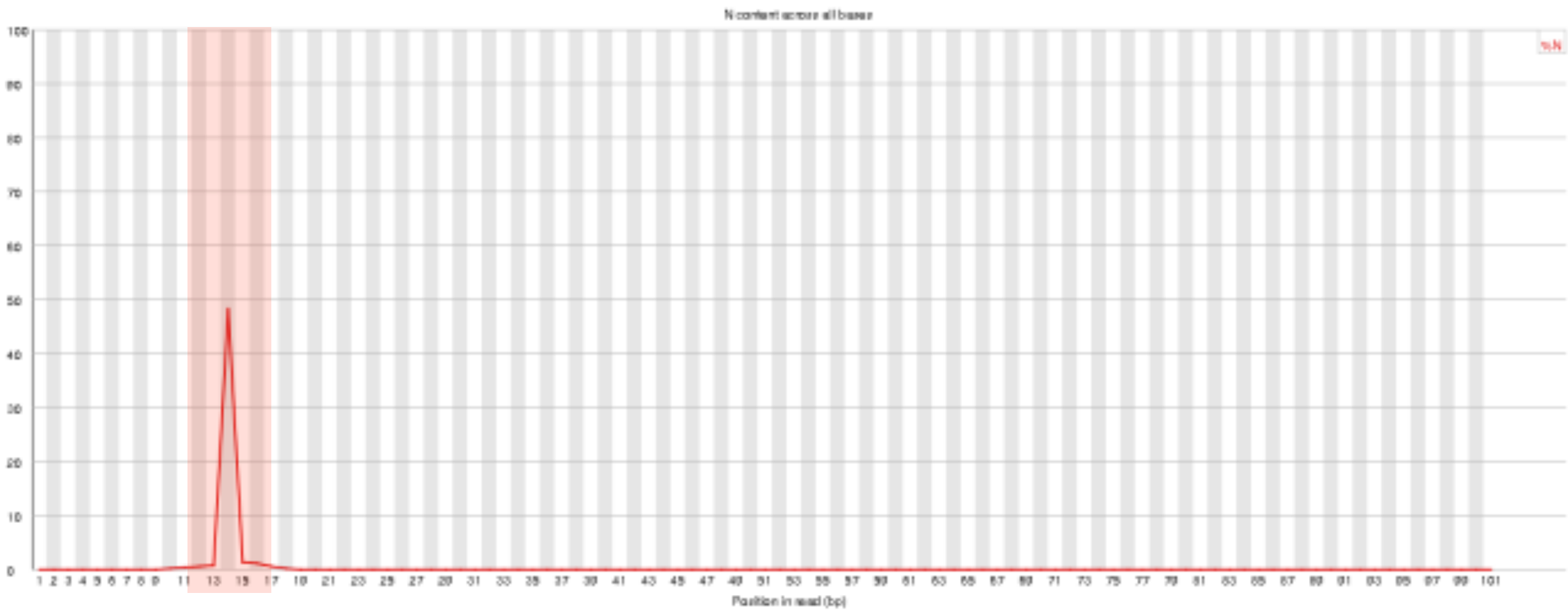

GC distribution over all sequences
Multiple GC clusters.
Host & Parasite?
GC count per read
Theoretical Distribution
500000
400000
300000
200000
100000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

NGS ▷ FastQC - N Content

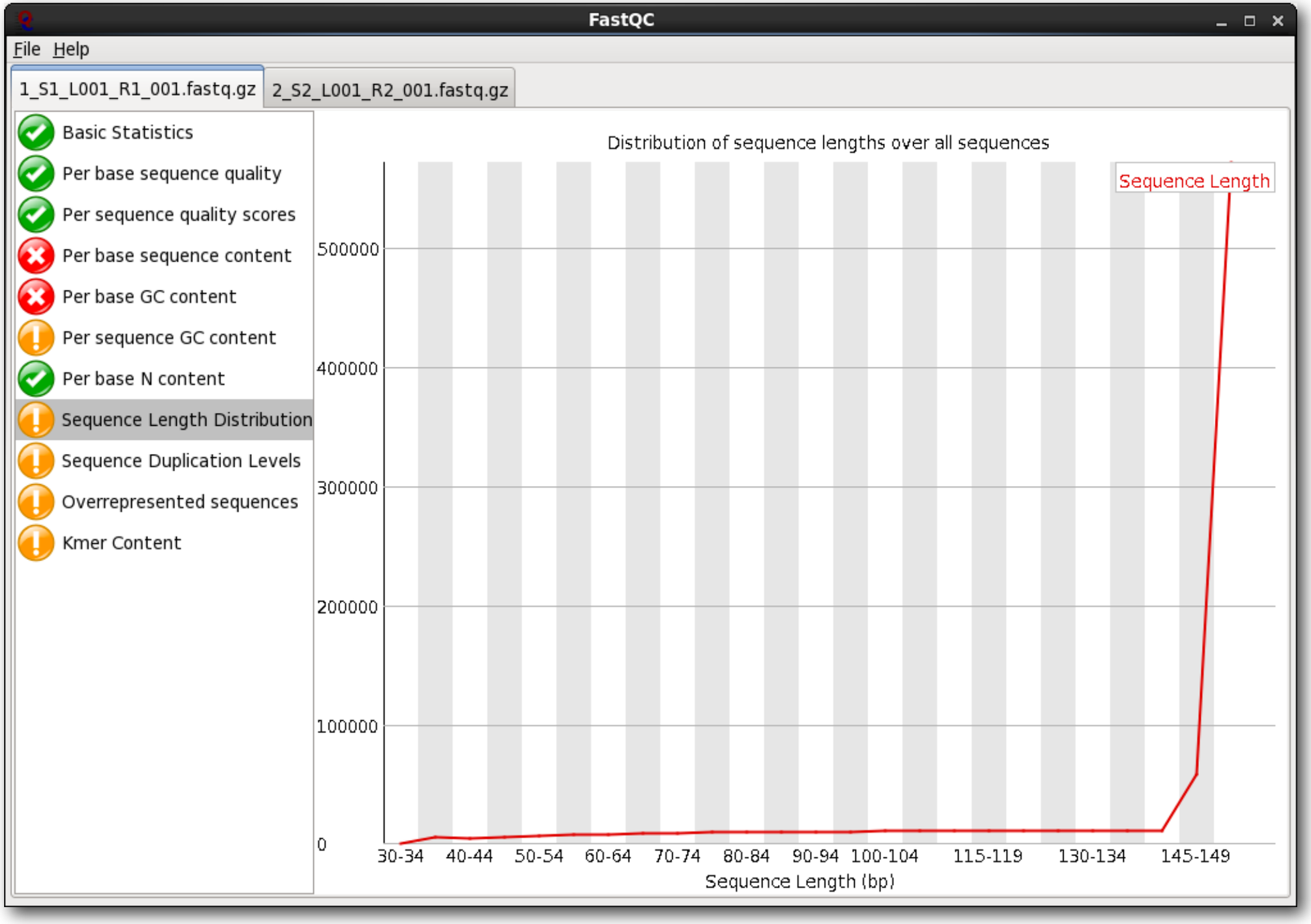


This module raises a warning if any position shows an N content of >5%.

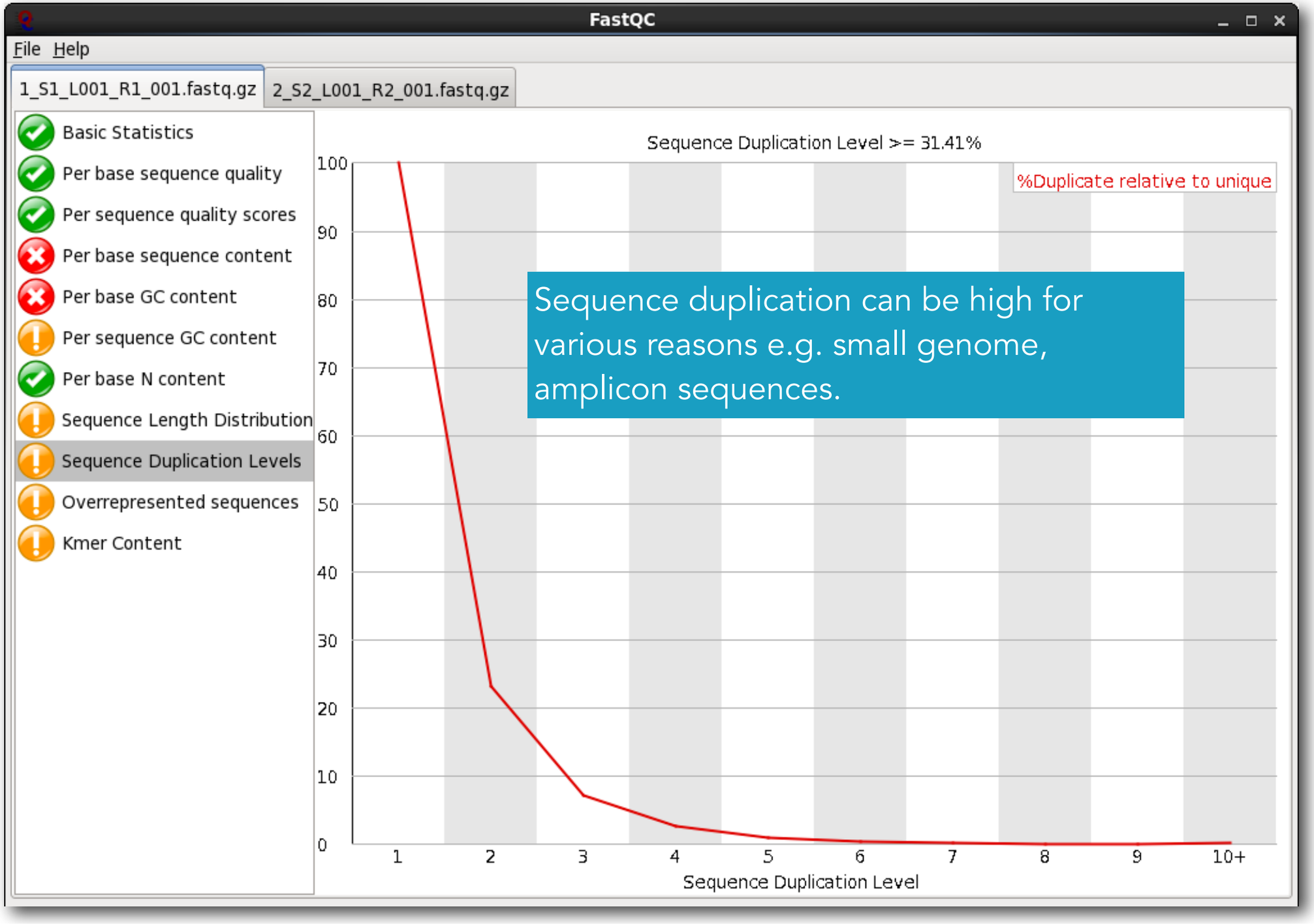
This module will raise an error if any position shows an N content of >20%.


N content across all bases
%N
100
90
80
70
60
50
40
30
20
10
0
Position in read (bp)

```
GTCGGGTTTTACCATTGGGTTTGGGTATTTCCACCCCCCGAATGGCTTGCGGTNTGTACNNNNNNNNNNNNGNNTNNNGTAAATAAACTTTCTGGATGGTGT
TCTTTAGTAAAAGGCGAAAGATTTATTCGTTCTGTATTGACGCCATGCCGGGTNNCTTTNNNNNNNNNNNNCNNTNNNTTATACAGCTTTGTGGAAATTTAC
AAAAATAGTACTTCAGAGGTAATAAATAAAATTATACCTCACCGTAAGCCTACNNCNANNNNNNNNNNNNNNNNNNNNTTGATAAGTTCCTTCACGTACTAC
TCCCGGGCACAGAAGTGCCCTTTCCACTGGGAAAGACTCATTTATTAATTACTNNTNTTNNNNNNNNNNNNNNNANNNAAATTAAATCATCTCACTCTTTAT
TGACAGAGTTACAATTGACTTCGTCCAAATCCAGAGAGGCATAGTTGACCATCAAAGACNNNNNANNNNNGNTGNNNGTCGTCGTCGGCCTAGGCACTGAC
CCCCGTCCCCTTAACAGTGAGTCATCAACAAAGAAATTGAGGTACACTTTTACNNTNTTNNNNNNNNNNNNNNNNNNAGCTTAGGTTCATCTGTGTCTTCT
TGTTGGCTTTGCCGAGATATACTGAGTTTATAGTTTCTGGCTTGGATGGCAGGNAATTTNNNNNCNNNNNTNCANNNTTCTCAATGCCTTTATTCCAGAAA
ACCTACAAGAGTTTTAAACTCTAAATGCAACTGGTTTCTAATTATTGAAAATANTTTATNNNNNCNNNNNGNCTNNNCTCAGTAAACTCAGGGGAGTAGAA
ATCGAGTTTTTGGACGAAGACGATGACGCAGCTGTGCATACTCAGACCACGCANNANCNNNNNNNNNNNNNNNNNNNNGGACTCTGCCTCGTCGTTGGTCGA
CTAAATATTGTTTGGAGTGGCAATGTGCCCAGCTGAAACACTGGGCACTTCAANCTGCCNNNNNGNNNNNGNTGNNNCTCTCTCATTAATGTTCAATGAGA
CGTCCAATTACGATCGCCAGTTTTTTTTTTTTATATATTTTTTTTTTTCTTATGTATTTNNNNNCNNNNNGTTCNNNTGCAGAACTTTTTATGTCTCCACT
GTGTGTCACACTCCAATGAAAAAGAGATAAAATCCTGGTGTAATTGAGACAATNNTNCANNNNNNNNNNNNNNNNNNCCATTGAGCAAGTATGGTAACGAT
TTCATGCTAGAAATGAAGGAGACTGCTGCCATTCTGAAACACGCCACCAAGCGTTCGTTNNNNNTNNNNNATTANNNTGGGCGAGGAACAAGGCCATCCGA
CCCCTTGGCCAAATACCTAGAATTGATCTTGAGCTTAGACGCATTCCGGCTTTNNTCCANNNNNNNNNNNNNNNANNNCTGTGCGTAGGCTGTGAGTAGGGT
GAGAACGTTATTATTTACGAACATAGTTACGAACTGCAGGATAATATTATTGTNGCTGANNNNNTNNNNNTNTCNNNTTCACCCGATAAGAACAAATTATT
CTTTAATCCAGGGGCTTTACGCGCCAATTGCCTCTGCTTTCCAATGGTATACTNNNNTNNNNNNNNNNNNNNNNNNNACCATTCTATGGACCATCTCTTGG
CTCTTTTTGAGCCCCTTTTGCCTTTTTTTAAACCTAACTATGGGACACTTATTNTTAAANNNNNNNNNNNANNANNNTTTTTAGCAGTAAAAGCTGTAACC
CCATAGAGAGAGGCATAAAGCTCAACAGCCGTTTGAAGAAAACTTTTTTGTGCNATTTTNNNNNNNNNNNANNTNNNAACGCAGTATCCTCGATATACTTA
GTGGTGAAGACGTTTACACTCGTCCGTTCCACATTCCTTTTCCTTCGTACACTNTGAACNNNNNNNNNNNTNNTNNNGACACCGGTAACAGCATCTTTTTT
CATCATCCACATCTGCTGCCGCAAGCATTGTTGGTCATCATCATCATCGCGTCNTATCTNNNNNNNNNNNTNNTNNNTATTTCATTGCCCGTGTCATGCAG
GAACAAACAGCTCCTATACGTGAAAATACCAAAAGGGTCGTTGCATCATTGAATGTTTATNNNNCANANGCCACTNNATATGGCATGAAGGTGCTTAACGA
CTTCCCCCAACTCCATATCATCATCTTCCAATATTTATAATACAACATCTTGCNNCNATNNNNNNNNNNNNNNTNNNATGTTTTCTAATACCAACATTATG
CTTGTAGACCTCGGCACCGAATGCTTCAGAGGTGACGAGAGACGTGAGAGCGATGAAGANNNNNGNNNNNTNTGNNNTTGAAGGGGTCATGCGGGGTTTCT
ACTCAACAGGATATGTACTCTGCAGAAGAATAAGAAGTTAAACAATATCACCANNANANNNNNNNNNNNNNNNNNNNATAAAACAATAAATTCAACGCGTT
```

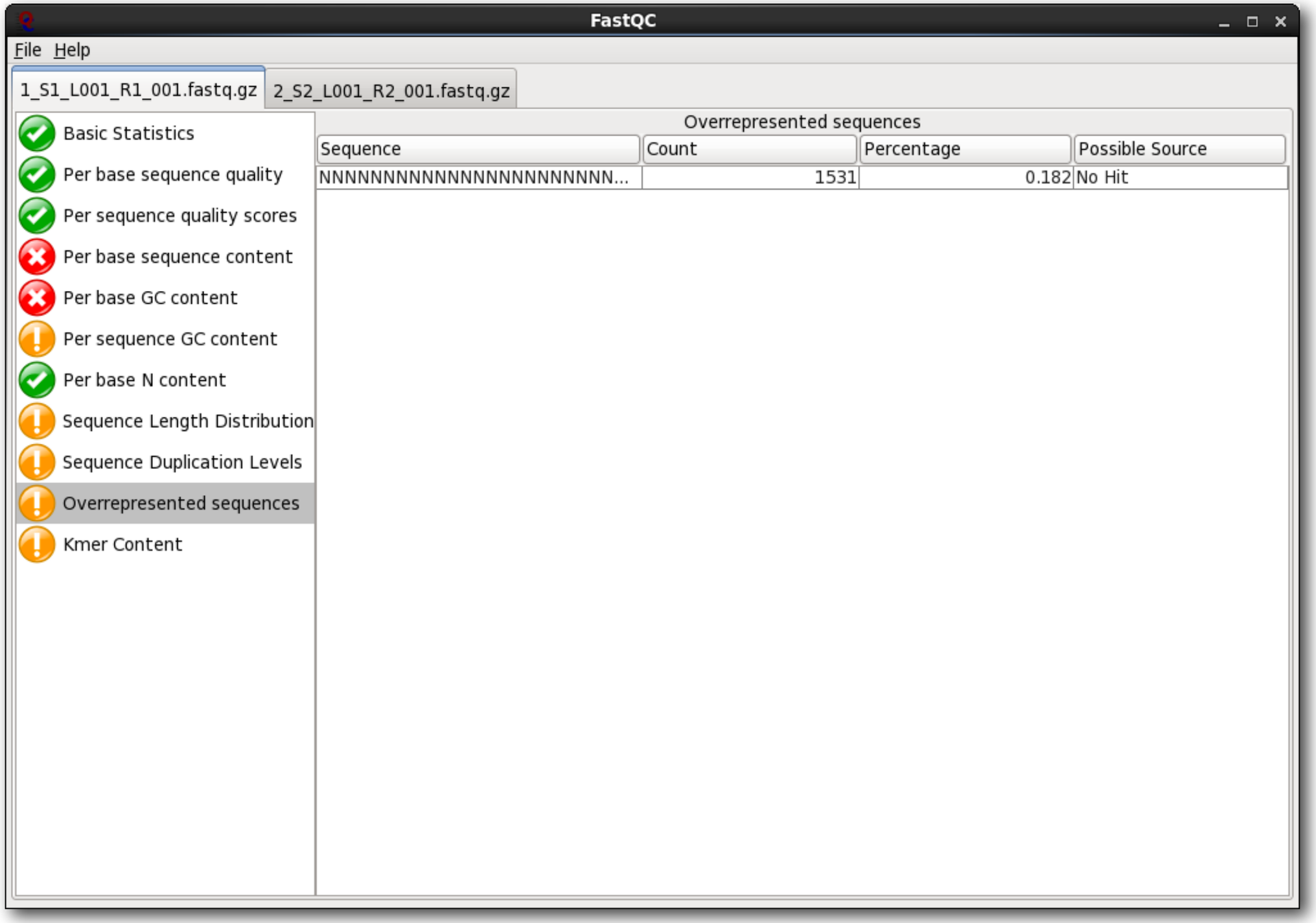


This module will raise a warning if all sequences are not the same length.
This module will raise an error if any of the sequences have zero length.

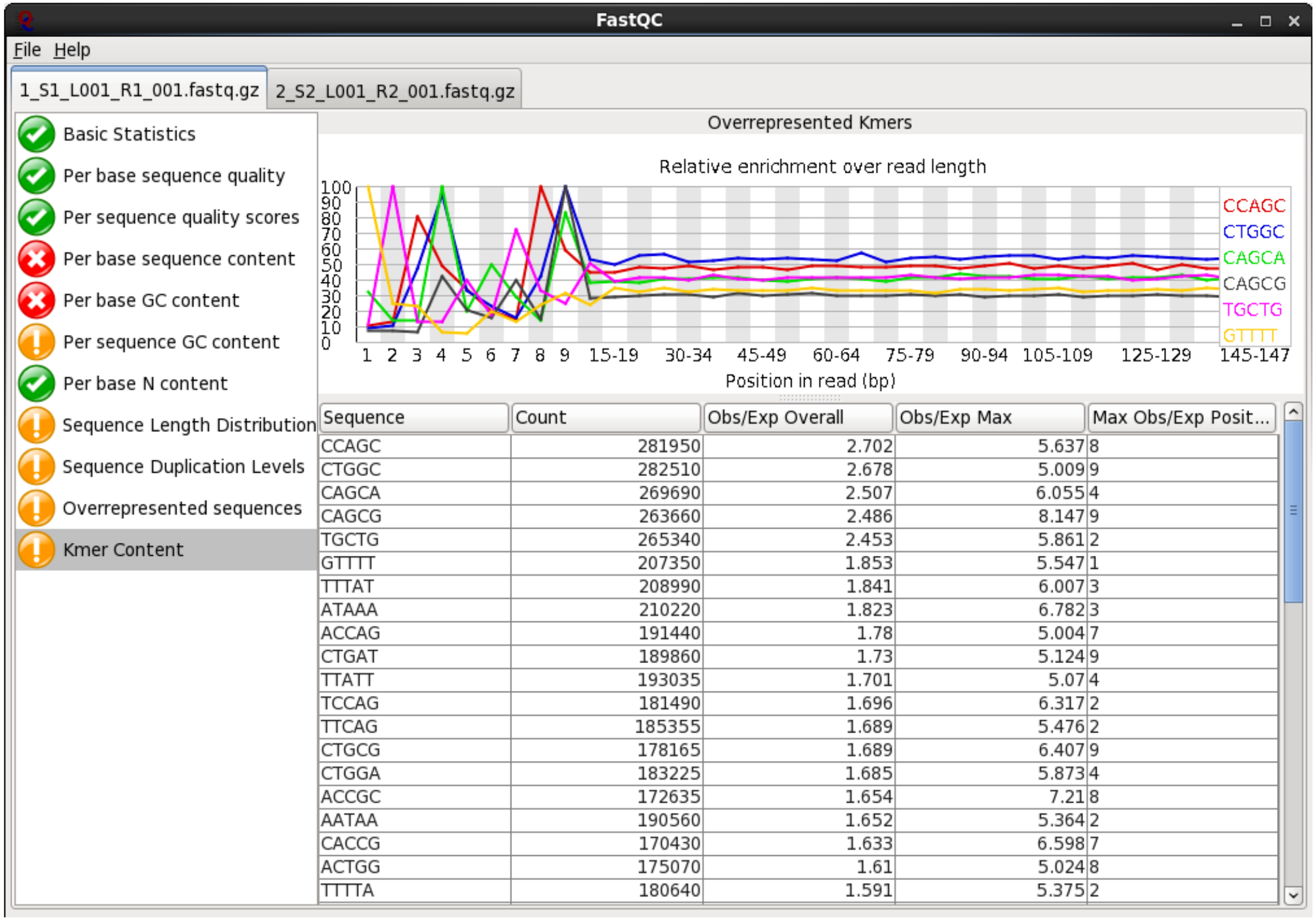


Sequence duplication can be high for various reasons e.g. small genome, amplicon sequences.

Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to underrepresent highly duplicated sequences.



This module lists all of the sequence which make up more than 0.1% of the total. To conserve memory only sequences which appear in the first 200,000 sequences are tracked to the end of the file. It is therefore possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module.



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Posit...
CCAGC	281950	2.702	5.637	8
CTGGC	282510	2.678	5.009	9
CAGCA	269690	2.507	6.055	4
CAGCG	263660	2.486	8.147	9
TGCTG	265340	2.453	5.861	2
GTTTT	207350	1.853	5.547	1
TTTAT	208990	1.841	6.007	3
ATAAA	210220	1.823	6.782	3
ACCAG	191440	1.78	5.004	7
CTGAT	189860	1.73	5.124	9
TTATT	193035	1.701	5.07	4
TCCAG	181490	1.696	6.317	2
TTCAG	185355	1.689	5.476	2
CTGCG	178165	1.689	6.407	9
CTGGA	183225	1.685	5.873	4
ACCGC	172635	1.654	7.21	8
AATAA	190560	1.652	5.364	2
CACCG	170430	1.633	6.598	7
ACTGG	175070	1.61	5.024	8
TTTTA	180640	1.591	5.375	2

This module will issue a warning if any k-mer is enriched more than 3 fold overall, or more than 5 fold at any individual position.

1. Choose the NGS technology and sample design according to your needs.

2. Keep your raw data safe and submit it as early as possible.

3. Coping one file (archive) is safer than coping multiple files.

NGS Data Submission

Data Corruption

- Upload one file per sample.
- Upload only archived files.
- Use md5sum to confirm data transfer.

A word of advice,

copy one file (archive) is safer than coping multiple files.

The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. A typical workflow includes the isolation and preparation of material for sequencing, a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatic analysis pipeline. ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Data Deposition Example from the Literature:

Mushegian *et al.* (2018) **Environmental sources of bacteria and genetic variation in behavior influence host-associated microbiota**. AEM doi:10.1128/AEM.01547-18.

Sequence data are deposited in the **European Nucleotide Archive of the EBI under accession number PRJEB30308** (http://www.ebi.ac.uk/ena/data/view/PRJEB30308). Data tables, OTUs sequences and code used for analysis can be found on **Github** at https://github.com/amusheg/Daphnia-microbiota-behavior and will be deposited in **Dryad** upon publication.

EMBL-EBI

Services | Research | Training | About us

ENA
European Nucleotide Archive

Search
Examples: BN000065, histone
Advanced
Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

Contact Helpdesk

Study: PRJEB30308

Microbiota associated with Daphnia exhibiting genetic variation in behavior

View: Project XML Study XML

Download: Project XML Study XML

Name
Microbiota of browsing Daphnia

Submitting Centre
Universitaet Basel

Secondary accession(s)
ERP112744

Description
In many organisms, host-associated microbial communities are acquired horizontally after birth. This process is believed to be shaped by a combination of environmental and host genetic factors. We examined whether genetic variation in animal behavior could affect the composition of the animal's microbiota in different environments. The freshwater crustacean Daphnia magna is primarily planktonic, but exhibits variation in the degree to which it browses in benthic sediments. We performed an experiment with clonal lines of D. magna showing different levels of sediment-browsing intensity exposed to either bacteria-rich or bacteria-poor sediment or whose access to sediments was prevented. We find that the bacterial composition of the environment and genotype-specific browsing intensity together influence the composition of the Daphnia-associated bacterial community. Exposure to more diverse bacteria did not lead to a more diverse microbiome, but greater abundances of environment-specific bacteria were found associated with host genotypes that exhibited greater browsing behavior. Our results indicate that, although there is a great deal of variation between individuals, behavior can mediate genotype-by-environment interaction effects on microbiome composition.

Navigation | Read Files | Portal | Attributes

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1 - 512 of 512 results in TEXT

Select columns

Showing results 1 - 10 of 512 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
PRJEB30308	SAMEA5166093	ERS2973813	ERX2993334	ERR2990925	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		
PRJEB30308	SAMEA5166094	ERS2973814	ERX2993335	ERR2990926	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		
PRJEB30308	SAMEA5166095	ERS2973815	ERX2993336	ERR2990927	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		
PRJEB30308	SAMEA5166096	ERS2973816	ERX2993337	ERR2990928	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		

NCBI Resources How To Sign in to NCBI

BioProject BioProject PRJEB30308 Search

Create alert Advanced Browse by Project attributes Help

Display Settings: Send to:

Microbiota of browsing Daphnia

Accession: PRJEB30308 ID: 516850

Microbiota associated with Daphnia exhibiting genetic variation in behavior

In many organisms, host-associated microbial communities are acquired horizontally after birth. More...

Accession	PRJEB30308
Scope	Monoisolate
Submission	Registration date: 24-Jan-2019 **Universitaet Basel**

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	512
OTHER DATASETS	
BioSample	512

SRA Data Details

Parameter	Value
Data volume, Gbases	22
Data volume, Mbytes	14805

Related information

BioSample

SRA

Recent activity

Turn Off Clear

PRJEB30308 (1)
BioProject

Microbiota of browsing Daphnia
BioProject

The European Nucleotide Archive in 2017

A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived

Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects

See more...

OPEN ACCESS Freely available online PLoS one

High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples

Hirokazu Toju[1,2*], Akifumi S. Tanabe[2], Satoshi Yamamoto[2], Hirotoshi Sato[3]

1 The Hakubi Center for Advanced Research, Kyoto University, Sakyo, Kyoto, Japan, **2** Division of Biological Science, Graduate School of Science, Kyoto University, Sakyo, Kyoto, Japan, **3** Kansai Research Center, Forestry and Forest Products Research Institute, Nagaikyutaro-68, Momoyama, Fushimi, Kyoto, Japan

Table 2. Seven ascomycete and seven basidiomycete fungi are shown with the accession numbers of their ITS sequences.

ID	Species	Family	Order	Phylum	GenBank accession
A1	*Podostroma cornu-damae* (Pat.) Boedijn	Hypocreaceae	Hypocreales	Ascomycota	AB509797
A2	*Leotia lubrica* (Scop.) Pers.: Fr.	Leotiaceae	Helotiales	Ascomycota	AB509686
A3	*Phillipsia domingensis* (Berk.) Berk.	Sarcoscyphaceae	Pezizales	Ascomycota	AB509610
A4	*Xylaria* sp.	Xylariaceae	Xylariales	Ascomycota	AB509642
A5	*Cordyceps nutans* Pat.	Cordycipitaceae	Hypocreales	Ascomycota	AB509505
A6	*Vibrissea truncorum* Fr.	Vibrisseaceae	Helotiales	Ascomycota	AB509599
A7	*Trichocoma paradoxa* Jungh.	Trichocomaceae	Eurotiales	Ascomycota	AB509823
B1	*Auricularia aff. auricula* (Hook.) Underw.	Auriculariaceae	Auriculariales	Basidiomycota	AB509633
B2	*Bjerkandera adusta* (Willd.: Fr.) Karst.	Meruliaceae	Polyporales	Basidiomycota	AB509484
B3	*Laccaria vinaceoavellanea* Hongo	Hydnangiaceae	Agaricales	Basidiomycota	AB509671
B4	*Geastrum mirabile* (Mont.) Fisch.	Geastraceae	Geastrales	Basidiomycota	AB509736
B5	*Boletus ornatipes* Peck	Boletaceae	Boletales	Basidiomycota	AB509727
B6	*Thelephora aurantiotincta* Corner	Thelephoraceae	Thelephorales	Basidiomycota	AB509809
B7	*Amanita farinosa* Schw.	Amanitaceae	Agaricales	Basidiomycota	AB509651

DNA was extracted from fruiting body specimens collected on Yakushima Island, Kagoshima Prefecture, Japan. The fruiting body specimens were deposited in Kyoto University Herbarium (KYO). See Kirk *et al.* [3] and NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/guide/taxonomy/) for the taxonomy of the specimens.
doi:10.1371/journal.pone.0040863.t002