Mansoura University
Faculty of Computers and Information
Department of Computer Science
First Semester: 2020-2021

[MED121] Bioinformatics: Sequencing Technologies I

Grade: Third Year (Medical Informatics Program)

Sara El-Metwally, Ph.D.

Faculty of Computers and Information,

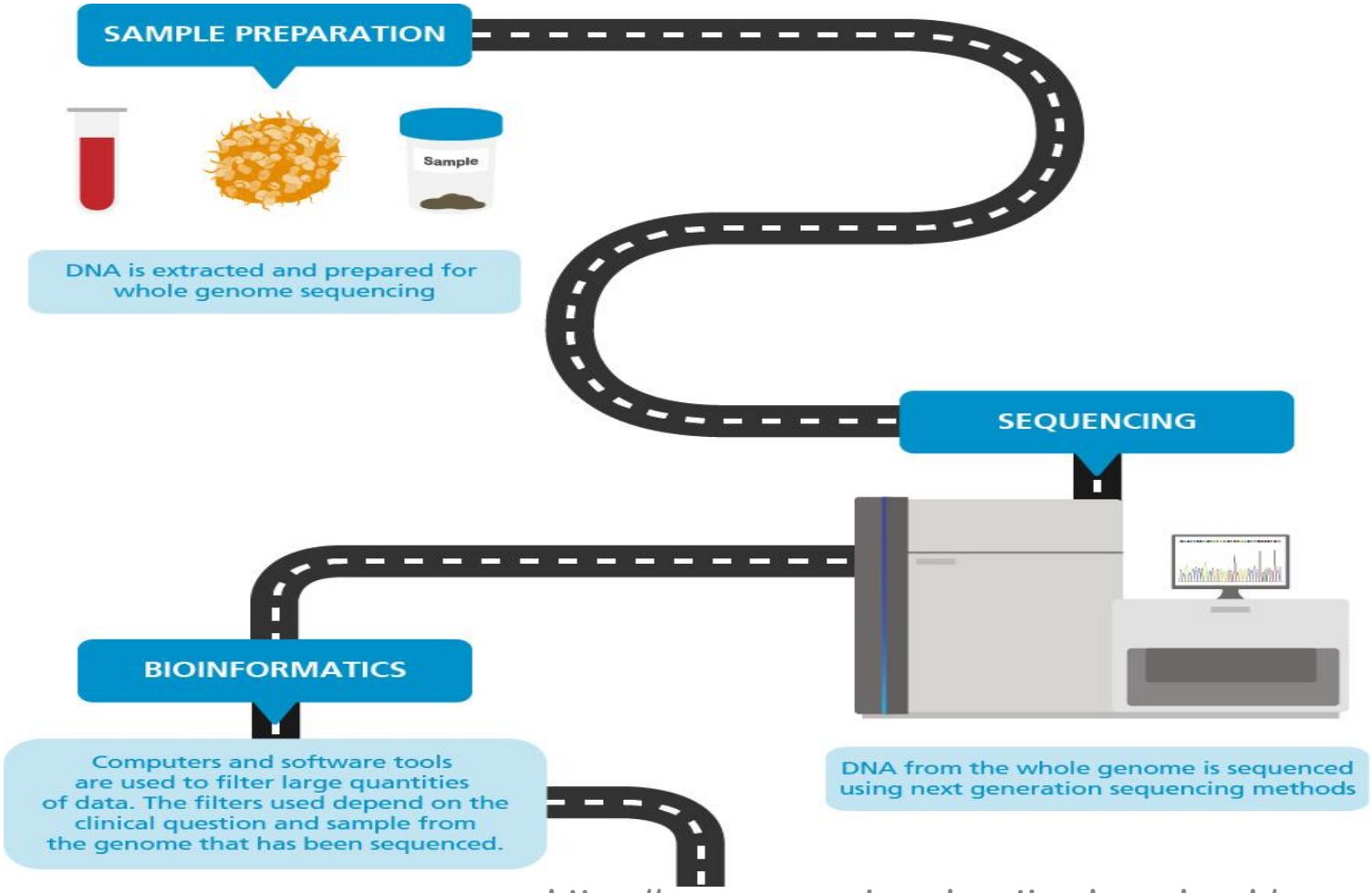
Mansoura University,

Egypt.

AGENDA

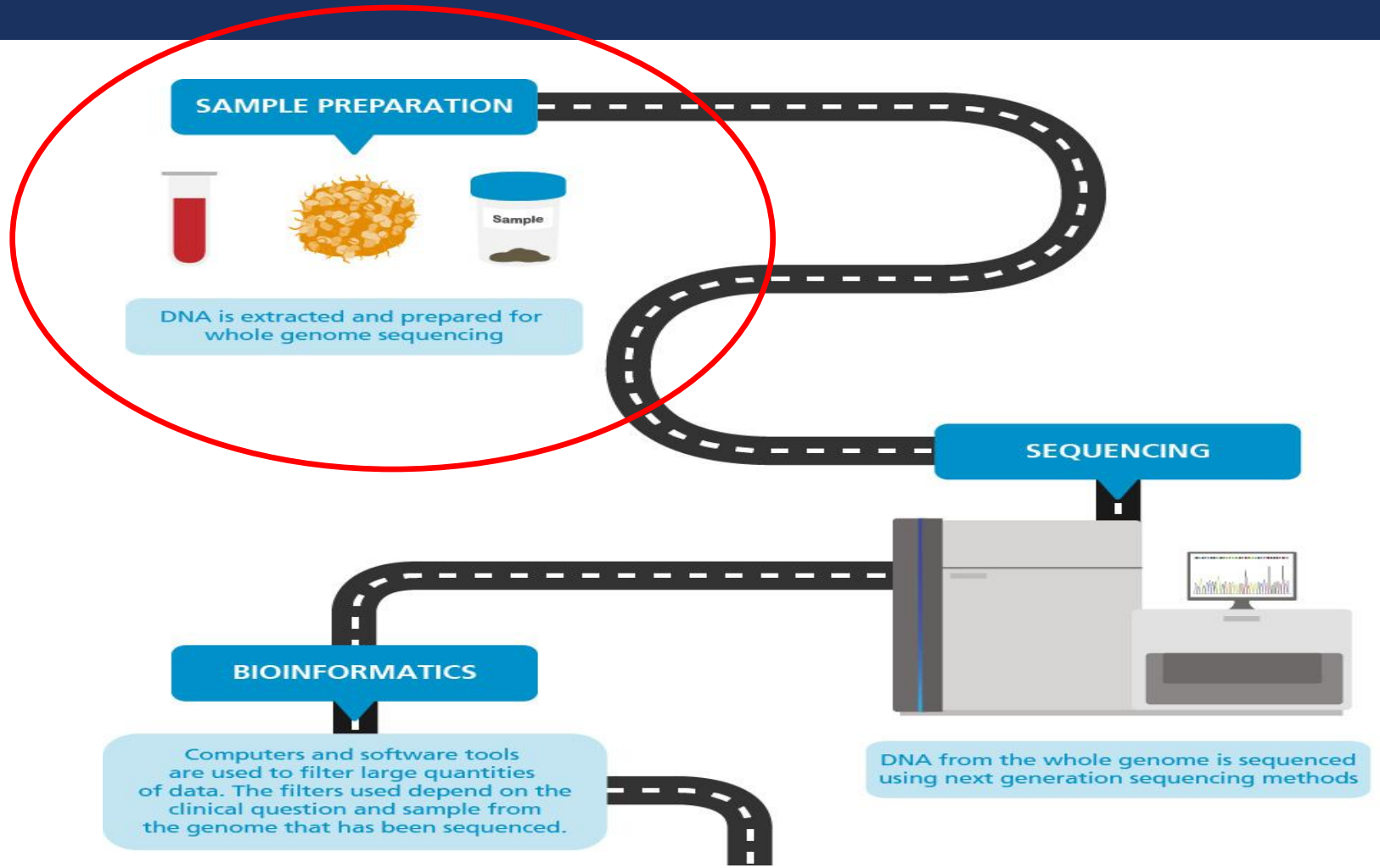
- Bioinformatics Road.
- Prepare a sequencing library.
- Sequencing Technologies
- Sequencing by Synthesis.
- Base Caller.
- Nanopore Sequencing .
- Single Molecule Real Time Sequencing.

BIOINFORMATICS ROAD



https://www.genomicseducation.hee.nhs.uk/

BIOINFORMATICS ROAD

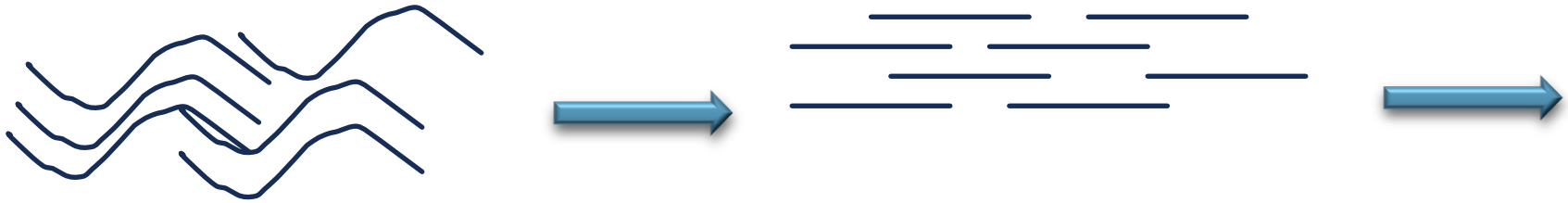


https://www.genomicseducation.hee.nhs.uk/

PREPARE A SEQUENCING LIBRARY

Step 1: Isolate the DNA

Step 2: Break the DNA into small fragments.



We do this because DNA is three billions bases long and the sequencing machines can only sequence short (200-300 bp) fragments.

PREPARE A SEQUENCING LIBRARY

Step 1: Isolate the DNA

Step 2: Break the DNA into small fragments.

Step 3: Add sequencing adapters.

The adapters do two things:

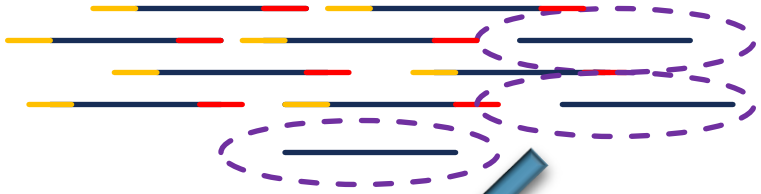
- ✓Allow the sequencing machine to recognize the fragments.
- ✓Allow you to sequence different samples at the same time, since different samples have different adapters. (save time and money)

PREPARE A SEQUENCING LIBRARY

Step 1: Isolate the DNA

Step 2: Break the DNA into small fragments.

Step 3: Add sequencing adapters.



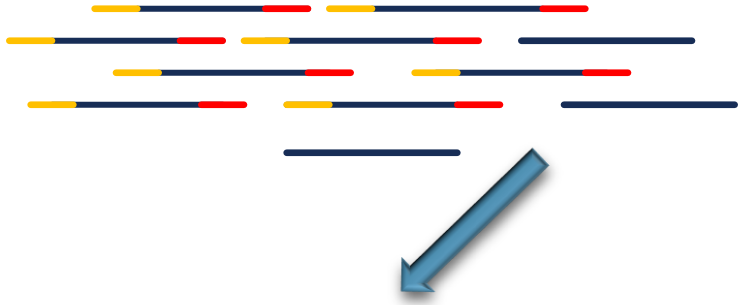
Note:
This step does not work 100% of the time (some fragments are not recognized by machine, so they are not represented in the sample)

PREPARE A SEQUENCING LIBRARY

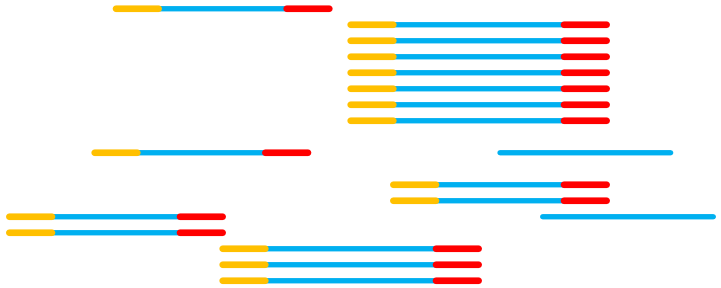
Step 1: Isolate the DNA

Step 2: Break the DNA into small fragments.

Step 3: Add sequencing adapters.



Step 4: PCR amplification

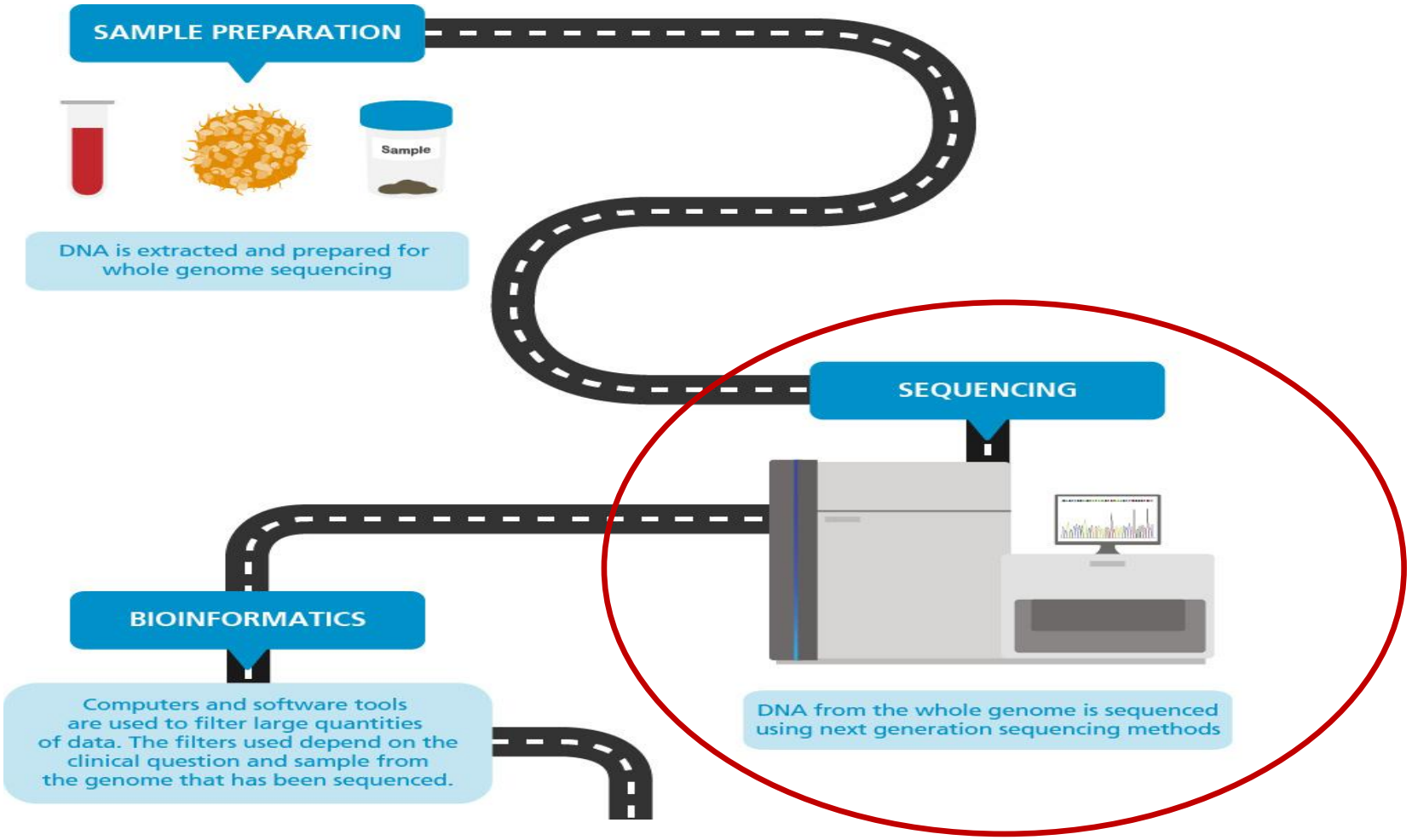


Step 5: QC

- ✓ Verify Library concentration
- ✓ Verify Library fragment lengths.

Only fragments with sequencing adapters are amplified, they are enriched.

BIOINFORMATICS ROAD



https://www.genomicseducation.hee.nhs.uk/

SEQUENCING METHODS

- **Sequencing:** the process of determining the precise order of nucleotides (A,C,T and G) or amino acid sequence within a DNA fragment or a protein using chemical and enzymatic reactions.

- DNA sequencing machines are evolved through three generations.

- Each generation is characterized by some factors such as the sequencing technology, cost, the quantity and the quality of the sequencing data.

INTRODUCTION TO SEQUENCING TECHNOLOGIES

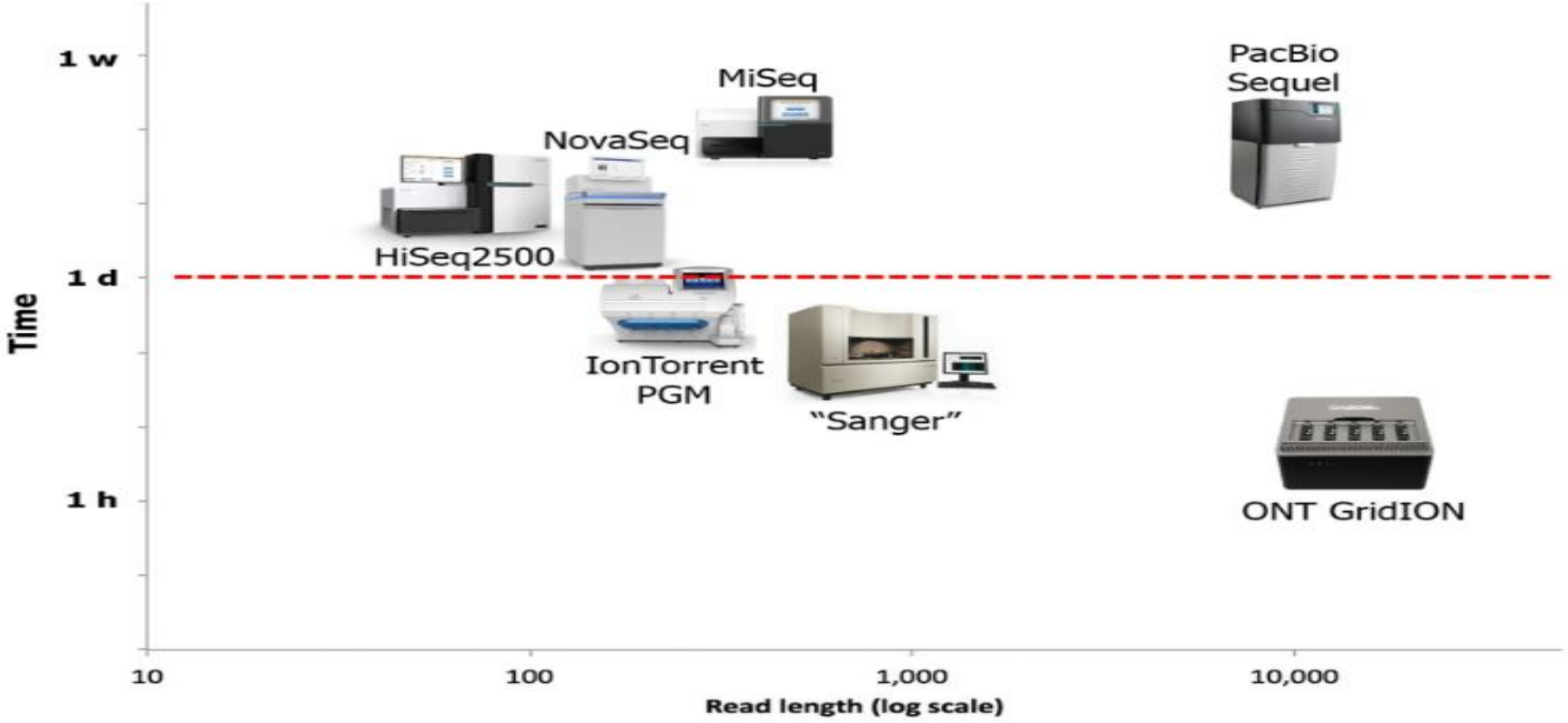
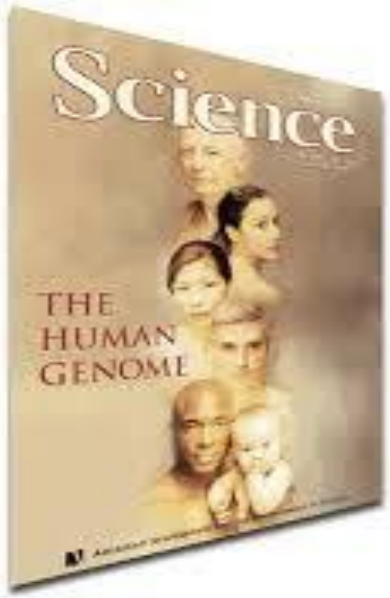
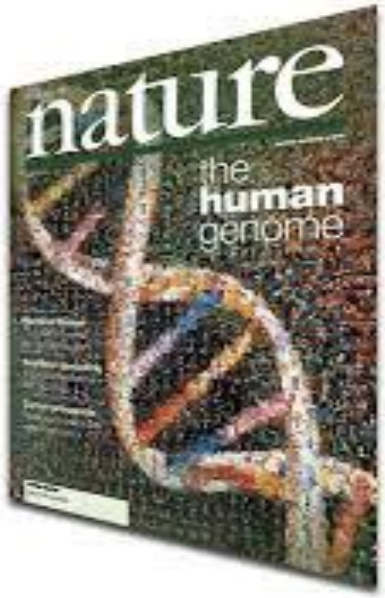


image credits : https://academic.oup.com/femsle/article/365/16/fny103/4982775

SEQUENCING TECHNOLOGIES

1990
13 years of Sequencing
3 B$





2008
5 months
1.5 M$

Now
1 day, 1000 $

HUMAN REFERENCE GENOME

GRCh38

⚠ **This assembly has been updated.** See current version

Description: Genome Reference Consortium Human Build 38
Organism name: Homo sapiens (human)
BioProject: PRJNA31257
Submitter: Genome Reference Consortium
Date: 2013/12/17
Synonyms: hg38
Assembly type: haploid-with-alt-loci
Assembly level: Chromosome
Genome representation: full
GenBank assembly accession: GCA_000001405.15 (replaced)
RefSeq assembly accession: GCF_000001405.26 (replaced)
RefSeq assembly and GenBank assembly identical: yes

Global statistics*

Number of regions with alternate loci or patches	207
Total sequence length	3,099,734,149
Total ungapped length	2,948,611,470
Gaps between scaffolds	349
Number of scaffolds	473
Scaffold N50	67,794,873
Scaffold L50	16
Number of contigs	999
Contig N50	57,879,411
Contig L50	18
Total number of chromosomes and plasmids	24
Number of component sequences (WGS or clone)	35,614

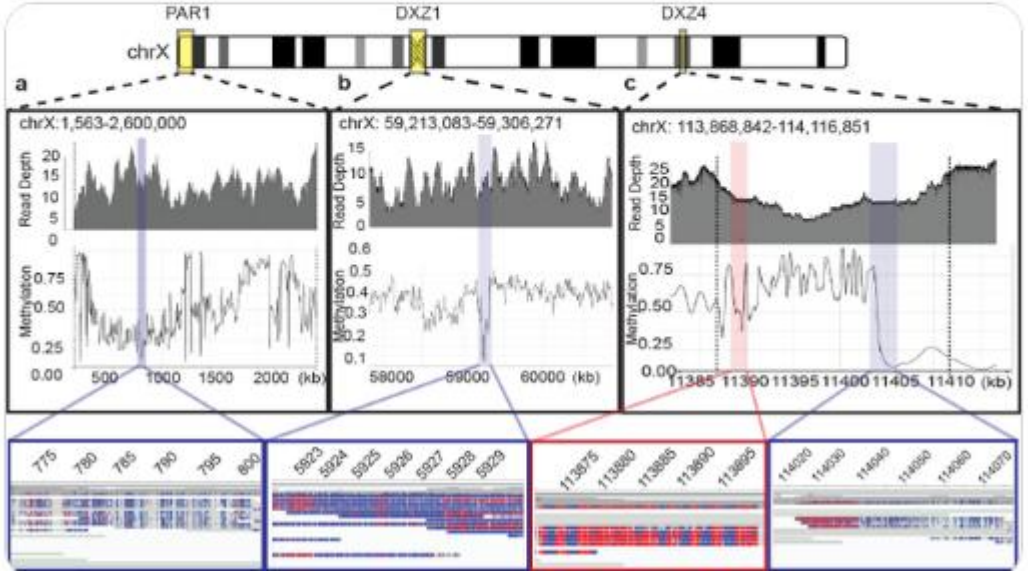
Image credits: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/

T2T CONSORTIUM

Ewan Birney and 8 others liked

Adam Phillippy @aphillippy · 16h

So proud of @khmiga @sergekoren and our entire #T2T consortium for the first-ever "Telomere-to-telomere assembly of a complete human X chromosome"! Fills all reference gaps and reveals methylation patterns of satellite arrays nature.com/articles/s4158...



nature

Explore our content Journal information

nature > articles > article

Article | Open Access | Published: 14 July 2020

Telomere-to-telomere assembly of a complete human X chromosome

Karen H. Miga, Sergey Koren, [...] Adam M. Phillippy

Nature **585**, 79–84(2020) | Cite this article

SEQUENCING BY SYNTHESIS

GATTACTTG

ATCTTGT

ATGCTT

GATC

ATGCTT

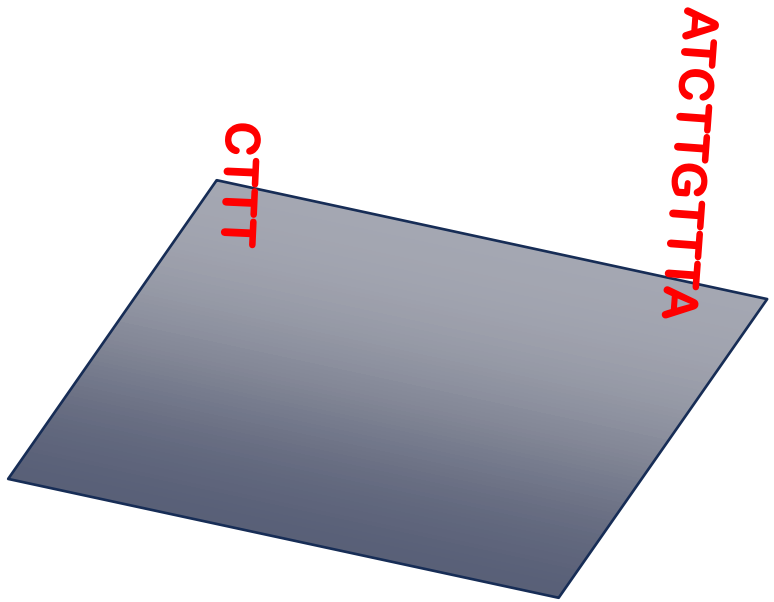
ATCTTGATTACTTGTT

CTTT

ATCTTGTTTA

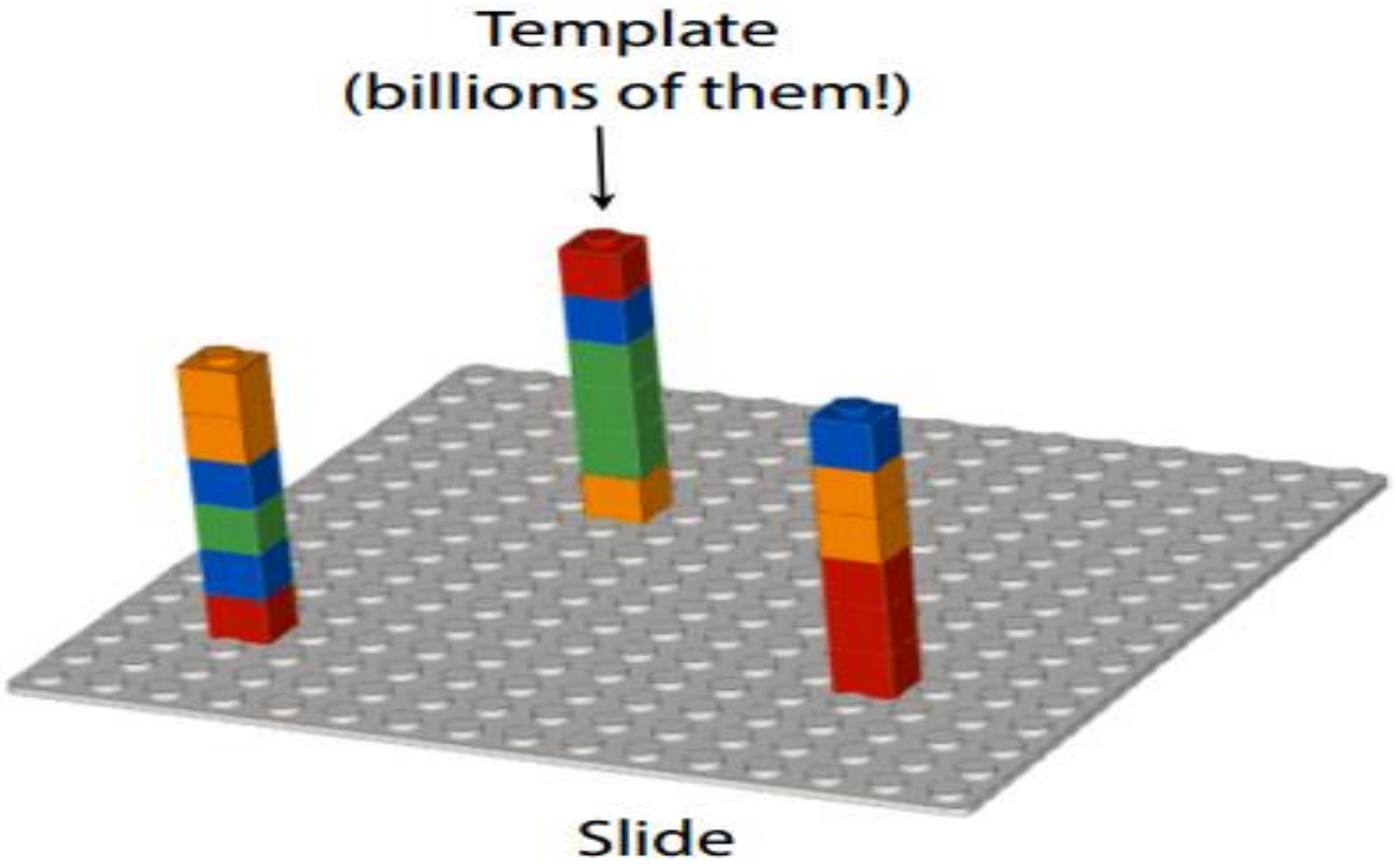
ATCTTGT

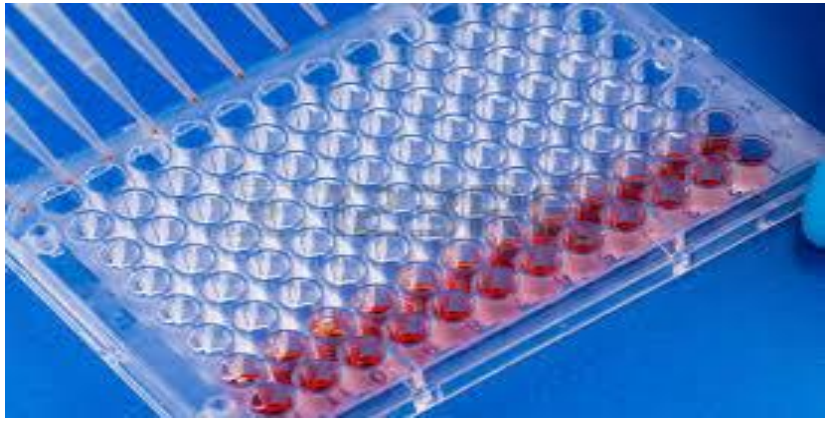
ATGCTT

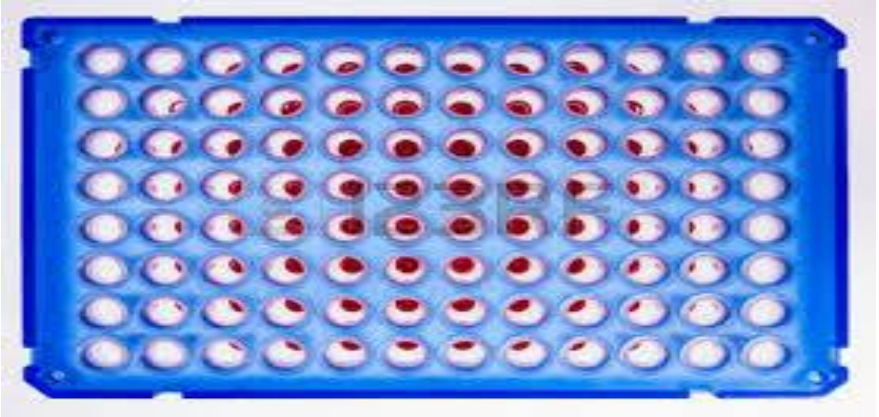


Deposit it on slide

SEQUENCING BY SYNTHESIS

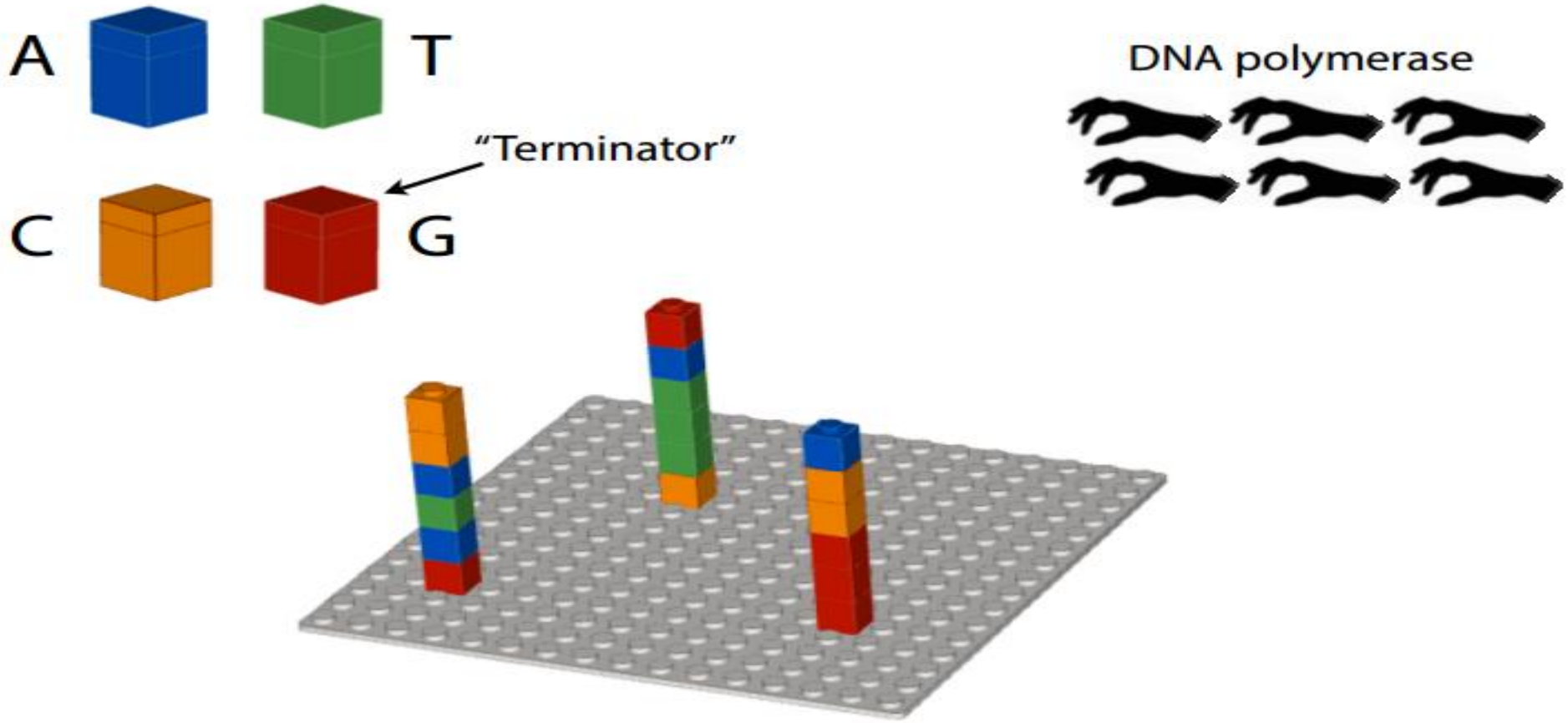






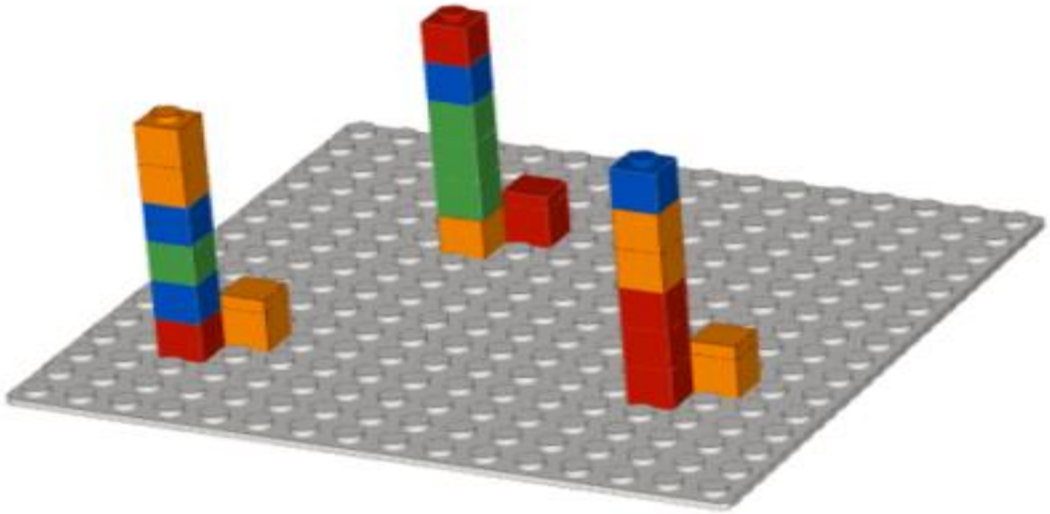
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

SEQUENCING BY SYNTHESIS



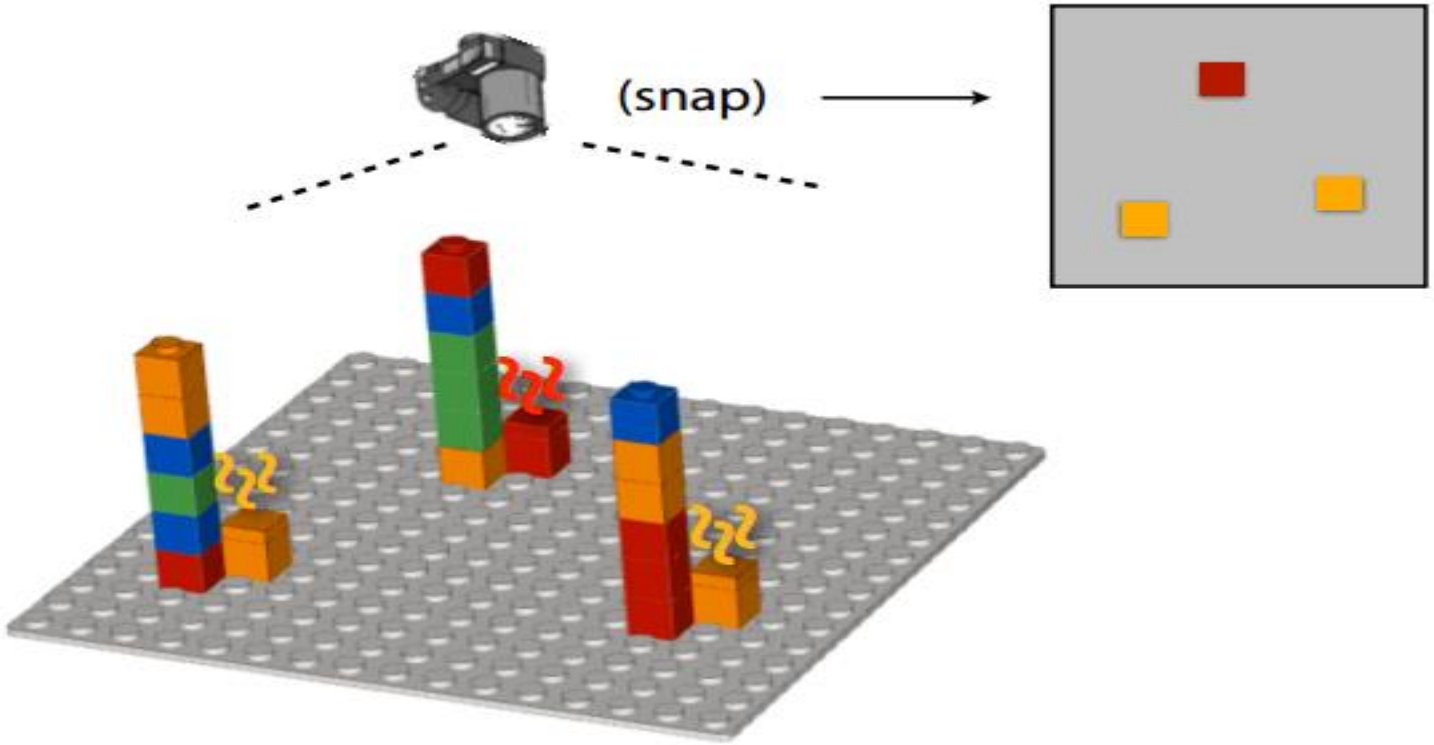
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

SEQUENCING BY SYNTHESIS



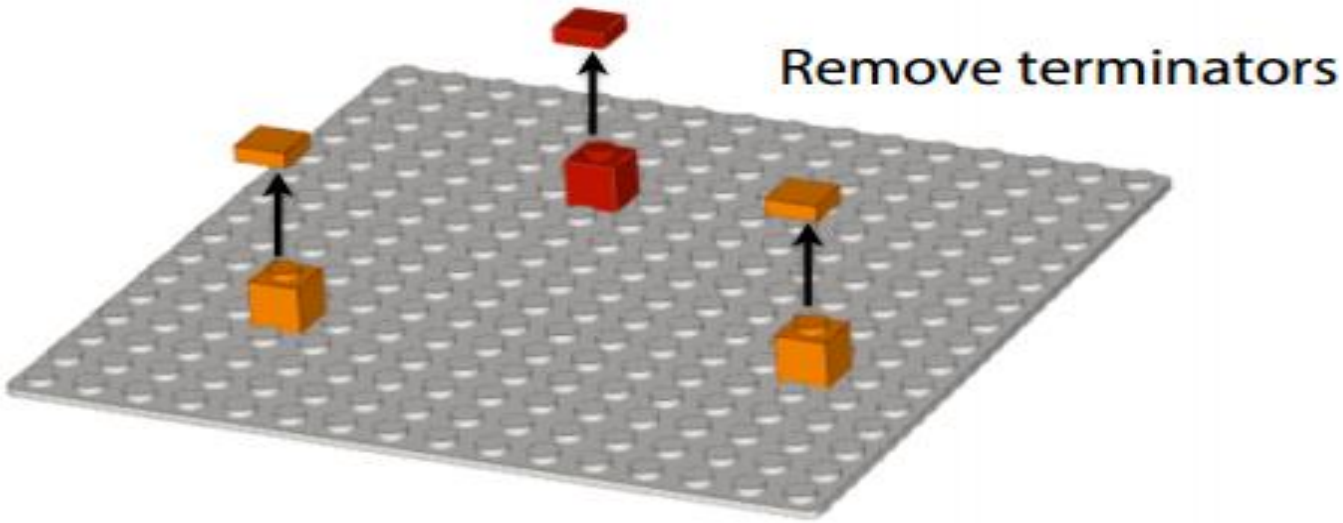
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

SEQUENCING BY SYNTHESIS



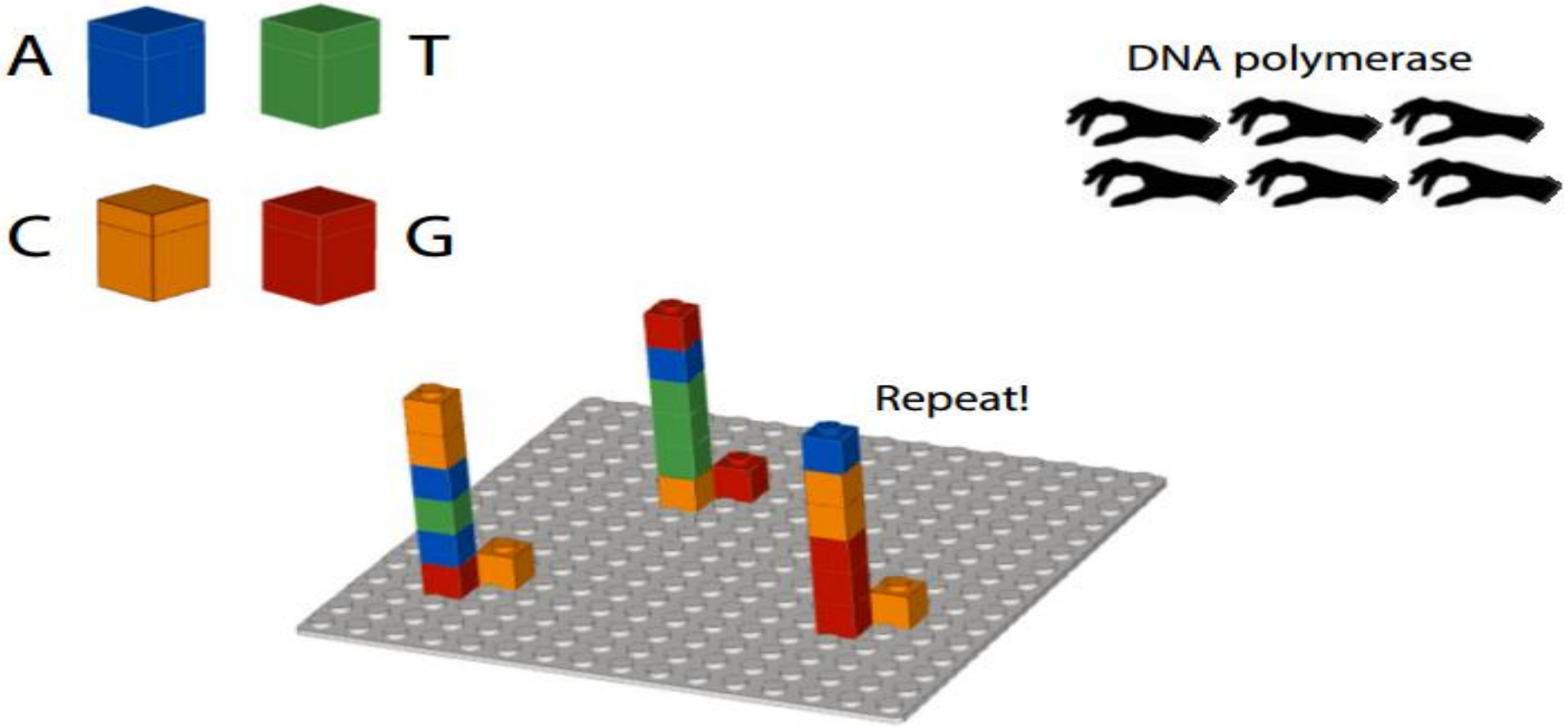
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

SEQUENCING BY SYNTHESIS

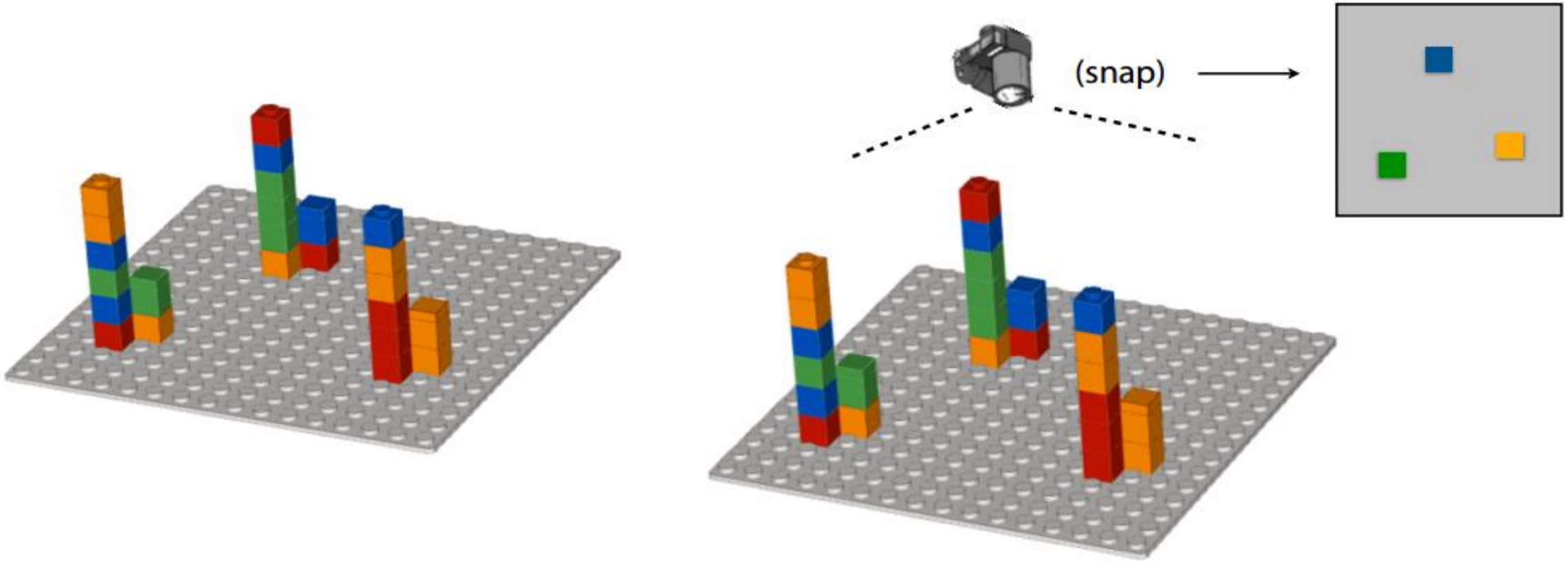


Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

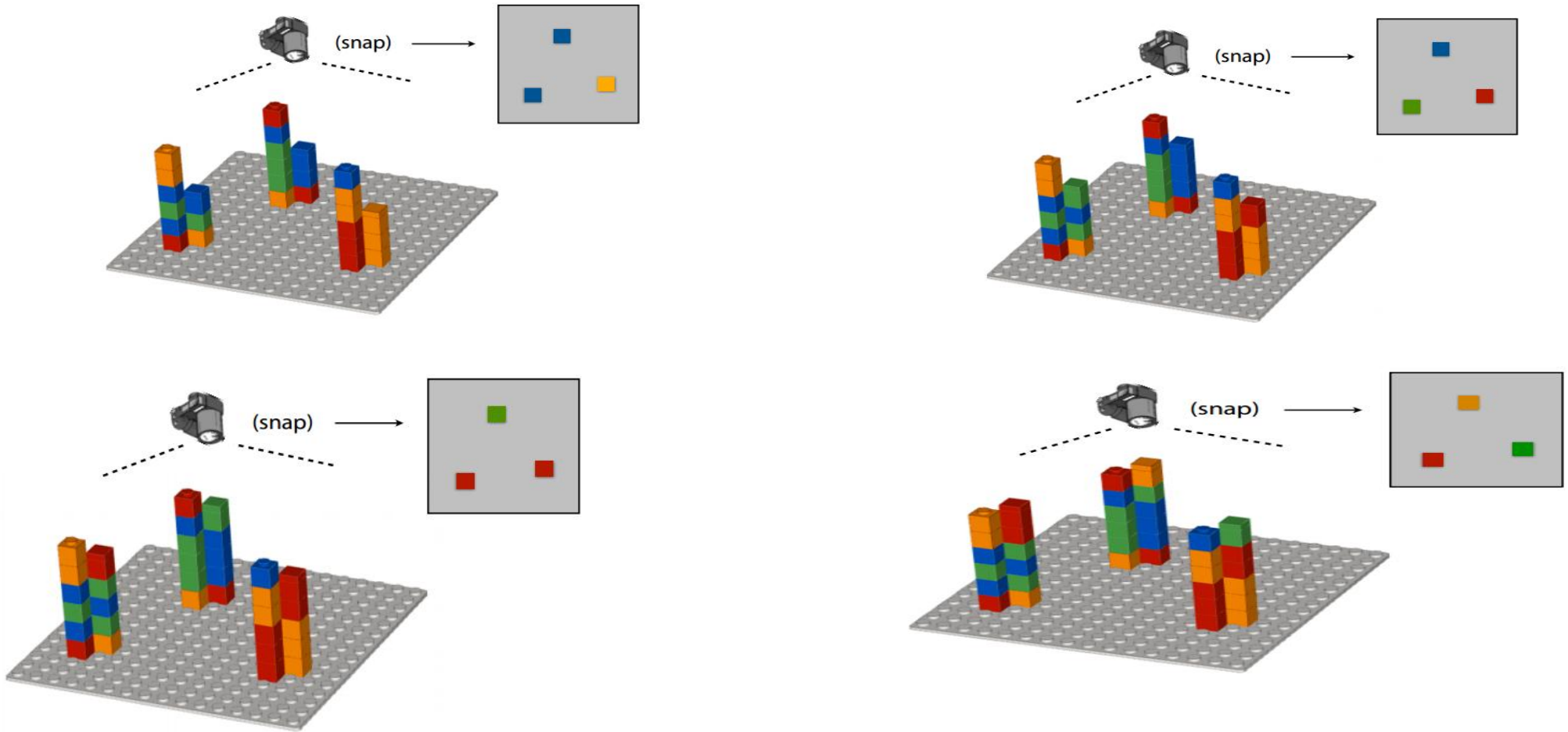
SEQUENCING BY SYNTHESIS



SEQUENCING BY SYNTHESIS

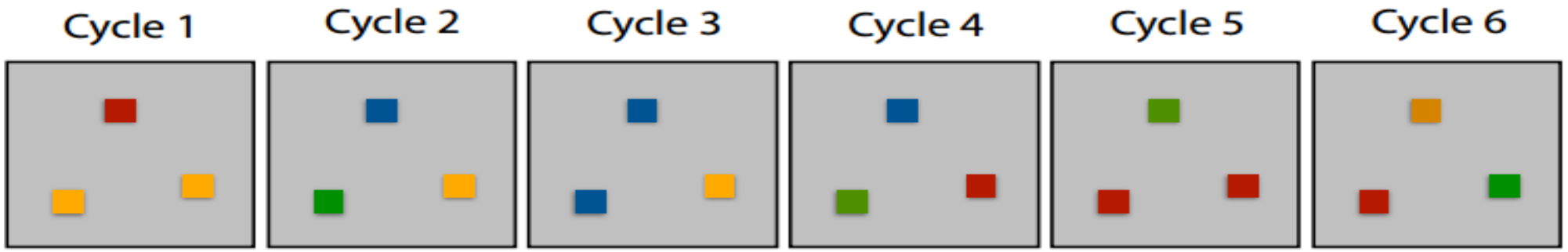


SEQUENCING BY SYNTHESIS



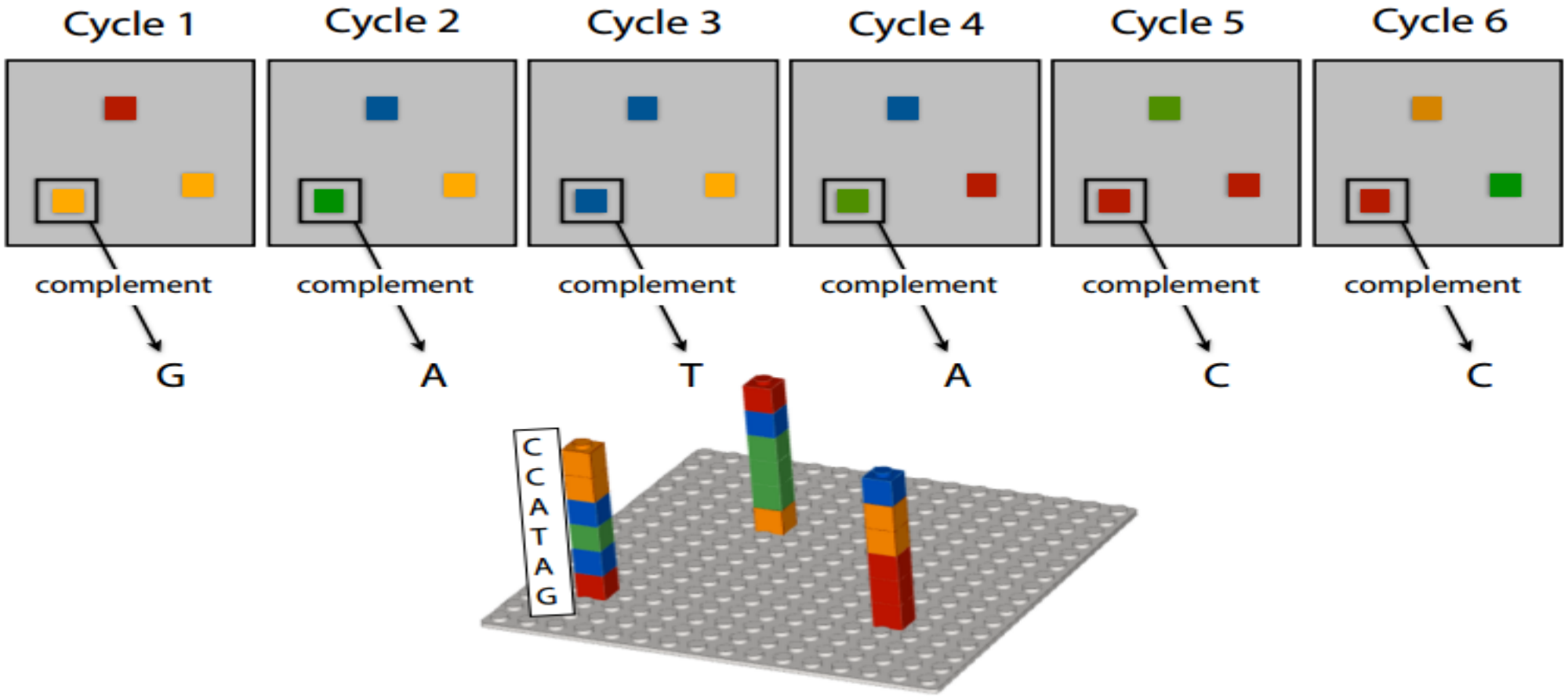
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

SEQUENCING BY SYNTHESIS



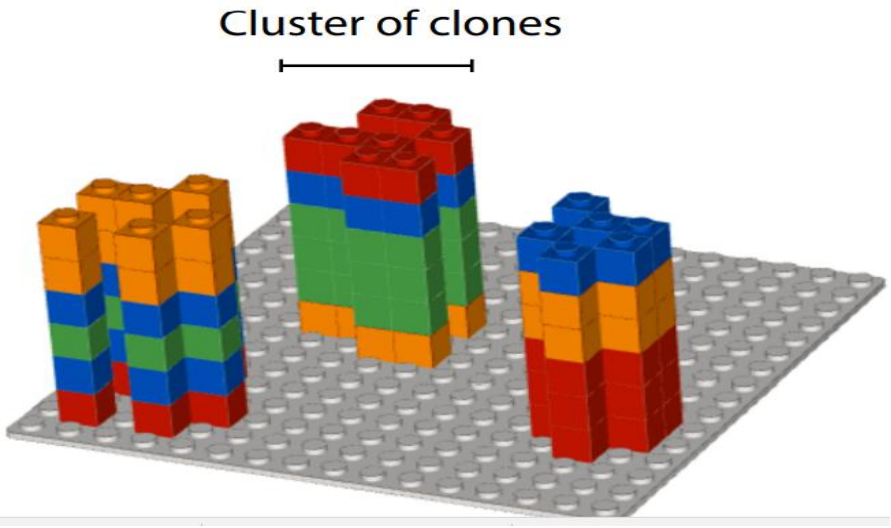
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

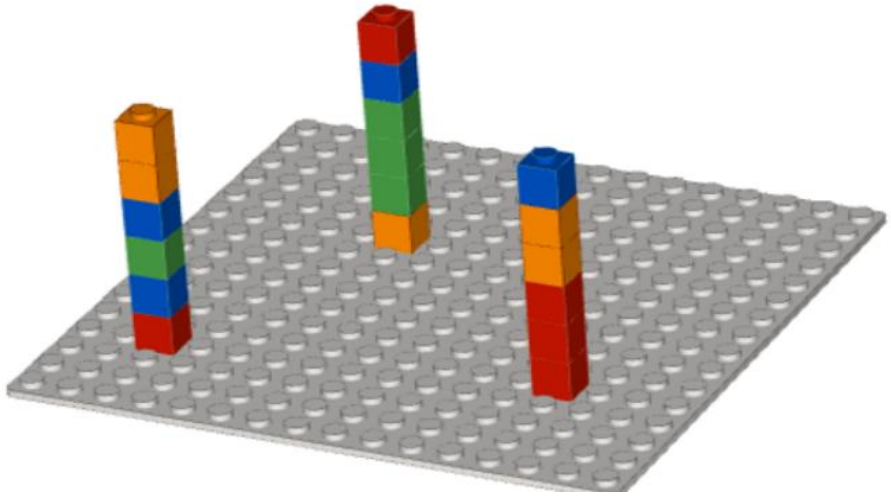
SEQUENCING BY SYNTHESIS



Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

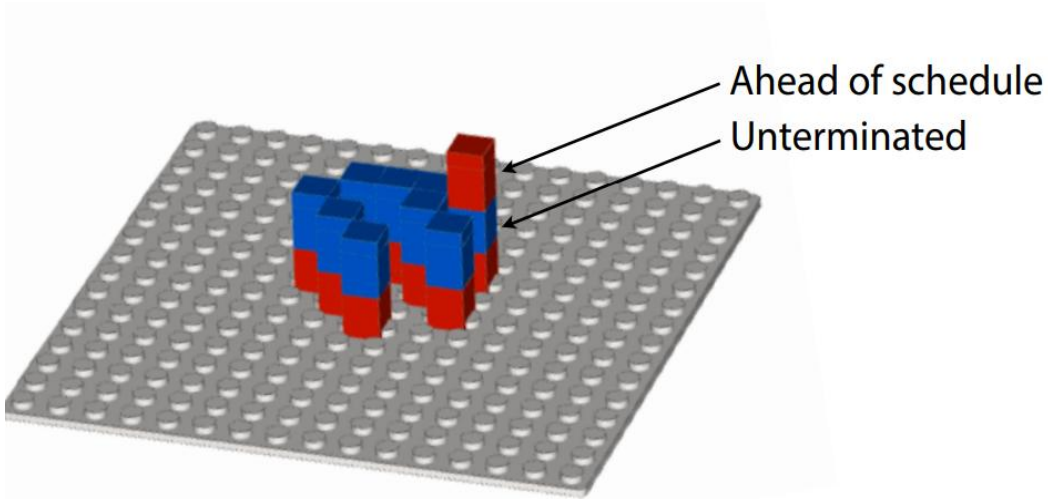
SEQUENCING BY SYNTHESIS

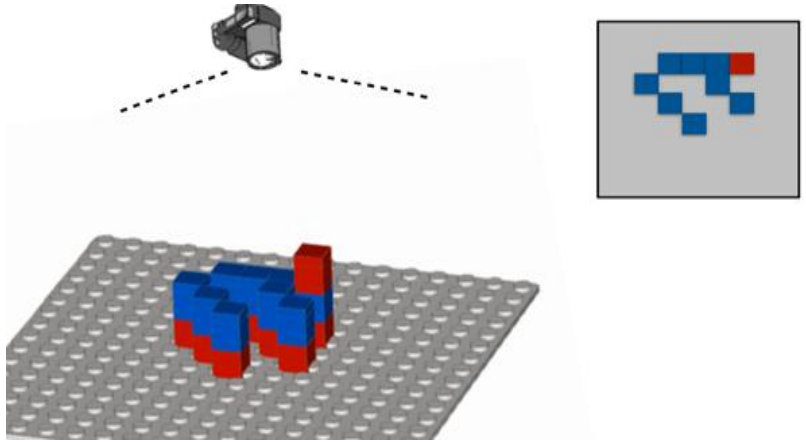




Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

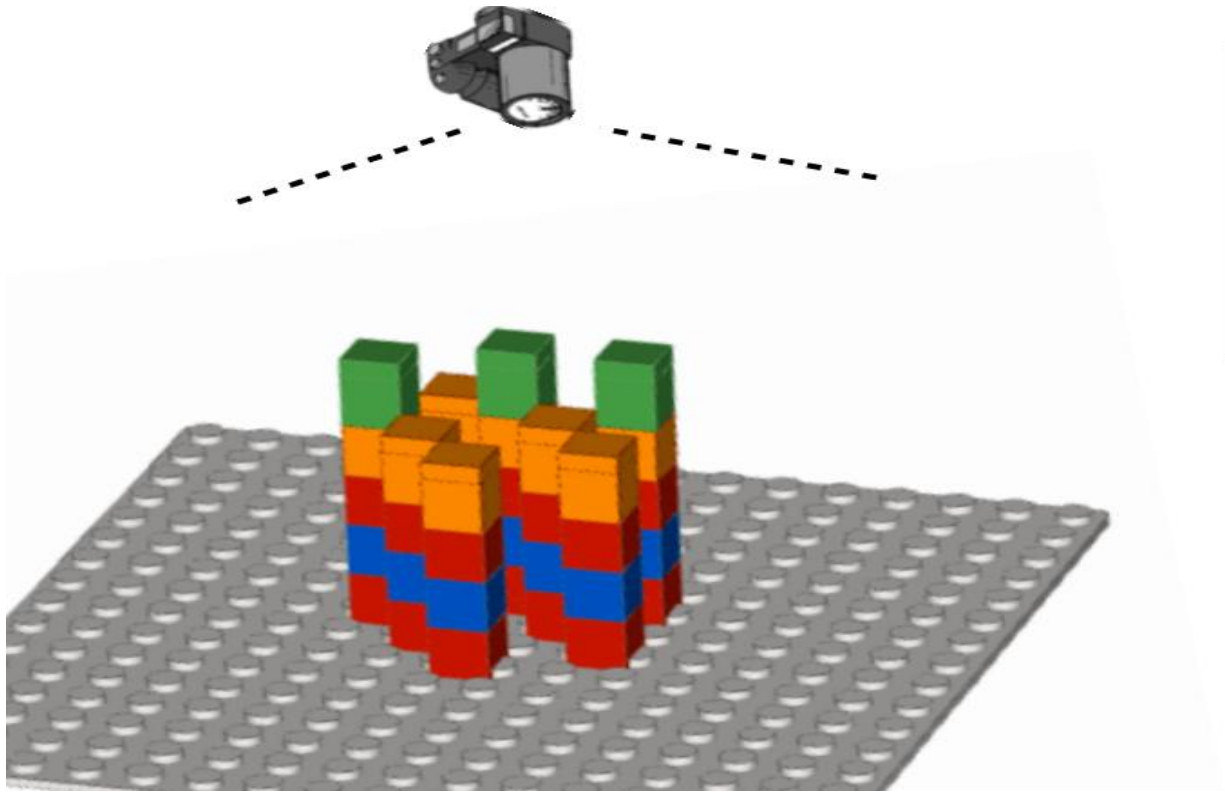
SEQUENCING BY SYNTHESIS

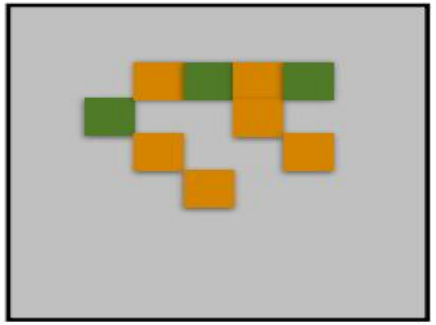




Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

SEQUENCING BY SYNTHESIS





Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

BASE CALLER

- **Base caller** is a program that assigns bases to chromatogram peaks.

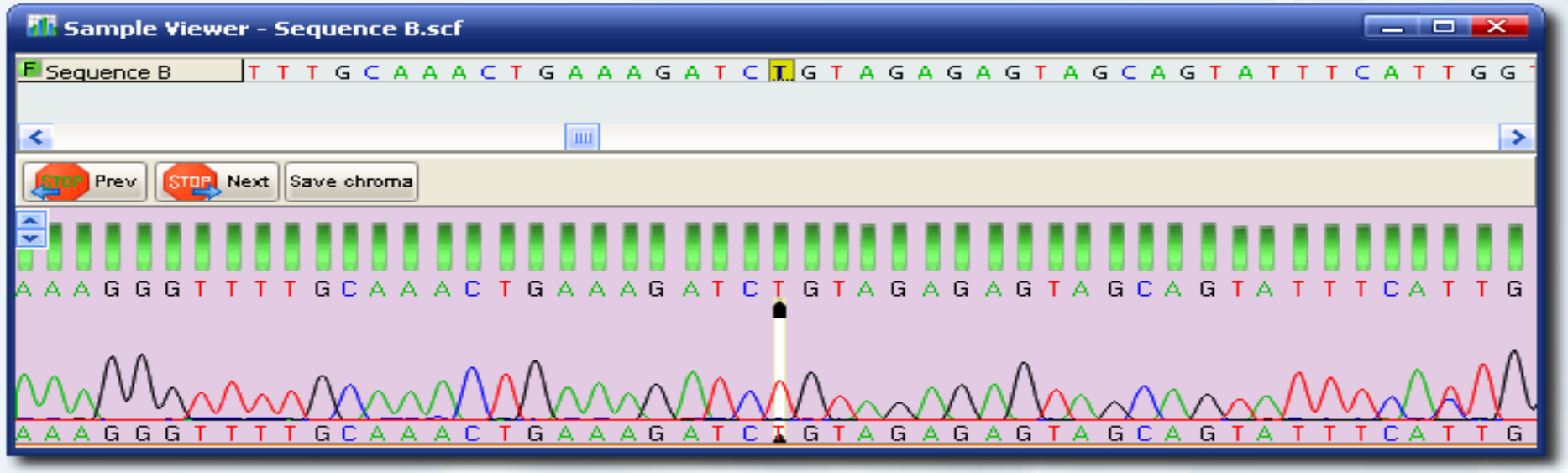


Fig. 1 - Example of chromatogram. The green bars above chromatogram peaks high confidence scores.

Image Credit: http://www.dnabaser.com/help/samples/what%20is%20a%20chromatogram.html

BASE CALLER

- For each base, base caller reports an important value " **base quality**".

- **Base Quality** is the base caller estimates of the probability that the base was called incorrectly.

BASE QUALITIES

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

Base quality

Probability that base call is incorrect

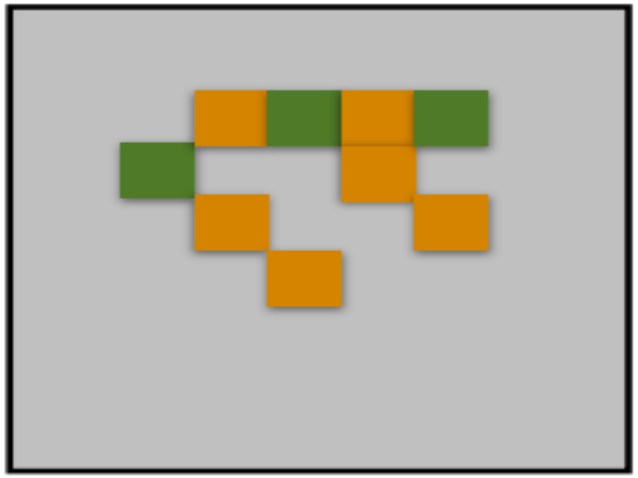
$Q = 10 \rightarrow$ 1 in 10 chance call is incorrect

$Q = 20 \rightarrow$ 1 in 100

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

Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

BASE QUALITIES



Suppose that base call is C "orange "

Estimate p?

Probability that base call is incorrect.

$$p = \frac{\text{non} - \text{orange light}}{\text{total light}}$$

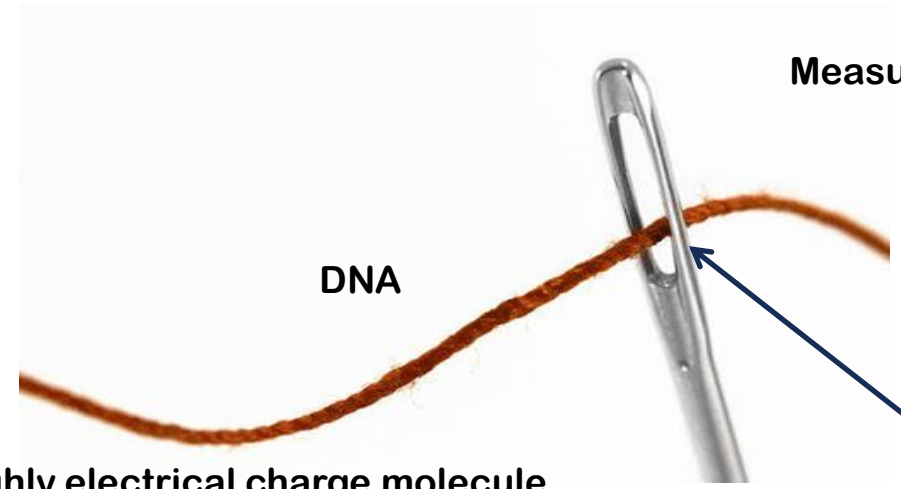
$$p = \frac{3}{9} = \frac{1}{3}$$

$$Q = -10 \times \log_{10} \mathrm{p}$$

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

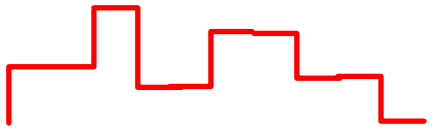
Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

NANOPORE SEQUENCING

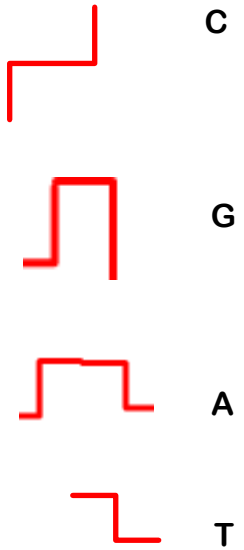


Highly electrical charge molecule

Measure differences in the magnitude of electrical voltages

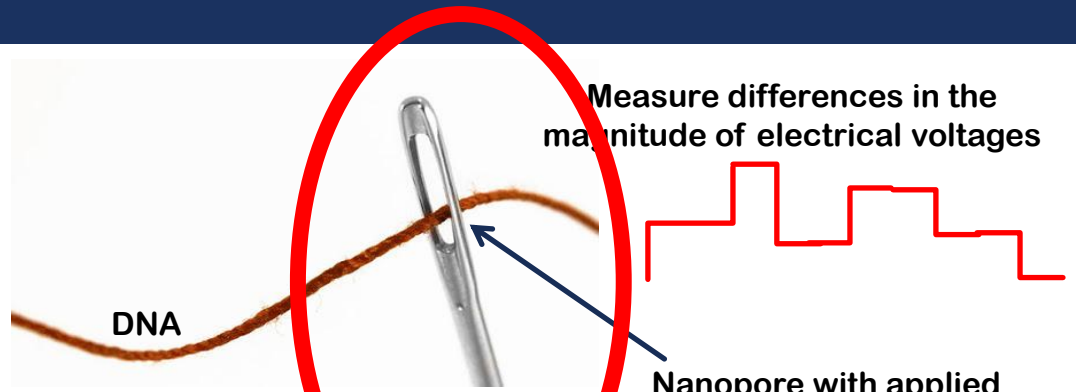


Nanopore with applied electrical voltages



NANOPORE SEQUENCING

Small diameter pore 1nm

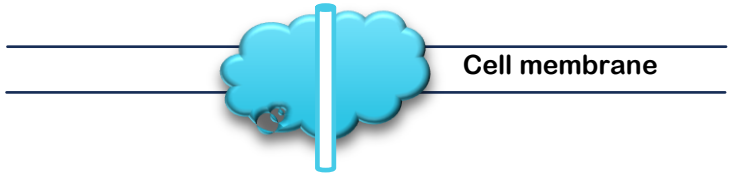


Measure differences in the magnitude of electrical voltages

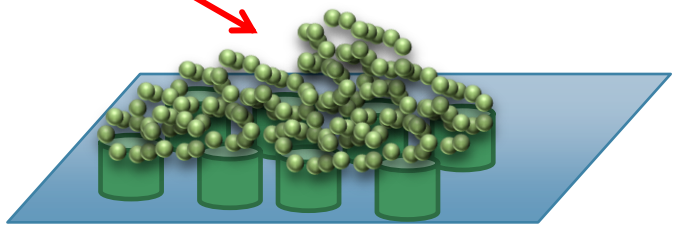
Nanopore with applied electrical voltages

Highly electrical charge molecule

Transmembrane protein

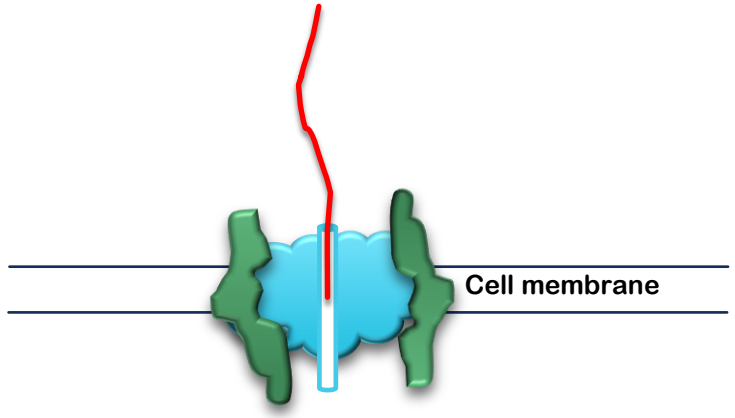


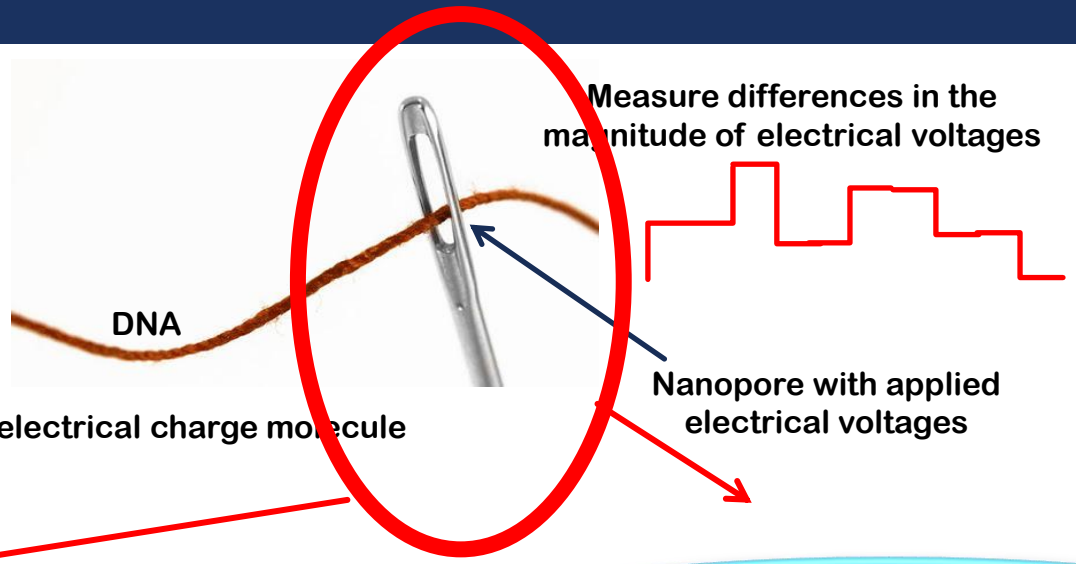
natural material extracted from biological cells

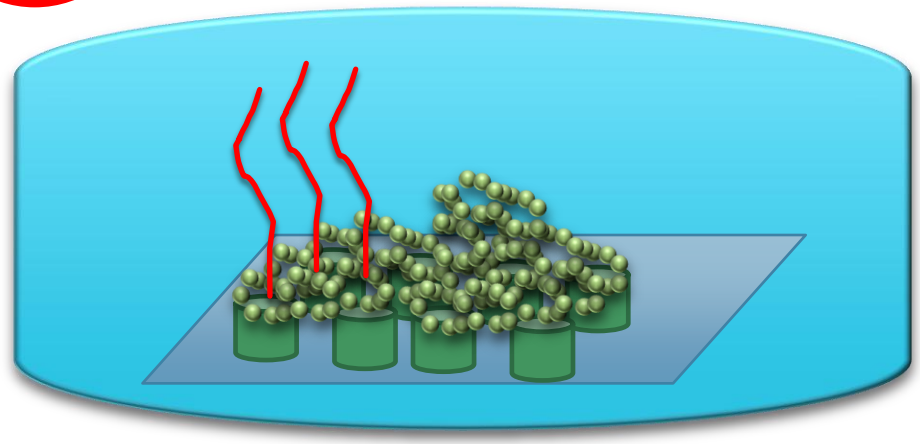


In Silicon

NANOPORE SEQUENCING







Electrophoresis solution

SEQUENCING TECHNOLOGIES

- https://www.youtube.com/watch?v=IT3NqhKD840 (SMART Sequencing)
- https://nanoporetech.com/applications/dna-nanopore-sequencing

SEQUENCING TECHNOLOGIES

Throughput

Read length

sequencing errors

Sequencing cost

Bioinformatics tools available for data analysis and processing

TOP TECHNOLOGIES IN THE SEQUENCING MARKET.

Company	Instruments
Illumina	MiniSeq; NextSeq; MiSeq; HiSeq; NovaSeq
Pacific biosciences	RSII; Sequel
Oxford Nanopore Technologies	SmidgION (under dev); MinION; GridION; PromethION (under dev)

Table credits : https://academic.oup.com/femsle/article/365/16/fny103/4982775

Sequencing Power for Every Scale

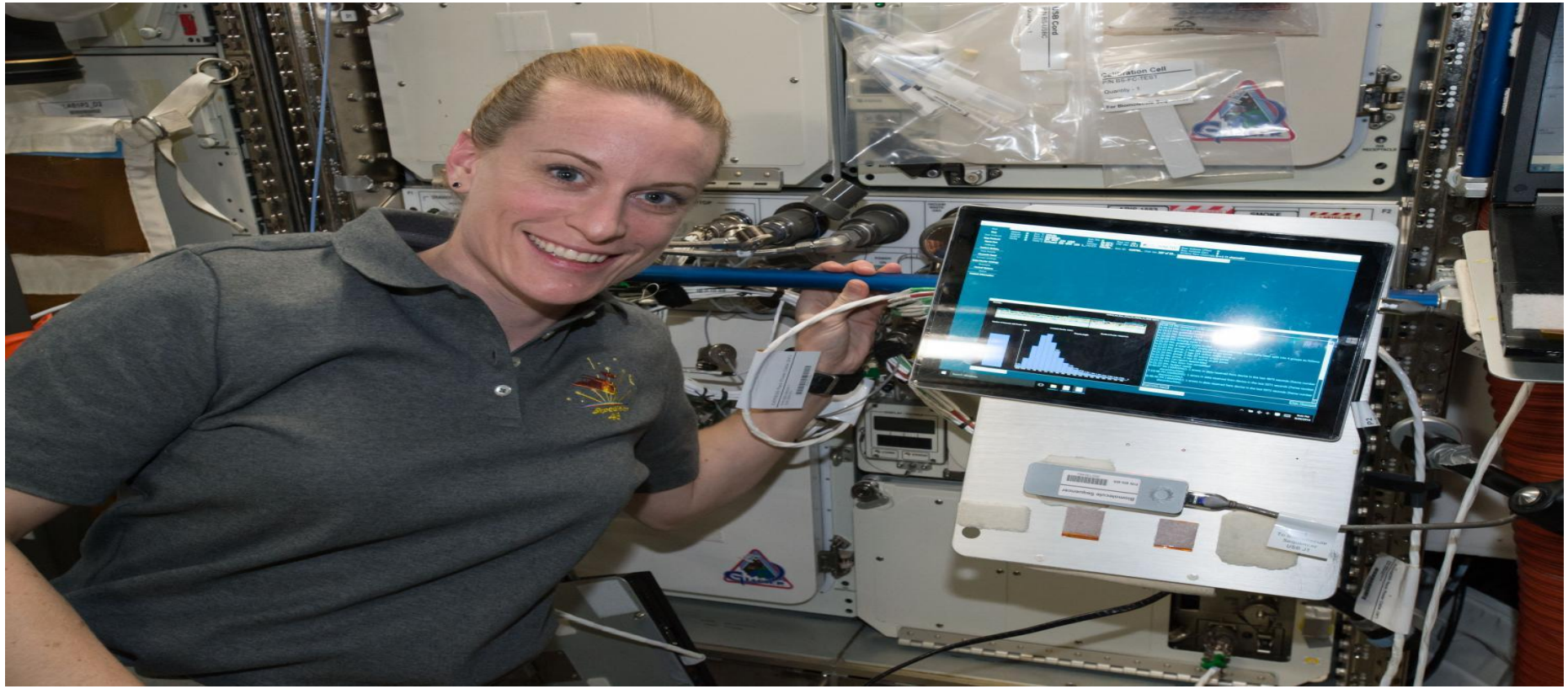
The broadest portfolio offering available

Sequencing System	iSeq™	MiniSeq™	MiSeq®	NextSeq®	HiSeq®	HiSeq® X	NovaSeq®
					4000	Five/Ten	6000
Output per run	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb[1]
Instrument price	$19.9K	$49.5K	$99K	$275K	$900K	$6M[2]/$10M[2]	$985K
Installed base[3]	NA	~600	~6,000	~2,400	~2,300[4]		~285

1. Output per run for the S1, S2 and S4 flow cells equal 1 Tb, 2 Tb and 6 Tb, respectively assuming two flow cells per run
2. Based on purchase of 5 and 10 units for HiSeq X Five and HiSeq X Ten, respectively
3. Based on end of fiscal year 2017
4. Combined HiSeq family

illumina®

SEQUENCING BY NANOPORE (FUN!)



Kate Rubins is pictured aboard ISS with the USB MinION sequencer (lower right) that was used in the first-ever DNA sequencing in space in August 2016.

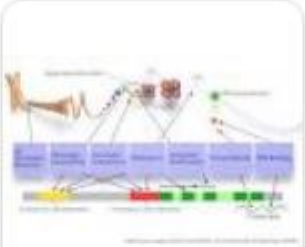
SEQUENCING BY NANOPORE (FUN!)

SEQUENCING PROJECTS

Human Genome Project

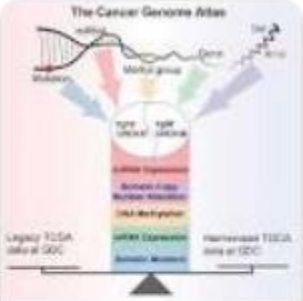
100,000 Genomes Project

1000 Genomes Project



ENCODE

Human Microbiome Project



The Cancer Genome Atlas

SEQUENCING PROJECTS

China—100,000 Genomes Project

United Kingdom—100,000 Genomes Project

Turkey—Turkish Genome Project

France—France Génomique (Médicine France Génomique 2025 or French Plan for Genomic Medicine 2025)

United States

Dubai, United Arab Emirates—Dubai Genomics

Saudi Arabia—Saudi Human Genome Program

Japan—Initiative on Rare and Undiagnosed Diseases

https://www.clinicalomics.com/topics/translational-research/biomarkers-topic/biobanking/10-countries-in-100k-genome-club/

SEQUENCING PROJECTS

أكاديمية البحث العلمي والتكنولوجيا المصرية
Yesterday at 12:26 AM ·

SCIDEV.NET
إطلاق مشروع جينوم للمصريين وقدماء المصريين
وقدماء المصريين

وزارة التعليم العالي والبحث العلمي المصرية
October 6 at 3:19 PM ·

مجلس أكاديمية البحث العلمي والتكنولوجى يوافق على برنامج الجينوم المصري

إطلاق مشروع جينوم للمصريين وقدماء المصريين

12 أكتوبر 2020 / الزيارات: 57

أعلن مجلس أكاديمية البحث العلمي والتكنولوجيا في مصر، بدء تنفيذ مشروع 'الجينوم البشري المرجعي للمصريين'، ضمن الخطة التنفيذية للأكاديمية لعام 2020-2021.

أُعلن عن المشروع يوم السادس من أكتوبر الجاري، مرتكزًا على ثلاثة محاور:

الأول: بناء جينوم مرجعي مصري يحمل المتغيرات الجينية الطبيعية والأكثر شيوعًا بين المصريين.

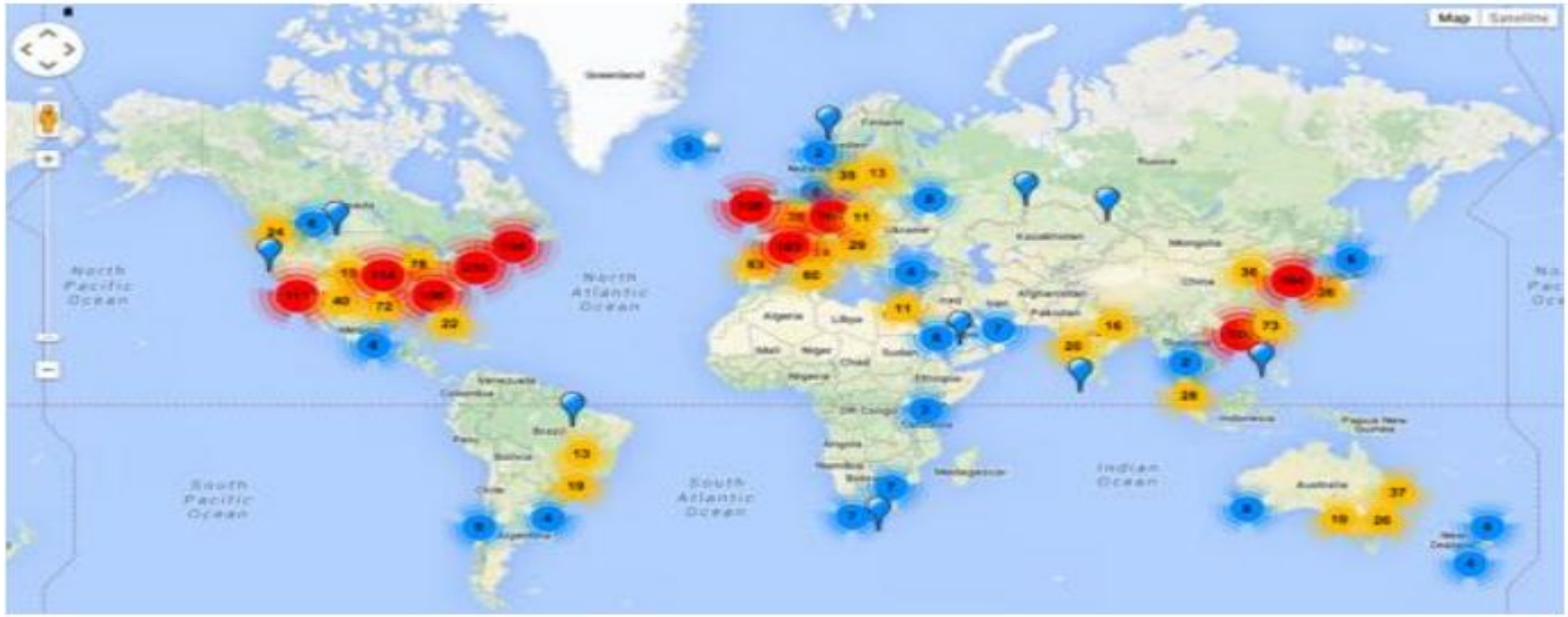
الثاني: هو دراسة جينوم المصريين القدماء،

الثالث: يكمن في البحث عن التغيرات الجينية المرتبطة بالأمراض الشائعة لدى الشعب المصري.

توفر الأكاديمية مليار جنيه مصري، تكفي لمعرفة المحتوى الجيني لنحو 20 ألف متطوع، يدرسها المشروع على مدار سنوات عمره الخمس، لكن المخطط زيادة مصادر التمويل كي يتسنى رسم التسلسل الوراثي لمئة ألف شخص.

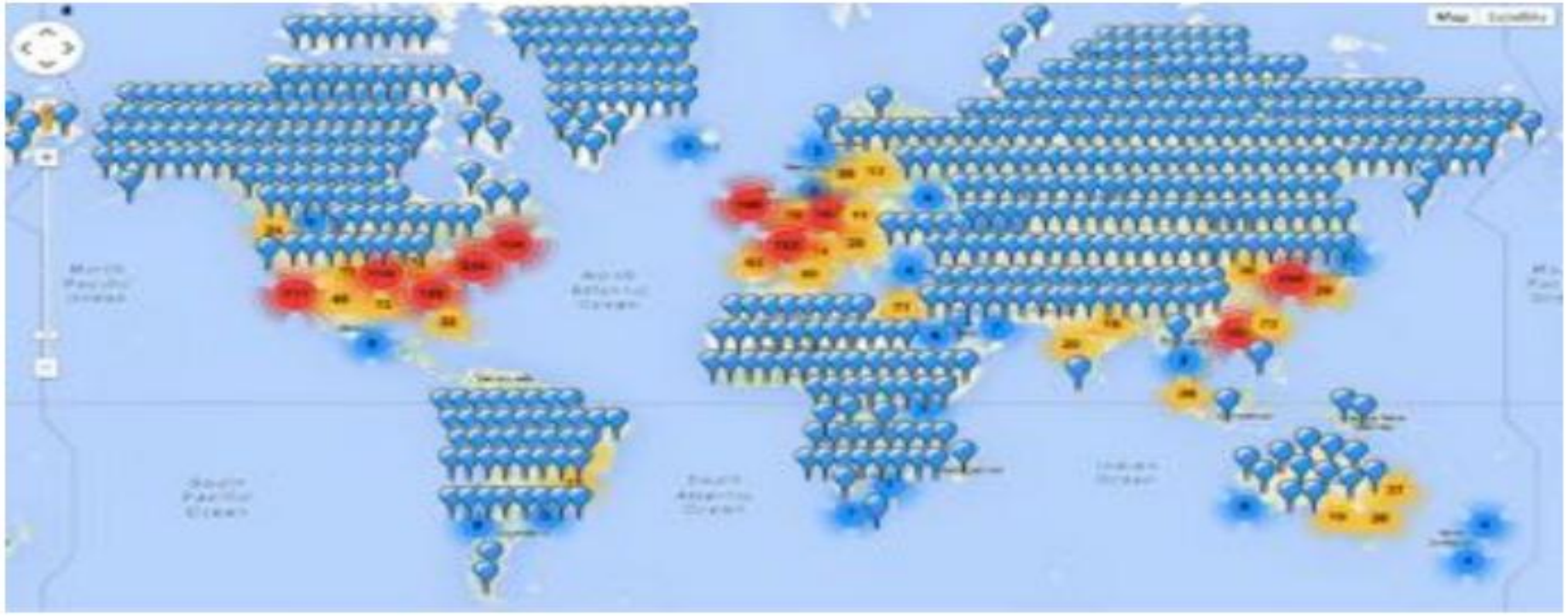
DATA DELUGE

Sequencing Centers 2018



DATA DELUGE

Sequencing Centers 2028



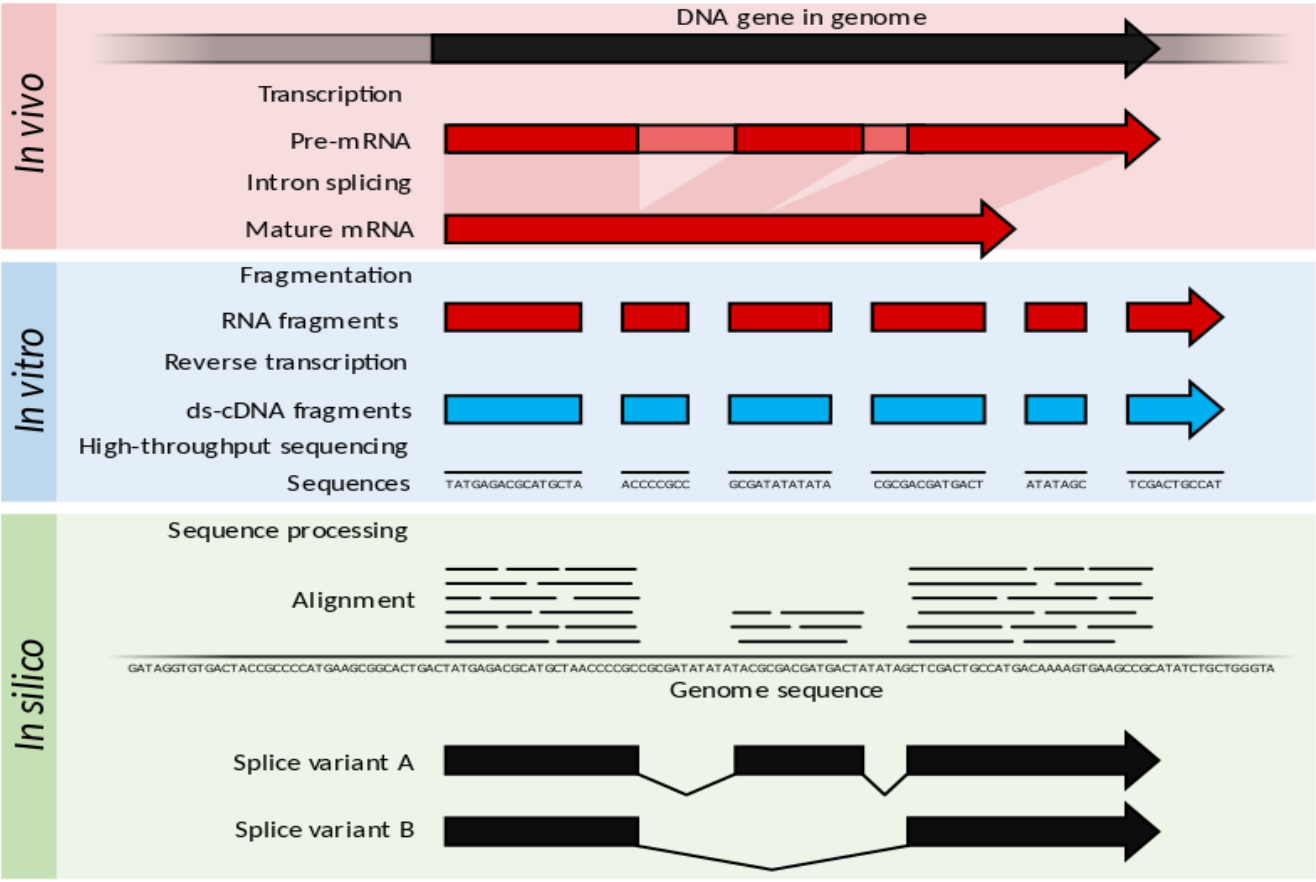
SEQUENCING SERVICES

https://www.abmgood.com/Whole-Genome-Sequencing-Service.html

For example:
the genes *KRAS* and *TP53* are often targeted across a range of cancer types, as they are commonly found to be mutated with a number of hotspots. *BRAF* and *EGFR* are also screened in many solid tumors, as they contain clinically relevant mutation

Image credits: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6861594/

RNA-SEQ



- RNA-seq is a particular technology-based sequencing technique which uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome.

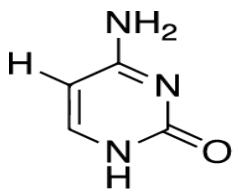
Image credit: https: https://en.wikipedia.org/wiki/RNA-Seq#/media/File:Summary_of_RNA-Seq.svg

NOTES

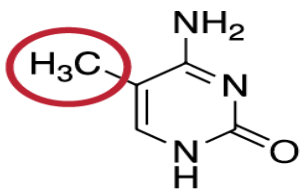
- Protein sequencing refers to methods for determining the amino acid sequence of proteins (or peptides) and analysis of the sequence, for example to infer protein conformation. Techniques include mass spectrometry and the Edman degradation reaction as well as prediction of the protein sequence from the encoding DNA or mRNA sequence.

SEQUENCING SERVICES

- Bisulfite Sequencing: is the use of bisulfite treatment of DNA before routine sequencing to determine the pattern of methylation.
- DNA methylation is a biological process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence. When located in a gene promoter, DNA methylation typically acts to repress gene transcription.



Cytosine



methylated Cytosine

Image credit: https://en.wikipedia.org/wiki/DNA_methylation#/media/File:DNA_methylation.png

SEQUENCING SERVICES

- In mammals, DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, repression of transposable elements, aging, and carcinogenesis.
- Treatment of DNA with bisulfite converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected. Therefore, DNA that has been treated with bisulfite retains only methylated cytosines. Thus, bisulfite treatment introduces specific changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single-nucleotide resolution information about the methylation status of a segment of DNA.

SEQUENCING SERVICES

- ChIP-sequencing, also known as ChIP-seq, is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest.
- Single cell sequencing examines the sequence information from individual cells with optimized next-generation sequencing (NGS) technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell .

SEQUENCING SERVICES

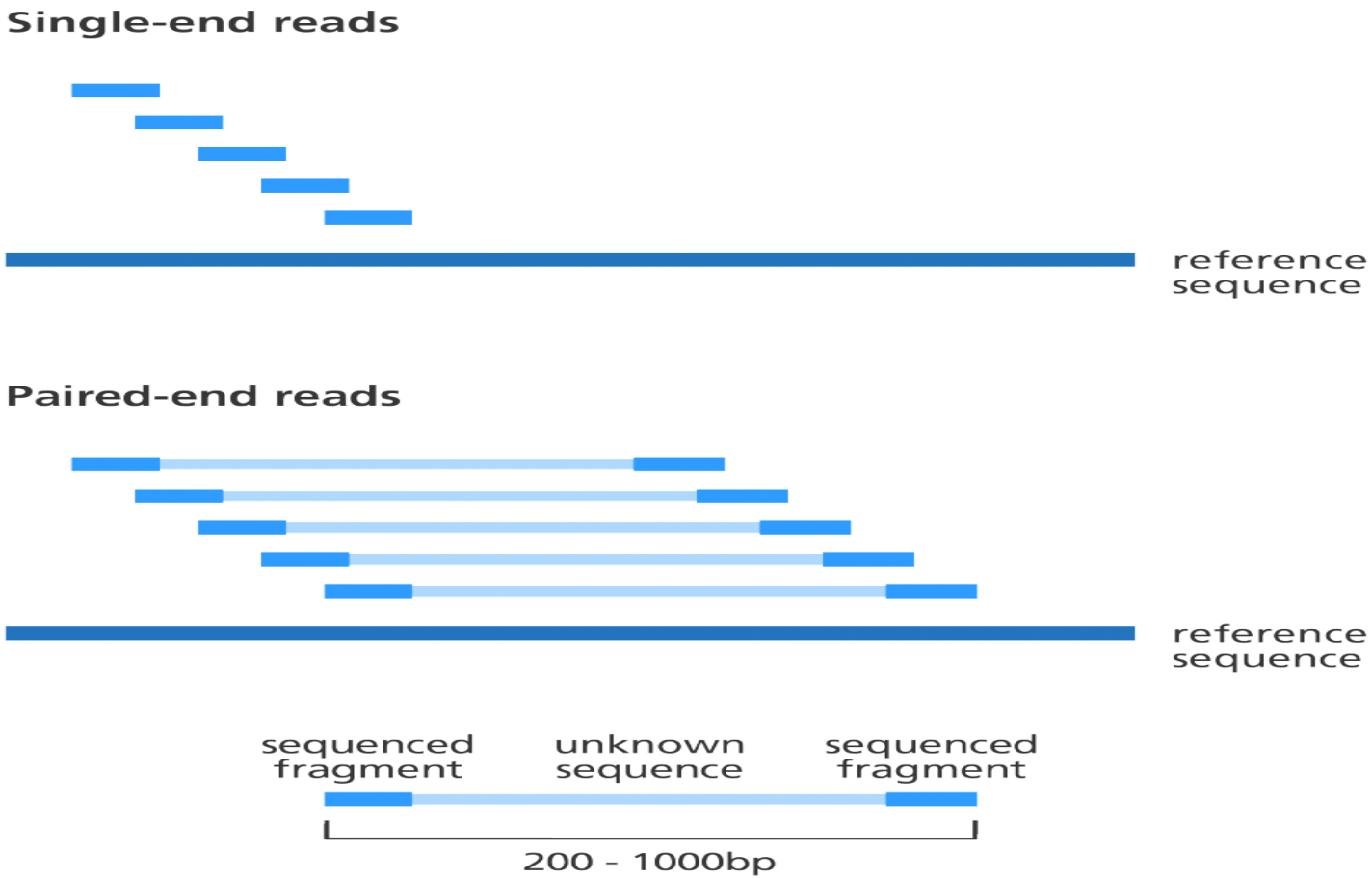


Image credit: https://www.biostars.org/p/267167/#267170

SEQUENCING SERVICES

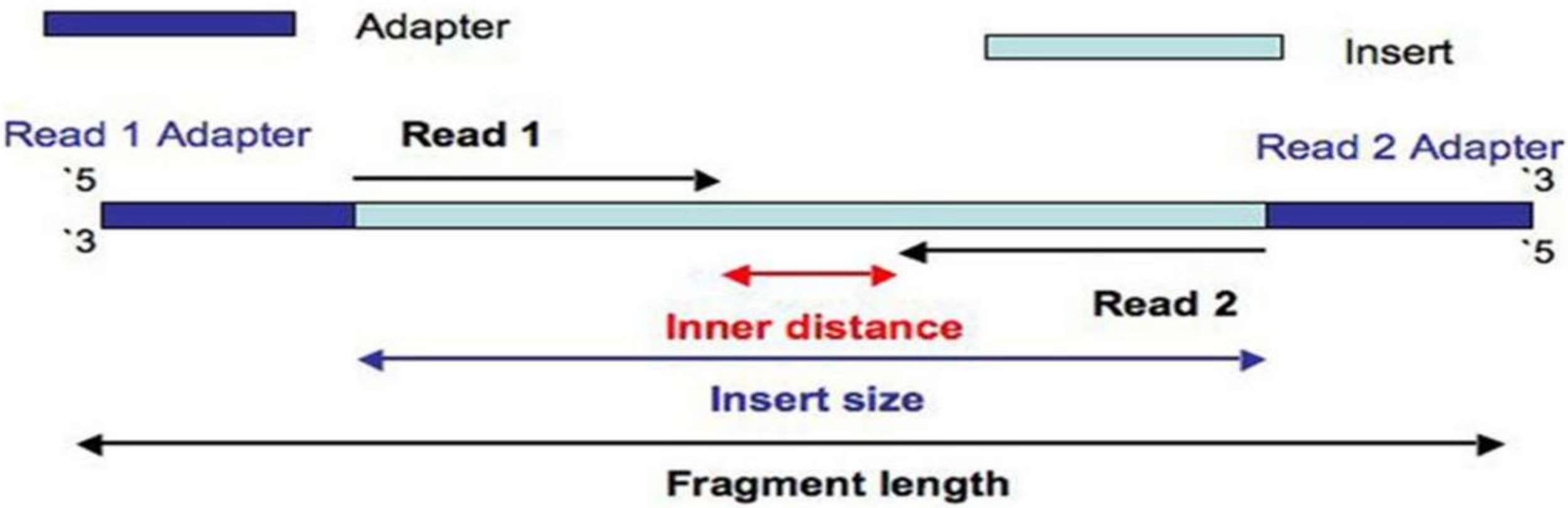


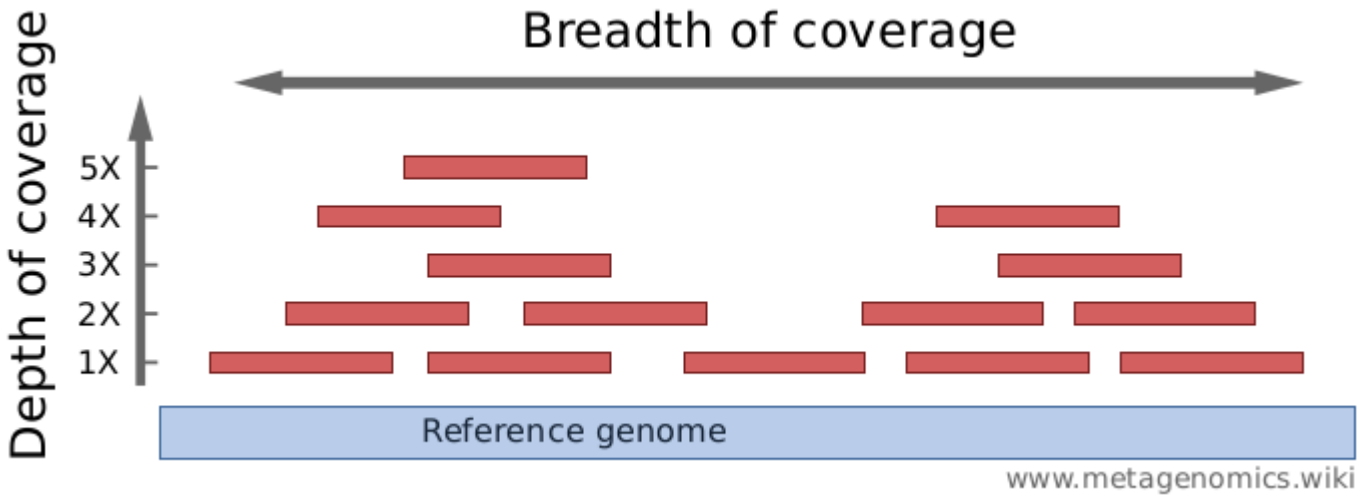
Image credit: https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/

COVERAGE DEPTH

- Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).
- coverage describes the average number of reads that align to, or "cover," known reference bases. The sequencing coverage level often determines whether variant discovery can be made with a certain degree of confidence at particular base positions.
- At higher levels of coverage, each base is covered by a greater number of aligned sequence reads, so base calls can be made with a higher degree of confidence.

https://www.ecseq.com/support/ngs/how-to-calculate-the-coverage-for-a-sequencing-experiment

COVERAGE DEPTH



Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: "90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth.“

COVERAGE DEPTH

Sequencing Method	Recommended Coverage
Whole genome sequencing (WGS)	30× to 50× for human WGS (depending on application and statistical model)
Whole-exome sequencing	100×
RNA sequencing	Usually calculated in terms of numbers of millions of reads to be sampled. Detecting rarely expressed genes often requires an increase in the depth of coverage.
ChIP-Seq	100×

Table Credit:
https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html

BASE QUALITIES

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
HHHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

Base quality is ASCII-encoded version of $Q = -10 \log_{10} p$

Usual ASCII encoding is "Phred+33":

take Q, rounded to integer, add 33, convert to character

Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

ASCII

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	¡	225	·
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	‚
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
4	<EOT>	36	$	68	D	100	d	132	Ñ	164	§	196	ƒ	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Ö	165	•	197	≈	229	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	Δ	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	…	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	´	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	¨	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Õ	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	–	240	
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	“	242	Ú
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	”	243	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	‘	244	Ù
21	<NAK>	53	5	85	U	117	u	149	ï	181	µ	213	’	245	ı
22	<SYN	54	6	86	V	118	v	150	ñ	182	∂	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	∑	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	∏	216	ÿ	248	¯
25	<EM>	57	9	89	Y	121	y	153	ô	185	π	217	Ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	∫	218	/	250	˙
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	°
28	<FS>	60	<	92	\	124	|	156	ú	188	º	220	‹	252	˚
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	›	253	˝
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	ﬁ	254	¸
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	ﬂ	255	ˇ

Example: Q=36.7

Phred+33= 37+33=70

= F

BASE QUALITIES

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
HHHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n

72-33=39

Slide Credit: Ben Langmead course of Algorithms for DNA Sequencing

GENOMIC DATA
(A READ IN FASTQ)

PHRED Score	Probability of Incorrect Base Call	Accuracy of Base Call
0	1 in 1	0%
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

- 10 corresponds to 10% error (1/10),
- 20 corresponds to 1% error (1/100),
- 30 corresponds to 0.1% error (1/1,000) and
- 40 corresponds to one error every 10,000 measurements (1/10,000) that is an error rate of 0.01%.

https://www.youtube.com/playlist?list=PL2mpR0RYFQsBiCWVJSvVAO3OJ2t7DzoHA

GENOMIC DATA
(A DATA IN FASTA)

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

Thank you!