

Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Massive **P**arallel **S**equencing

First Generation Sequencing

Sanger Sequencing

Second Generation Sequencing

Next Generation Sequencing (NGS)

Third Generation Sequencing

Single Molecule Sequencing

Fourth Generation Sequencing

??? Sequencing

First Generation Sequencing
Sanger Sequencing

Second Generation sequencing
Next Generation Sequencing (NGS)

Third Generation Sequencing
Single Molecule Sequencing

Fourth Generation Sequencing

Massive
Parallel
Sequencing

PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing

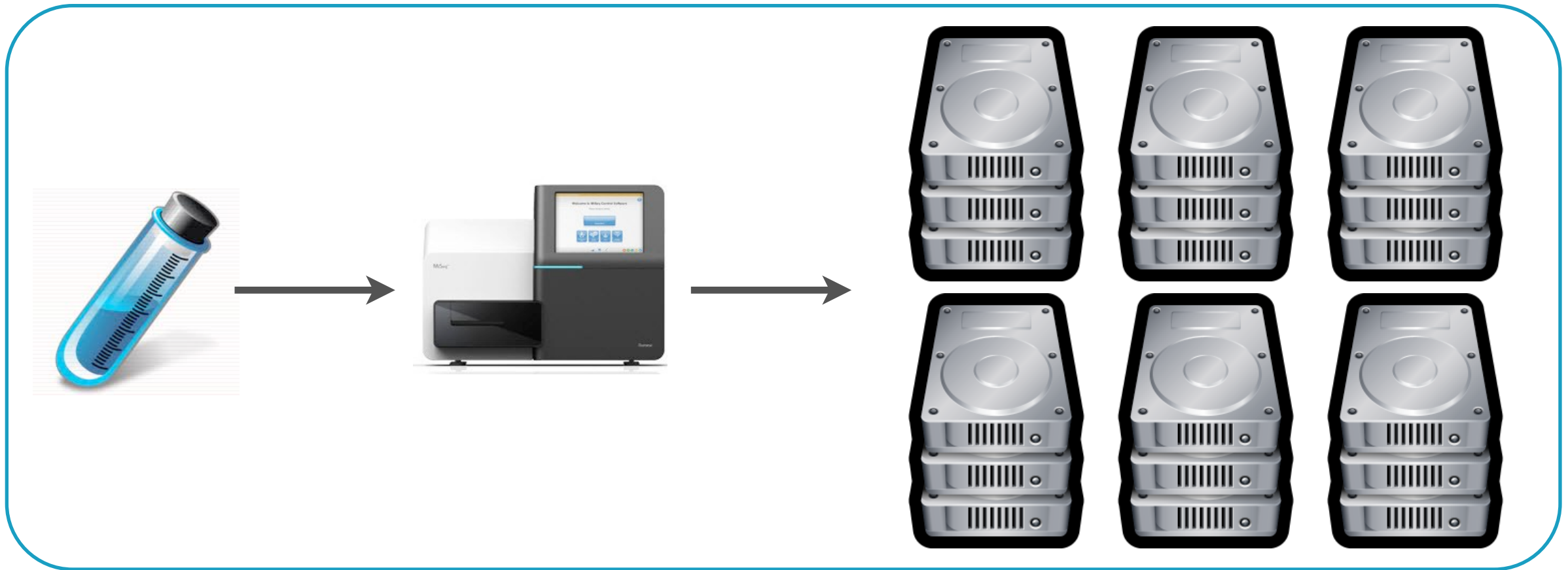


Actually, that's the coffee machine...this is the next-gen sequencer.

Next (Next) Generation Sequencing **Hype**



Next (Next) Generation Sequencing **Reality**

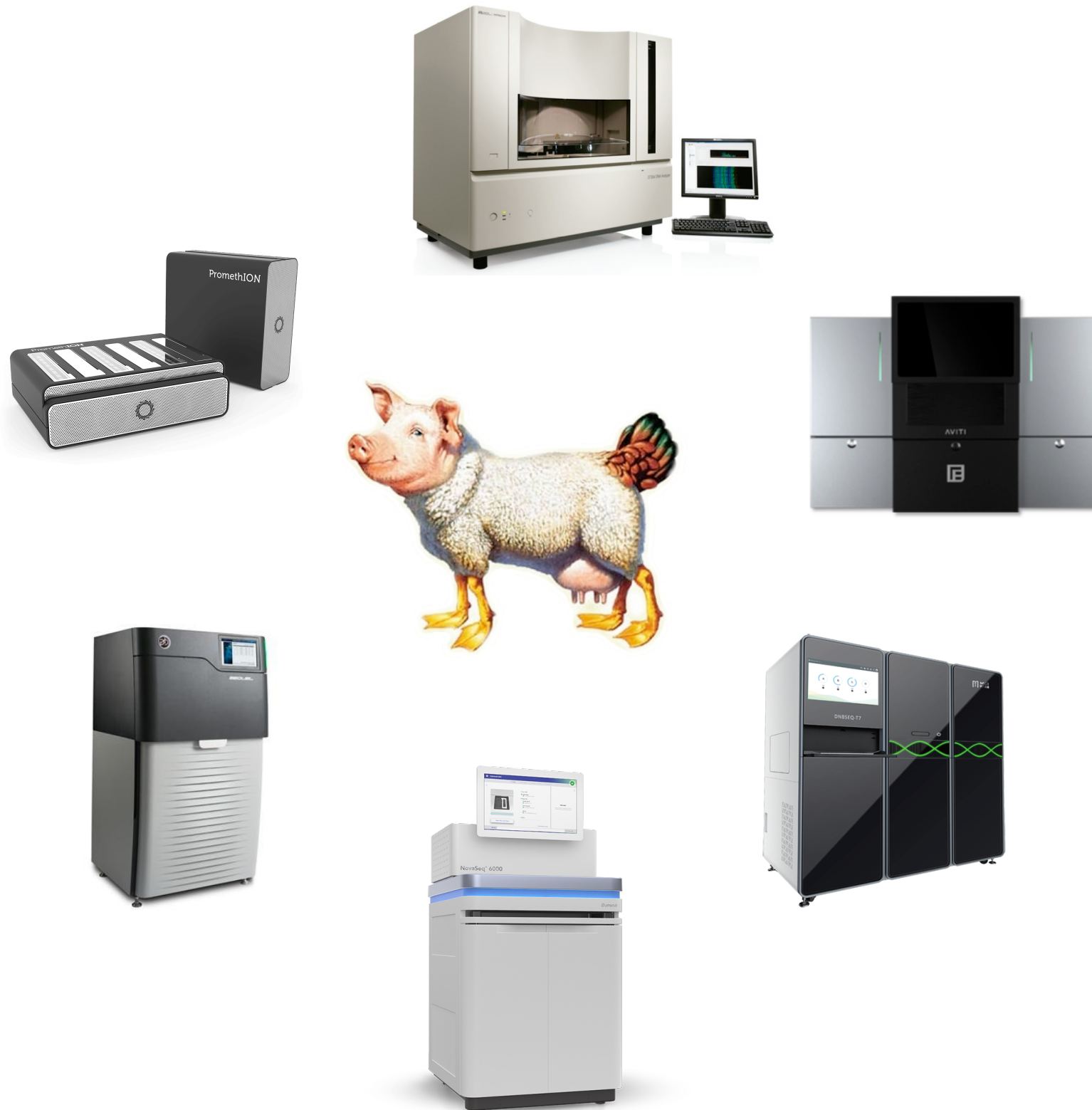


The **First Law of Technology** says we invariably **overestimate** the **short-term impact** of a truly transformational discovery, while **underestimating** its **longer-term effects**.

<https://www.scientificamerican.com/>

Sequencing Technologies

PopGen ► Massive Parallel Sequencing



1 Sanger (chain termination)

Roche 454 Pyrosequencing (pyrophosphate)

Ion Torrent (semiconductor technology)

2 **Illumina** (fluorescent - sequencing by synthesis (SBS))

Singular Genomics (fluorescent - rapid SBS)

Aviti (fluorescent - sequencing by avidity)

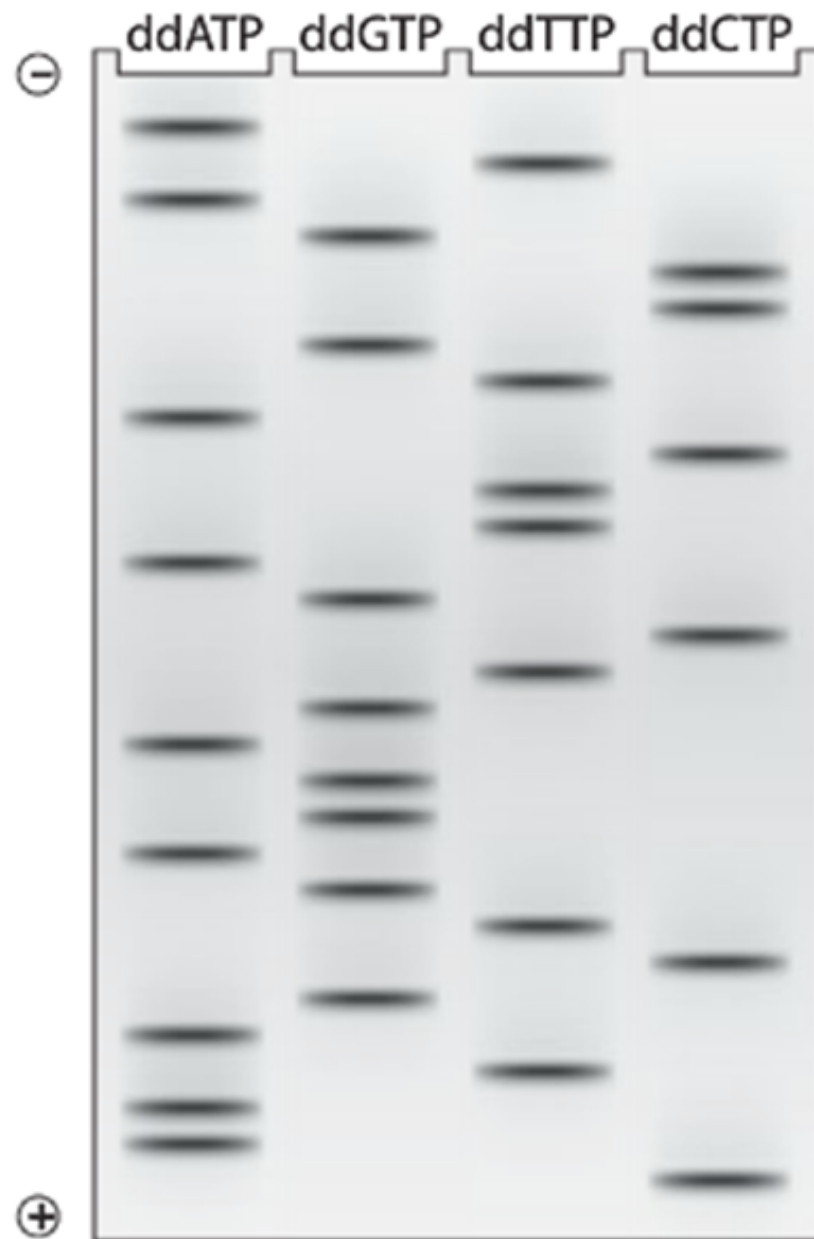
PacBio (fluorophore)

3 **Nanopore** (ionic current)

Helicos - SeqLL (fluorescent)

4 Bionano - Saphyr (third-generation optical mapping)

PopGen ► Massive Parallel Sequencing



The Nobel Prize in Chemistry 1980

Paul Berg, Walter Gilbert, Frederick Sanger



Paul Berg

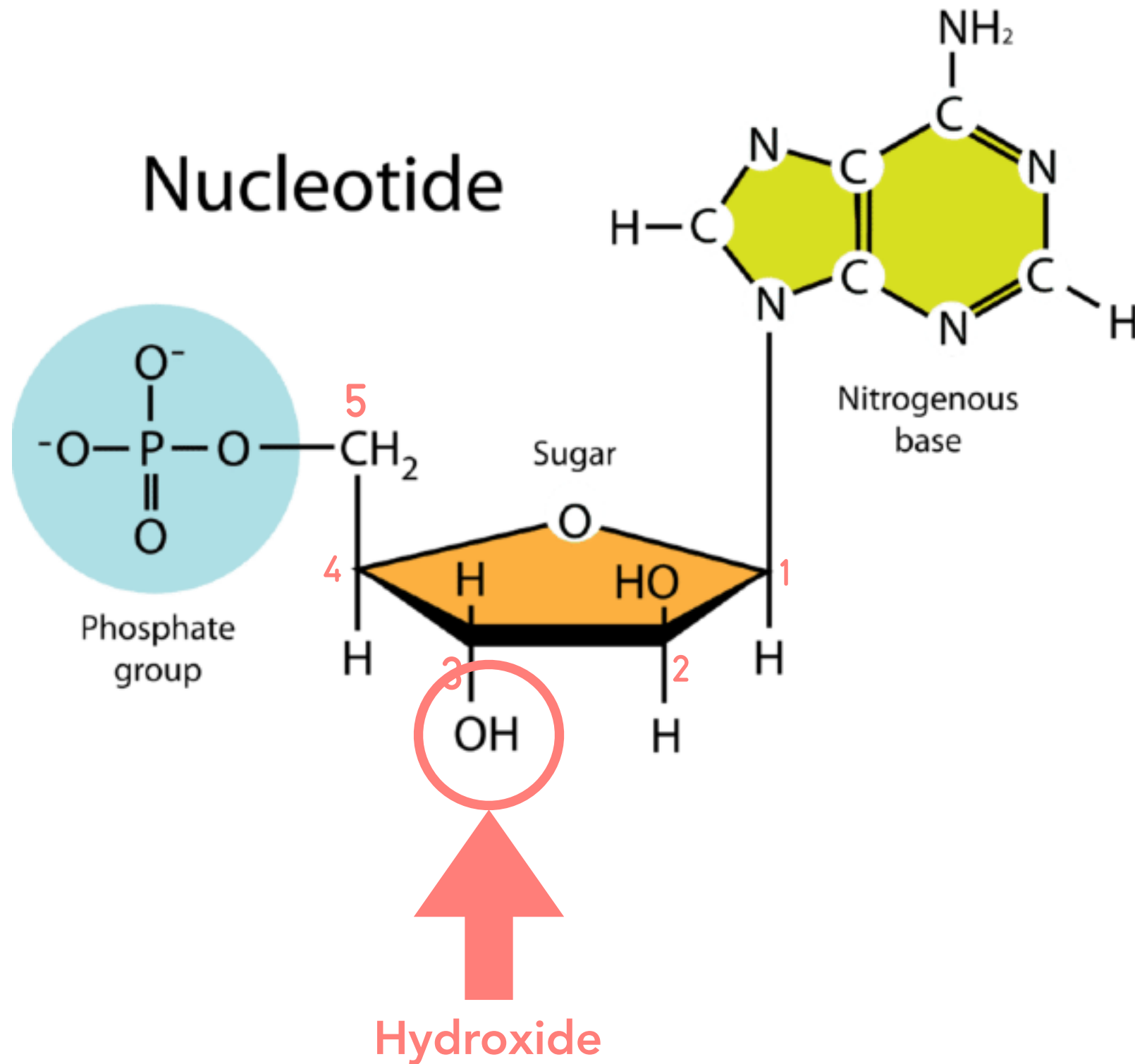


Walter Gilbert

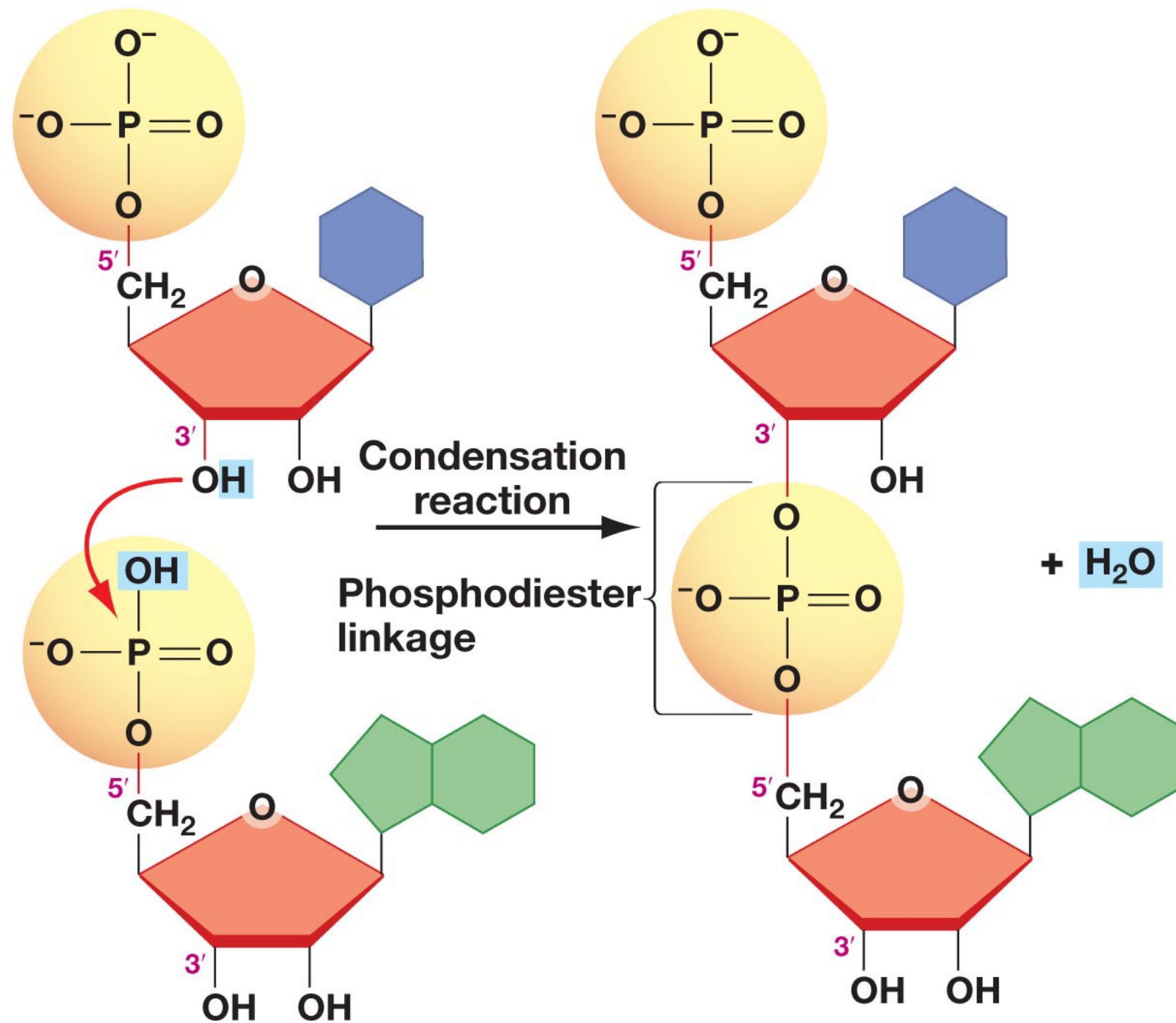


Frederick Sanger

The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg "*for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA*", the other half jointly to Walter Gilbert and Frederick Sanger "*for their contributions concerning the determination of base sequences in nucleic acids*".



PopGen ► Massive Parallel Sequencing



© 2011 Pearson Education, Inc.

DNA

(+) forward strand

5 ' –ATGCGTCACT...TGGCACTGGCA–3 '

3 ' –TACGCAGTGA...ACCGTGACCGT–5 '

(-) reverse strand

Primer Sites

forward primer site

reverse primer site

5 ' – ATGCGTCACT...TGGCACTGGCA – 3 '
3 ' – TACGCAGTGA...ACCGTGACCGT – 5 '

PCR Amplification

Reverse Primer

< ' 3-CGTGACCGT-5 '

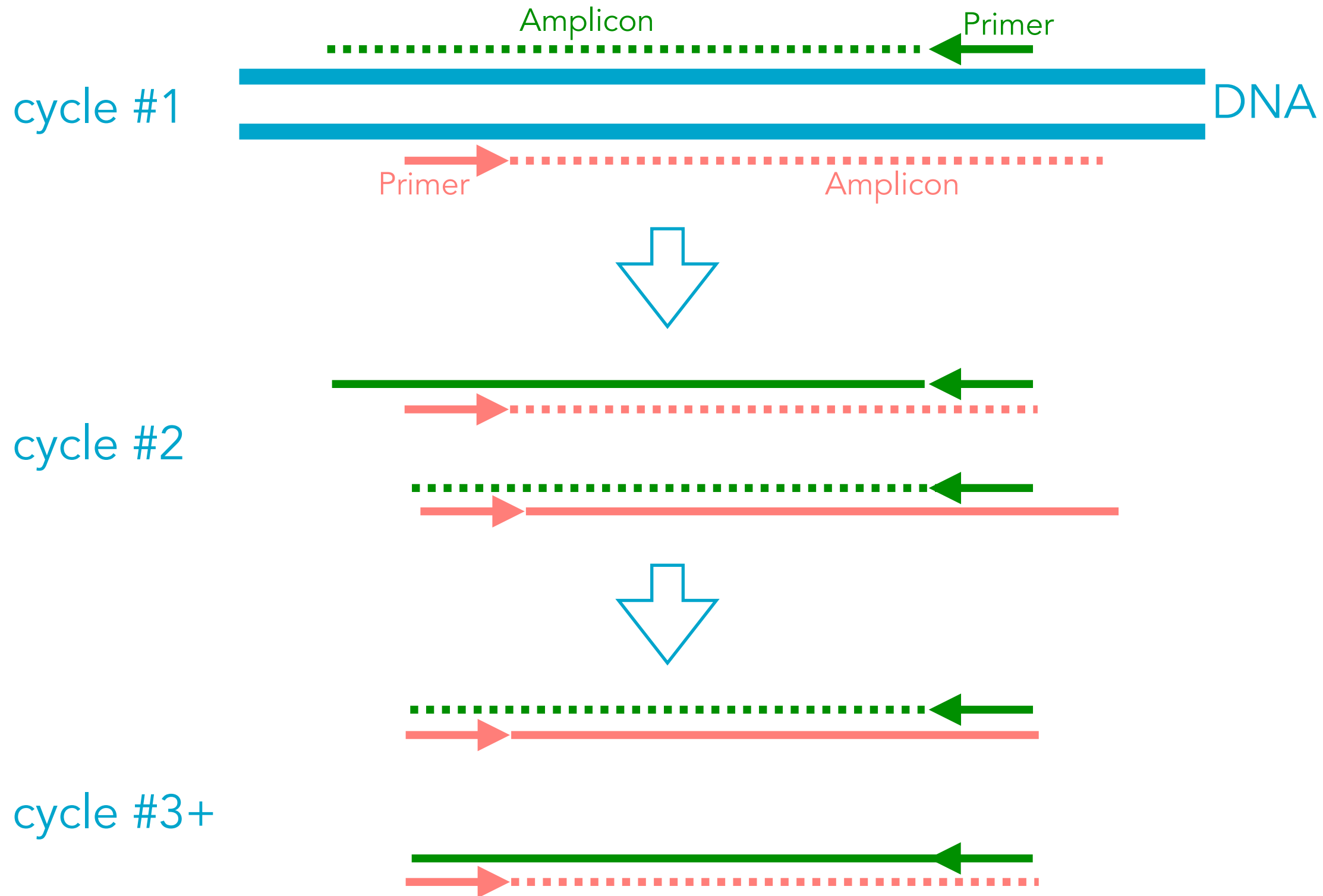
5 ' -ATGCGTCACT...TGCACTGGCA-3 '

3 ' -TACGCAGTGA...ACGTGACCGT-5 '

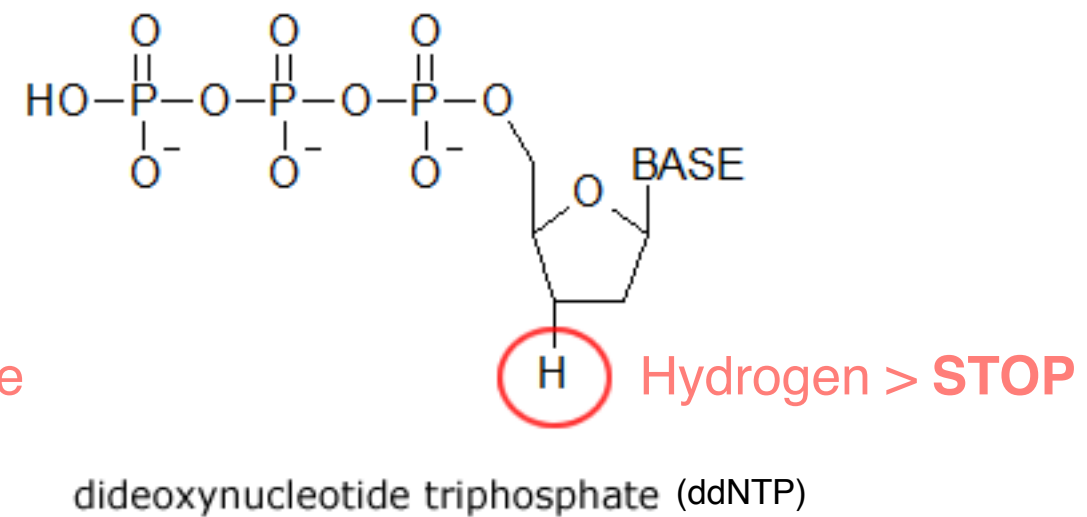
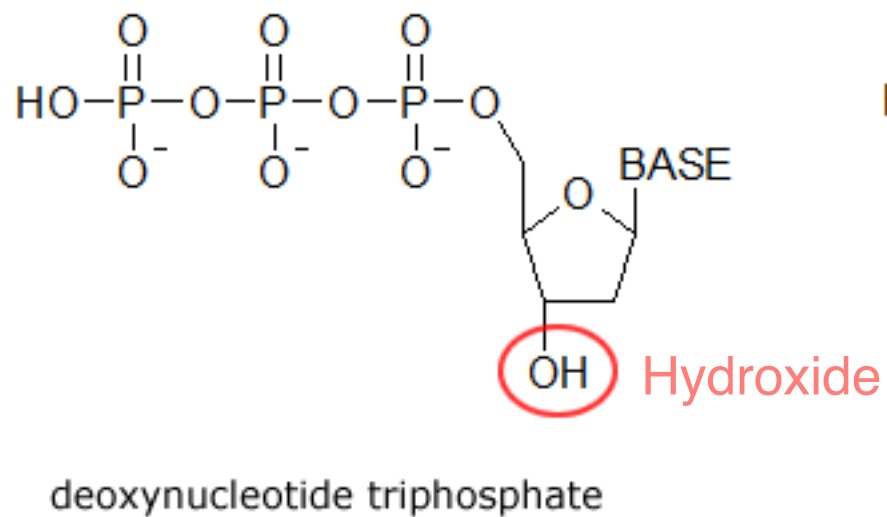
5 ' -ATGCGTCAC-3 ' >

Forward Primer

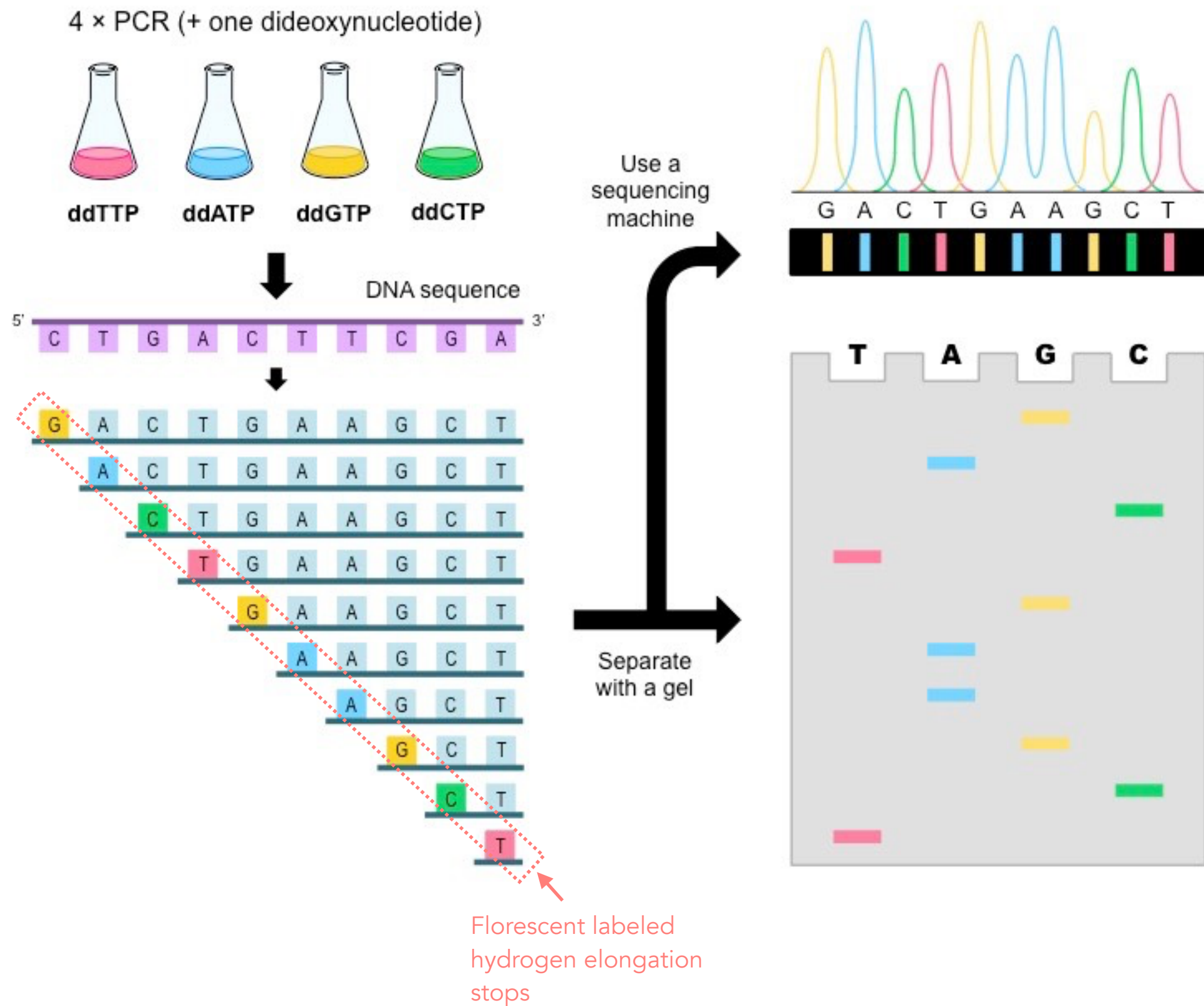
PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing



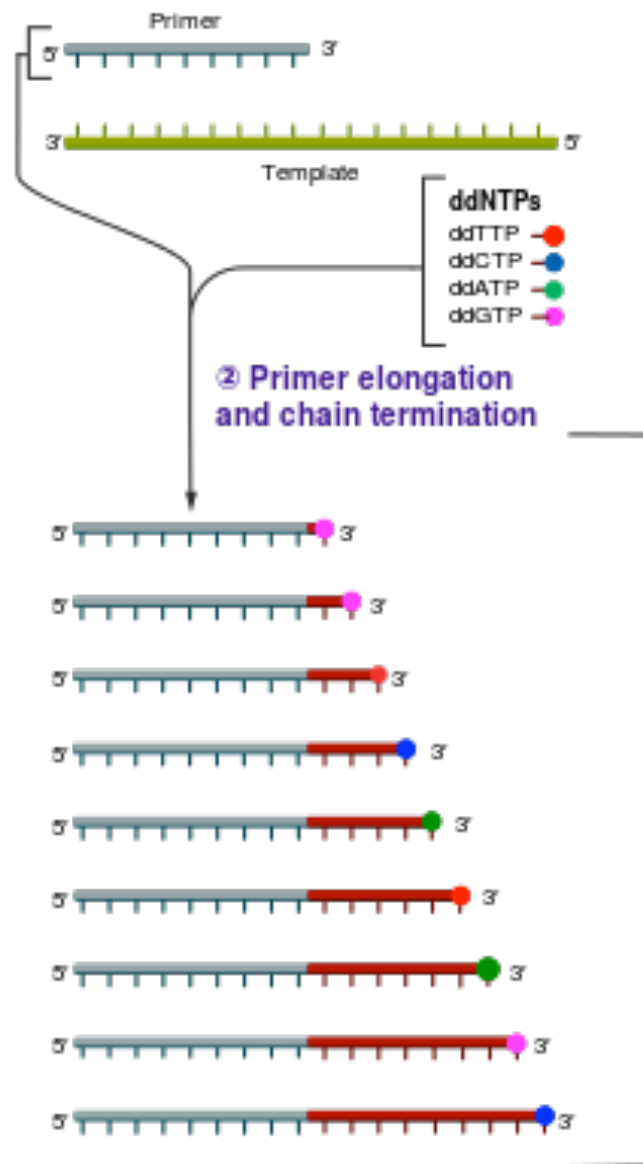
PopGen ► Massive Parallel Sequencing



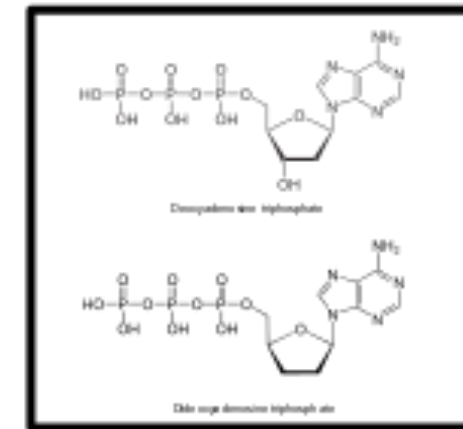
PopGen ► Massive Parallel Sequencing

① Reaction mixture

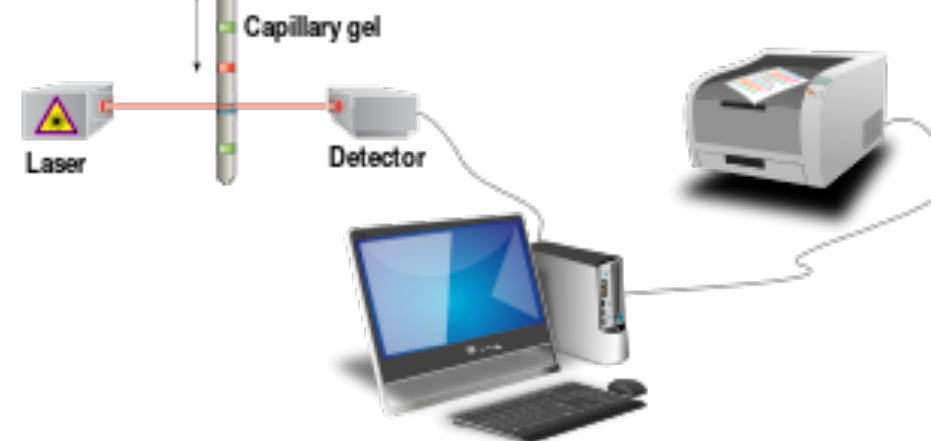
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)



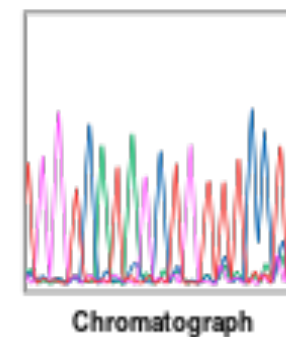
② Primer elongation and chain termination



③ Capillary gel electrophoresis separation of DNA fragments



④ Laser detection of flourochromes and computational sequence analysis



Pyrosequencing



GS Junior

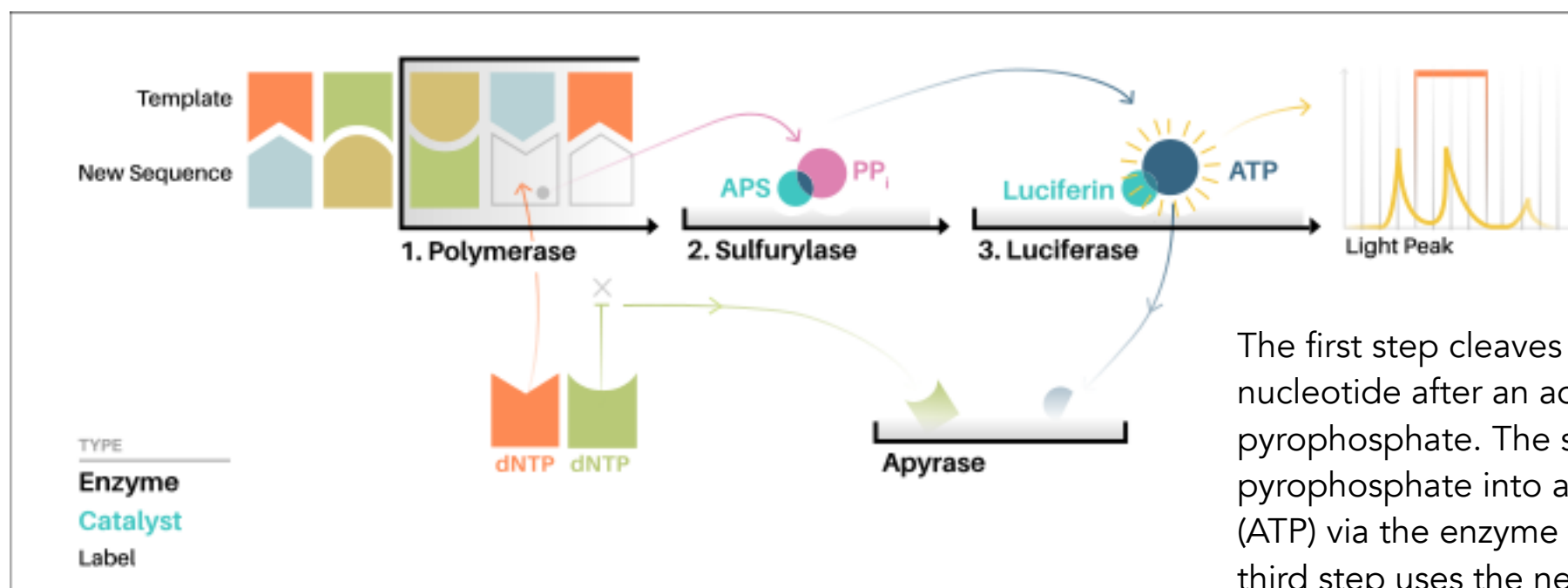


Roche 454



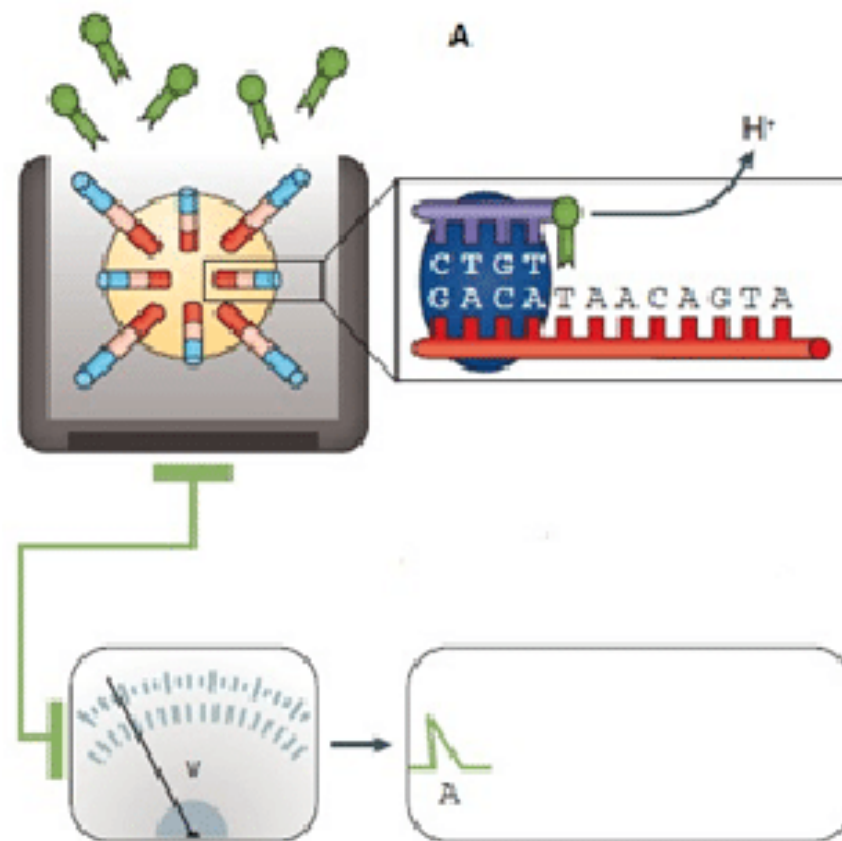
The **PyroMark** uses Pyrosequencing technology for real-time, sequence-based detection and quantification of sequence variants and epigenetic methylation. The PyroMark Q24 is highly suited for the analysis of CpG methylation, SNPs, insertion/deletions, STRs, variable gene copy number, as well as for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

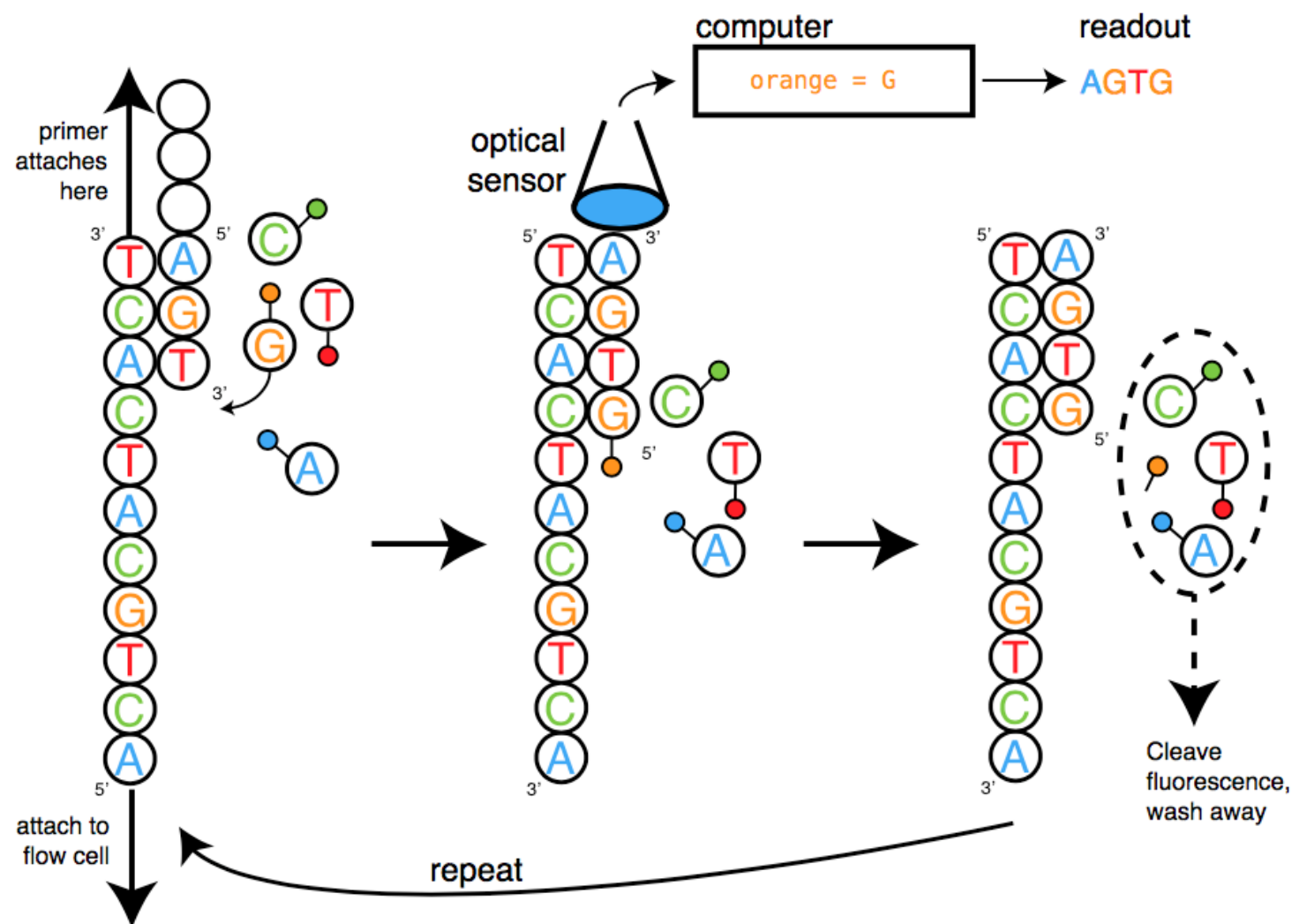


Illumina Sequencing by Synthesis (fluorescent)

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

<http://www.illumina.com>

Sequencing by Synthesis (SBS) (fluorescent)



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

Singular Genomics (fluorescent - rapid SBS)

G4

A highly versatile benchtop sequencer targeting applications where accuracy, speed, flexibility, and scale matter most



Key milestones

- Launched 2021
- First units expected ship in Q2 2022

PX

Integrated in situ platform for multiomic analysis in single cells and tissues



Key milestones

- Early access program to begin 2022
- Commercial launch expected in 2023

PX image is for illustrative purposes only

Element Biosciences (fluorescent - sequencing by avidity)

A benchtop sequencer with
unprecedented performance,
cost, and flexibility.

Meet AVITI



Performance

%Q30 > 90 at 2x150
1B reads/flow cell*
600 Gb+ output/run

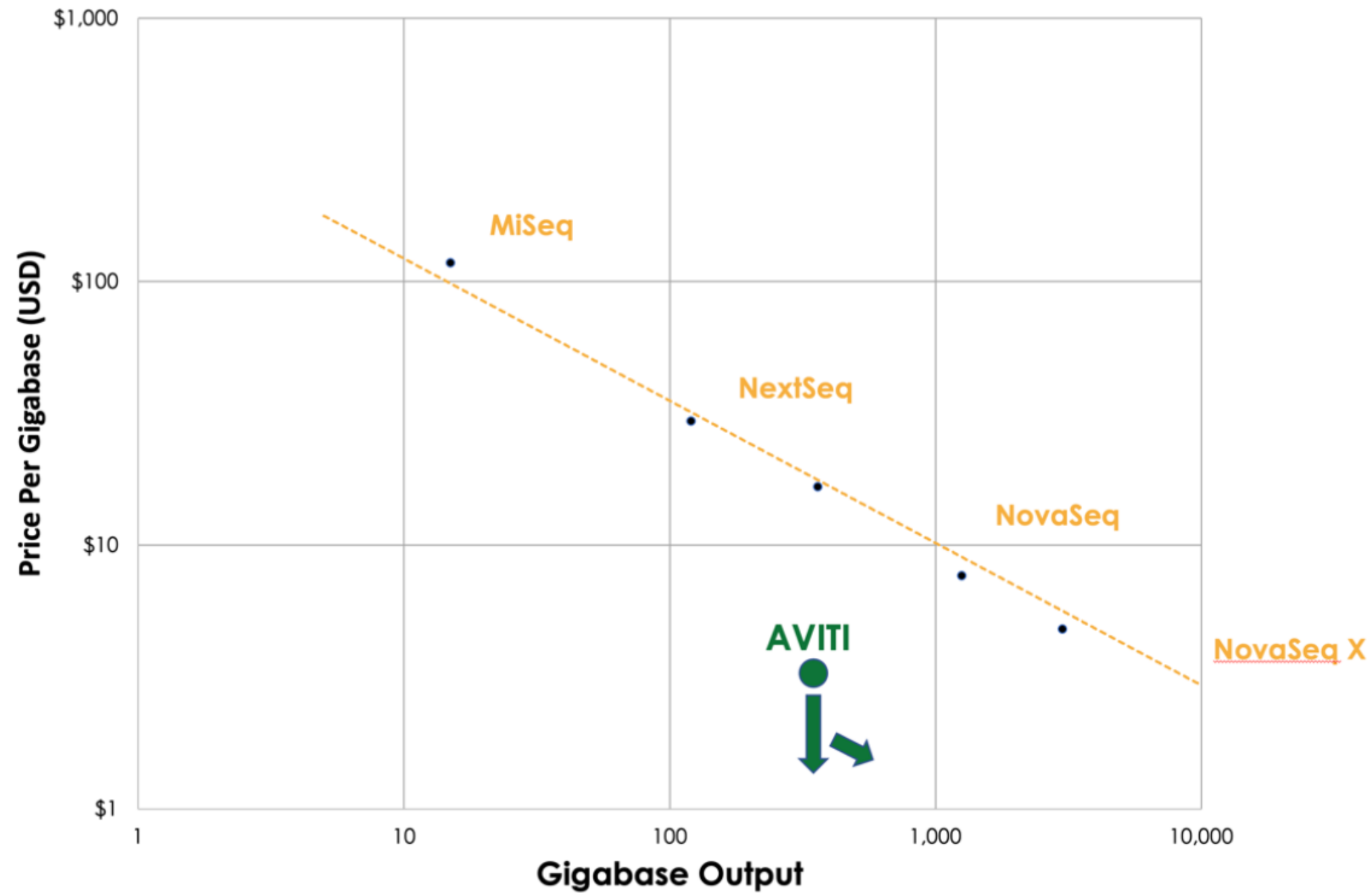
Flexibility

Dual flow cells
Flexible start
Tunable read throughput

Cost

\$289K/instrument
\$1680/300 cycles
\$1080/150 cycles
Leasing/financing options
available

PopGen ► Massive Parallel Sequencing



Pacific Biosciences

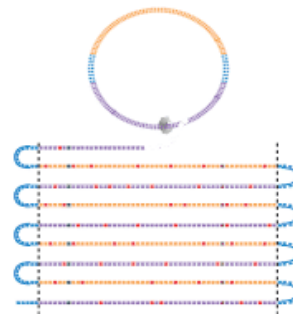
Over a decade of on market technological innovation



RS I /RS II
75k/150k ZMW



Sequel
1M ZMW



HiFi



Sequel II/IIe
8M ZMW



Revio
25M x4 ZMW

2011-2013: PacBio launches groundbreaking single-molecule sequencing platform; named top innovation by *The Scientist* magazine

2015: PacBio launches the Sequel system

2018: HiFi enables ultra-high accuracy at long read lengths and establishes PacBio as leader in sequencing accuracy

2019 | 2020: PacBio launches Sequel II and Sequel IIe

2023: PacBio expected to launch Revio enabling the sub-\$1k long-read genome at scale

Increased throughput **>10,000-fold**

Increased read length **>100-fold**

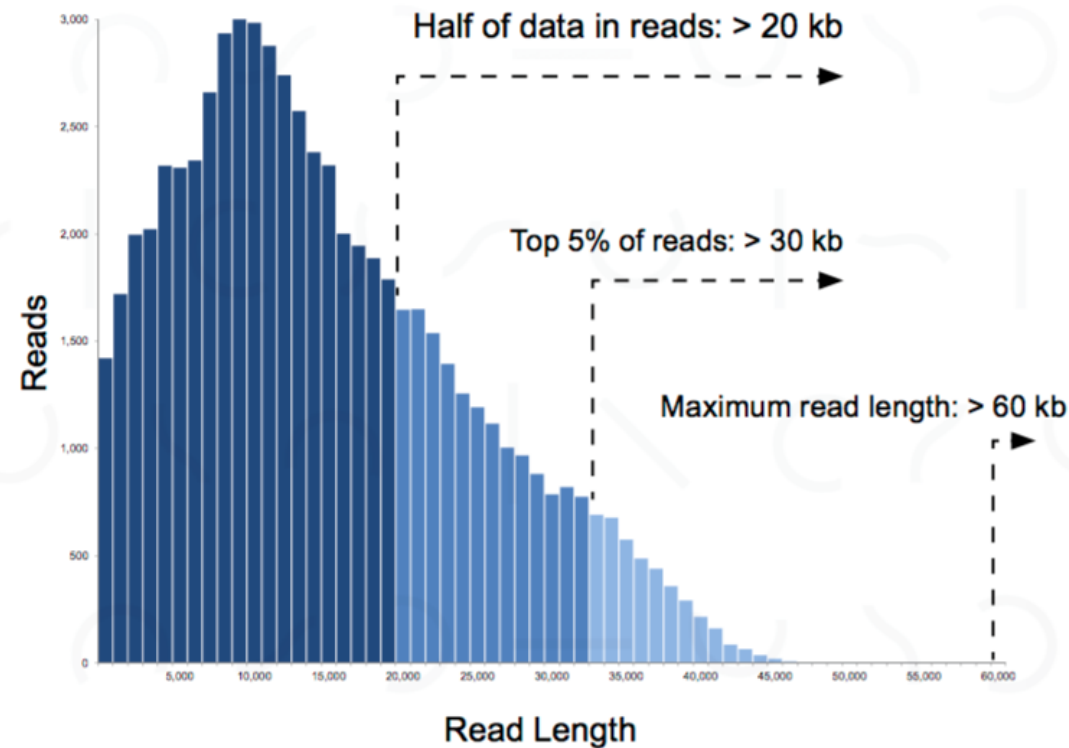
PacBio (fluorophore)



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.



PopGen ► Massive Parallel Sequencing

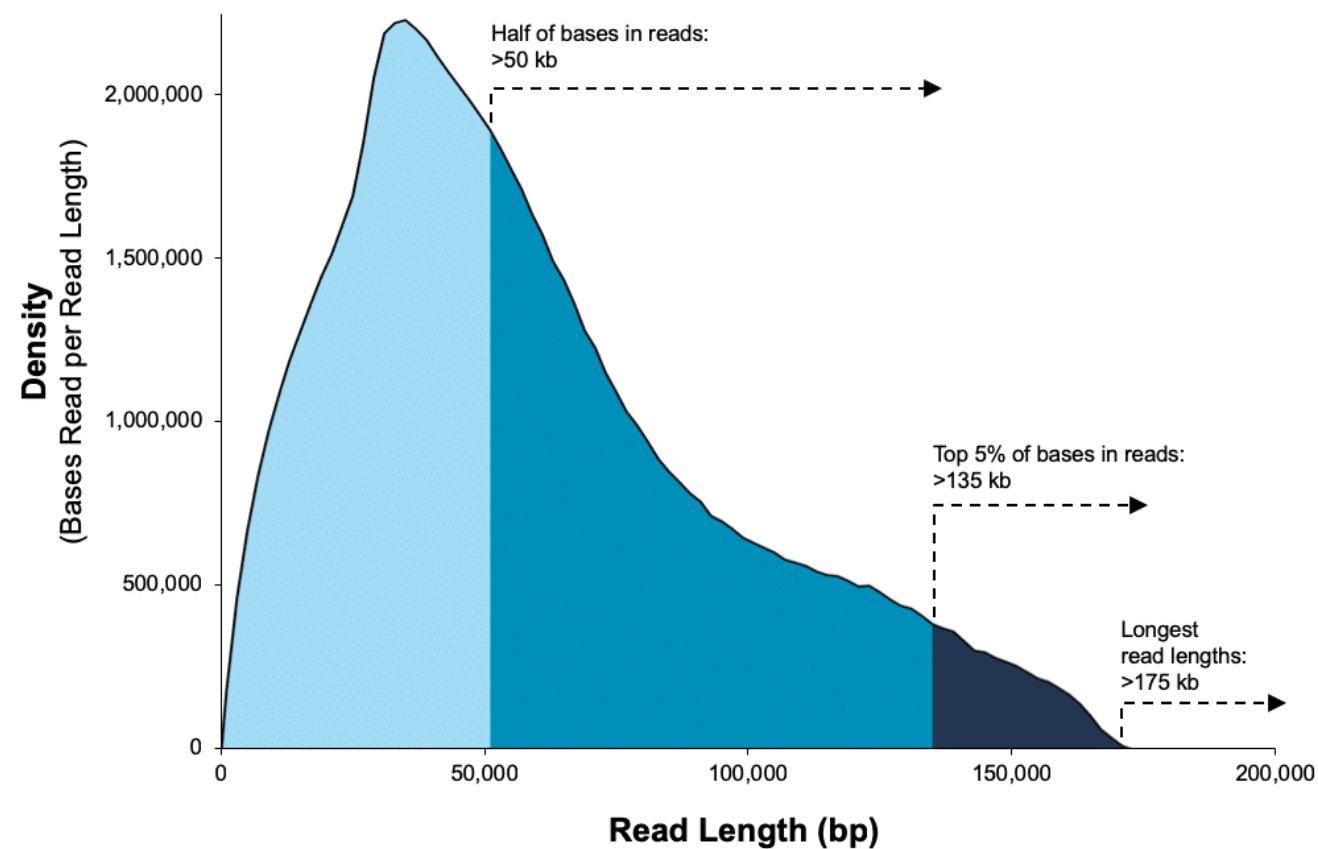


(Pacific Biosciences)

<https://www.pacb.com>

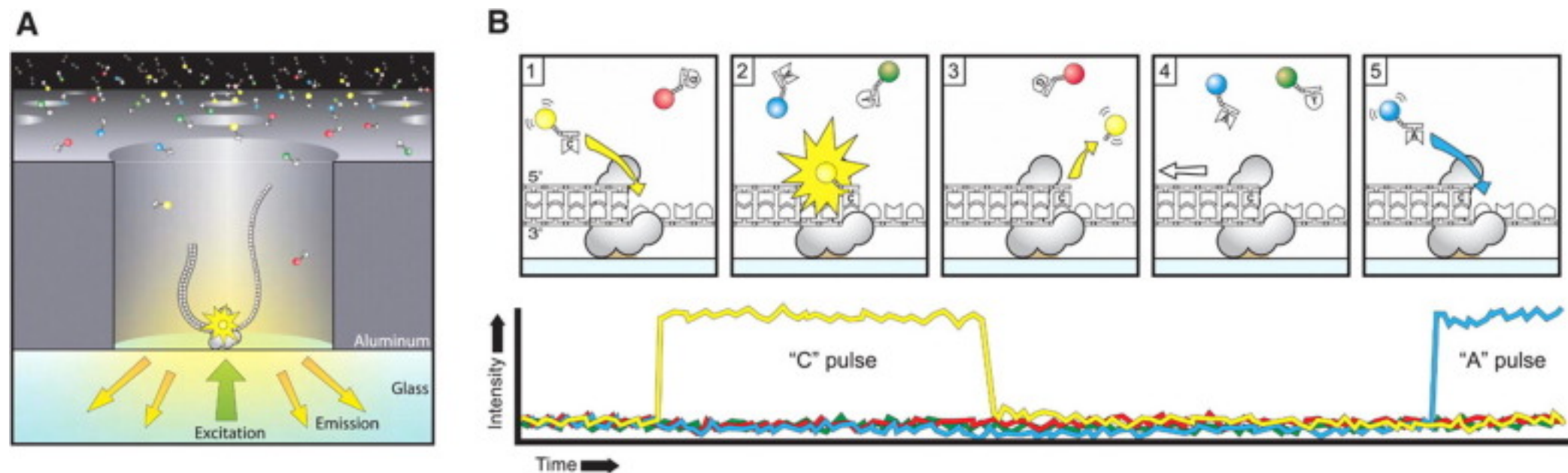


Sequel

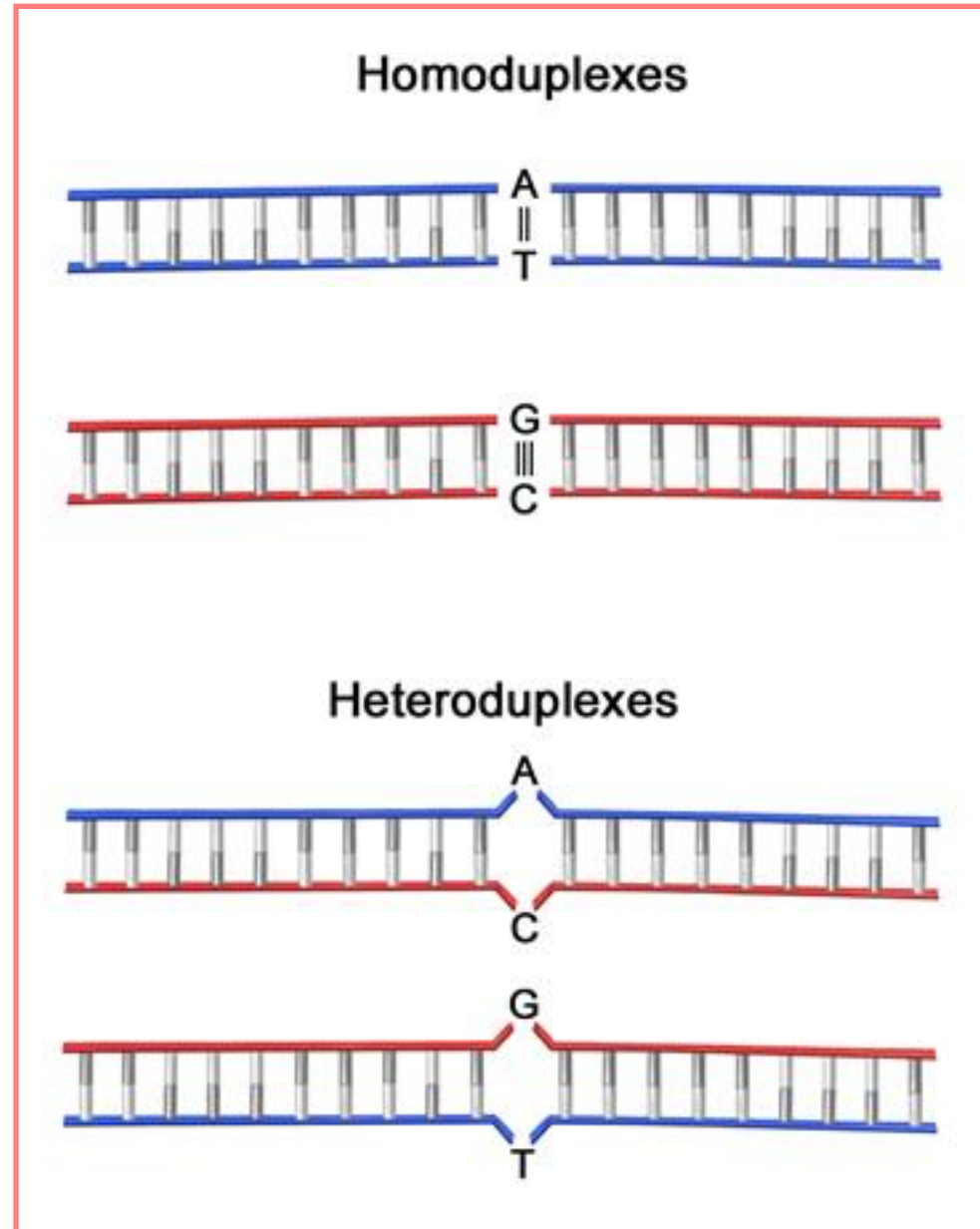
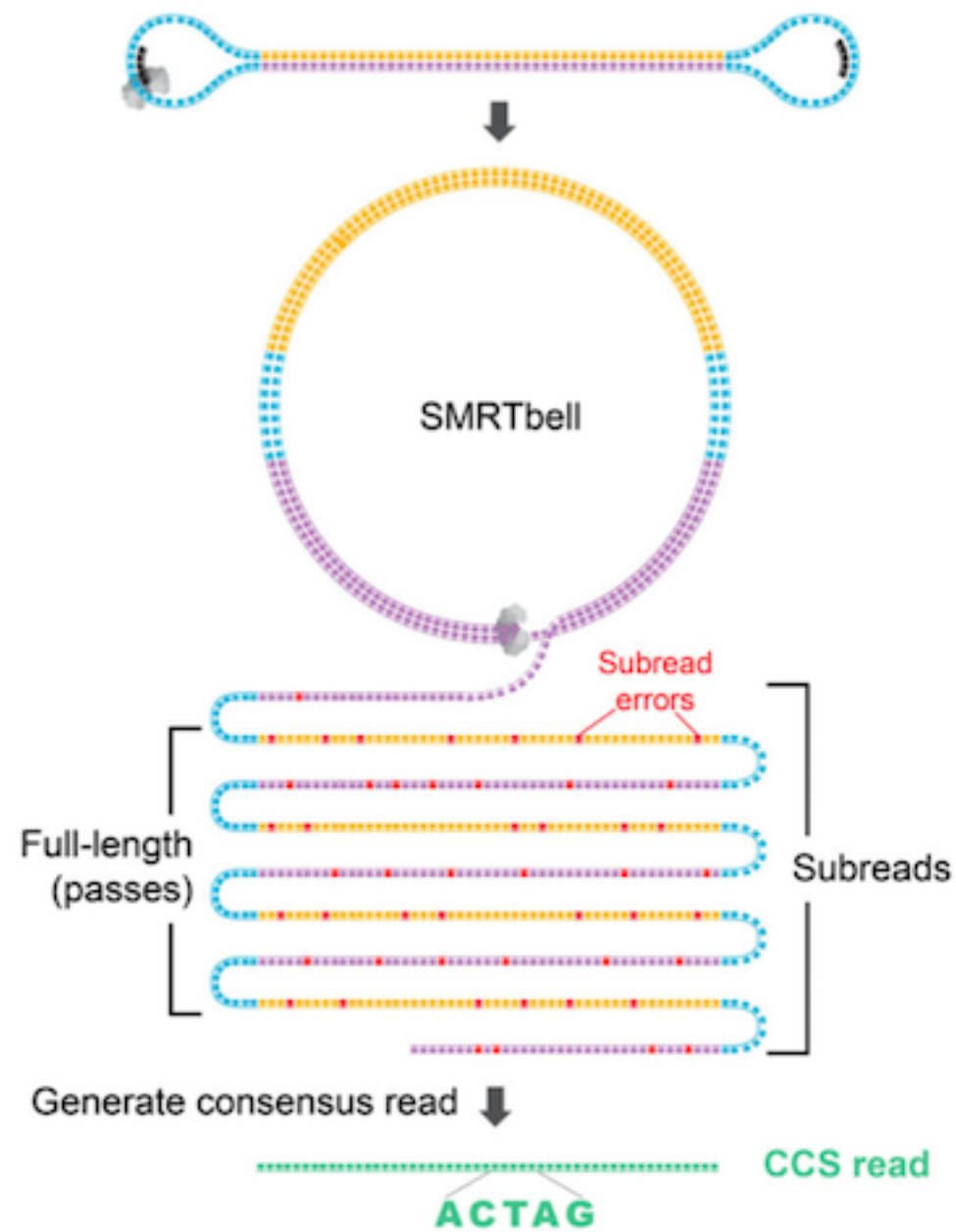


Data from a 35 kb size-selected *E. coli* library using the SMRTbell Express Template Prep Kit 2.0 on a Sequel II System (1.0 Chemistry, Sequel II System Software v7.0, 15-hour movie)*.

PacBio (fluorophore)



PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing



The **Onso** system is an advanced benchtop platform for short-read DNA sequencing, known for achieving exceptionally high accuracy through PacBio's sequencing by binding (SBB) technology. It generates over 100 Gb of ultra-high-quality Q40+ data, empowering researchers to surpass traditional detection limits.



Vega provides individual labs with 60 Gb of long-read data per run, offering best-in-class, highly accurate HiFi reads. This platform enables comprehensive, high-resolution sequencing suited for complex analyses, such as de novo genome assembly, transcript isoform detection, and full epigenetic profiling, giving researchers a robust tool to explore intricate biological questions with confidence.

PopGen ► Massive Parallel Sequencing



SmidgION



Flongle



MinION



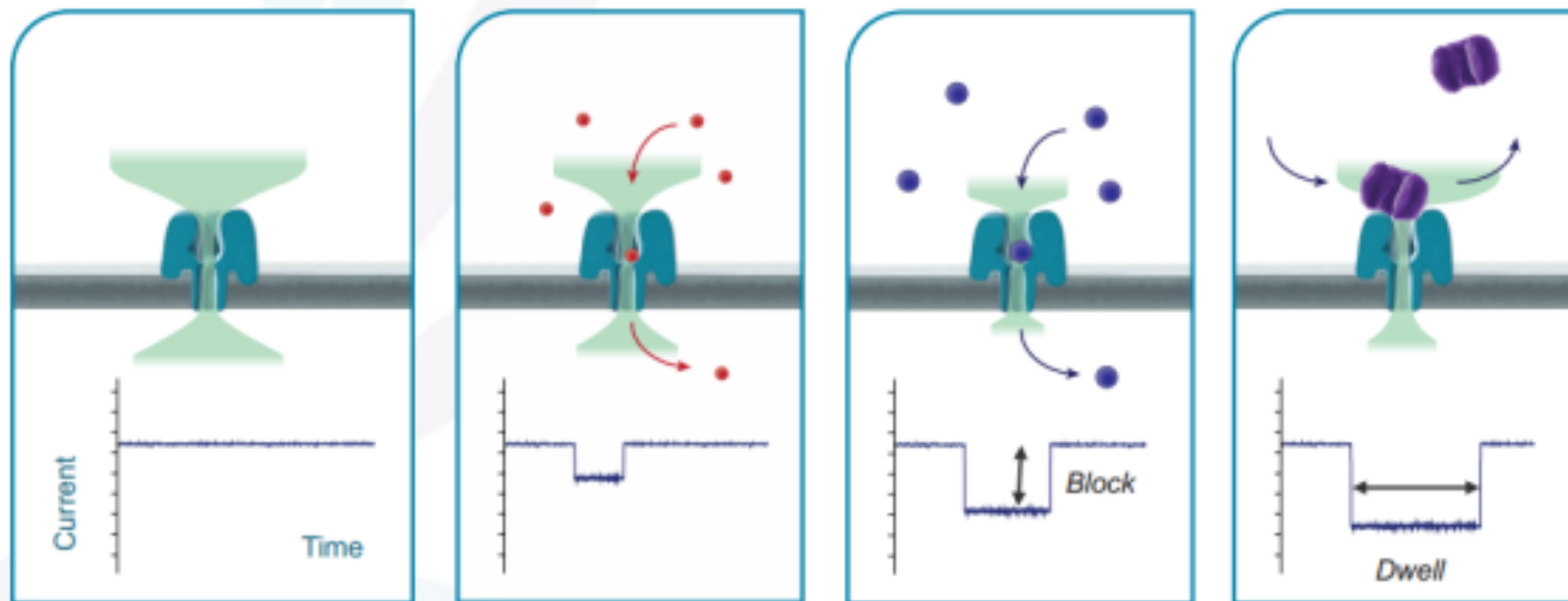
GridION



PromethION

<https://www.nanoporetech.com>

PopGen ► Massive Parallel Sequencing



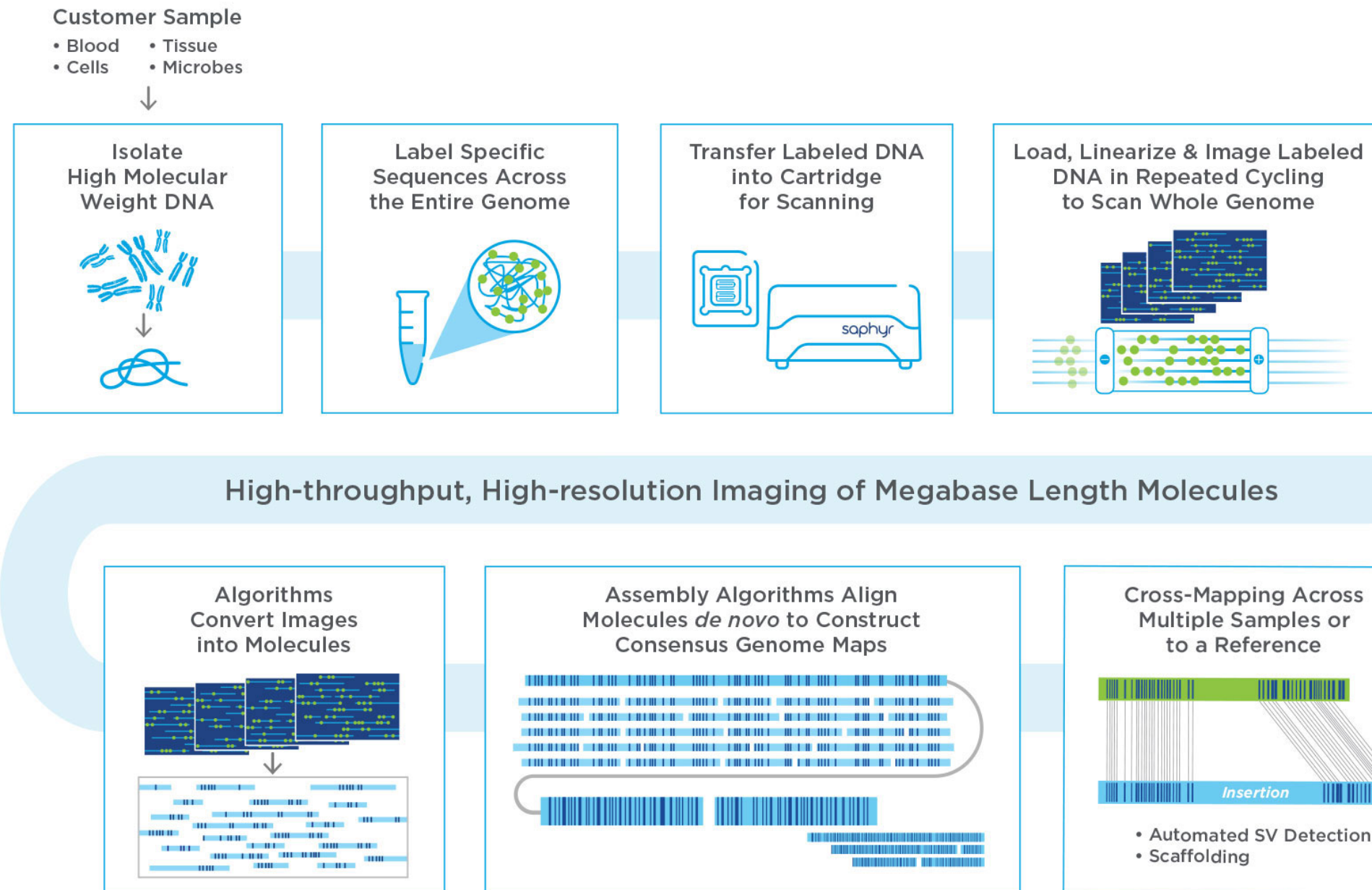
PopGen ► Massive Parallel Sequencing



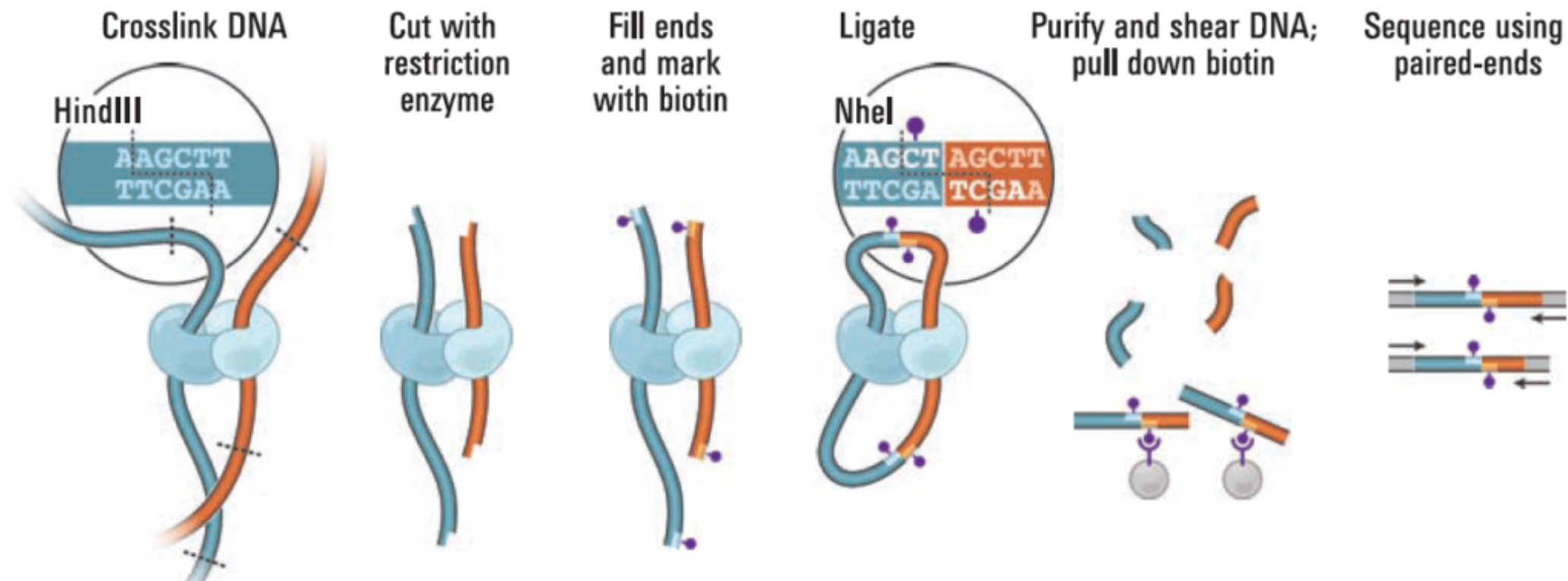
PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing



Proximity ligation methods connect sequence and structure



Source: Lieberman-Aiden et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289-293.

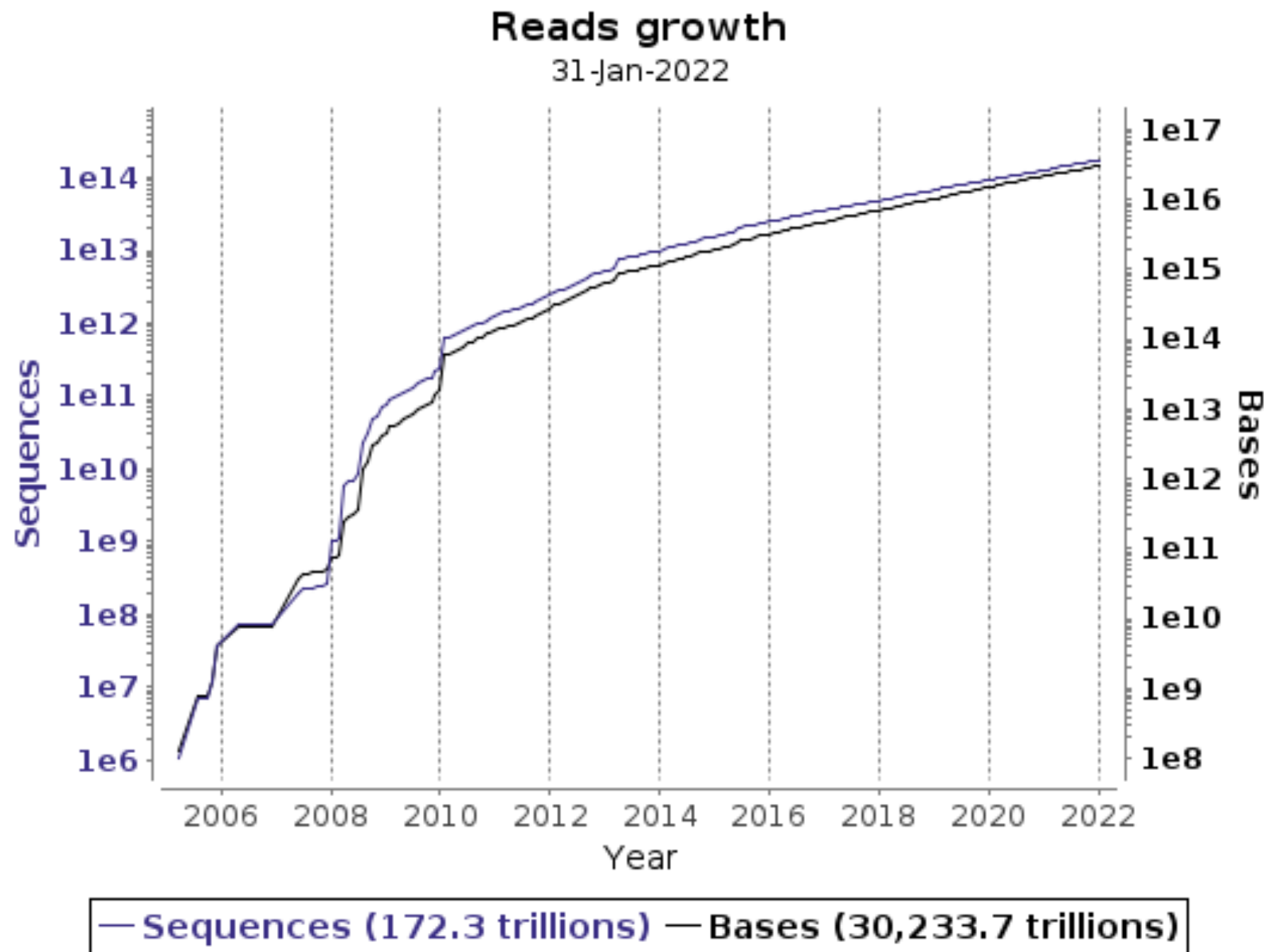
Hi-C sequencing, which captures the 3D organisation of chromatin, enables genome scaffolding by linking DNA sequences that are in close physical proximity. For example, Dovetail's Hi-C solutions work well with non-model organisms, helping to assemble chromosome-scale genomes without the need for high-molecular-weight DNA or species-specific protocols.

PopGen ► Massive Parallel Sequencing



MinION Mk1C

PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing

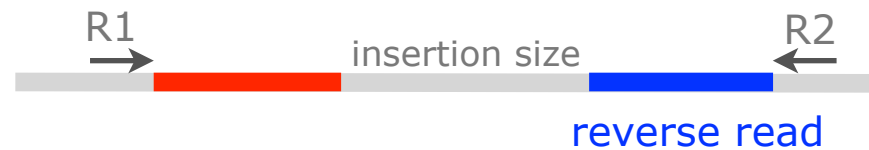


Sequencing Reads

PopGen ► Massive Parallel Sequencing



Single Reads (SR)



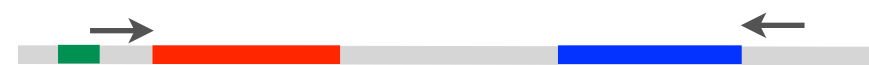
Paired-End (PE) Reads



Overlapping Paired-End (PE) Reads



Single Reads (SR) with Index



Paired-End (PE) Reads with Index



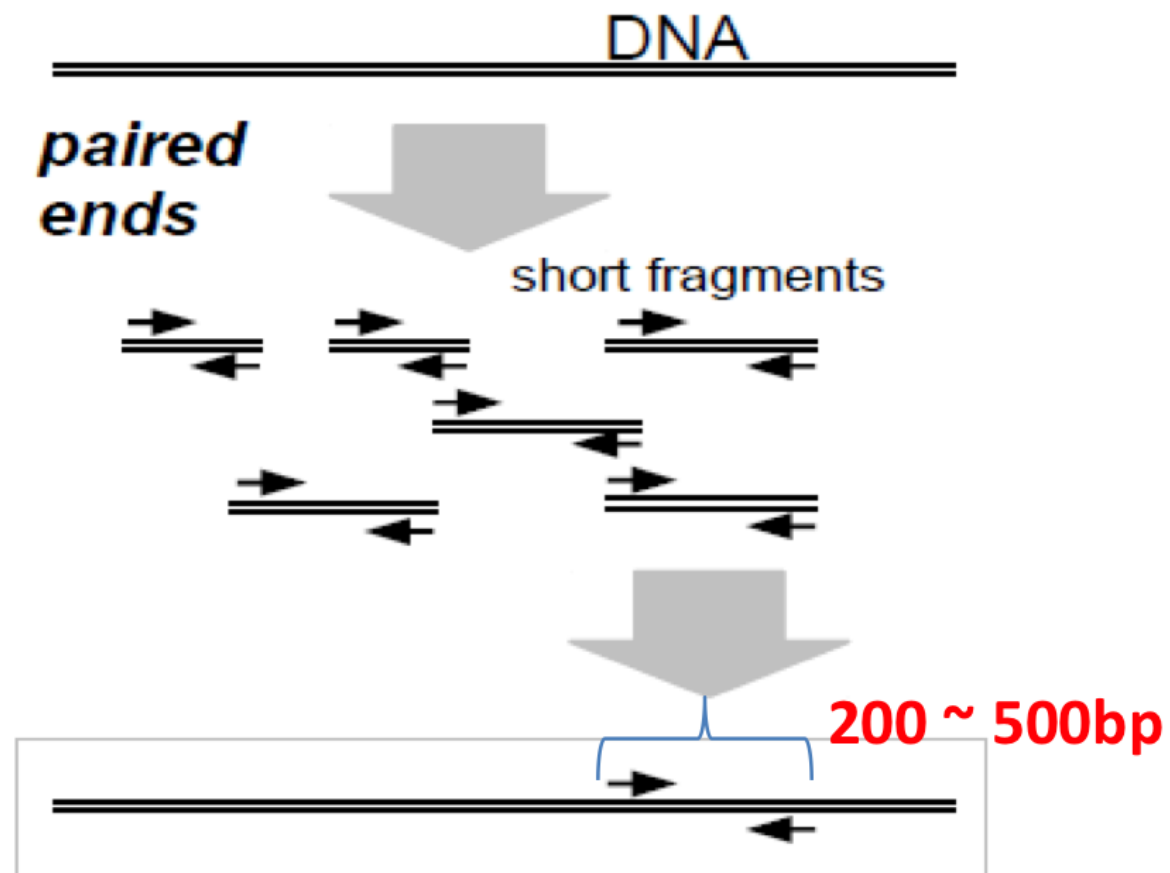
Paired-End (PE) Reads with Dual Indexing



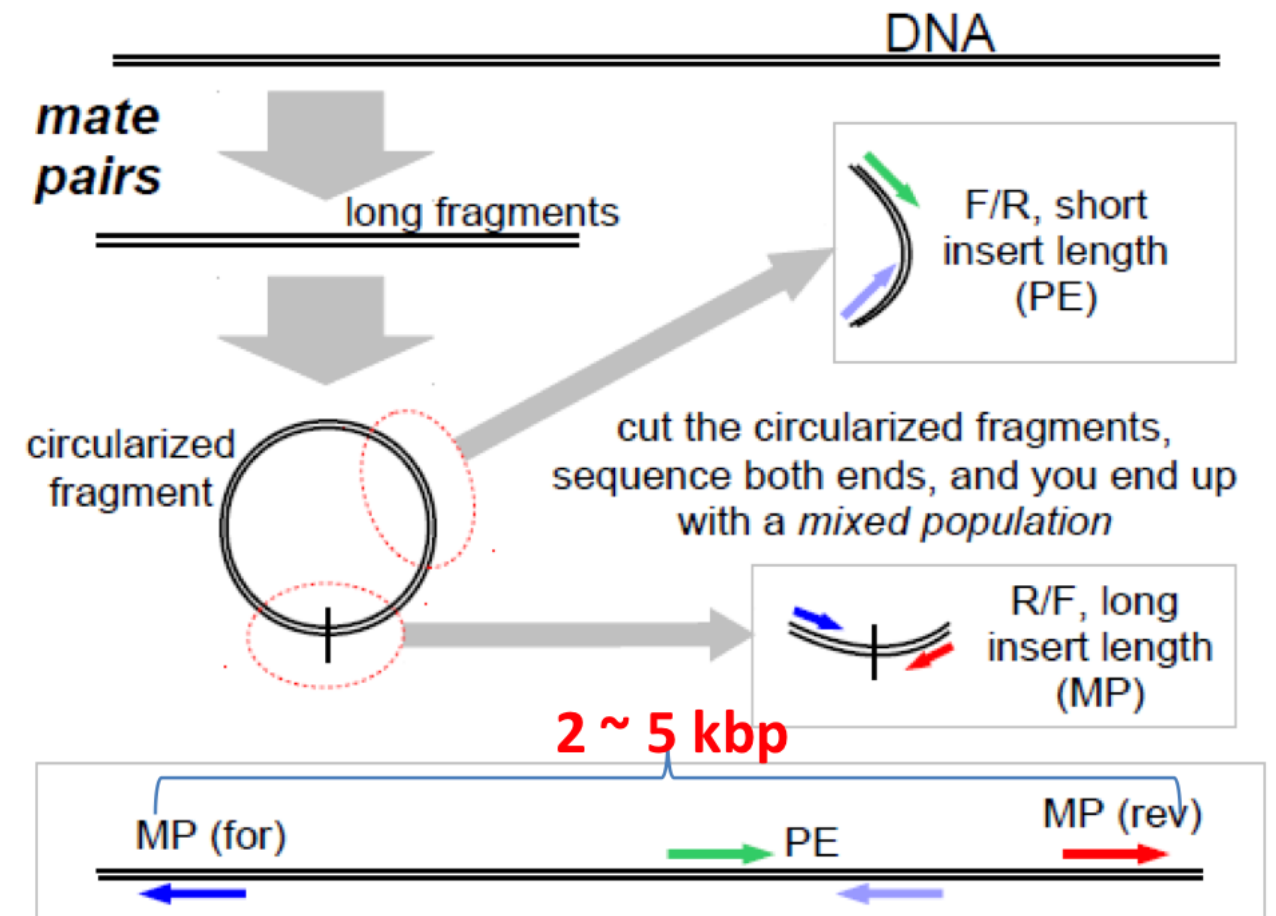
Extended Single Reads (SR) with Index

PopGen ► Massive Parallel Sequencing

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 - AAAATACCGCATGATTGTTTCGTTTCACAAACAAAGATATAGCTTGCCAGATAGCGTATGCCAGATTGCAA  
3 - GGAGATGTGATCATTGTGCAGCTTATGCTCATGAACTCCCAAGATATGGTGTCAAGGTCGGGTTGACCA  
4 - ACTATGCAGCTGCTTATTGCACTGGCCTCTTGCTCGCAAGAAGGCTCCTTTCAAAATTGAAATTGGCTGA  
5 - CACTTACAAAGGTTGTGAAGAAGTGAATGGTGATGAATACCTTGTGGAAGGAGAGGGACAGCCTGGA  
6 - CCTTTCCGTTGTTACCTTGATATTGGCCTTGCCAGAACCTCAACTGGTGCCAAGATCTTTGGTGCATTGA  
7 - AAGGTGCAGTTGATGGTGGACTTGACATCCACACAGCAACACGAGATTCCCTGGTTATGACAATGAAGC  
8 - AAAGGAATTTGACCCAGAGGTGCACAGACAACA...  
...
```

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 - BCCFFFFDFHHHFJJJJJJJJIIJJJJIIJJGIHHHHHHFFFDDDEDDDC
   @HWI-ST486:166:C06K9ACXX:7:1101:2519:1936 1:N:0:ACAGTG
   NTGCACACGAATTCGCCGTTGTGGCAGCTCGAGTACCTGTGGTTCGCCGAG
   +
   #1=DDDFHFFHFFIGIIJJJJIIJJIIJ;?*?.?/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. Fields², Naohisa Goto³, Michael L. Heuer⁴ and Peter M. Rice⁵

ASCII encoded quality scores

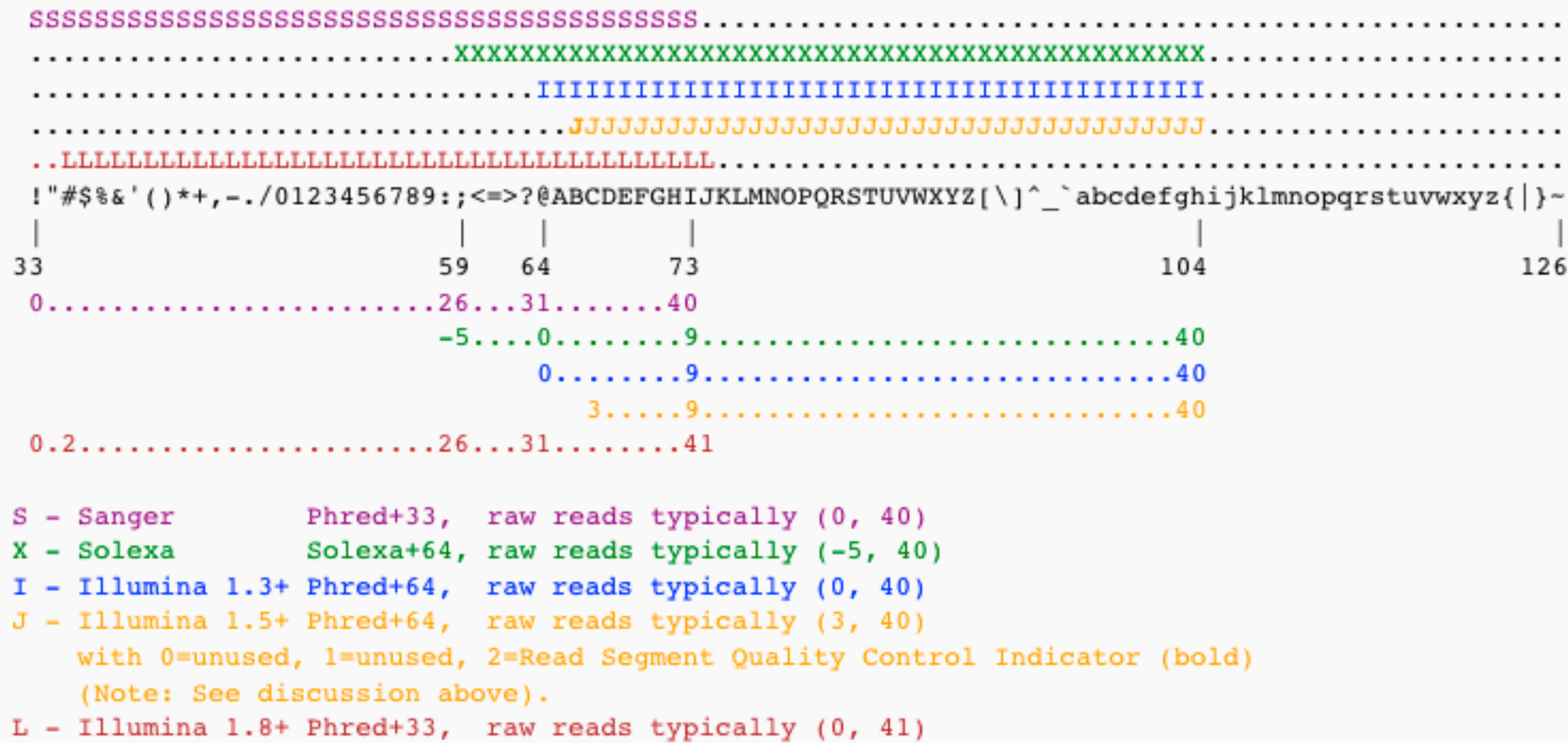
```
@HWI-ST741_0085:1:1101:1444:1939#0/1
ATAGTTACAATCGATCCATTTGCAGAGTACAGATACATGATACGGGAAT
+HWI-ST741_0085:1:1101:1444:1939#0/1
ffffdfdfdfdfdfgggfaafffcdfcfffbfdddeaegfgfgafaffW^a]
```

HiSeq score

39 33 38 33 38 38 23 30 33 29

Phred score

PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing

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'Amore³, Umer Z. Ijaz⁴, Neil Hall³ and Christopher Quince⁵

Abstract

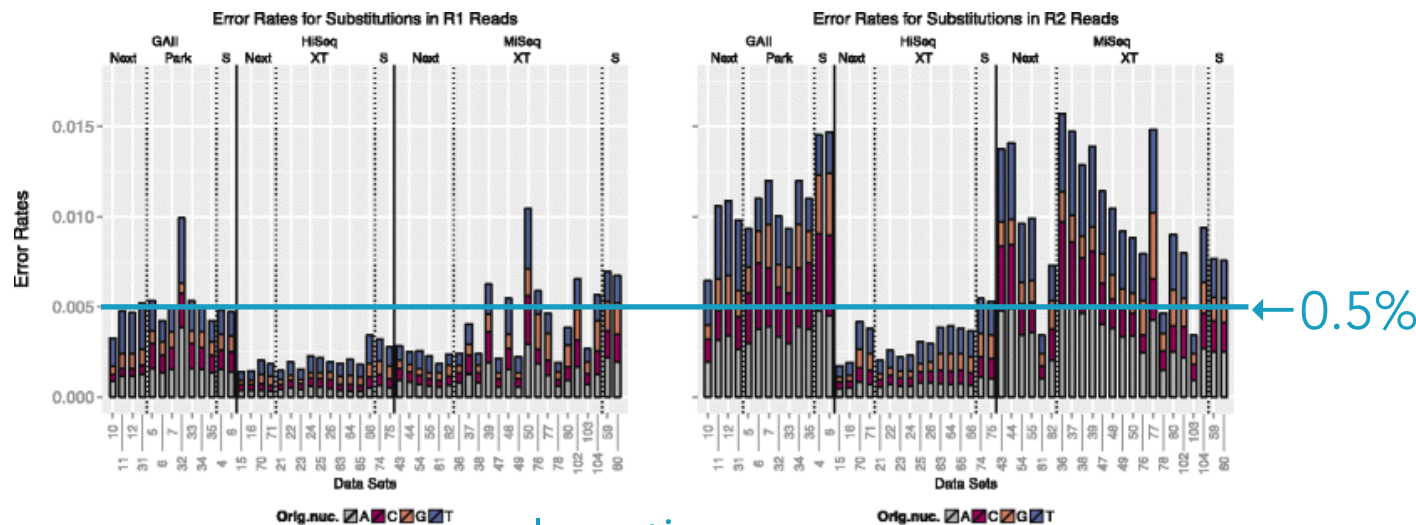
Background: Illumina's sequencing platforms are currently the most utilised sequencing systems worldwide. The technology has rapidly evolved over recent years and provides high throughput at low costs with increasing read-lengths and true paired-end reads. However, data from any sequencing technology 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.

PopGen ► Massive Parallel Sequencing

Substitutions

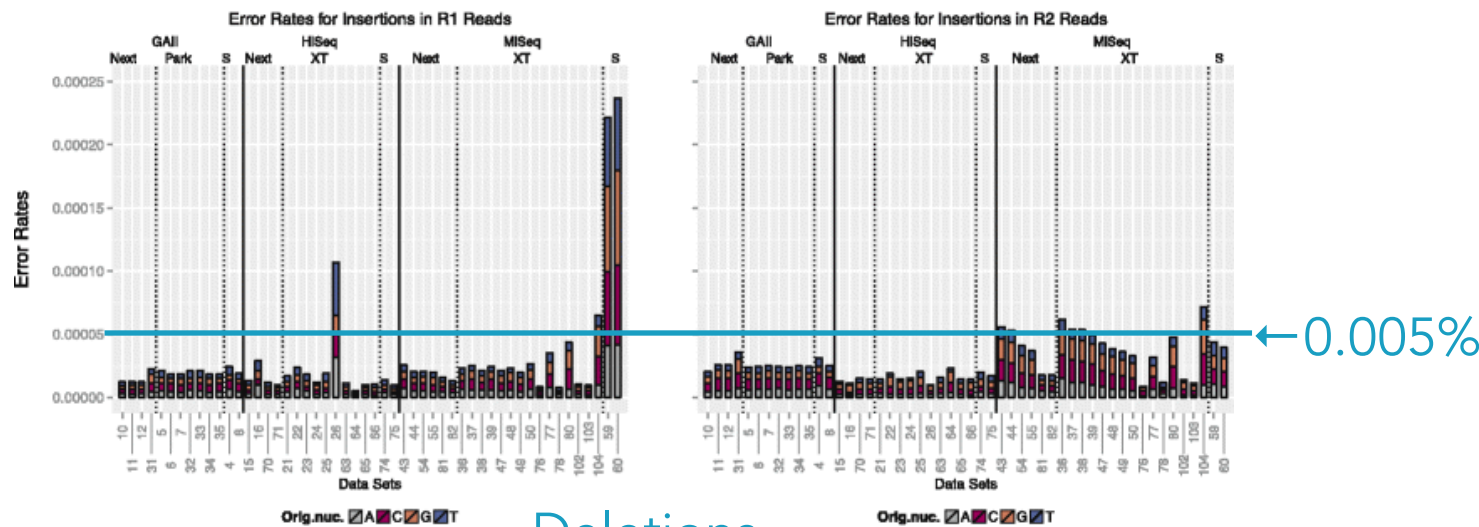


Illumina

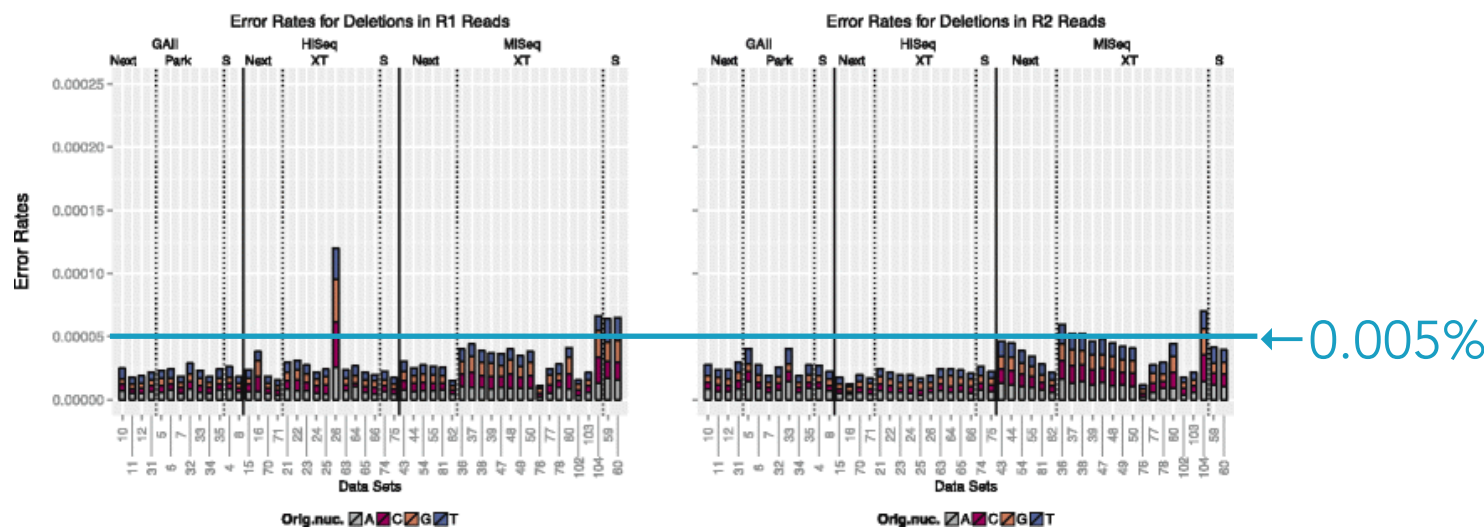
Average substitution rates

Platform	R1/R2	A	C	G	T
GAI	R1	0.0015	0.0010	0.0008	0.0018
GAI	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

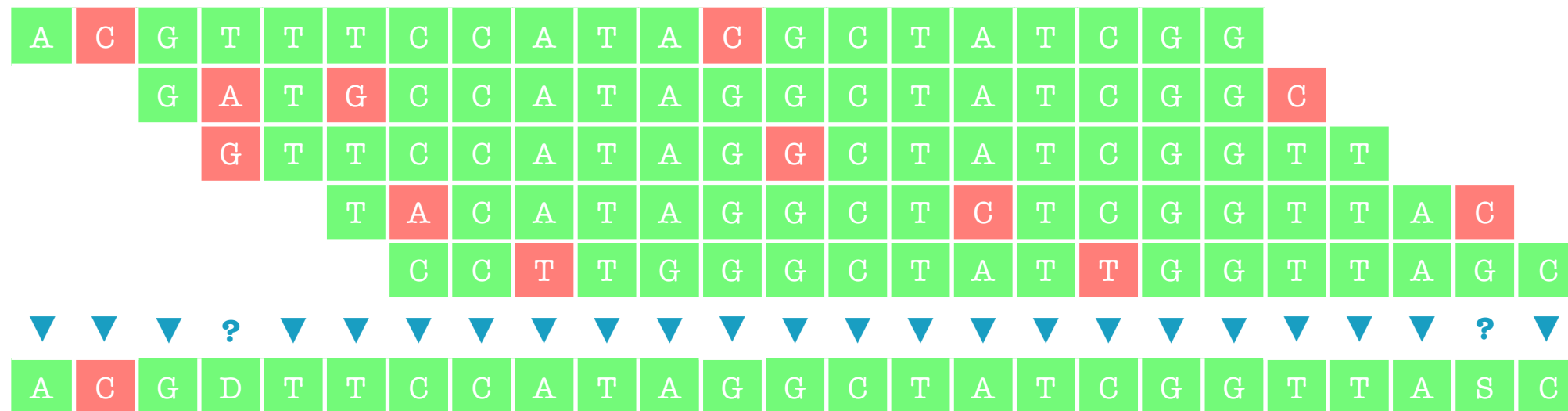
Insertions



Deletions



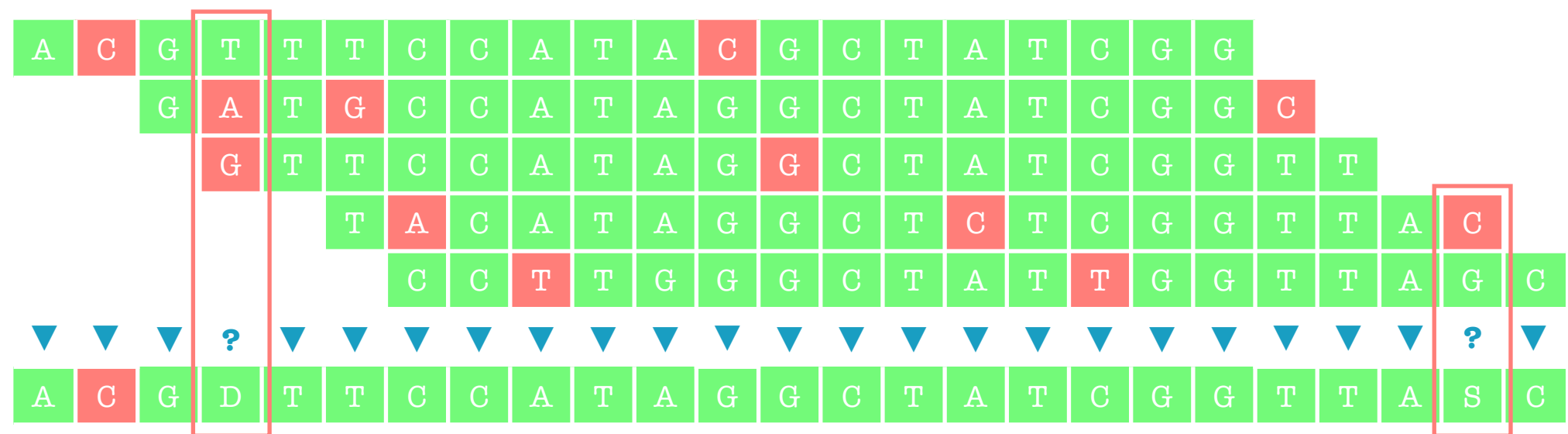
Error Correction



Read quality

Number of reads (coverage)

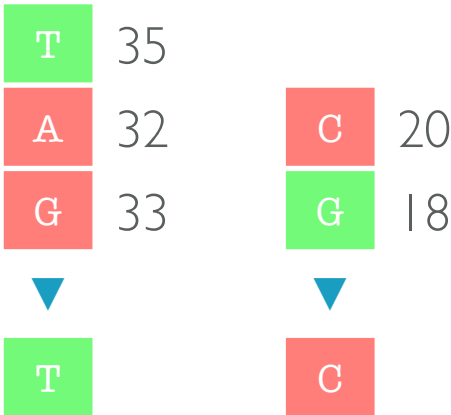
Error Correction



Read quality

Number of reads (coverage)

Phred score



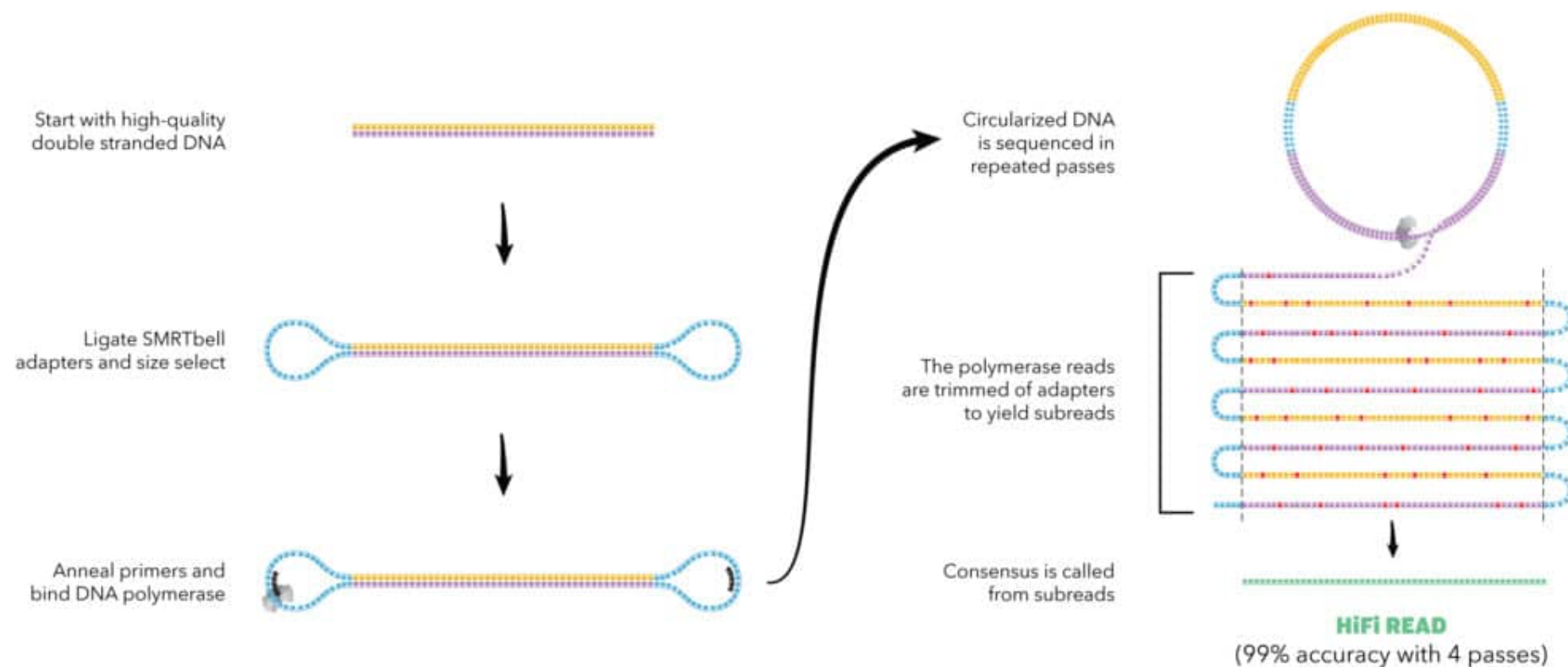


Error Rate

BAM → FASTQ

BAM → CCS.FASTX

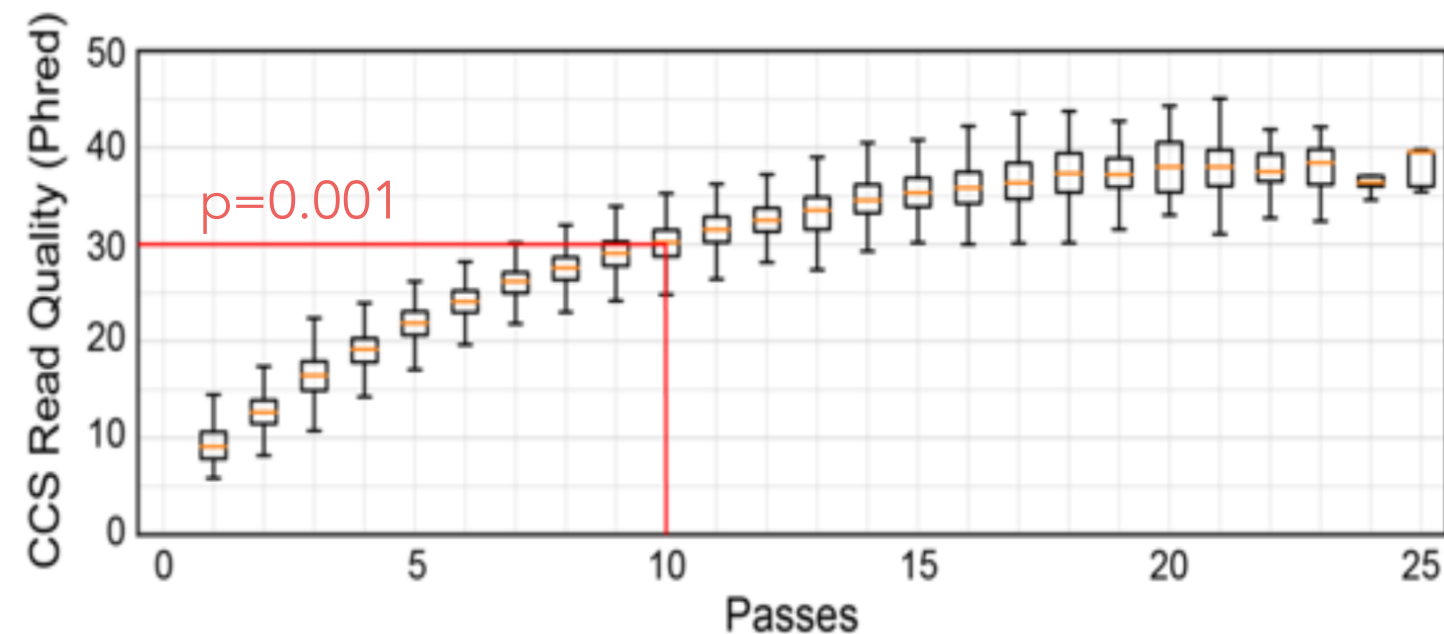
Circular Consensus Sequences (CCS)



Advanced scientific discoveries require sequencing data that is both accurate and complete. Single Molecule, Real-Time (SMRT) Sequencing technology has evolved to a different type of long read, known as highly accurate long reads, or HiFi reads.



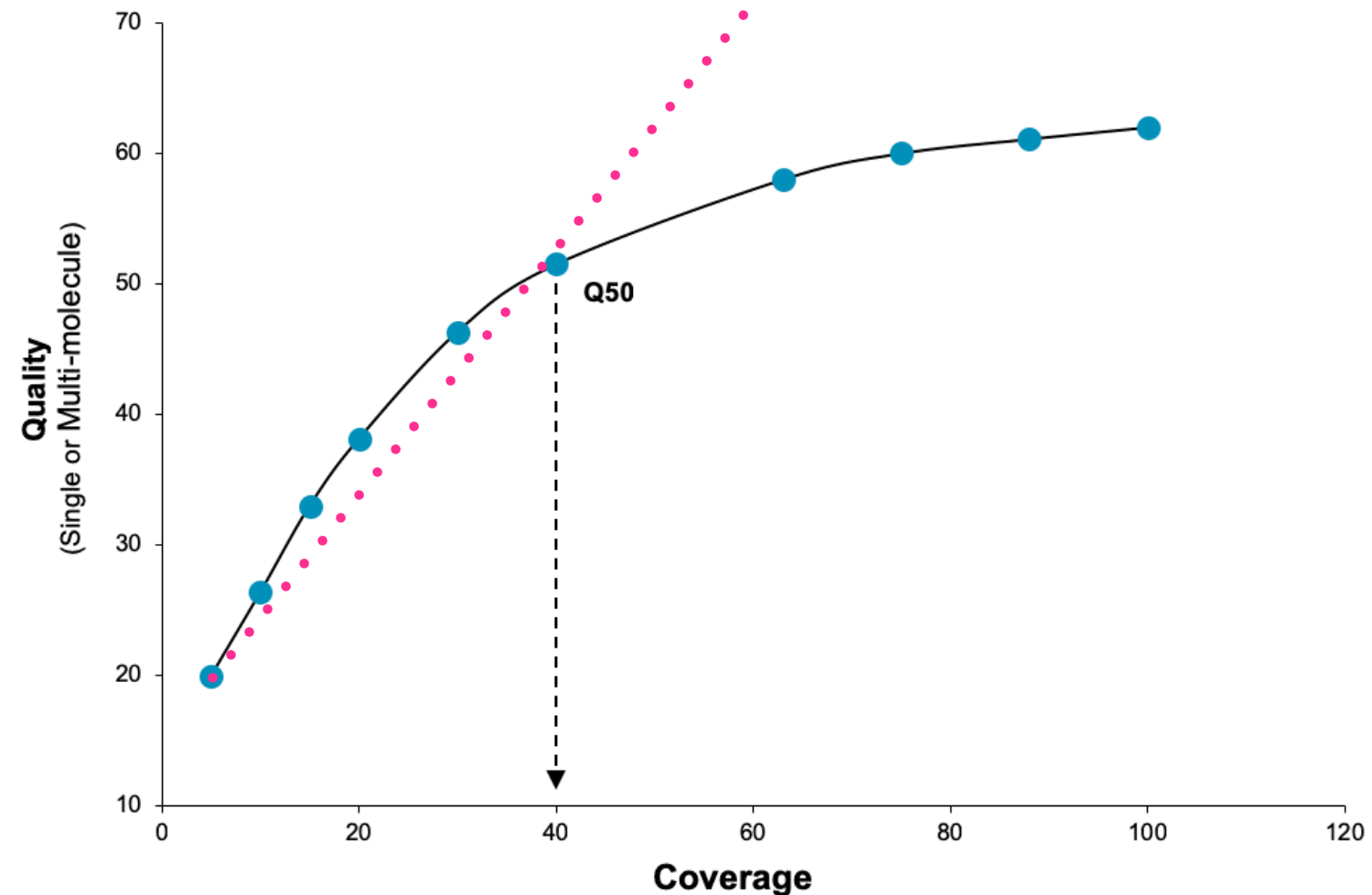
Wenger et al. (2019) Highly-accurate long-read sequencing improves variant detection and assembly of a human genome.



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 $\pm$ 1.2 kb. The predicted accuracy of the CCS reads has a median of Q30 (99.9%) and a mean of Q27 (99.8%).

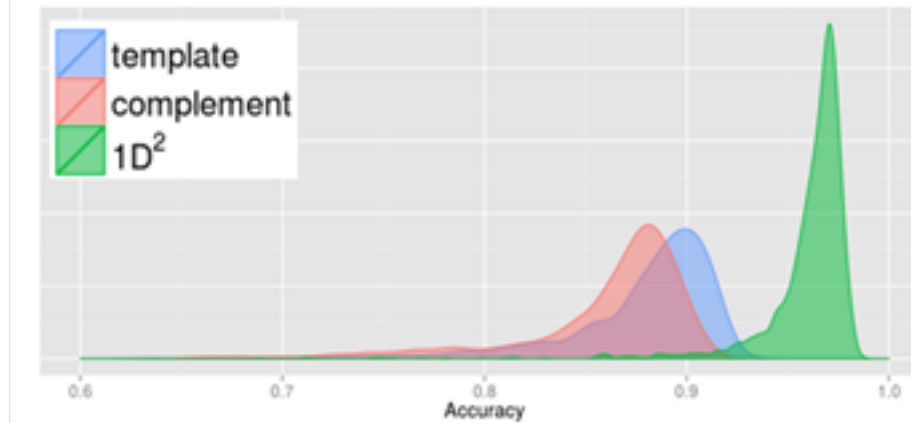
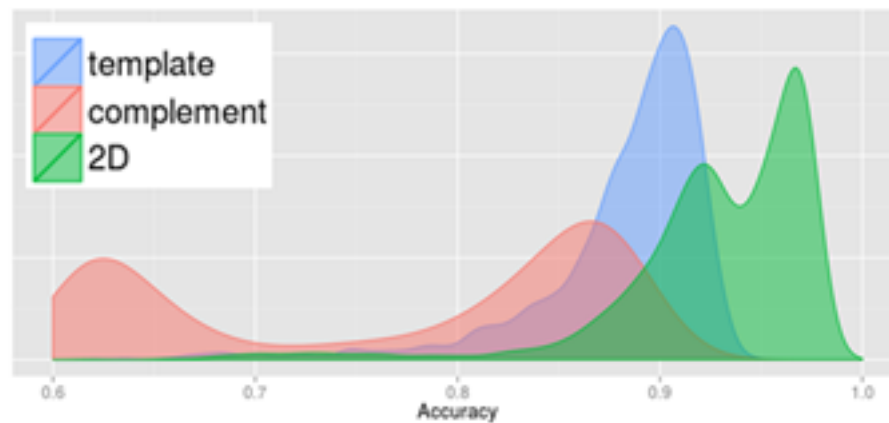


Why does it not improve anymore?



Q50 → $p=1e-05$ → 40 x coverage

PopGen ► Massive Parallel Sequencing



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>



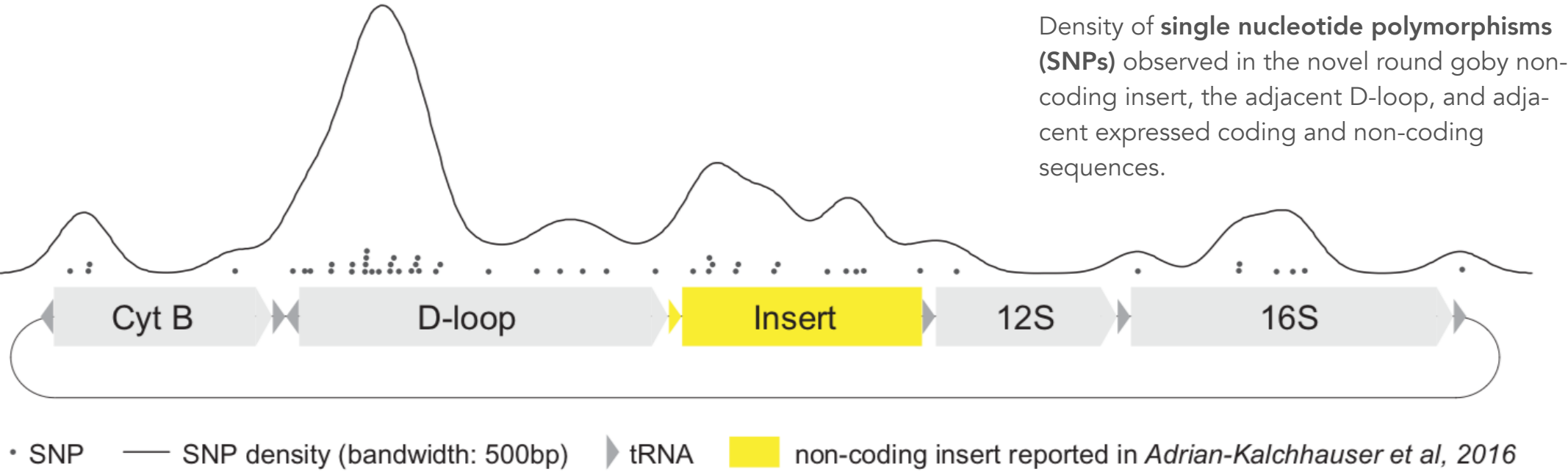
ARTICLE

OPEN ACCESS

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

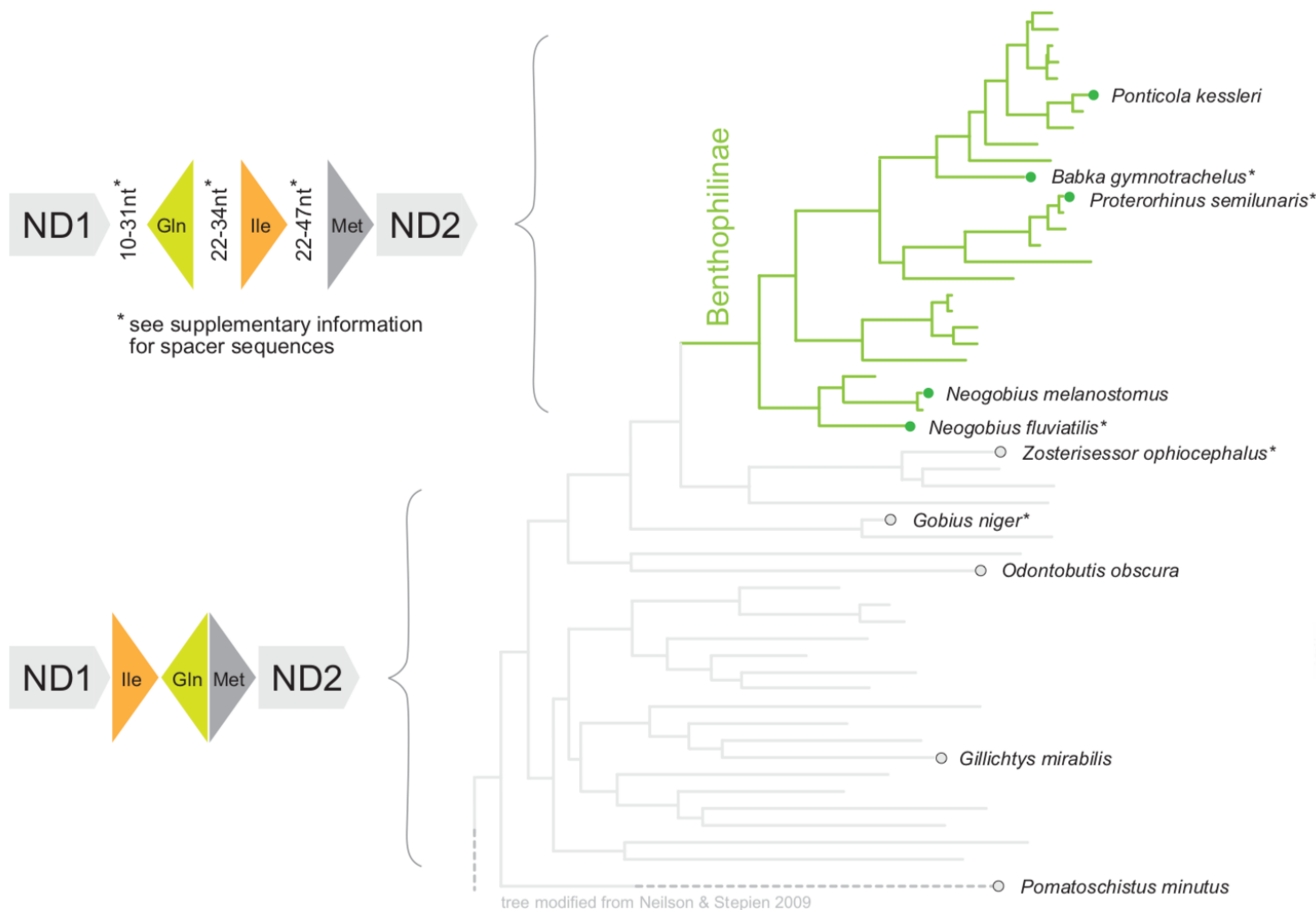
^aBiozentrum, Department Growth & Development, University of Basel, Basel, Switzerland; ^bGenetic Diversity Centre Zurich, ETH Zurich, Zurich, Switzerland; ^cProgram Man-Society-Environment, Department of Environmental Sciences, University of Basel, Basel, Switzerland



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

^aBiozentrum, Department Growth & Development, University of Basel, Basel, Switzerland; ^bGenetic Diversity Centre Zurich, ETH Zurich, Zurich, Switzerland; ^cProgram 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.

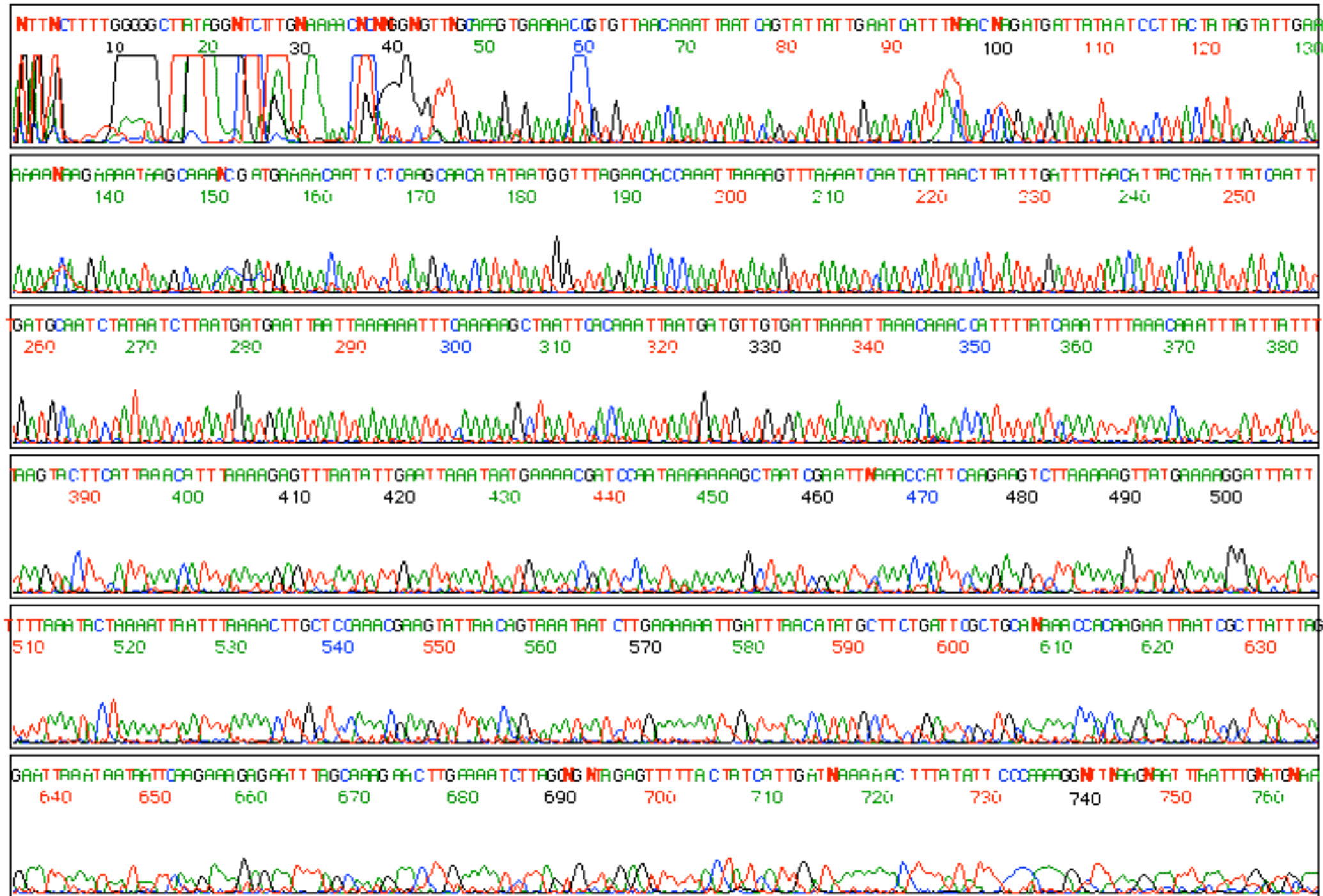


FASTQ quality control

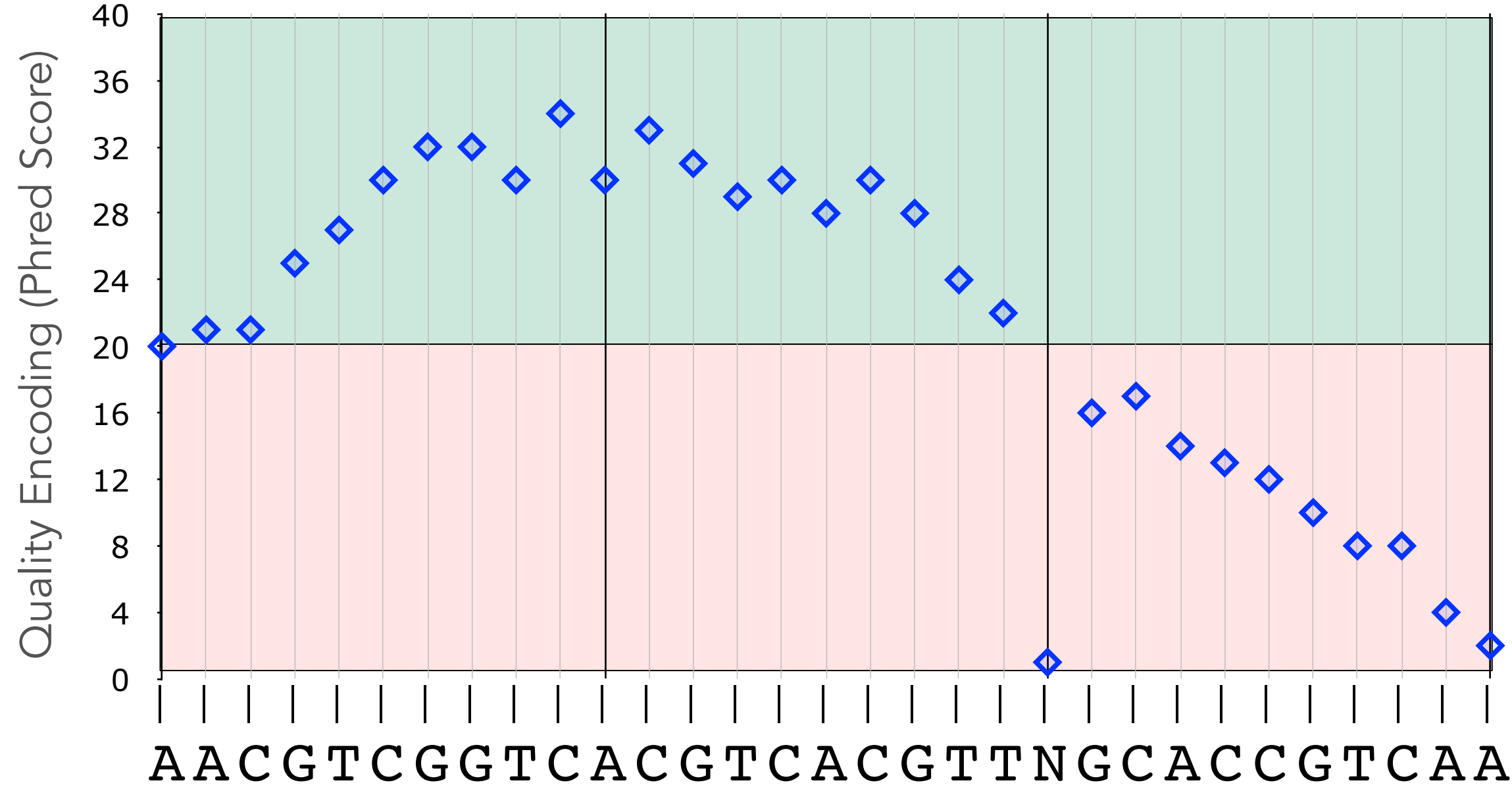
PopGen ► Massive Parallel Sequencing



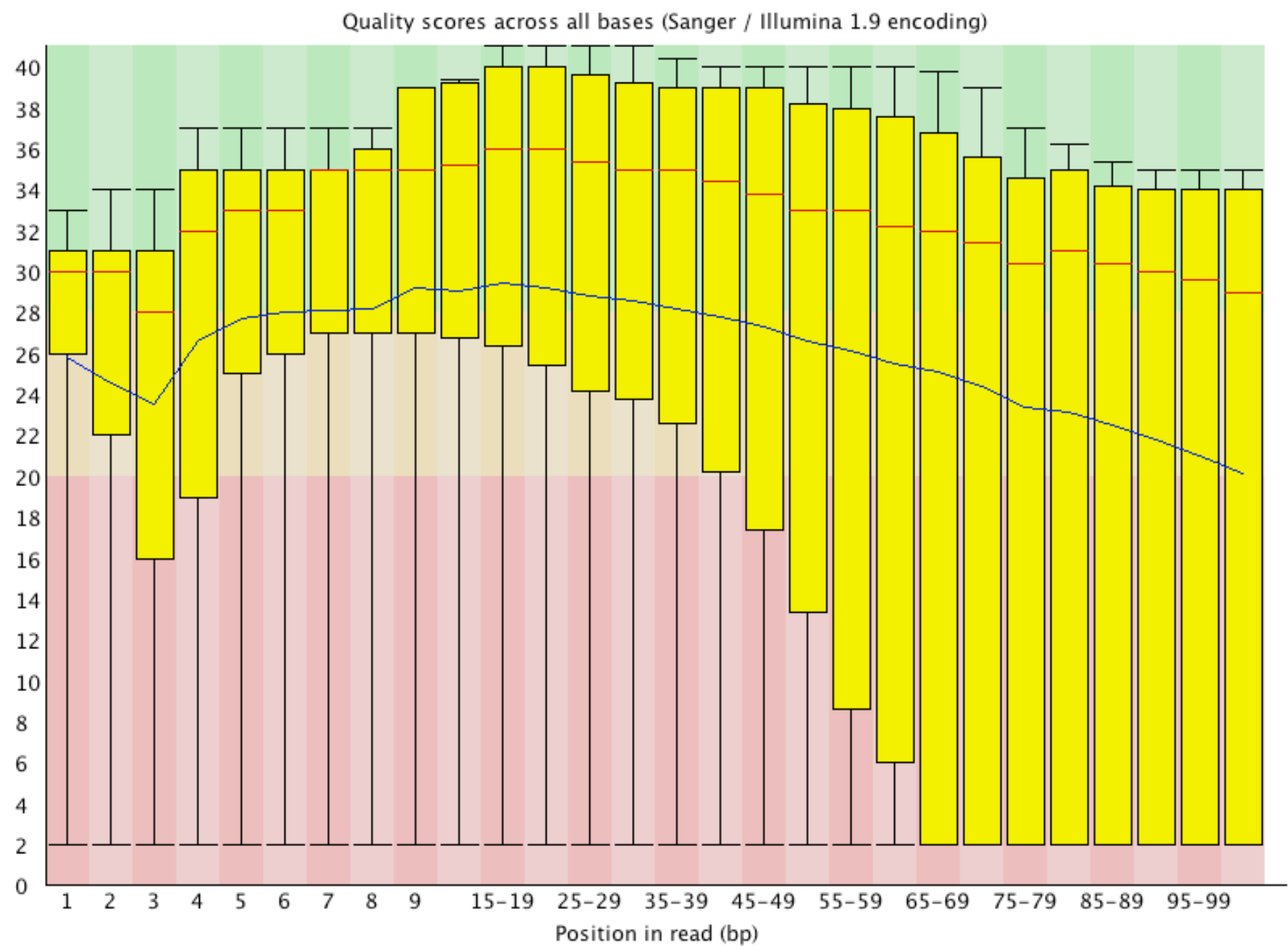
Model 377
Version 3.0
SemiAdaptive
Version 3.0



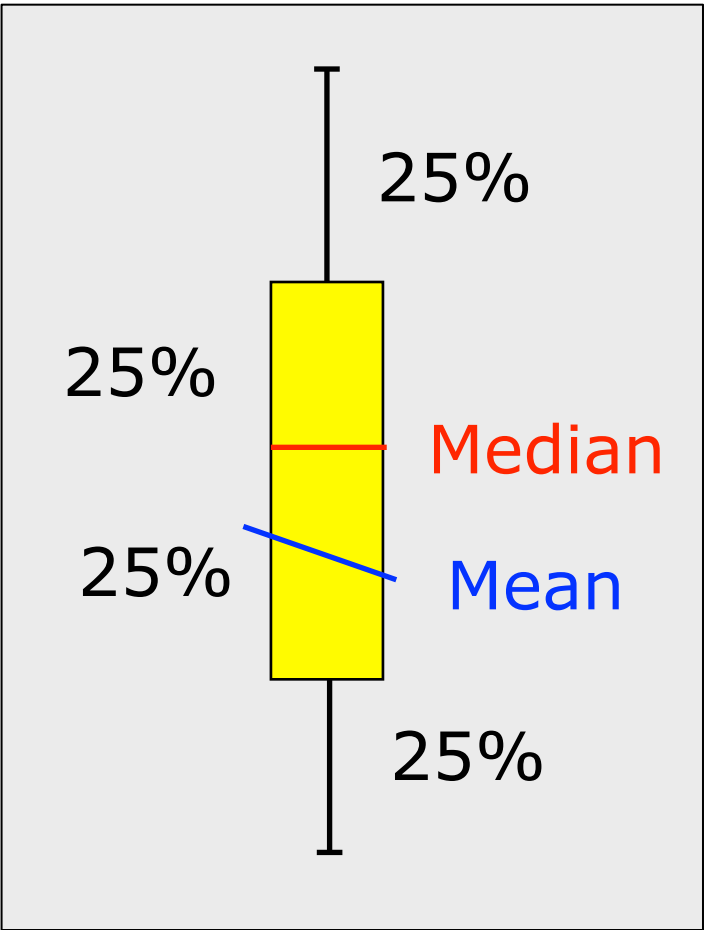
Quality Scores per Base



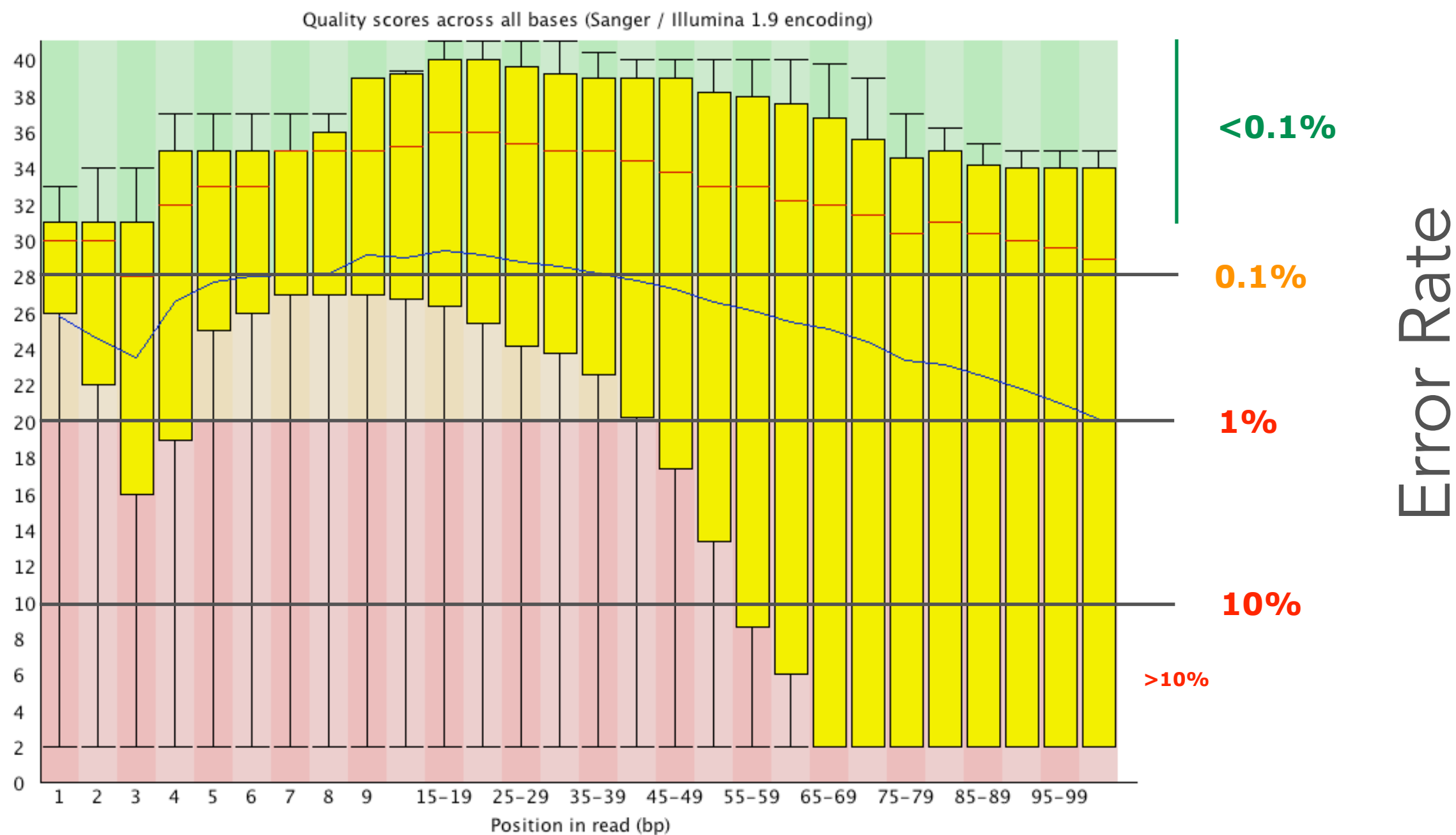
PopGen ► Massive Parallel Sequencing



Boxplot



PopGen ► Massive Parallel Sequencing



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/>)

PopGen ► Massive Parallel Sequencing

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



FastQC High Throughput Sequence QC Report Version: 0.11.2

www.bioinformatics.babraham.ac.uk/projects/

© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewels 2011-14,

Picard BAM/SAM reader ©The Broad Institute, 2013

BZip decompression ©Matthew J. Francis, 2011

Base64 encoding ©Robert Harder, 2012

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

PopGen ► Massive Parallel Sequencing

1_S1_L001_R1_001.fastq.gz2_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

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

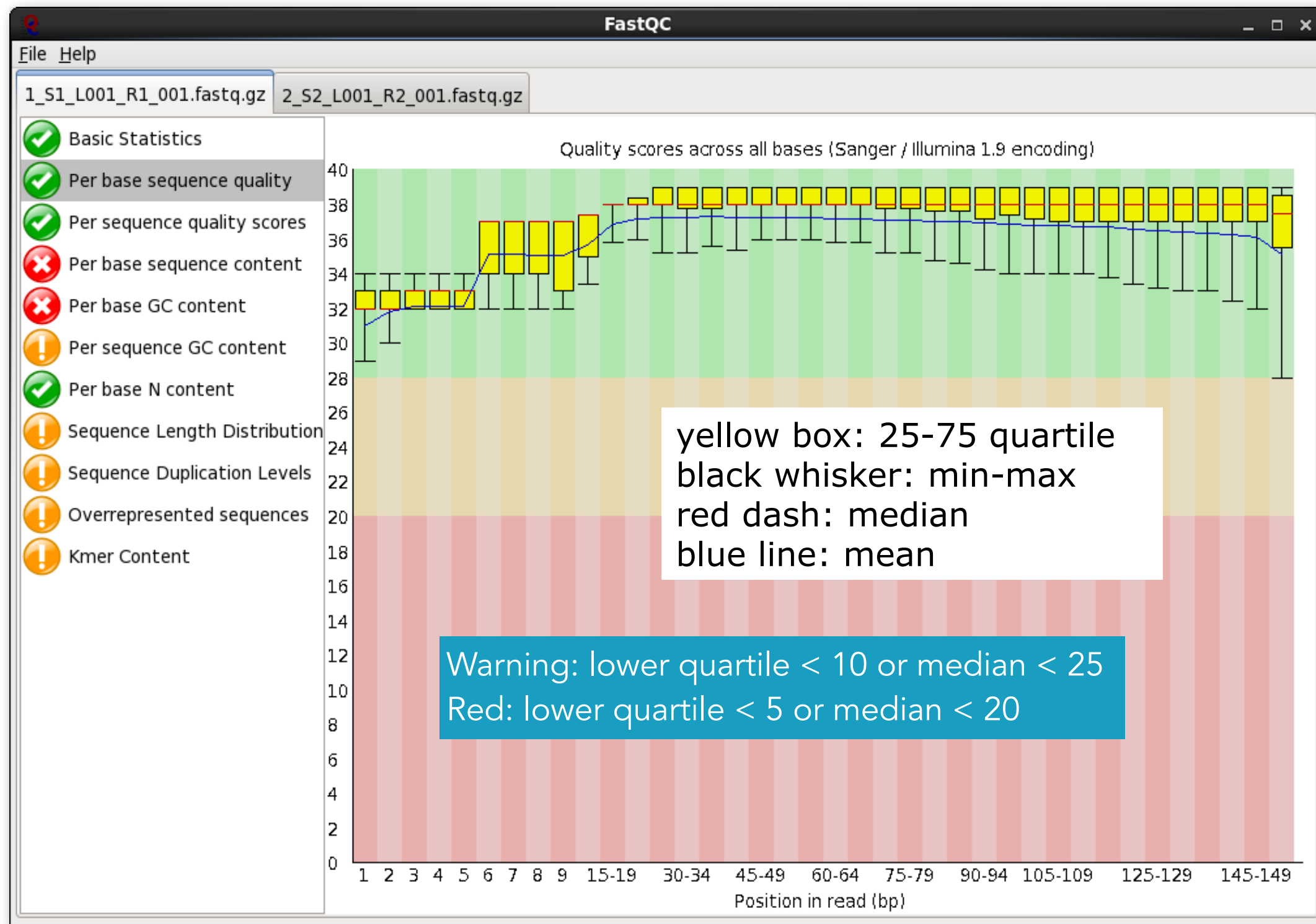
✗ Error

The traffic lights are context dependend!

79

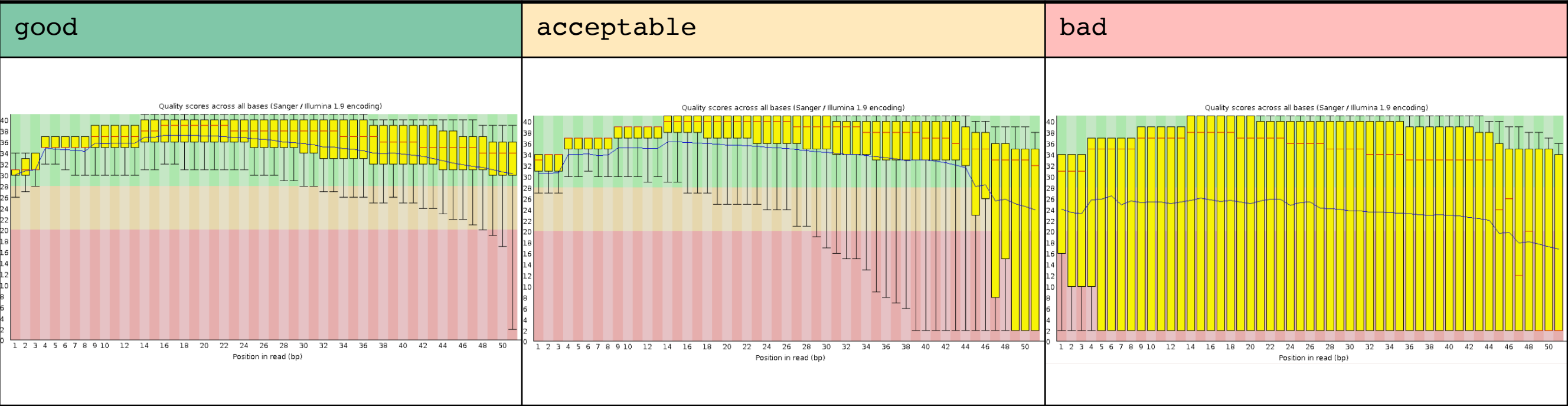
UniBS | EvoGen | JCW

PopGen ► Massive Parallel Sequencing

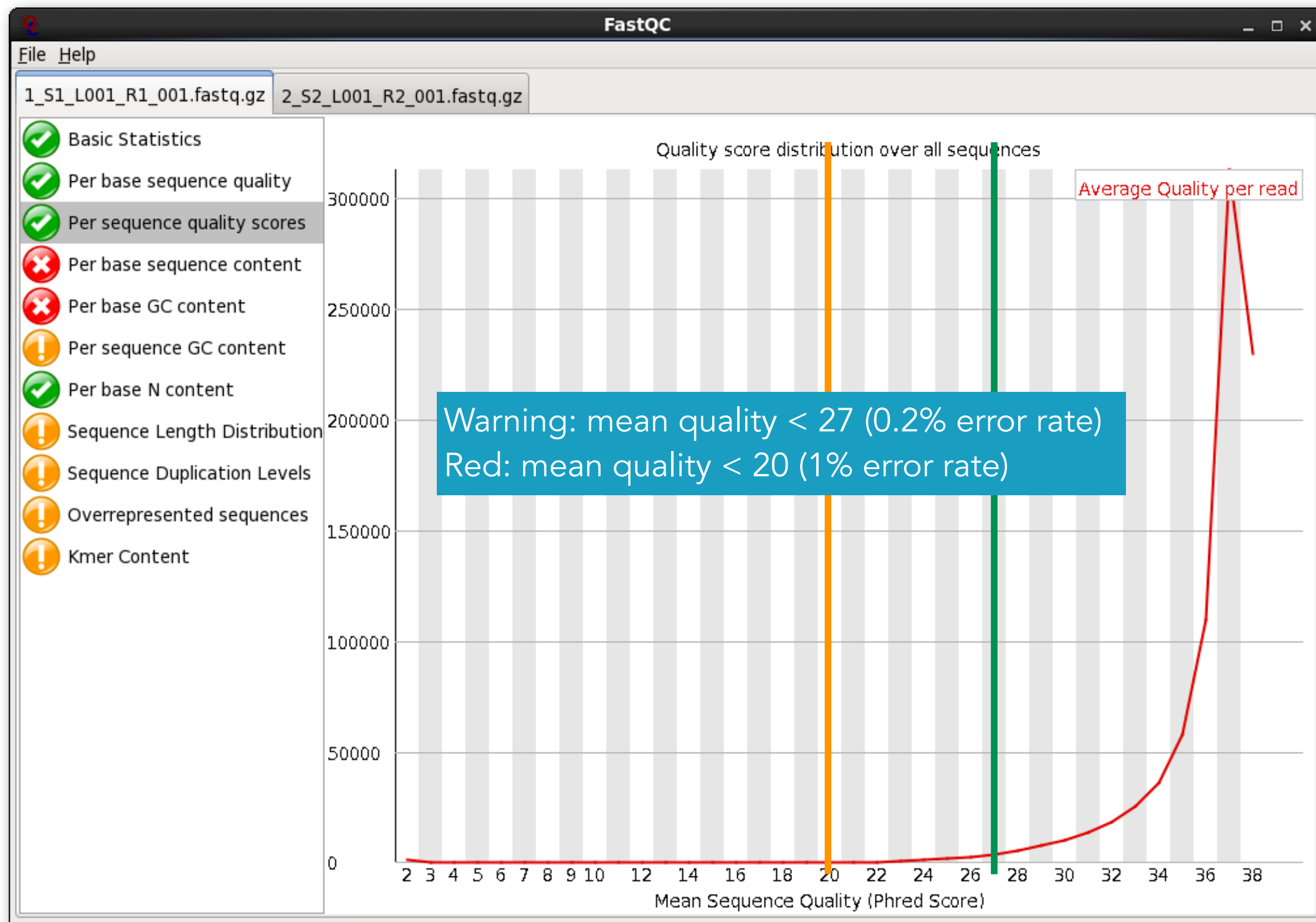


A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25. This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

PopGen ► Massive Parallel Sequencing



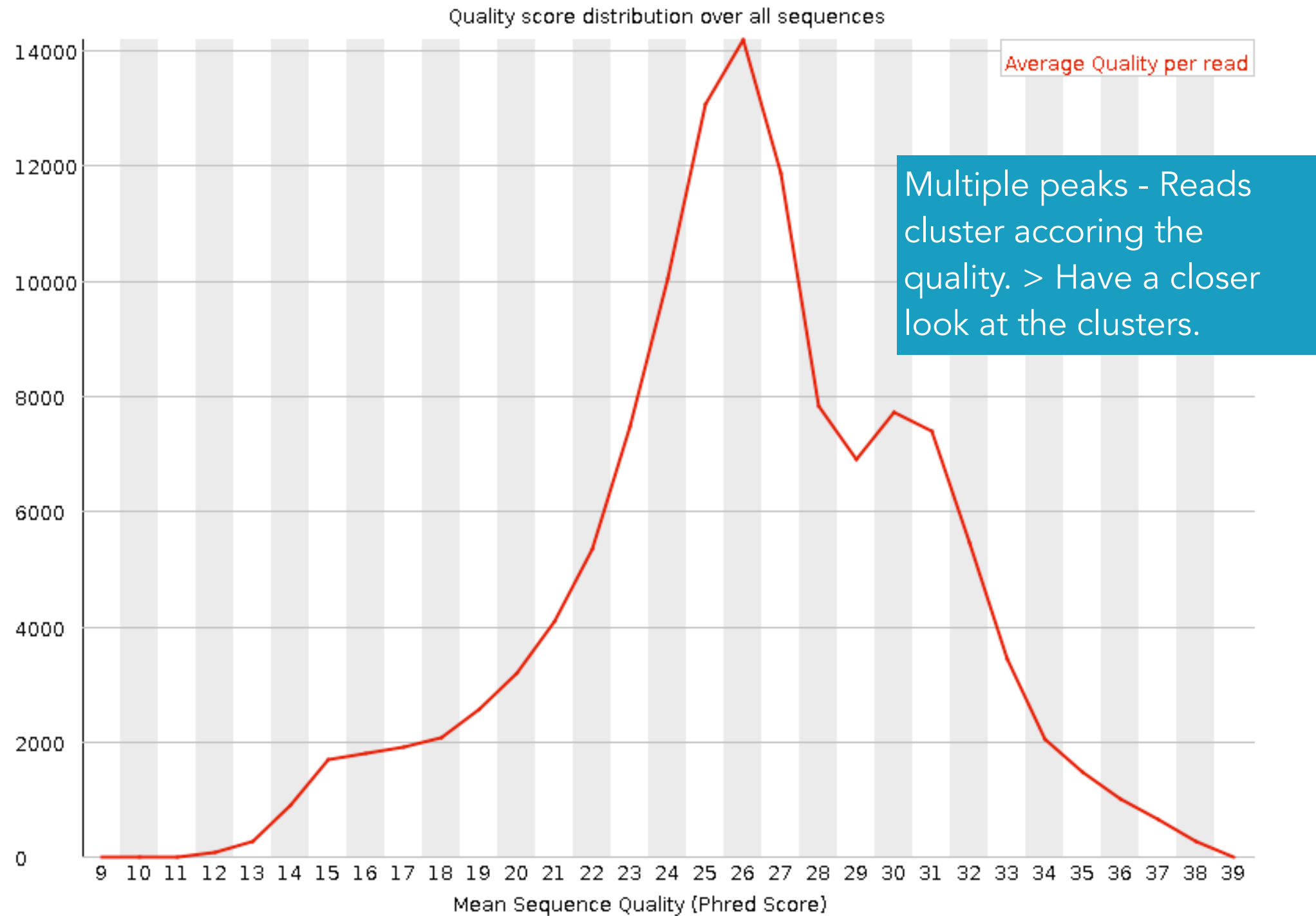
PopGen ► Massive Parallel Sequencing



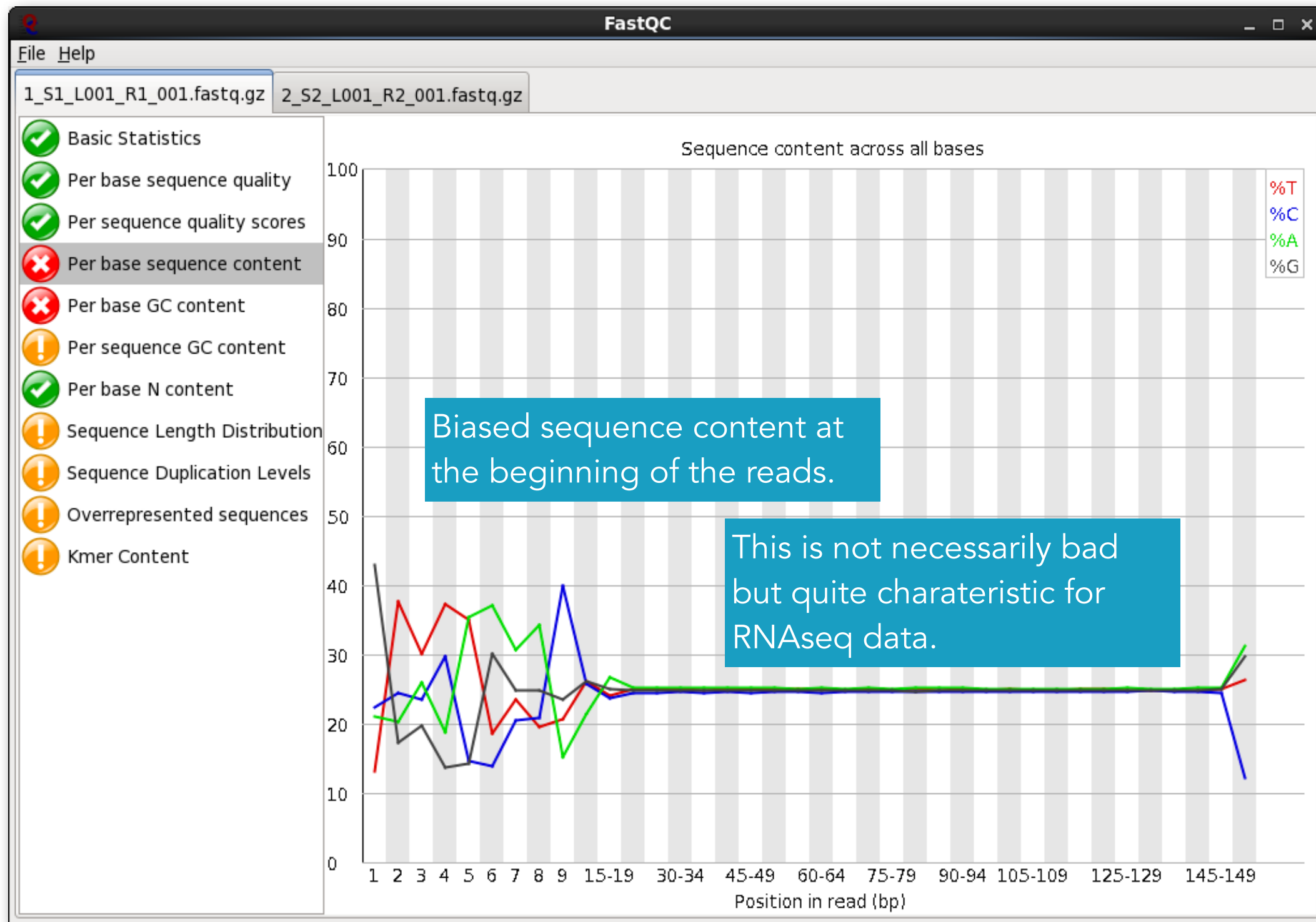
A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate.

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

PopGen ► Massive Parallel Sequencing

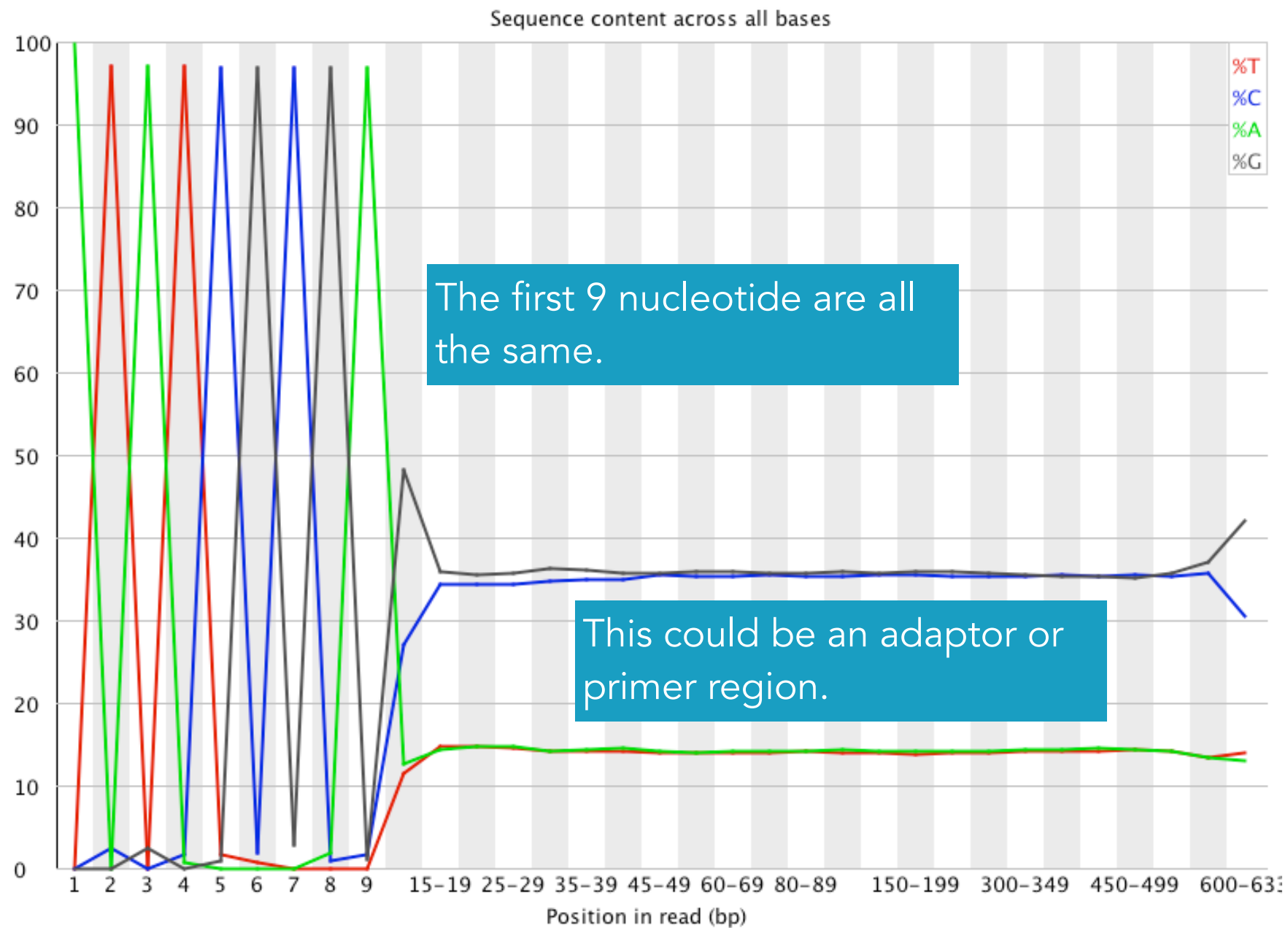


PopGen ► Massive Parallel Sequencing

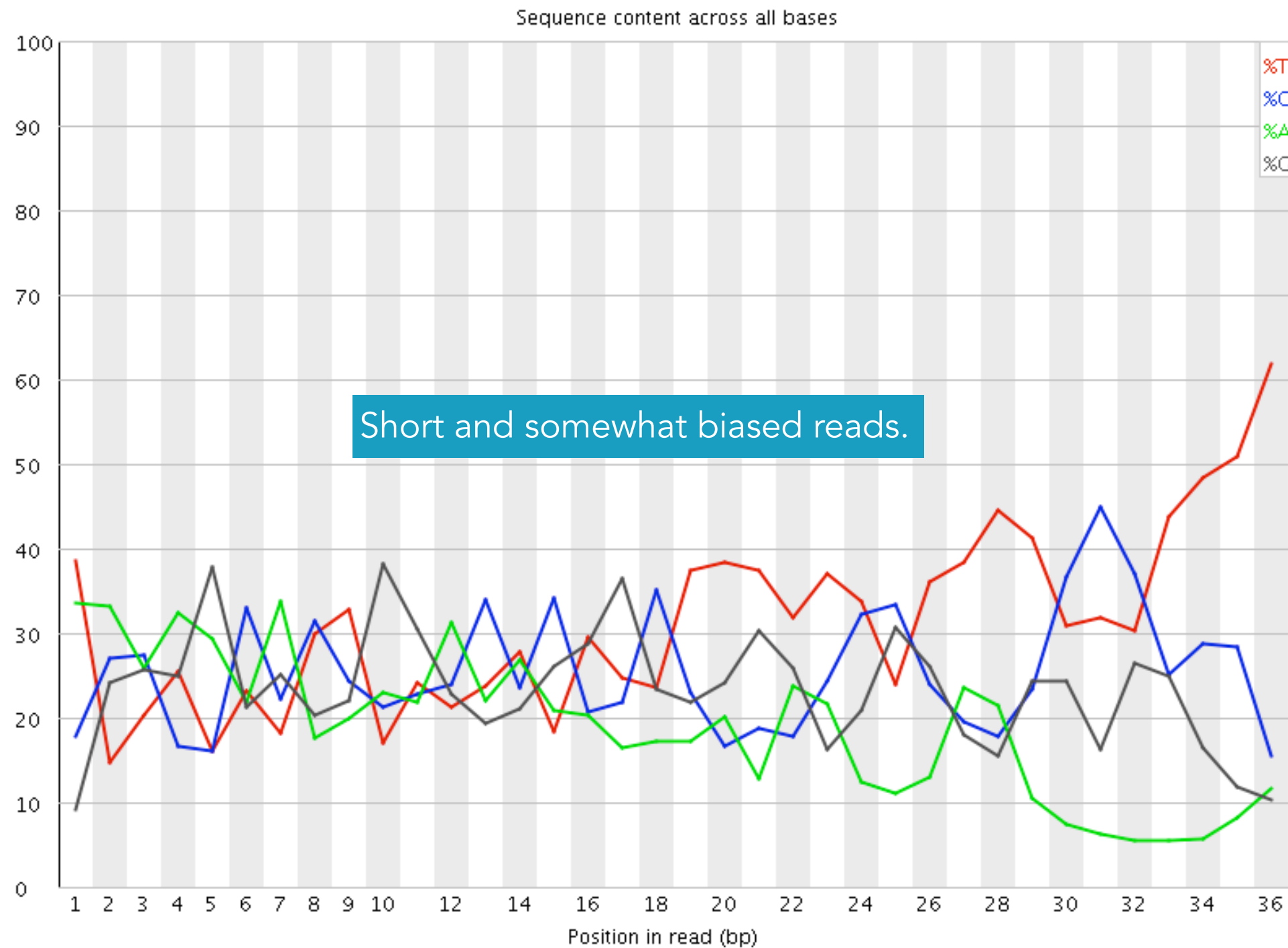


This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position.
This module will fail if the difference between A and T, or G and C is greater than 20% in any position.

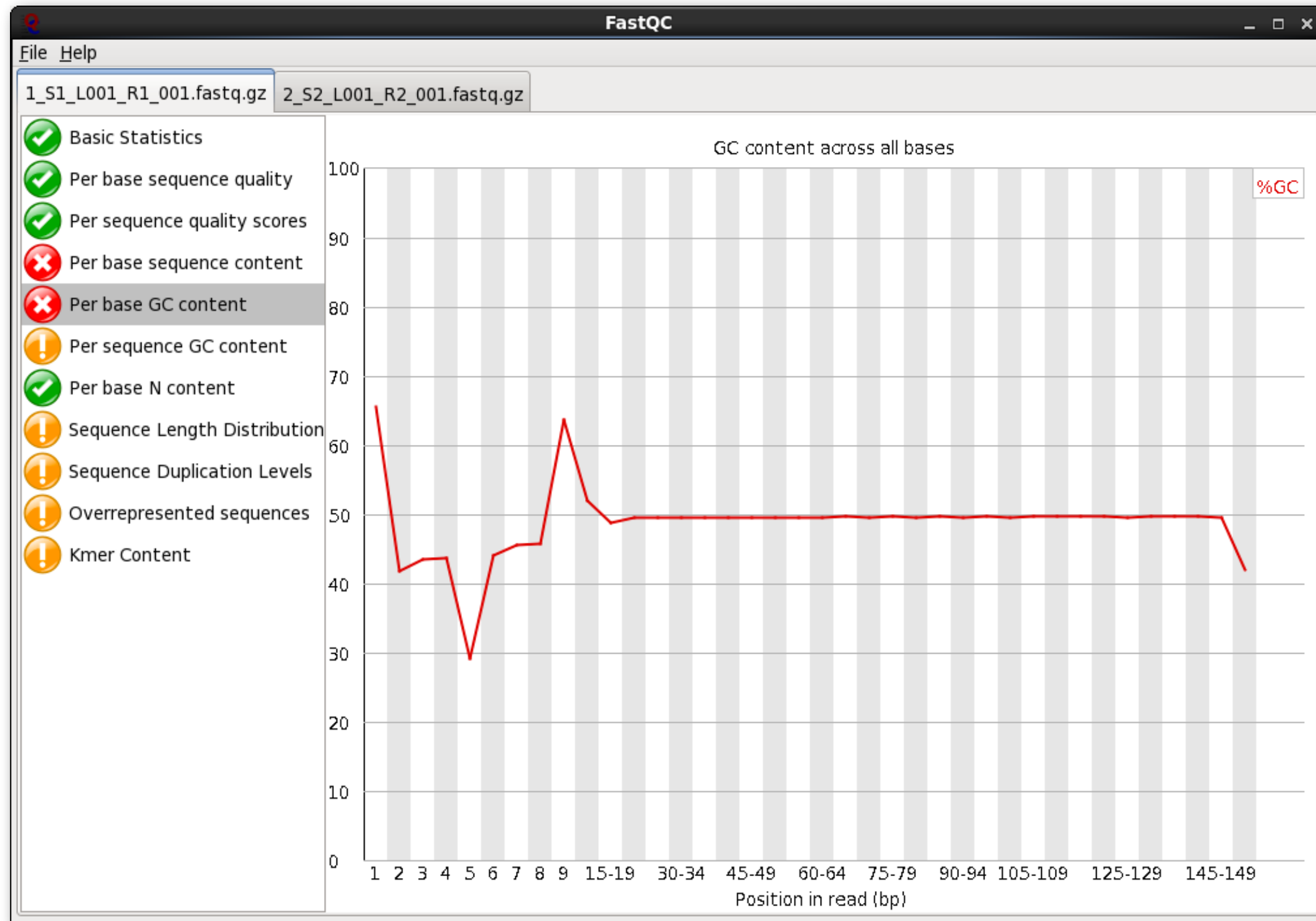
PopGen ► Massive Parallel Sequencing



PopGen ► Massive Parallel Sequencing

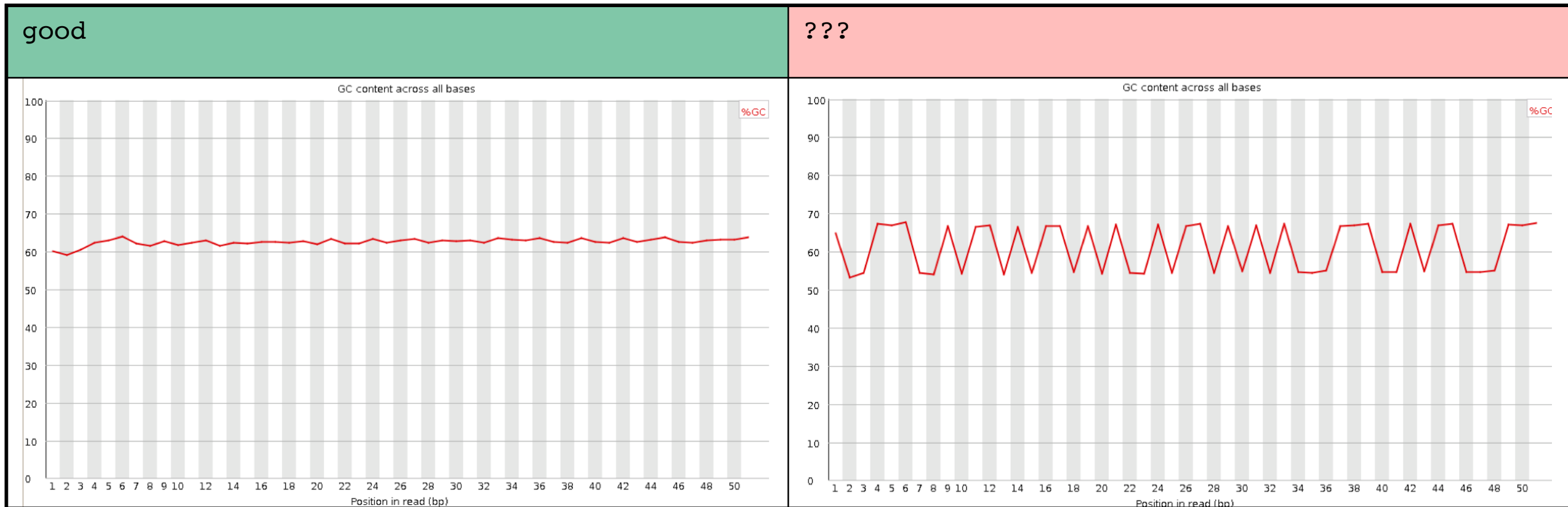


PopGen ► Massive Parallel Sequencing

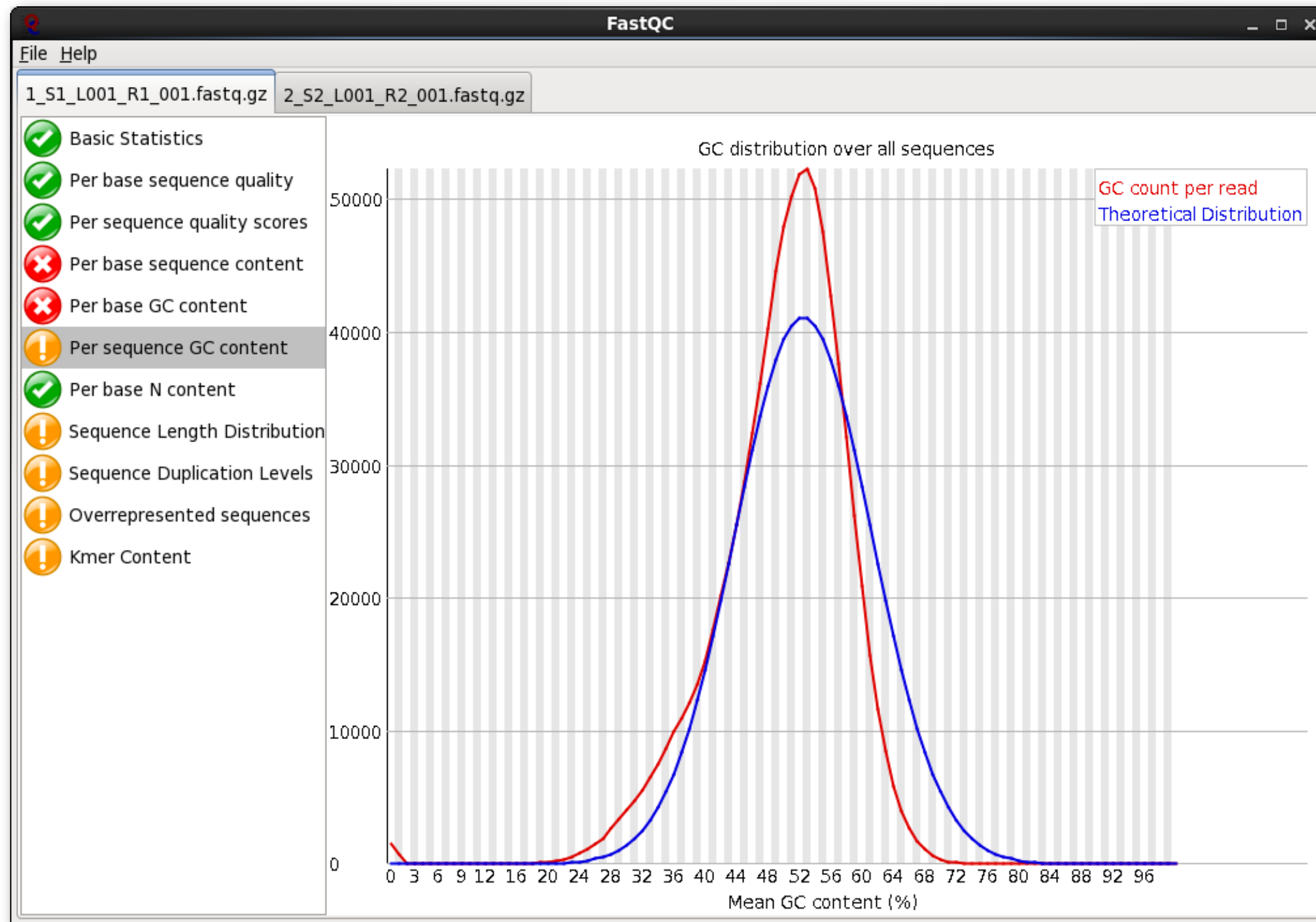


In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the line in this plot should run horizontally across the graph. The overall GC content should reflect the GC content of the underlying genome.

GC content per base:



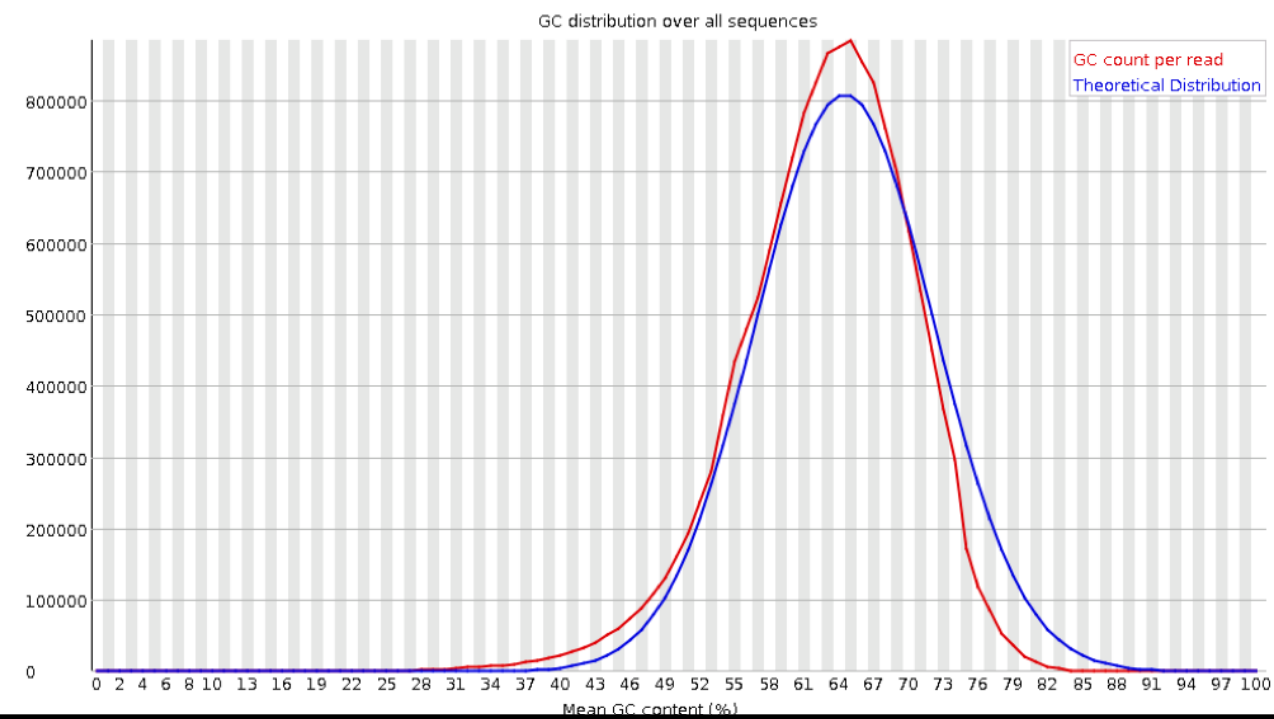
PopGen ► Massive Parallel Sequencing



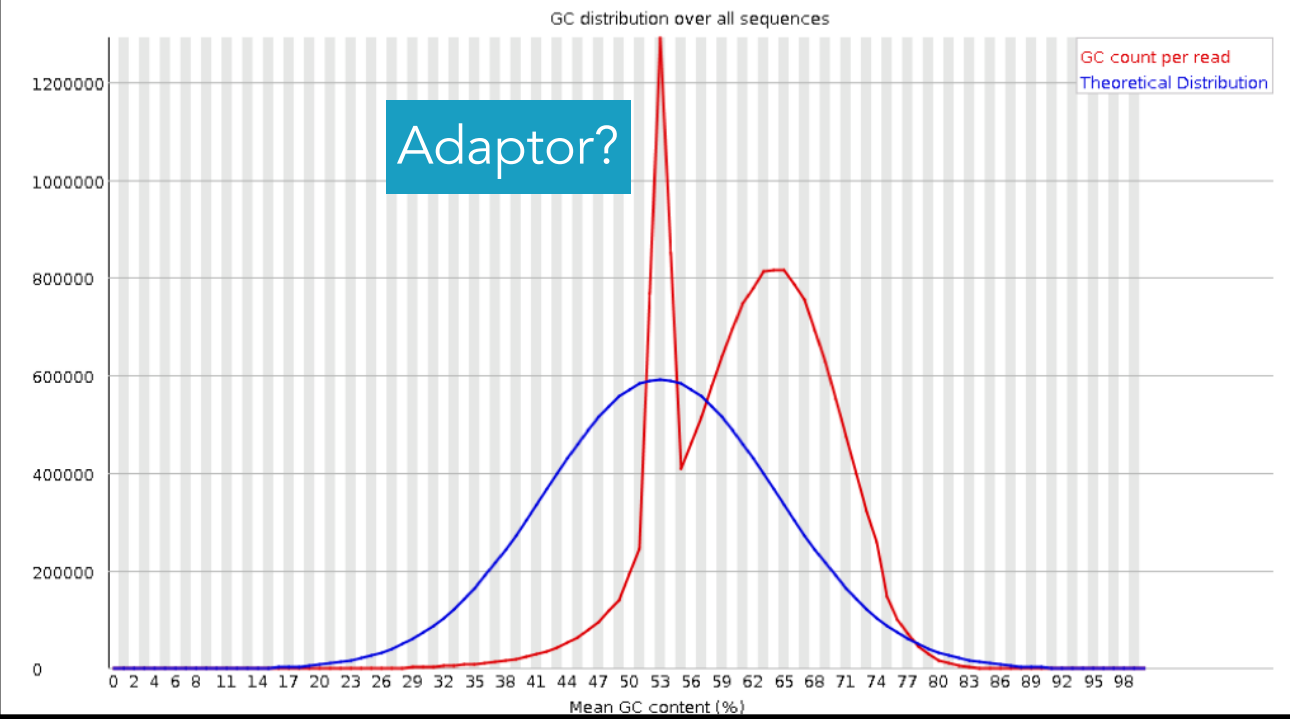
In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used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

PopGen ► Massive Parallel Sequencing

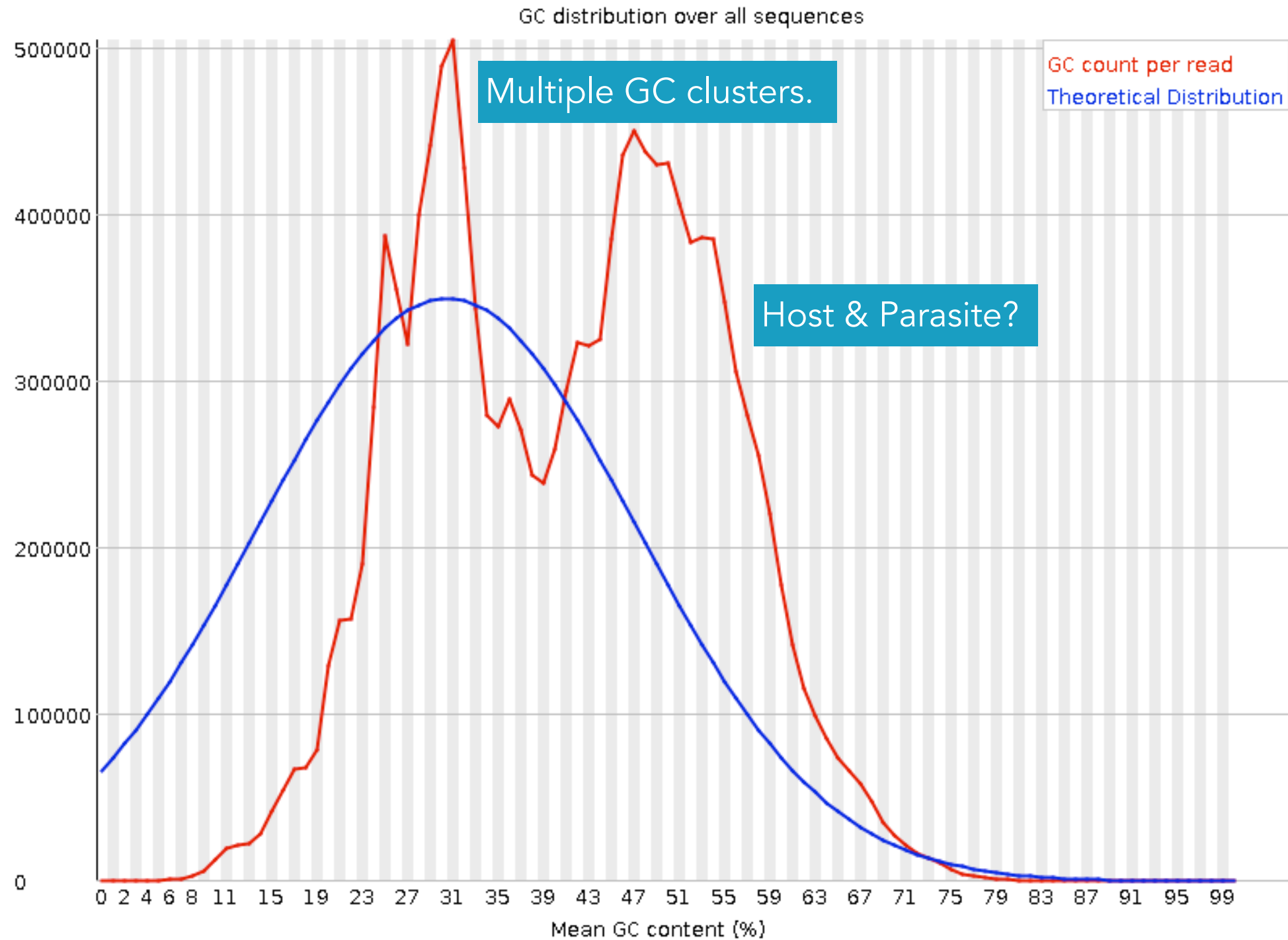
good



bad

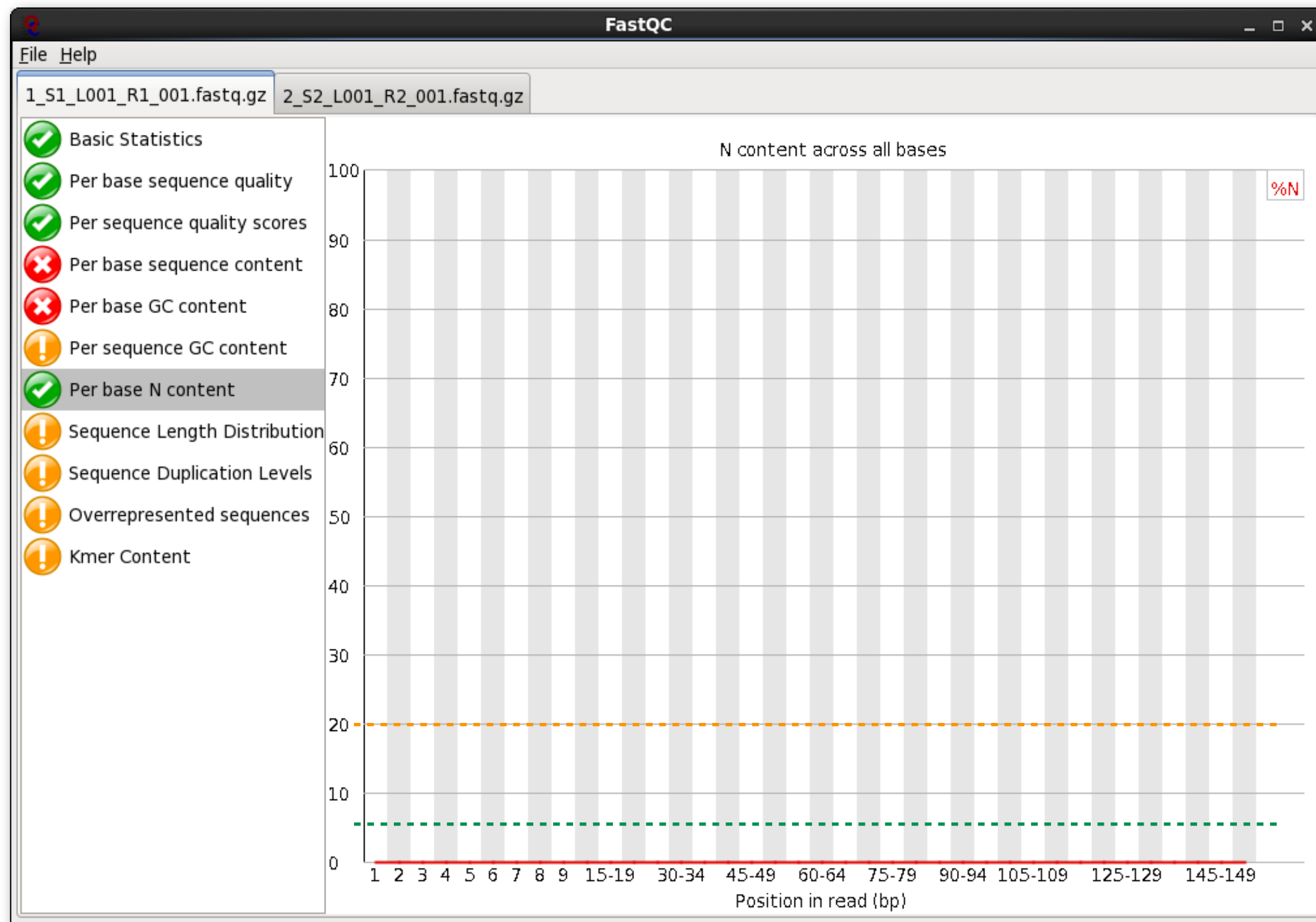


PopGen ► Massive Parallel Sequencing



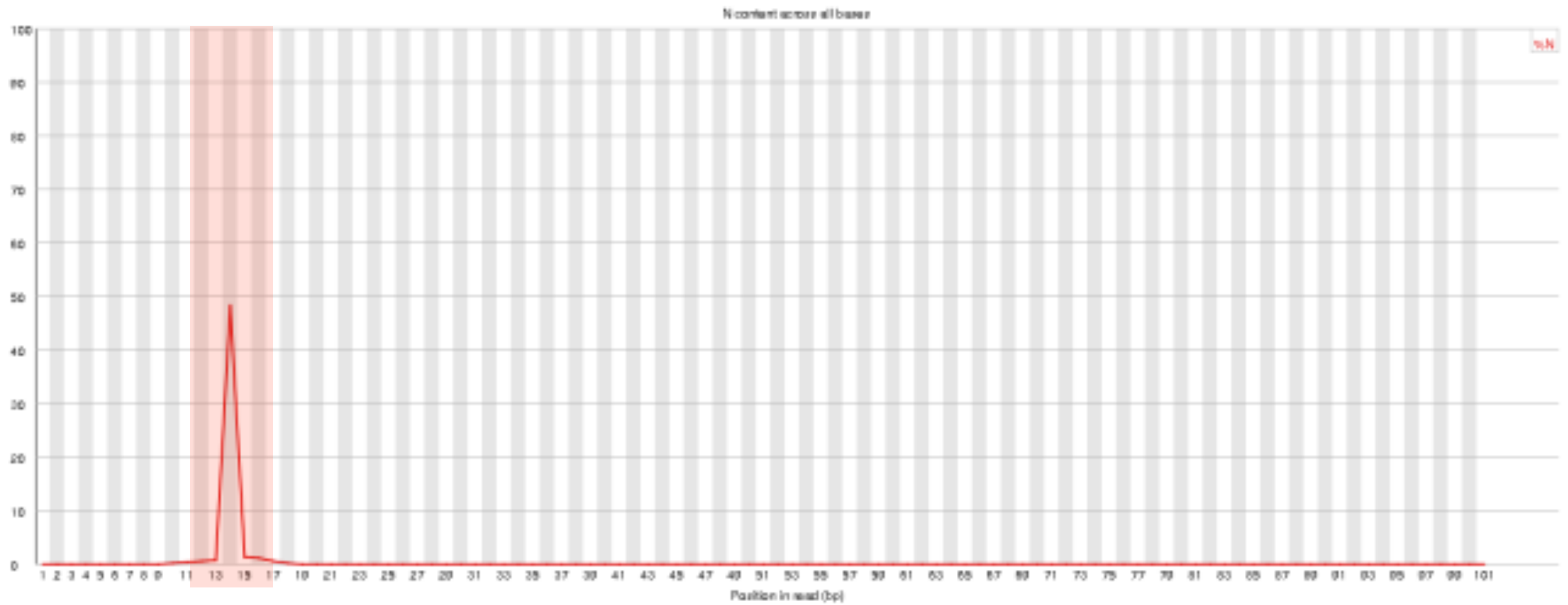
NGS ▶ FastQC - N Content

PopGen ▶ Massive Parallel Sequencing



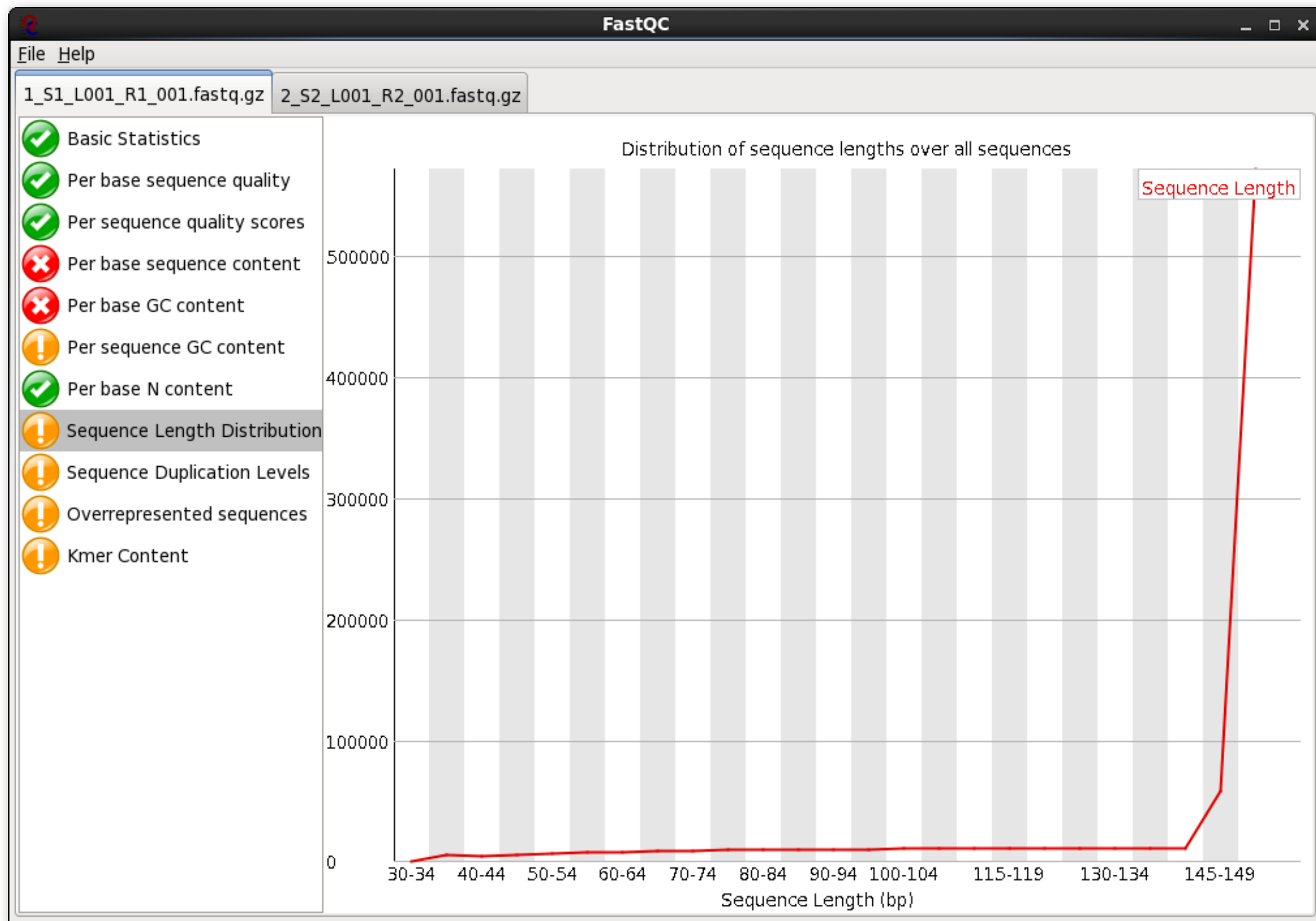
This module raises a warning if any position shows an N content of >5%.
This module will raise an error if any position shows an N content of >20%.

PopGen ► Massive Parallel Sequencing



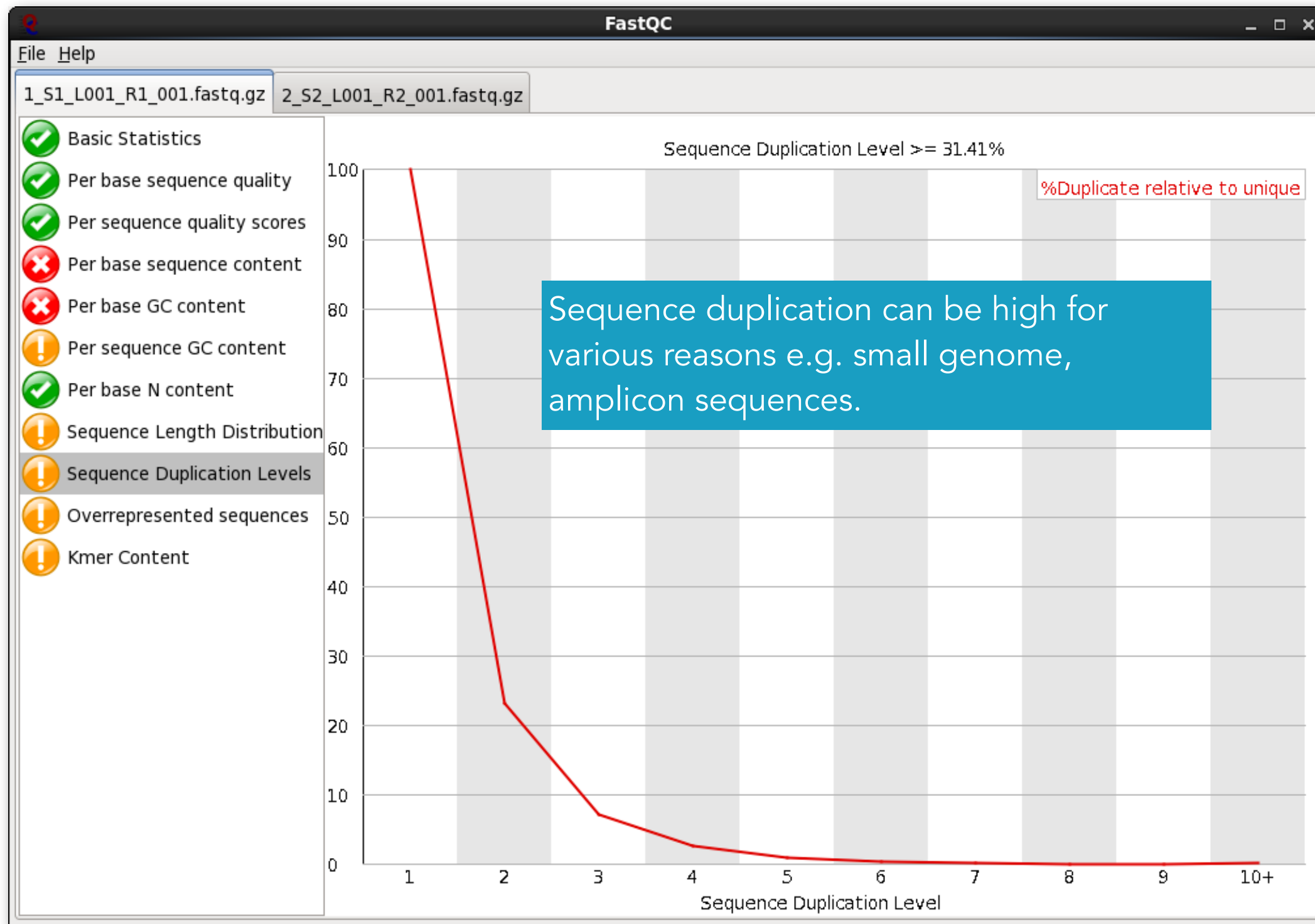
GTCGGGTTTTACCATTGGGTTTGGGTATTTCCACCCCCGAATGGCTTGCGGTNTGTACNNNNNNNNNNNGNNTNNNGTAAATAAACTTTCTGGATGGTGT
TCTTTAGTAAAAGGCGAAAGATTTATTCGTTCTGTATTGACGCCATGCCGGGTNNCTTTNNNNNNNNNNNNNCNNTNNNTTATACAGCTTTGTGGAAATTTAC
AAAAATAGTACTTCAGAGGTAATAAATAAAATTATACCTCACCGTAAGCCTACNNCNANNNNNNNNNNNNNNNNNNNNTTGATAAGTTCCTTCACGTACTAC
TCCCGGGCACAGAAGTGCCCTTTCCACTGGGAAAGACTCATTTATTAATTACTNNTNTTNNNNNNNNNNNNNNNNNNNANNNAAATTAAATCATCTCACTCTTTAT
TGACAGAGTTACAATTGACTTCGTCCAAATCCAGAGAGGCATAGTTGACCATCAAAGACNNNNNANNNNNNGNTGNNNGTCGTCGTCGGCCTAGGCACTGAC
CCCCGTCCCCTTAACAGTGAGTCATCAACAAAGAAATTGAGGTACACTTTTACNNTNTTNNNNNNNNNNNNNNNNNNNNAGCTTAGGTTTATCTGTGTCTTCT
TGTTGGCTTTGCCGAGATATACTGAGTTTATAGTTTCTGGCTTGGATGGCAGGNAATTTNNNNNCNNNNNTNCANNNTTCTCAATGCCTTTATTCCAGAAA
ACCTACAAGAGTTTTTAACTCTAAATGCAACTGGTTTCTAATTATTGAAAATANTTTATNNNNNCNNNNNGNCTNNNCTCAGTAAACTCAGGGGAGTAGAA
ATCGAGTTTTTTGGACGAAGACGATGACGCAGCTGTGCATACTCAGACCACGCANNANCNNNNNNNNNNNNNNNNNNNNNGGACTCTGCCTCGTCGTTGGTCGA
CTAAATATTGTTTGGAGTGGCAATGTGCCCAGCTGAAACACTGGGCACTTCAANCTGCCNNNNNGNNNNNGNTGNNNCTCTCTCATTAATGTTCAATGAGA
CGTCCAATTACGATCGCCAGTTTTTTTTTTTTTATATATTTTTTTTTTTCTTATGTATTTNNNNNCNNNNNGTTCNNNTGCAGAACTTTTTATGTCTCCACT
GTGTGTCACACTCCAATGAAAAAGAGATAAAATCCTGGTGTAATTGAGACAATNNTNCANNNNNNNNNNNNNNNNNNNNNCCATTGAGCAAGTATGGTAACGAT
TTCATGCTAGAAATGAAGGAGACTGCTGCCATTCTGAAACACGCCACCAAGCGTTCGTTNNNNNTNNNNNATTANNNTGGGCGAGGAACAAGGCCATCCGA
CCCCTTGGCCAAATACCTAGAATTGATCTTGAGCTTAGACGCATTCCGGCTTTNNTCCANNNNNNNNNNNNNNNNNNNANNNCTGTGCGTAGGCTGTGAGTAGGGT
GAGAACGTTATTATTTACGAACATAGTTACGAACTGCAGGATAATATTATTGTNGCTGANNNNNNTNNNNNTNTCNNNTTCACCCGATAAGAACAAATTATT
CTTTAATCCAGGGGCTTTACGCGCCAATTGCCTCTGCTTTCCAATGGTATACTNNNNNTNNNNNNNNNNNNNNNNNNNNNACCATTCTATGGACCATCTCTTGG
CTCTTTTTGAGCCCCTTTTGCCTTTTTTTTAAACCTAACTATGGGACACTTATTNTTAAANNNNNNNNNNNNNANNNNNNTTTTTAGCAGTAAAAGCTGTAACC
CCATAGAGAGAGGCATAAAGCTCAACAGCCGTTTGAAGAAAACTTTTTTGTGCONATTTTNNNNNNNNNNNNNANNTNNNAACGCAGTATCCTCGATATACTTA
GTGGTGAAGACGTTTACACTCGTCCGTTCCACATTCCTTTTCCTTCGTACACTNTGAACNNNNNNNNNNNNNTNNTNNNGACACCGGTAACAGCATCTTTTTT
CATCATCCACATCTGCTGCCGCAAGCATTGTTGGTCATCATCATCATCGCGTONATATCTNNNNNNNNNNNNNTNNTNNNTATTTTCATTGCCCGTGTGCATGCAG
GAACAAACAGCTCCTATACGTGAAAATACCAAAGGGTCGTTGCATCATTTGAATGTTTATNNNNCANANGCCACTNNATATGGCATGAAGGTGCTTAACGA
CTTCCCCCAACTCCATATCATCATCTTCCAATATTTATAATAACAACATCTTGCNNCNATNNNNNNNNNNNNNNNNNNNTNNNATGTTTTCTAATAACCAACATTATG
CTTGTAGACCTCGGCACCGAATGCTTCAGAGGTGACGAGAGACGTGAGAGCGATGAAGANNNNNGNNNNNTNTGNNNTTGAAGGGGTCATGCGGGGTTTCT
ACTCAACAGGATATGTACTCTGCAGAAGAATAAGAAGTTAAACAATATCACCANNANANNNNNNNNNNNNNNNNNNNNNATAAAACAATAAATTCAACGCGTT

PopGen ► Massive Parallel Sequencing



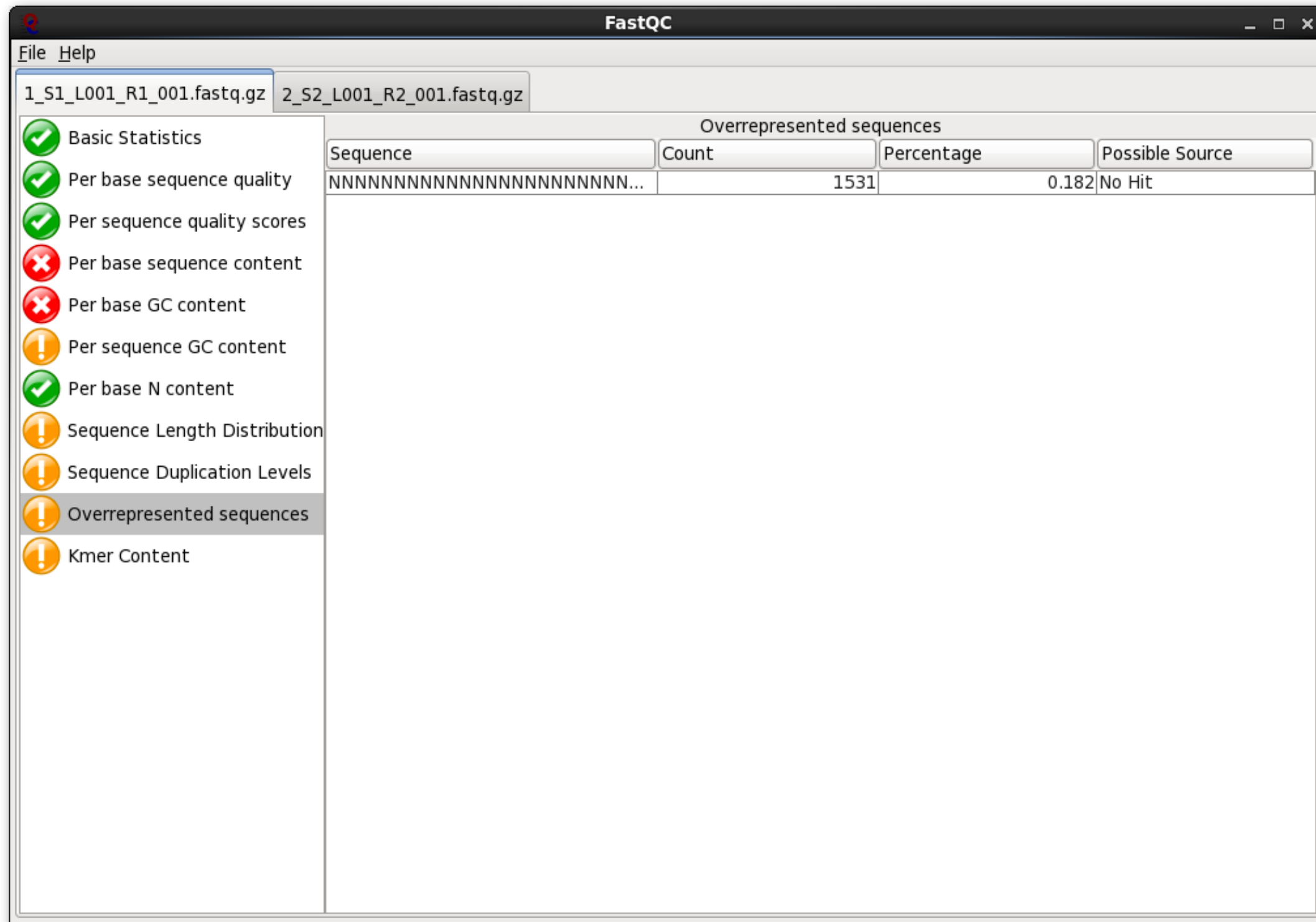
This module will raise a warning if all sequences are not the same length.
This module will raise an error if any of the sequences have zero length.

PopGen ► Massive Parallel Sequencing



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.

PopGen ► Massive Parallel Sequencing



FastQC

File Help

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

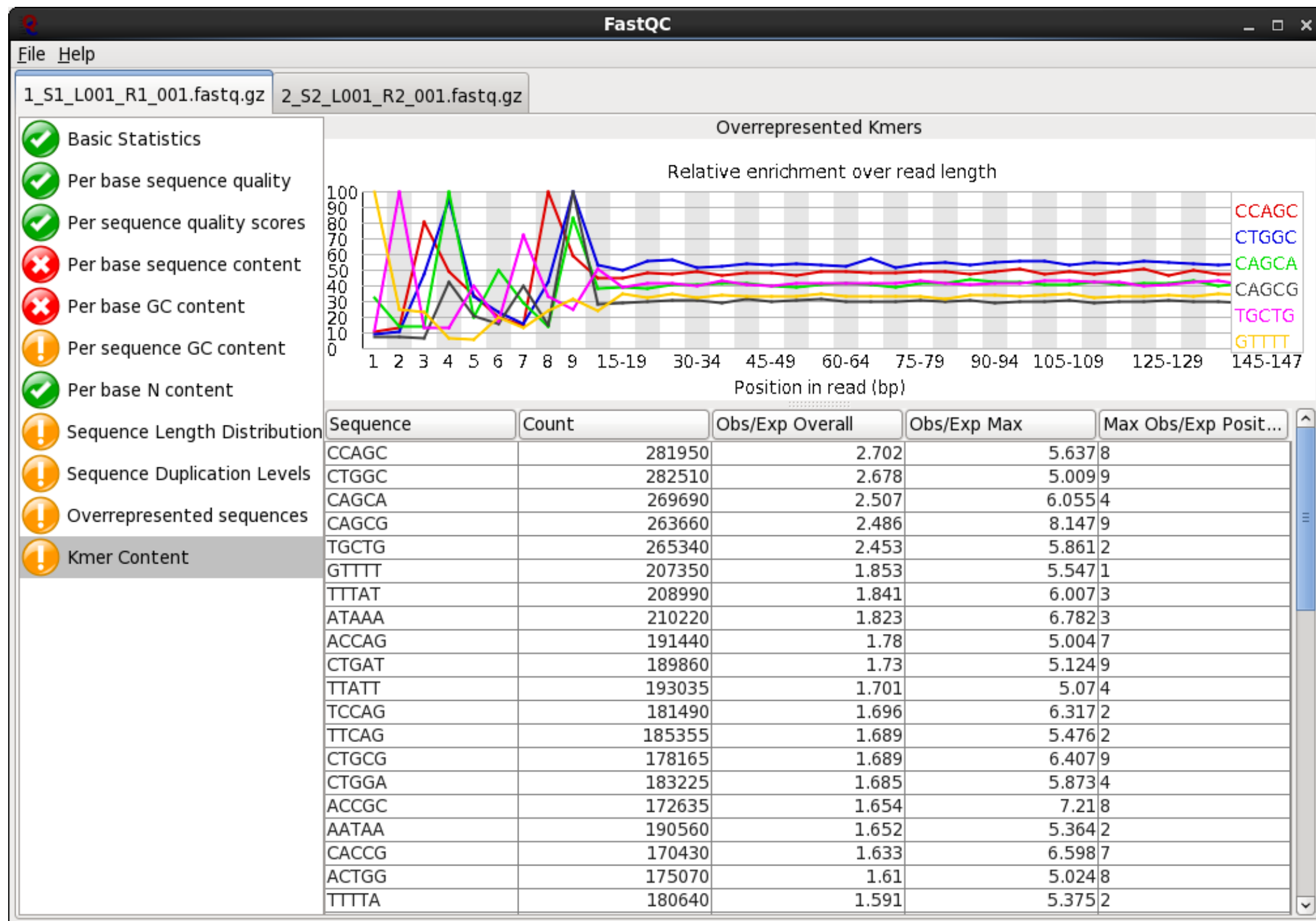
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNN...	1531	0.182	No Hit

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.

PopGen ► Massive Parallel Sequencing



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.