

DNA Sequencing and Data Analysis

Prof Noam Shomron
Hadas Volkov

Lecture 3, April 11, 2024

DNA Sequencing and Data Analysis

Modern DNA Sequencing, 2nd wave File Formats, tools.

Thursday 18:30 to 21:00
Hangar H2

nshomron@gmail.com

hadas.volcov@post.runi.ac.il

Class & Home Assignment

github.com/ShomronLab/CompGenomicsWS

**Home Assignments are submitted via
Moodle up to two weeks!!!**

The screenshot shows a GitHub repository page for 'ShomronLab / CompGenomicsWS'. The repository has 1 branch and 0 tags. The main file listed is 'README.md' under the 'hadasvilk Lesson 5' commit. The README content includes:

Computational Genomics WorkShop

Instructions and assignments for the computational genomics workshop given at RUNI 2022/3

General information and requirements

Don't forget to bring your laptop to class

We start our exploration by getting deeply familiar with the most predominant method of sequencing to date, Next Generation Sequencing (NGS). We will learn how high-throughput sequencing is performed and get our computational tools configured in order to search for novel cancerous mutations in the genome of a cancer patient.

For our mission to be successful we trust you already have your Unix-like OS ready, preferably a Linux distro. MacOS users should be fine to install all software and Windows users are encouraged to enable WSL and install one of the Linux distributions offered on the Microsoft Store.

To facilitate easy installation of bioinformatic tools please consider installing the Anaconda package manager (the miniconda variant is fine) and configuring the bioconda channel. A detailed (short and simple) guide on how to achieve it can be found [here](#). Class and home work will assume a functional conda environment, but tech savvy students should be fine with whatever method they choose to get binaries for their system. Please make sure you have ~20Gb of available storage on your machine, mainly for the sequencing data and auxiliary files.

In the following weeks class time will be split into two sections, we will start with a lecture followed by hands-on time for an assignment you'd have to submit up to a week after, so don't forget to bring your laptop to class.

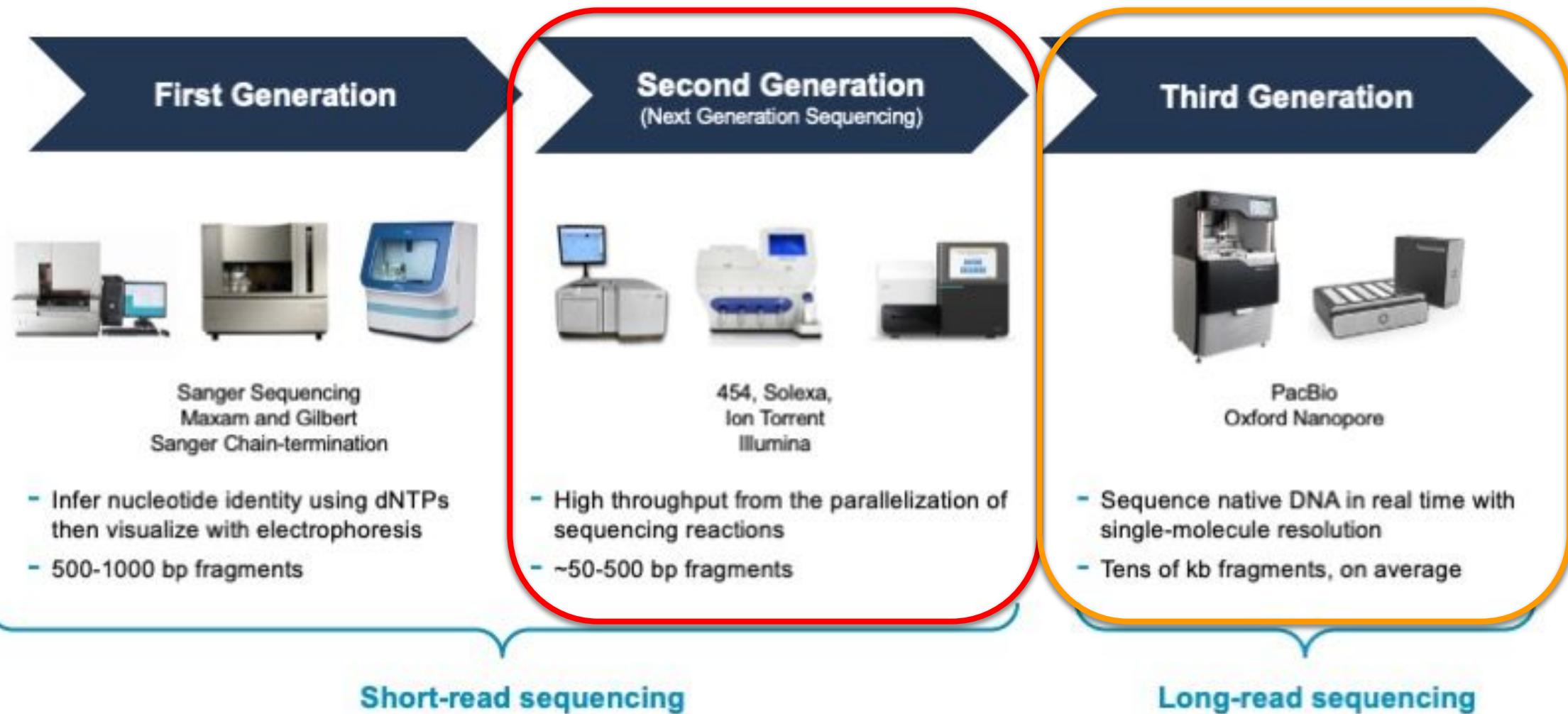
Lesson 5

[Class Slides](#)

Tools to be used today

The right sidebar contains sections for About, Releases, and Packages, each with their respective details.

The Evolution of DNA Sequencing Tools



The First Sequencing Efforts

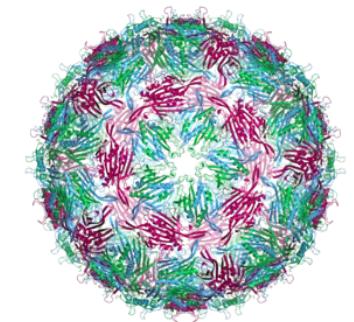
Highly abundant ssRNA - rRNA, tRNA, phages

1965 - first full sequence of yeast tRNA¹

1972 - first protein-coding sequence²

1976 - first full (RNA) phage genome³ - 3,569 nucleotides

Main technology - 2D fractionation of radioactive nucleotides⁴

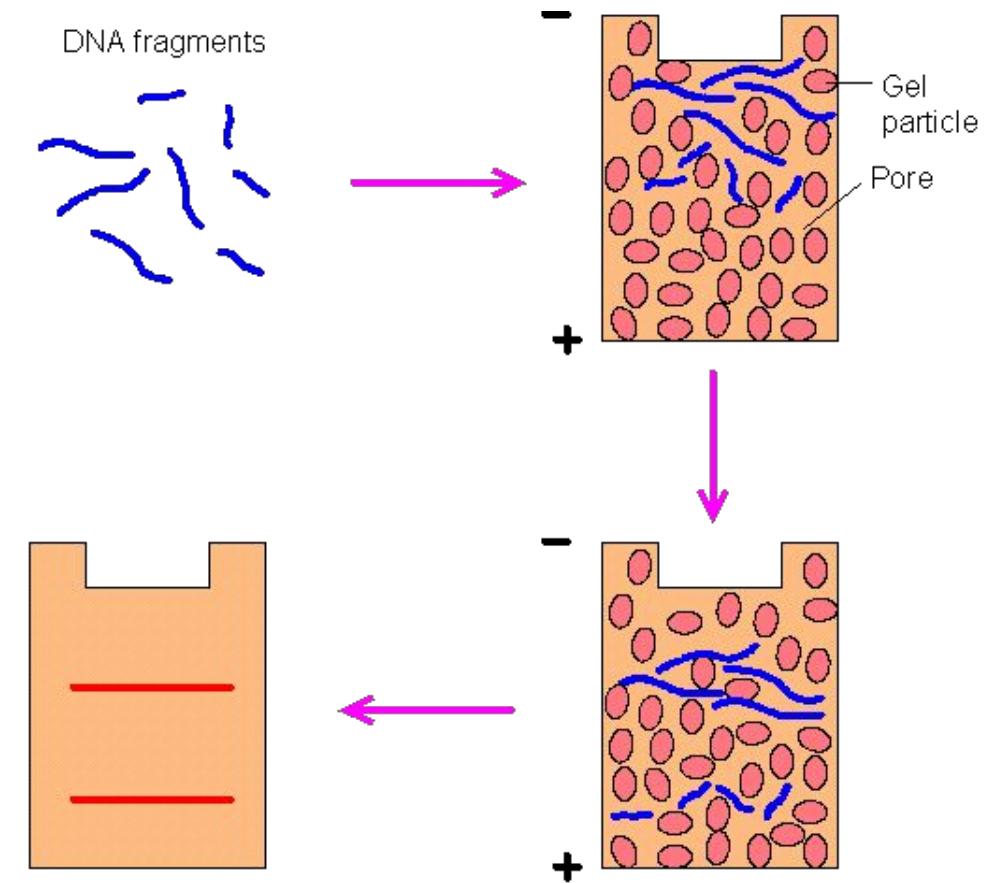


1. Holley, Robert W., et al. "Structure of a ribonucleic acid." *Science* (1965): 1462-1465.
2. Jou, W. Min, et al. "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein." *Nature* 237.5350 (1972): 82.
3. Fiers, Walter, et al. "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene." *Nature* 260.5551 (1976): 500.
4. Sanger, F., G. G. Brownlee, and B. G. Barrell. "A two-dimensional fractionation procedure for radioactive nucleotides." *Journal of molecular biology* 13.2 (1965): 373-IN4.

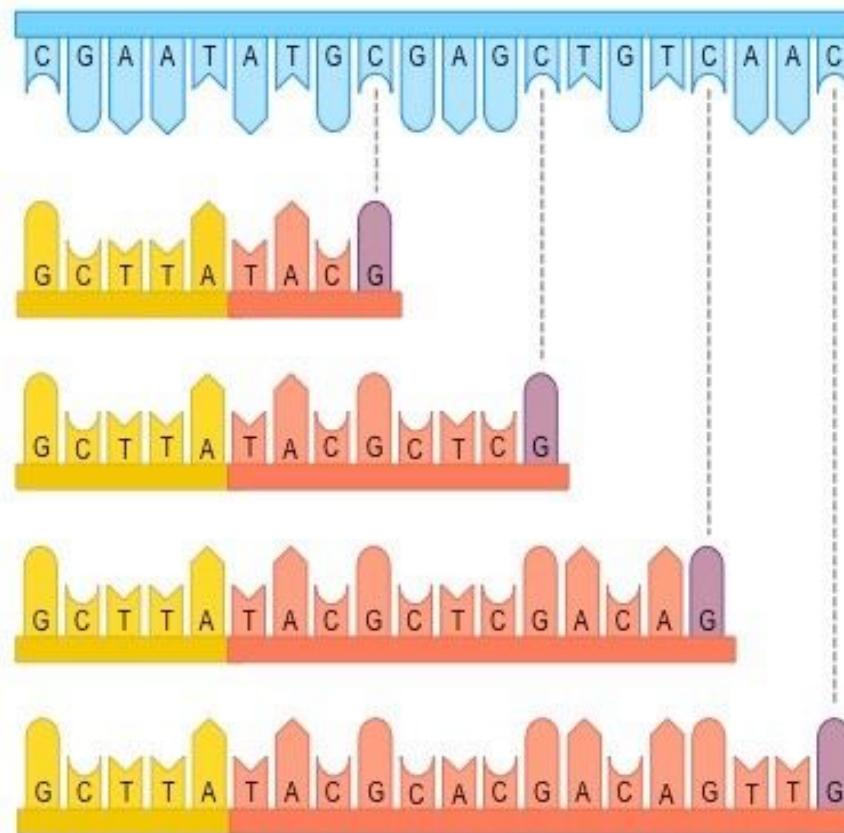
1970's - 1st Generation Sequencing

- Chain termination method¹
- Chemical cleavage method²
- Both used polyacrylamide gels
- First DNA phages sequenced

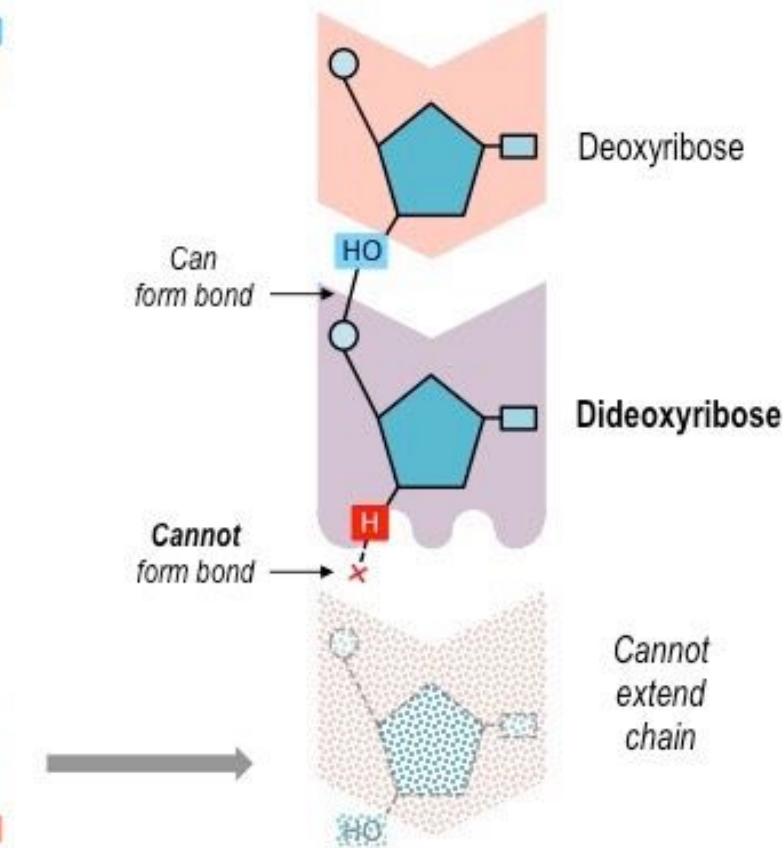
1. Sanger, Frederick, Steven Nicklen, and Alan R. Coulson. "DNA sequencing with chain-terminating inhibitors." *Proceedings of the national academy of sciences* 74.12 (1977): 5463-5467.
2. Maxam, Allan M., and Walter Gilbert. "A new method for sequencing DNA." *Proceedings of the National Academy of Sciences* 74.2 (1977): 560-564.



1970's - 1st Generation Sequencing



Sequence terminates when the ddNTP is incorporated
Fragment lengths reflect base position in sequence



Chain termination by
dideoxynucleotides

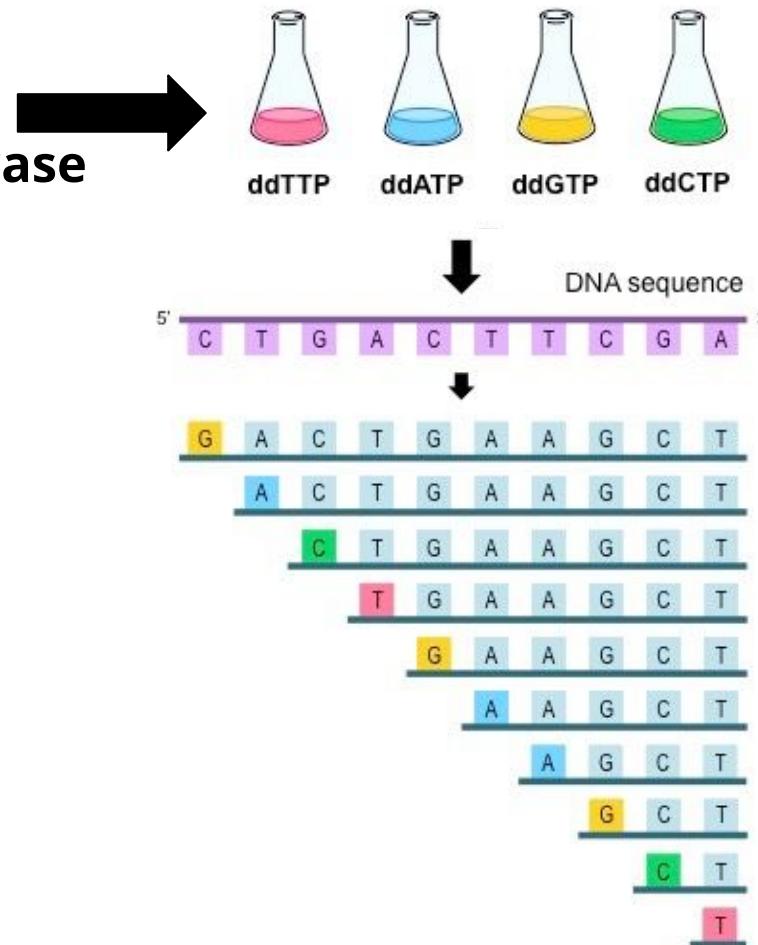
The Chain Termination Method - Sanger Sequencing

Target DNA

Primer

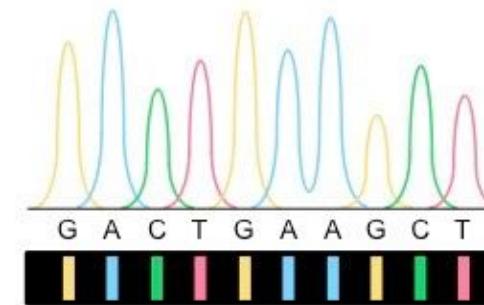
DNA Polymerase

4 dNTPs



Use a sequencing machine

Separate with a gel



The Chain Termination Method - Sanger Sequencing

SBS - sequencing by synthesis

Produces reads 500-1,000 bp long

First commercial machines

Still used today!

Can sequence ~70k bp/hour

Cost: ~\$500/1Mbp

Used for various whole genome projects

Shotgun sequencing



The Human Genome Project

Reminder: human genome size is ~3Gb

Started 1990

First genome draft - 2000

Project completion - 2003

Largest ever biological collaborative project

20 sequencing centers around the world

Entirely based on Sanger sequencing

Estimated cost: \$5 billion



2000s - 2nd Generation Sequencing

Pyrosequencing¹ - “454 sequencing”

Illumina (Solexa) method

Allow massively parallel sequencing

Produce short reads - usually 50-250 bp

Deep sequencing - each genomic region is sequenced multiple times

1. Ronaghi, Mostafa, Mathias Uhlén, and Pål Nyrén. "A sequencing method based on real-time pyrophosphate." *Science* 281.5375 (1998): 363-365.
2. Canard, Bruno, and Robert S. Sarfati. "DNA polymerase fluorescent substrates with reversible 3'-tags." *Gene* 148.1 (1994): 1-6.

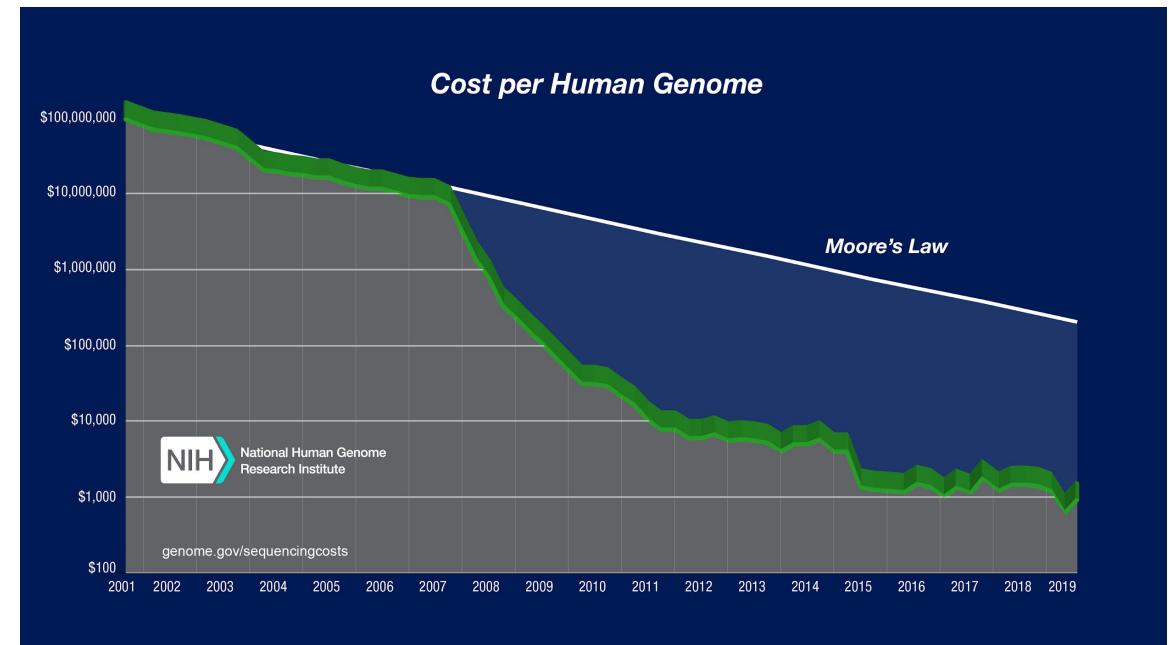
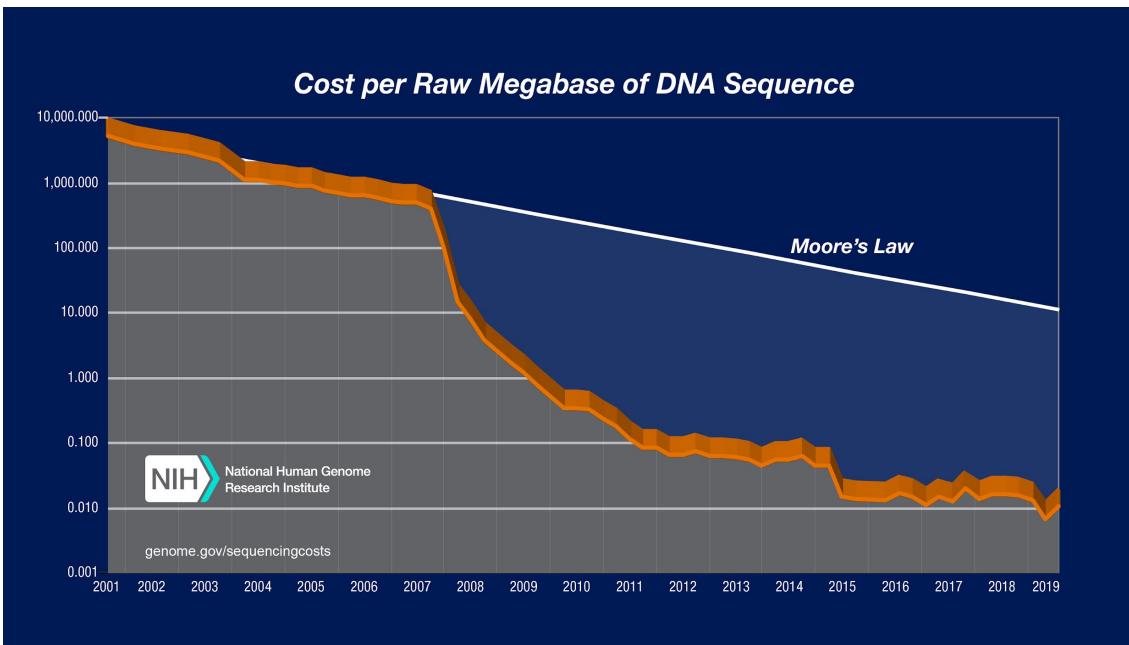
Illumina Technologies - Standard for NGS

By far the most popular sequencing technology

Up to 120 Gb/hour

Significantly reduced sequencing costs

Allowed the sequencing of numerous species and samples



 Ultima Genomics reposted this

• • •

Ultima Genomics



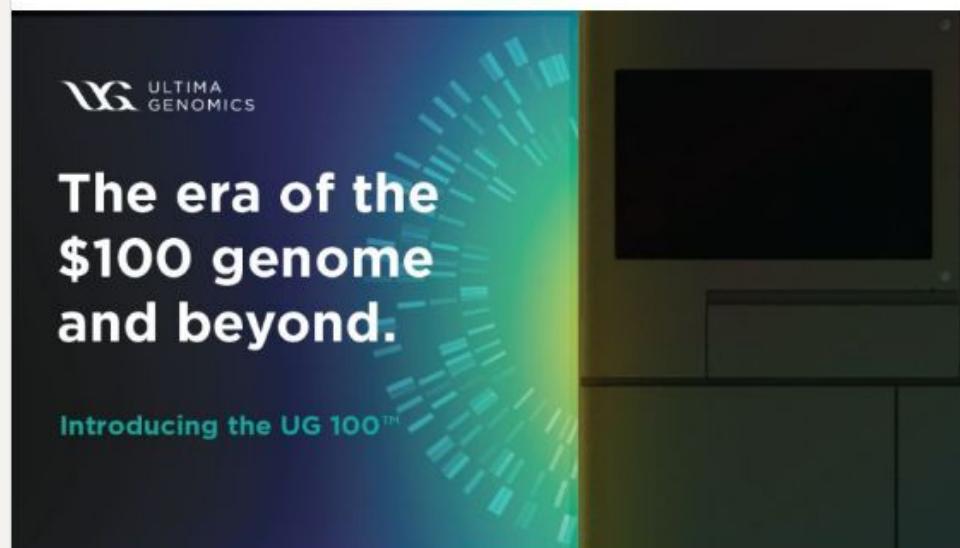
14,006 followers

2d • Edited • 

Ultima is gearing up for the full commercial availability of the UG 100™ at the upcoming **AGBT - Advances in Genome Biology and Technology** (AGBT) on February 5, 2024. We will also be unveiling our ppmSeq™ Technology with industry-leading raw read accuracy for calling SNVs for rare event detection applications.

Hear why our early access customers are excited about the launch of the UG 100™ and the era of the \$100 genome and beyond.

https://lnkd.in/gANE_9Cv



Ultima Genomics Announces End of Early Access and Upcoming Launch of UG 100™

[businesswire.com](#) • 5 min read

 David Bogumil and 224 others

9 comments • 34 reposts

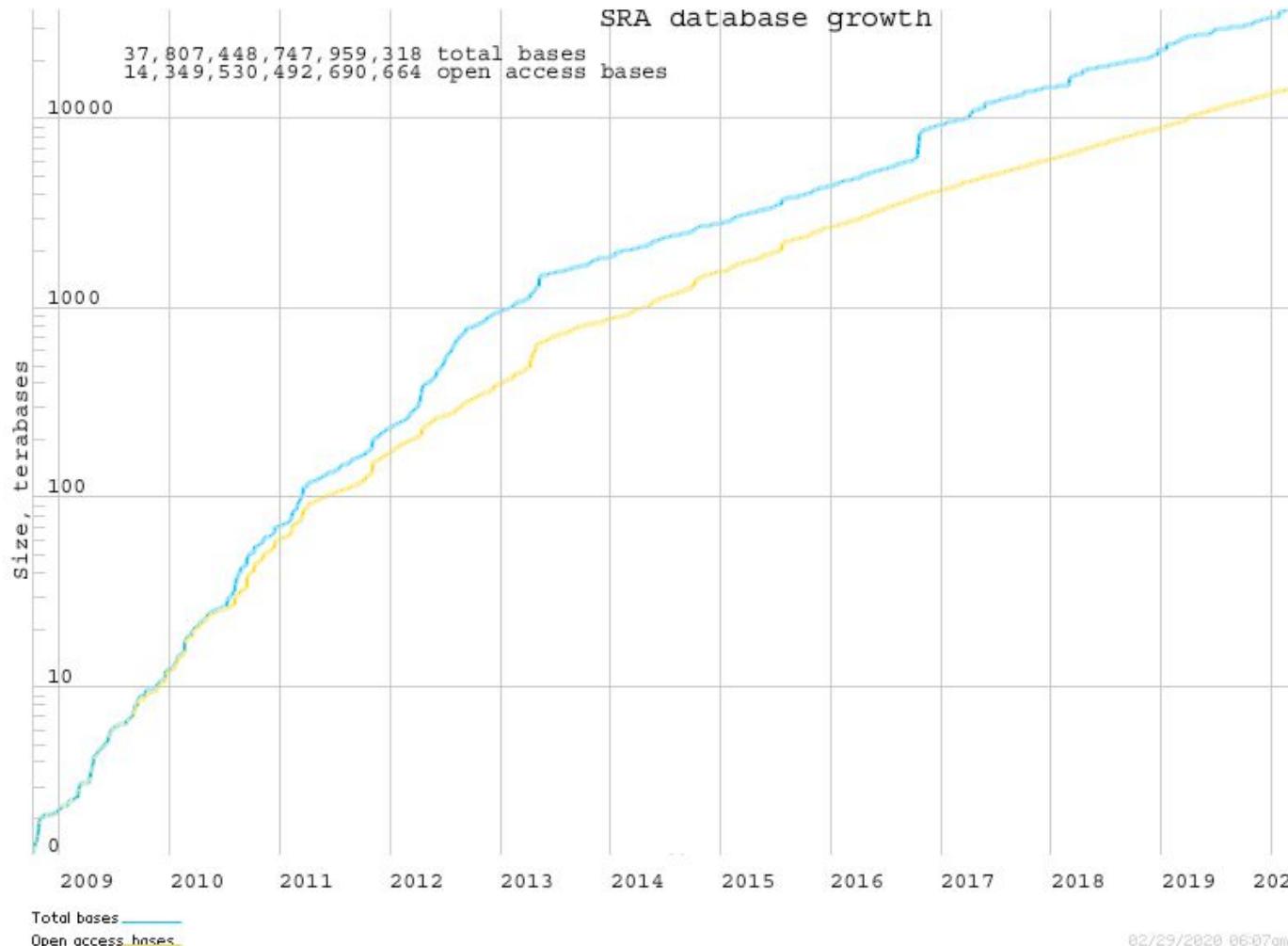
 Like

 Comment

 Repost

 Send

Number of Bases Deposited in SRA



Some Sequencing Projects From 2022

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature communications](#) > [articles](#) > article

Article | Open Access | Published: 04 March 2022

Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil



nature genetics

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature genetics](#) > [articles](#) > article

Article | Open Access | Published: 22 September 2022

Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics



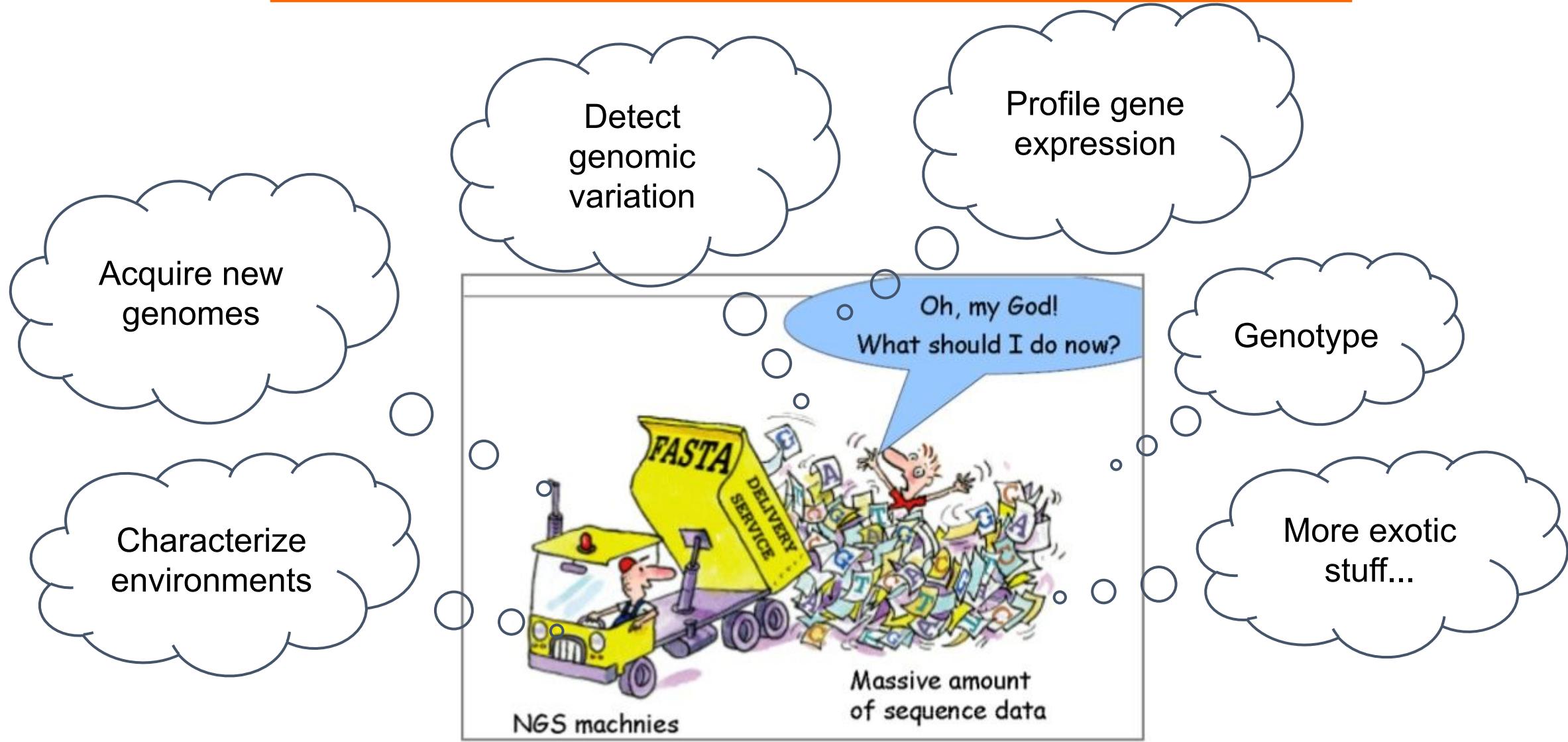
[Zool Res.](#), 2022 Jan 18; 43(1): 78–80.
doi: [10.24272/j.issn.2095-8137.2021.266](https://doi.org/10.24272/j.issn.2095-8137.2021.266)

PMCID: PMC8743251
PMID: [34877831](https://pubmed.ncbi.nlm.nih.gov/34877831/)

Whole-genome resequencing infers genomic basis of giant phenotype in Siamese fighting fish (*Betta splendens*)

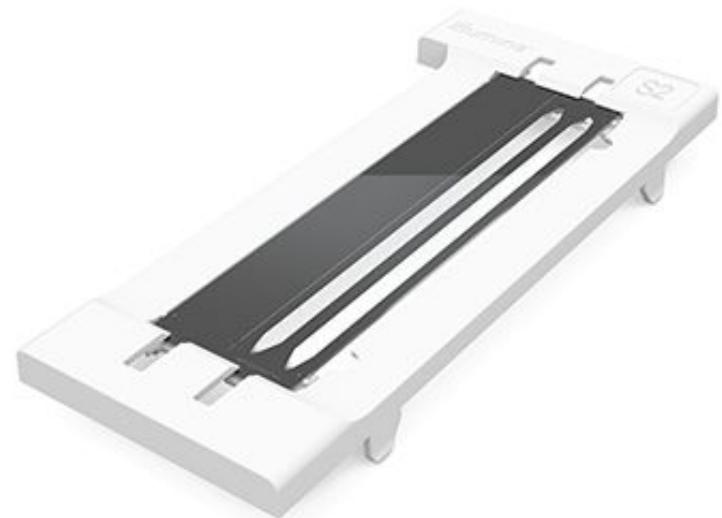


What can we do with NGS?



The Illumina Sequencing Method

The Illumina Sequencing Machine



DNA/RNA Extraction

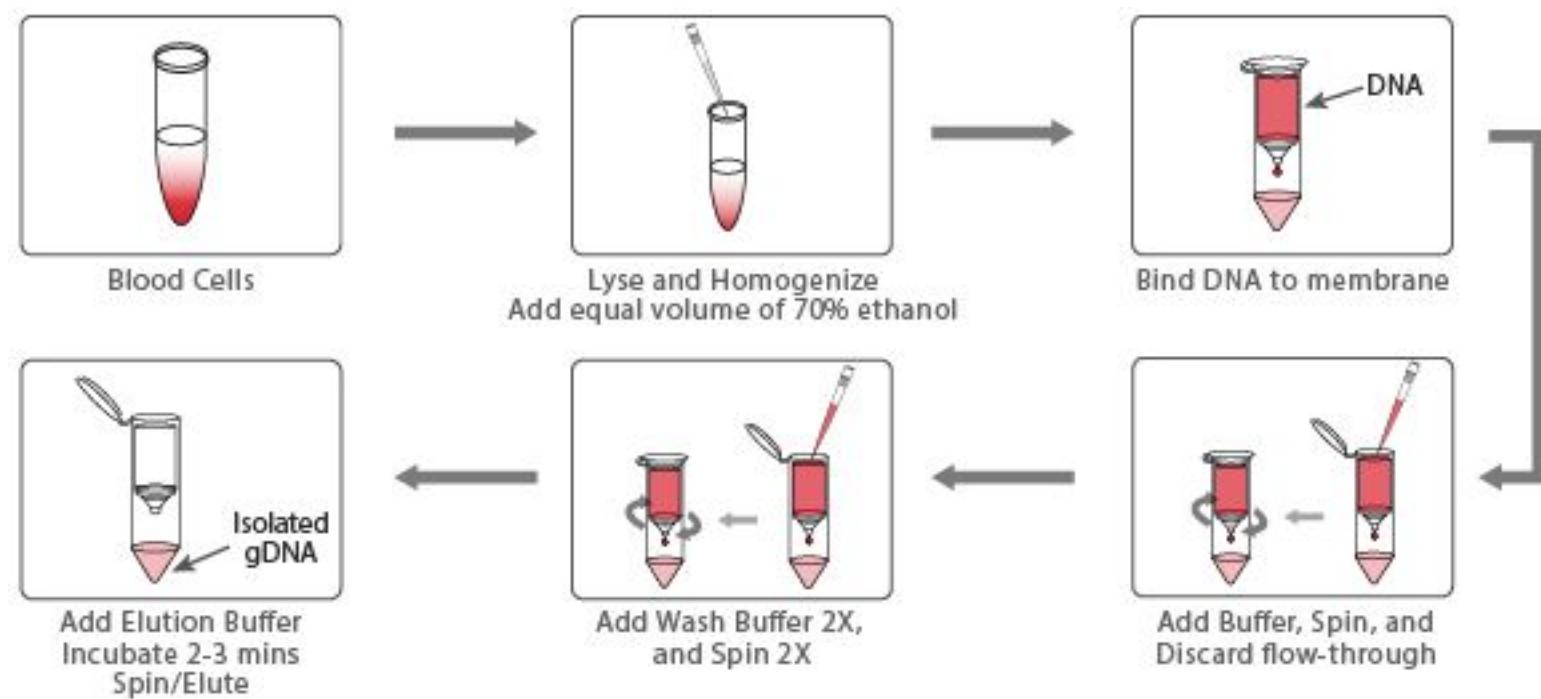
Protocol depends on organism and tissue

Required DNA amount depends on application

DNA should be as intact as possible

- Fragment sizes
- Single strand breaks

May be challenging!



Library Prep

Many protocols exist

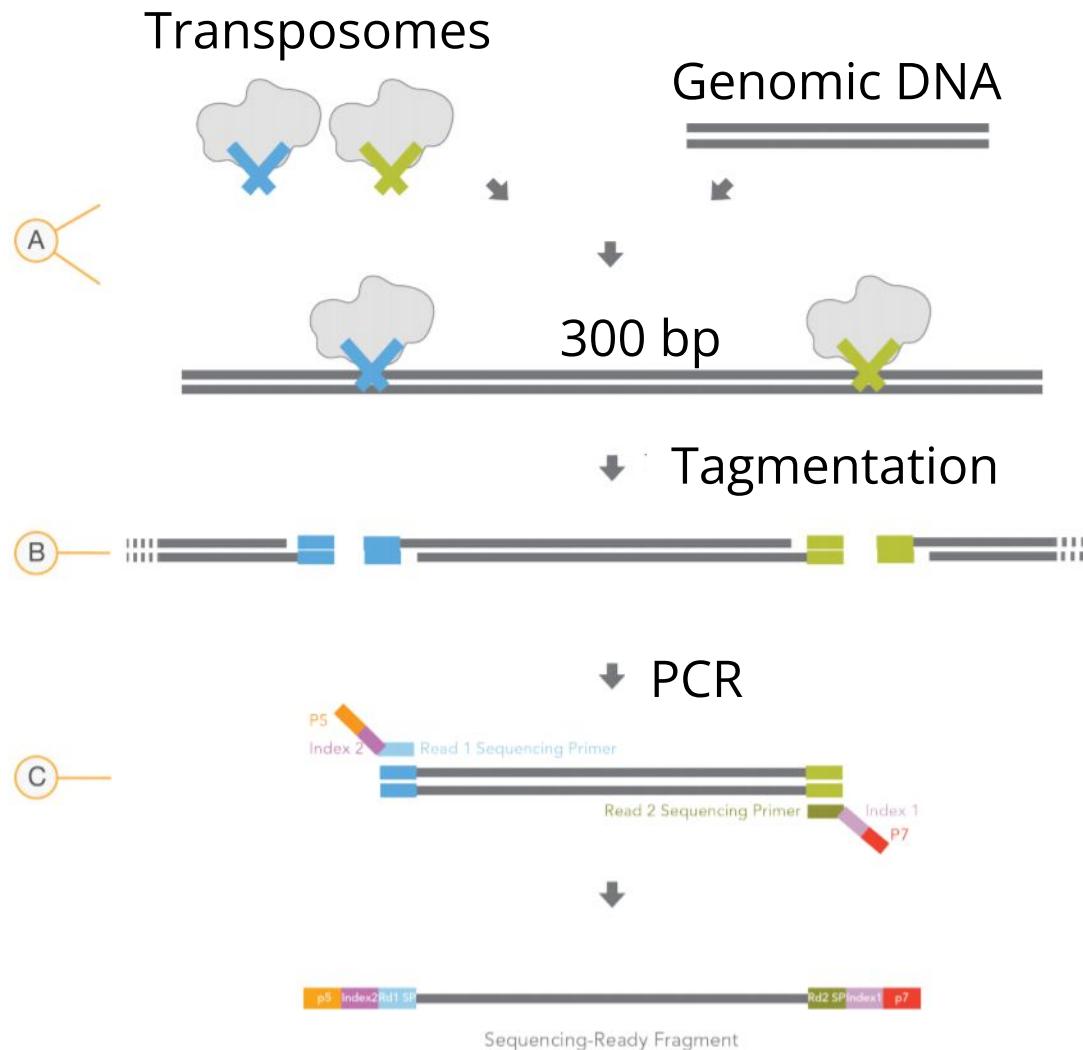
Main goal - fragment and add adapters

Single-end or paired-end?

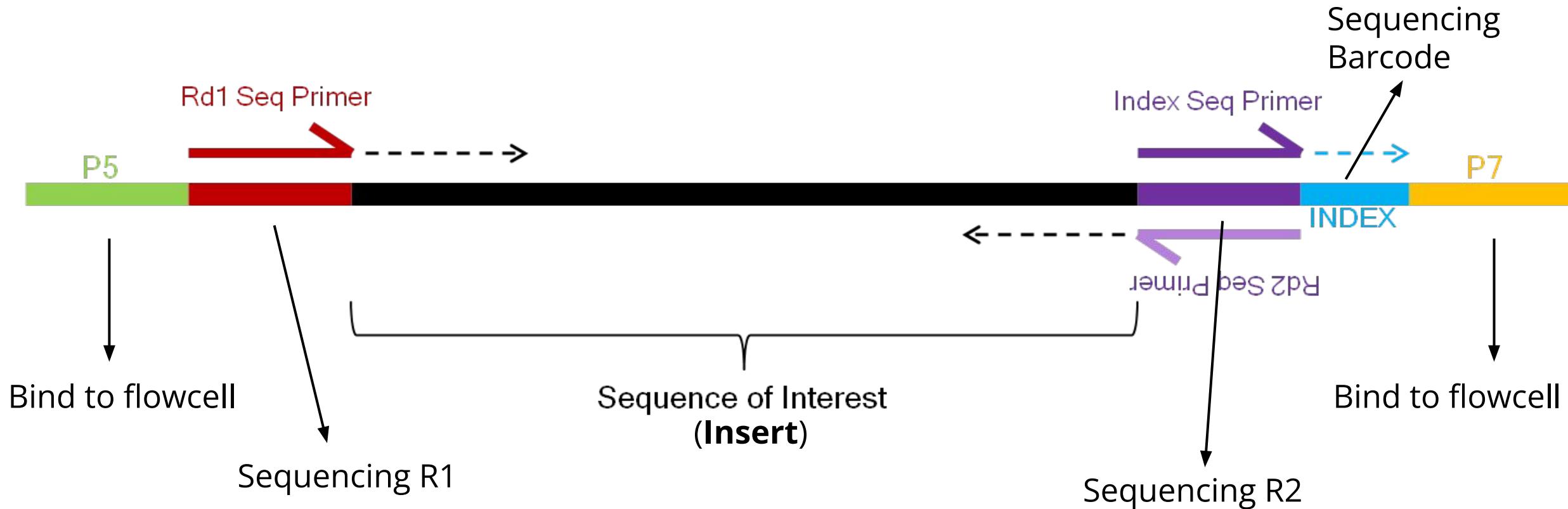
Whole genome sequencing (**WGS**) or **targeted** sequencing?

Insert size is determined

Basic WGS Library Prep Example

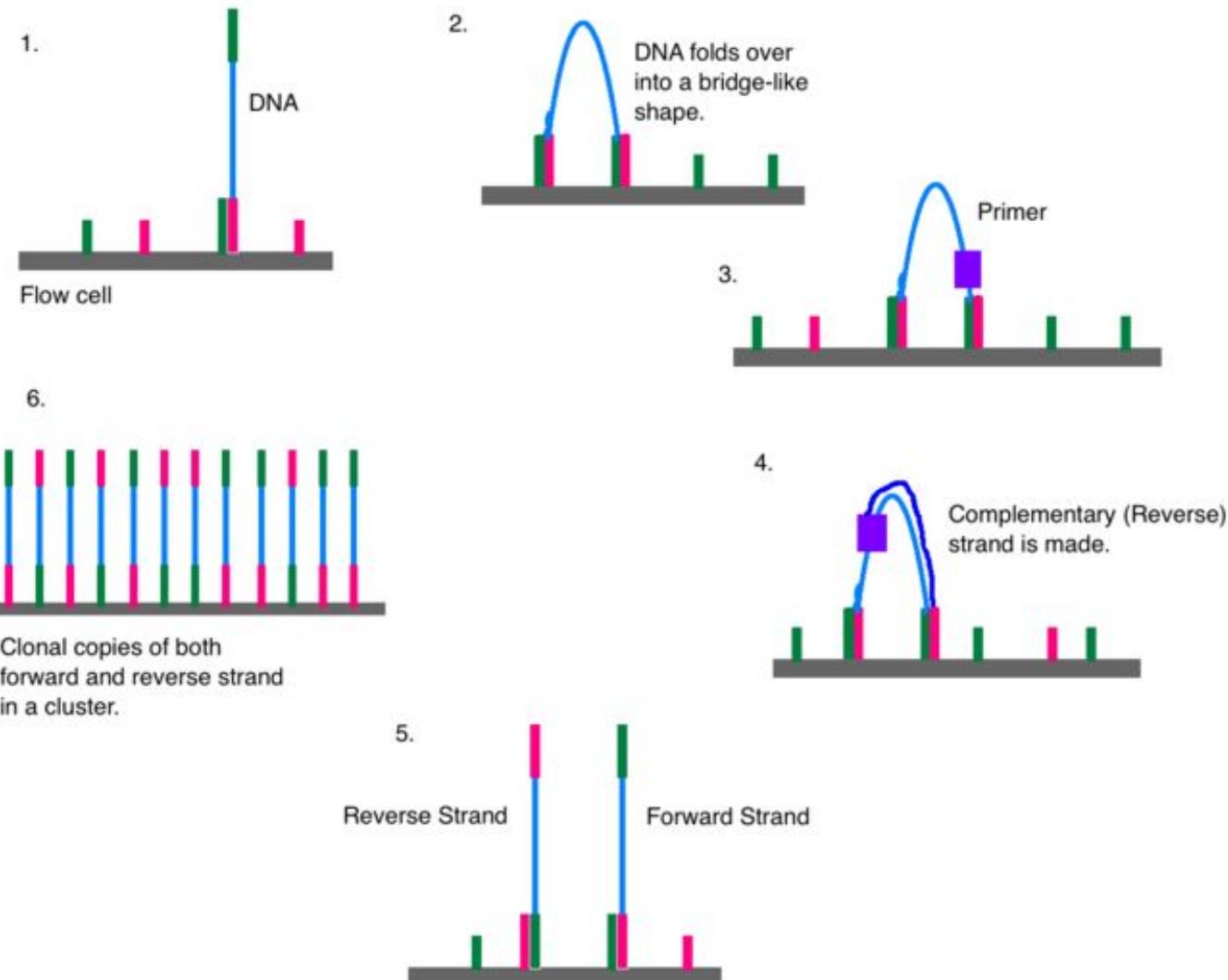


Library Prep What Do We Get



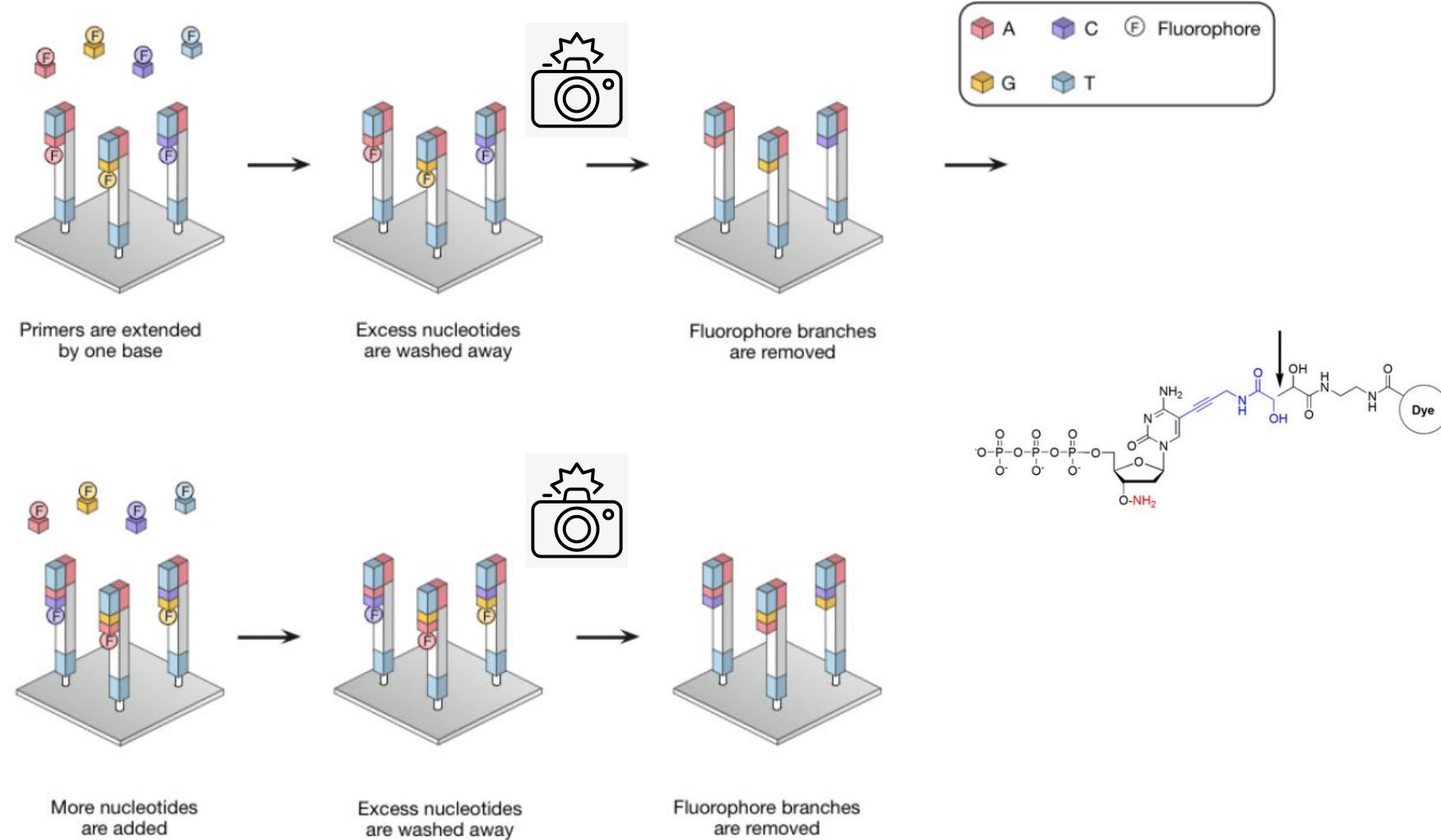
Main Steps in Sequencing

STEP 1 - Cluster amplification
No sequencing

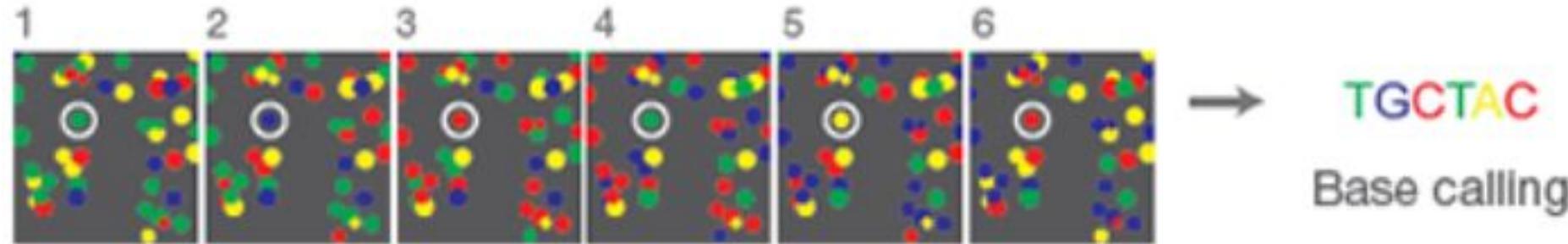


Main Steps in Sequencing

STEP 2 - Sequencing



Output of the Sequencing Machine



- Each flow cell cluster results in a read or read pair

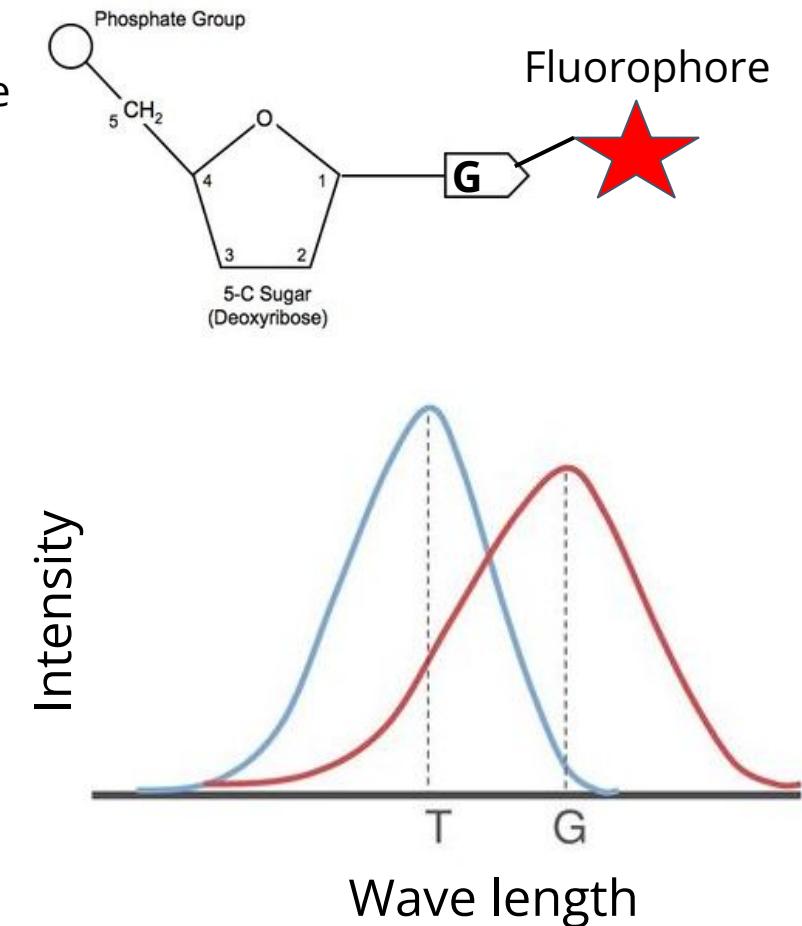
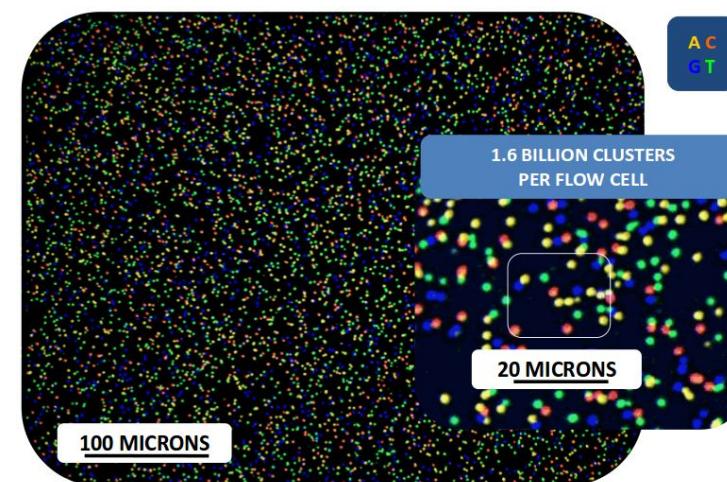
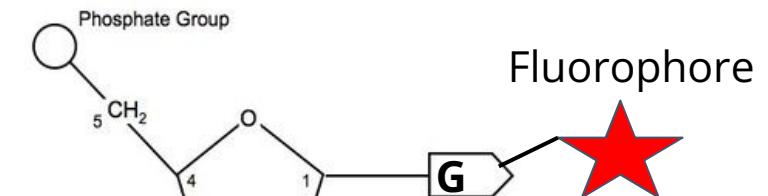
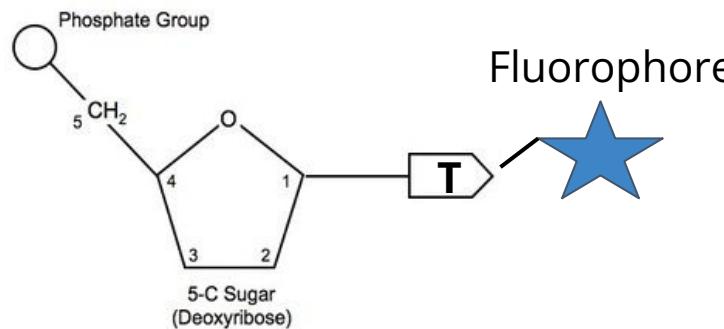
Low Quality Base Calling

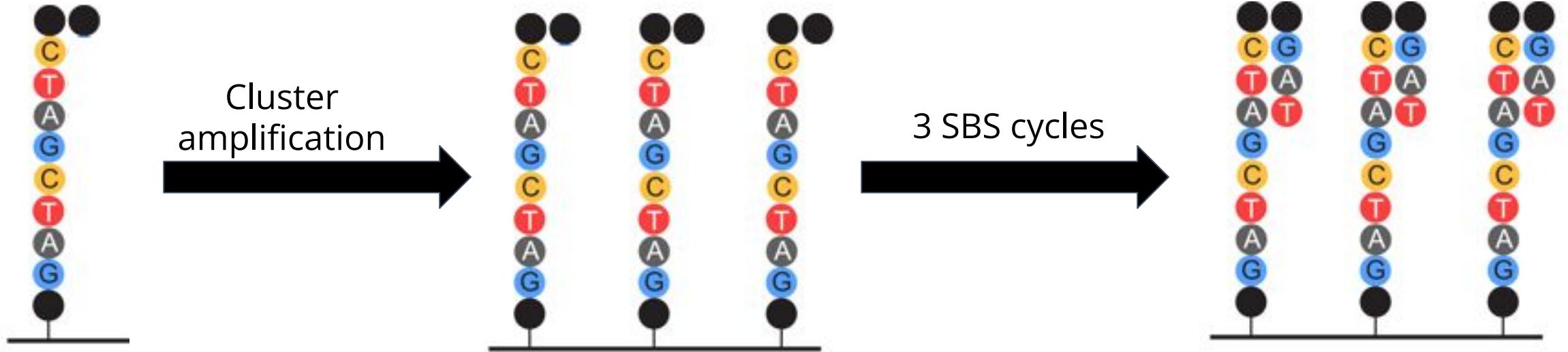
- Illumina machines make mistakes!

- Rate of errors: ~1/1000

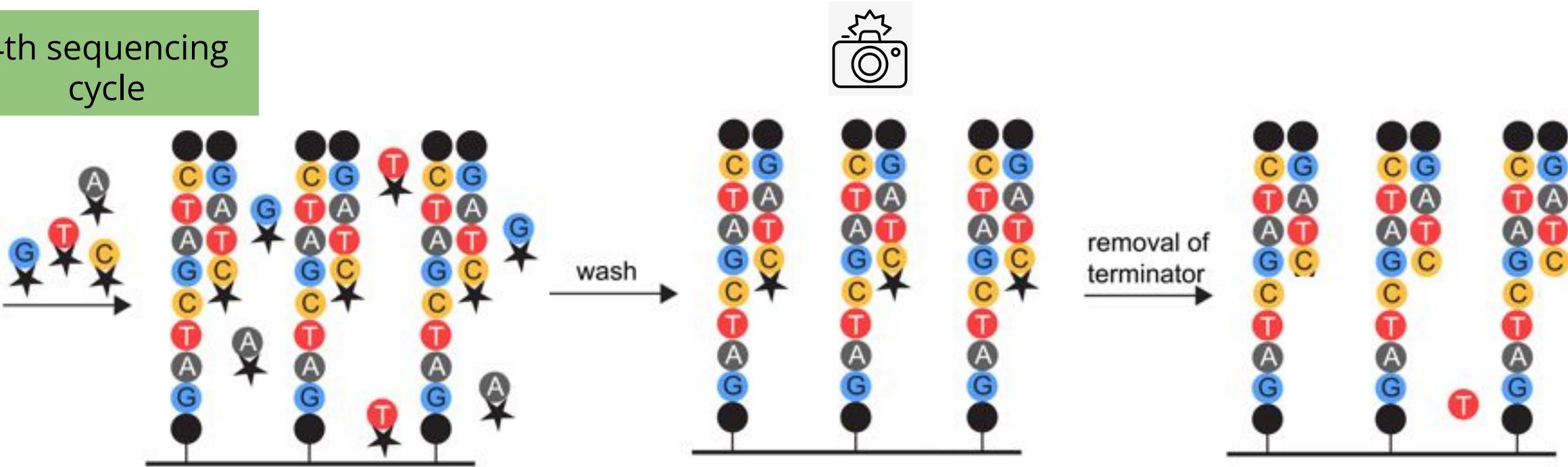
- Reasons:

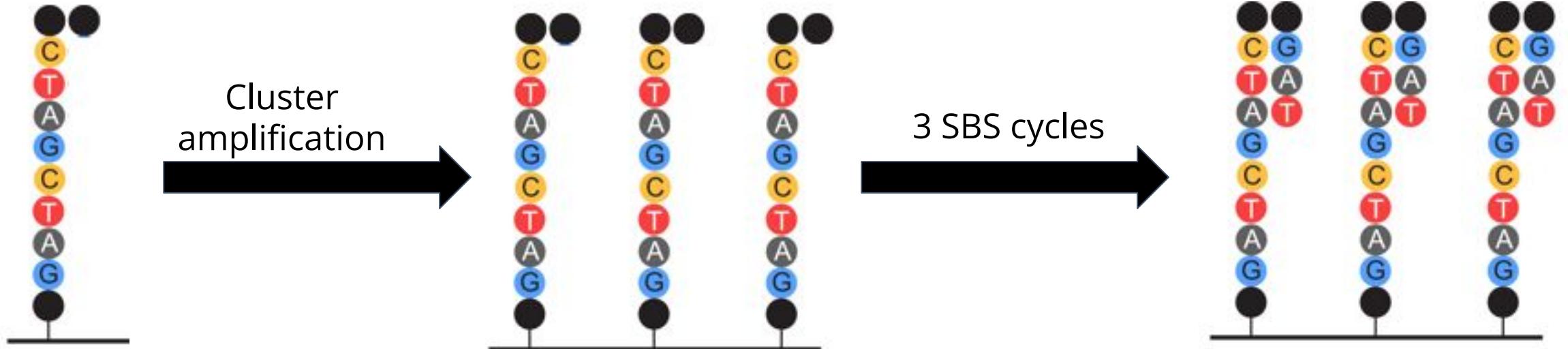
- Color cross-talk
- Clusters cross-talk
- Phasing



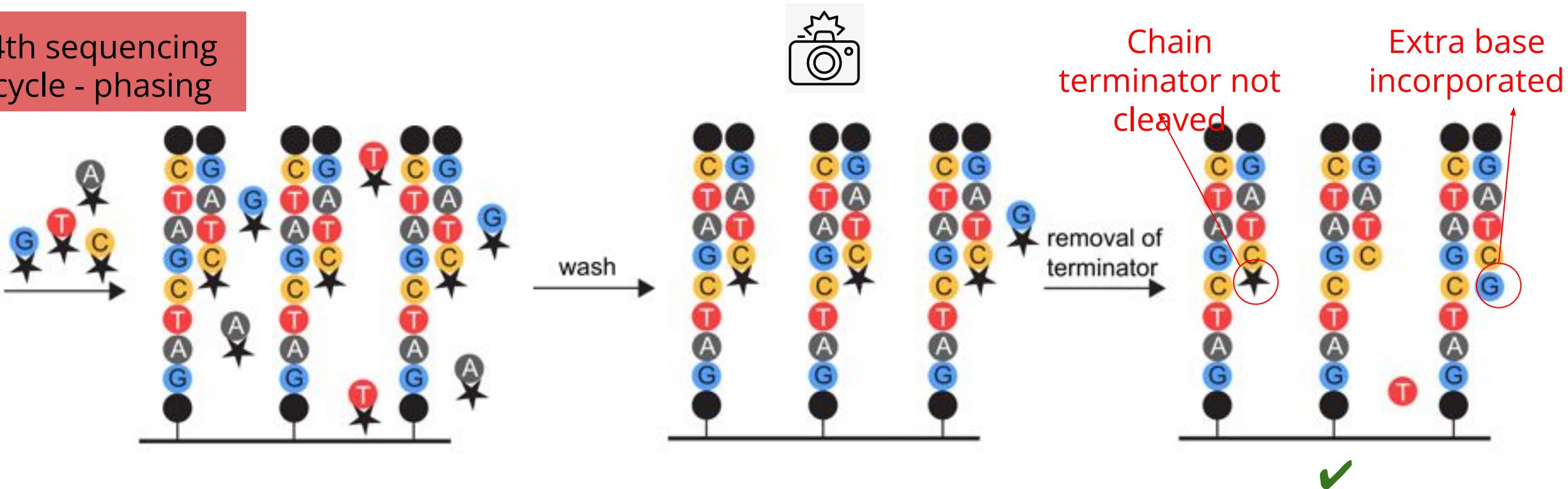


4th sequencing cycle

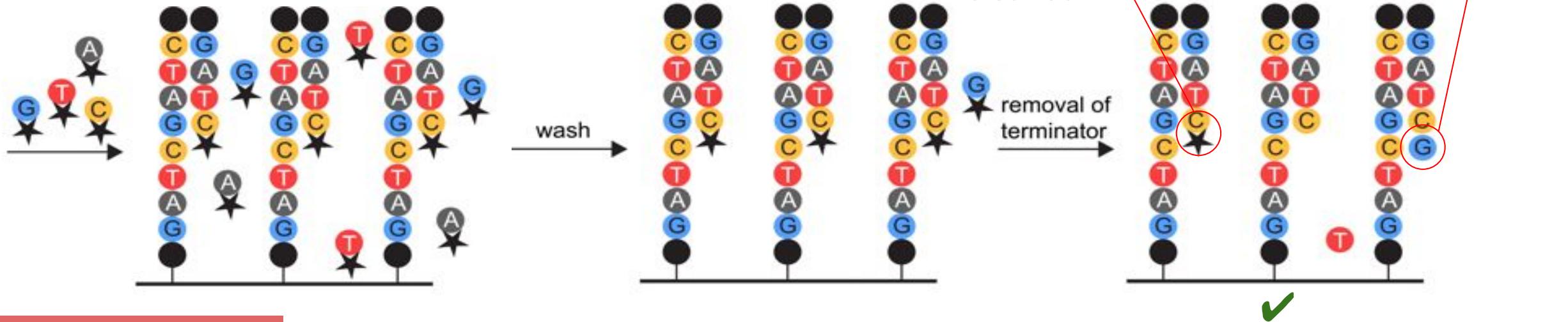




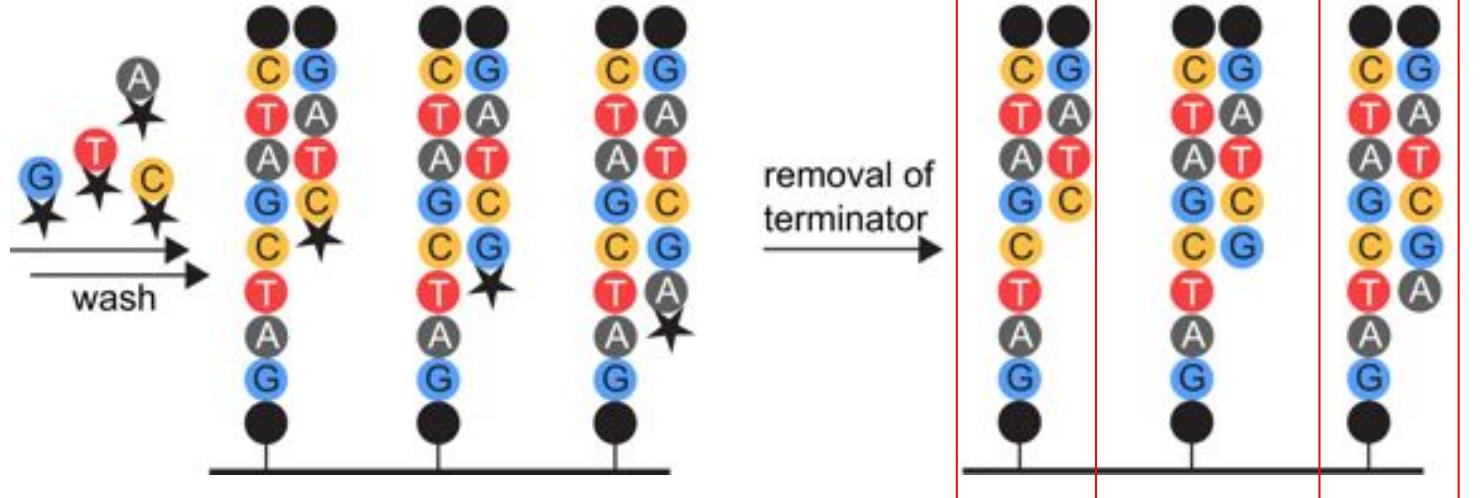
4th sequencing cycle - phasing



4th sequencing cycle - phasing



5th sequencing cycle - phasing



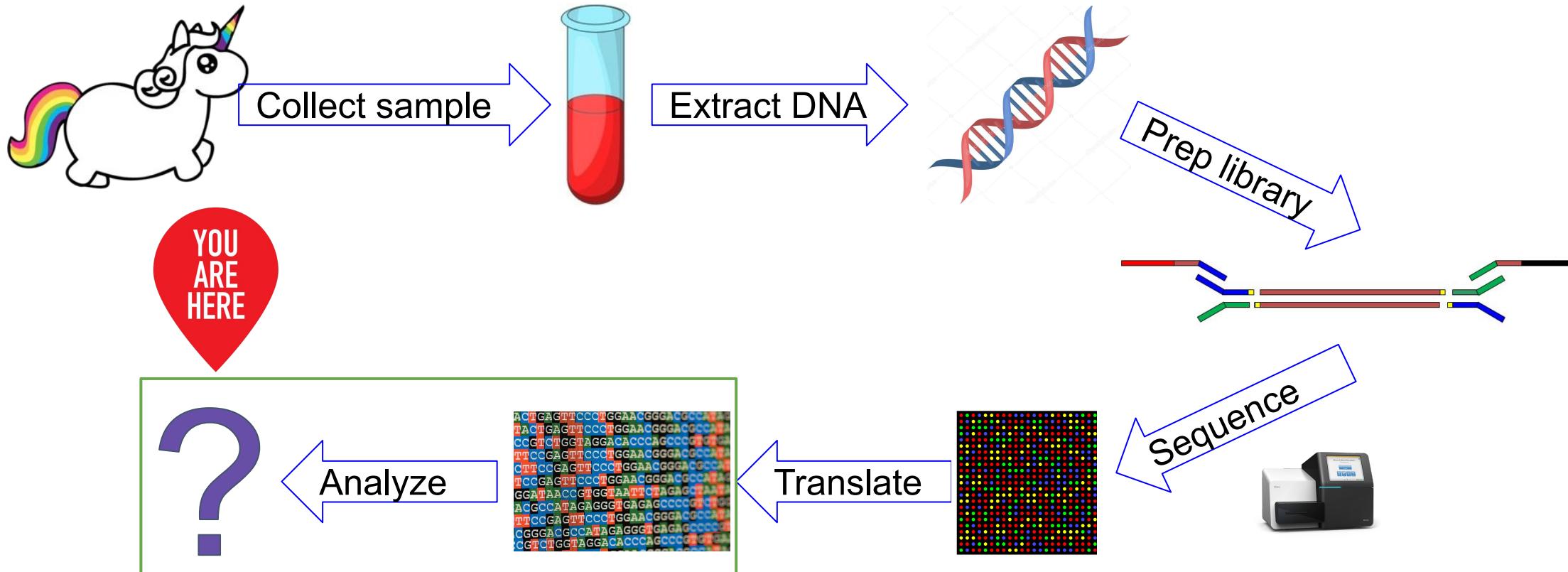
Low or Biased Yield

- Too few reads
Usually caused by non-optimal cluster density
- Some genomic fragments sequenced more often than others
Usually caused by library prep issues or PCR duplication

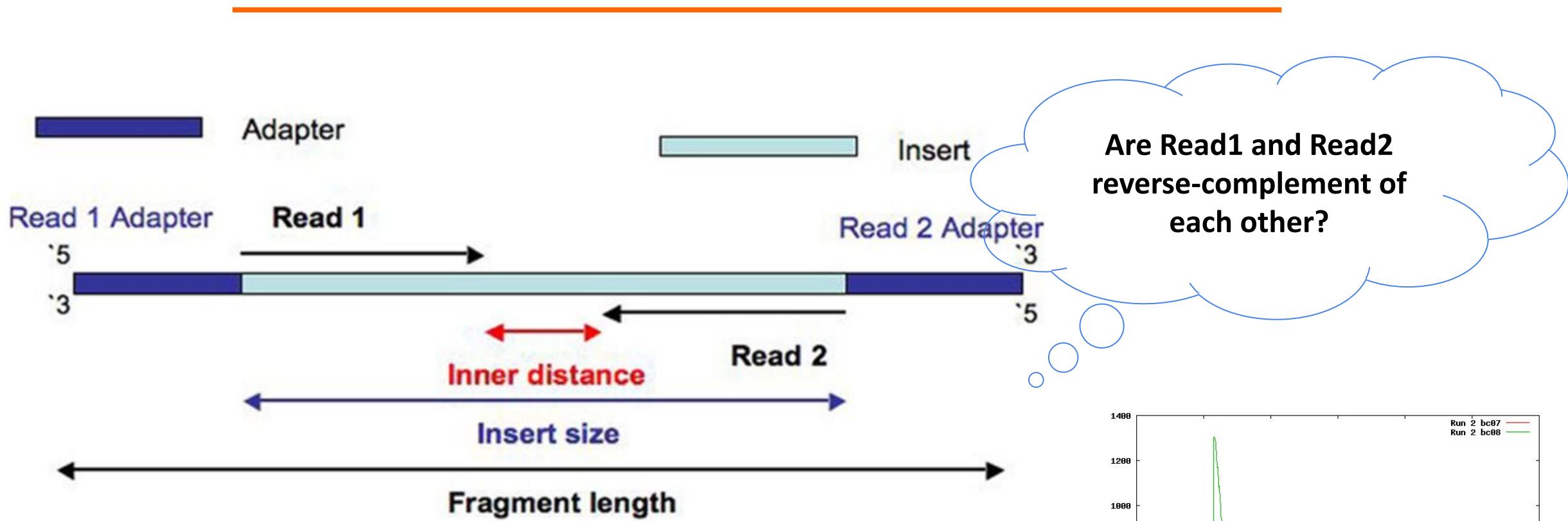
Illumina Sequencing Instruments

	iSeq	MiniSeq	MiSeq	NextSeq	HiSeq 4000	HiSeq X	NovaSeq
Run Time	9–17.5 hrs	4–24 hours	4–55 hours	12–30 hours	< 1–3.5 days	< 3 days	13–44 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb	6000 Gb
Maximum Reads Per Run	4 million	25 million	25 million	400 million	5 billion	6 billion	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 250

NGS - From Organism to Sequence

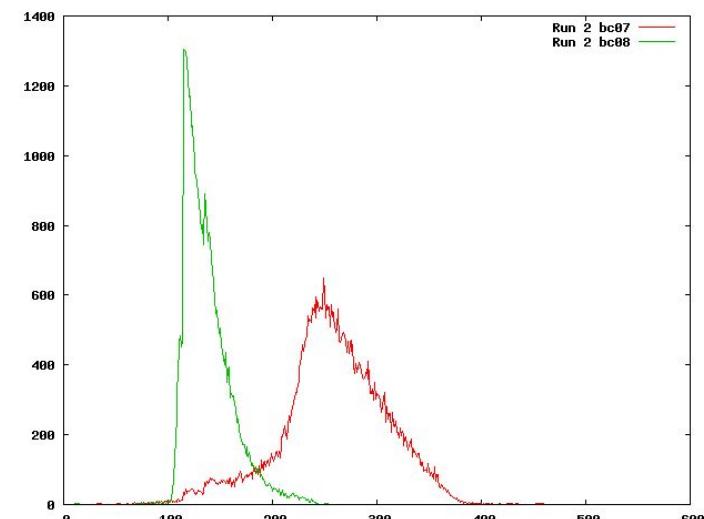


Terminology

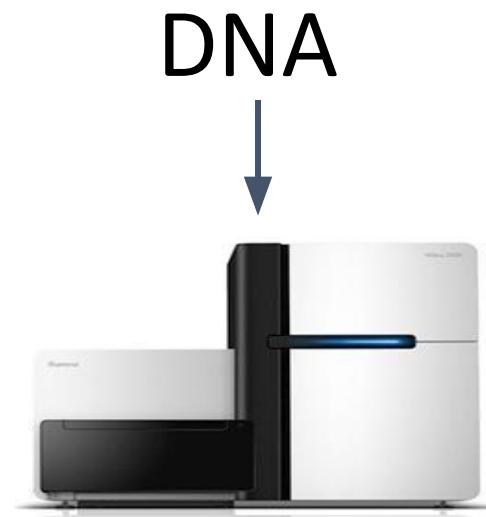


Read length - fixed - typically $50 < L < 250$

Insert size - distributed - limitation - up to 700 bp



Sequencing Yield DNA Sequences in FASTQ Format



@seq1
ACCTTCGAACGGCGGGGGG**G**TTACAA
+
! ' ' *((((*++))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
! ' ' *!!!!***+))+++++.1***
And so on...

The FASTA Format

FASTA format is the most basic format for reporting a sequence and is accepted by almost all sequence analysis program

Sequence ID >1 dna : chromosome chromosome : GRCh38 : 1 : 1 : 248956422 : 1 REF
Sequence CGGTGGCTCACGCCTGTAATCCCAACACTTGGGAGGCCAAAGCAGGTGGATTACCTGAG
ATCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCTGTCTACTAAAAATACA

en.wikipedia.org/wiki/FASTA_format

The FASTQ Format

A “standard” format for storing and defining sequences from next-generation sequencing technologies.

Sequence ID @SEQ_ID

Sequence GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT

<separator> +

Quality scores ! ' ' * (((***+)) % % % ++) (% % % %) . 1 *** - + * ' ')) ** 55CCF >>>>> CCCCCCCC65

en.wikipedia.org/wiki/FASTQ_format

The FASTQ Format's Sequence Identifier

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (<i>paired-end or mate-pair reads only</i>)
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

FASTQ Quality Scores: Estimate of Confidence in Each Base

Sequence ID @SEQ_ID

Sequence GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT

<separator> +

Quality scores ! ' ' * (((***+)) % % % ++) (% % % %) . 1 *** - + * ' ')) **55CCF>>>>>CCCCCCC65



Qualities are based on the Phred scale and are *encoded*

$$Q = -10 \cdot \log_{10}(P_{\text{err}})$$

The Ph in Phred comes from Phil Green, [the inventor of the encoding](#)

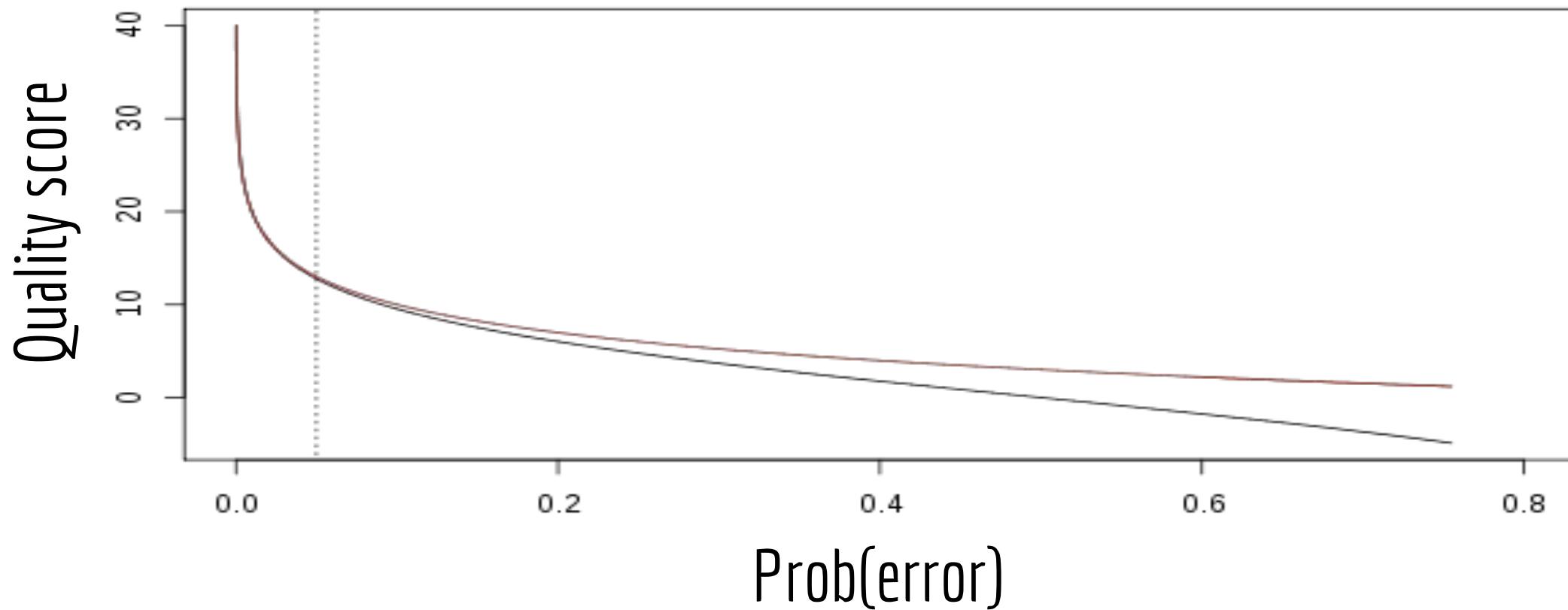
Phred Quality Score Calculation

$$Q = -10 * \log_{10}(P_{\text{err}})$$

Error probability (P_{err})	$\log_{10}(P_{\text{err}})$	Phred quality score
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40

A Higher Quality Score is Better

≥ 20 is considered "good"



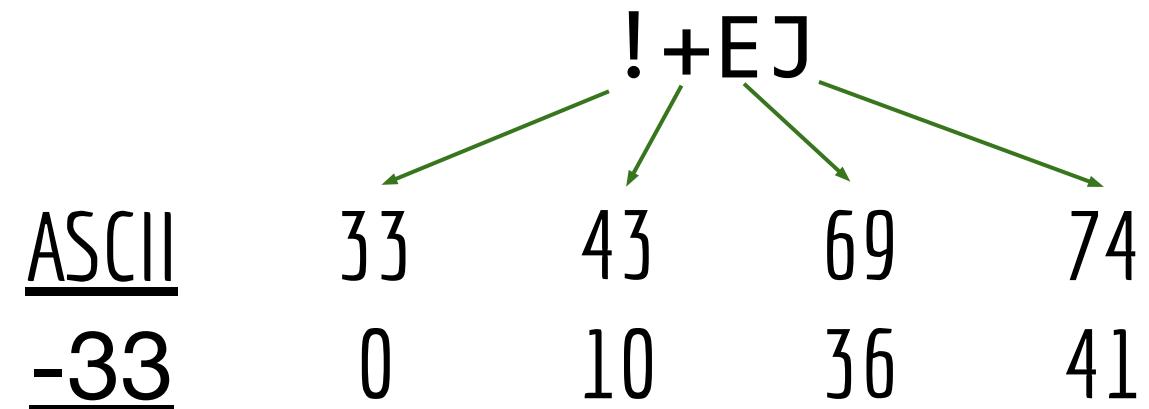
Q-Score Encoding - ASCII Table

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	Ø	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	\$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(72	48	H	104	68	h
9	09	Horizontal tab	41	29)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	Ø	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	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End 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	Substitution	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	Ø

Formula for getting PHRED quality from encoded quality:

$$Q = \text{ascii}(\text{char}) - 33$$

Example:



Is My Data Any Good?

FastQC

A common first step when getting your fastq data

Generates a simple HTML report

Can help detect issues with the data

Running:

```
fastqc <file1.fastq> <file2.fastq> ... <file n.fastq>
```

For more options:

```
fastqc -h | less
```

The FastQC Report

View using your favorite web browser

Contains multiple analysis modules

Issues “Warning” and “Failure” messages per module

Remember to look both at R1 and R2

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



Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)

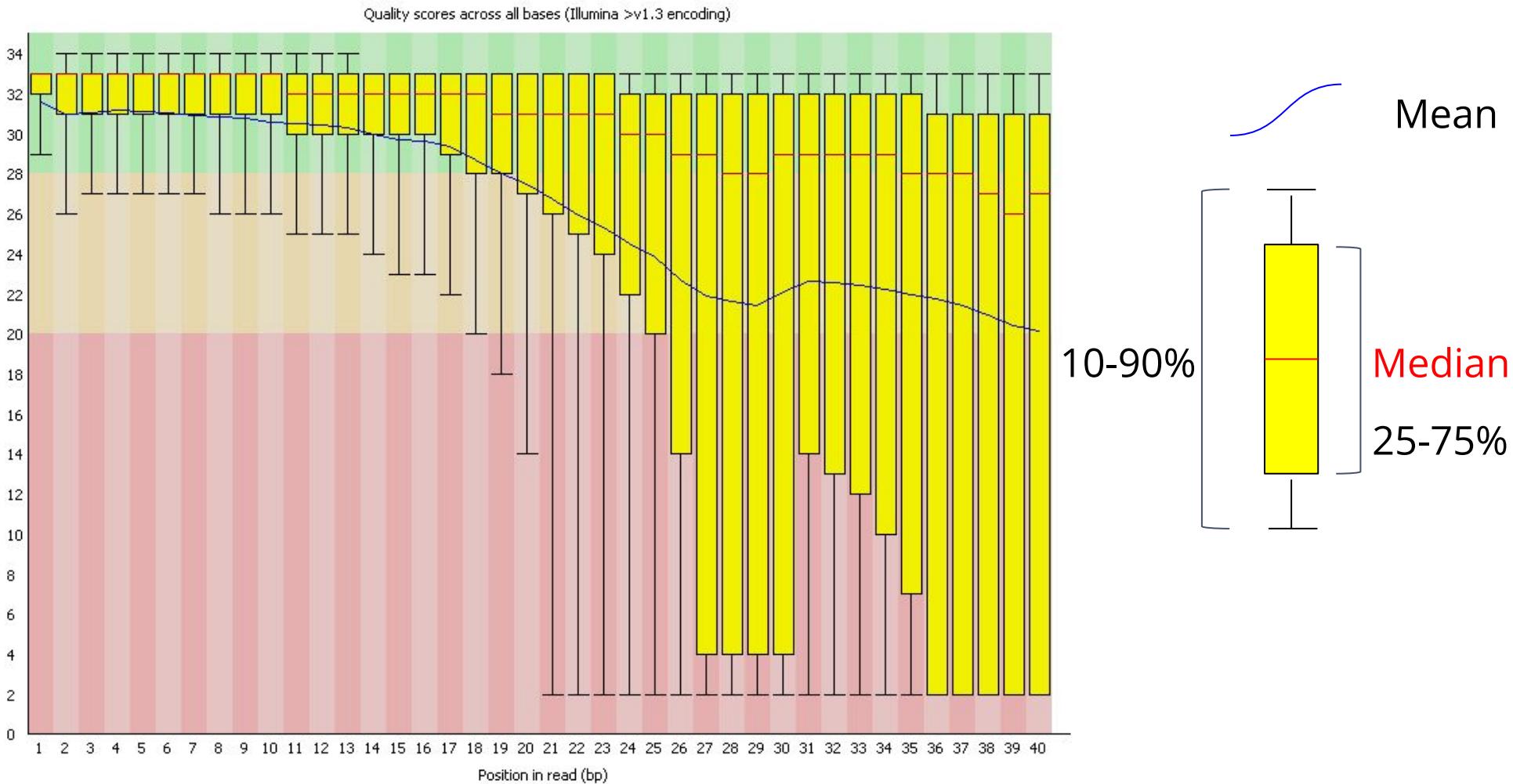
Basic Statistics



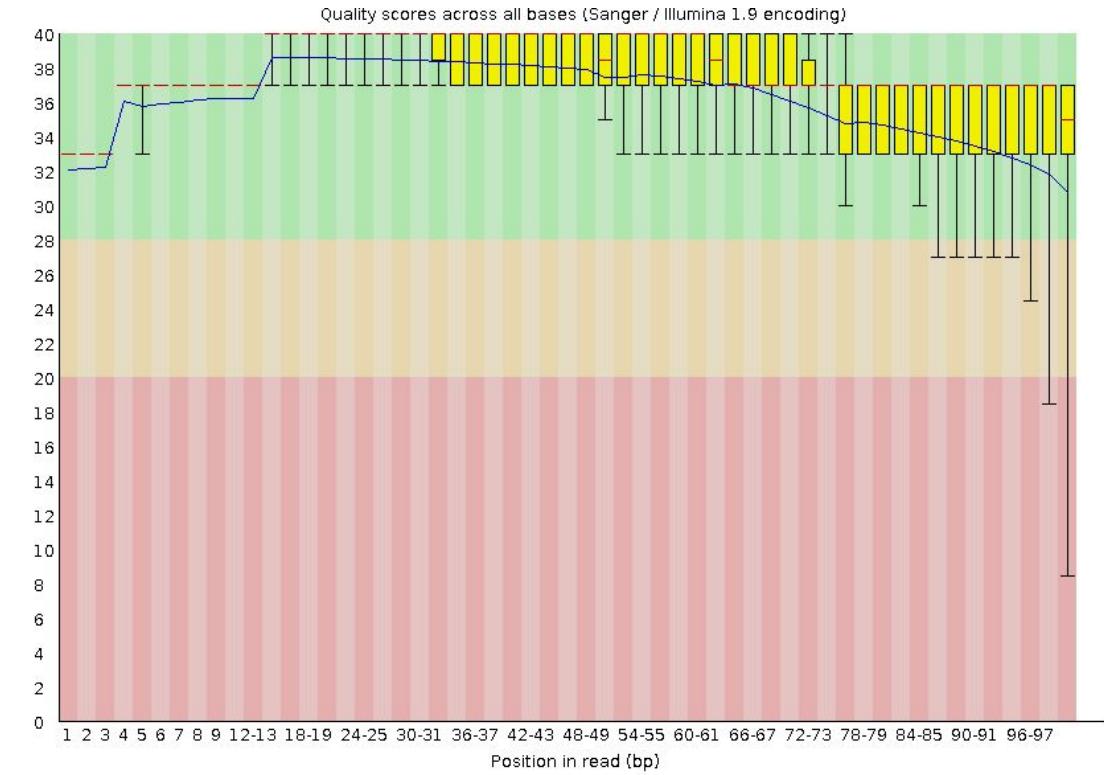
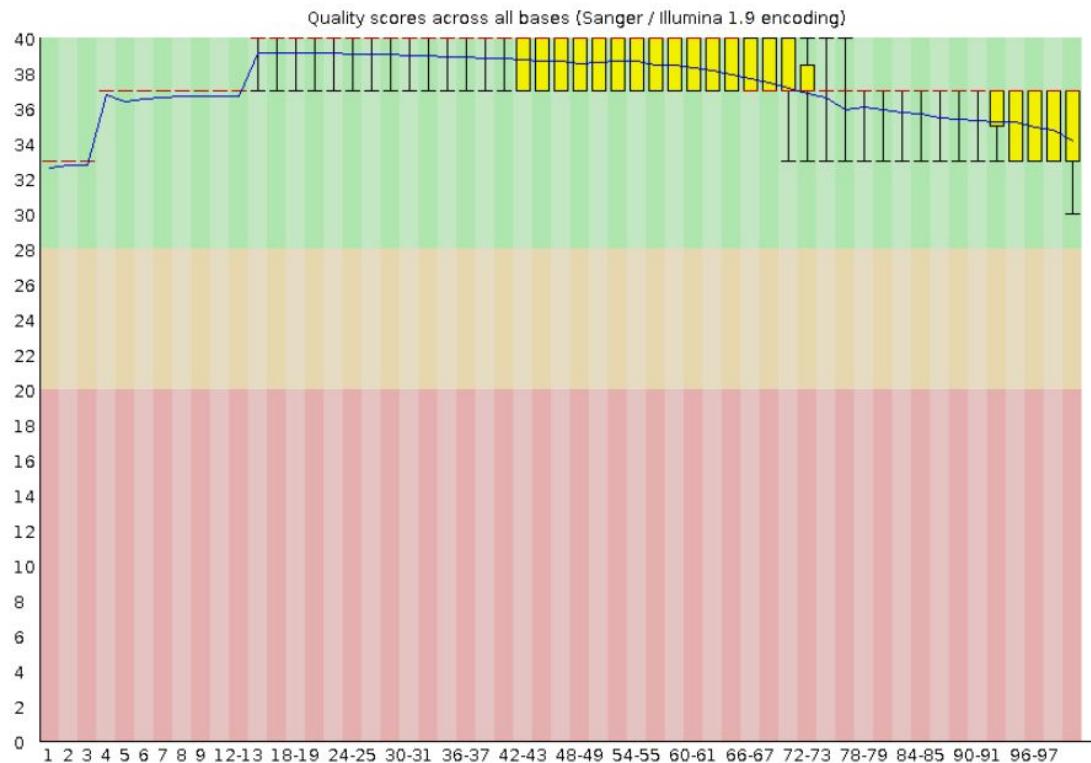
Basic Statistics

Measure	Value
Filename	ERR2834525_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3161562
Sequences flagged as poor quality	0
Sequence length	101
%GC	41

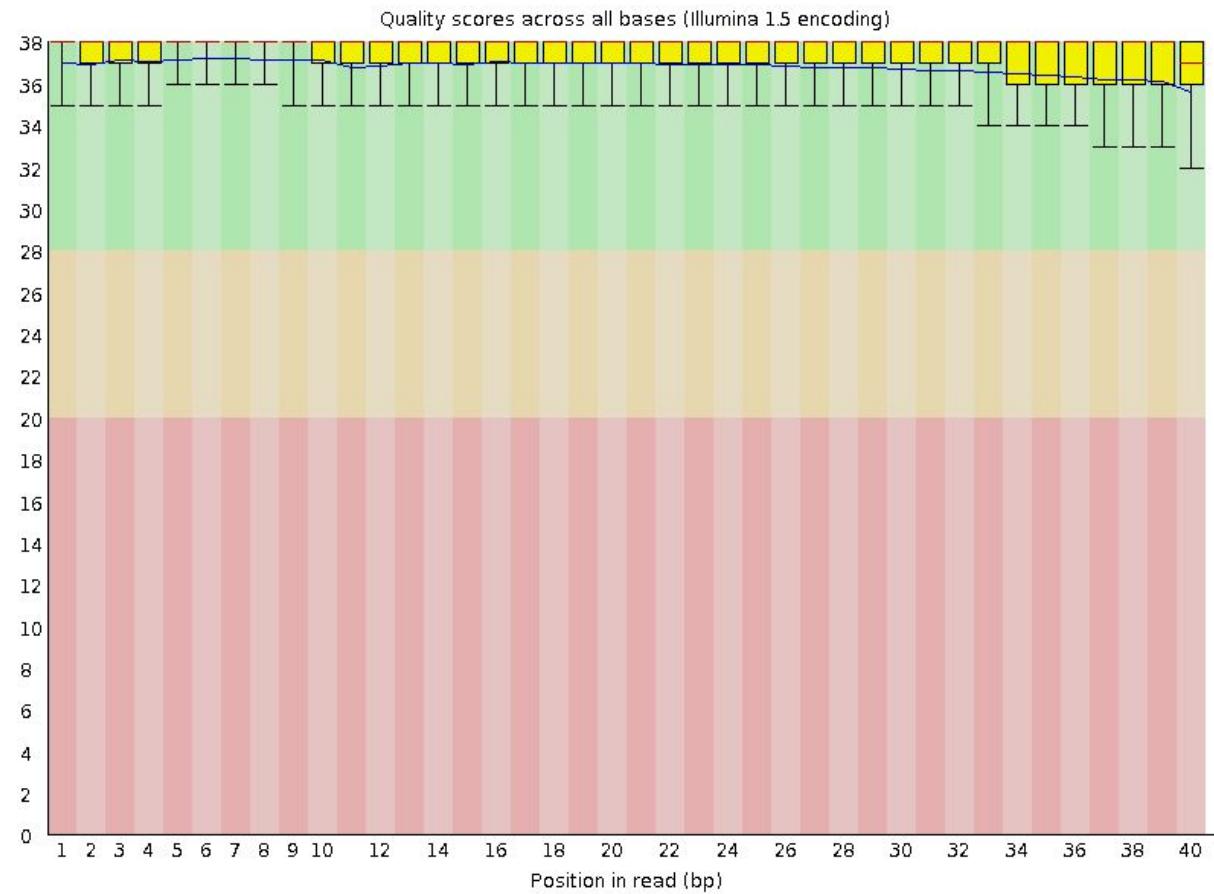
Per Base Sequence Quality



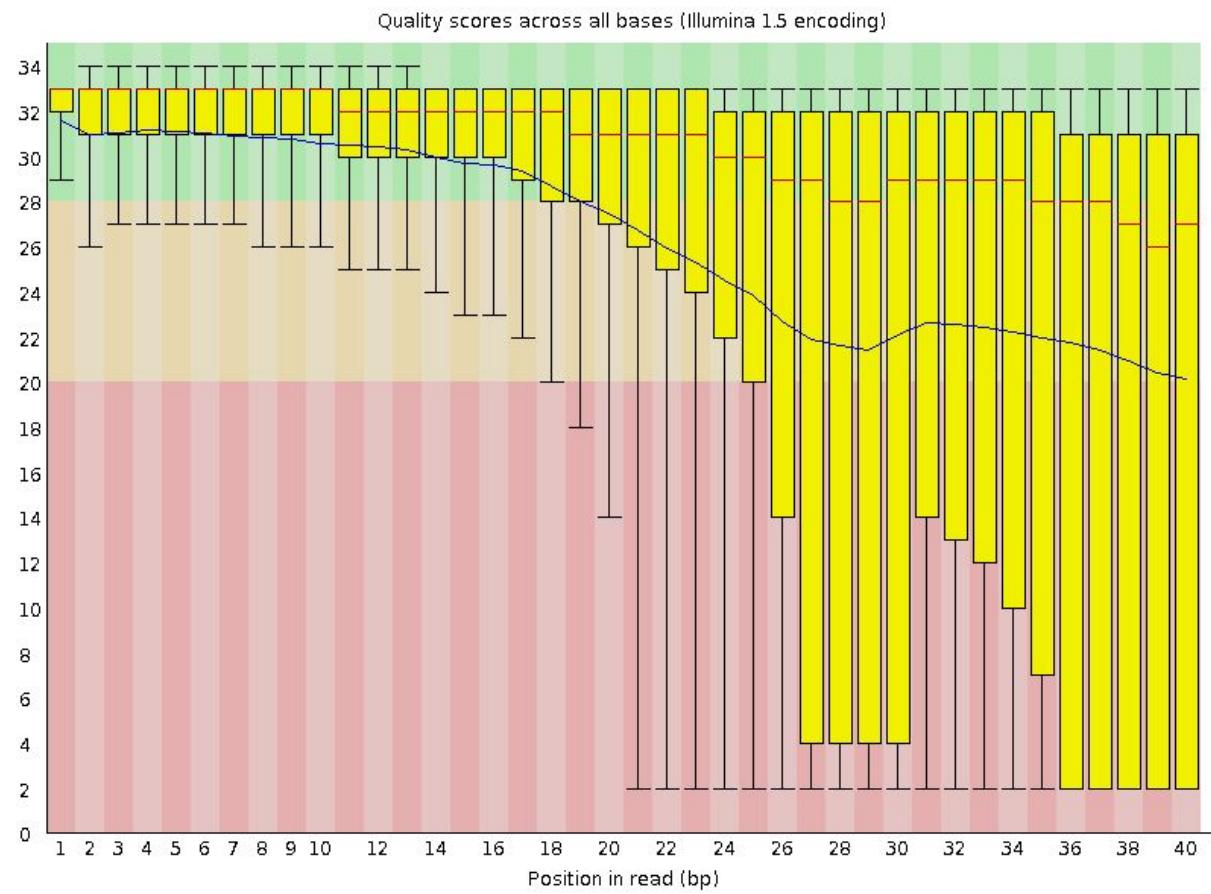
Per Base Sequence Quality - R1 vs. R2



Good

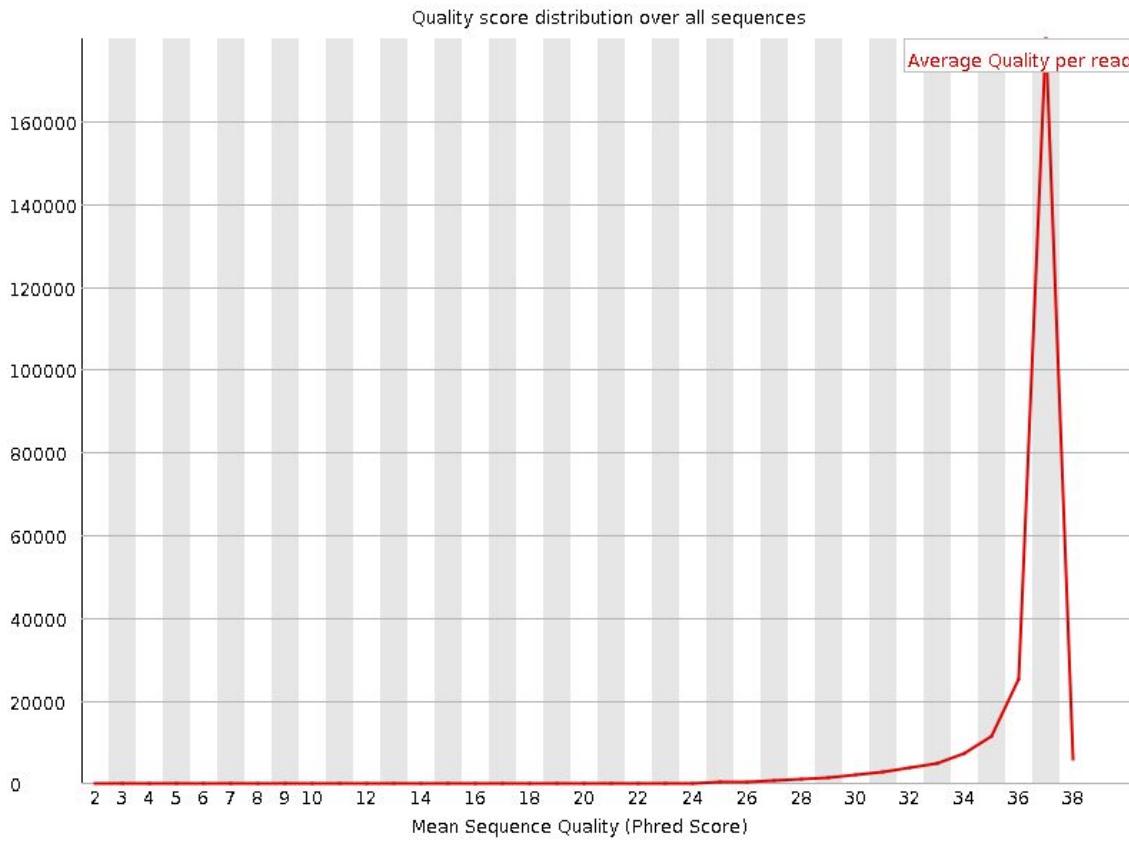


Bad

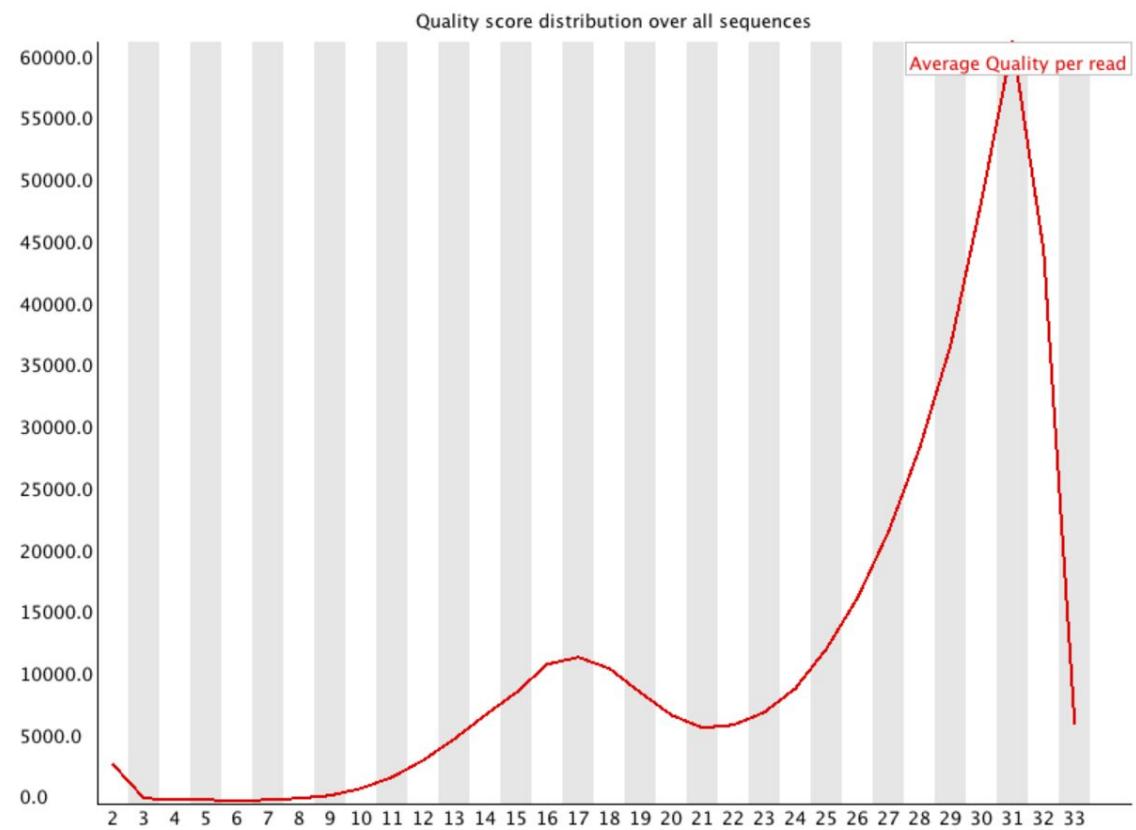


Mean Read Quality

Good

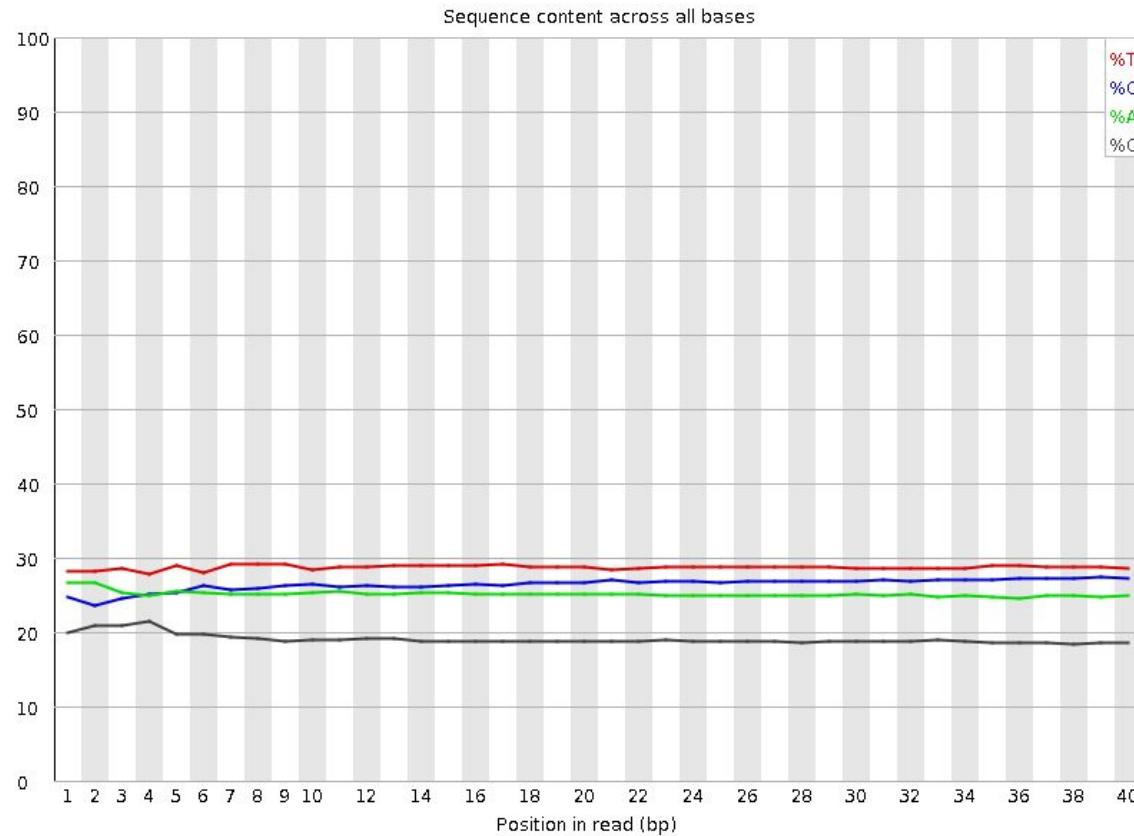


Bad

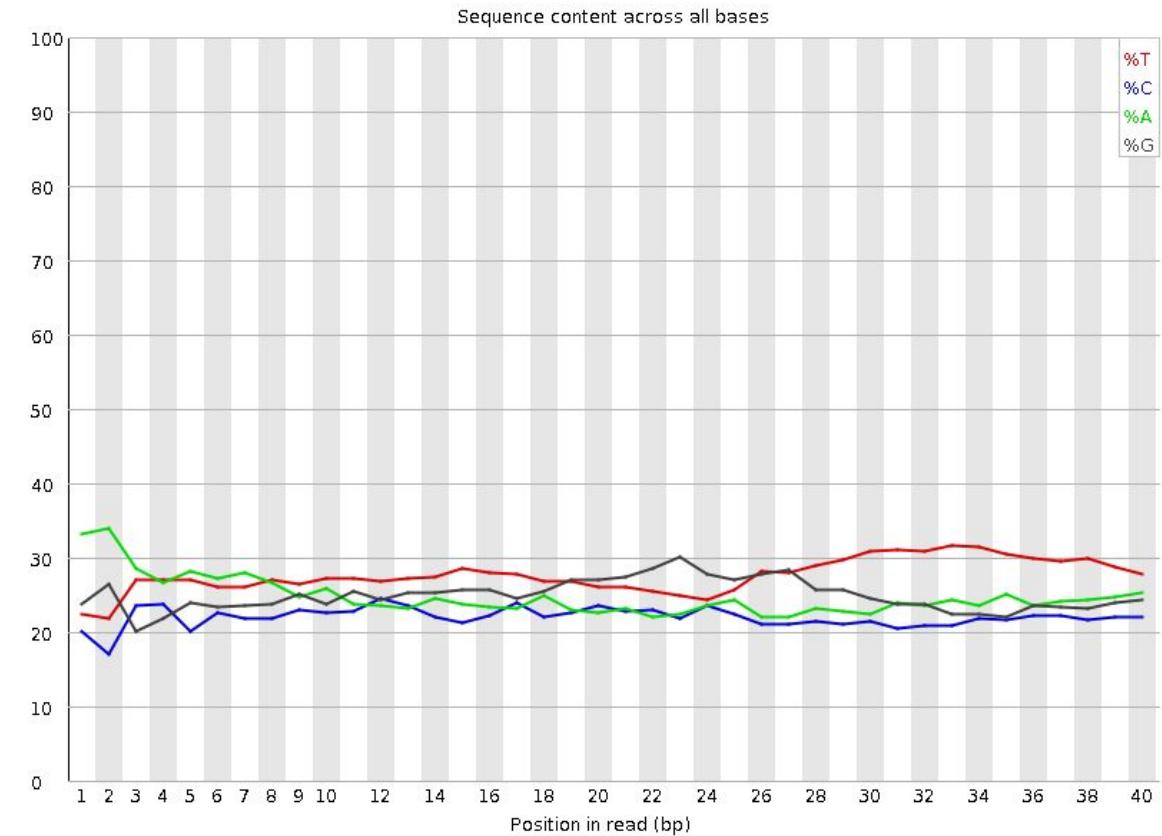


Per Base Sequence Content

Good

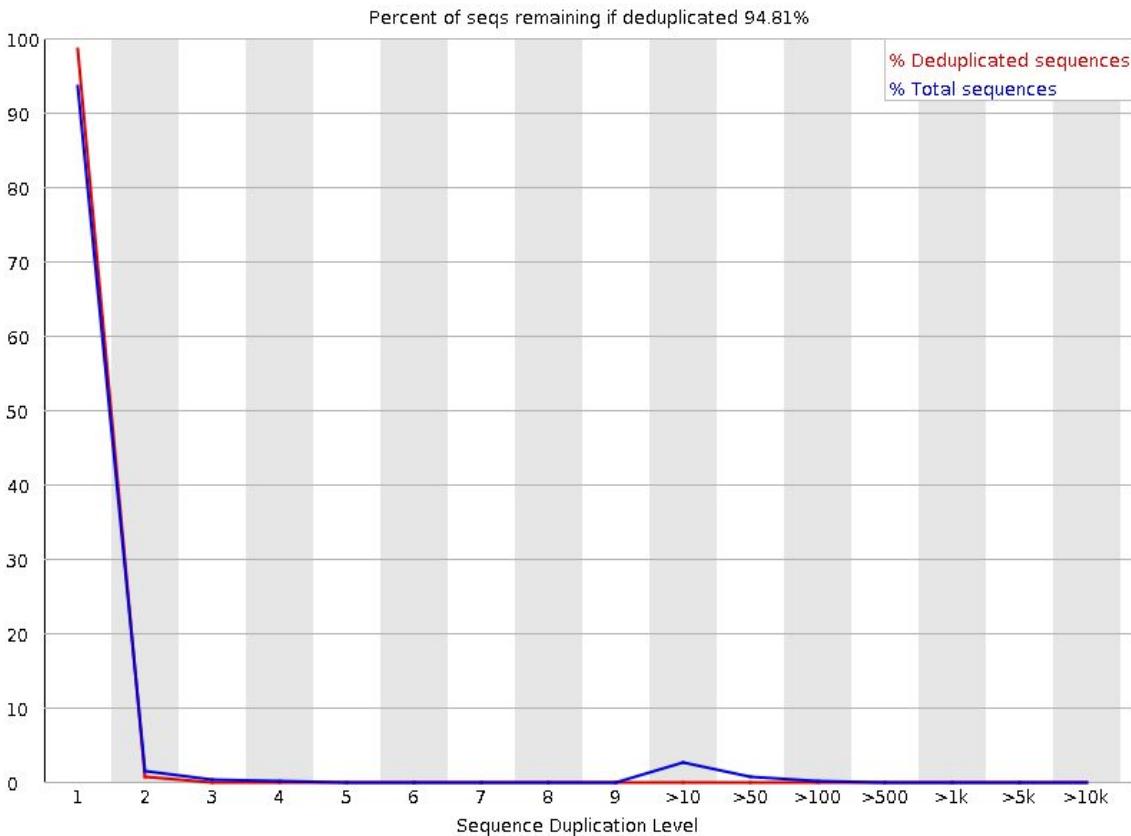


Bad

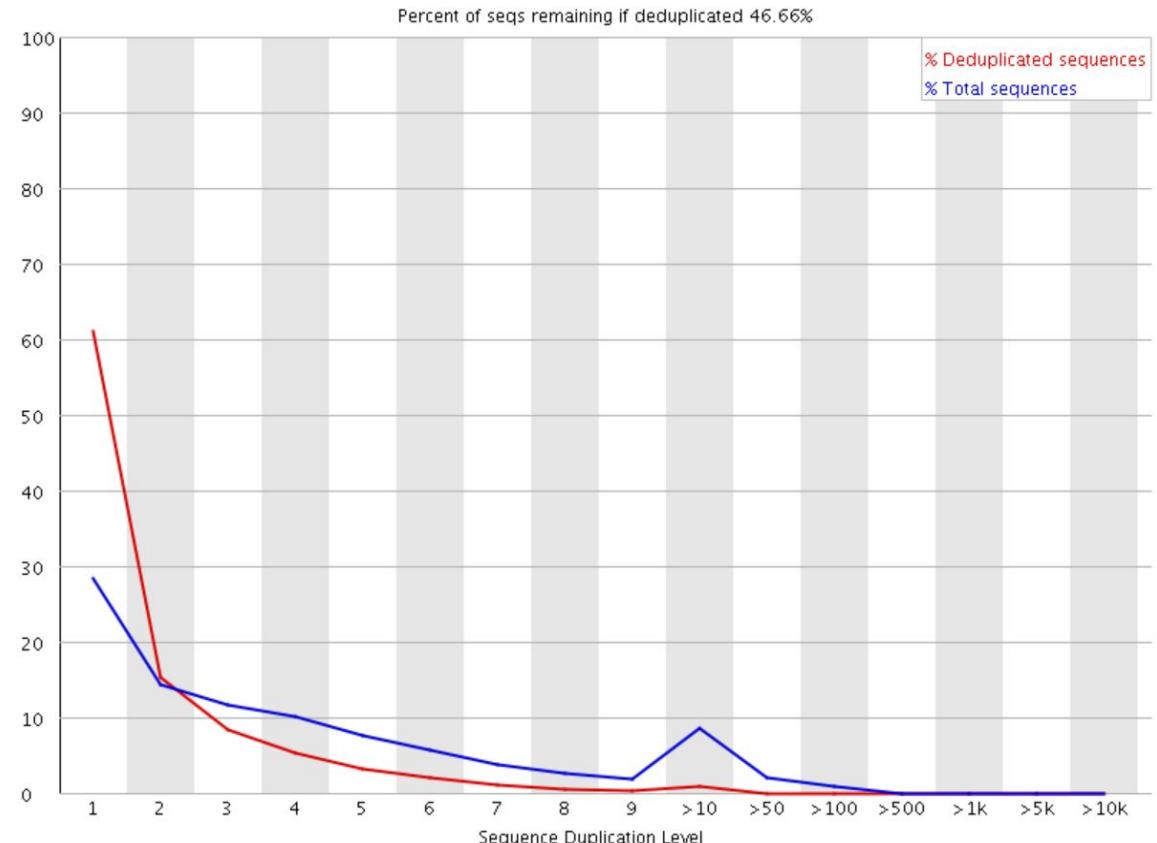


Duplicate Reads

Good

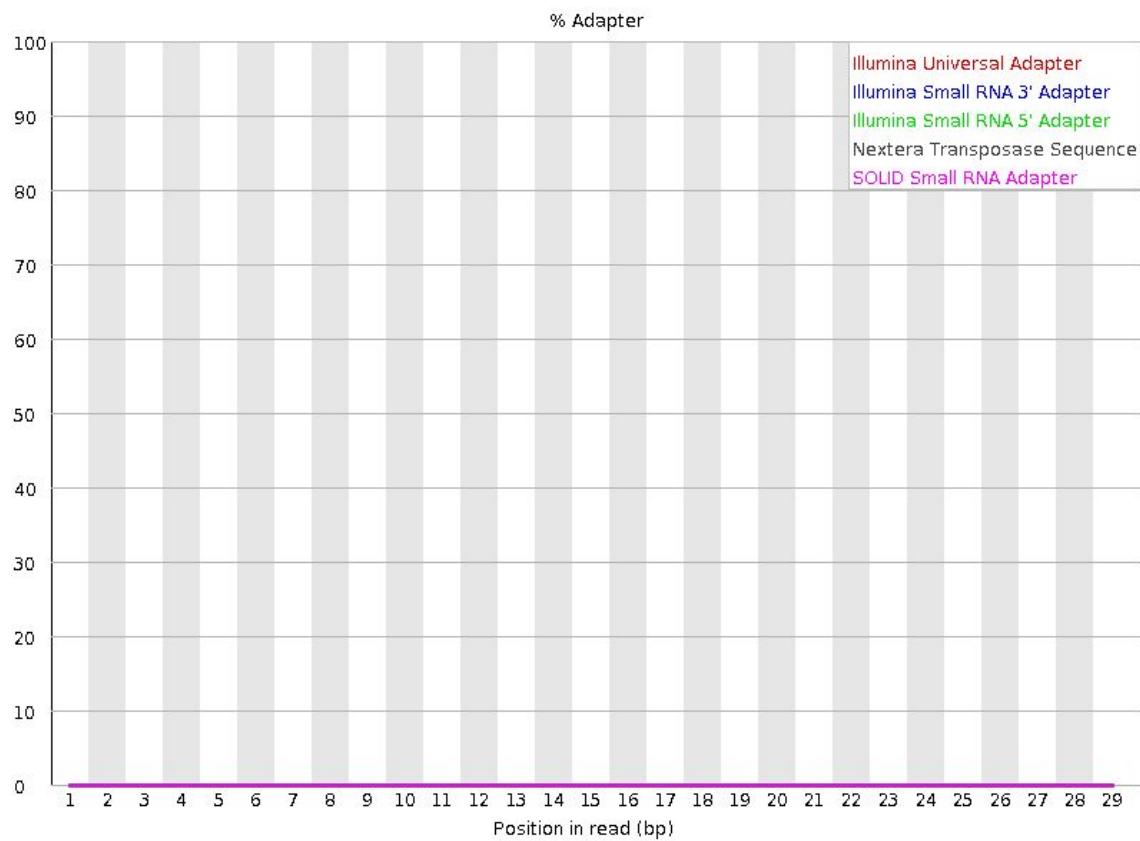


Bad

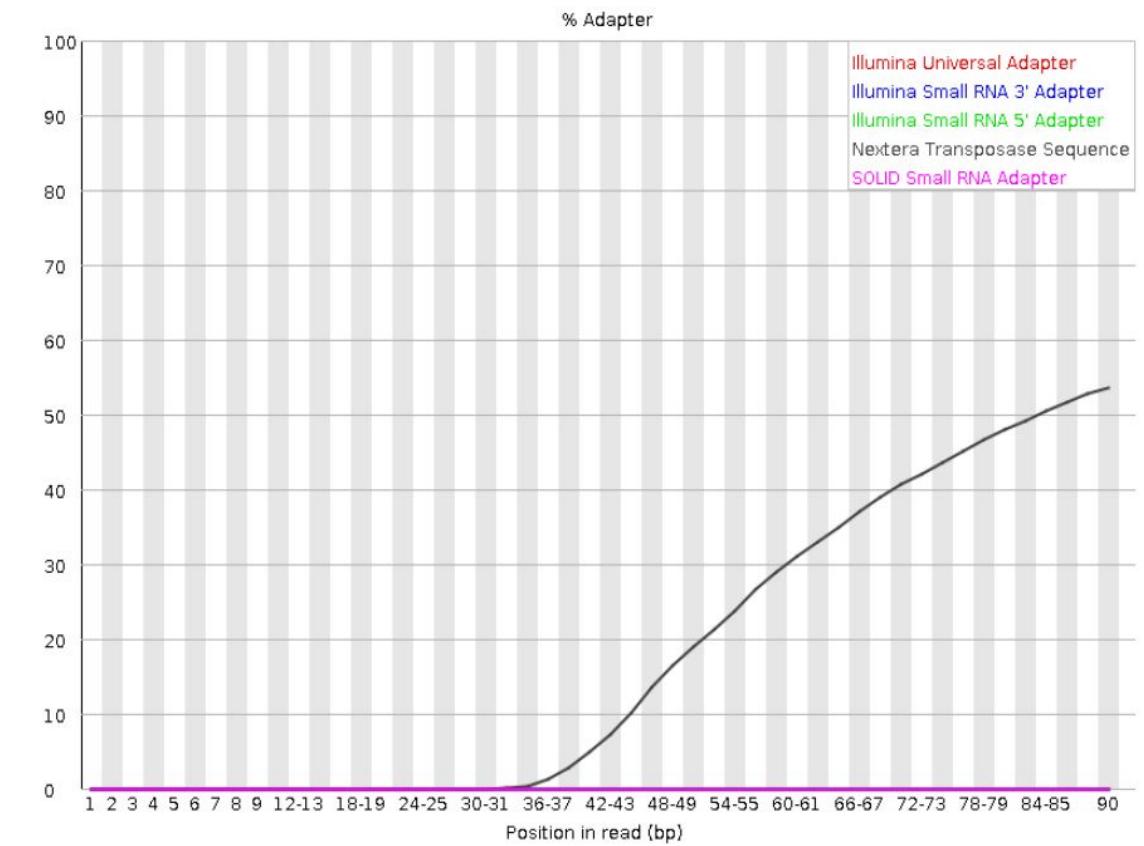


Adapter Sequence

Good



Bad



Warnings and Alerts

Use FastQC report to decide on next steps

Warnings might not affect downstream analysis

Mainly useful when comparing results



Our Toy Data

Cervical Adenocarcinoma Whole Exome Sequencing

Matched Normal Tumor Samples of a
55 year old female patient

<https://www.ncbi.nlm.nih.gov/sra/?term=ERS5326207>

[ERX4703404](#): NextSeq 550 paired end sequencing; Cervical Adenocarcinoma Whole Exome sequencing
1 ILLUMINA (NextSeq 550) run: 28.9M spots, 5.8G bases, 2.3Gb downloads

Design: Cervical Adenocarcinoma Whole Exome sequencing

Submitted by: ADVANCED CENTRE FOR TREATMENT RESEARCH AND EDUCATION IN CANCER
(ADVANCED CENTRE FOR TREATMENT RESEARCH AND EDUCATI)

Study: Cervical Adenocarcinoma Whole Exome sequencing

[PRJEB41309](#) • [ERP125057](#) • [All experiments](#) • [All runs](#)

[hide Abstract](#)

Raw data of exome sequencing of 17 paired samples and 1 orphan tumor sample, total 18 samples

Sample: Sample 8

[SAMEA7569583](#) • [ERS5326207](#) • [All experiments](#) • [All runs](#)

Organism: *Homo sapiens*

Library:

Name: Sample 8_p

Instrument: NextSeq 550

Strategy: WXS

Source: GENOMIC

Selection: RANDOM

Layout: PAIRED

Construction protocol: 17 paired and 1 orphan tumor were collected from ACTREC-TMC. DNA was extracted from tumor tissue and the matched normal tissue or blood using DNeasy tissue extraction kit (Qiagen) and QIAamp DNA blood mini kit (Qiagen) following manufacturer's instruction. Genomic DNA was sheared using covaris to generate 150-500 bp fragment size. The fragment ends were repaired followed by adenylation at 3'end and sample was purified using AMPure XP beads. The fragments were ligated to adaptor and amplified by PCR. The generated library is then hybridized with SureselectTarget Enrichment system kit and hybrids are separated using streptavidin coated magnetic beads. Then the sample was PCR amplified using indexing primers and purified. The prepared libraries were loaded on Illumina flowcell to generate clusters

Experiment attributes:

Experimental Factor: *genotype*: germline genotype

Experimental Factor: *sampling site*: normal tissue adjacent to neoplasm

Runs: 1 run, 28.9M spots, 5.8G bases, [2.3Gb](#)

Run	# of Spots	# of Bases	Size	Published
ERR4833597	28,927,775	5.8G	2.3Gb	2022-02-13

Basic Command Line & Conda

Live Session Azure hadas@74.249.40.65

A word cloud composed of numerous Linux command names, each rendered in a different color. The commands include: alias, service, sort, whoami, uname, pwd, cp, zip, du, less, cat, sudo, vim, ln, rpm, rm, mkdir, ls, tail, curl, git, touch, zcat, gzip, dnf, apt, chown, mv, grep, sed, ssh, clear, awk, head, htop, nano, diff, unzip, wget, cd, man, dnfls, ps, chmod, df, killall.