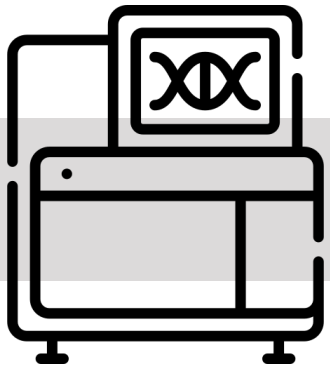
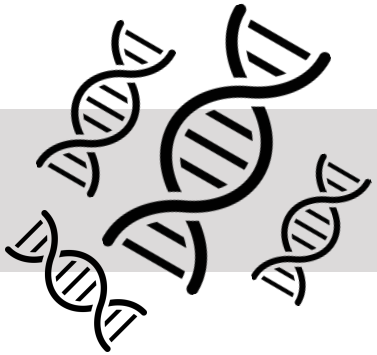
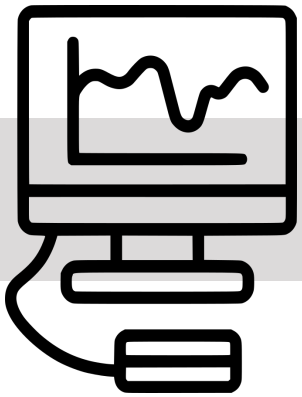
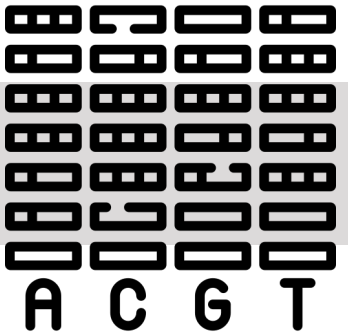
Whole-Genome sequencing technologies









Learning objectives

1. Understand the principles behind the different sequencing technologies
2. Library preparation methods and their effect on the data
3. Perform quality control (QC) on long-read sequencing data
4. Align long-reads to a reference genome
5. Call SNPs in long-read sequencing data
6. Analyze differences between sequencing technologies
7. Compare and intersect VCF files
8. Use common benchmarking metrics

Why is DNA sequencing important

A profound implication of the central dogma is that nearly all the information necessary to construct and operate a living thing is contained in its DNA.[2] We call the complete complement of DNA (and therefore the collection of all the genes) in a particular species its *genome*. That is why genome sequencing projects, which determine the exact sequence of all the DNA in an organism, are so important.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.

History of genome sequencing

1977 DNA Sequencing by Chemical Degradation" by Allan Maxam and Walter Gilbert

1978 DNA Sequencing by Enzymatic Synthesis" is published by Fred Sanger

1980 Fred Sanger and Walter Gilbert receive the Nobel Prize in Chemistry

1982 GenBank starts as a public repository of DNA sequences

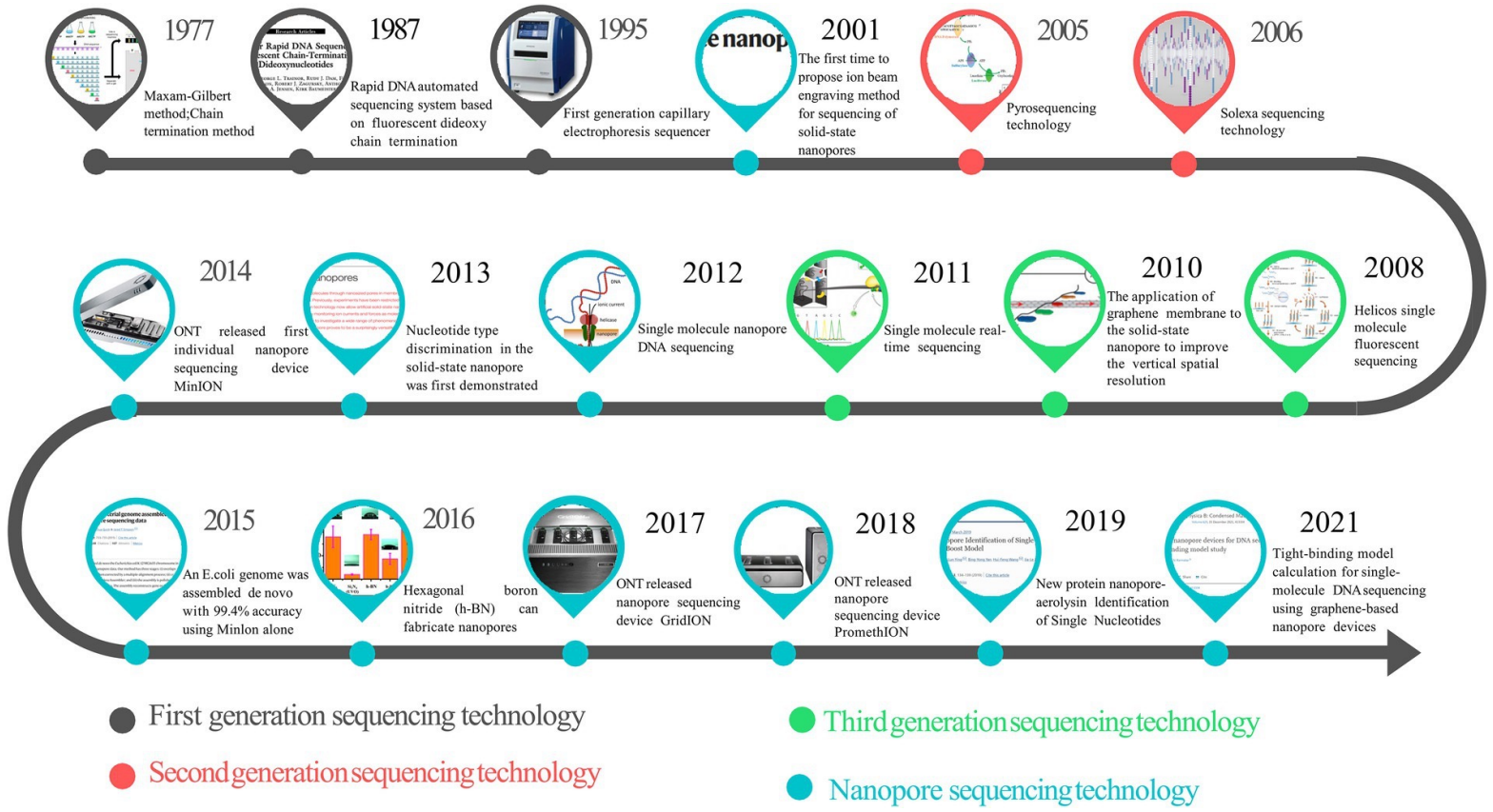
1986 Leroy Hood's laboratory at the California Institute of Technology announces the first semi-automated DNA sequencing machine.

1977 Genome sequence of *E. coli* is published

2001 Draft sequence of the Human genome is published

2004 First next generation sequencing technologies become available

2009 Genome-wide single-cell sequencing becomes available



Genomics technology

First-Generation Sequencing

1. Low throughput
2. Radio- or fluorescently-labelled dNTPs or oligonucleotides before electrophoretic analysis
3. Eg: Sanger sequencing

Second-Generation Sequencing

1. High throughput
2. Parallelization of a large number of reactions
3. Eg: Illumina.

Third-Generation Sequencing

1. Sequencing single molecules, so they don't require DNA amplification.
2. PacBio (Pacific Biosciences) and ONT (Oxford Nanopore).

Fourth-Generation Sequencing?

1. Spatial Genomics (eg: 10x Genomics, NanoString, Bio-Rad…)

Genomics technology

1st era

Sanger sequencing, Genome Projects

2nd era

Short reads

3rd era

Early Long Reads

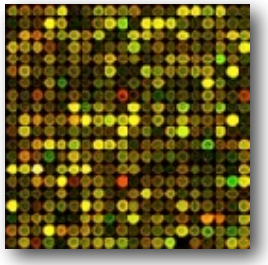
4th era

Telomere-to-telomere

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

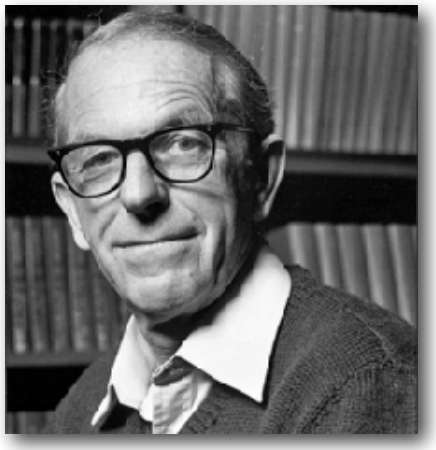
Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010



Fred Sanger
1918-2013

"Chain termination" sequencing

Sanger sequencing

Sanger sequencing
1977-1990s

Fred Sanger in episode 3 of PBS documentary "DNA"

Not-so-high-throughput Sanger sequencing

First practical method invented by Fred Sanger in 1977. Initially used to sequence shorter genomes, e.g. viral genomes 10,000s of bases long.

Sanger sequencing

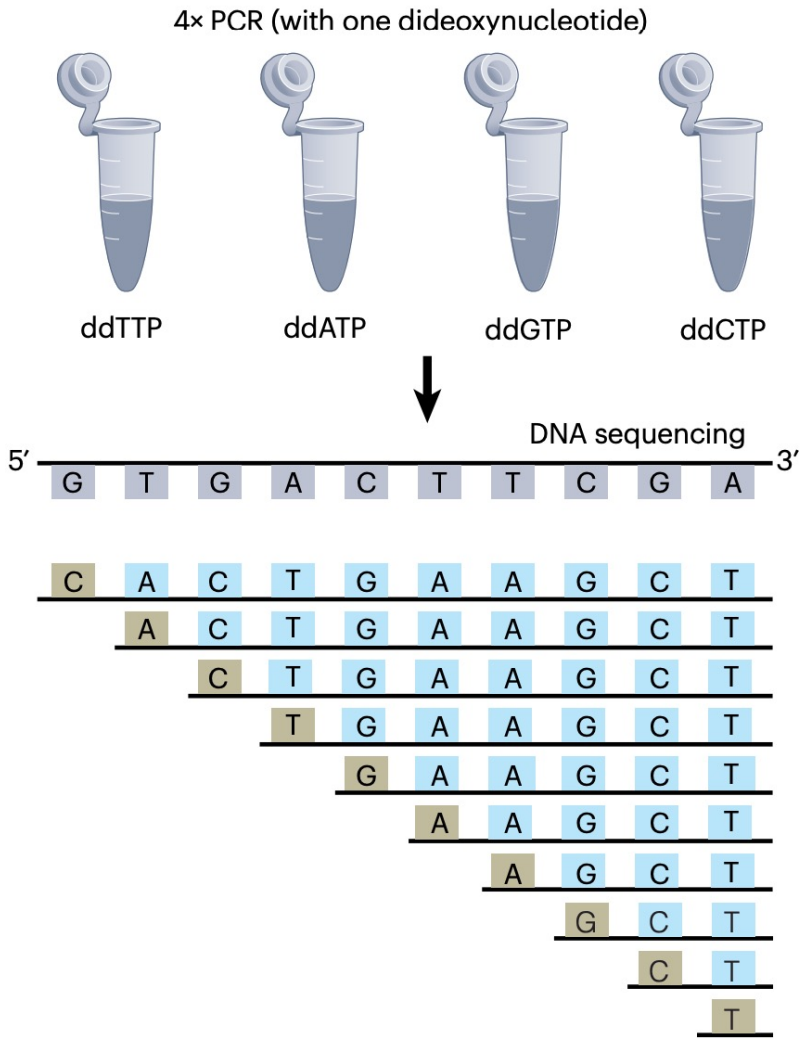
DNA template ATGCAGCGTTACCATG...

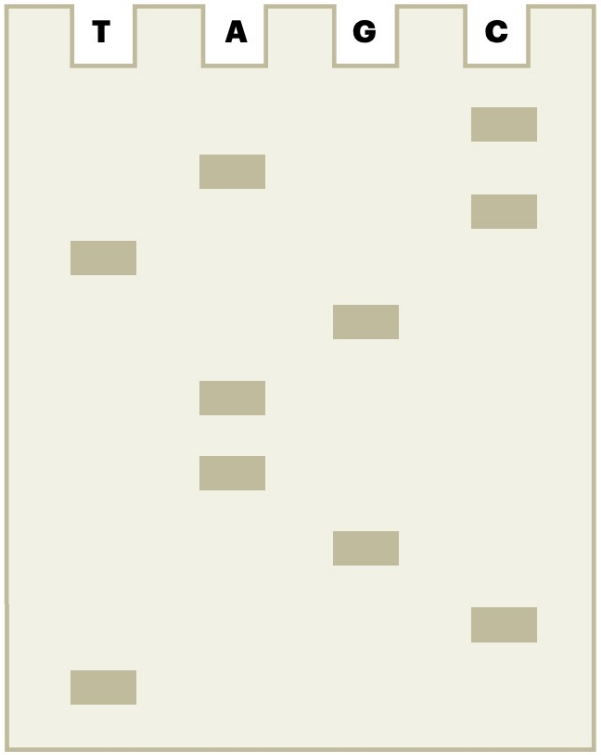
Primer ATGC

Flourescent Dideoxy nucleotides (ddNTPs) – No hydroxyl group A C G T

Normal nucleotides A C G T

Sanger sequencing

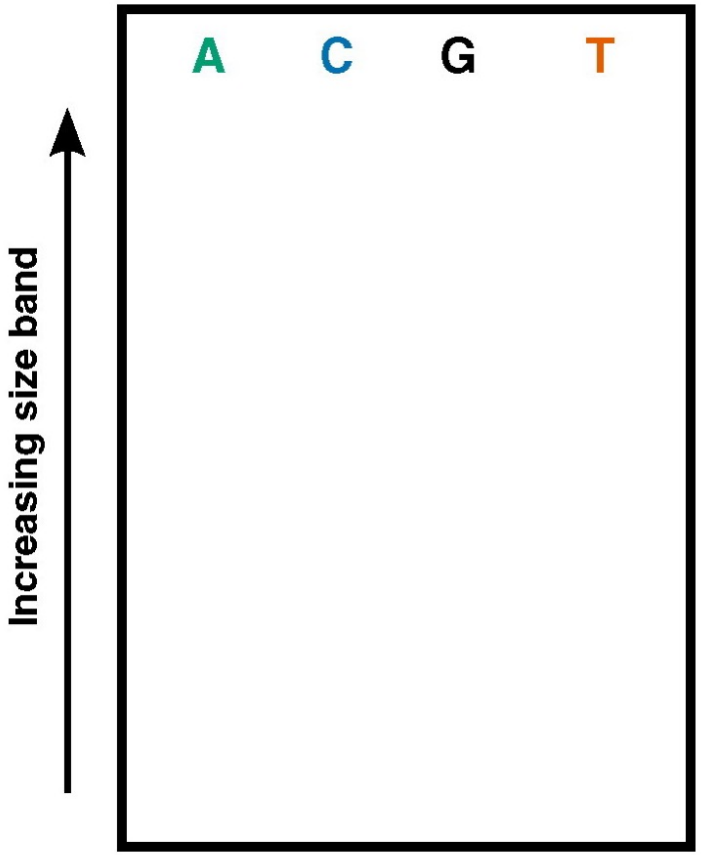




https://www.nature.com/articles/s41587-023-01986-3

Sanger sequencing

ATGCAGCGTTACCATG...



A

A
ATGCA
ATGCAGCGTTA
ATGCAGCGTTACCA

C

ATGC
ATGCAGC
ATGCAGCGTTAC
ATGCAGCGTTACC

G

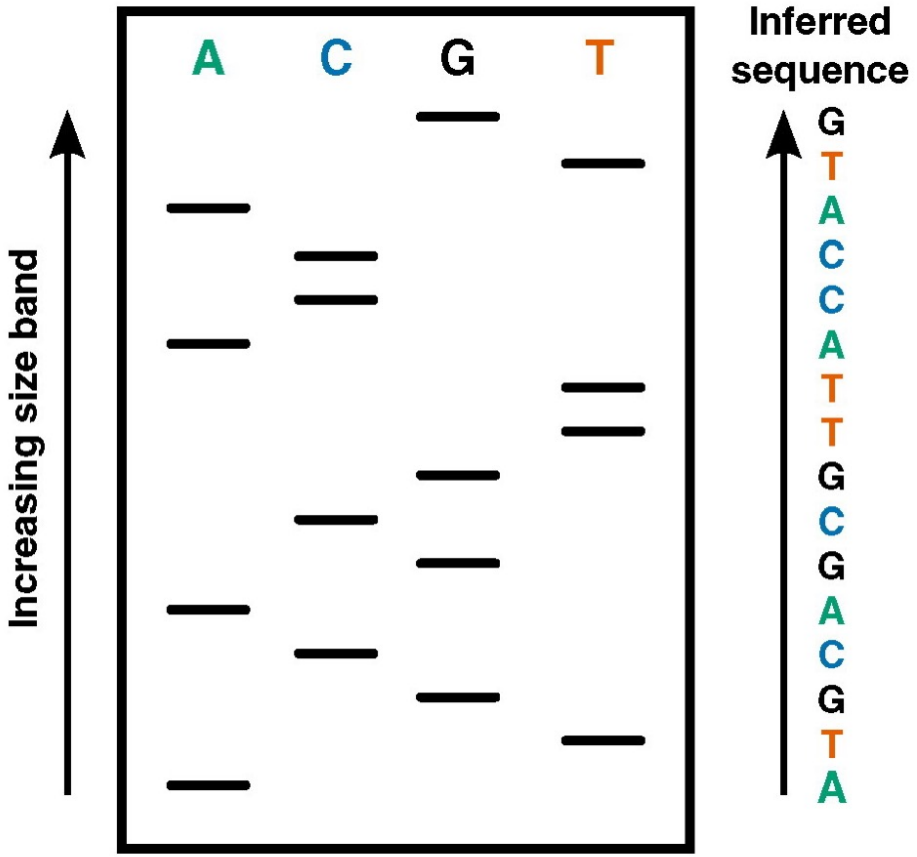
ATG
ATGCAG
ATGCAGCG
ATGCAGCGTTACCATG

T

AT
ATGCAGCGT
ATGCAGCGTT
ATGCAGCGTTACCAT

Sanger sequencing

ATGCAGCGTTACCATG...



A

A
ATGCA
ATGCAGCGTTA
ATGCAGCGTTACCA

C

ATGC
ATGCAGC
ATGCAGCGTTAC
ATGCAGCGTTACC

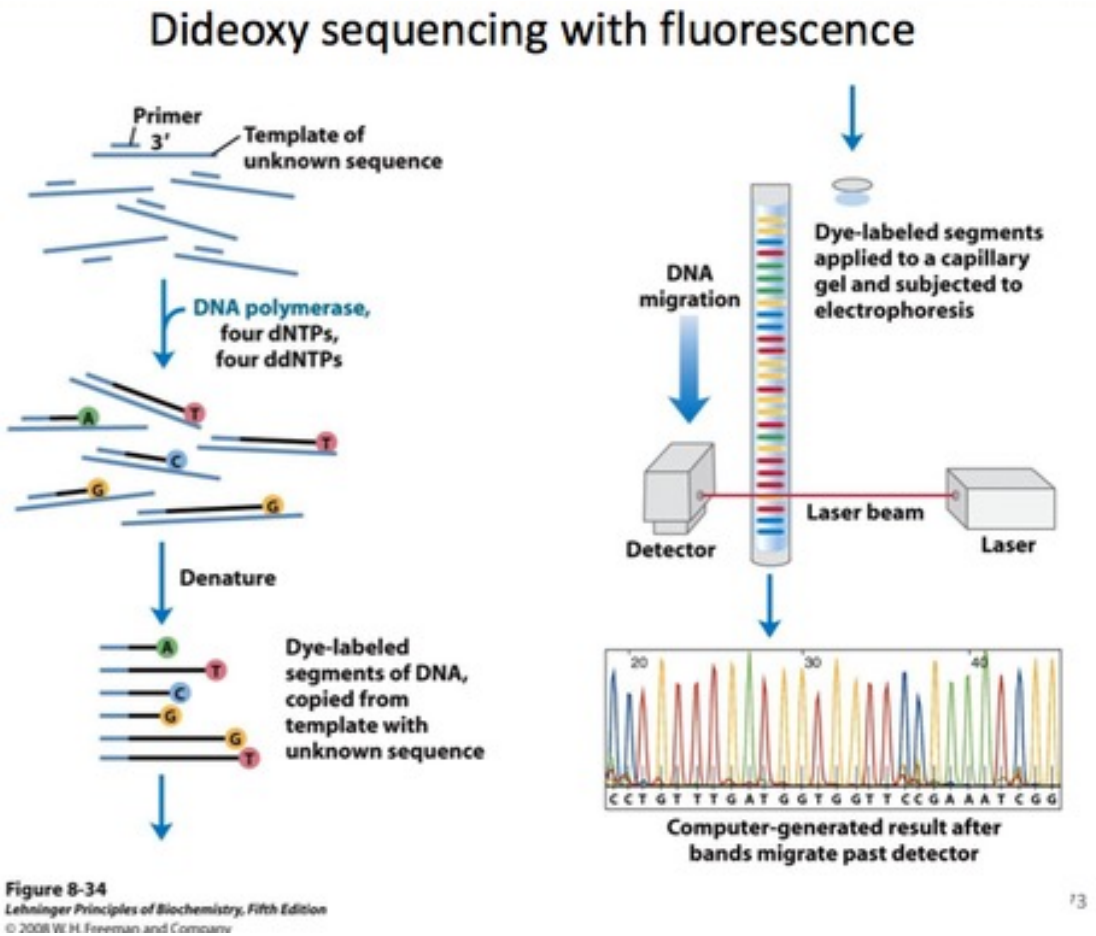
G

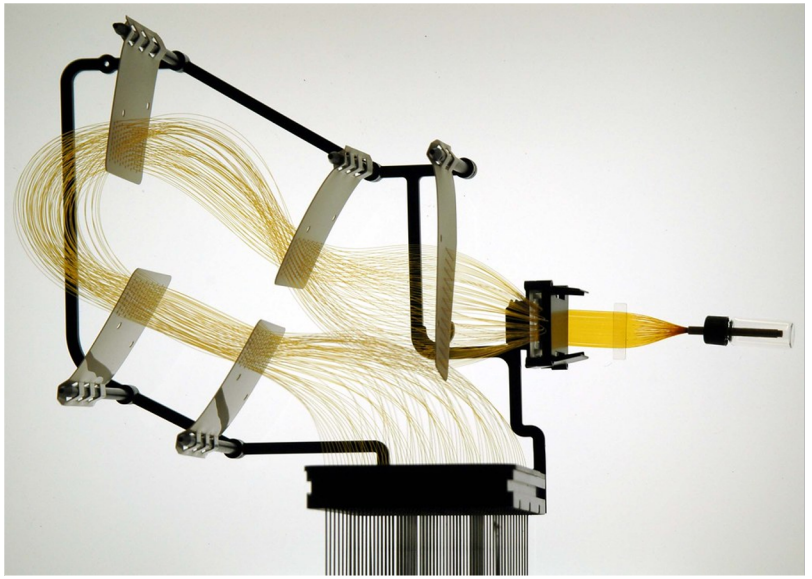
ATG
ATGCAG
ATGCAGCG
ATGCAGCGTTACCATG

T

AT
ATGCAGCGT
ATGCAGCGTT
ATGCAGCGTTACCAT

Sanger sequencing





Sanger sequencing throughput

1977: Sanger et al invents method → 700 bases per day = 12,000 years to sequence the human genome

1985: ABI 370 (first automated sequencer) → 5000 bases per day = 1,600 years

1995: ABI 377 (Bigger gels, better chemistry & optics, more sensitive dyes, faster computers) → 19,000 bases per day = 430 years

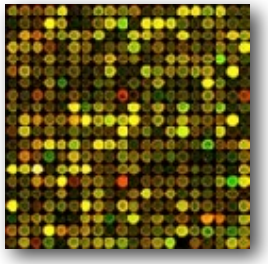
1999: ABI 3700 (96 capillaries, 96 well plates, fluid handling robots) → 400,000 bases per day = 20.5 years

https://journals.sagepub.com/cms/10.1177/221106829800300503/asset/images/large/10.1177_221106829800300503-fig2.jpeg

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010

Solexa

illumina®

Second Generation Sequencing

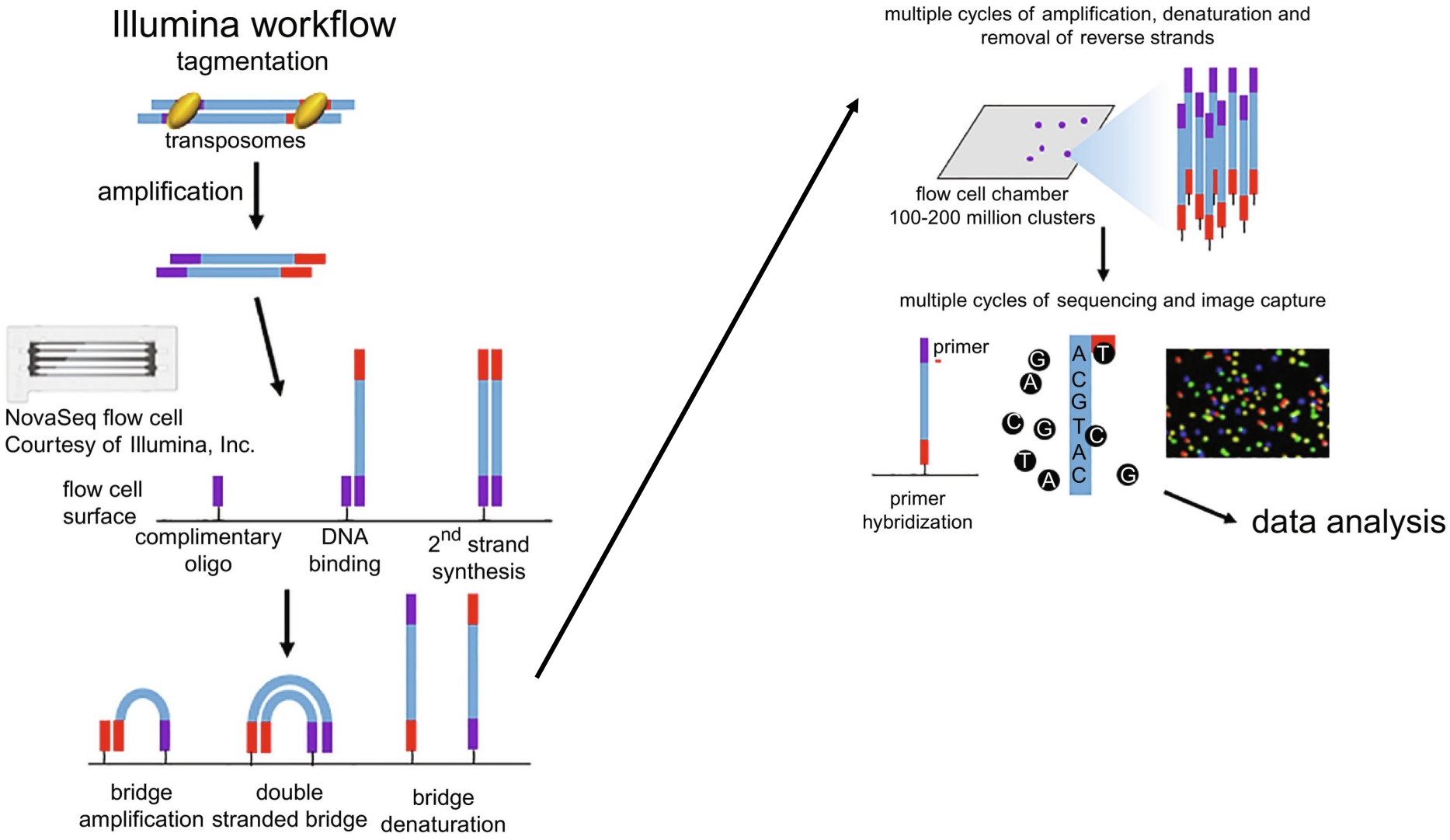
Common components

- Flow cells as reaction chambers
- Iterative sequencing process
- Massive parallelization
- Clonally amplified or single molecule templates

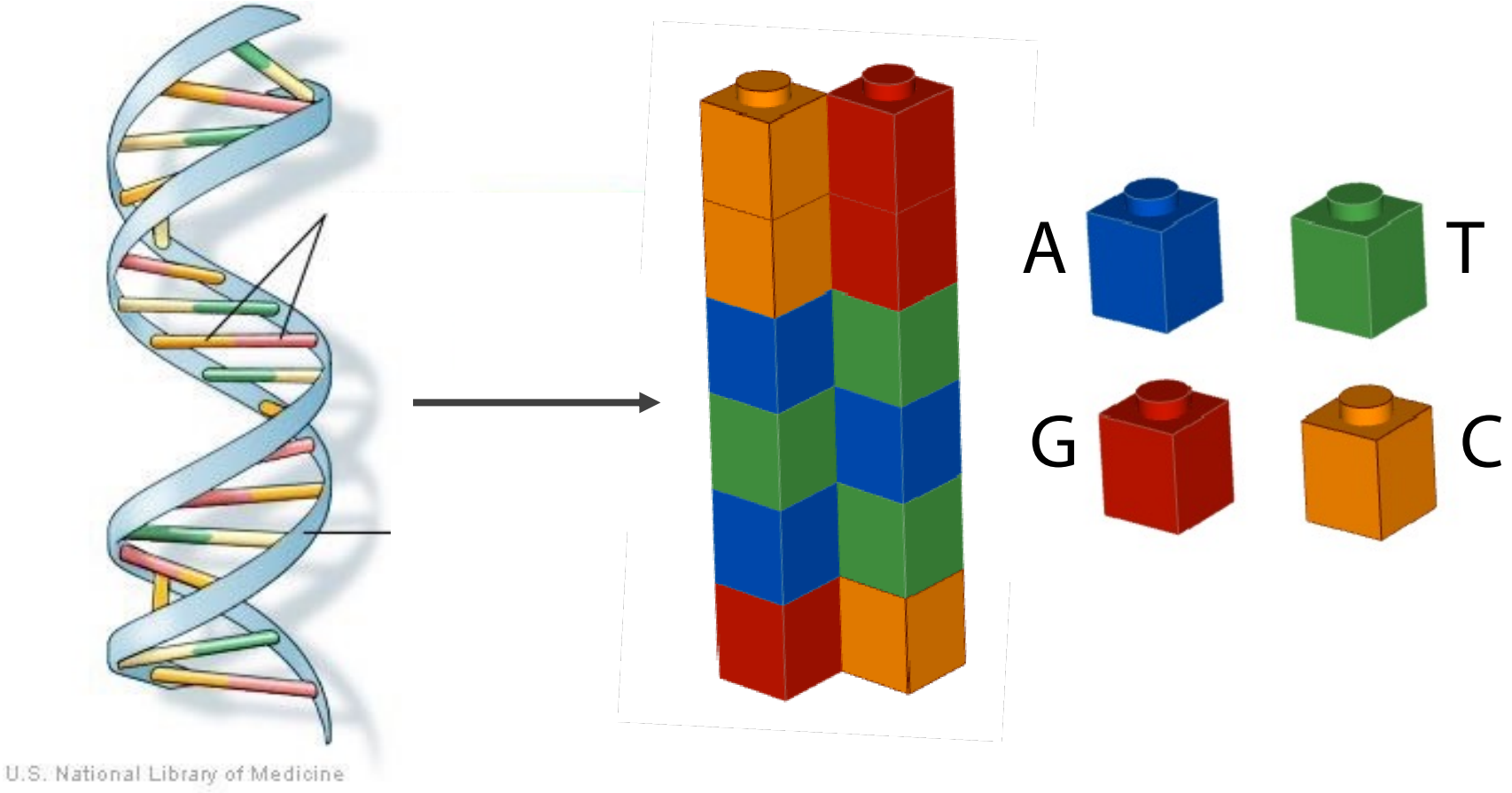
Differences

- Template preparation
- Sequencing chemistry
- Flow cell configuration

Illumina – Sequencing by synthesis



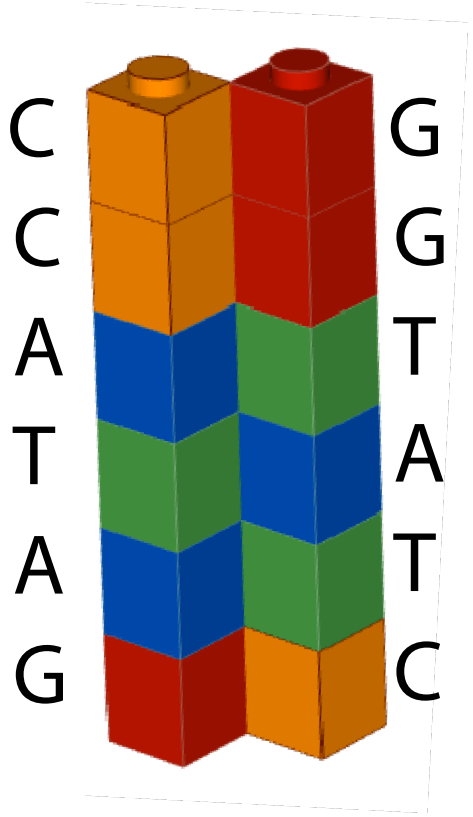
Illumina – Sequencing by synthesis

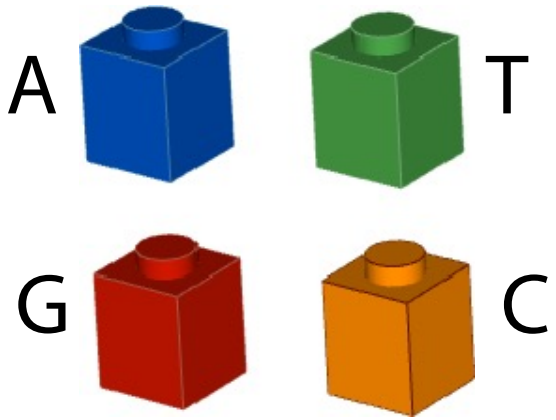


Double stranded DNA (double helix)

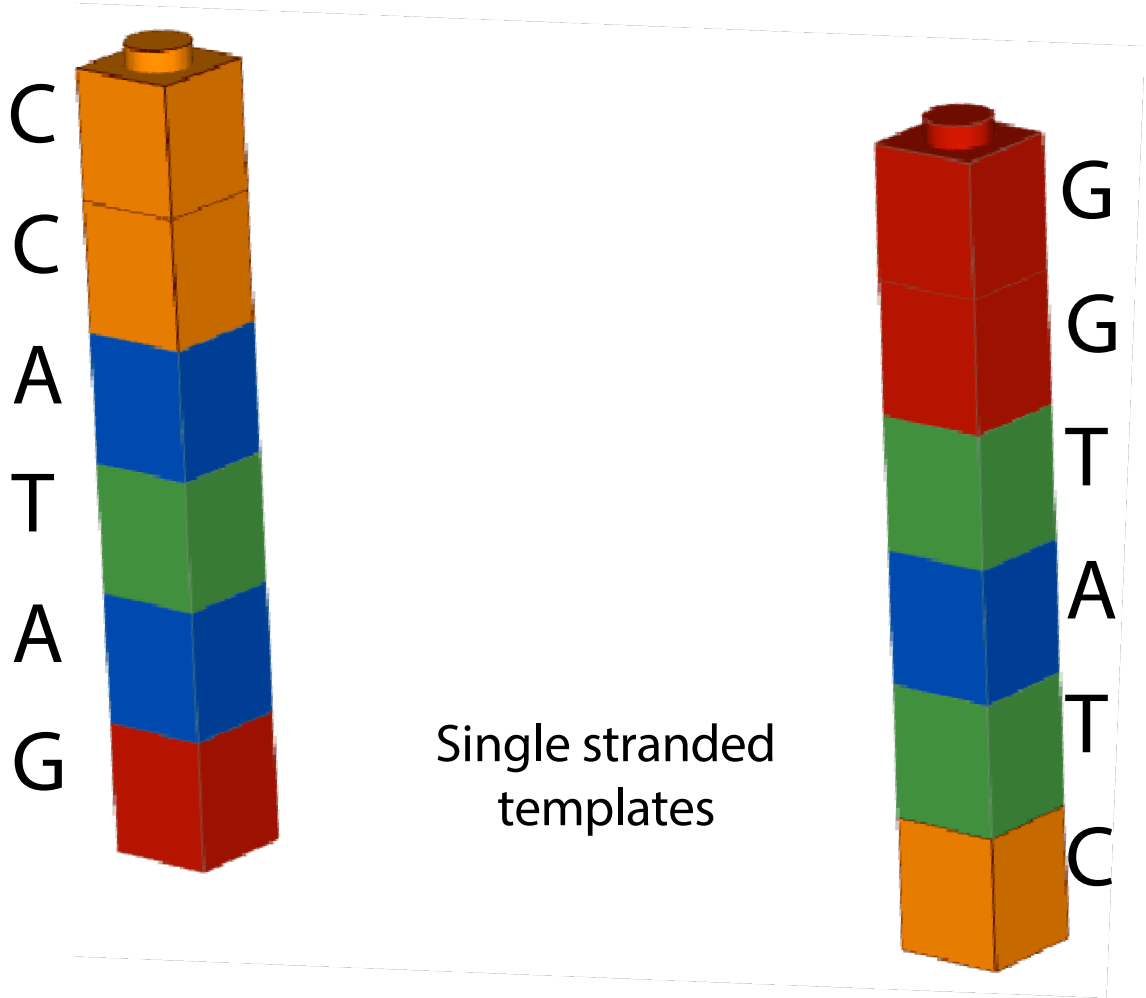
Double stranded DNA (lego version)

Illumina – Sequencing by synthesis

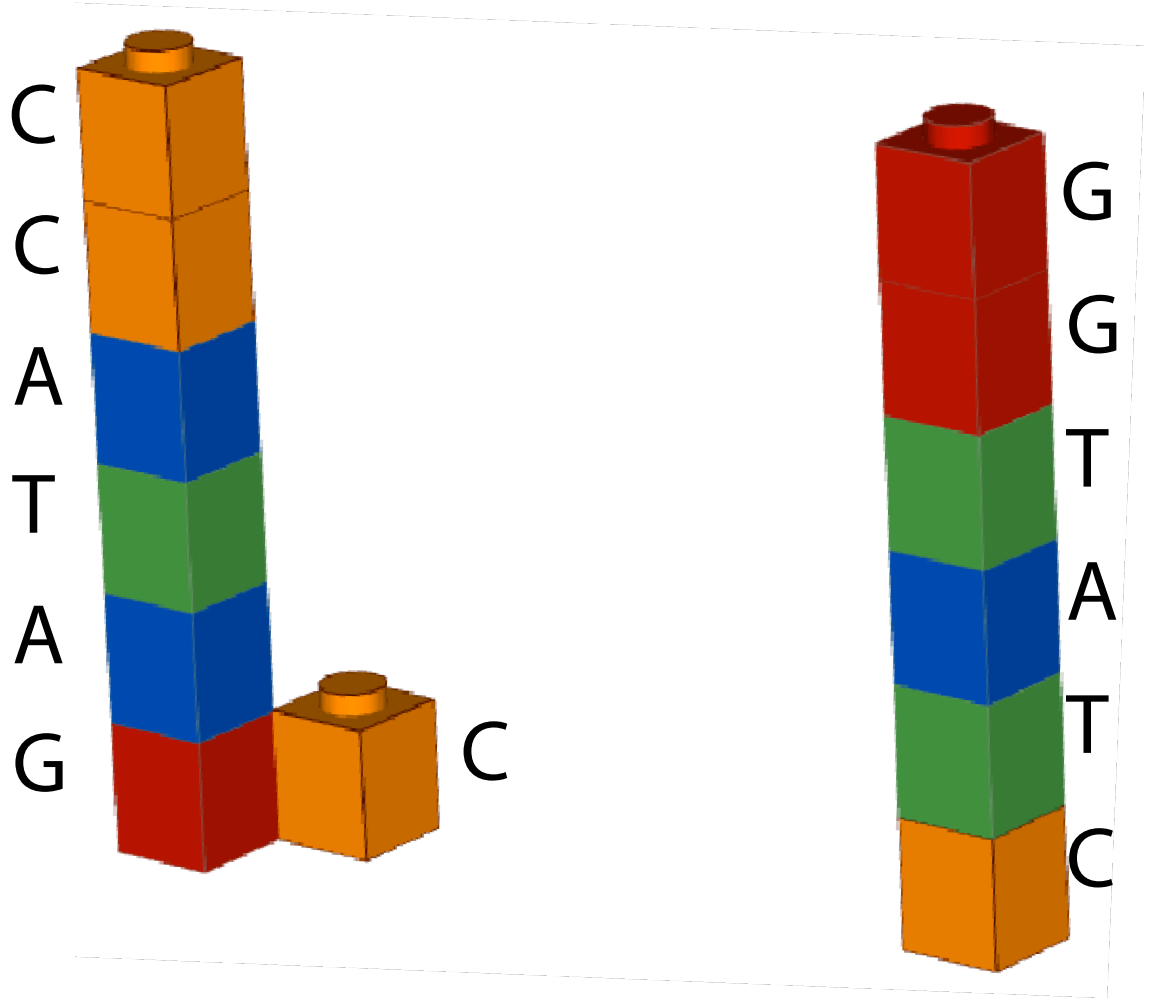




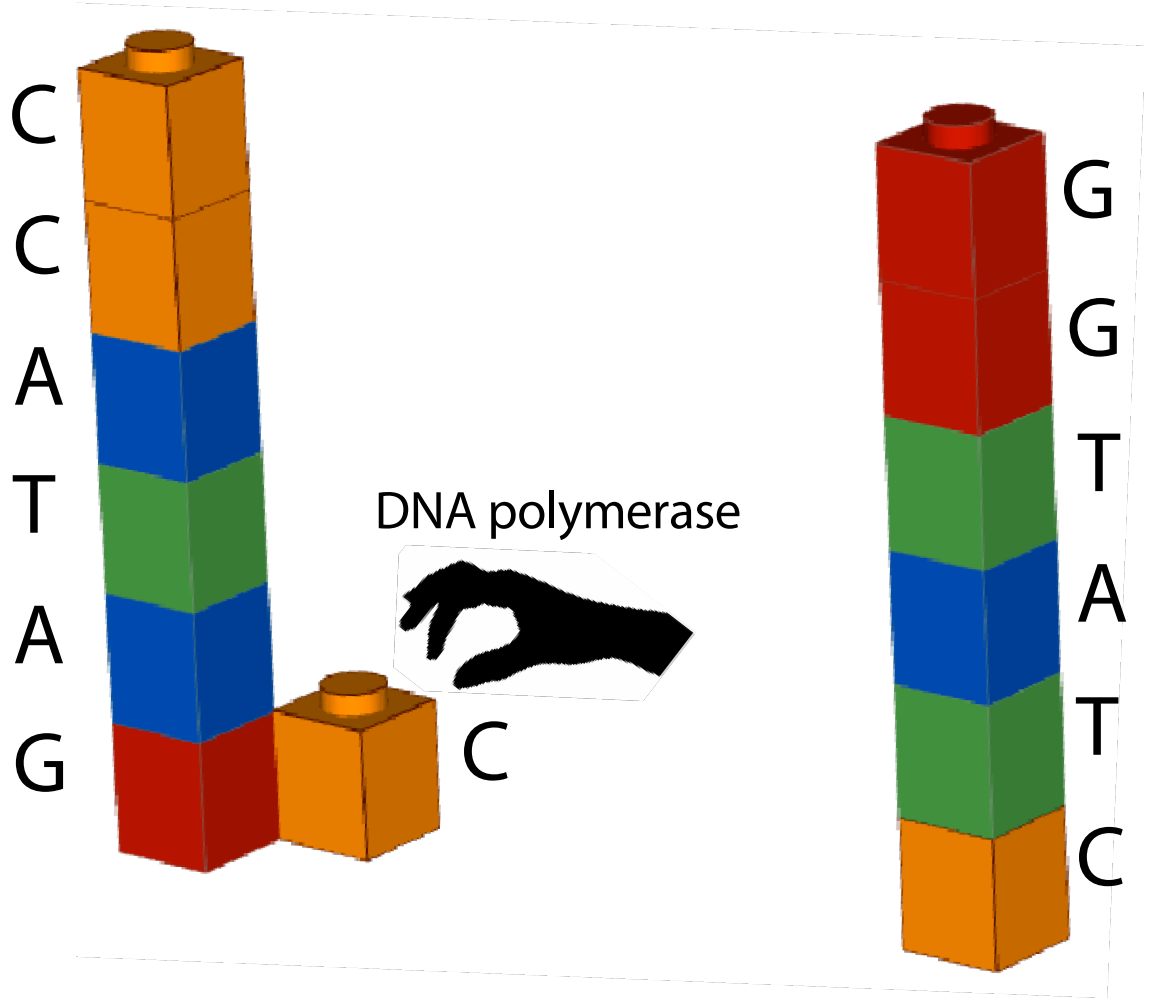
Illumina – Sequencing by synthesis



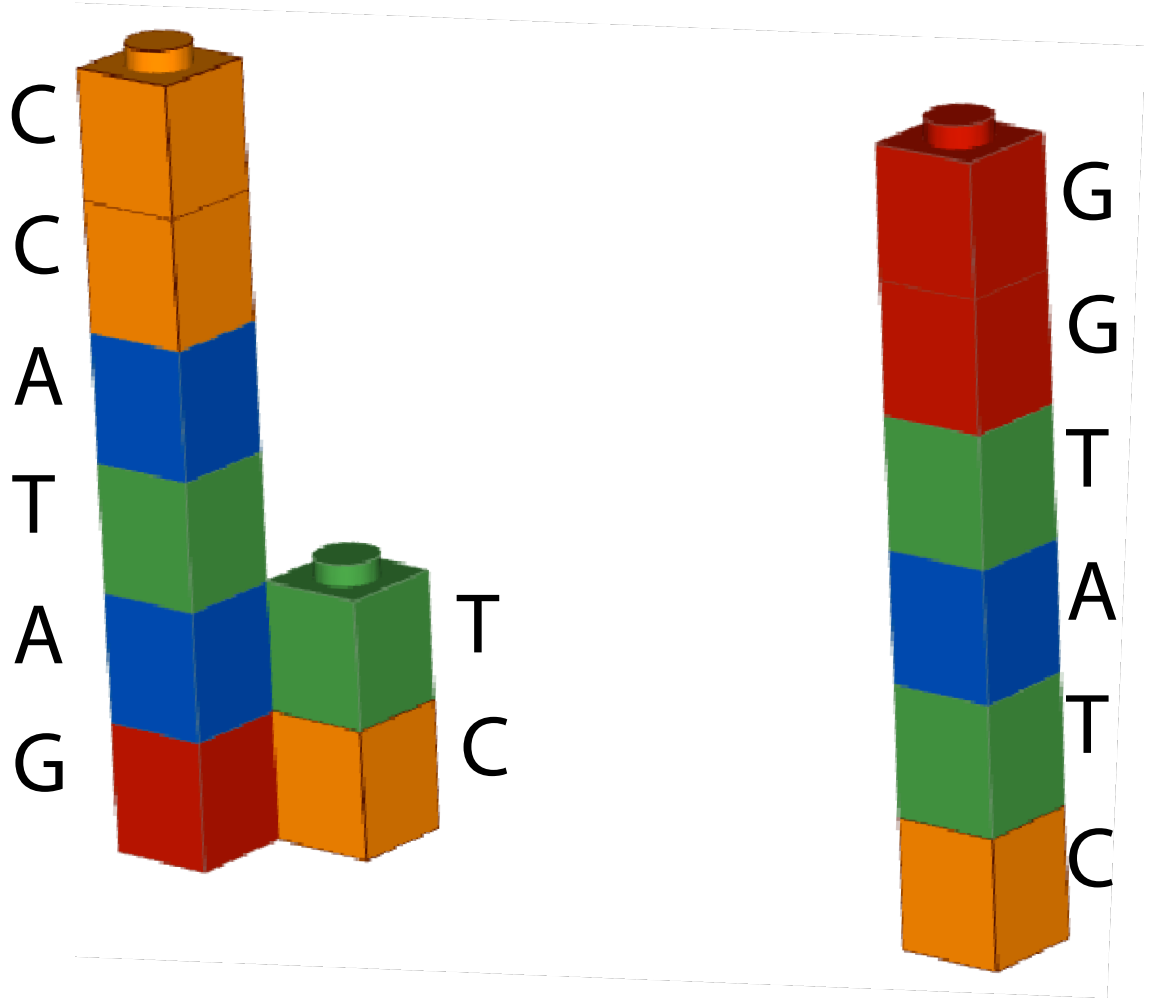
Illumina – Sequencing by synthesis



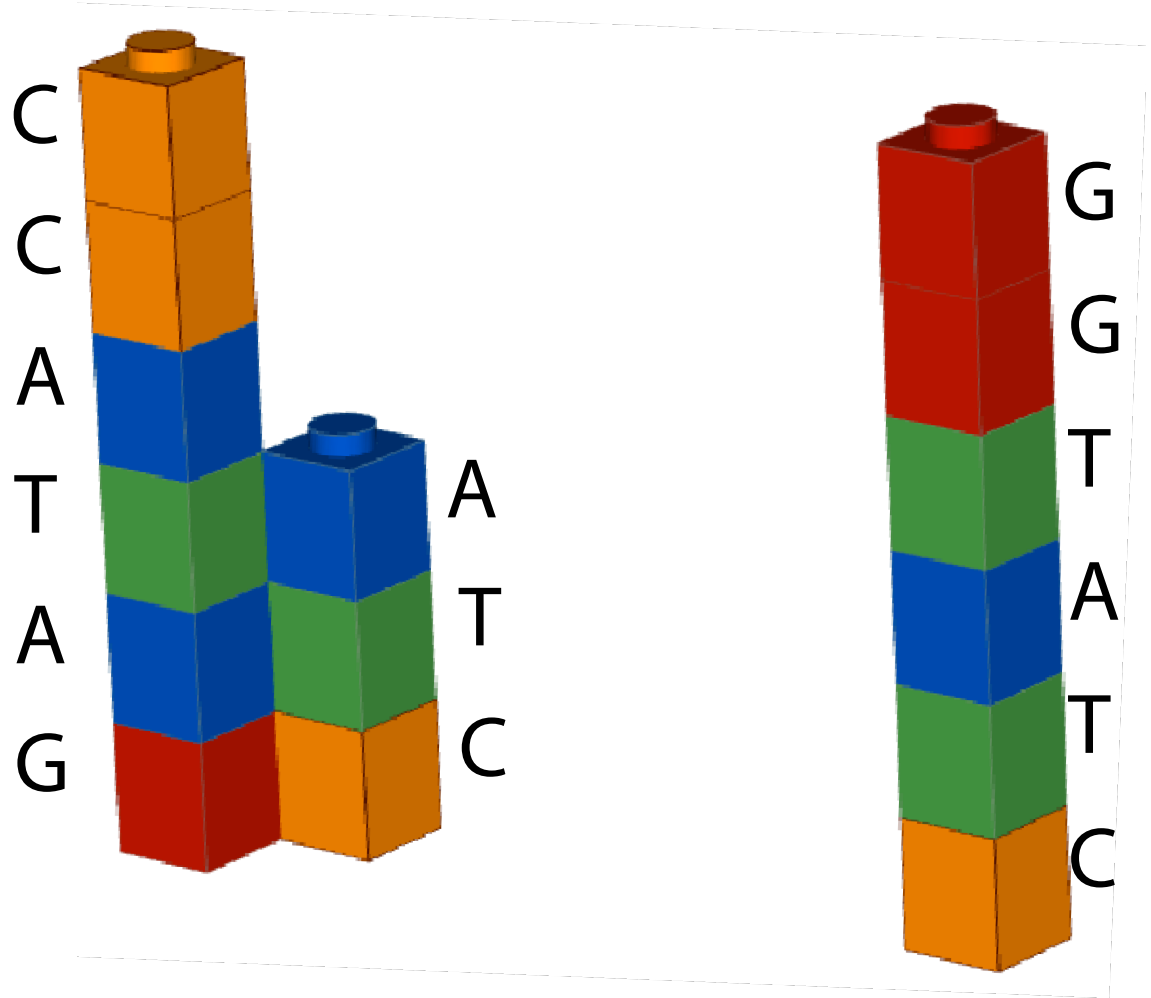
Illumina – Sequencing by synthesis



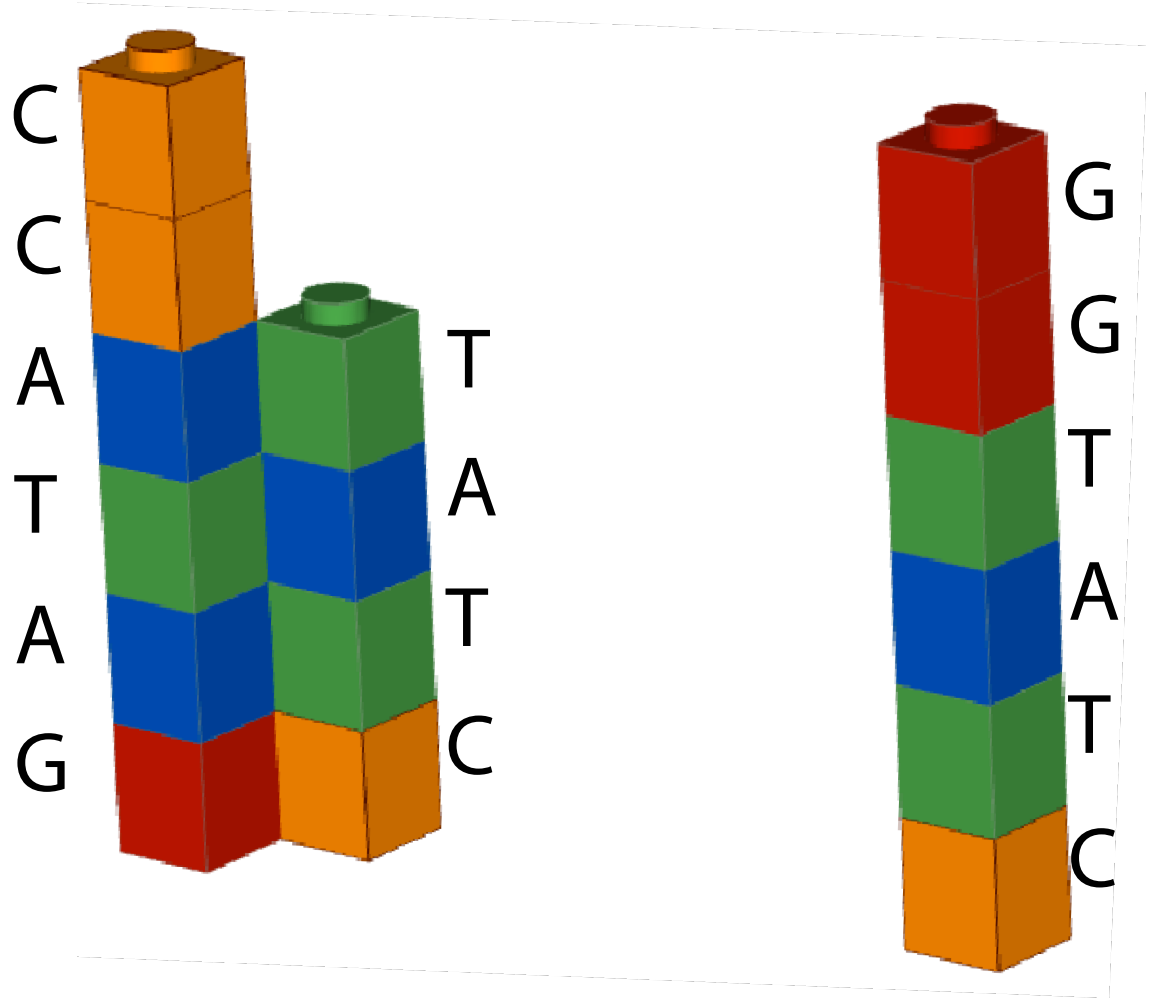
Illumina – Sequencing by synthesis



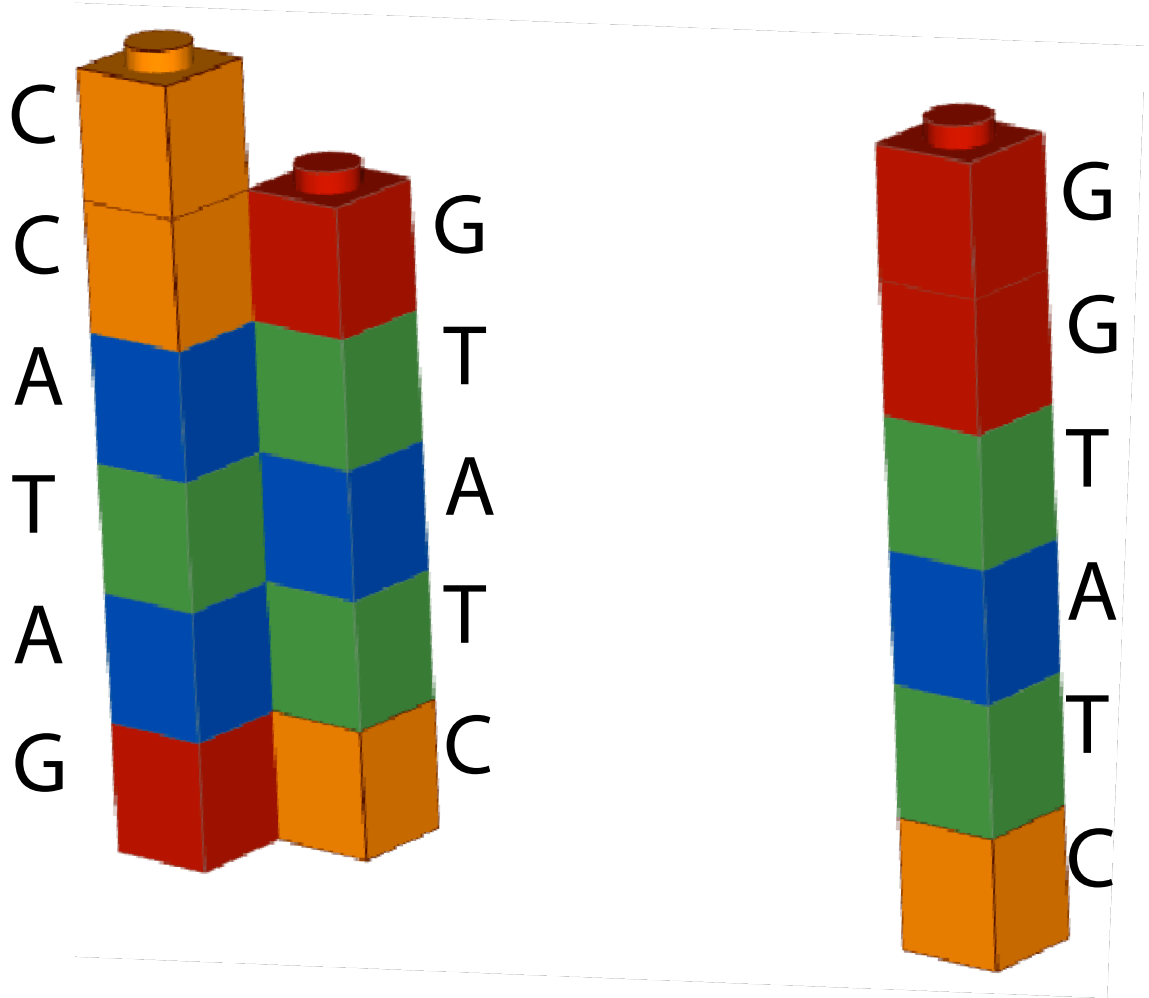
Illumina – Sequencing by synthesis



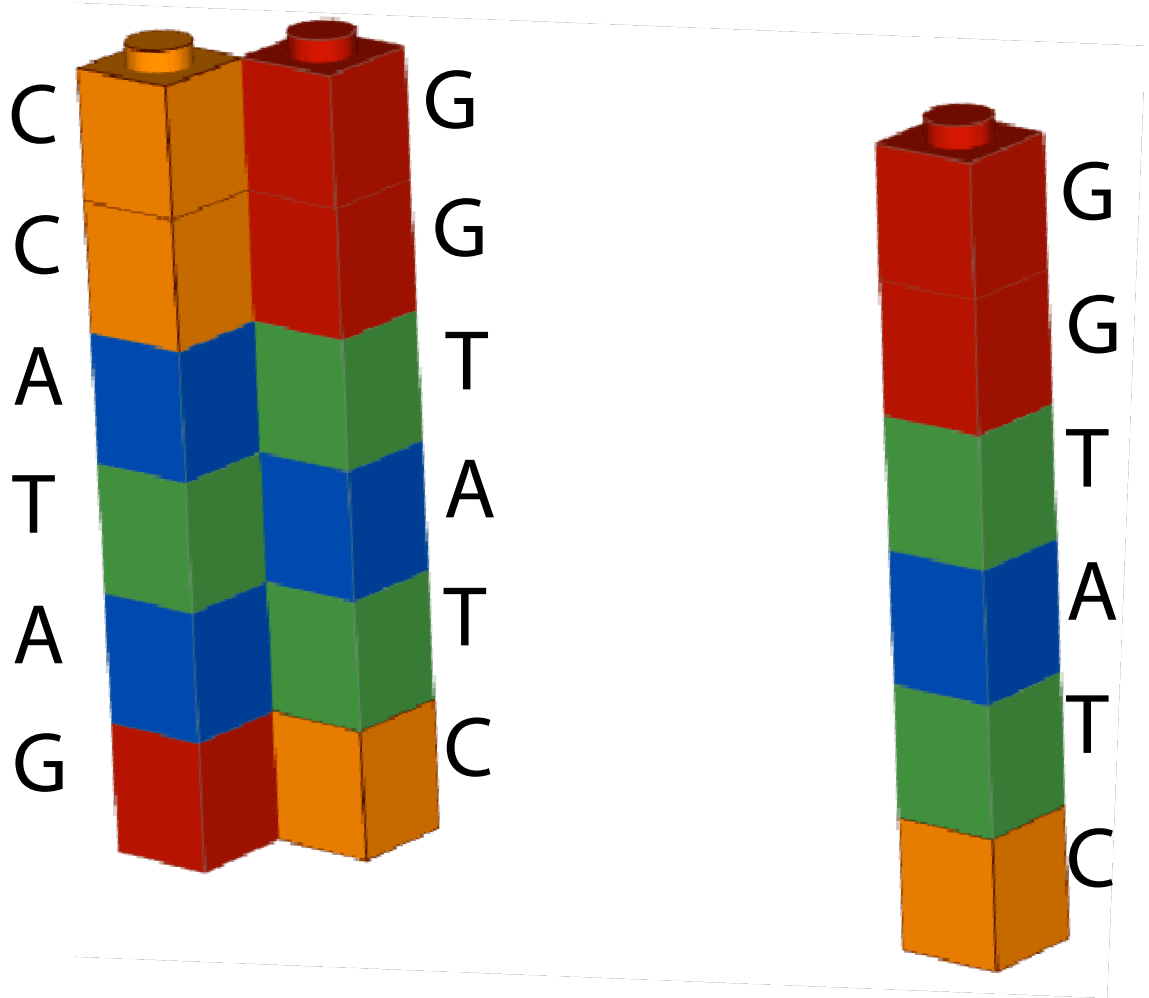
Illumina – Sequencing by synthesis



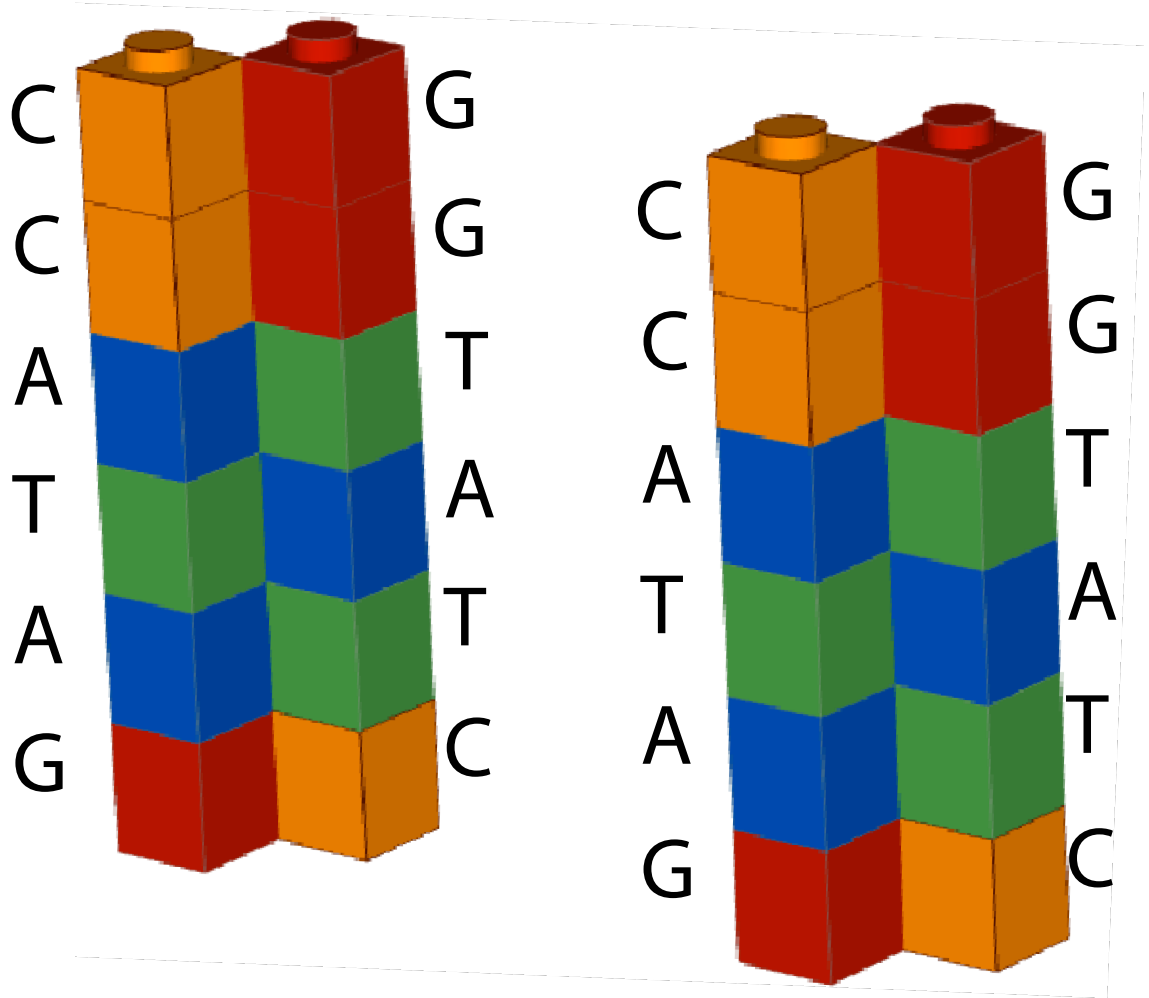
Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis



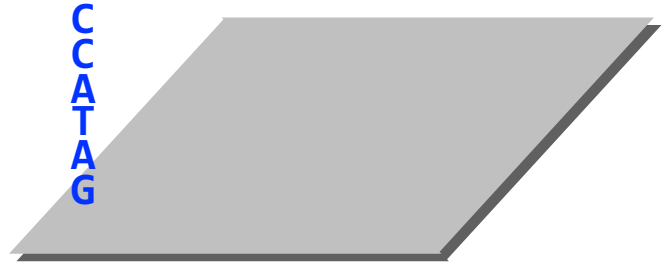
Input DNA

CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT

Cut into snippets

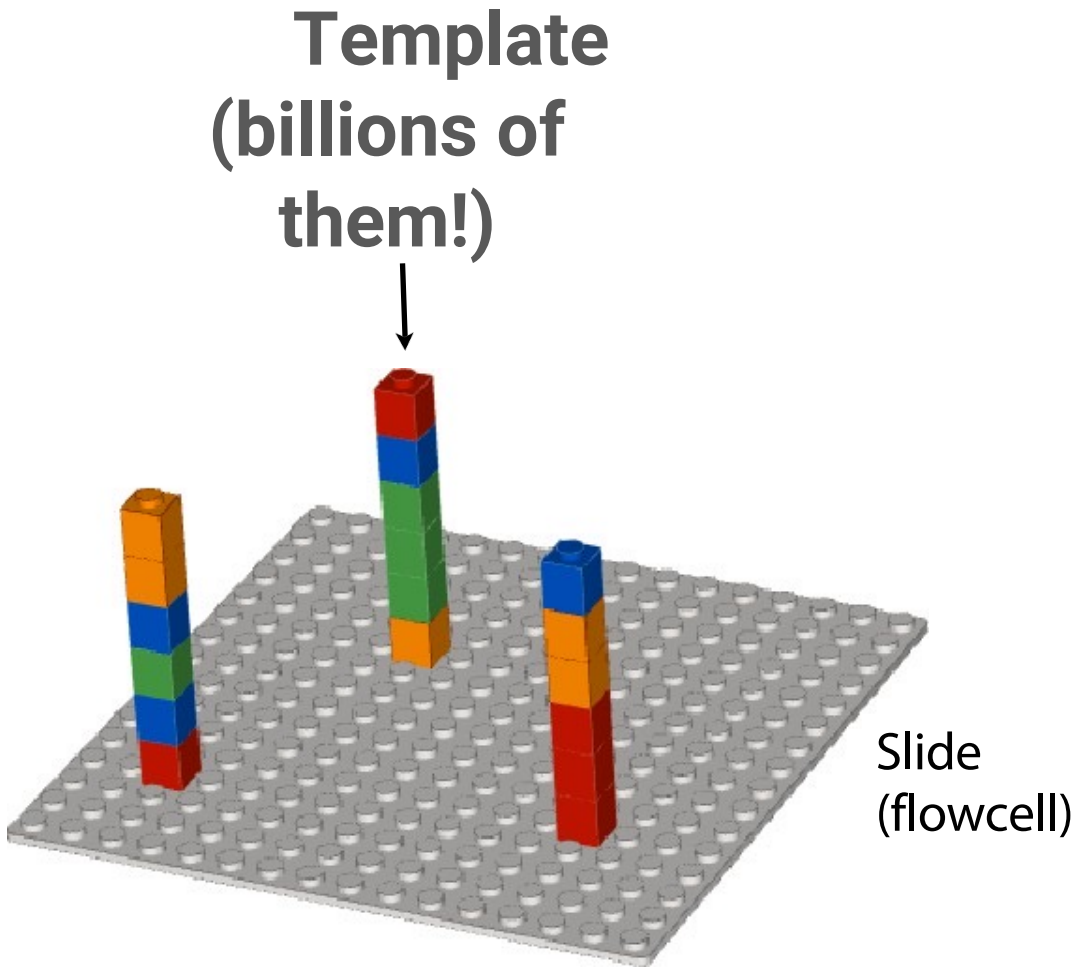
CCATAGTA TATCTCGG CTCTAGGCCCTC ATTTTTT
CCA TAGTATAT CTCGGCTCTAGGCCCTCA TTTTTT
CCATAGTAT ATCTCGGCTCTAG GCCCTCA TTTTTT
CCATAG TATATCT CGGCTCTAGGCCCT CATTTTTT

Deposit on slide

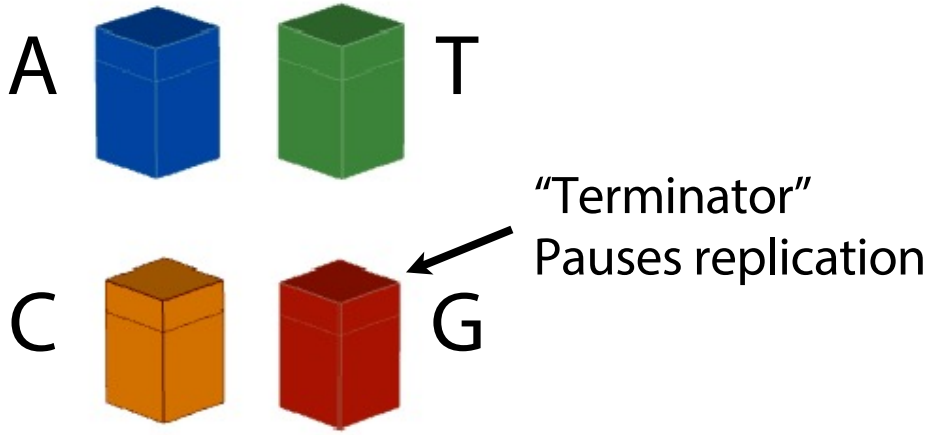


More details: Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6;456(7218):53-9

Illumina – Sequencing by synthesis

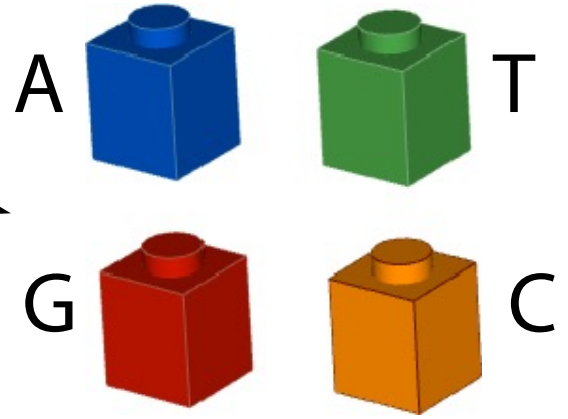


Illumina – Sequencing by synthesis

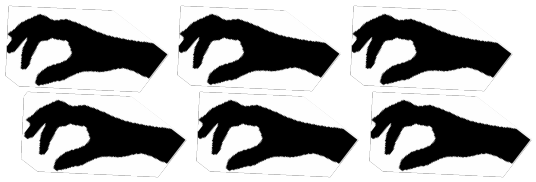


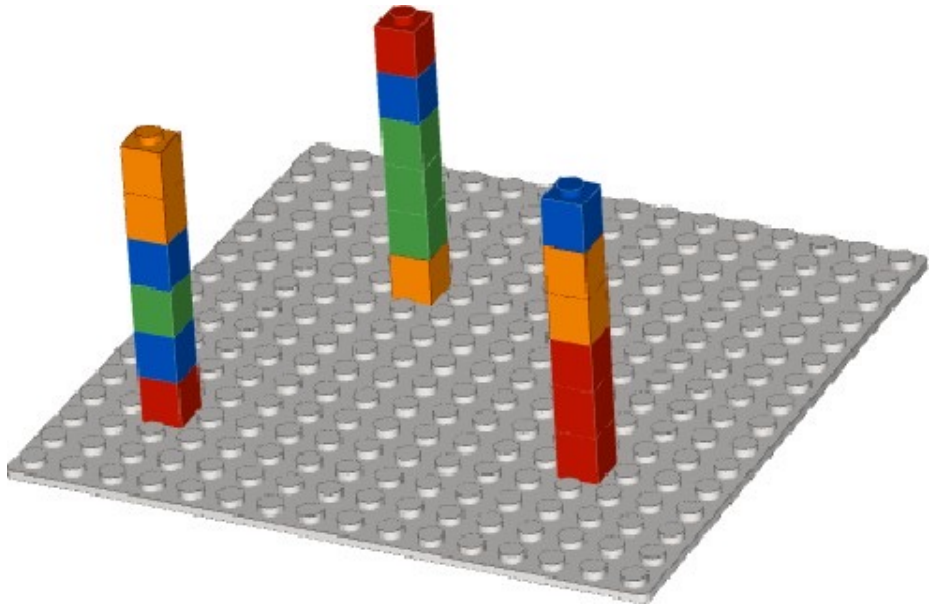
"Terminator"
Pauses replication

Normal
nucleotides

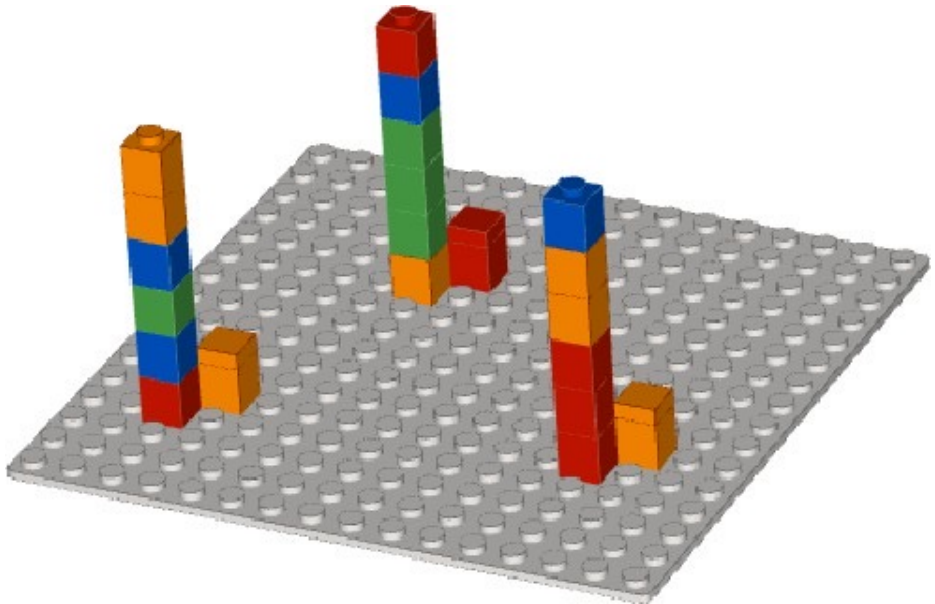


DNA polymerase

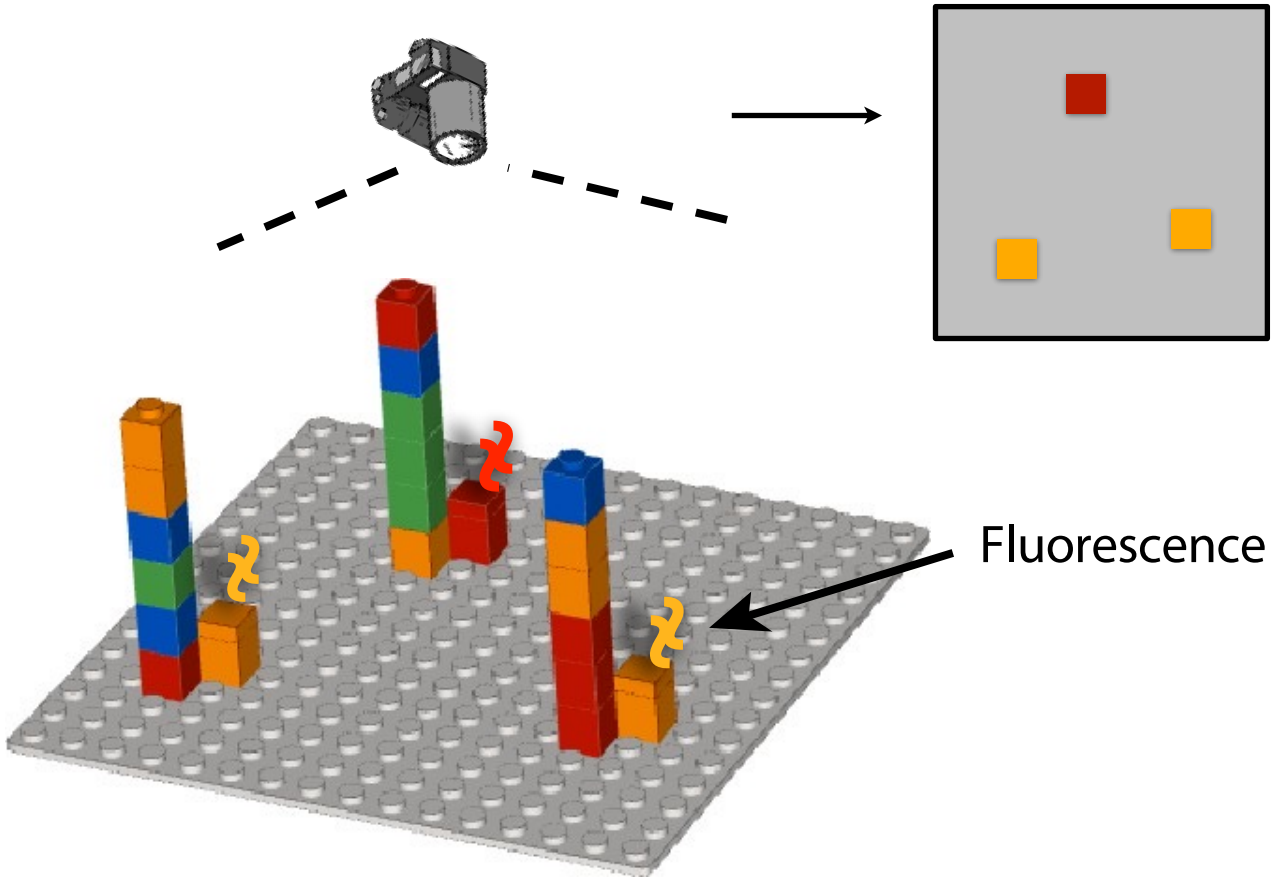




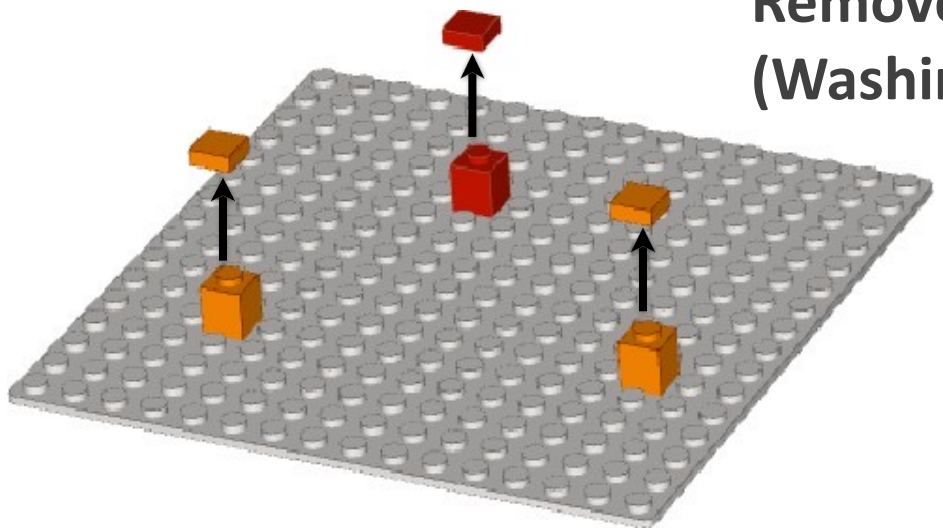
Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis

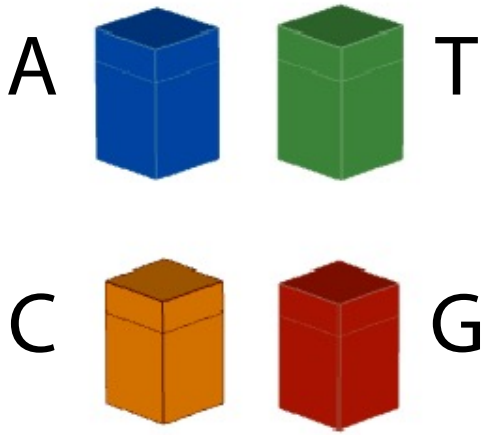


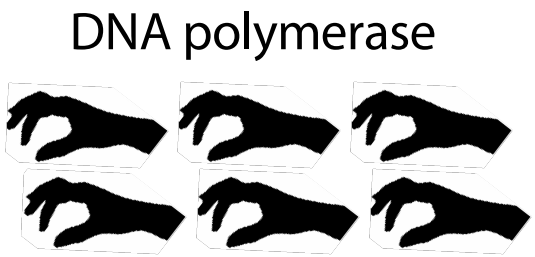
Illumina – Sequencing by synthesis

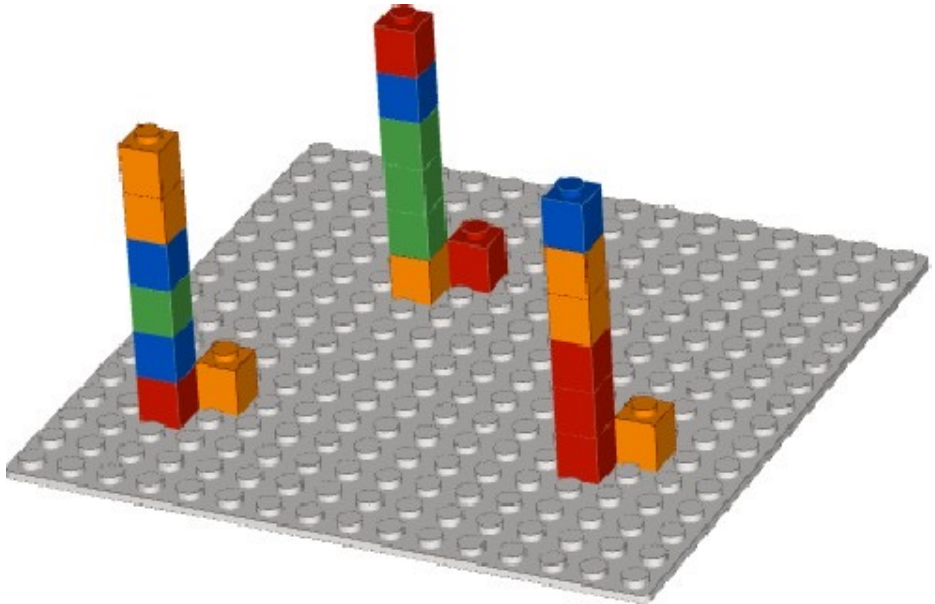


Remove terminators
(Washing step)

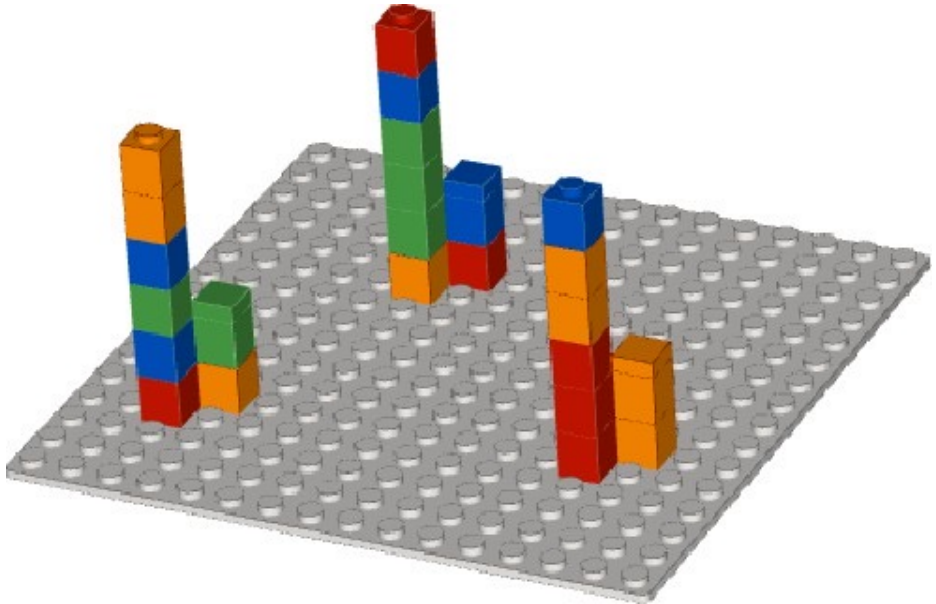
Illumina – Sequencing by synthesis



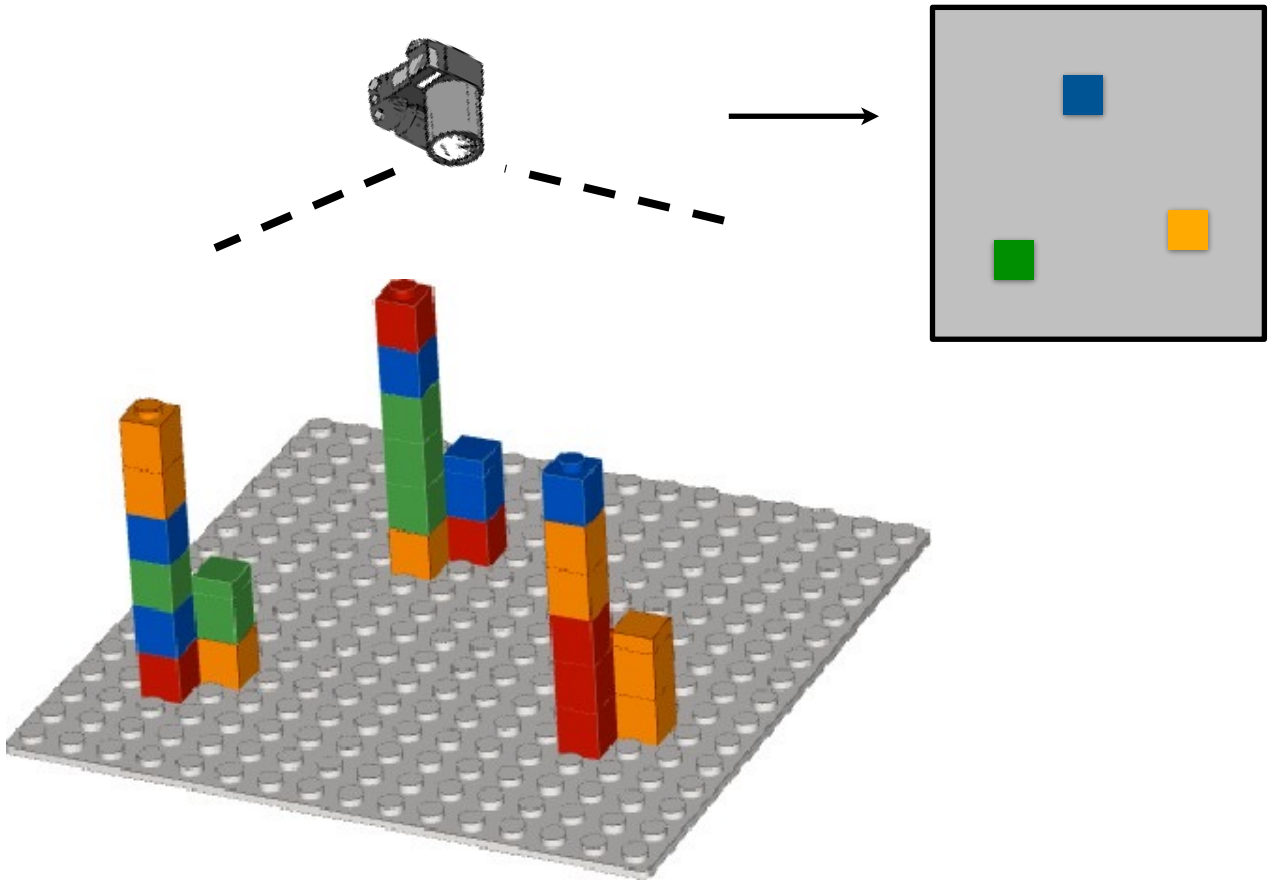




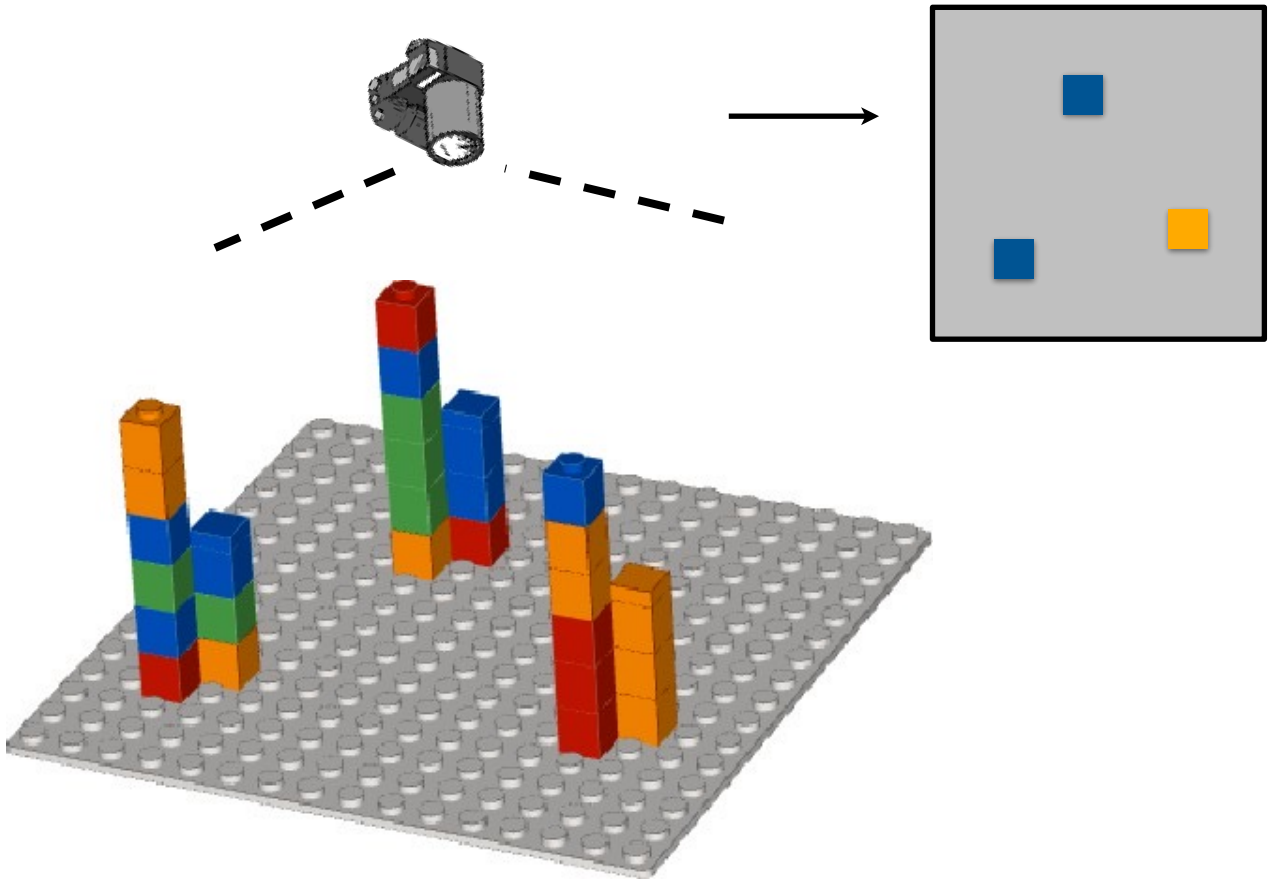
Illumina – Sequencing by synthesis



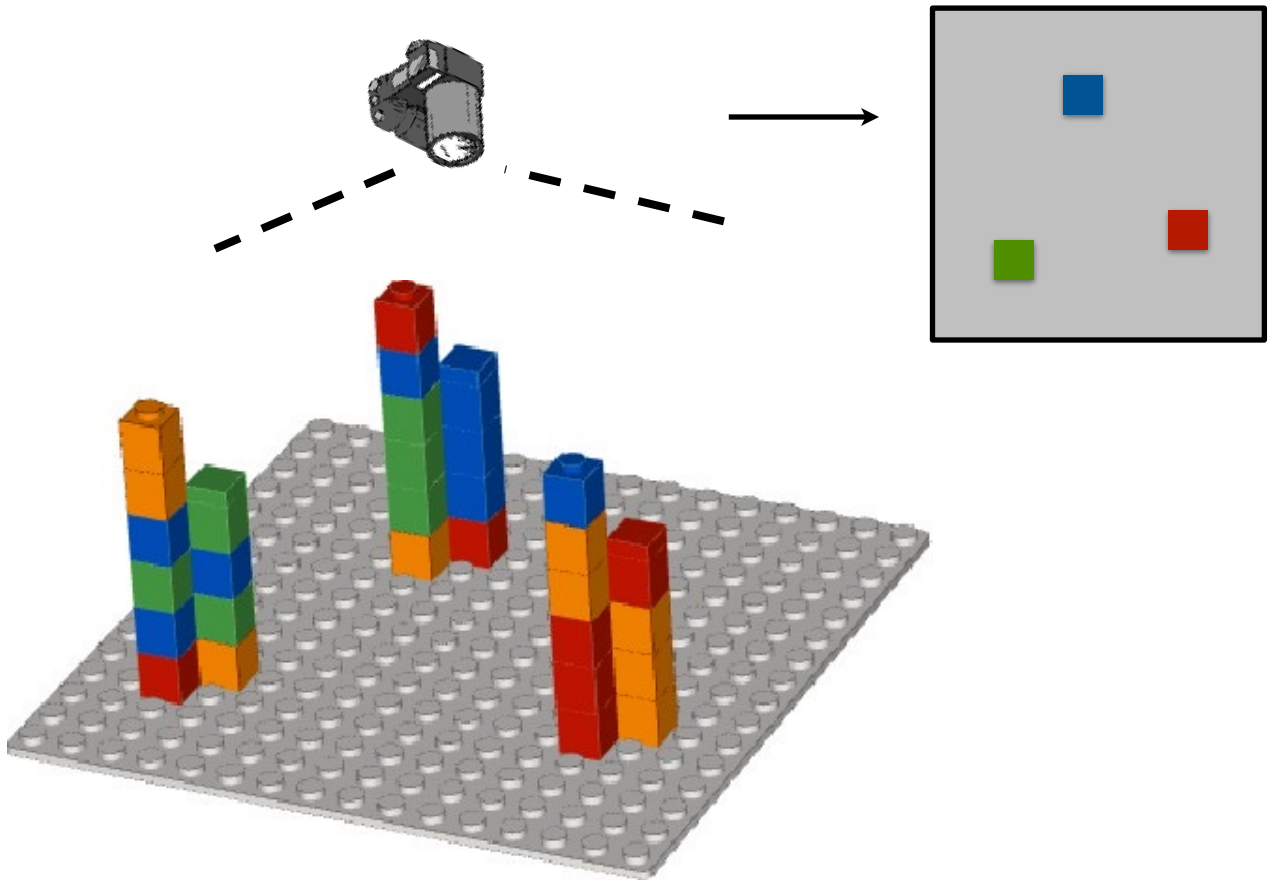
Illumina – Sequencing by synthesis



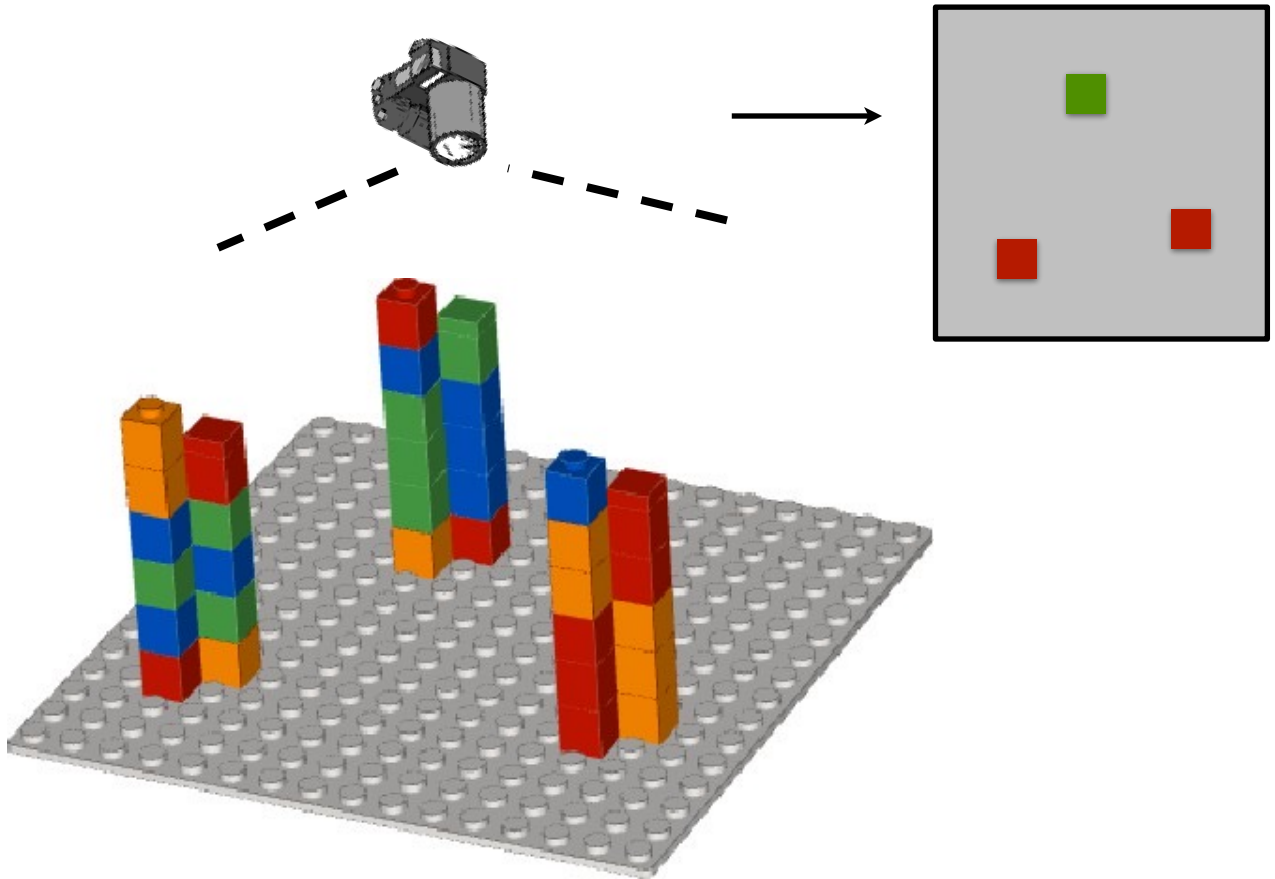
Illumina – Sequencing by synthesis



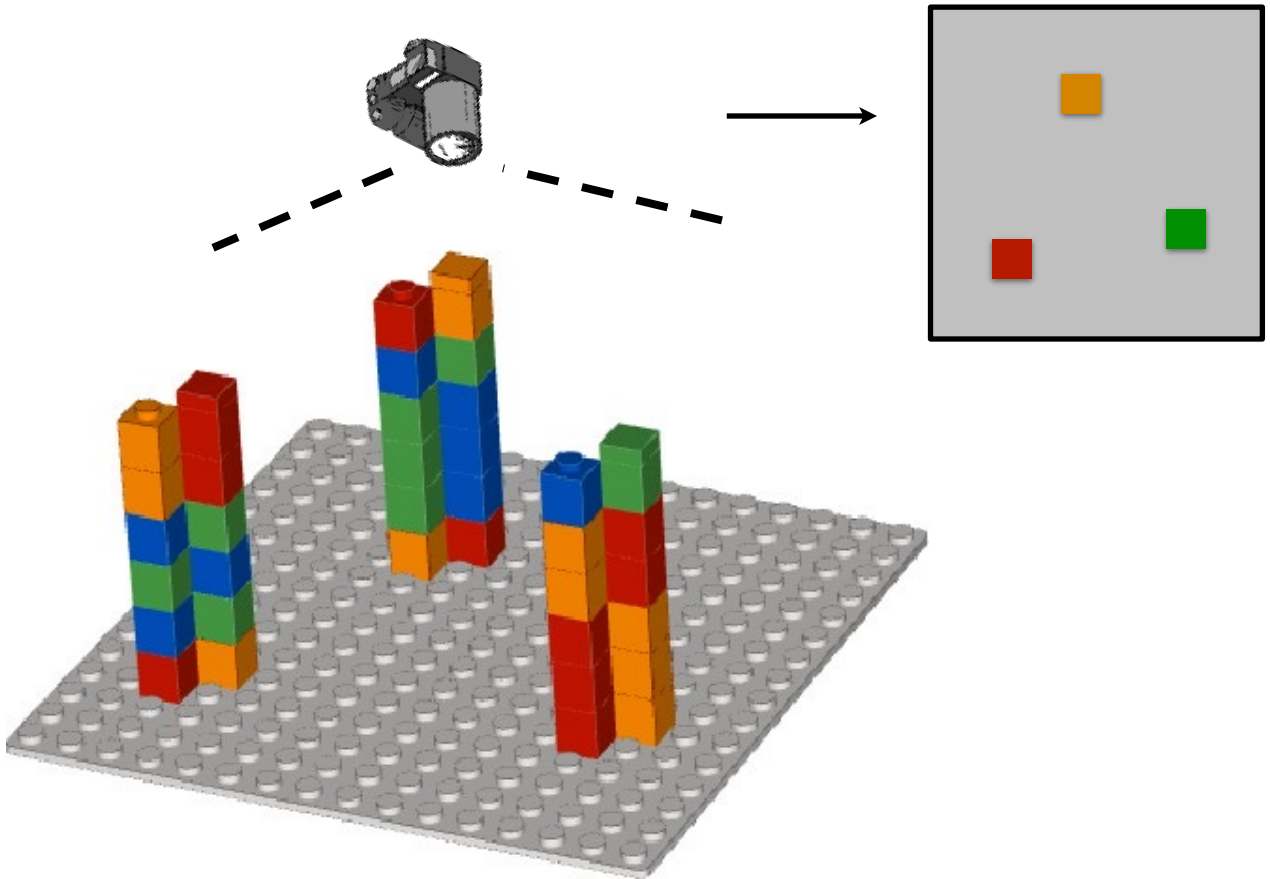
Illumina – Sequencing by synthesis



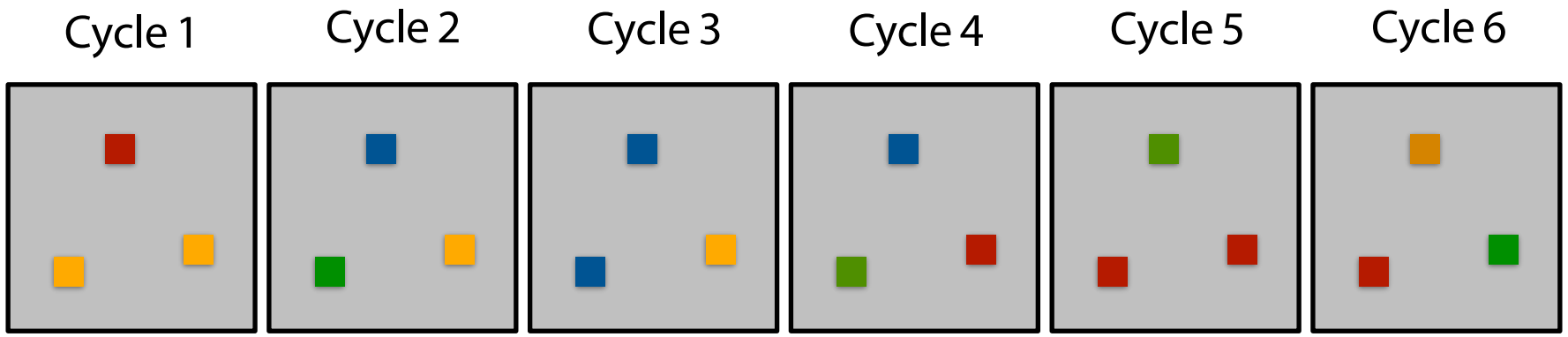
Illumina – Sequencing by synthesis



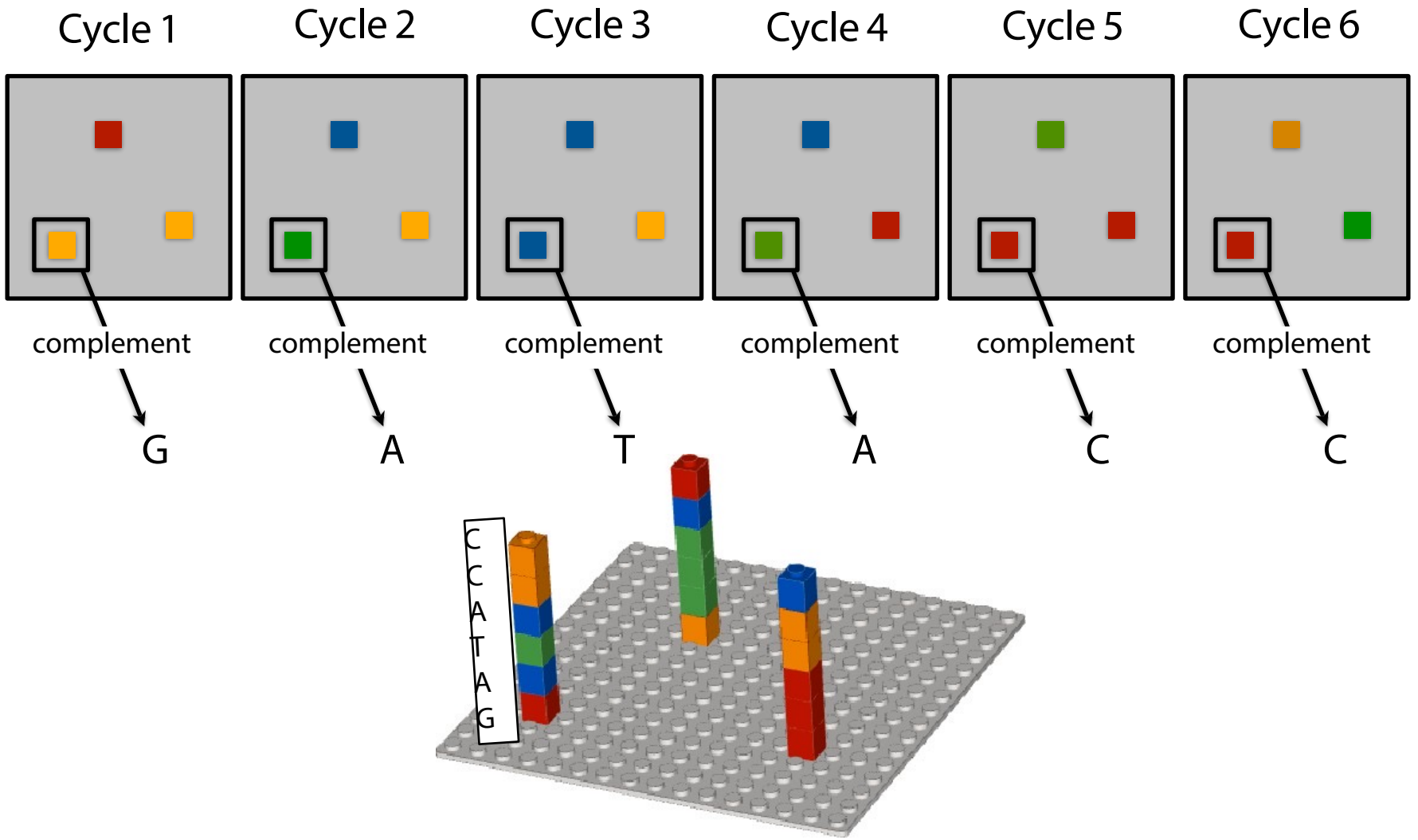
Illumina – Sequencing by synthesis



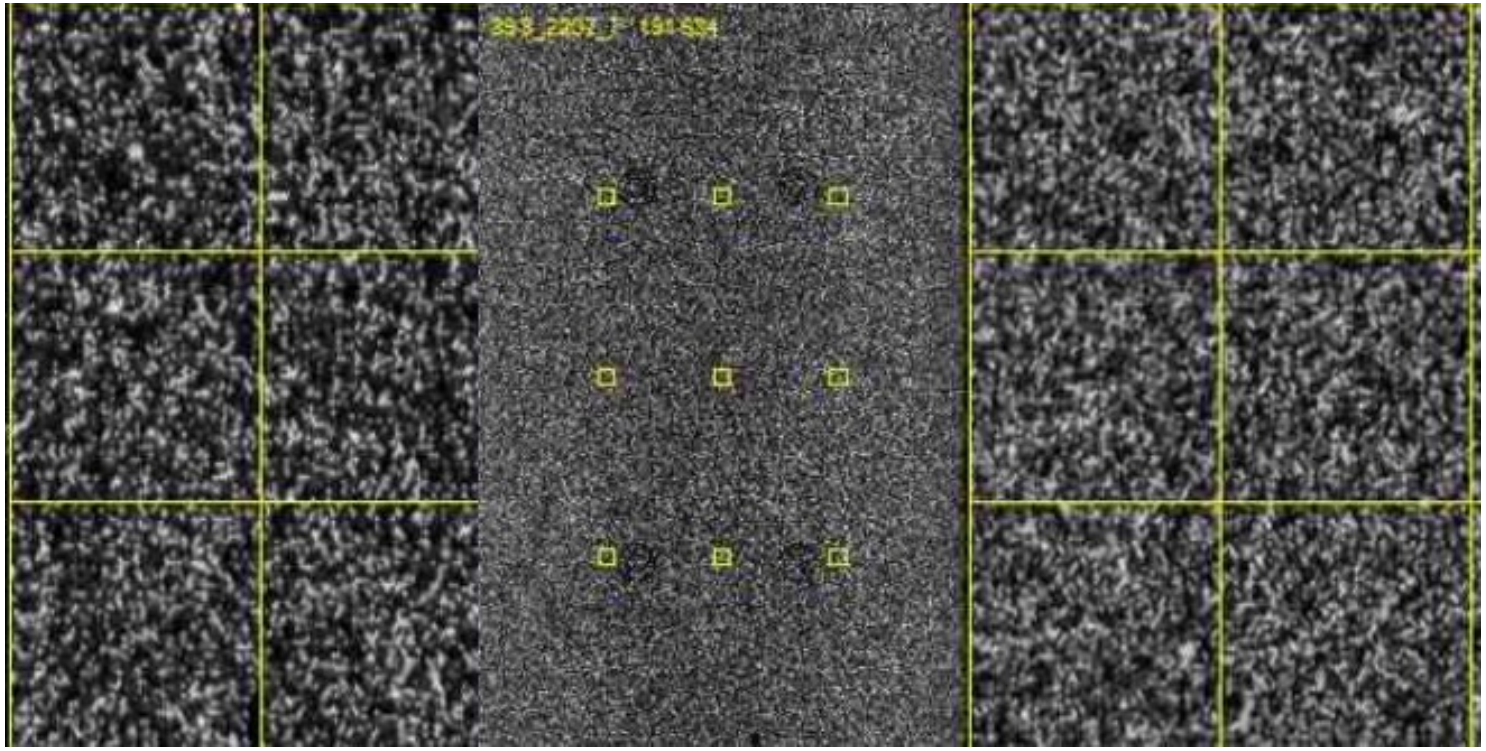
Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis



Actual Illumina HiSeq 3000 image

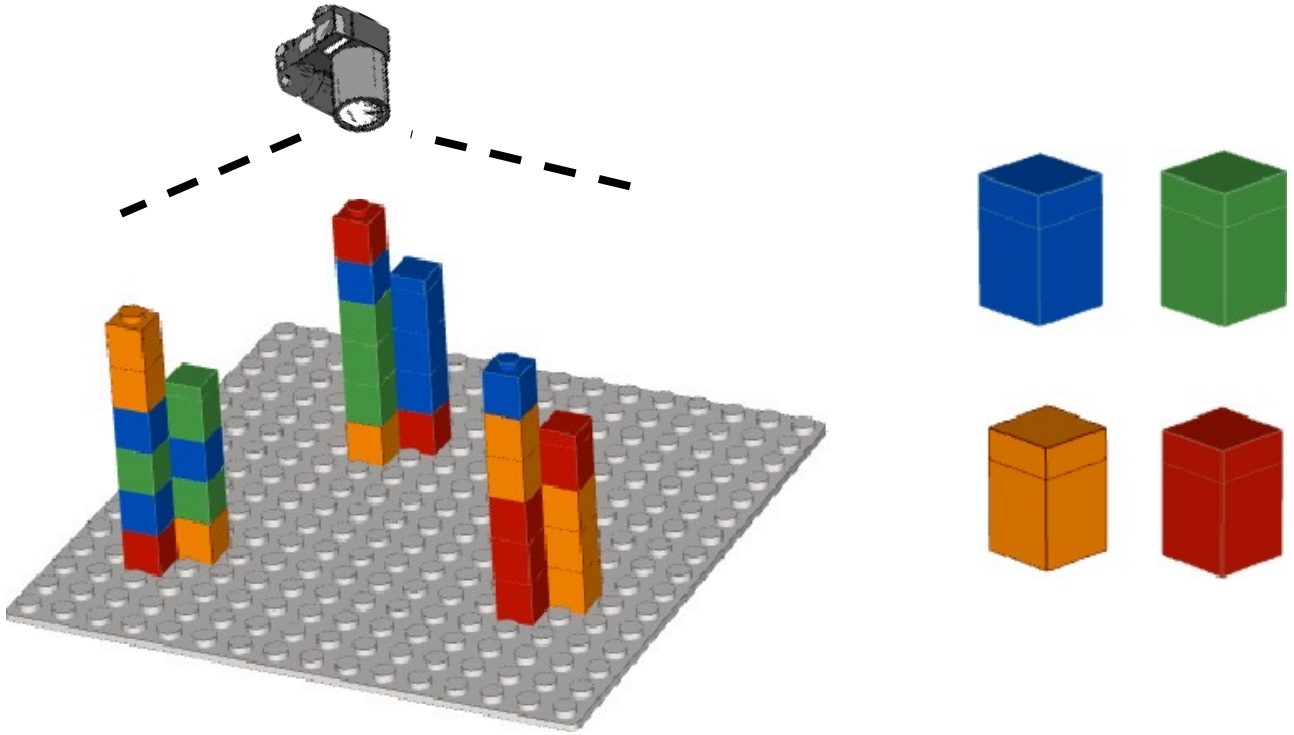
http://dnatech.genomecenter.ucdavis.edu/2015/05/07/first-hiseq-3000-data-download/

Illumina – Sequencing by synthesis

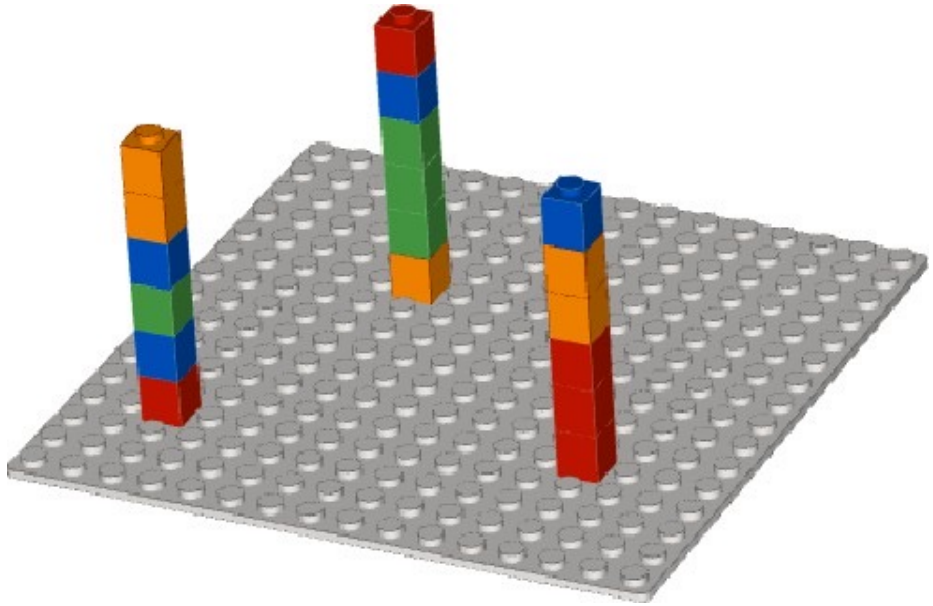
Billions of templates on a slide

Massively parallel: photograph captures all templates simultaneously

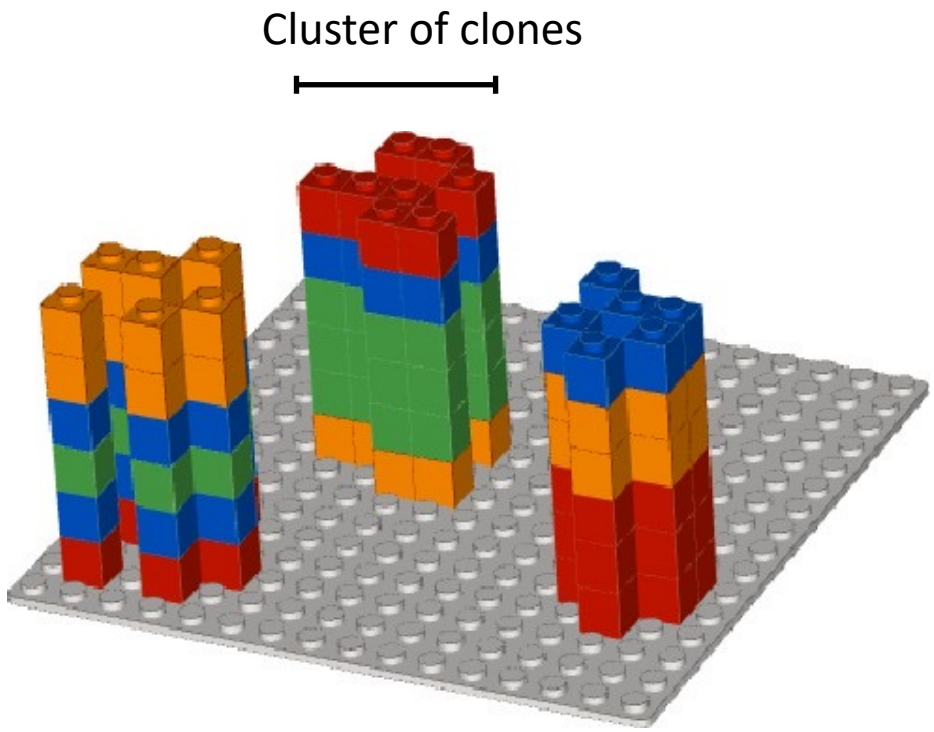
Terminators are "speed bumps," keeping reactions in sync



Illumina – Sequencing by synthesis

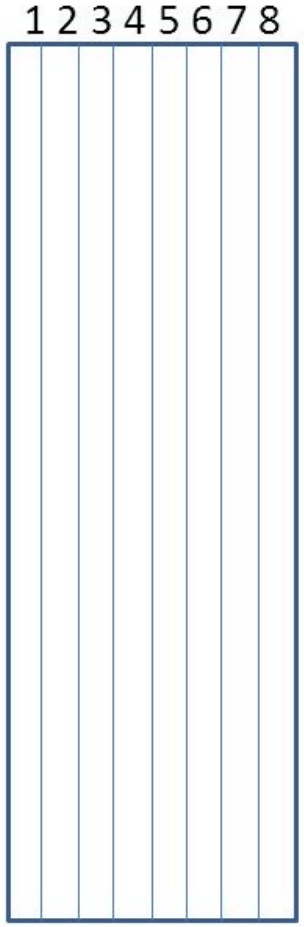


Illumina – Sequencing by synthesis



The Illumina Flowcell

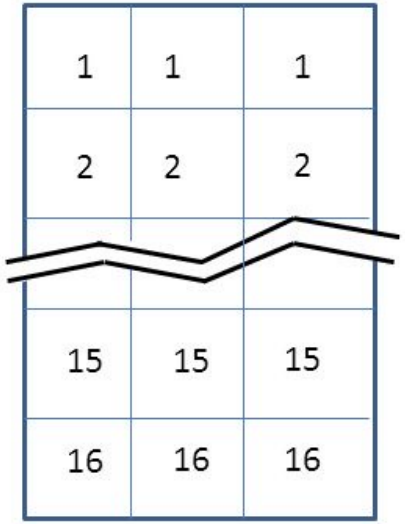
Illumina flowcell geometry (HiSeq)



A flowcell has 8 lanes, which are physically separated. Each surface (upper and lower) of each lane is imaged during each cycle of sequencing in 3 separate "swaths", and 16 images or 'tiles', are collected from each swath, for a total of 96 tiles per lane. The swaths and tiles are not physically separated.

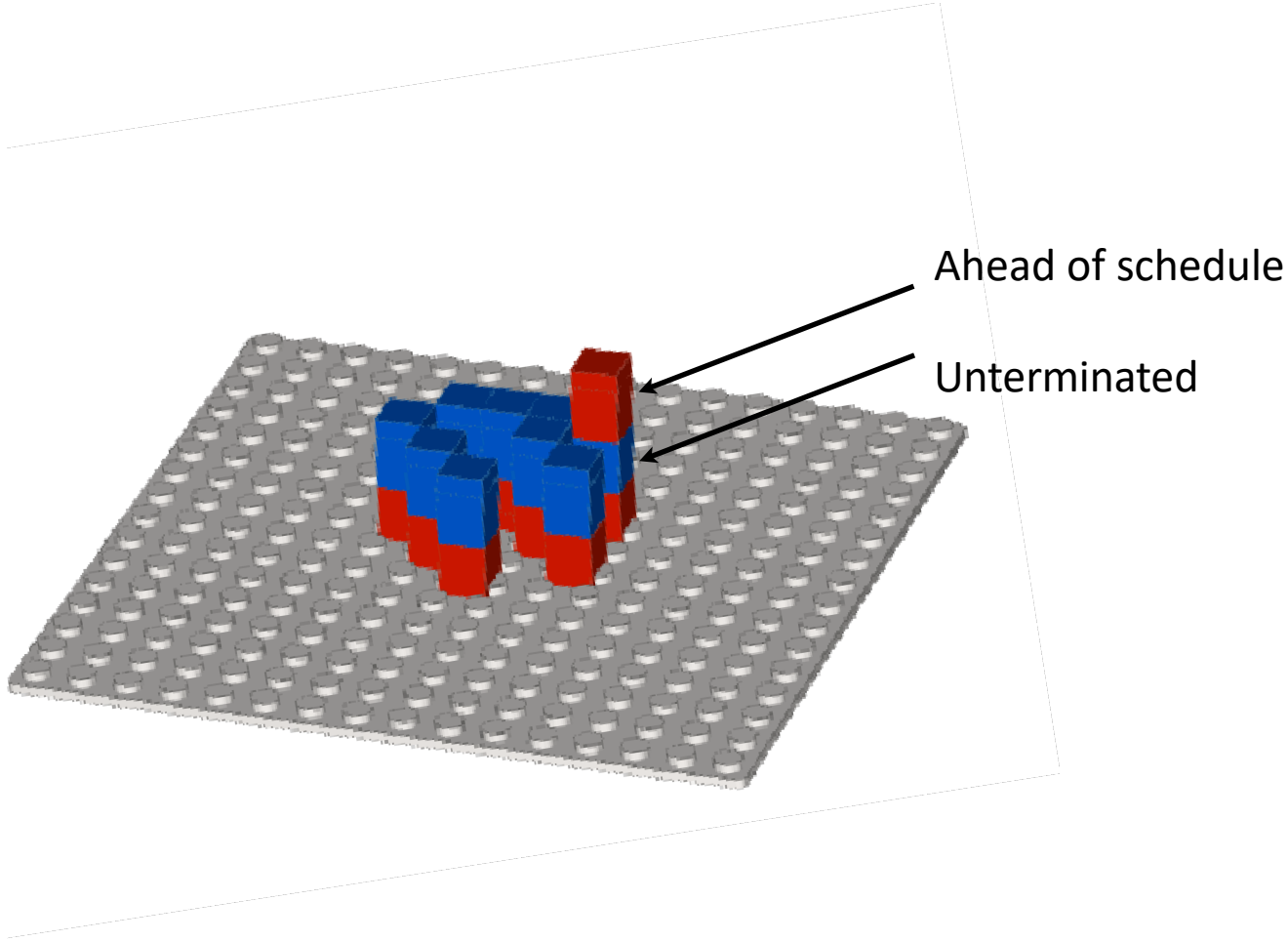
Tiles within a lane are numbered from 1 to 16 down (from outflow end to inflow end), and swaths are numbered from left to right.

The top surface is 1, and the bottom surface is 2. Each tile ID is expressed as a 4-digit number, organized as Surface-Swath-Tile
[12] [123] [01..16]

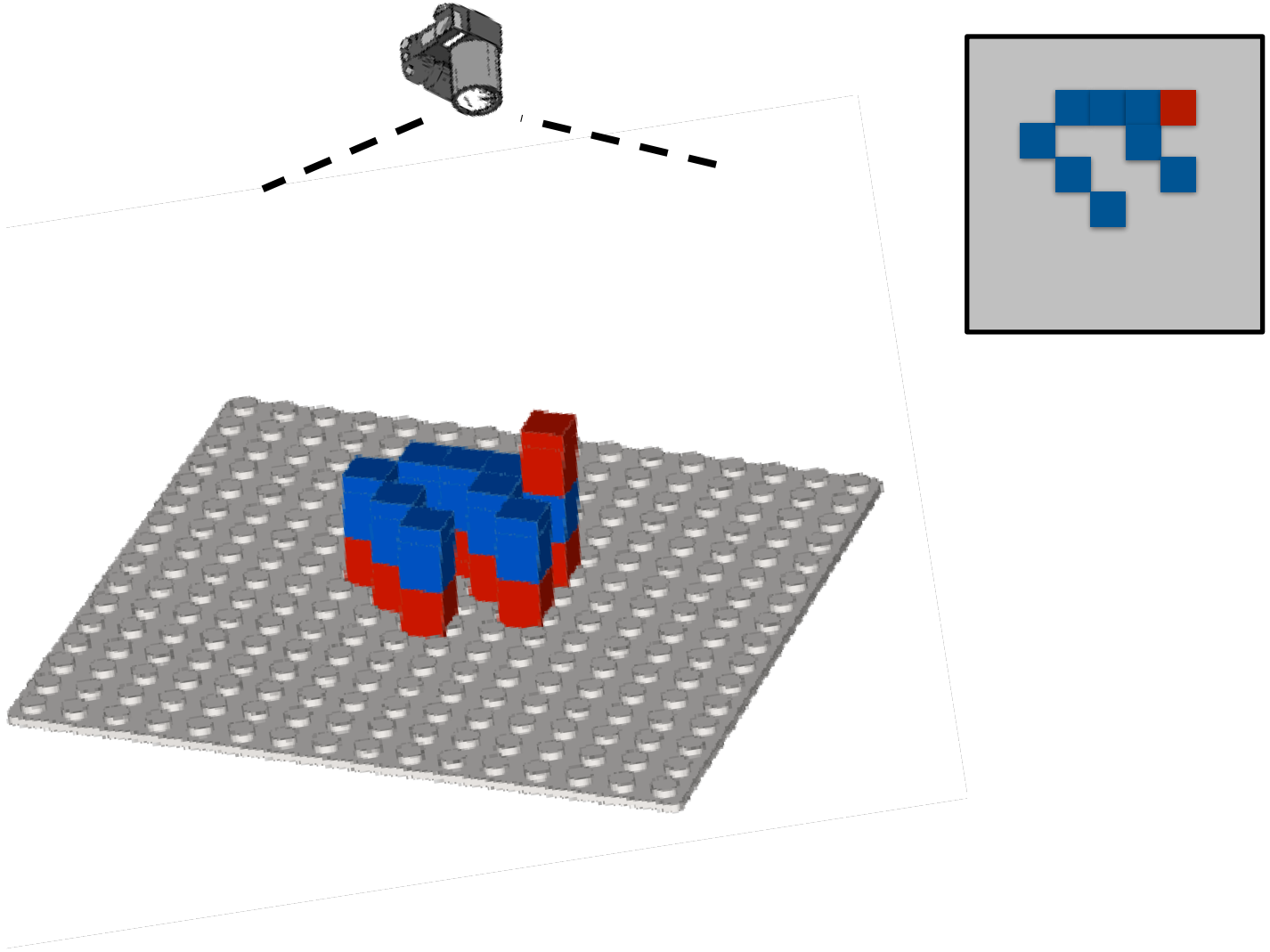


http://seqanswers.com/forums/showthread.php?t=19259

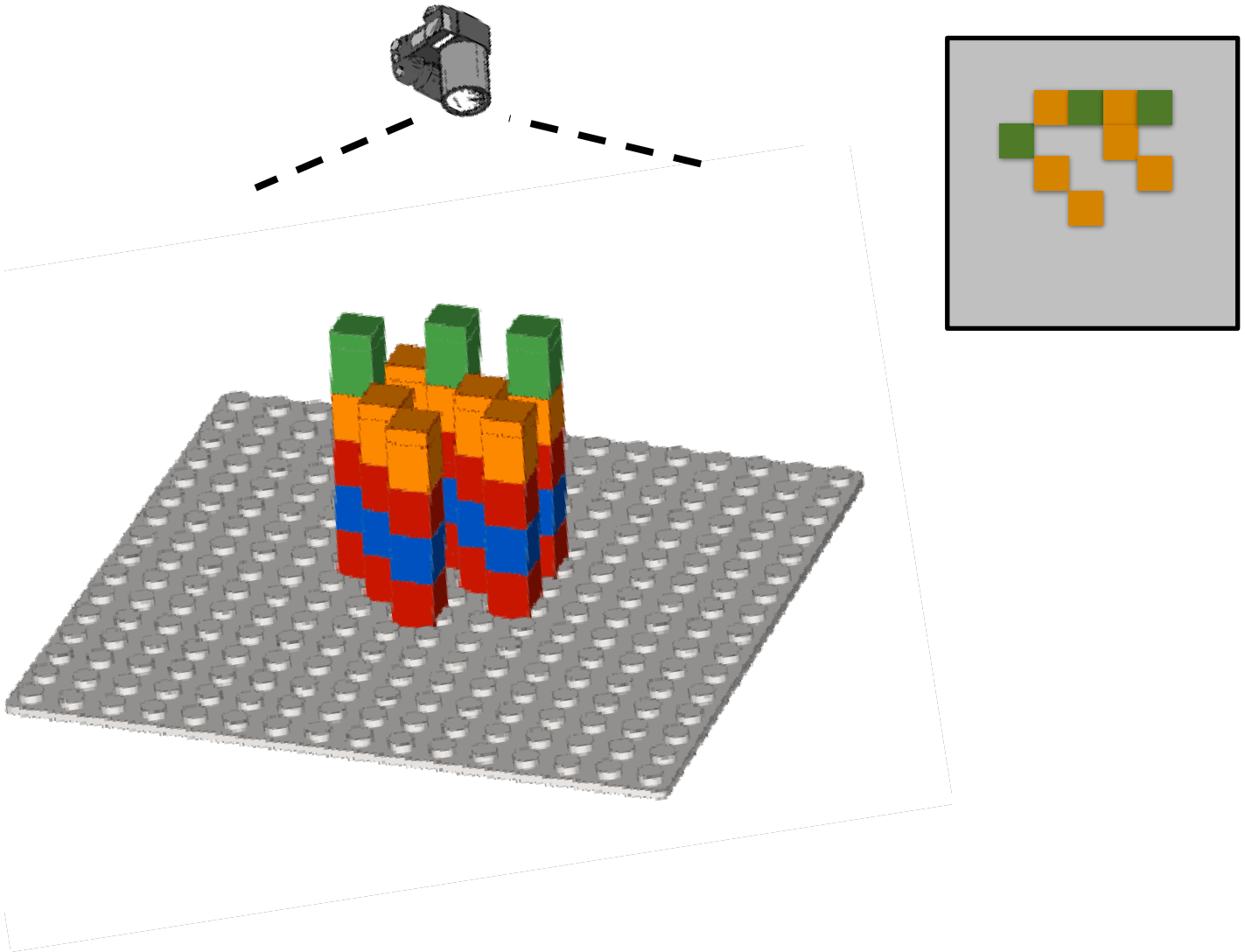
Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis



Illumina – Sequencing by synthesis

$$Q = -10 \cdot \log_{10} p$$

Base quality

Probability that base call is incorrect

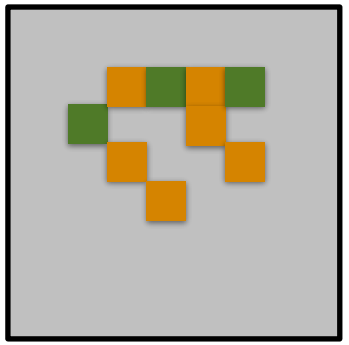
$Q = 10 \rightarrow$ 1 in 10 chance call is incorrect

$Q = 20 \rightarrow$ 1 in 100

$Q = 30 \rightarrow$ 1 in 1,000

Illumina – Sequencing by synthesis

Call: orange (C)



Estimate *p*, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

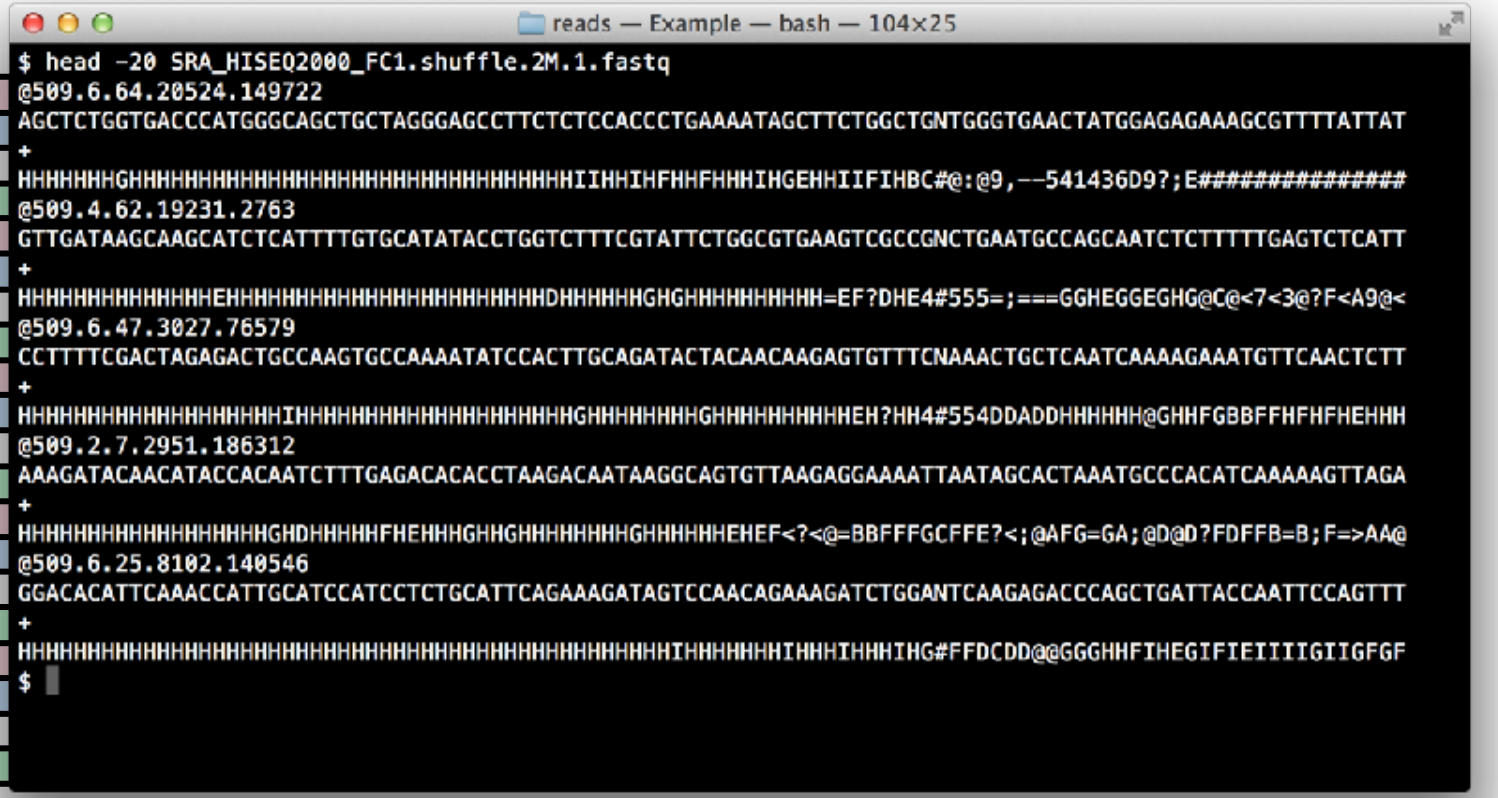
$Q = -10 \log_{10} 1/3 = 4.77$

FASTQ format – Sequencing reads

Name `@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1`
Sequence `ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT`
(ignore) `+`
Base qualities `?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEG:G`

FASTQ format – Sequencing reads

Read 1	Name
	Sequence
	(placeholder)
	Base qualities
Read 2	Name
	Sequence
	(placeholder)
	Base qualities
Read 3	Name
	Sequence
	(placeholder)
	Base qualities
Read 4	Name
	Sequence
	(placeholder)
	Base qualities
Read 5	Name
	Sequence
	(placeholder)
	Base qualities



FASTQ format – Sequencing reads

- Bases and qualities line up:
  - AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
  - !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
    HHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
- Base quality is ASCII-encoded version of $Q = -10 \log_{10} p$

FASTQ format – Sequencing reads

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

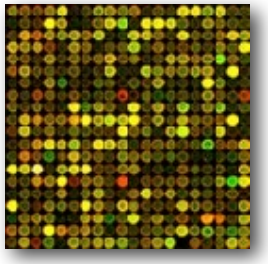
ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010

PacBio

Oxford NANOPORE Technologies

Third Generation Sequencing

Common components

- Single molecule sequencing
- Long-reads
- Do not require amplification
- Sequencing results in real-time

Differences

- Different detection mechanisms
- Library preparation chemistry
- Flow cell configuration

Long read platforms comparison

PacBio
Pros: accuracy, read length, detect base modifications
Cons: instrument cost

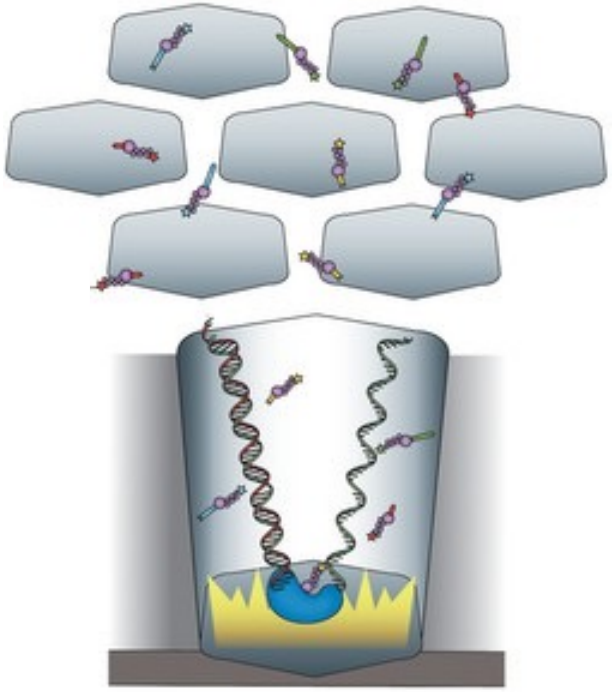
Oxford Nanopore
Pros: ultra long reads, cheap(er) instruments, direct RNA, detect base modifications
Cons: slightly lower accuracy

Long read sequencing

Pacific Biosciences sequencing

DNA templates are captured by DNA polymerase at the bottom of a tiny well

As the polymerase copies the template strand the incorporation events are recorded as flashes of light

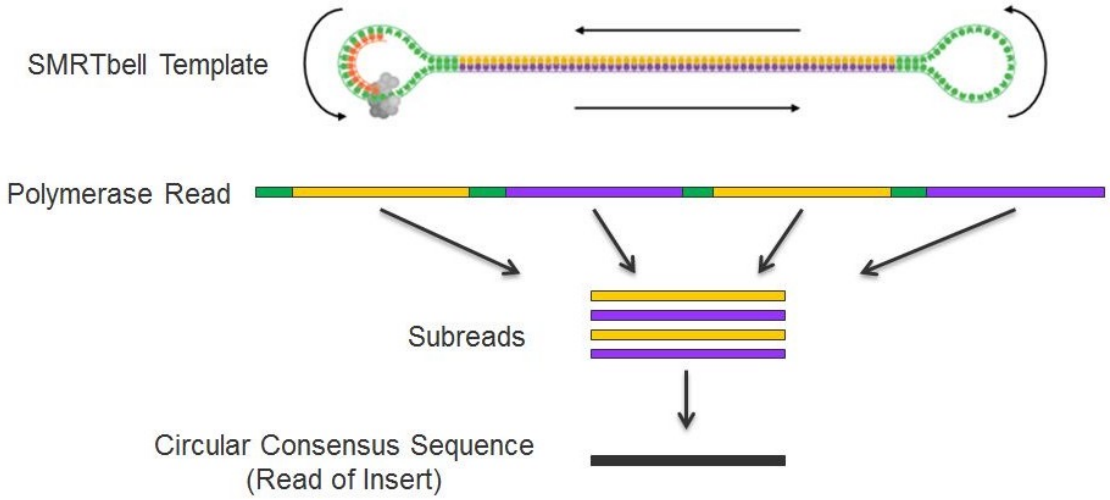


PacBio sequencing

Pacific Biosciences HiFi sequencing

Clever molecular biology trick: add hairpins to either end of DNA molecule allowing sequencing to go around in a circle

Perform multiple passes to create a *consensus* read with ~99.9% accuracy



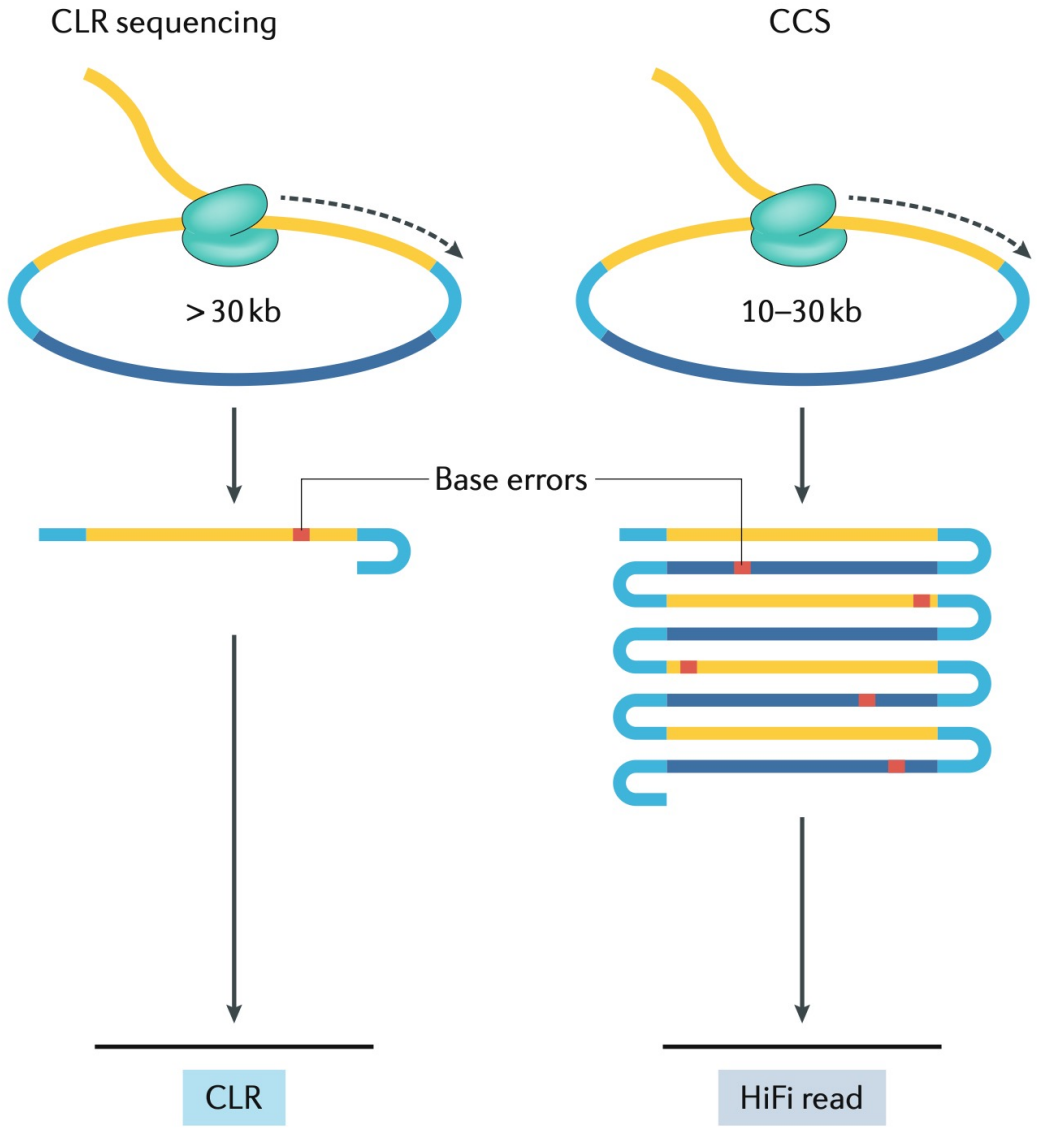
PacBio sequencing

CLR:

- It sequences the DNA strand once.
- Reads are long (>30,000bp) but fairly error prone (10-15% errors)
- Using this approach as single molecule sequencing has low signal-to-noise ratio

CCS:

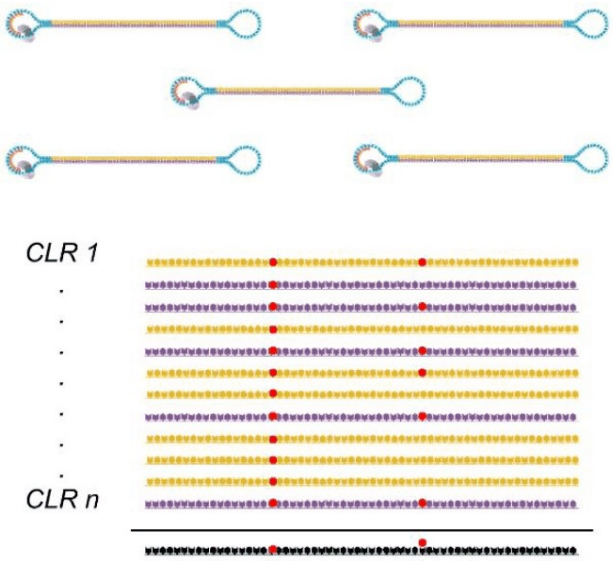
- It sequences the strand multiple times.
- Reads are shorter (~10,000bp) but of high quality (0.1% errors)



PacBio sequencing

Continuous Long Read (CLR) Sequencing Mode

Inserts >25 kb, up to 175 kb



LONG READS

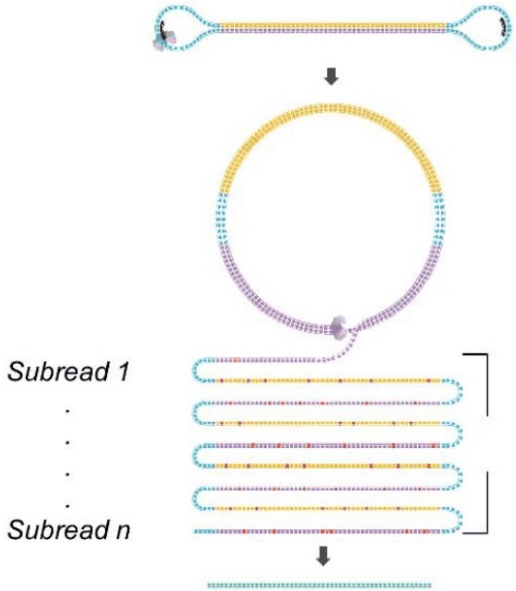
Multi-molecule consensus sequence

90% Accurate

Multiple molecules; single reads

Circular Consensus Sequencing (CCS) Mode

Inserts 10-20 kb



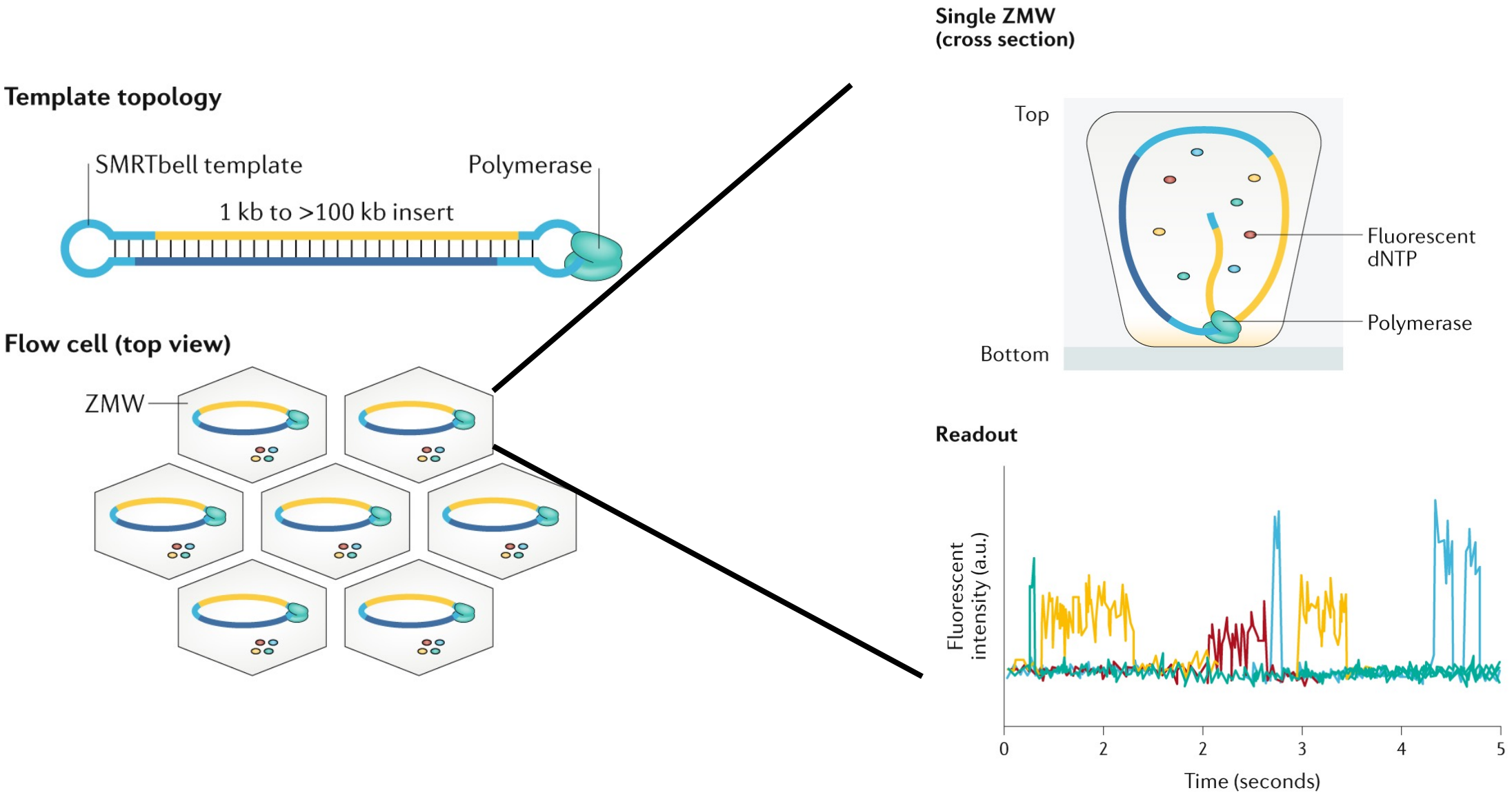
HiFi READS

Single-molecule consensus sequence

99% Accurate

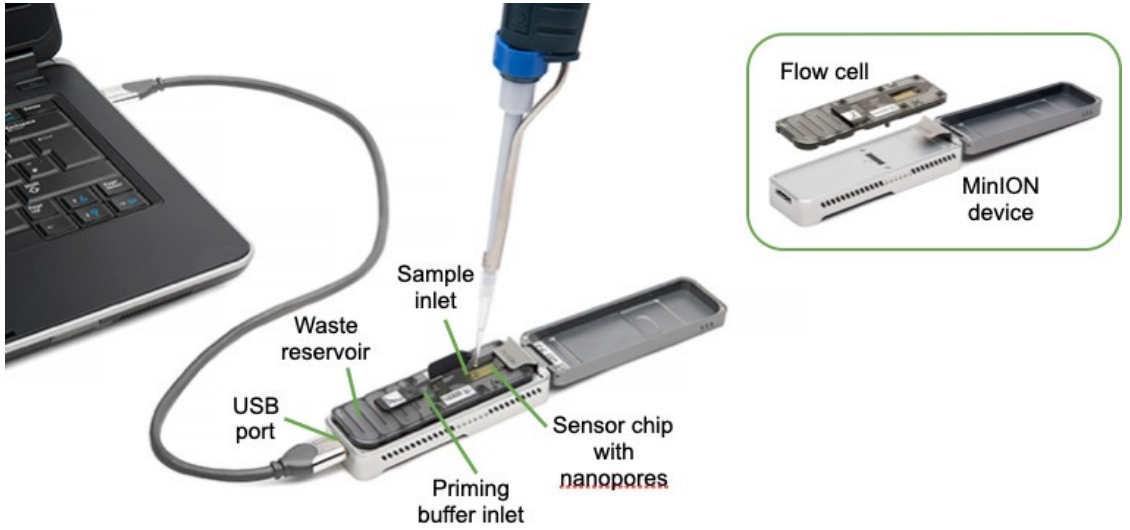
Single molecule; multiple reads

Pacific Biosciences - PacBio



Nanopore sequencing

Miniaturized sequencing device that connects to a standard laptop



Nanopore sequencing

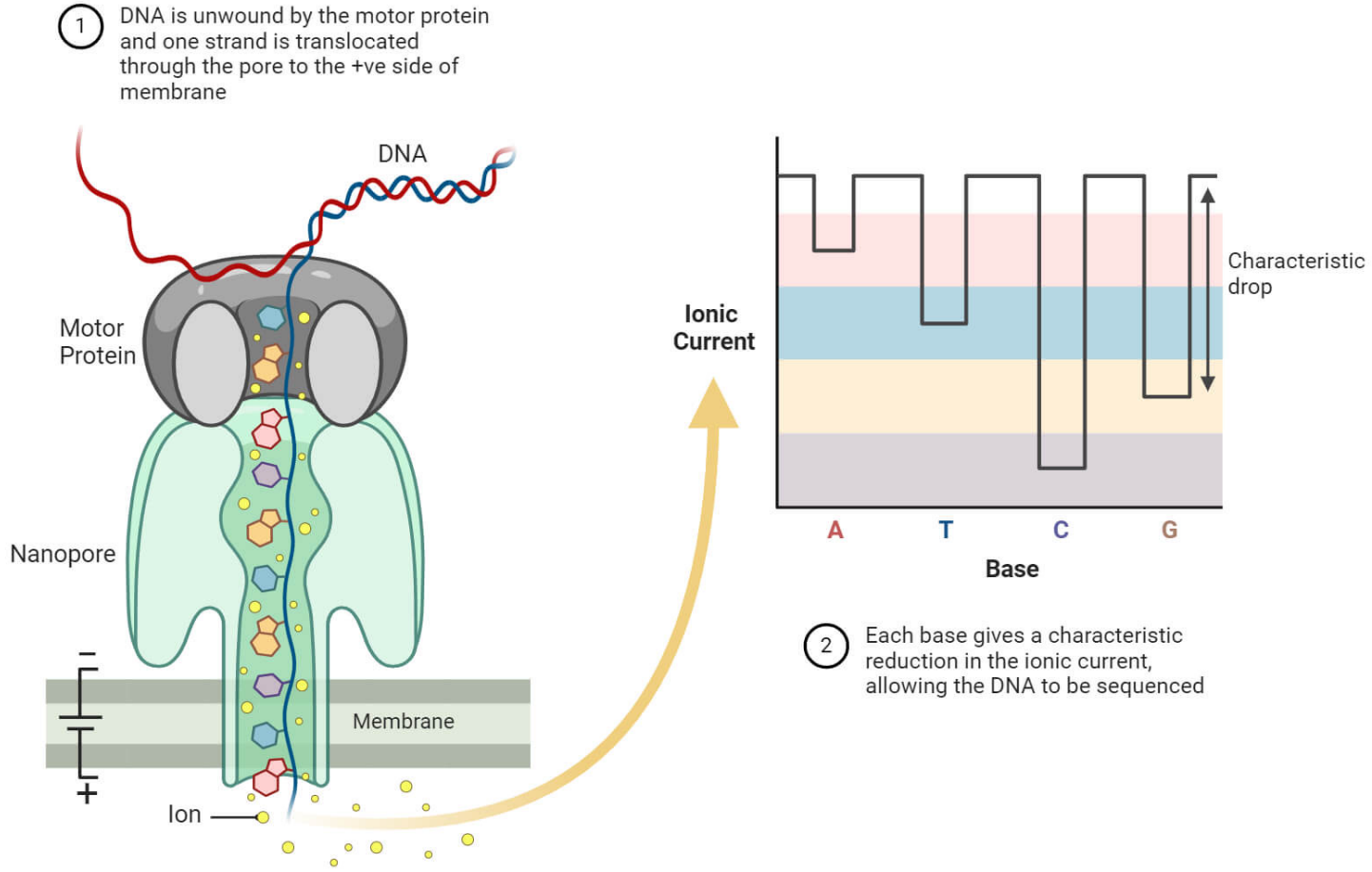
Oxford Nanopore Sequencing

DNA is pushed through a *nanopore* (a transmembrane pore protein) using a motor protein

The system senses electrical signal, and every nucleotide has a different electric current

Machine learning is used to decode “squiggle” into a predicted read sequence

Reads can be very long (>100kb), ~99% accurate



Nanopore sequencing

Oxford Nanopore Sequencing

DNA is pushed through a *nanopore* (a transmembrane pore protein) using a motor protein

The system senses electrical signal, and every nucleotide has a different electric current

Machine learning is used to decode “squiggle” into a predicted read sequence

Reads can be very long (>100kb), ~99% accurate

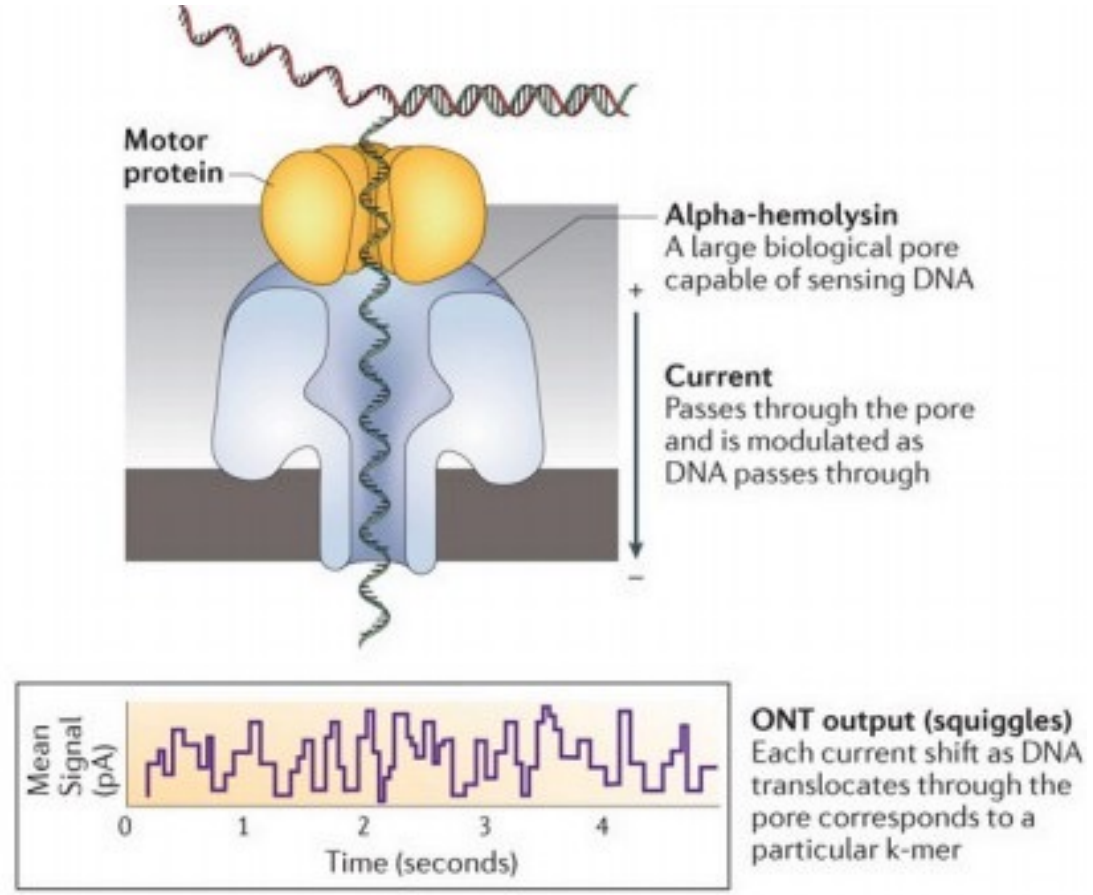
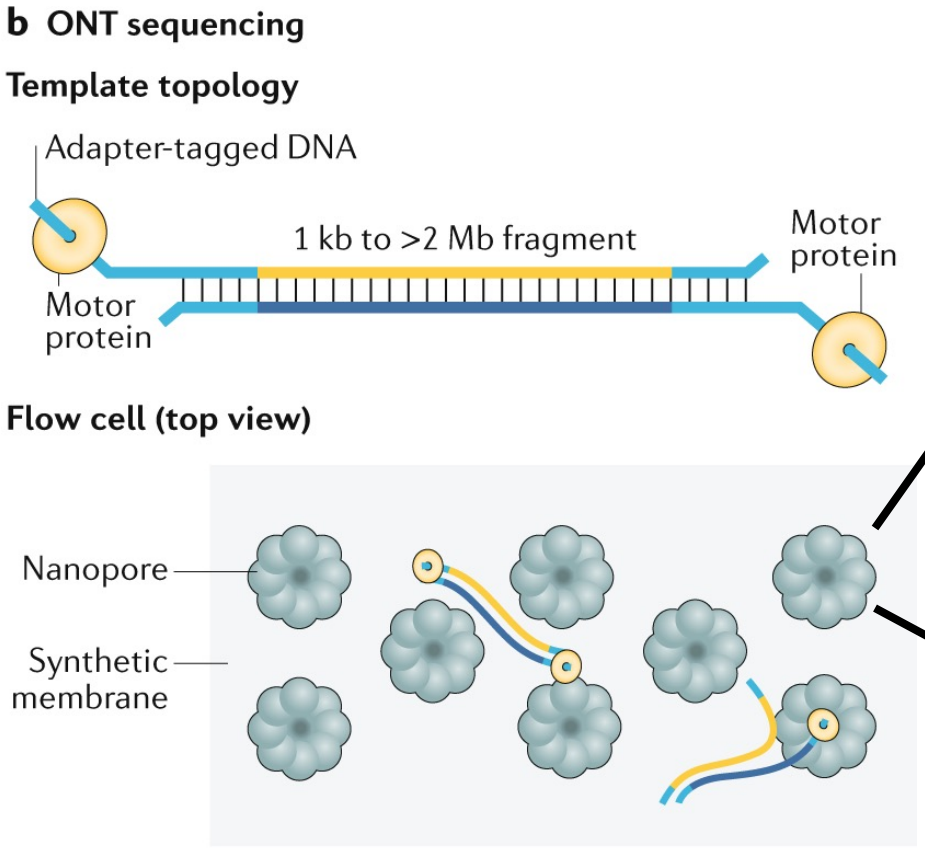
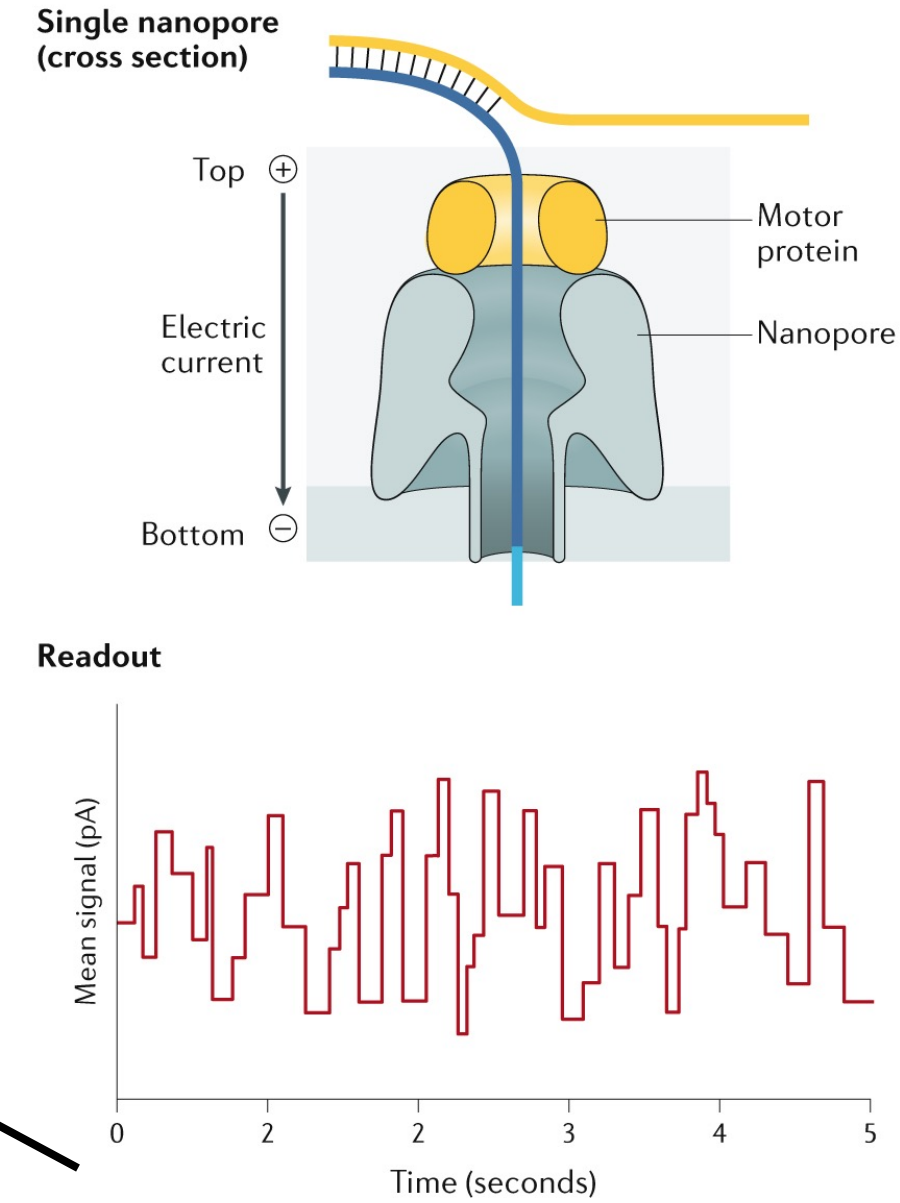


Figure from Goodwin et al.

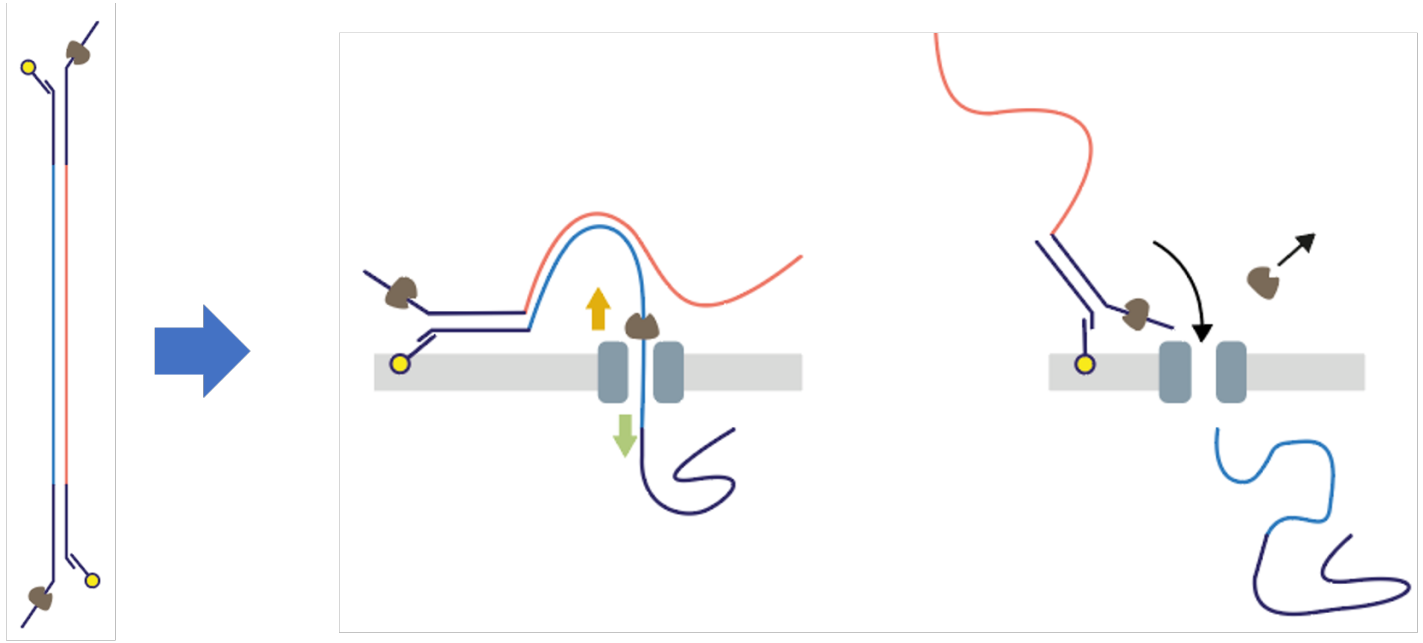
Oxford Nanopore - ONT





Oxford Nanopore - ONT

- Duplex sequencing – Passing both DNA strands through the pore -> Double accuracy



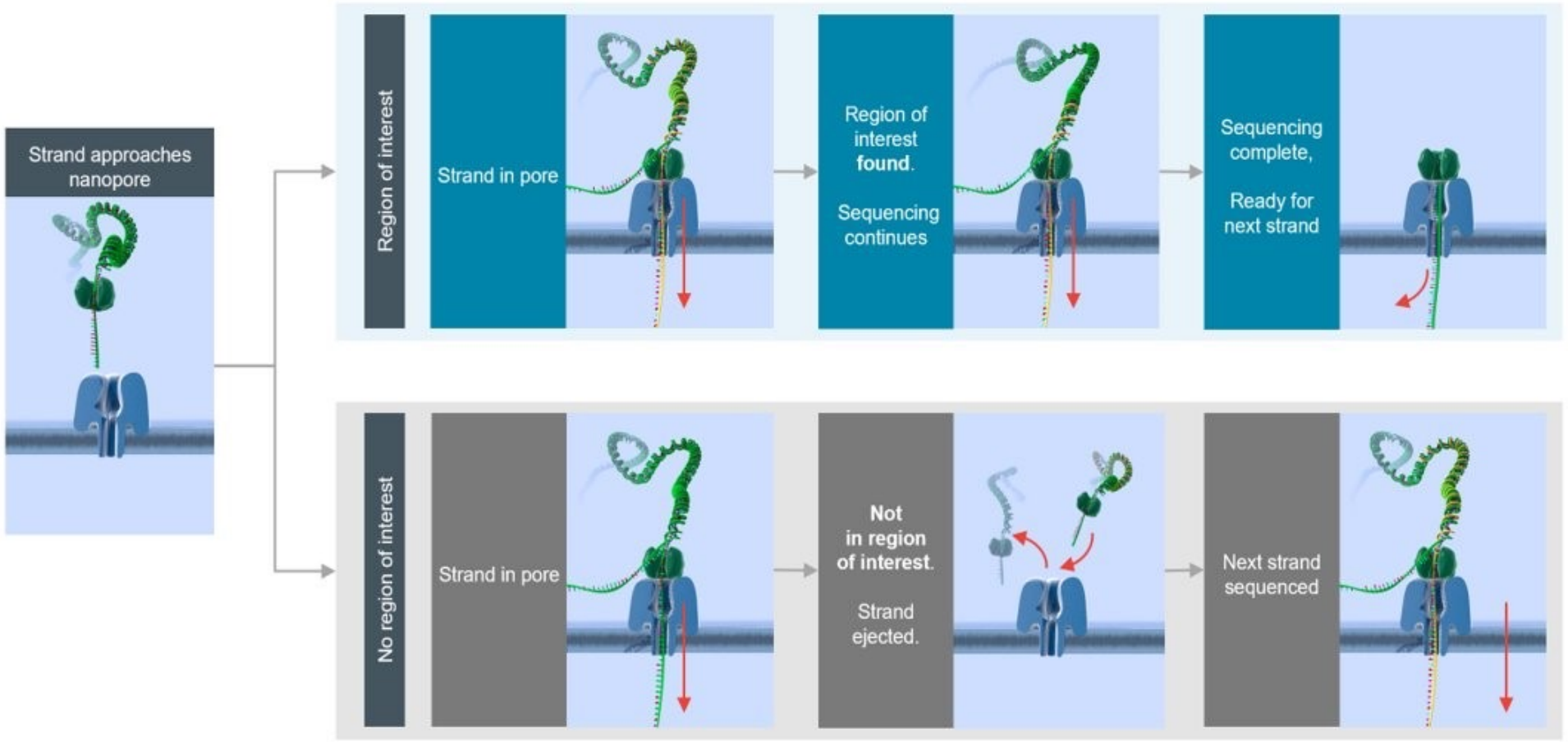
Linear dsDNA molecule adapted on both ends

First strand sequenced

Second strand captured and sequenced immediately after

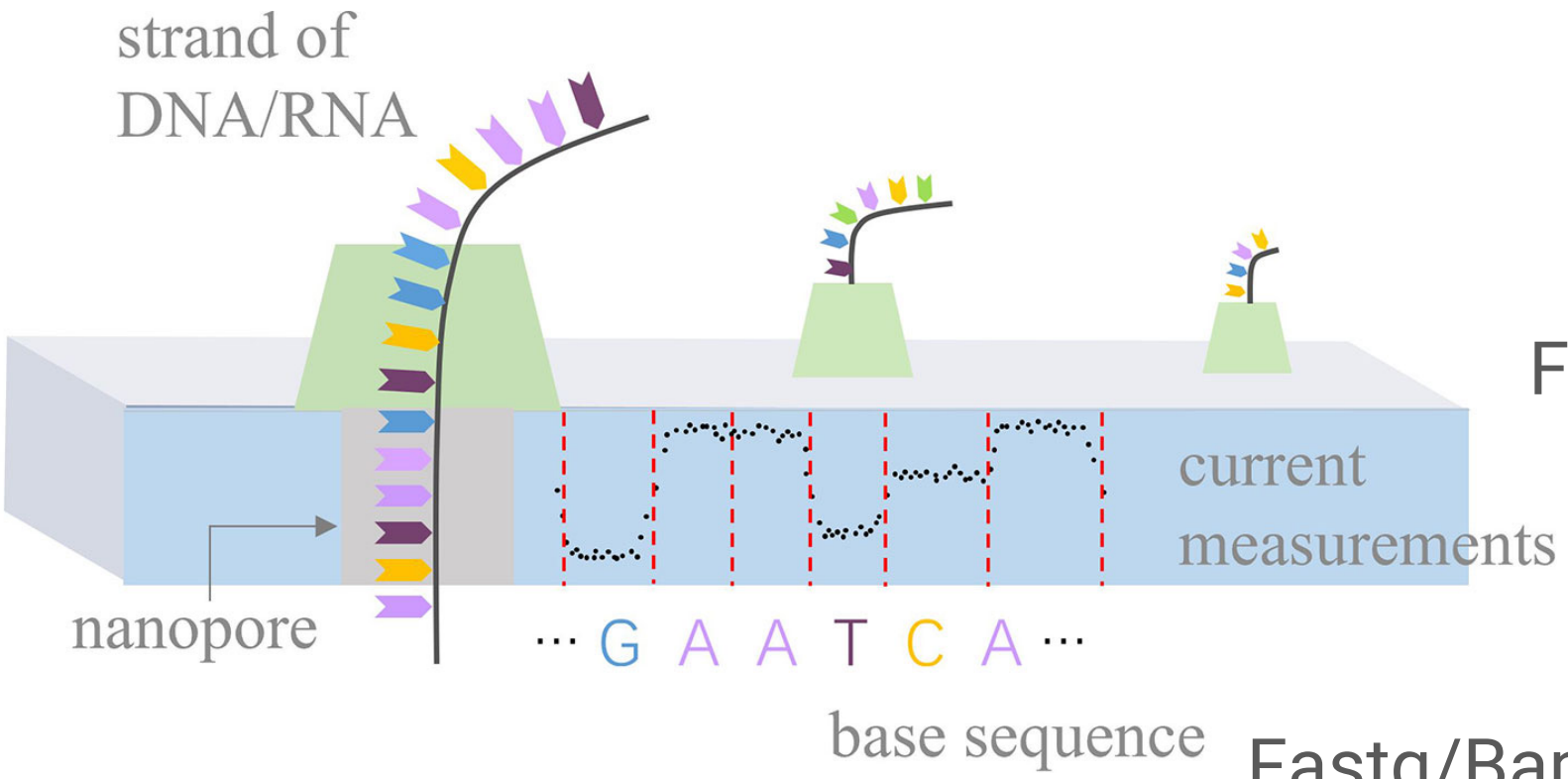
Oxford Nanopore - ONT

- Adaptive sequencing – Depleting or enriching sequences on the fly



Oxford Nanopore - ONT

- Nanopore raw current files – Pod5

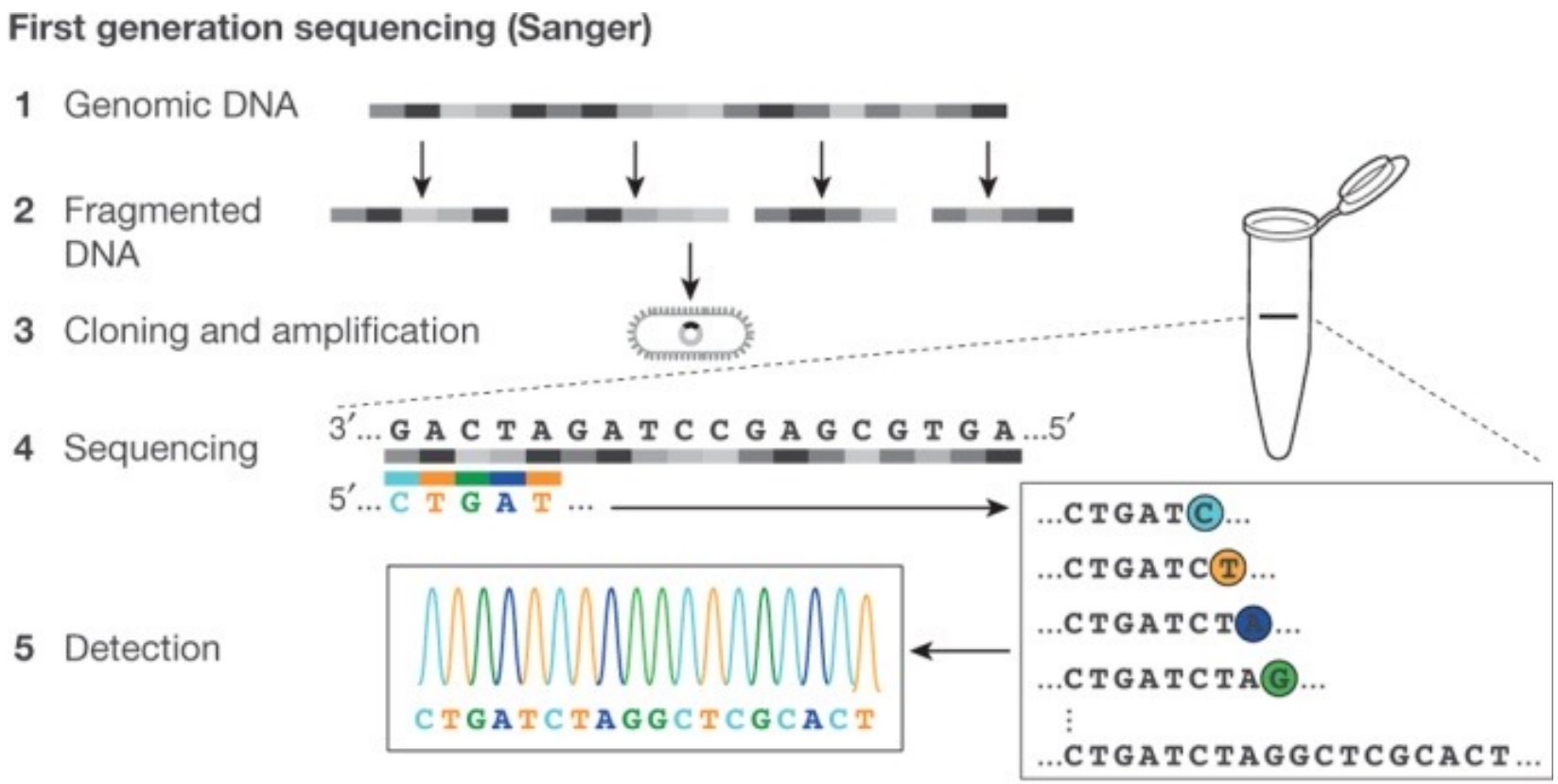


Fast5/Pod5 file format

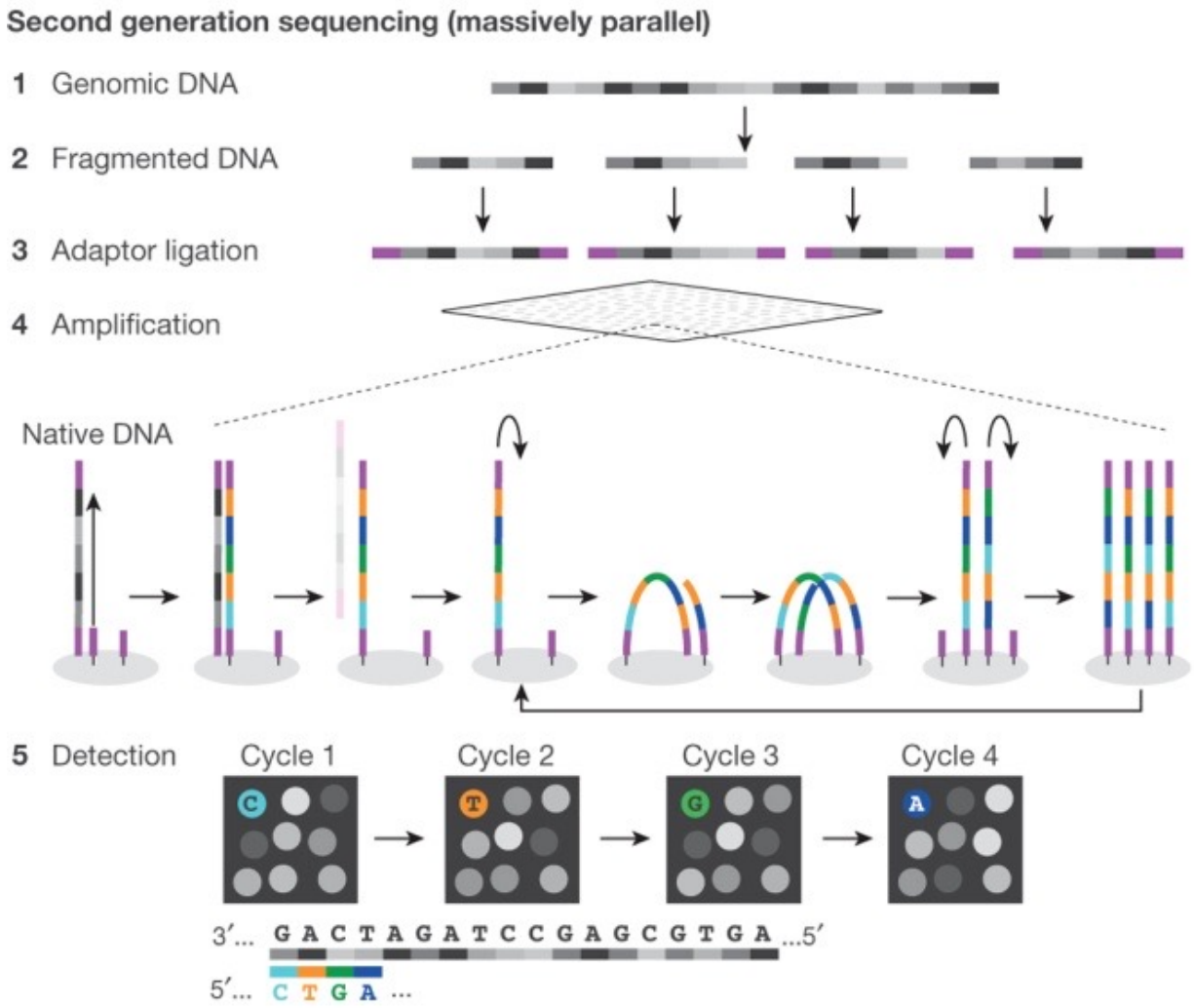
Basecalling

Fastq/Bam format

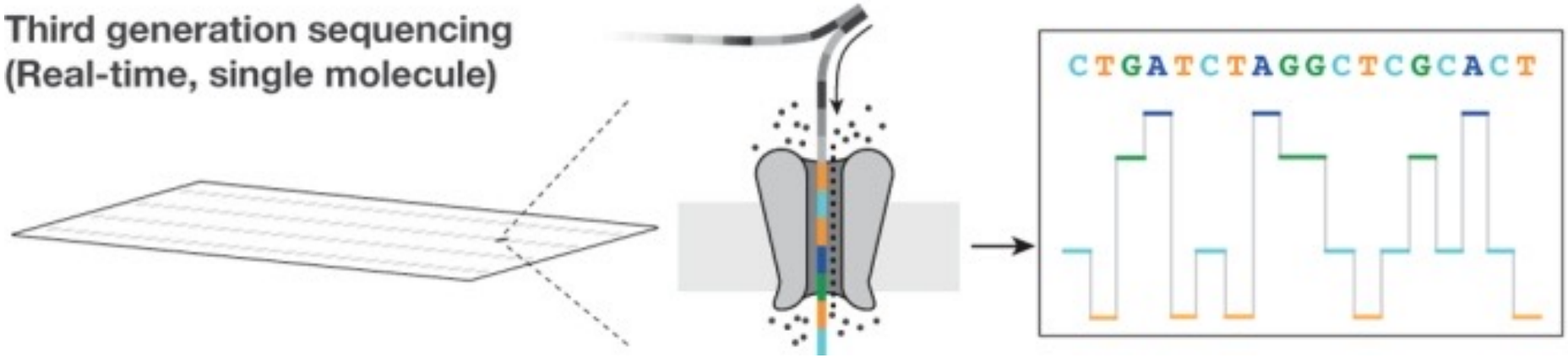
First Generation Sequencing



Second Generation Sequencing



Third Generation Sequencing



Review Sequencing Technologies

Generation	Sequencing Technology	Year	Company	Average Read Length	Cost per Gigabase
First Generation Sequencing	Sanger Sequencing	1977	Frederick Sanger	~800 bp	Very High (>$1000)
Second-Generation Sequencing	454 Sequencing	2005	Roche	~400 bp	High (100−500)
	Illumina Sequencing (Solexa/Solexa II)	2006	Illumina	~150-300 bp	Low (1−10)
	Ion Torrent Sequencing	2010	Thermo Fisher Scientific	~200 bp	Medium (10−50)
Third-Generation Sequencing	PacBio SMRT	2009	Pacific Biosciences	10,000+ bp	High (50−200)
	Oxford Nanopore	2014	Oxford Nanopore Technologies	10,000+ bp	Medium-High (10−100)

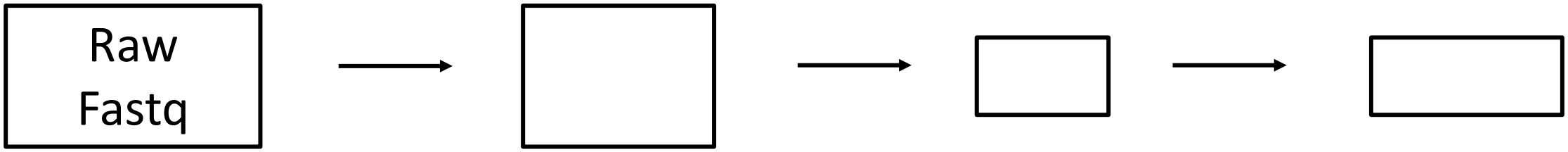
Assignment 1

We will compare the characteristics and SNP calls for three sequencing technologies:

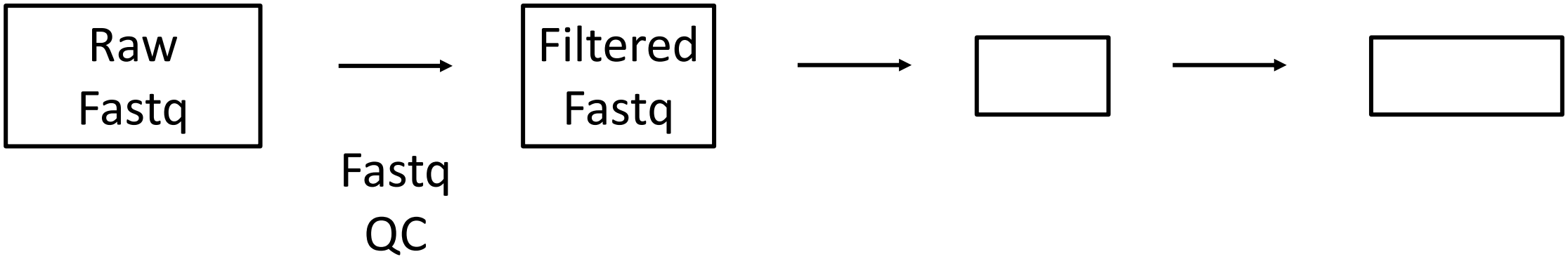
1. Illumina
2. PacBio
3. Nanopore

The benchmarking will be done using the well-characterized HG002

Assignment 1

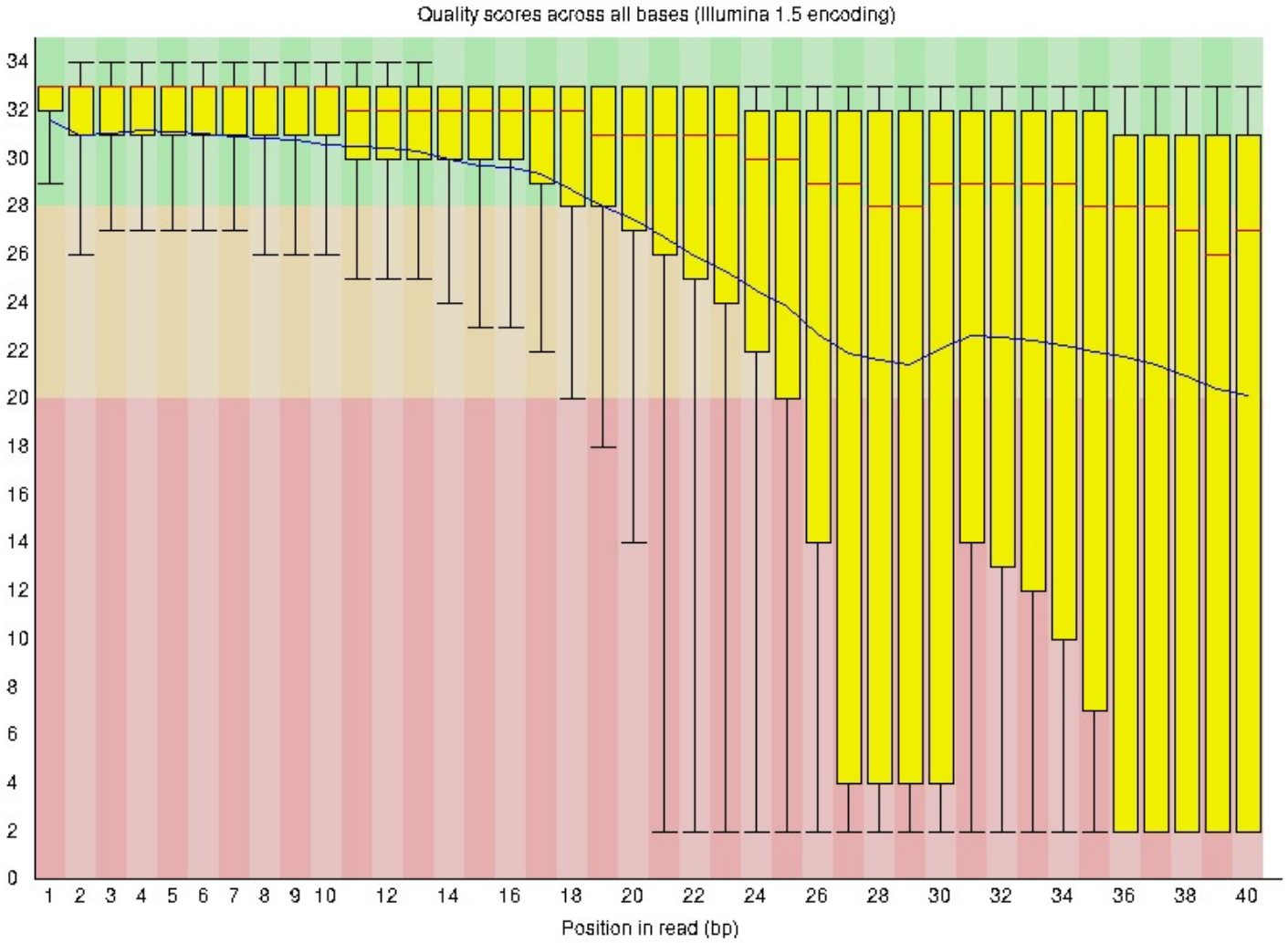


Assignment 1



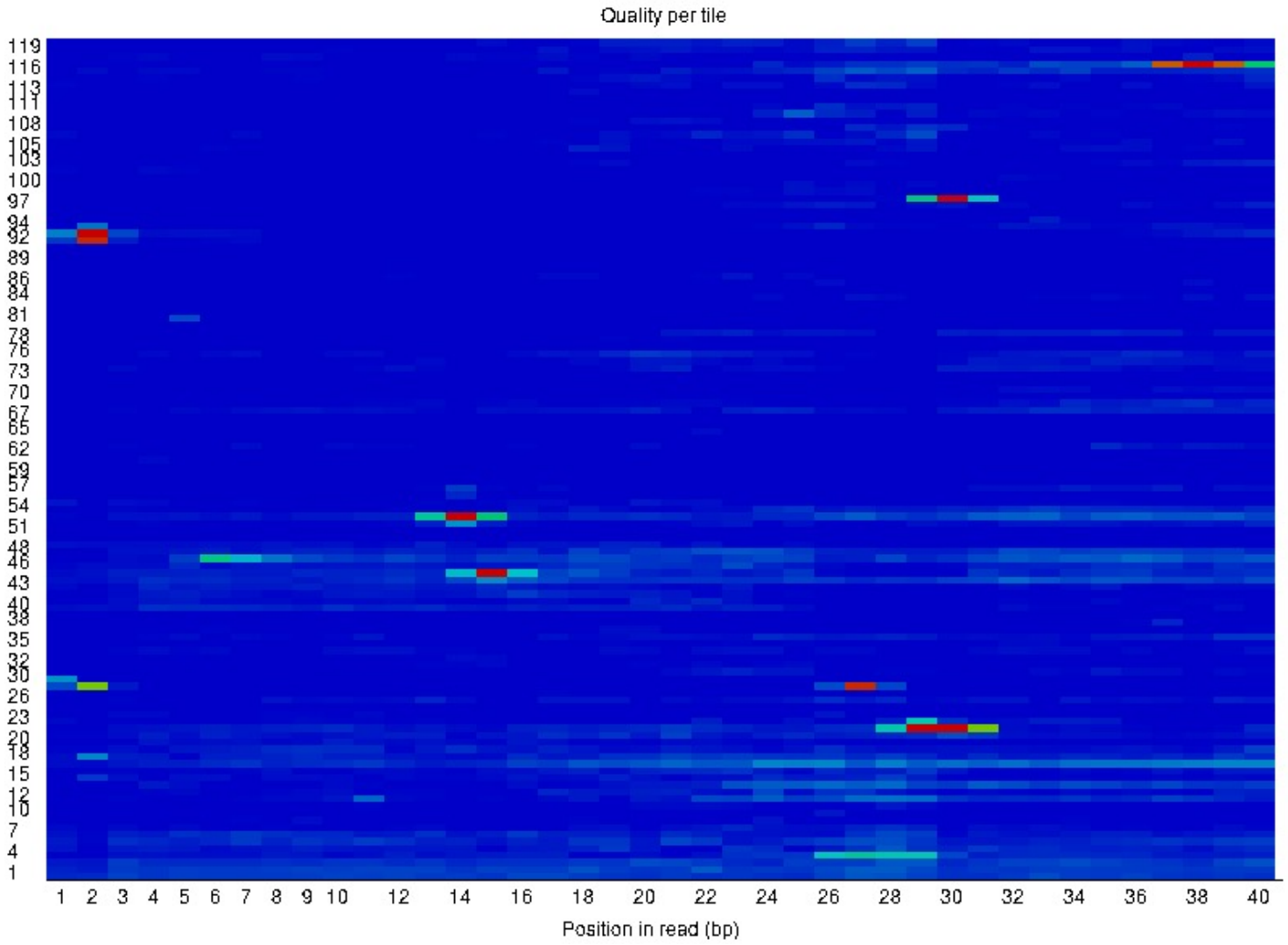
Quality check

Per base sequence quality



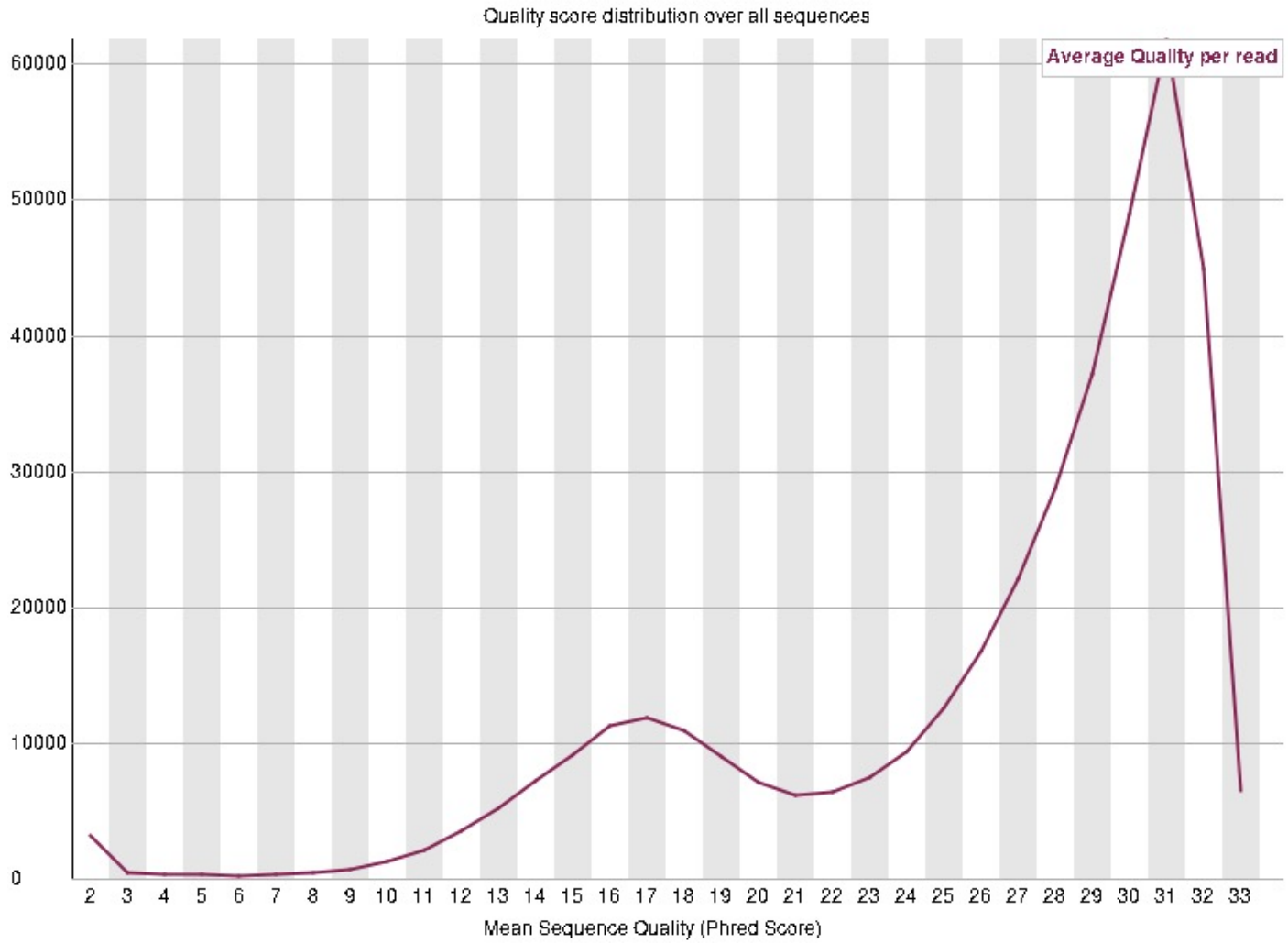
Quality check

Per tile sequence quality



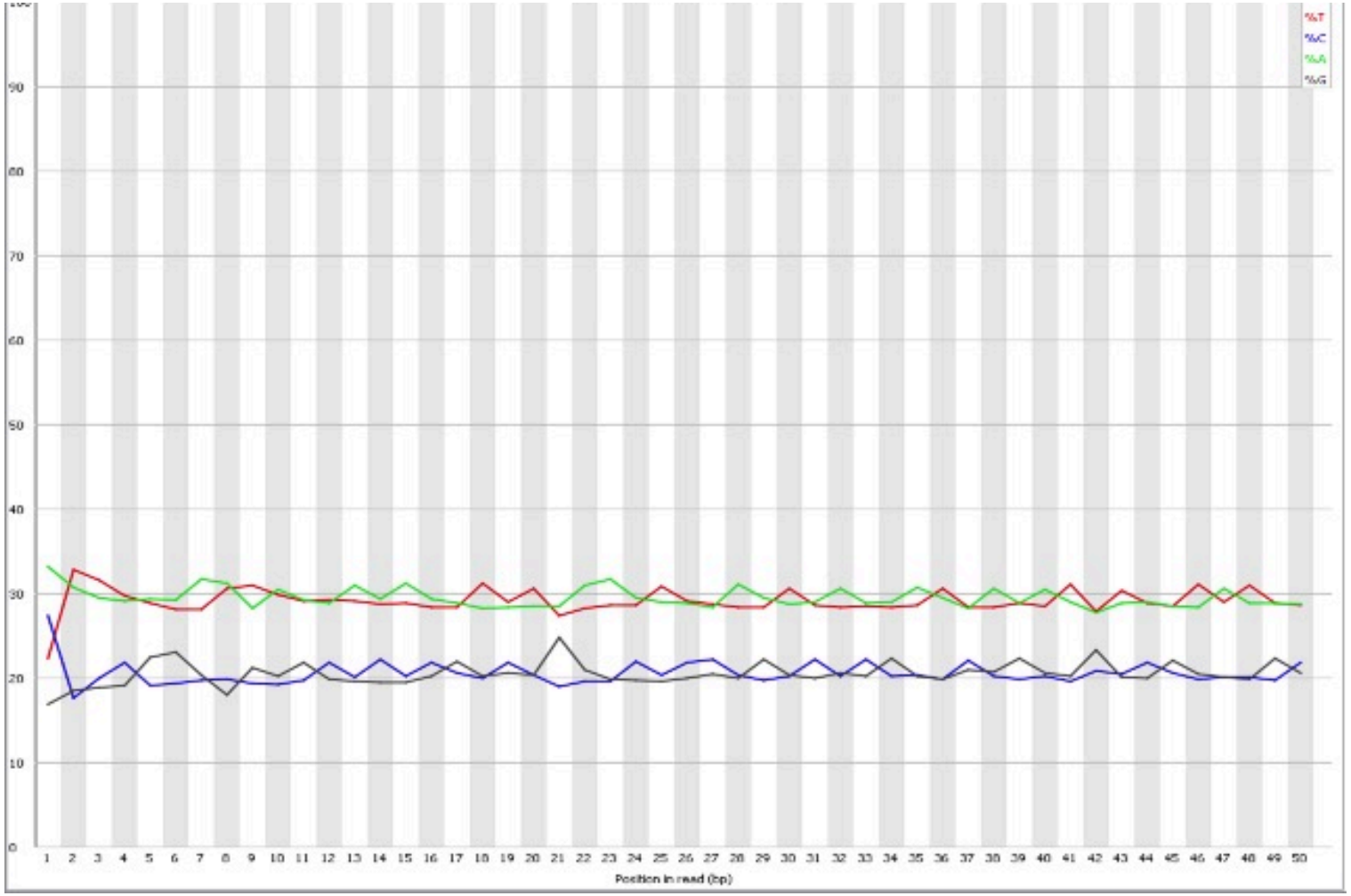
Quality check

Per sequence quality scores



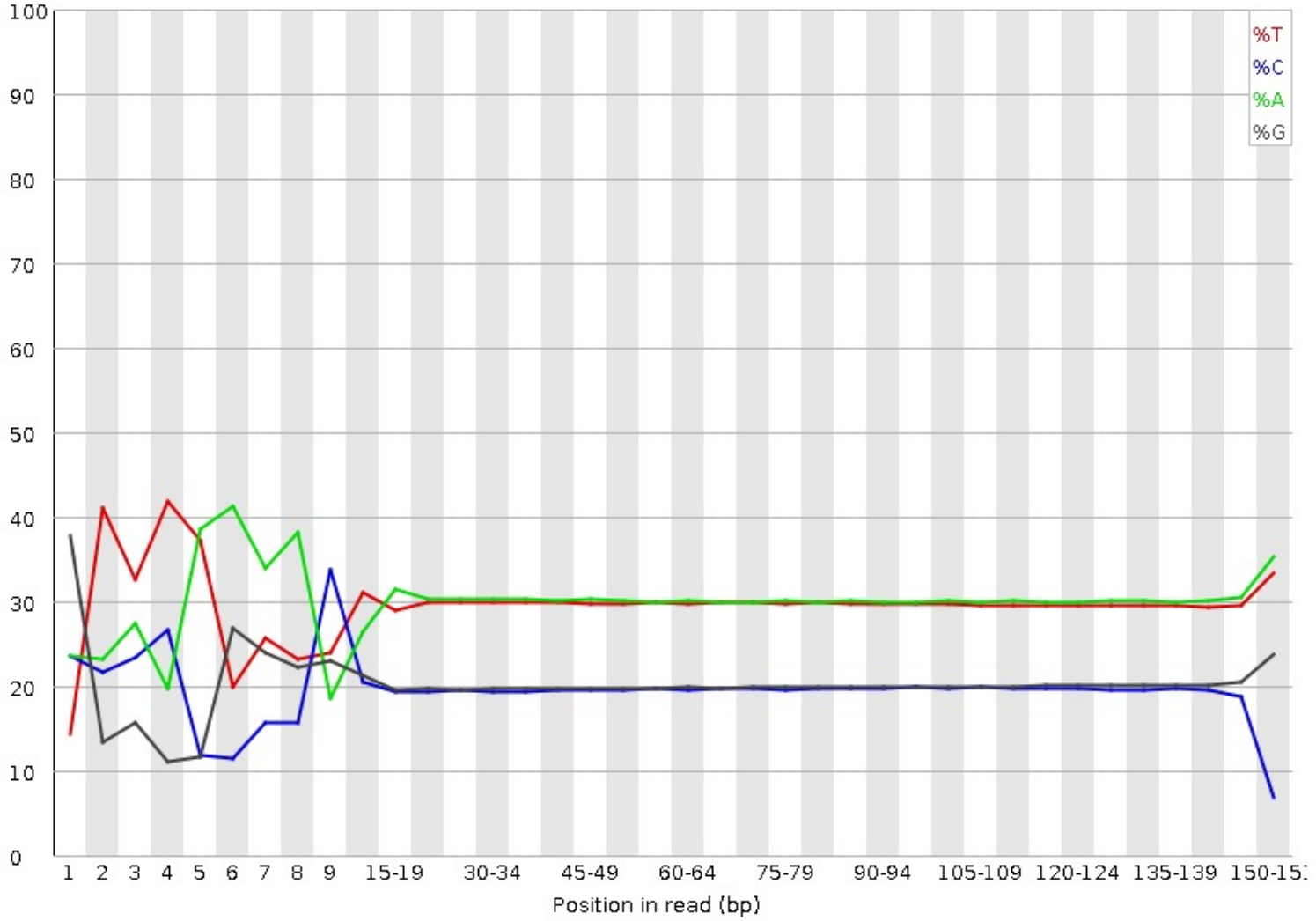
Quality check

Per base sequence content



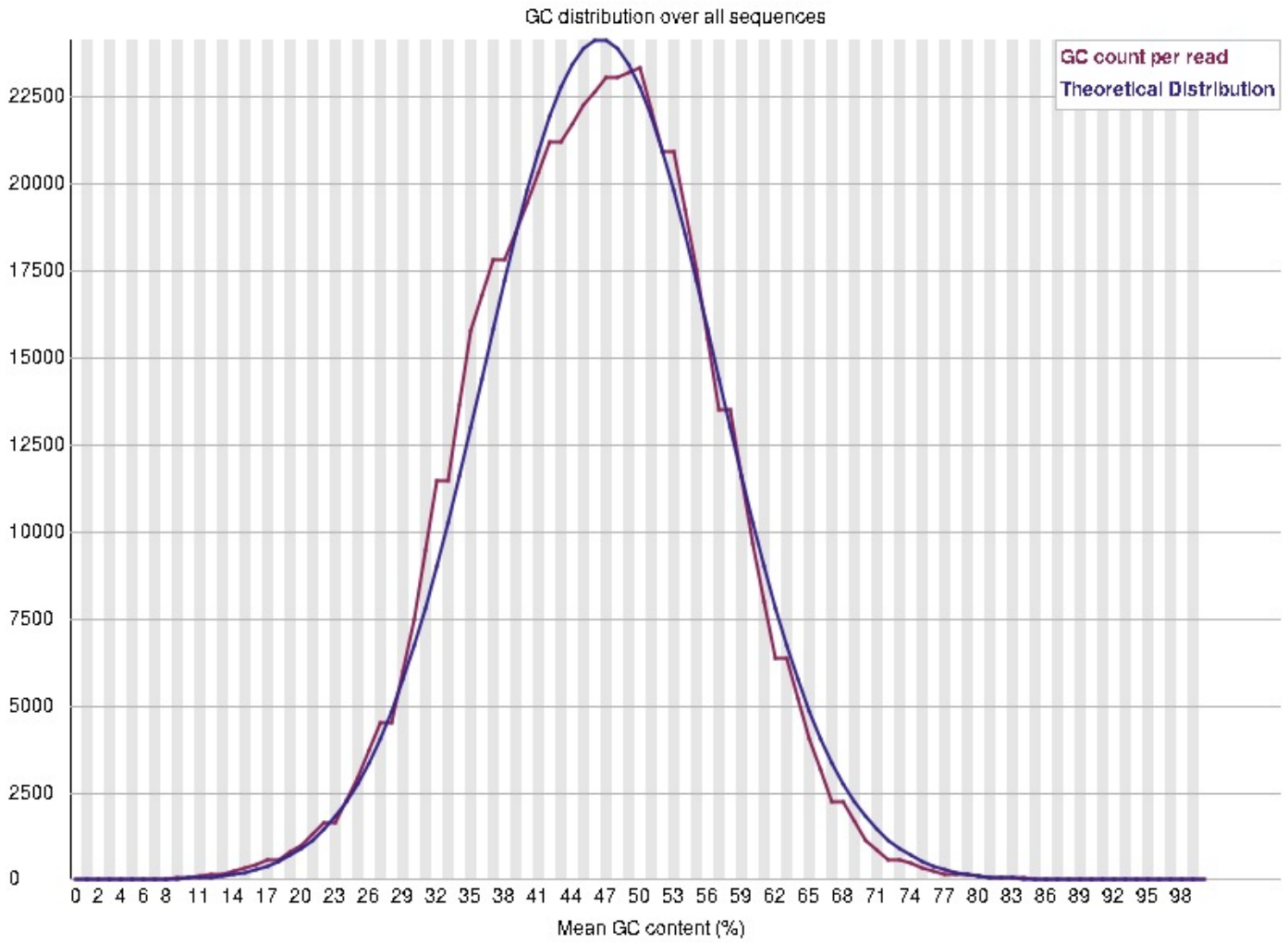
Quality check

Per base sequence content



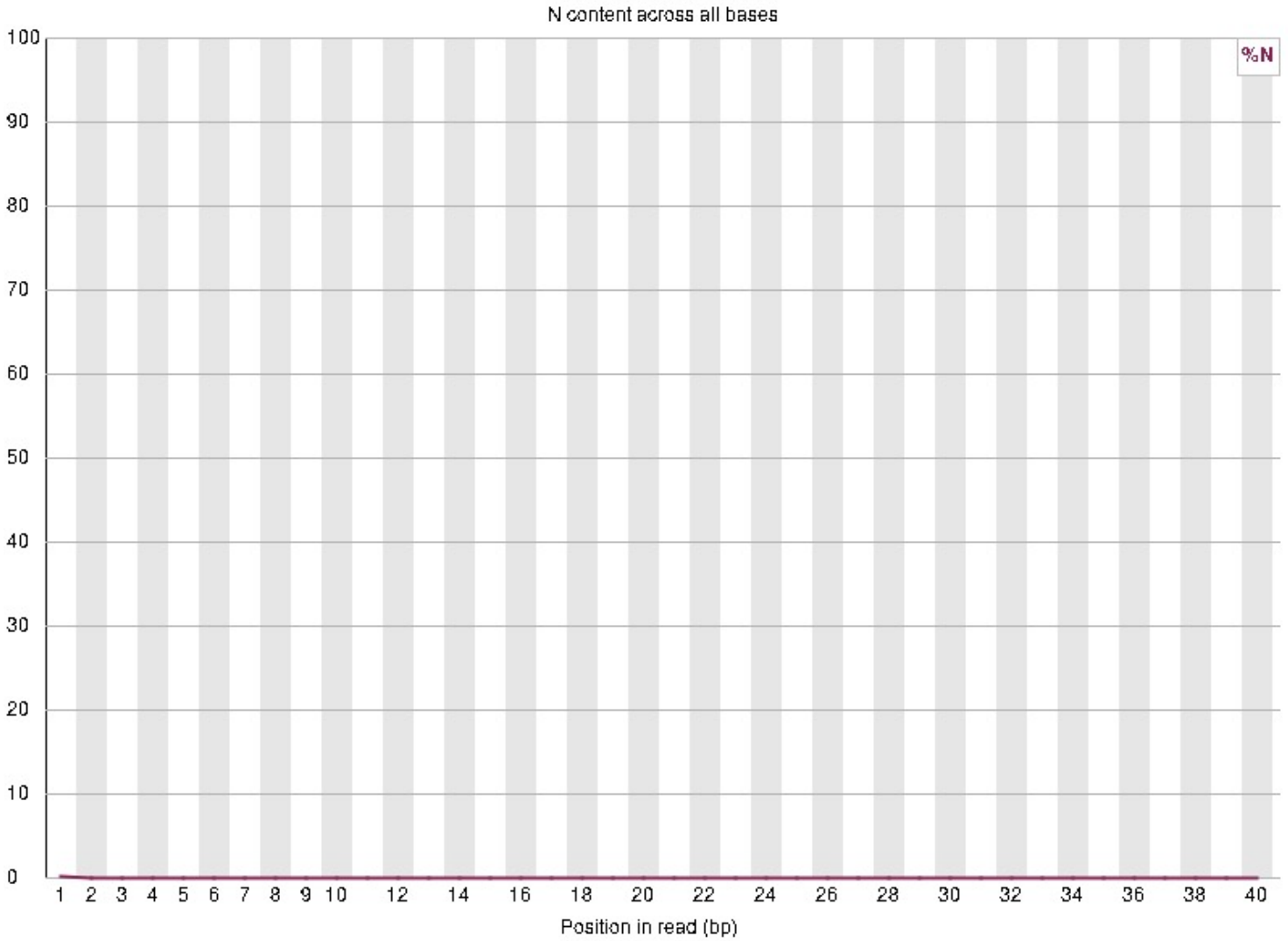
Quality check

Per sequence GC content



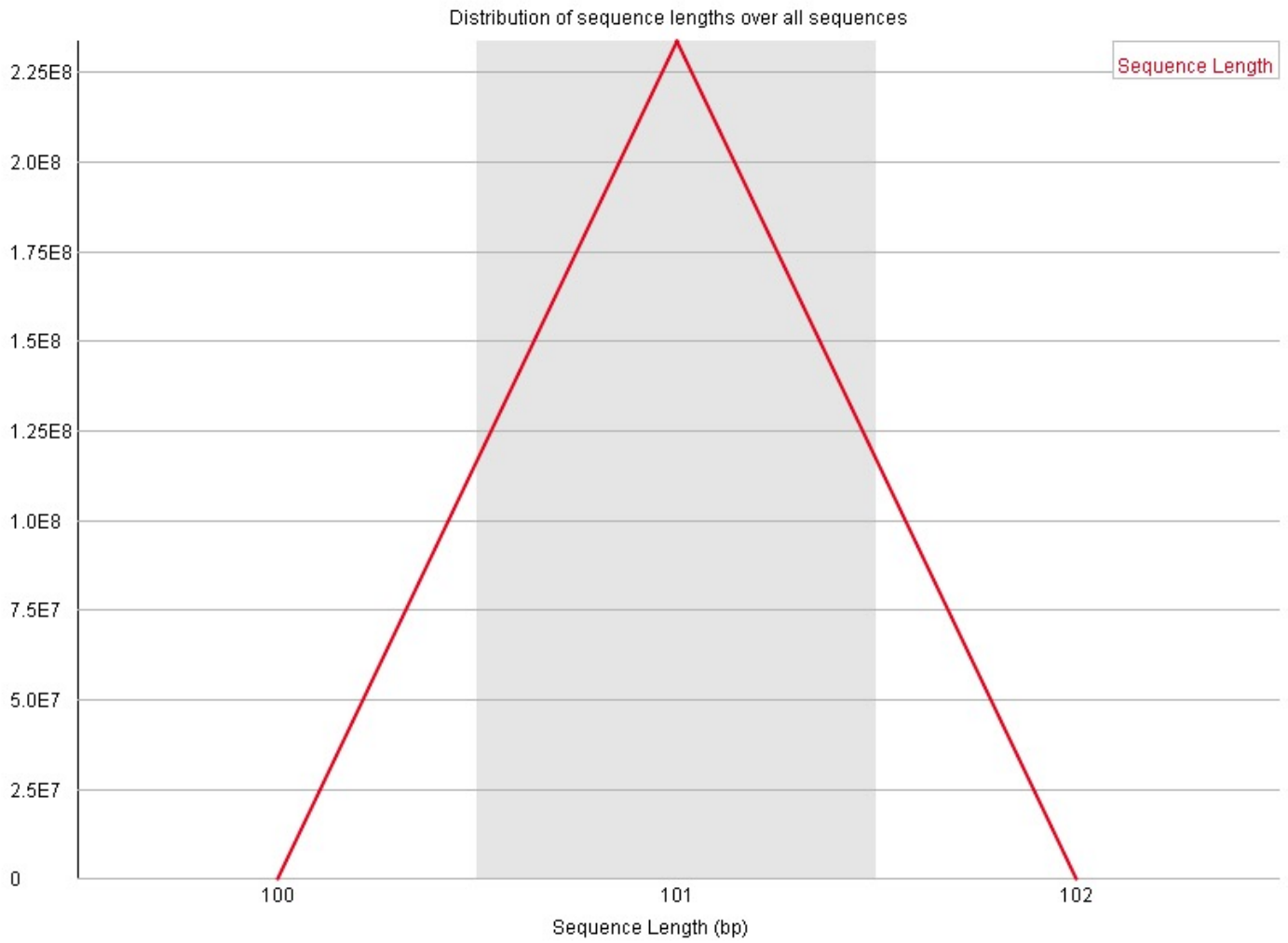
Quality check

Per base N content



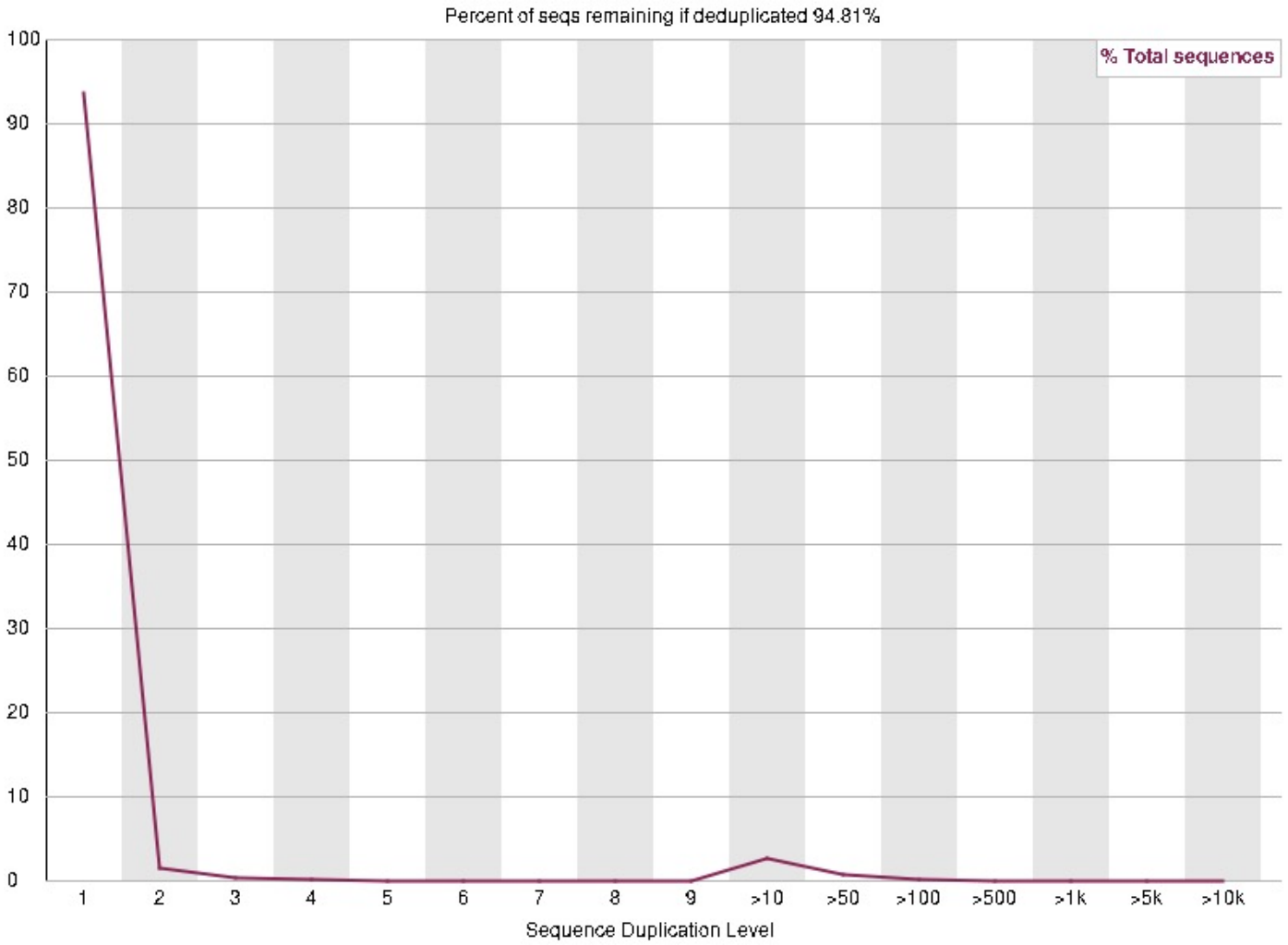
Quality check

Sequence Length Distribution



Quality check

Sequence Duplication Levels



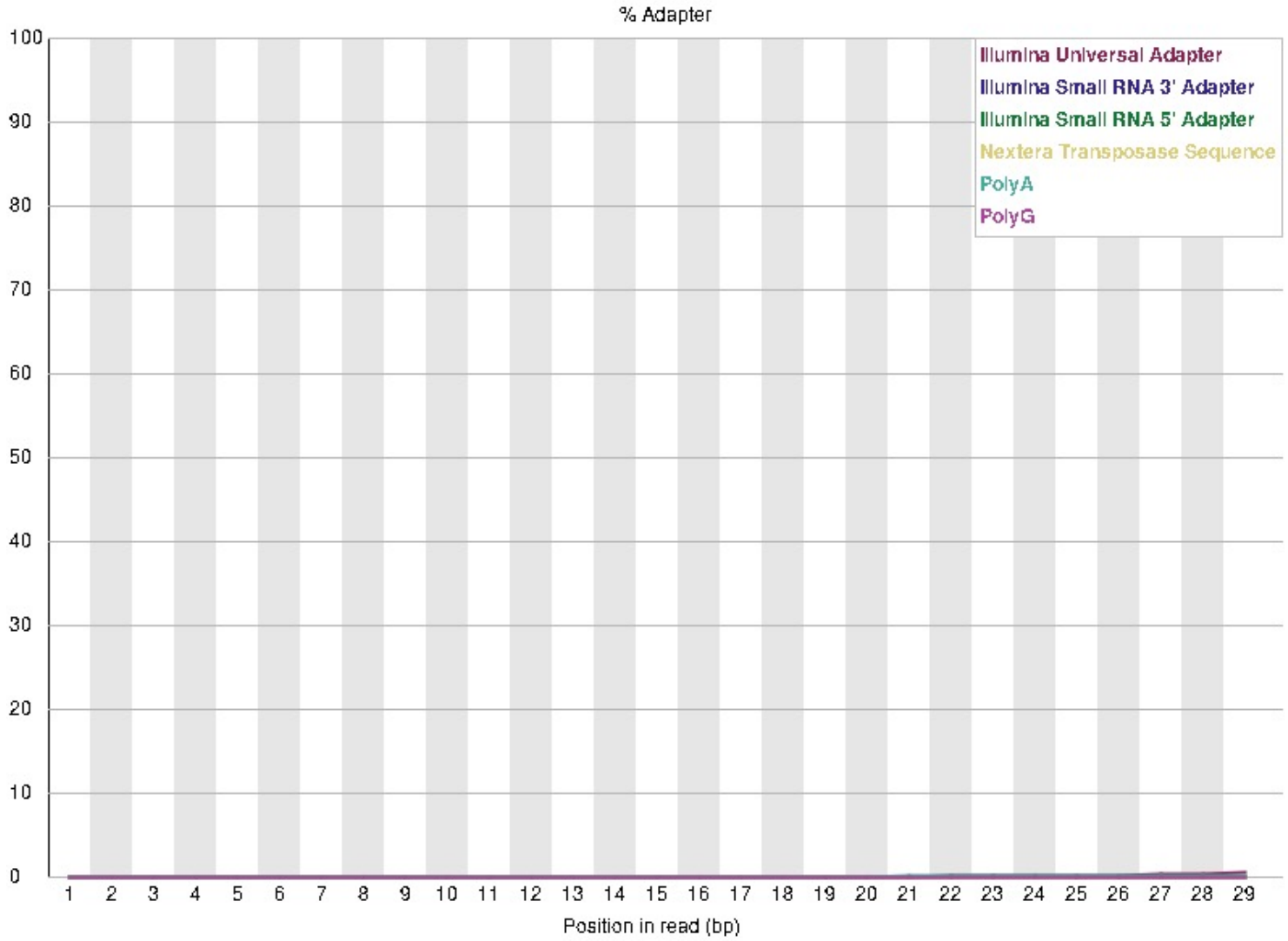
Quality check

Overrepresented sequences

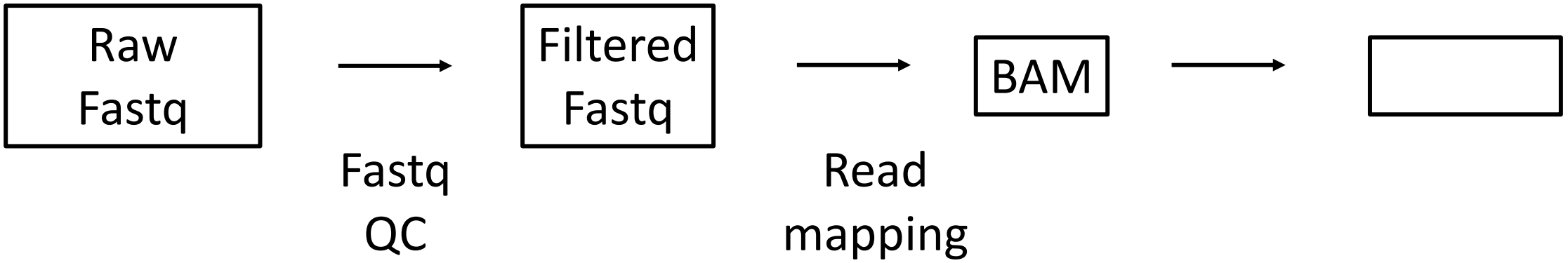
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End Sequencing Primer 2 (100% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End Sequencing Primer 2 (100% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End Sequencing Primer 2 (100% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Quality check

Adapter Content



Assignment 1



Mapping and variant calling

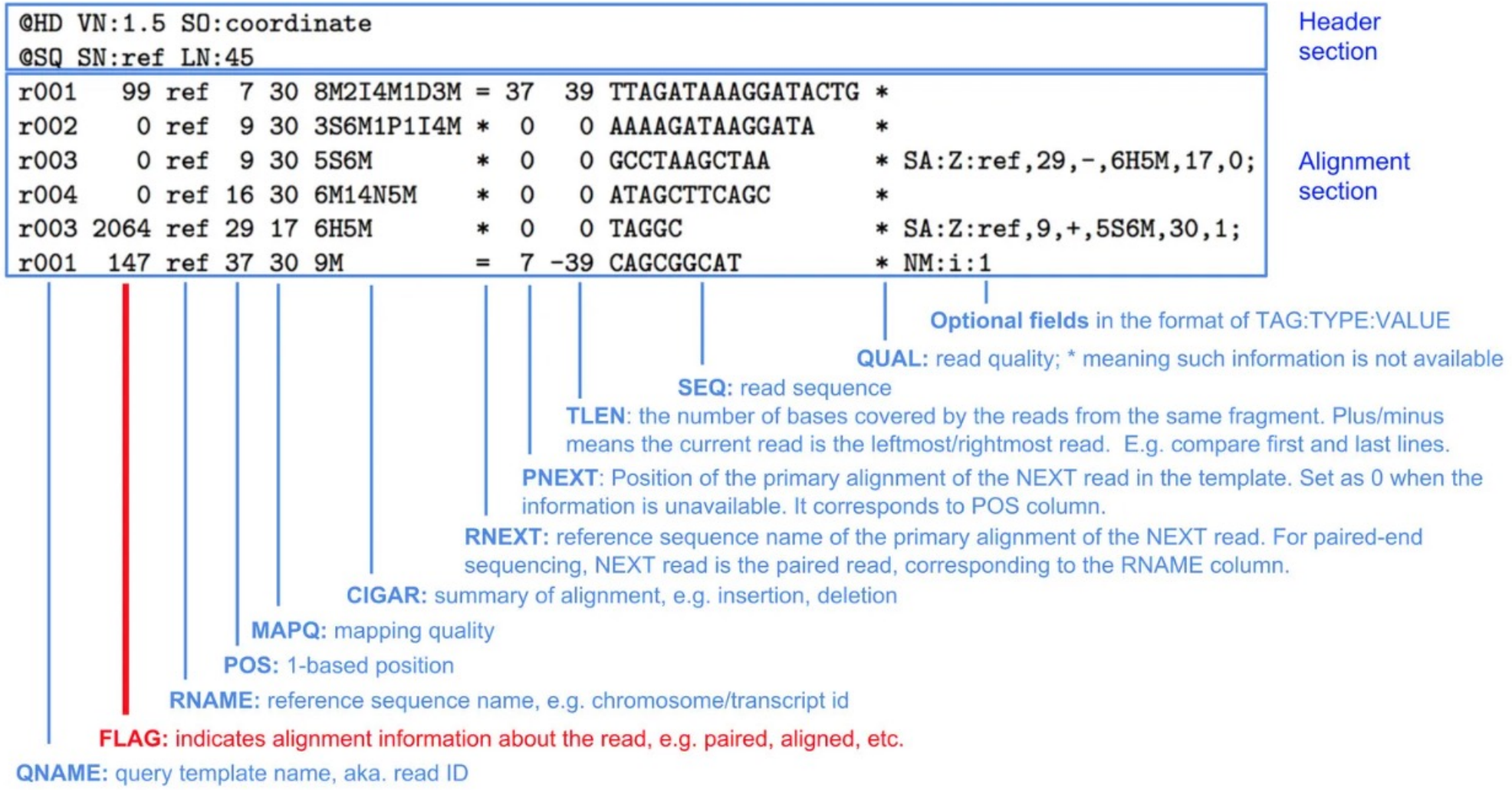
- BWA
- SAMtools
- BCFtools

wellcome sanger institute

The SAM/BAM format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[12]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHP=X])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

The SAM/BAM format

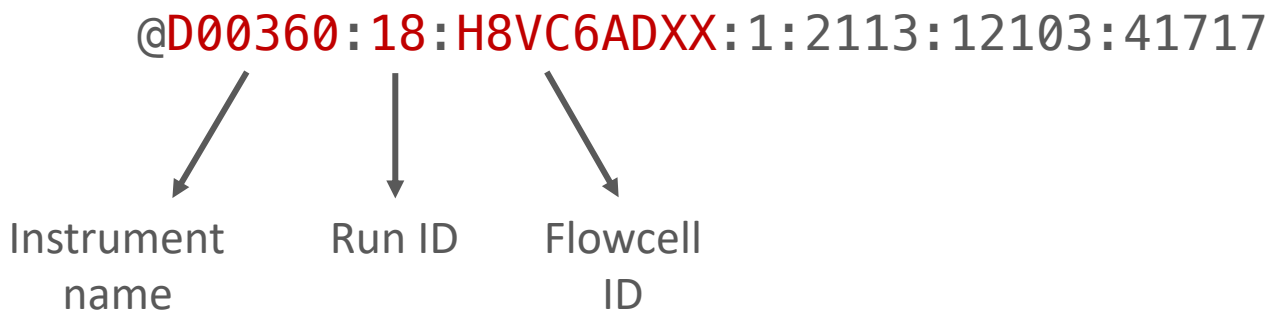


The SAM/BAM format

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Mapping – Read Groups

- Read groups – Provide technical information about flowcell and multiplexing of illumina reads

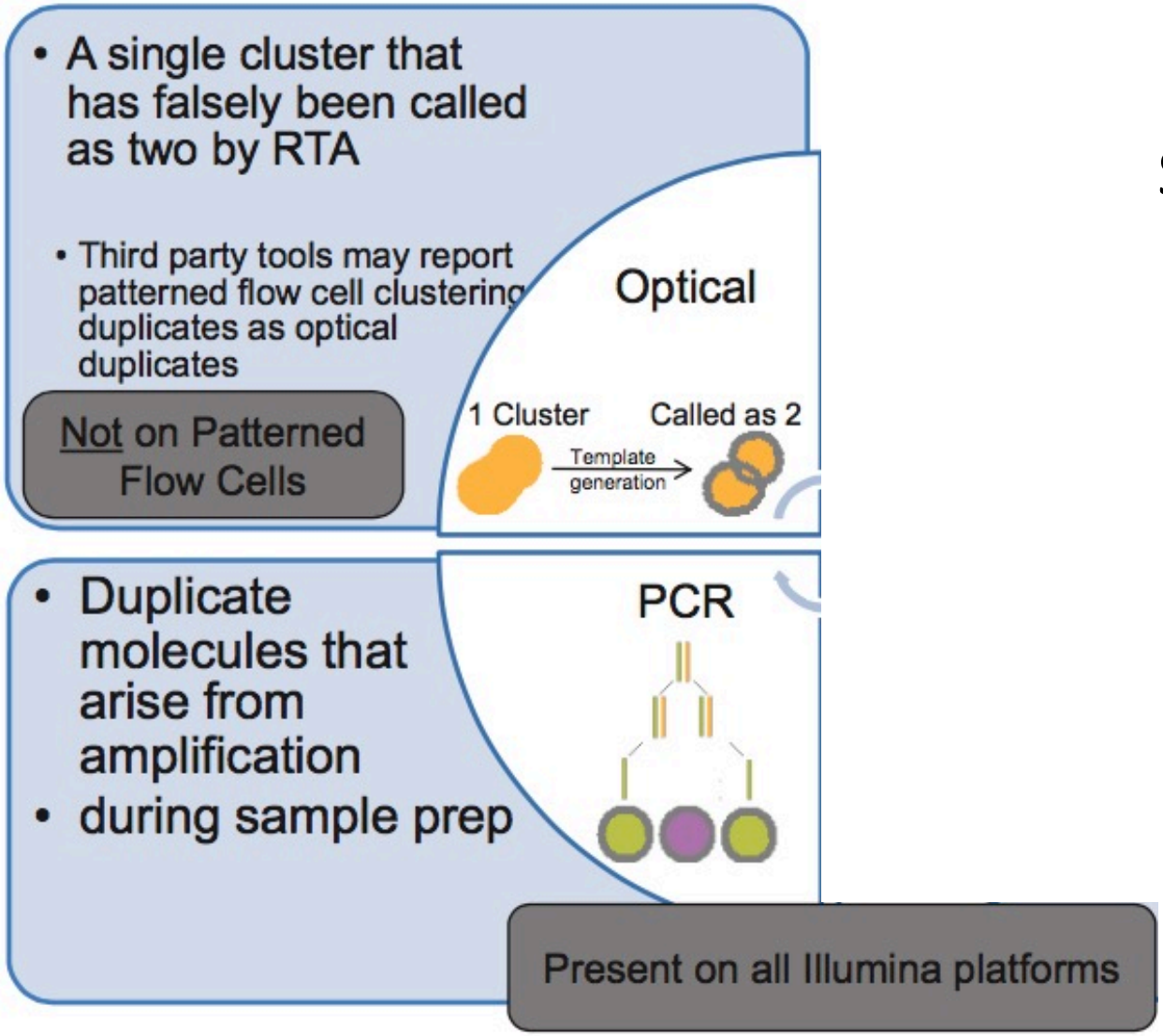


ID = {FLOWCELL_BARCODE}.{LANE}

PU = {FLOWCELL_BARCODE}.{LANE}.{SAMPLE_BARCODE}

```
bwa mem -R "@RG\tID:H8VC6ADXX.1\tH8VC6ADXX.1.sample1\tPL:HiSeq\tSM:Sample" reference fwd rev
```

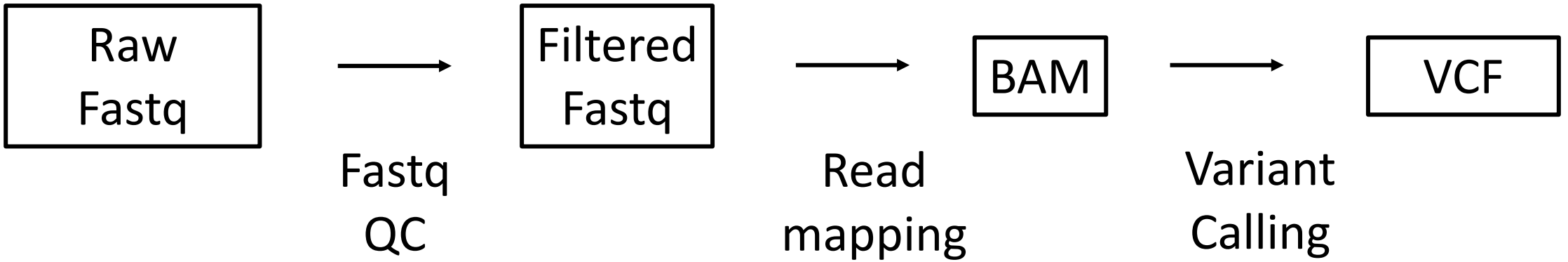
Marking duplicates



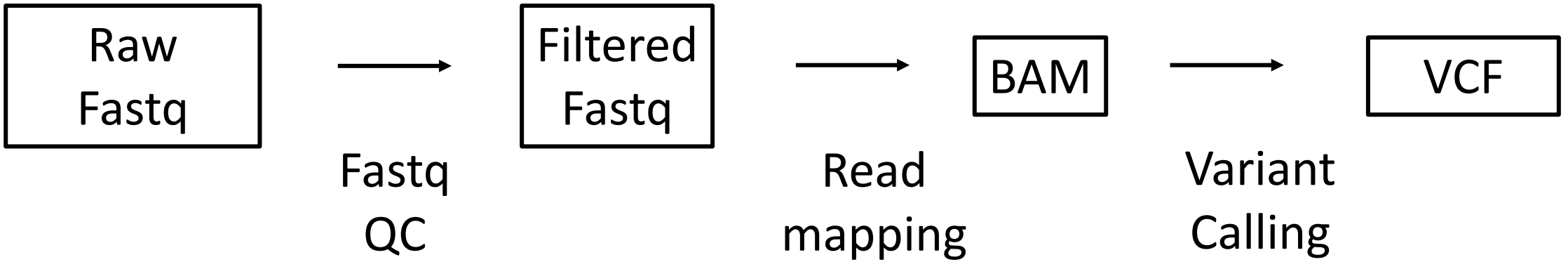
Software:

- Picard
- sambamba

Assignment 1



Assignment 1



Assignment 1

Final output

Sample VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	1	.	T	G	60	PASS	…	GT	1/1
chr1	2	.	T	.	60	PASS	…	GT	0/0
chr1	3	.	G	C	60	PASS	…	GT	1/1
chr1	4	.	A	G	60	FAIL	…	GT	1/1
chr1	5	.	C	A	60	PASS	…	GT	1/1

Benchmark VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	1	.	T	G	60	PASS	…	GT	1/1
chr1	2	.	T	A	60	PASS	…	GT	1/1
chr1	3	.	G	C	60	PASS	…	GT	0/1
chr1	4	.	A	G	60	PASS	…	GT	1/1
chr1	5	.	C	T	60	PASS	…	GT	1/1

Assignment 1

Final output

True Positive

Sample VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	1	.	T	G	60	PASS	…	GT	1/1
chr1	2	.	T	.	60	PASS	…	GT	0/0
chr1	3	.	G	C	60	PASS	…	GT	1/1
chr1	4	.	A	G	60	FAIL	…	GT	1/1
chr1	5	.	C	A	60	PASS	…	GT	1/1
chr1	6	.	T	C	60	PASS	…	GT	1/1
chr1	7	.	T	.	60	PASS	…	GT	0/0

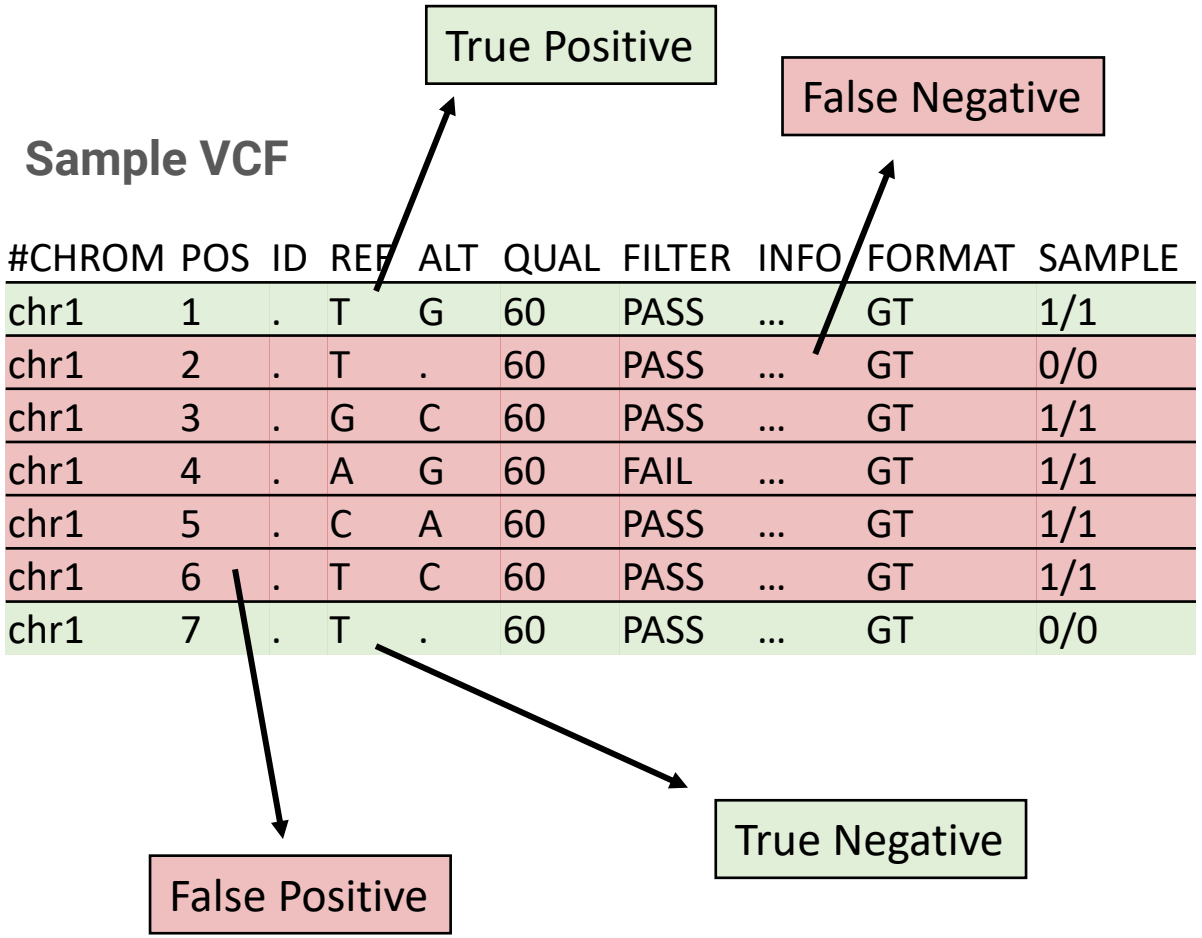
True Negative

Benchmark VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	1	.	T	G	60	PASS	…	GT	1/1
chr1	2	.	T	A	60	PASS	…	GT	1/1
chr1	3	.	G	C	60	PASS	…	GT	0/1
chr1	4	.	A	G	60	PASS	…	GT	1/1
chr1	5	.	C	T	60	PASS	…	GT	1/1
chr1	6	.	T	.	60	PASS	…	GT	0/0
chr1	7	.	T	.	60	PASS	…	GT	0/0

Assignment 1

Final output



Benchmark VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr1	1	.	T	G	60	PASS	…	GT	1/1
chr1	2	.	T	A	60	PASS	…	GT	1/1
chr1	3	.	G	C	60	PASS	…	GT	0/1
chr1	4	.	A	G	60	PASS	…	GT	1/1
chr1	5	.	C	T	60	PASS	…	GT	1/1
chr1	6	.	T	.	60	PASS	…	GT	0/0
chr1	7	.	T	.	60	PASS	…	GT	0/0

Assignment 1

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Assignment 1

Benchmarking

- F1 score $$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

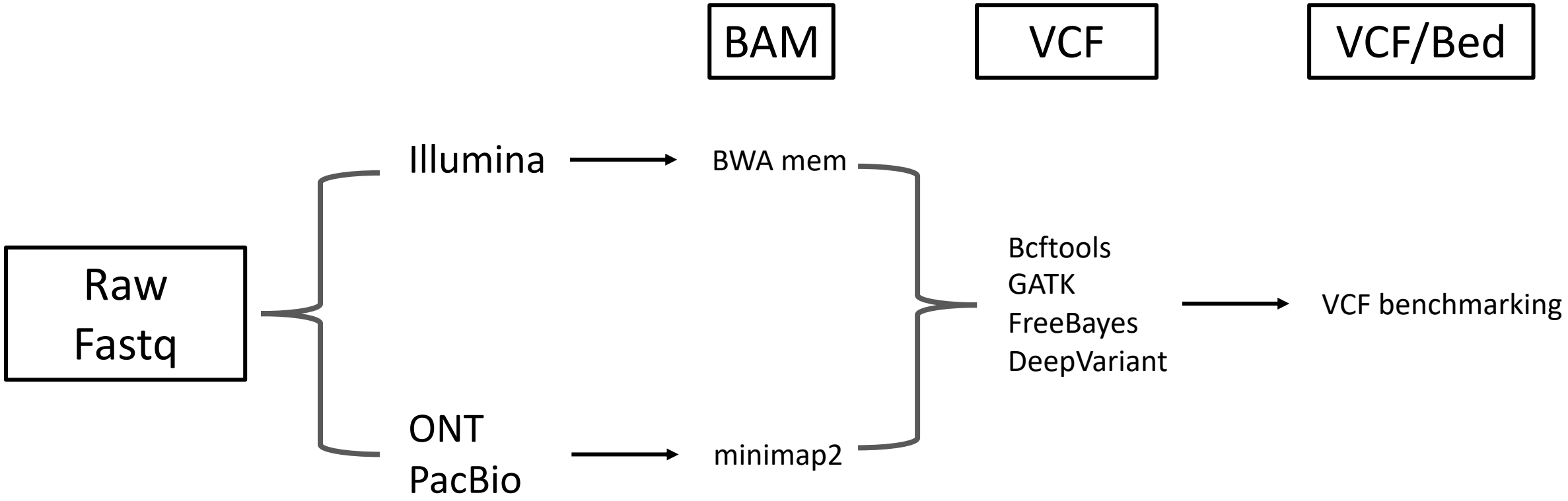
Assignment 1

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

True Positive Rate (TPR)

Specificity = 1 - FPR

Assignment 1



Assignment 1

```
##fileformat=VCFv4.3
##contig=<ID=chr1,length=249250621>
##INFO=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

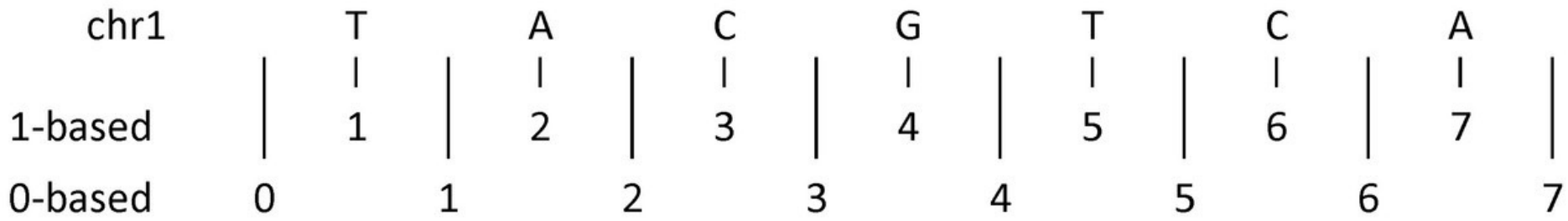
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample
chr1	100	rs123	A	G	30	PASS	…	GT	0/1
chr1	200	rs456	C	T	40	PASS	…	GT	1/1
chr1	300	.	G	A	20	PASS	…	GT	1/1
chr2	150	rs789	T	C	50	PASS	…	GT	1/1

VCF

chr1	99	100	rs123
chr1	199	200	rs456
chr1	299	300	.

BED

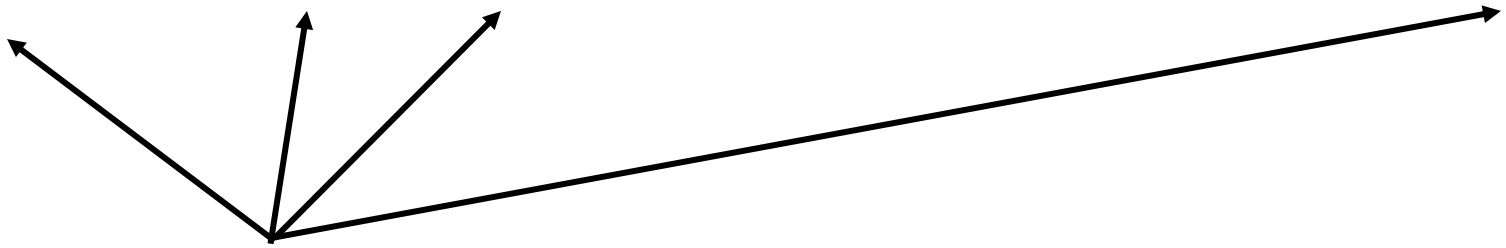
Assignment 1



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

Assignment 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Benchmark	T	C	C	A	G	C	C	C	T	C	A	G	C	G	T	C	A	T	G	C
Sample	T	C	C	T	G	C	A	C	G	C	A	G	C	G	T	C	A	T	C	C



False Positives

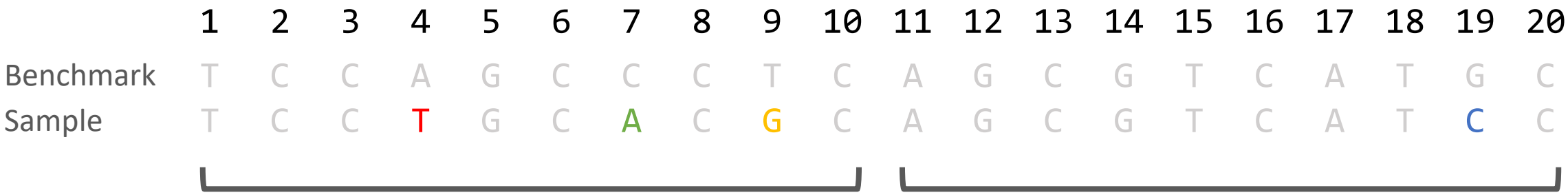
Assignment 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Benchmark	T	C	C	A	G	C	C	C	T	C	A	G	C	G	T	C	A	T	G	C
Sample	T	C	C	T	G	C	A	C	G	C	A	G	C	G	T	C	A	T	C	C

1. Get list of FP into bed file

Chr	3	4
Chr	6	7
Chr	8	9
Chr	18	19

Assignment 1



1. Get list of FP into bed file

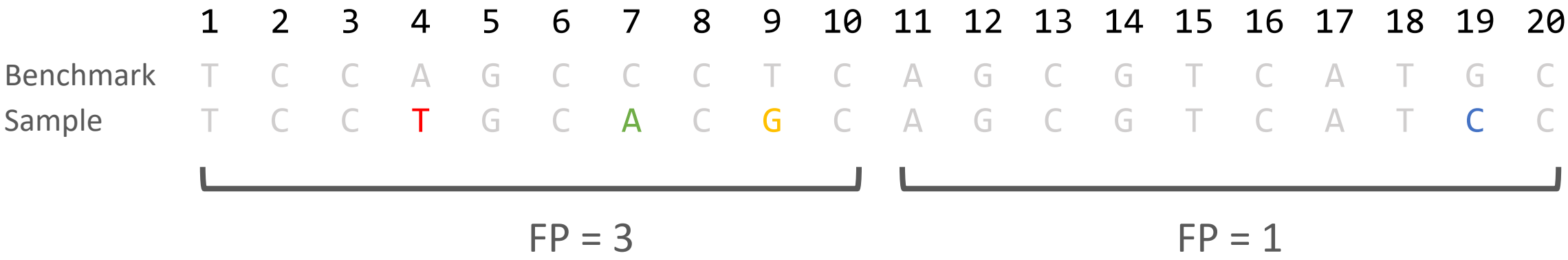
Chr	3	4
Chr	6	7
Chr	8	9
Chr	18	19

2. Make genomic windows for your chromosome

Chr 20 →

Chr	0	10
Chr	10	20

Assignment 1



1. Get list of FP into bed file

Chr	3	4
Chr	6	7
Chr	8	9
Chr	18	19

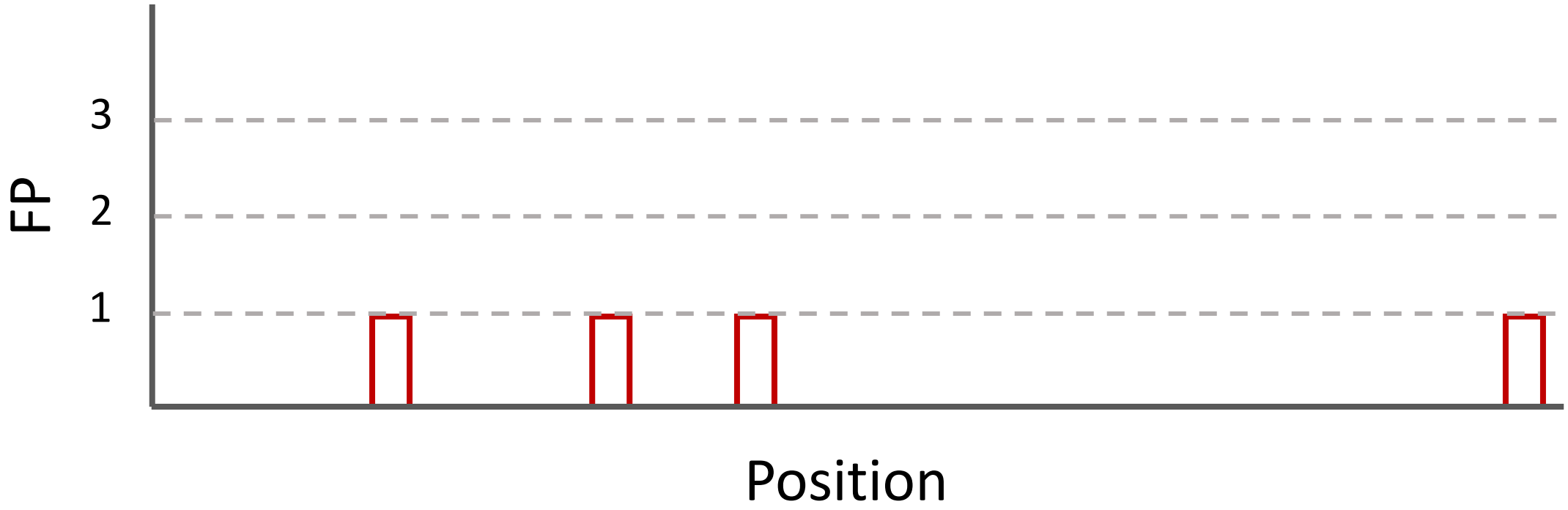
2. Make genomic windows for your chromosome

Chr	20	→	Chr	0	10
			Chr	10	20

3. Count number of FP in windows

Assignment 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Benchmark	T	C	C	A	G	C	C	C	T	C	A	G	C	G	T	C	A	T	G	C
Sample	T	C	C	T	G	C	A	C	G	C	A	G	C	G	T	C	A	T	C	C



Assignment 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Benchmark	T	C	C	A	G	C	C	C	T	C	A	G	C	G	T	C	A	T	G	C
Sample	T	C	C	T	G	C	A	C	G	C	A	G	C	G	T	C	A	T	C	C

FP = 3

FP = 1