Genômica Computacional

Introdução em Genômica e Sequenciamento

Professor: Ricardo A. Vialle

CS31 - Genômica Computacional [11,18,25/10 e 1,8/11/23 - 12h00-14h00 - 4ªs. feiras]

11 de Outubro de 2023

Cronograma

Data	Tema
11-Oct	Introdução a Genômica, Sequenciamento (teórica)
18-Oct	Bioinformática - Linux - Processamento de dados de sequenciamento (teórico-prática)
25-Oct	Montagem de genomas (teórico-prática)
1-Nov	Anotação de genomas (teórico-prática)
8-Nov	Analise de variabilidade genética (teórico-prática)

Avaliação

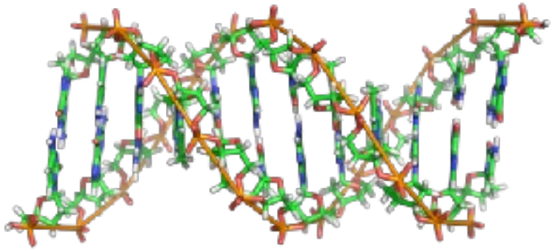
Nota final será determinada de acordo com presença

Genome

“The complete set of genes or genetic material present in a cell or organism.”

Oxford dictionaries

“Blueprint” or “recipe” of life



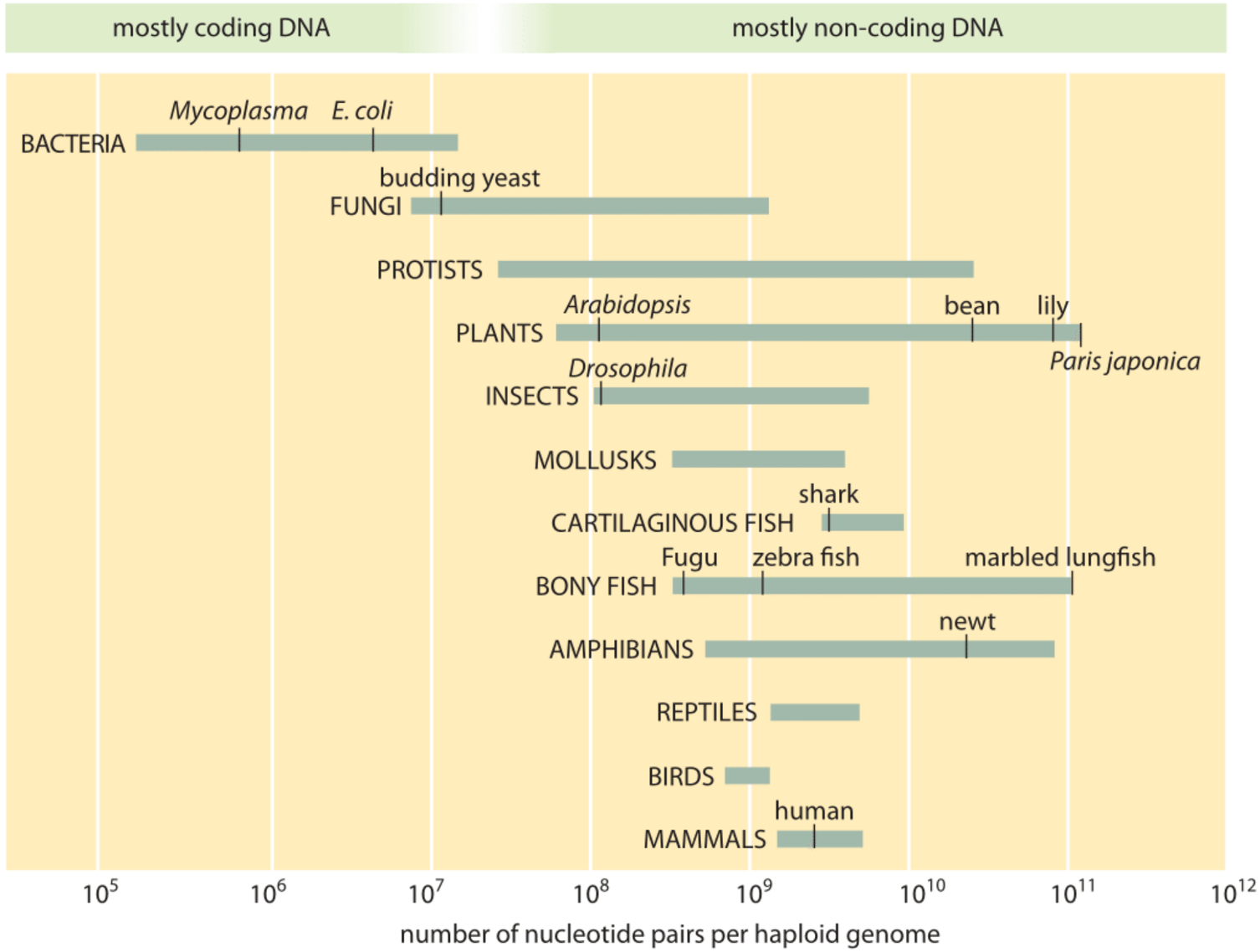
TAGCCCGACTTG

Self-copying store of read-only information about how to develop and maintain an organism

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Genome

http://book.bionumbers.org/how-big-are-genomes/



Genomics and other -omics

The branches of science known informally as **omics** are various disciplines in biology whose names end in the suffix *-omics*

Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.

Traditional Terms

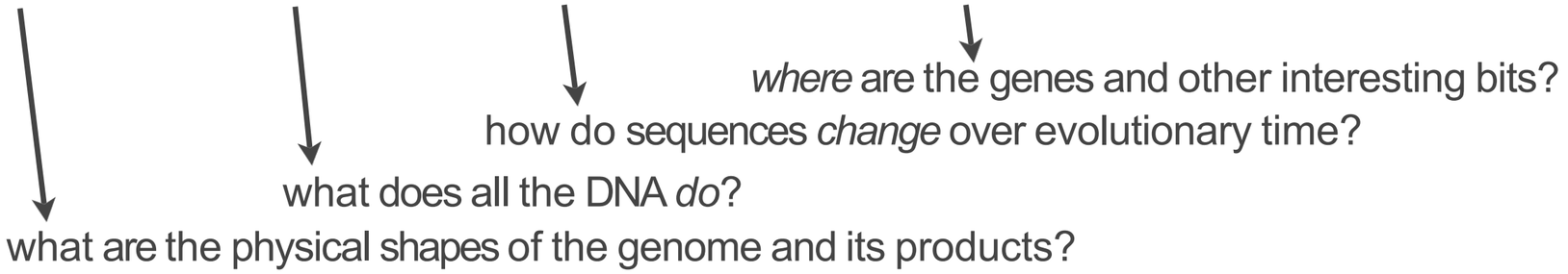
Term	***Definition: "the study of .."***
Genomics	All of the genes in the genome; a general term for all studies involving many genes and their products
Transcriptomics	The expressed mRNAs
Proteomics	The proteins in the organism
Metabolomics	The metabolites that result from the expression of proteins and genes

Adapted from: Phil McClean Lab
https://www.ndsu.edu/pubweb/~mcclean/plsc411/index.htm

Genomics

Oxford dictionaries

"The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping** of **genomes**."



Collins English Dictionary

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Genomics: tools for basic science

> "The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping** of genomes." Oxford dictionaries

Structure / mapping

What is the DNA sequence of the genome?

Where are the genes?

What is the genome's three-dimensional shape in the cell?

Function

What does all the DNA in the genome do?

What genes interact with what other genes? How does the cell know what DNA is on/off?

Evolution

How did history shape our ethnicities and populations?

What big events shaped our current genetics?

Which portions of the genome are conserved by evolution?

Slide from: Bem Langmead Lab

https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Genomics: tools for medicine

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

Collins English Dictionary

How is genotype related to health phenotypes?

What's the difference between DNA in a tumor vs DNA in healthy tissue?

Can genomic data help predict what drugs might be appropriate for:

- a particular cancer patient?
- a particular genetic disorder?

Can genomic data reveal weaknesses in the defenses of pathogens?

Can genomic data help us predict what flu strains will prevail next year?

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Computational Genomics

Addresses crucial problems at the intersection of genomics and computer science

The intersection:

Key biological models are straight out of computer science: **circuits** and **networks** for molecular interactions, **trees** for evolution and pedigrees, **strings** for DNA, RNA and proteins

Thanks to sequencers and microarrays, research bottlenecks increasingly hinge on computational issues: **speed**, **scalability**, **energy**, **cost**

With large, noisy, biased high-throughput datasets comes a critical need for **machine learning** and **statistical reasoning**

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Computational Genomics: Computation

How to efficiently analyze the huge quantities of fragmentary evidence that come from DNA sequencers

How to model biological phenomena and make predictions

How to combine data from disparate datasets to reach new conclusions in the presence of error and systematic bias

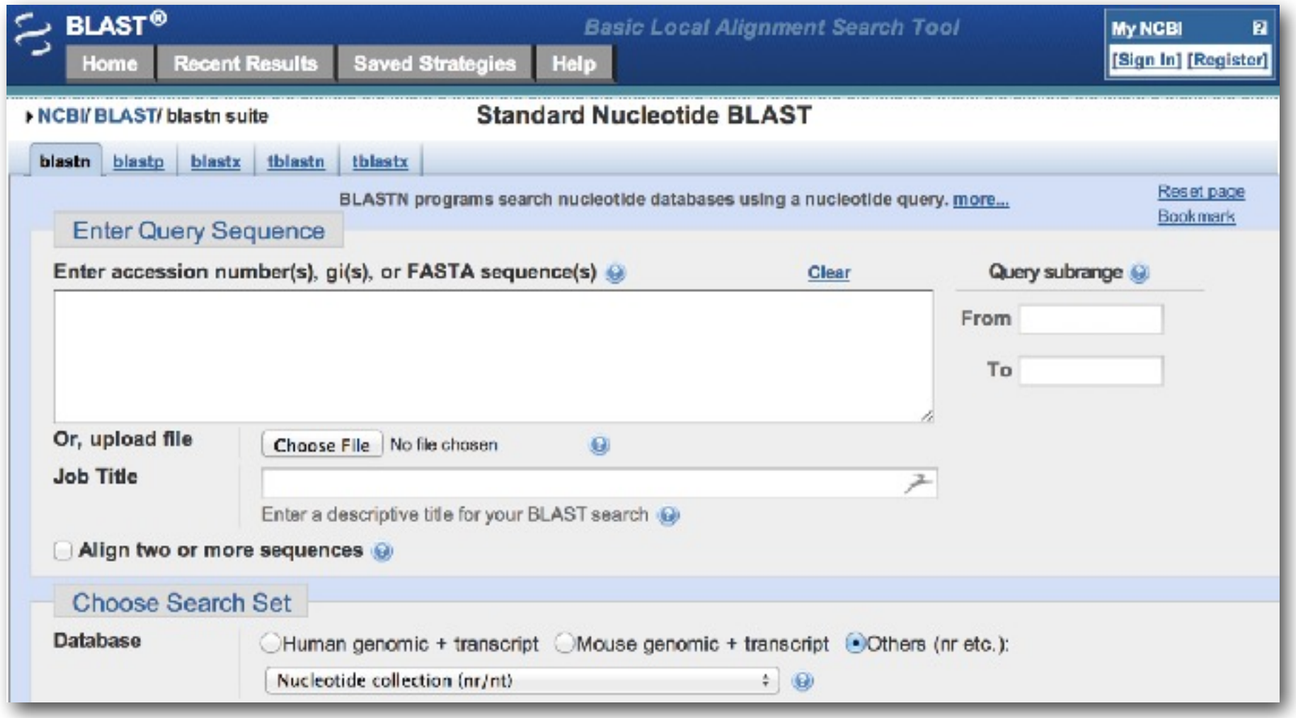
How to store huge quantities of data economically and securely while also allowing it to be queried

How to visualize large, complicated datasets

Draws on: Algorithms, data structures, pattern matching, indexing, compression, information retrieval, distributed and parallel computing, cloud computing, machine learning, ...

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Computational Genomics: success stories



http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

The BLAST sequence alignment program is a hugely successful tool, a fixture of biological analysis and cited over 50,000 times

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Computational Genomics: success stories

The Human Genome Project depended crucially on contributions by computer scientists, especially new methods for assembling DNA fragments into chromosomes.

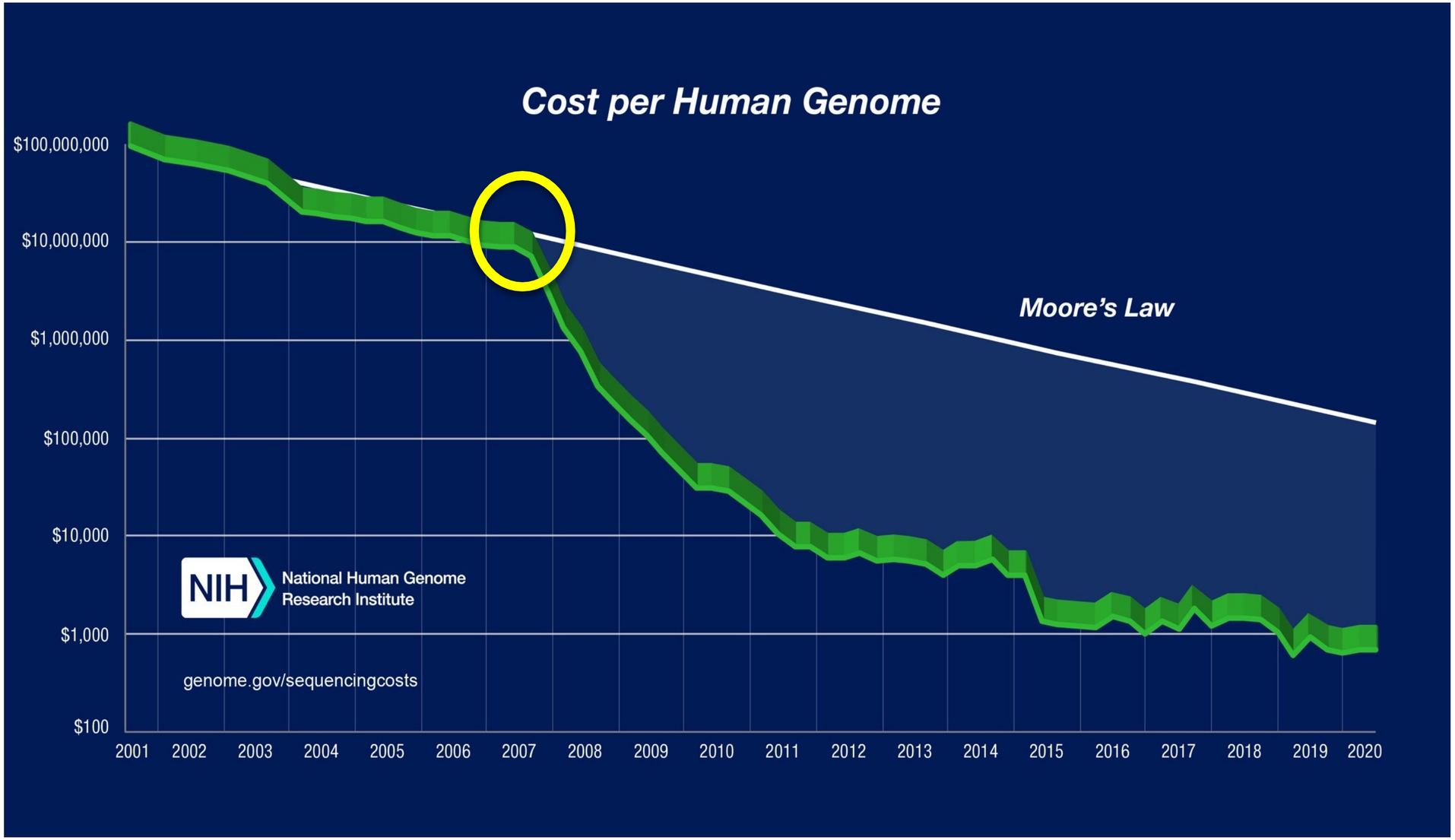
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Computational Genomics: success stories

The idea of using high-throughput DNA sequencing in medical settings is only possible because of novel, extremely efficient software developed in the years after second-generation sequencers arrived.

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

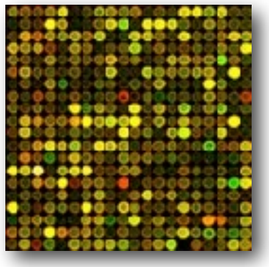
Genomics: shaped by technology



Genomics: shaped by technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

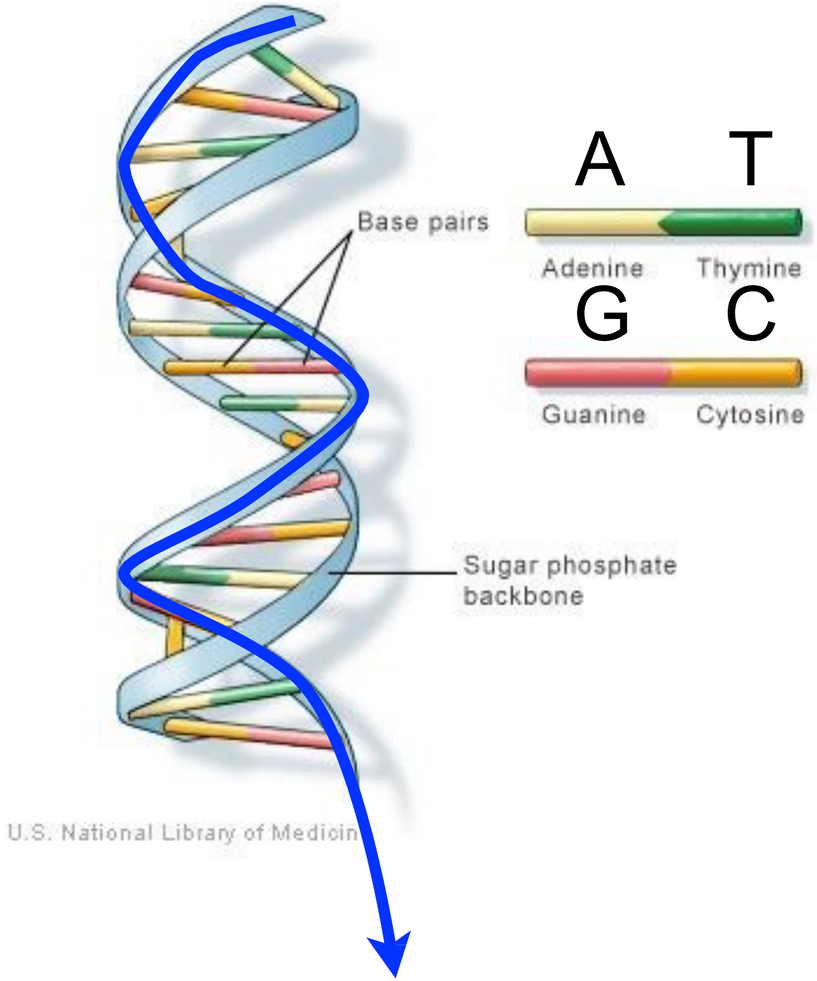
Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010

These provide very high-resolution snapshots of the world of nucleic acids (not just DNA)

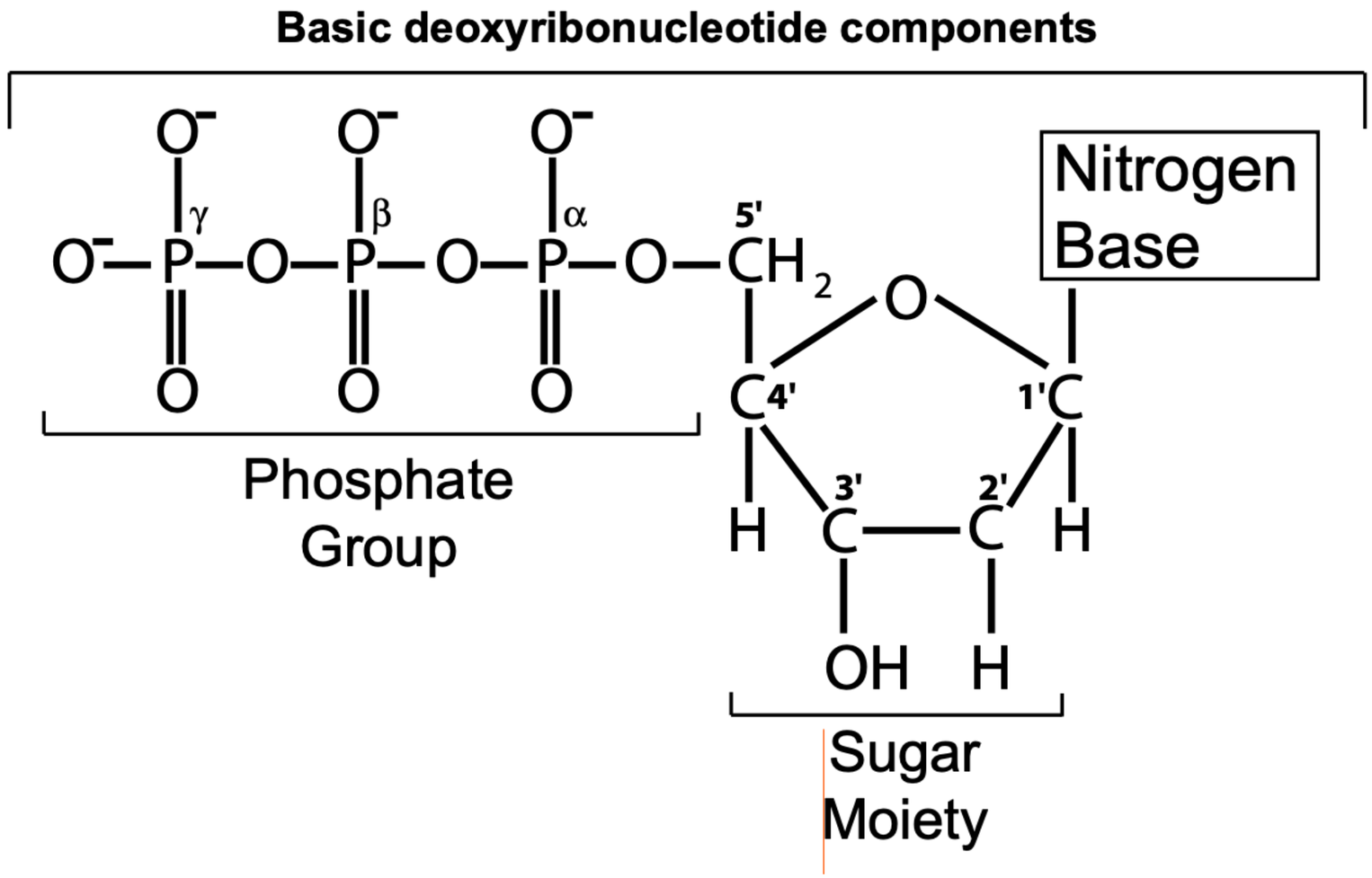
DNA



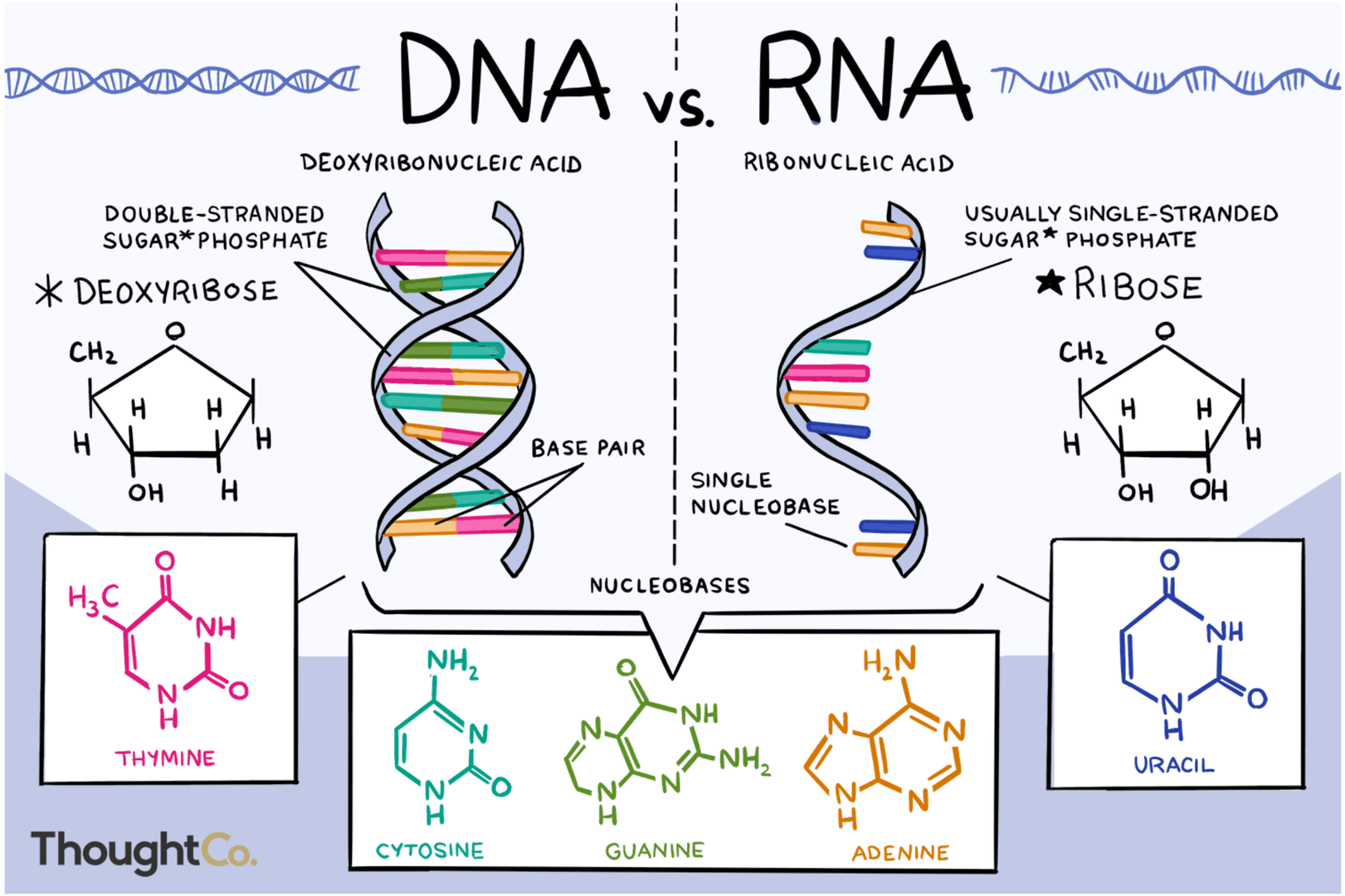
TCACACTGAGCGTGCTG

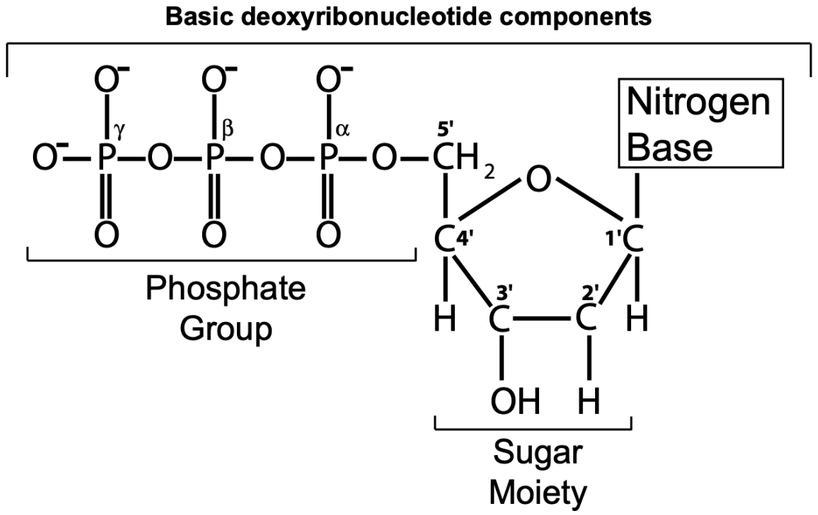
http://ghr.nlm.nih.gov/handbook/basics/dna

DNA

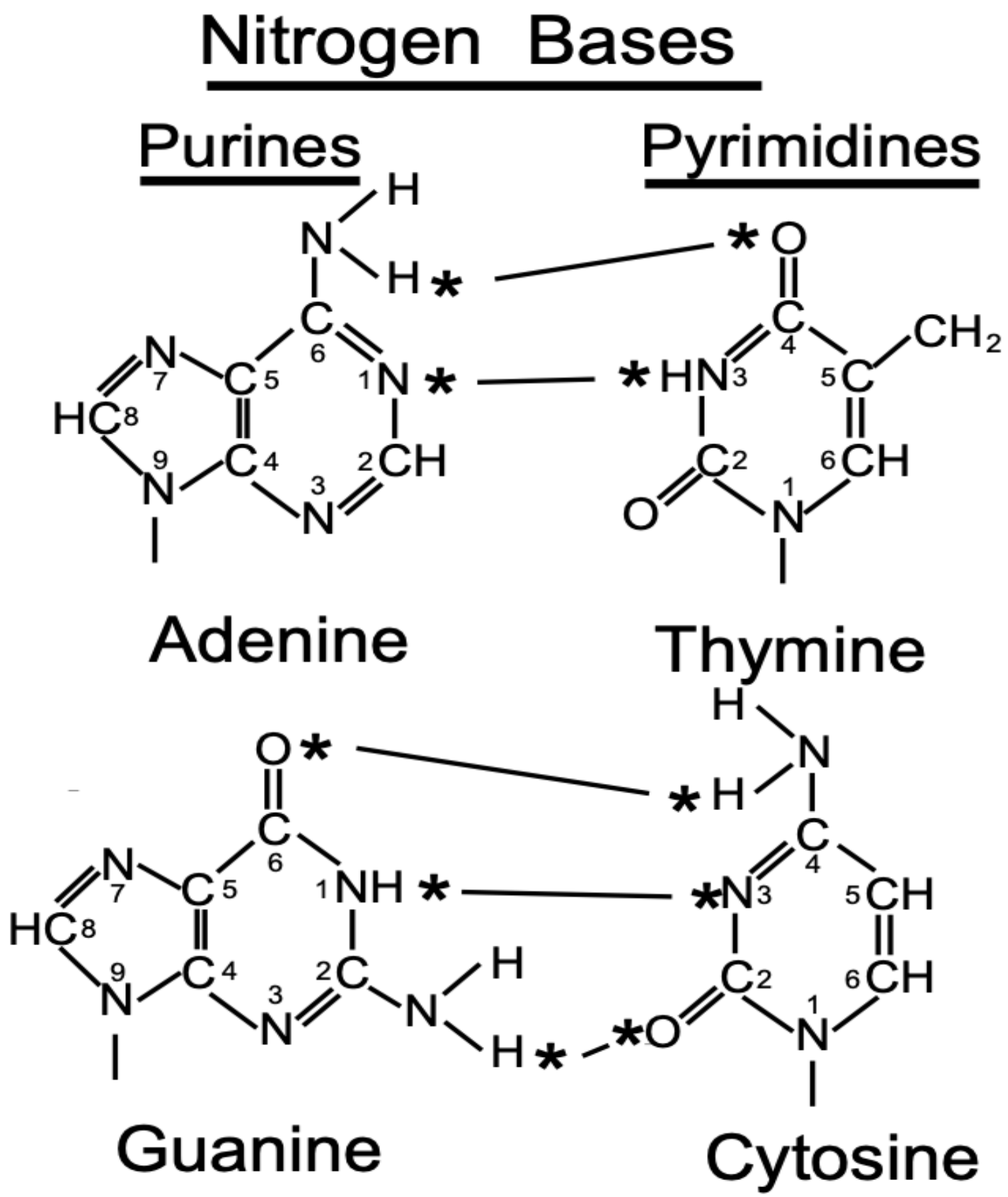

Basic deoxyribonucleotide components
Nitrogen Base
Phosphate Group
Sugar Moiety
γ
β
α
5'
4'
1'
3'
2'
OH
H

DNA


DNA vs. RNA
DEOXYRIBONUCLEIC ACID
RIBONUCLEIC ACID
DOUBLE-STRANDED SUGAR*PHOSPHATE
* DEOXYRIBOSE
BASE PAIR
USUALLY SINGLE-STRANDED SUGAR* PHOSPHATE
★ RIBOSE
SINGLE NUCLEOBASE
NUCLEOBASES
THYMINE
CYTOSINE
GUANINE
ADENINE
URACIL
ThoughtCo.

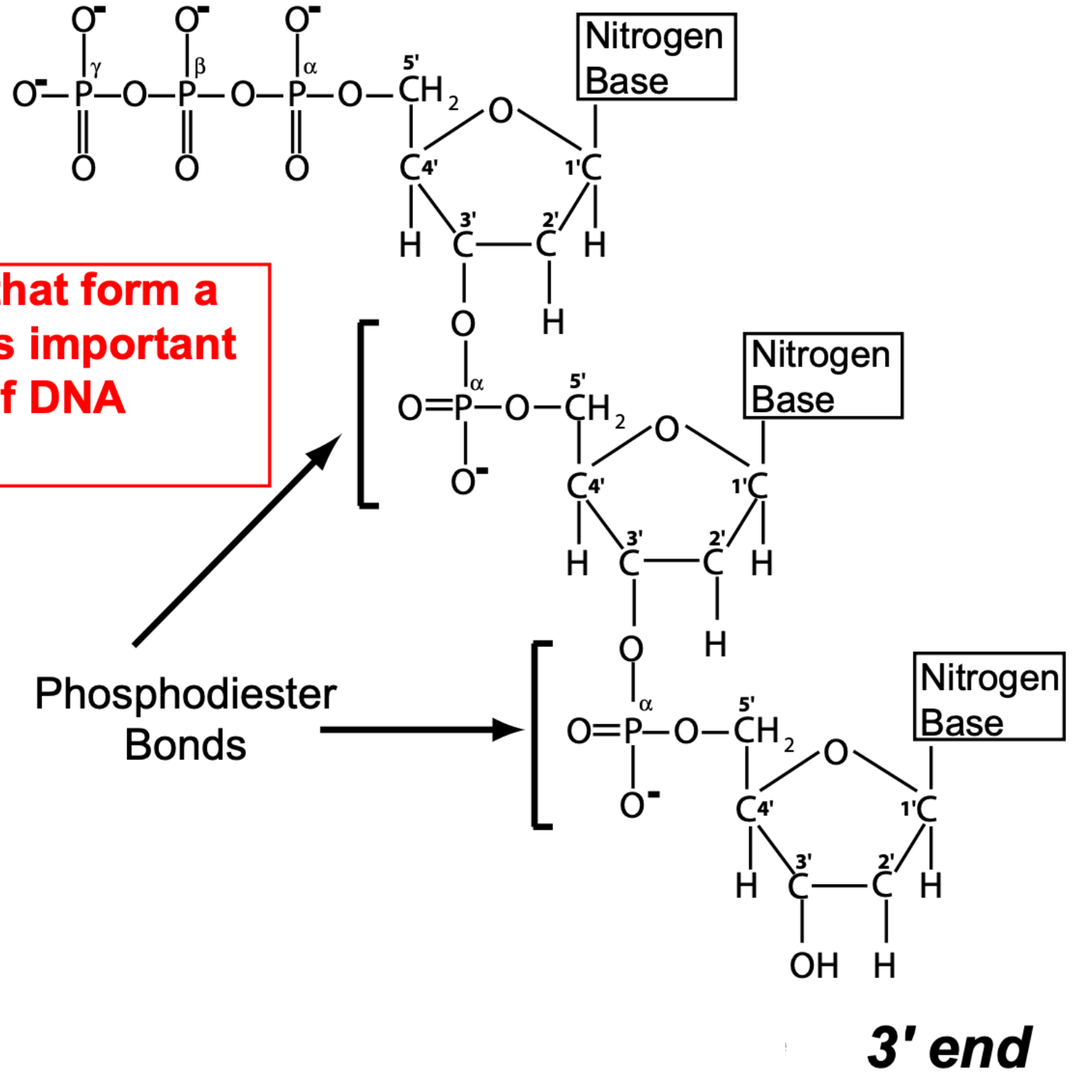

Basic deoxyribonucleotide components
Nitrogen Base
Phosphate Group
Sugar Moiety

DNA


Nitrogen Bases
Purines
Pyrimidines
Adenine
Thymine
Guanine
Cytosine

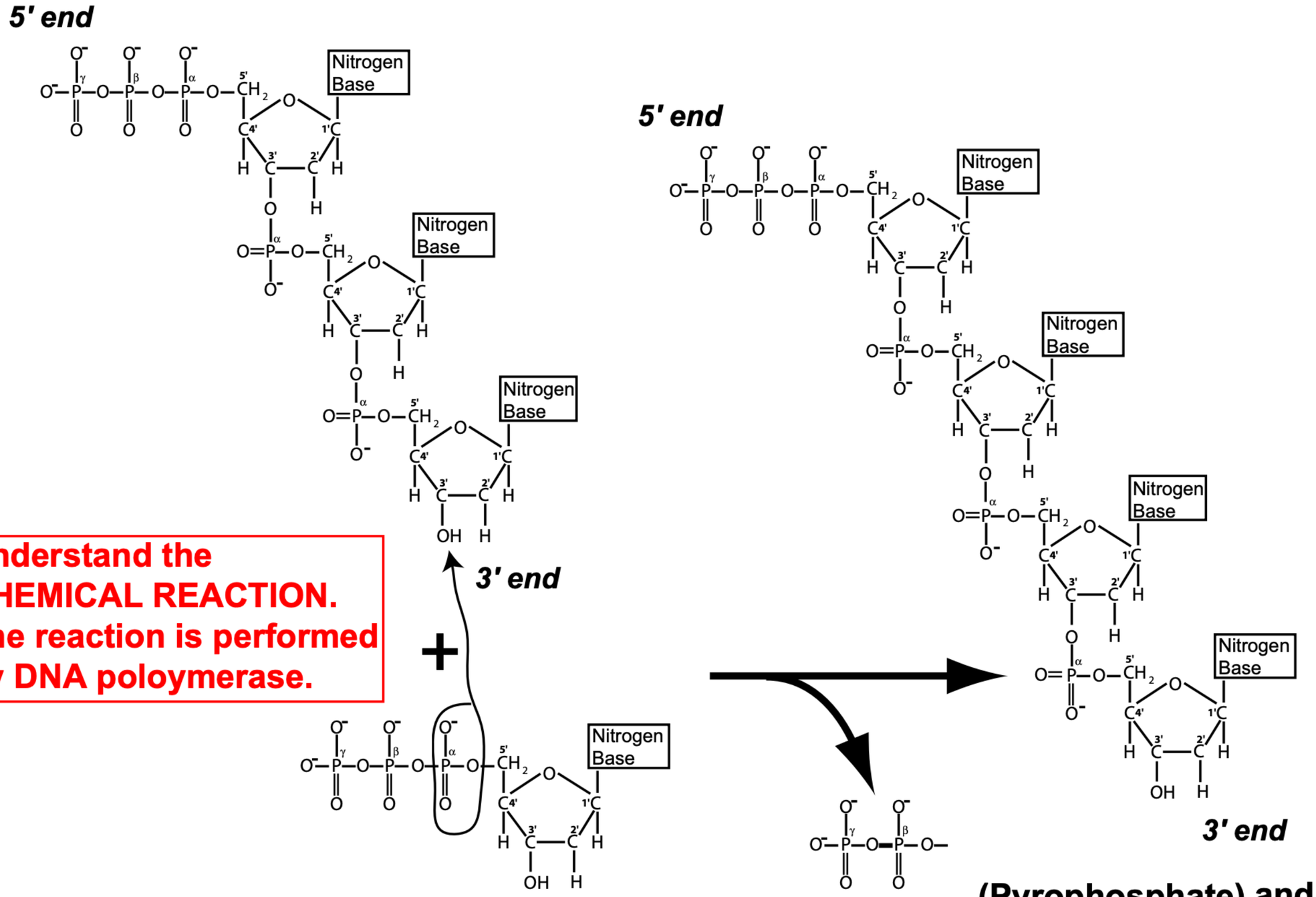
DNA

5' end



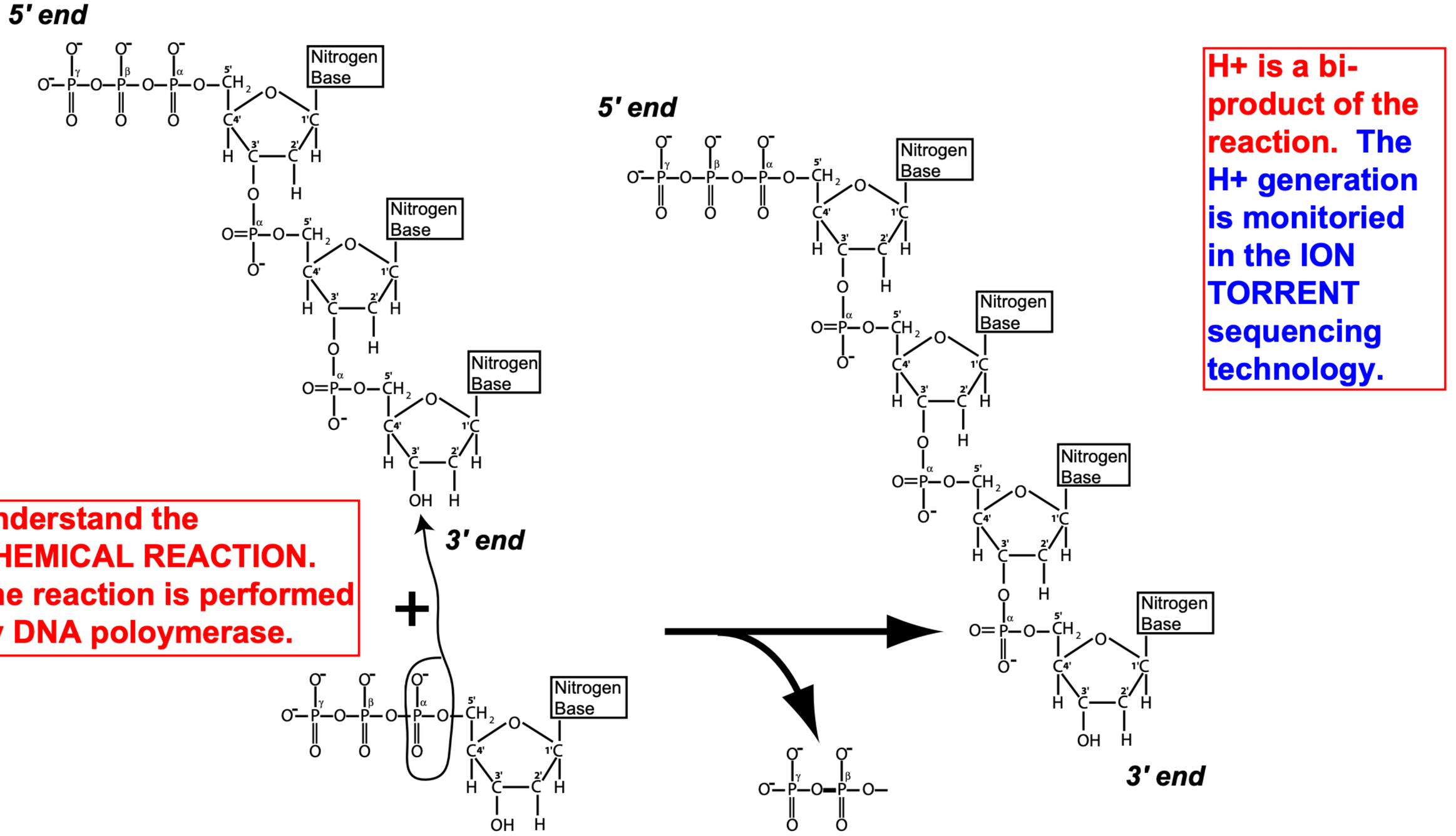
3' end

Know the BONDS that form a DNA molecule. It is important for the principles of DNA sequencing.



Understand the CHEMICAL REACTION. The reaction is performed by DNA poloymerase.

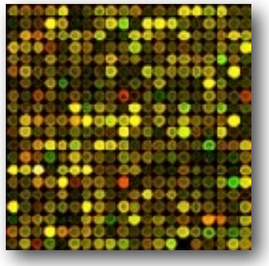
(Pyrophosphate) and H+ molecule


5' end
Nitrogen Base
3' end
+
5' end
Nitrogen Base
3' end
Understand the CHEMICAL REACTION. The reaction is performed by DNA poloymerase.
H+ is a bi-product of the reaction. The H+ generation is monitoried in the ION TORRENT sequencing technology.

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

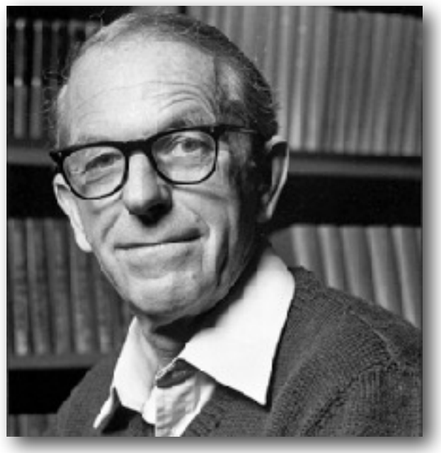
Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

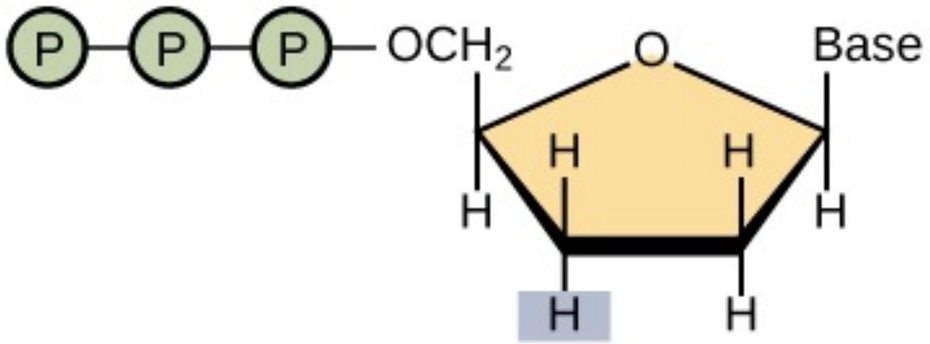
Since ~2010



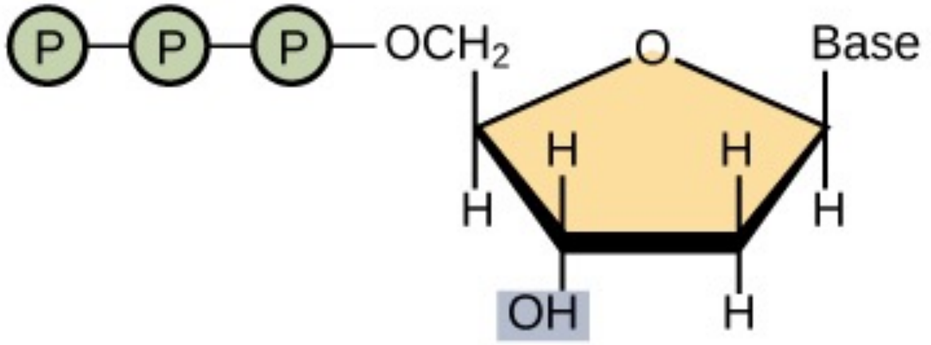
Fred Sanger
1918-2013

"Chain termination" sequencing

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

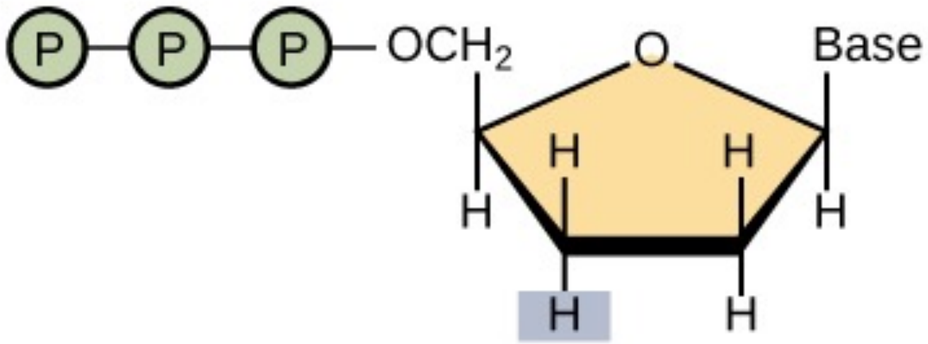


Dideoxynucleotide (ddNTP)

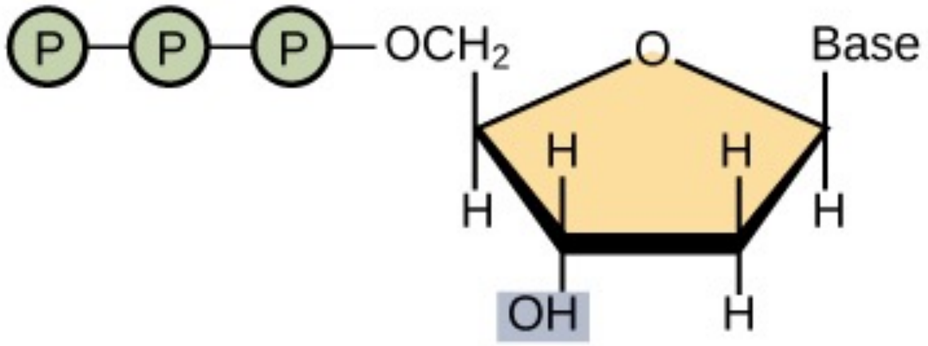


Deoxynucleotide (dNTP)

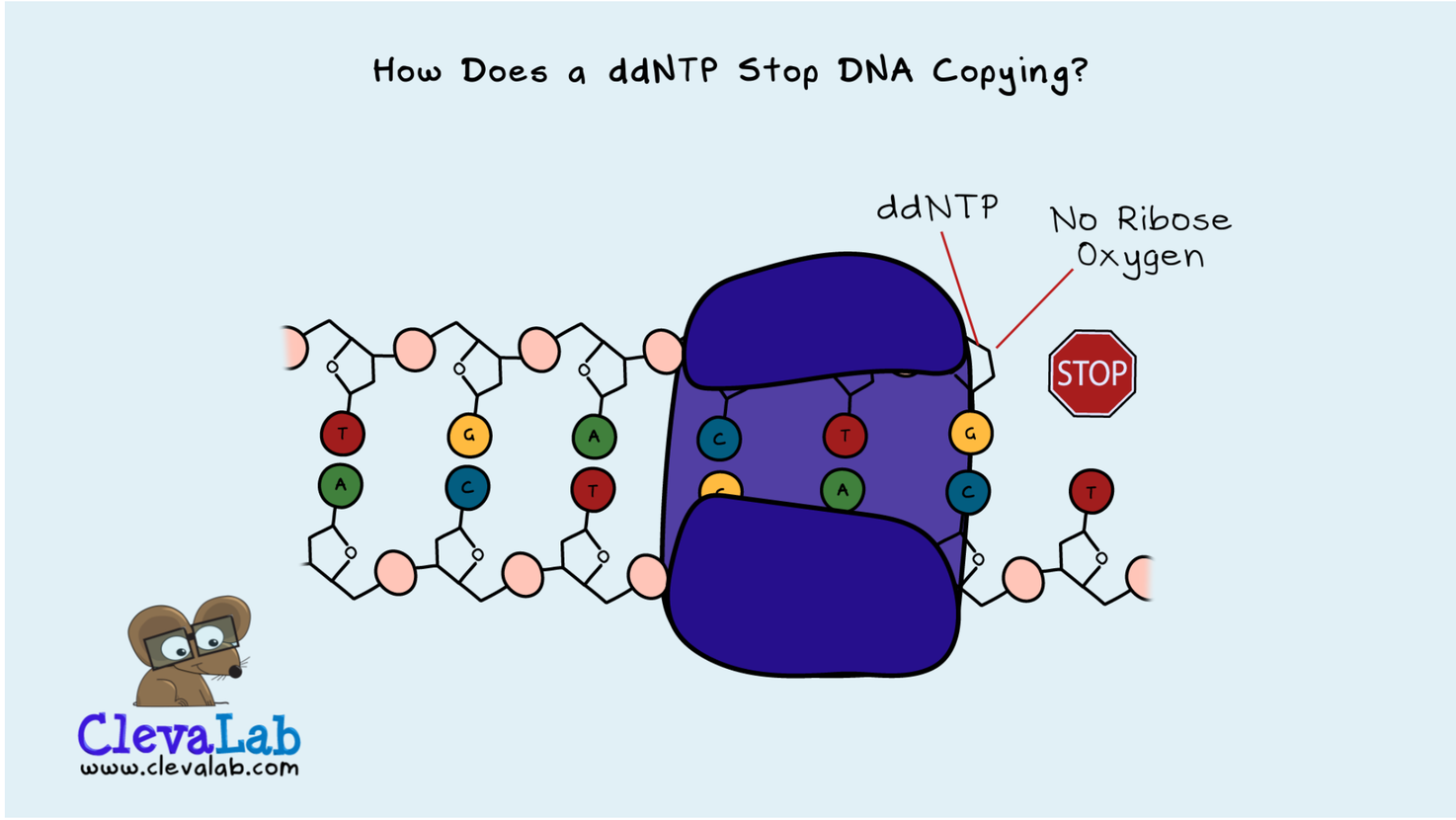
https://courses.lumenlearning.com/trident-boundless-microbiology/chapter/bioinformatics/



Dideoxynucleotide (ddNTP)



Deoxynucleotide (dNTP)



https://courses.lumenlearning.com/trident-boundless-microbiology/chapter/bioinformatics/

Sequencing of DNA by the Sanger method

https://onlinesciencenotes.com/dna-sequencing-by-sangers-method/

Sequencing of DNA by the Sanger method

Single-stranded DNA with unknown sequence (blue) serves as a template

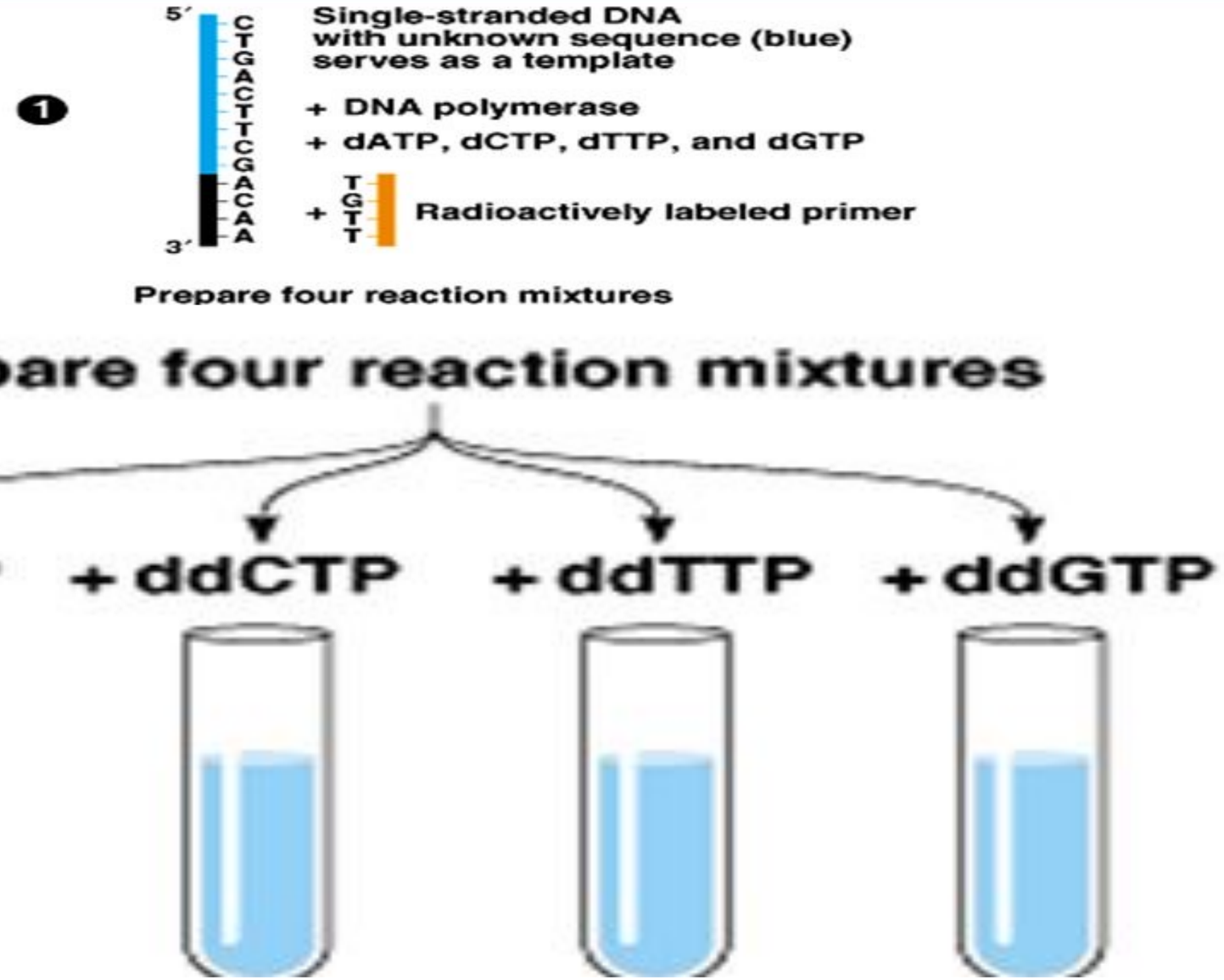
\+ DNA polymerase

\+ dATP, dCTP, dTTP, and dGTP

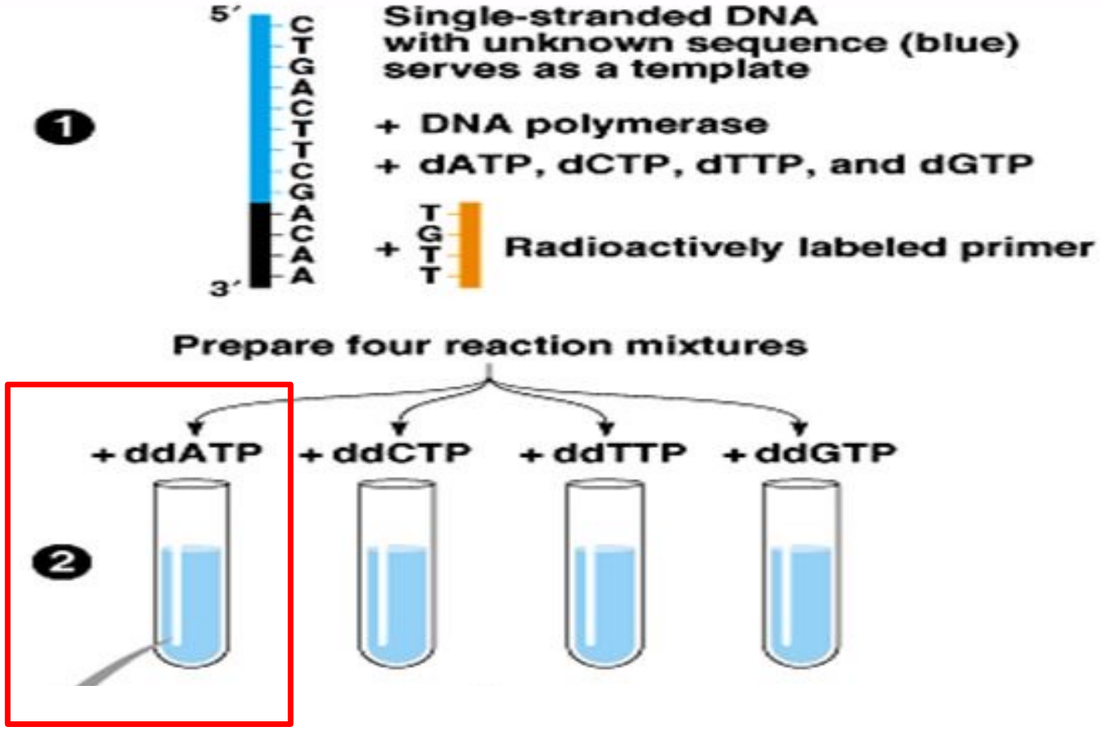
Radioactively labeled primer

https://onlinesciencenotes.com/dna-sequencing-by-sangers-method/

Sequencing of DNA by the Sanger method

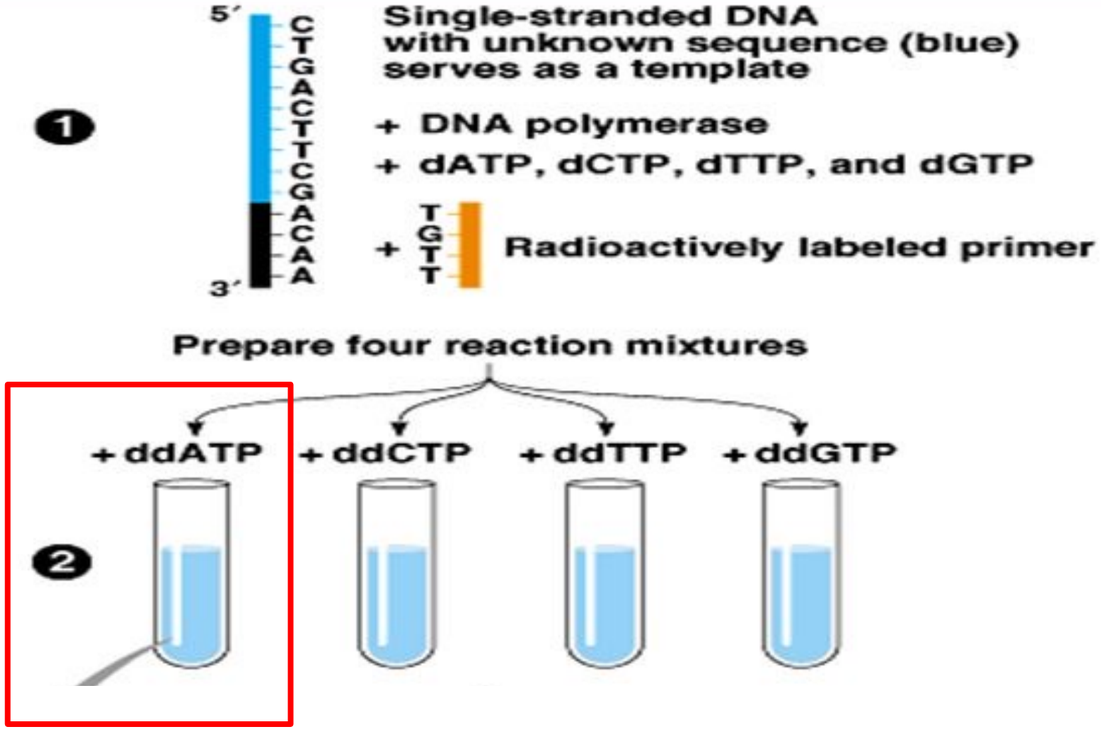


https://onlinesciencenotes.com/dna-sequencing-by-sangers-method/



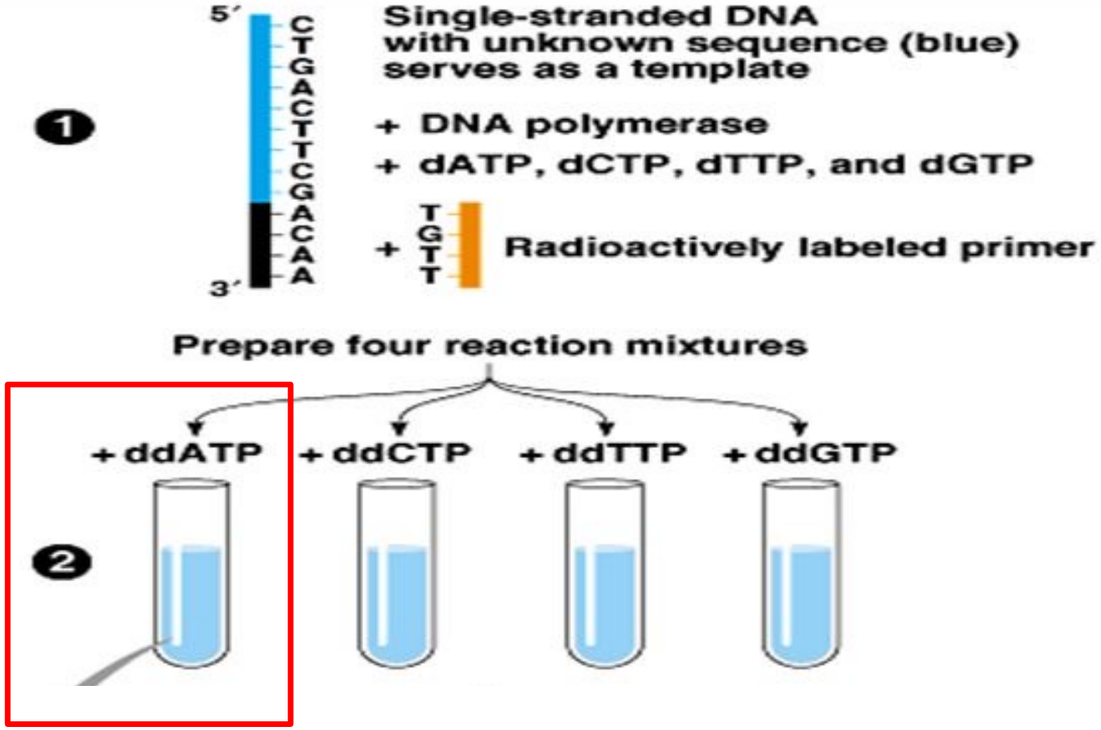
AACAGCTTCAGTC

TTGT



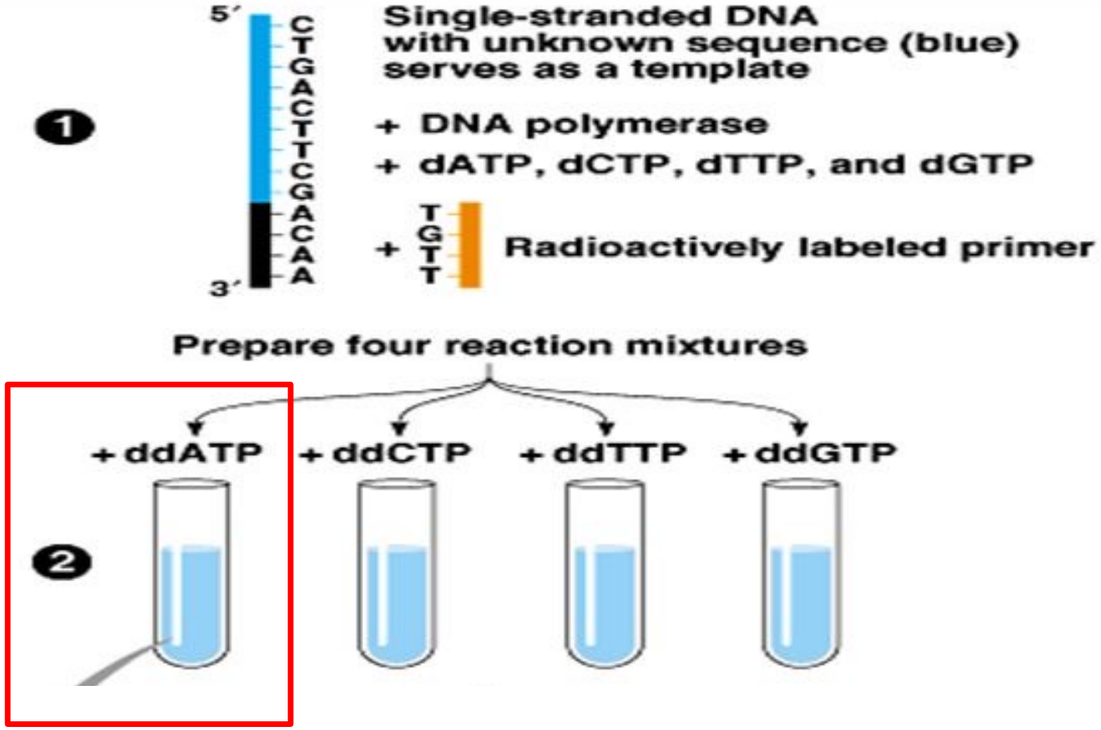
AACAGCTTCAGTC

TTGTC



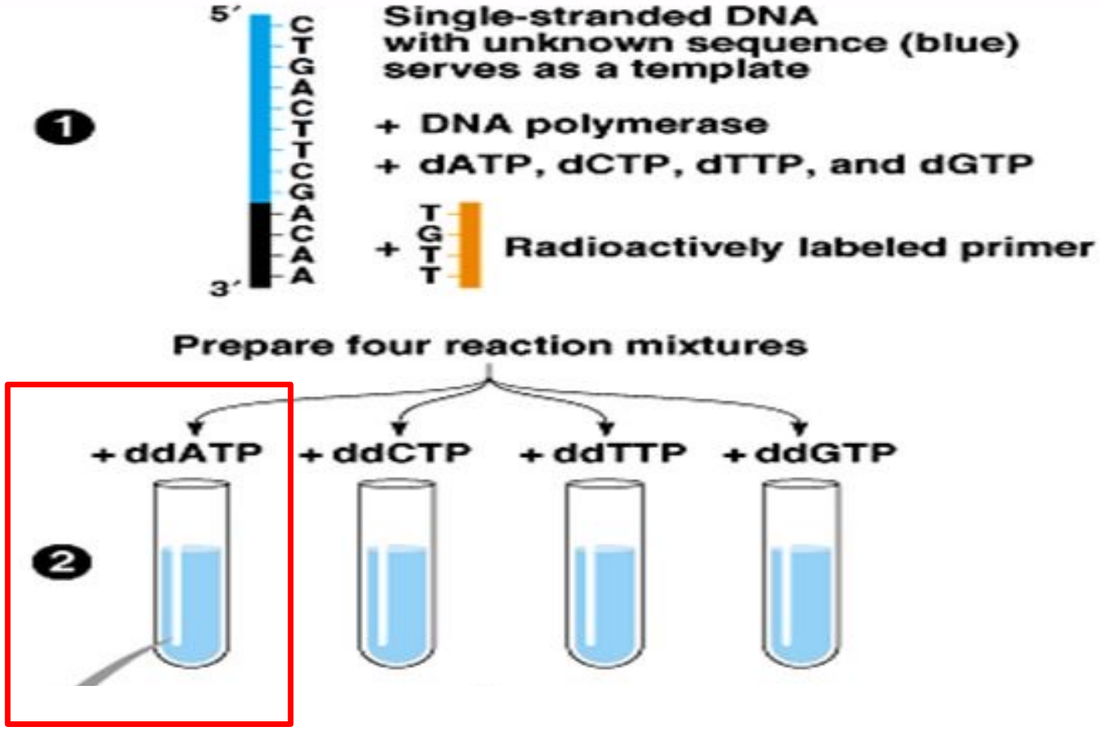
AACAGCTTCAGTC

TTGTCG



AACAGCTTCAGTC

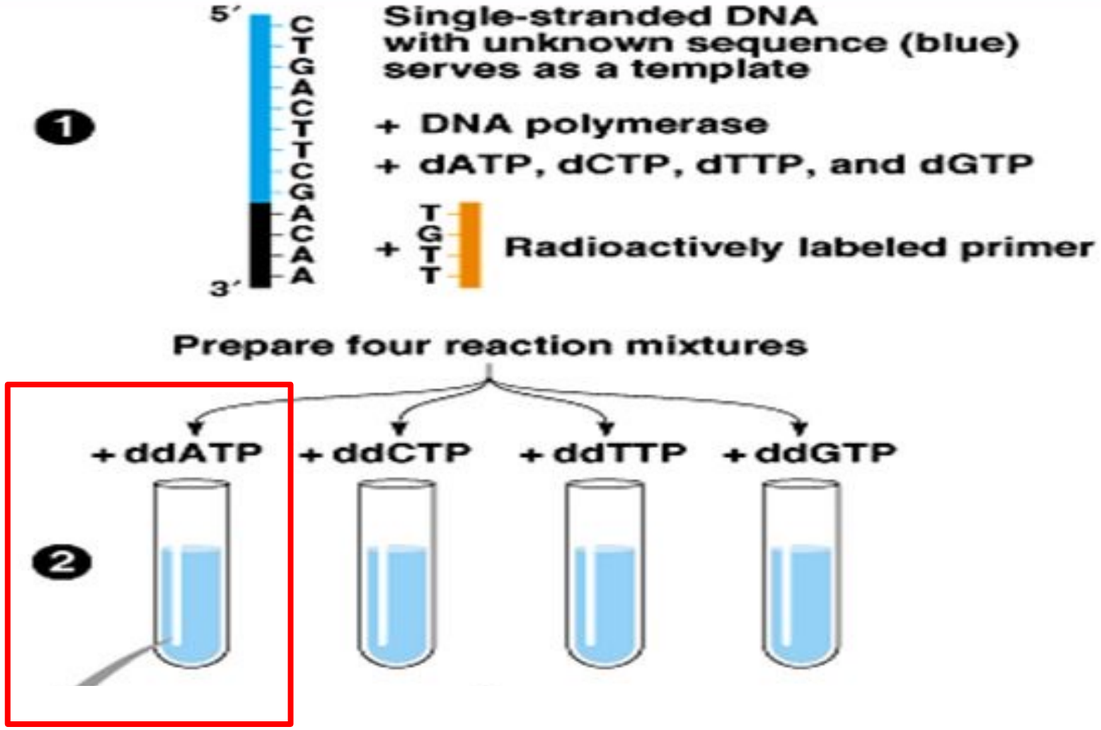
TTGTCGA



AACAGCTTCAGTC

TTGTCGA

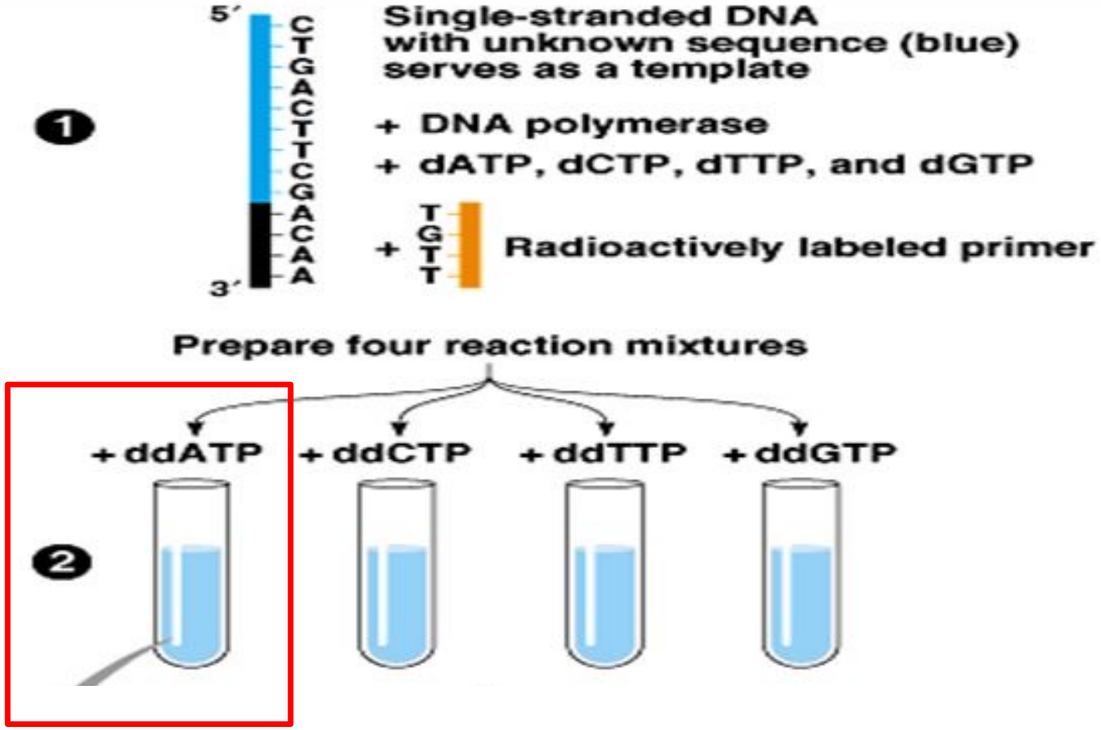
TTGT



AACAGCTTCAGTC

TTGTCGA

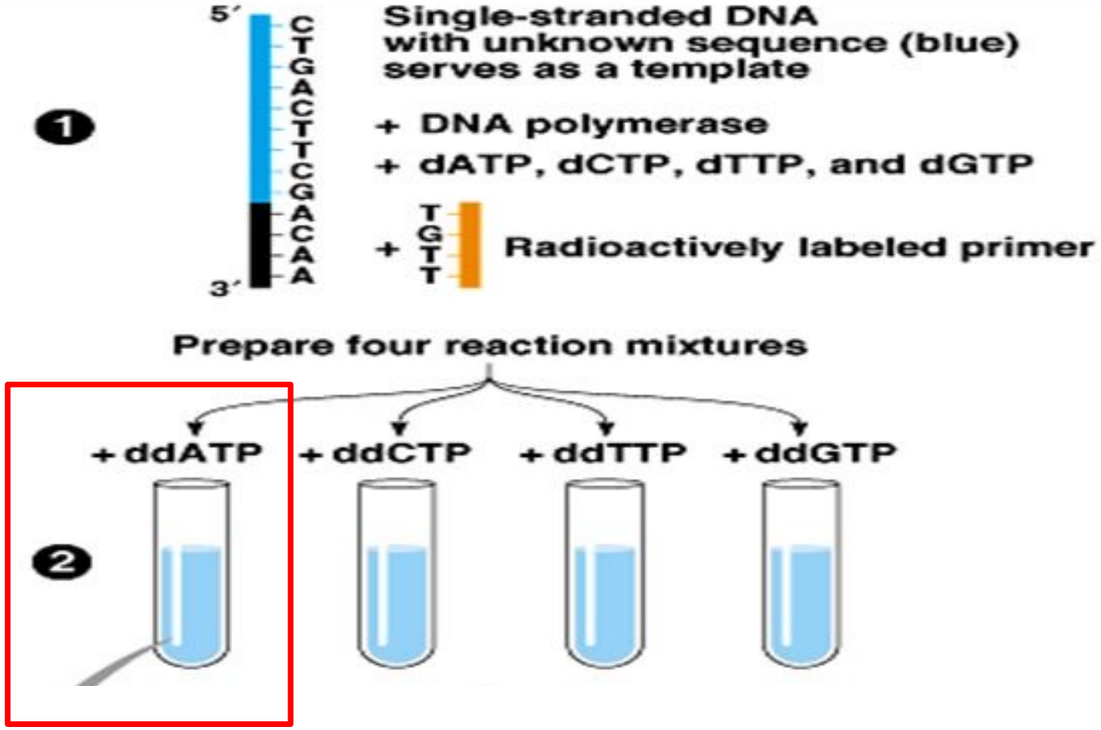
TTGTC



AACAGCTTCAGTC

TTGTCGA

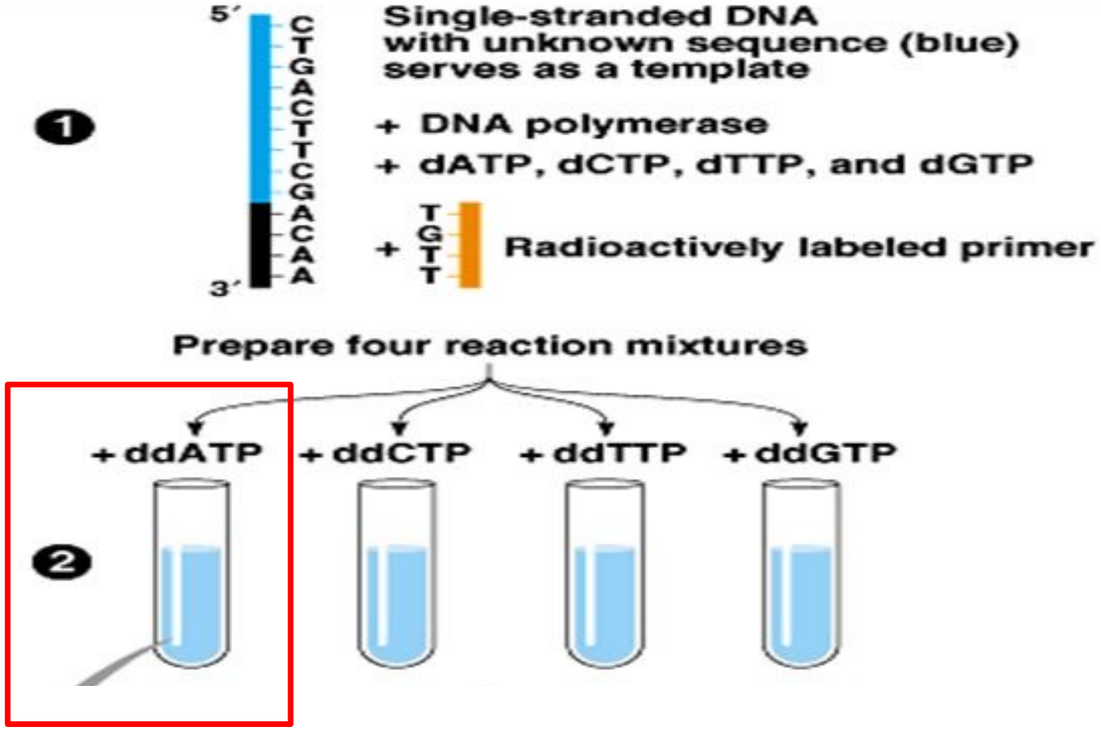
TTGTCG



AACAGCTTCAGTC

TTGTCGA

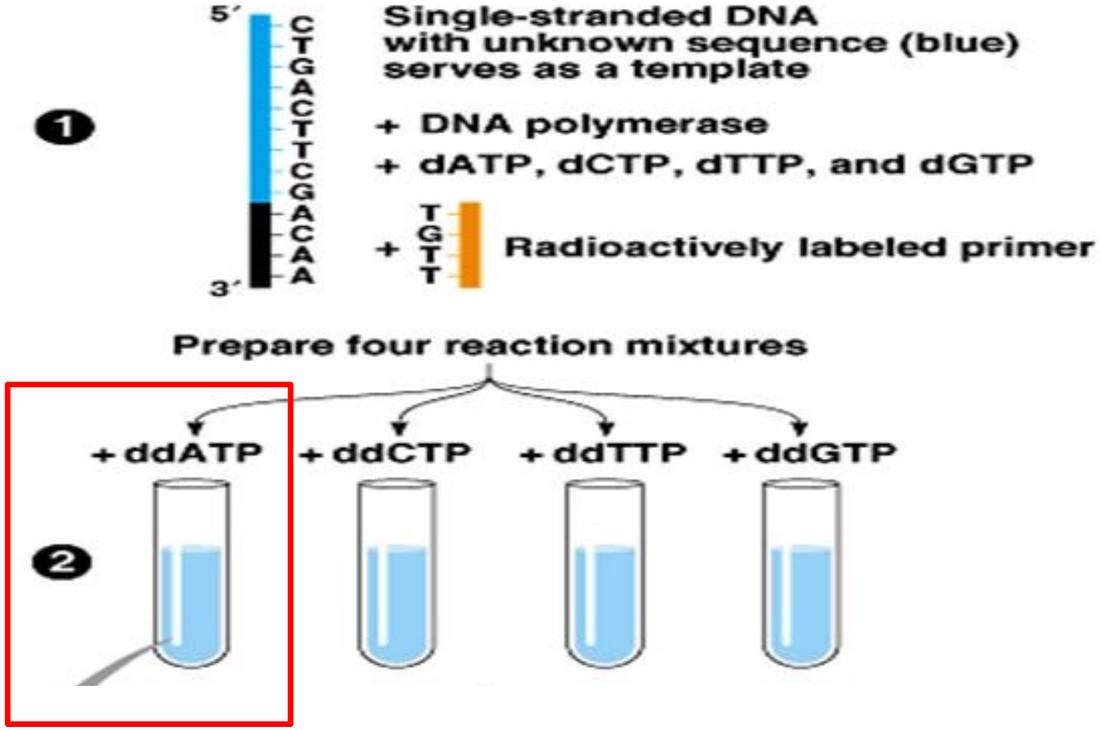
TTGTCGA



AACAGCTTCAGTC

TTGTCGA

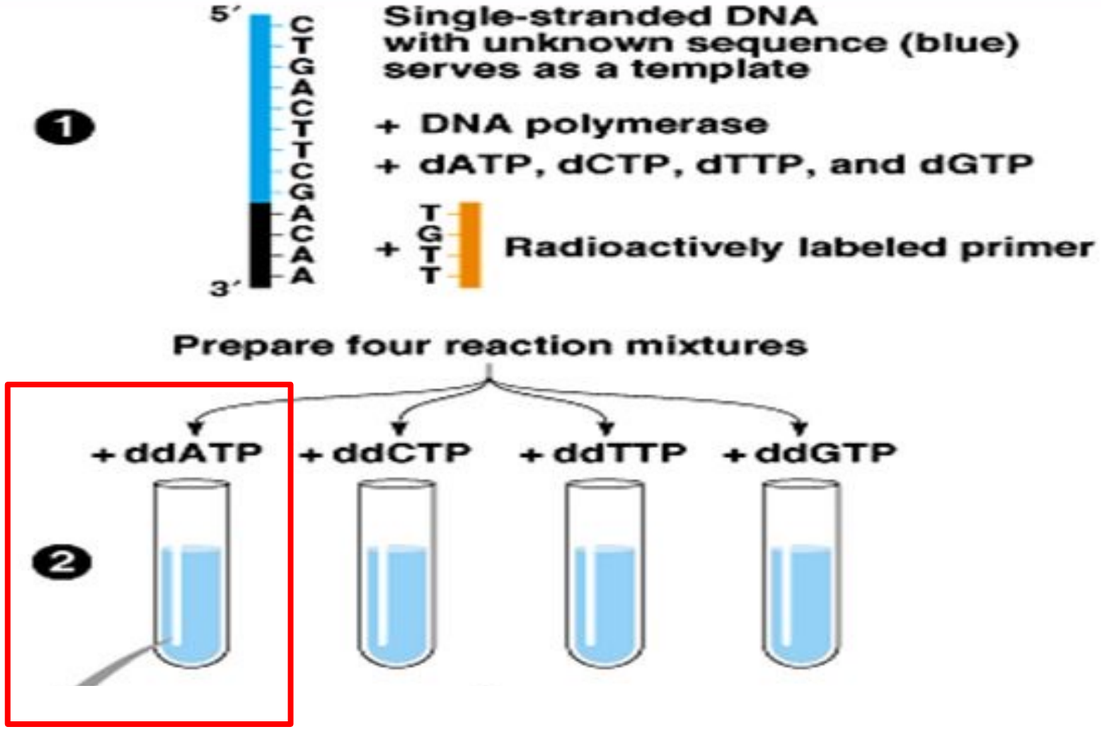
TTGTCGAA



AACAGCTTCAGTC

TTGTCGA

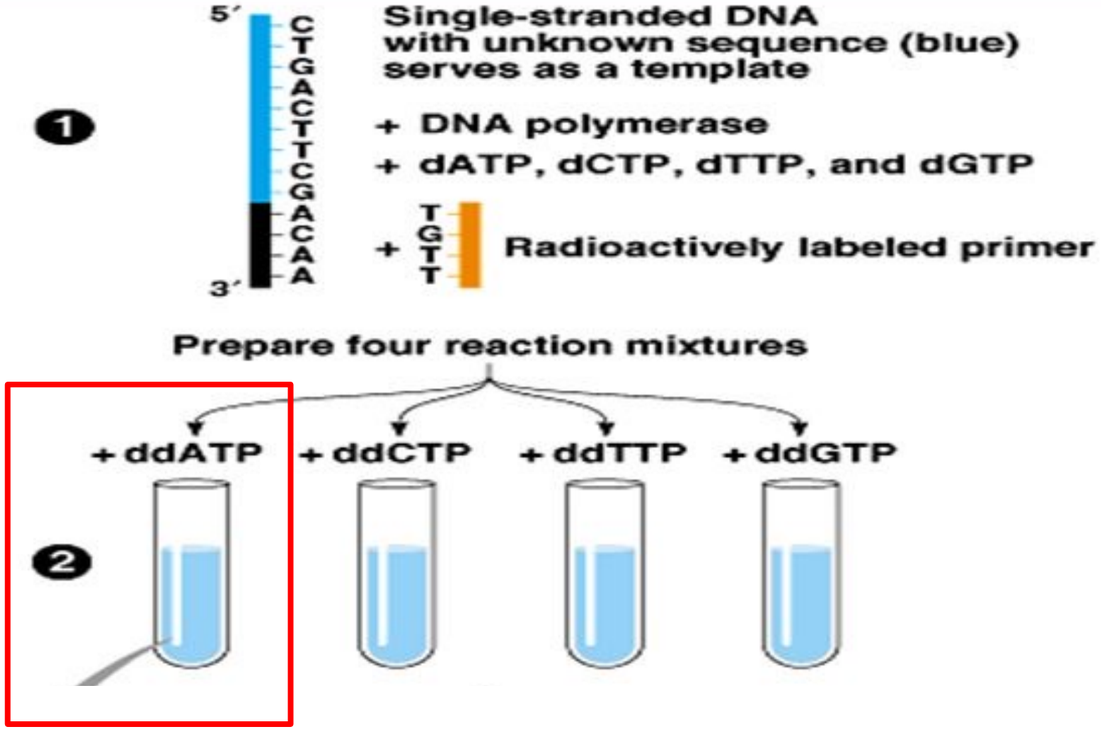
TTGTCGAAG



AACAGCTTCAGTC

TTGTCGA

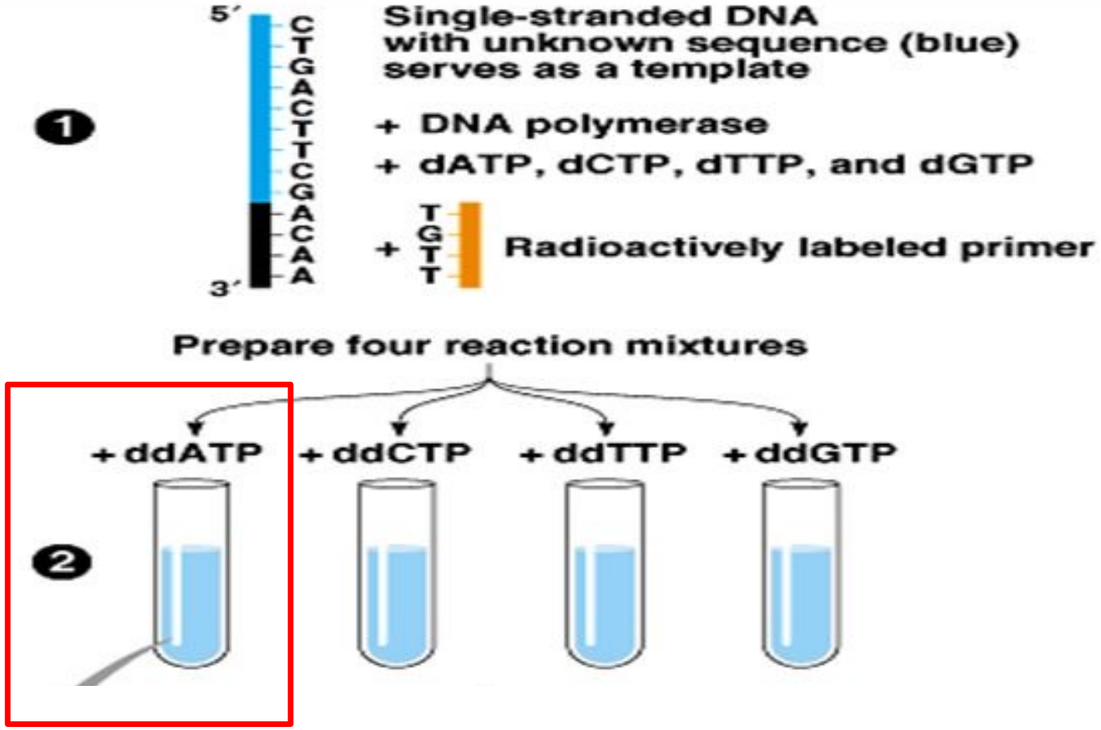
TTGTCGAAGT



AACAGCTTCAGTC

TTGTCGA

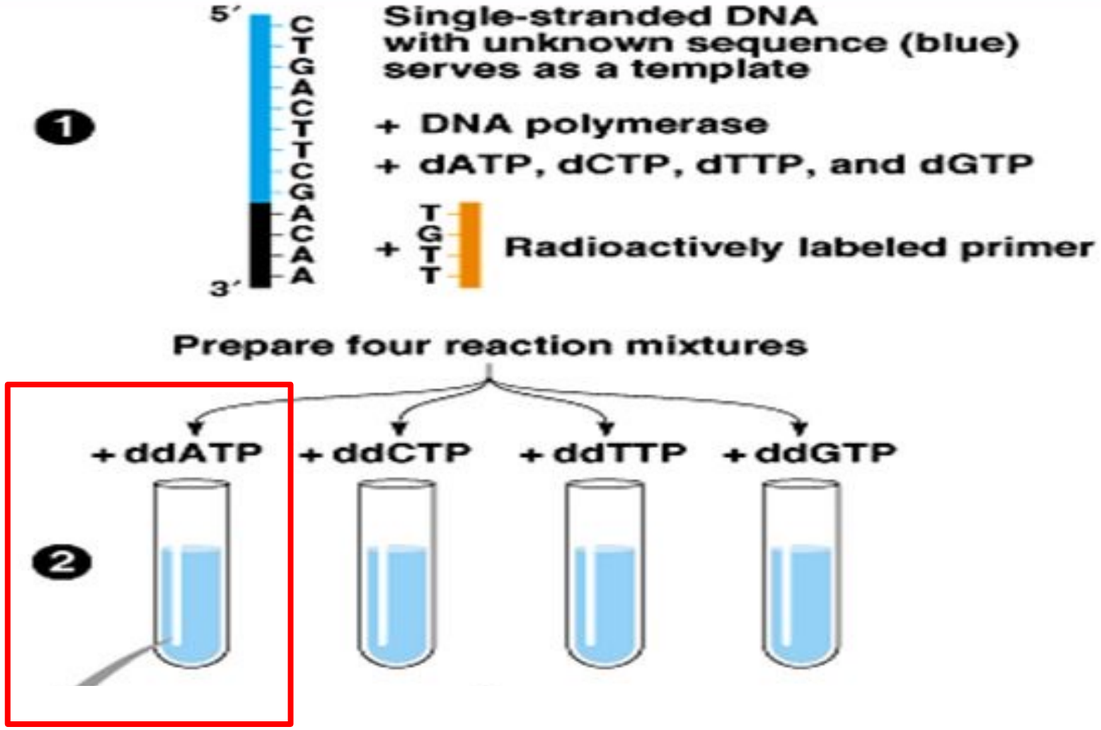
TTGTCGAAGTC



AACAGCTTCAGTC

TTGTCGA

TTGTCGAAGTCA

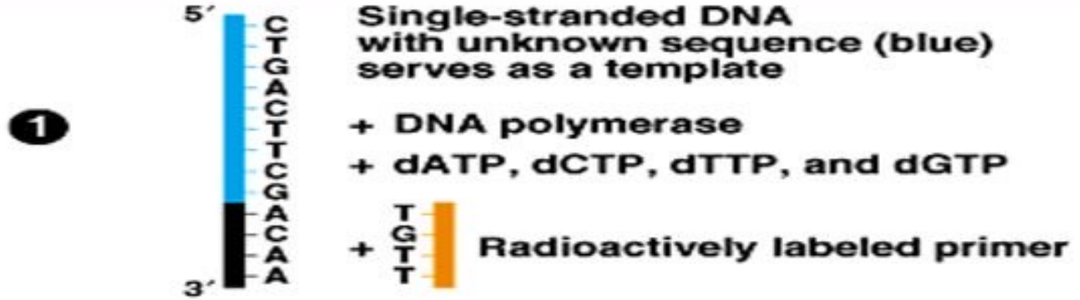


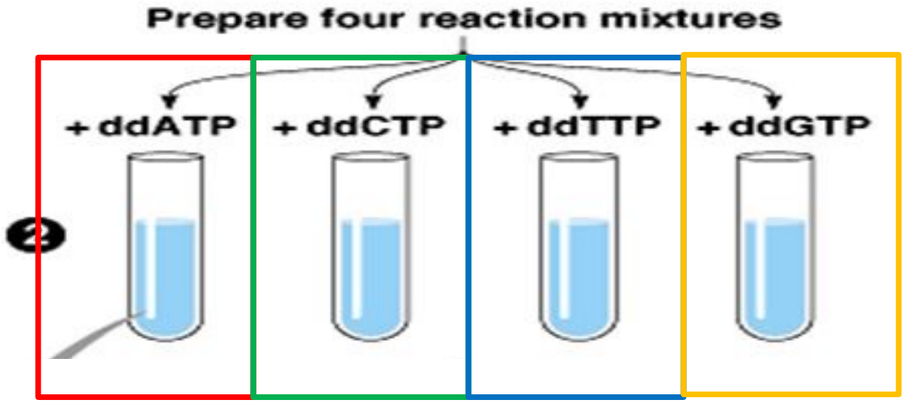
AACAGCTTCAGTC

TTGTCGA

TTGTCGAAGTCA

TTGTCGAA





AACAGCTTCAGTC
TTGTCGA
TTGTCGAAGTCA
TTGTCGAA

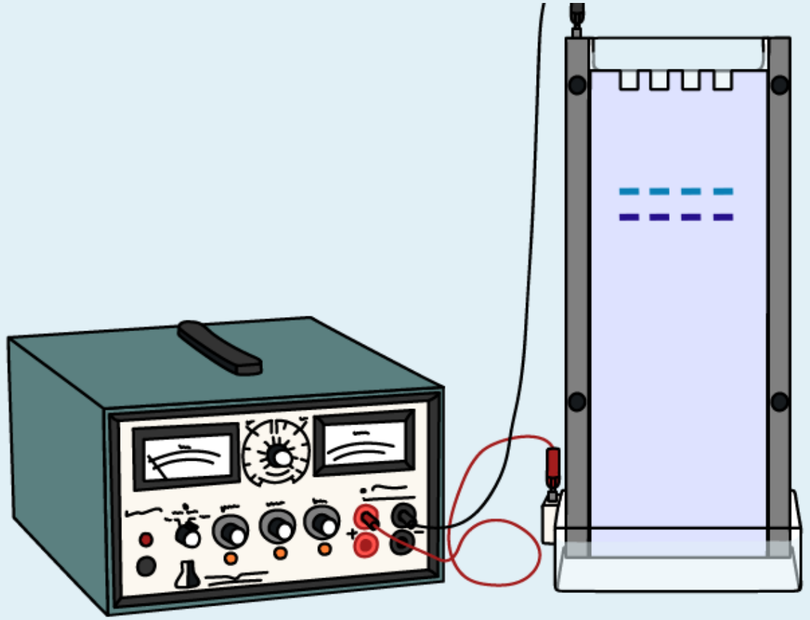
AACAGCTTCAGTC
TTGTC
TTGTCGAAGTC

AACAGCTTCAGTC
TTGTCGAAGT

AACAGCTTCAGTC
TTGTCGAAG
TTGTCG

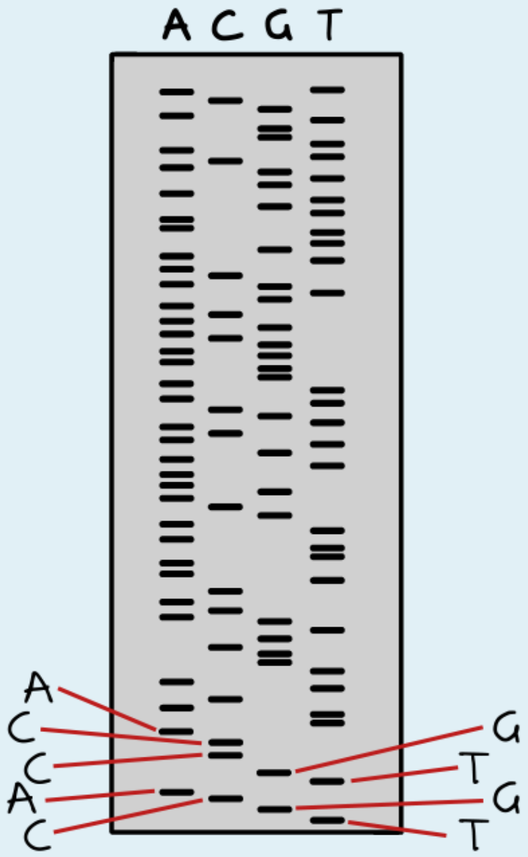
ClevaLab
www.clevalab.com

Bases Are Called From Small to Large



Large

Small



Base Calling

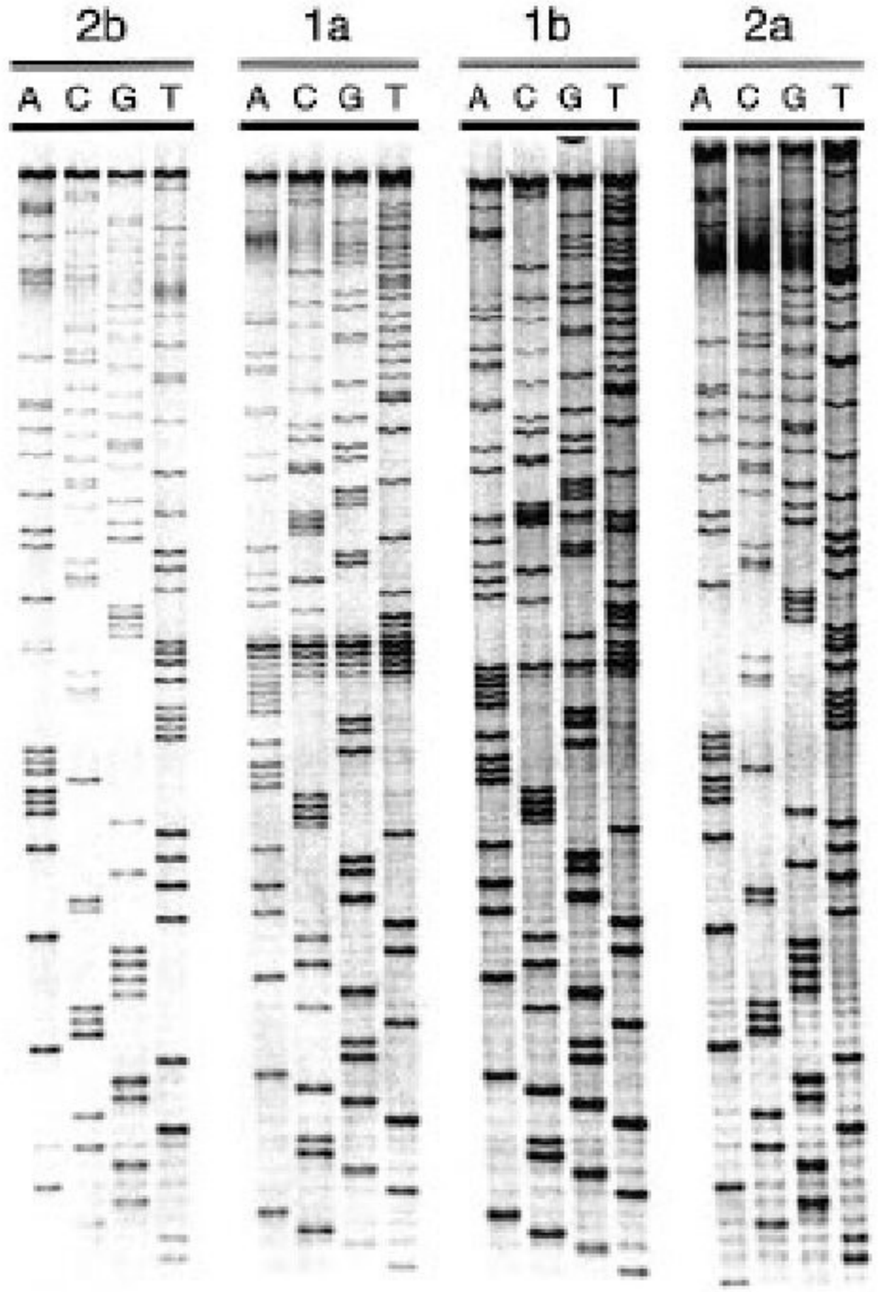
5' to 3'

TGCATGCCA

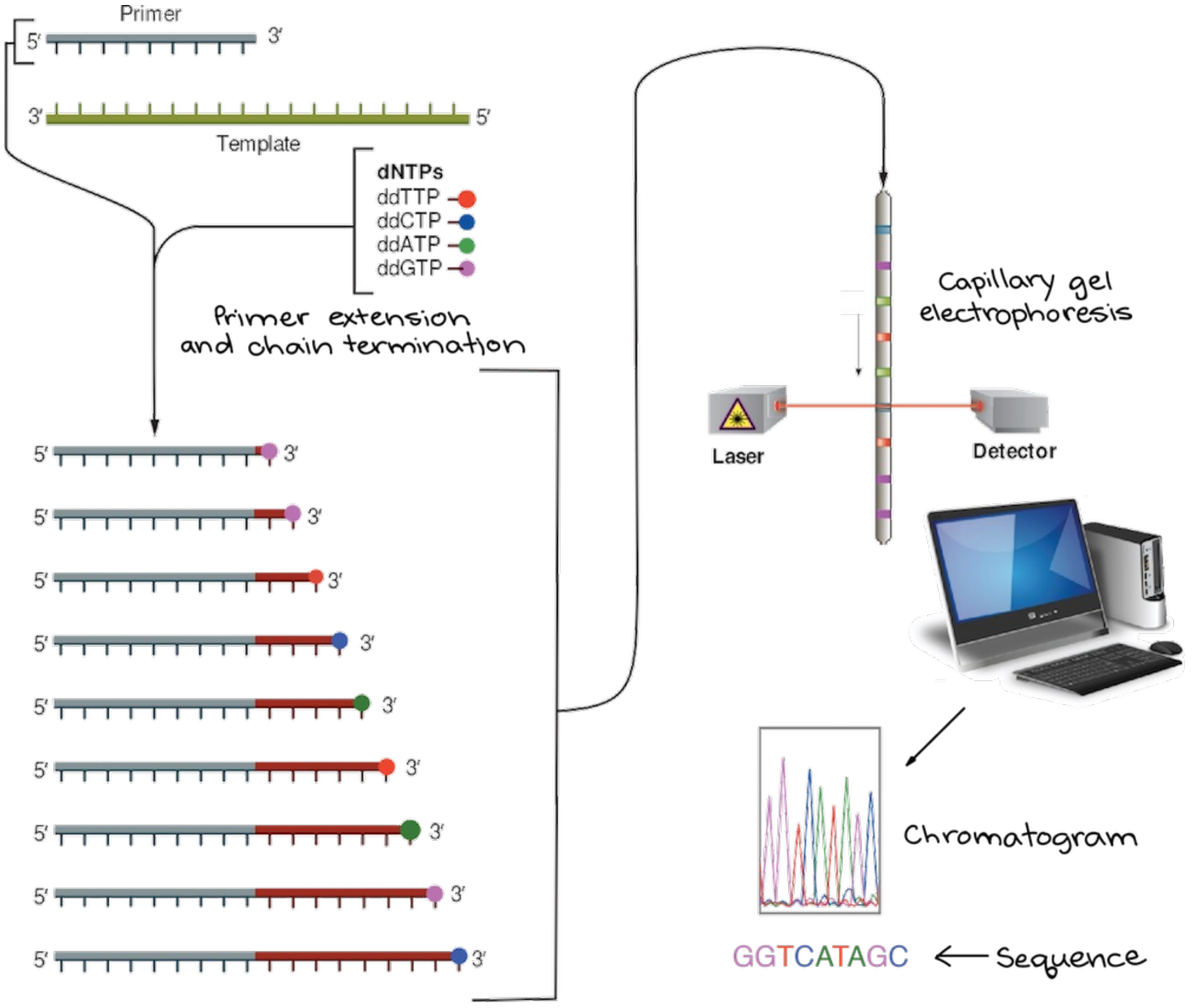
Large

Small

Autoradiogram

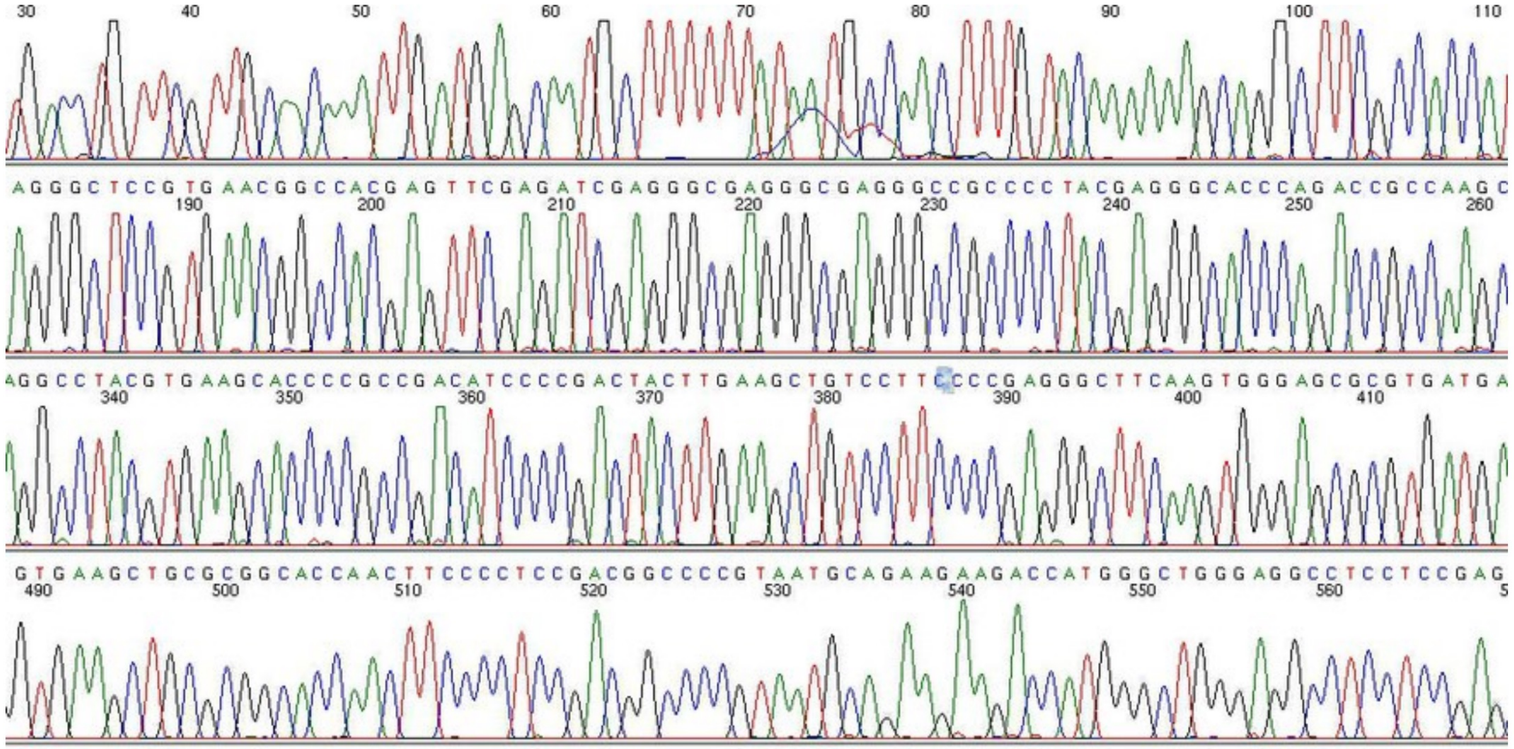


http://dx.doi.org/10.1093/nar/24.19.3806


Primer
5′
3′
3′
5′
Template
dNTPs
ddTTP
ddCTP
ddATP
ddGTP
Primer extension
and chain termination
Capillary gel
electrophoresis
Laser
Detector
Chromatogram
GGTCATAGC
Sequence

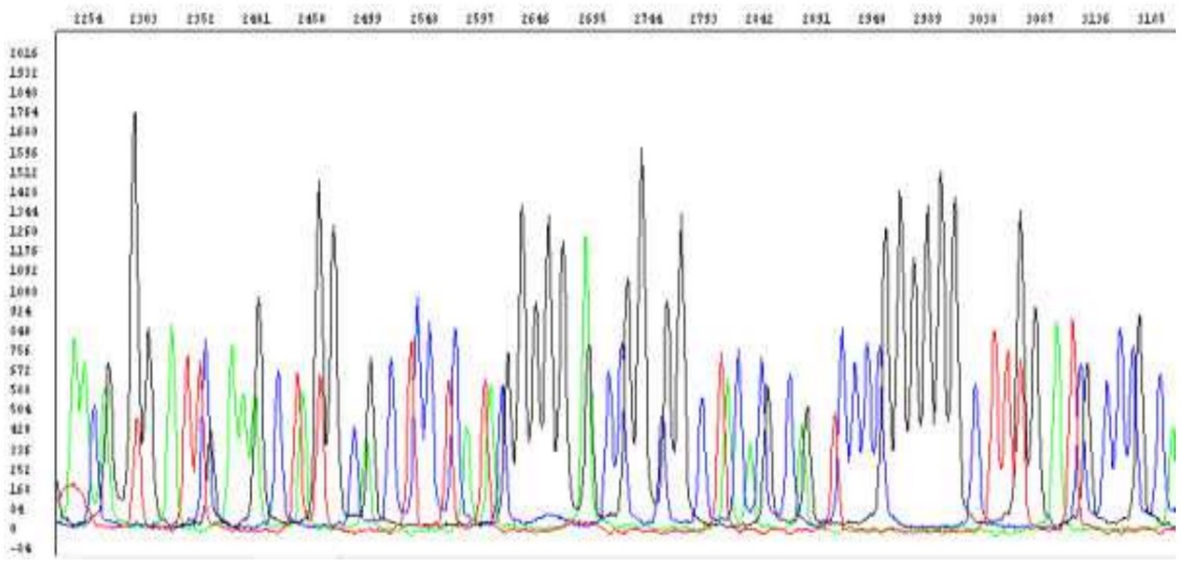
Output from Automated DNA Sequencer

Sanger sequencing throughput
Originally
****96 samples per ~two hours**
Then
****384 samples per ~two hours**
Fragment Length
****500-750 nt**

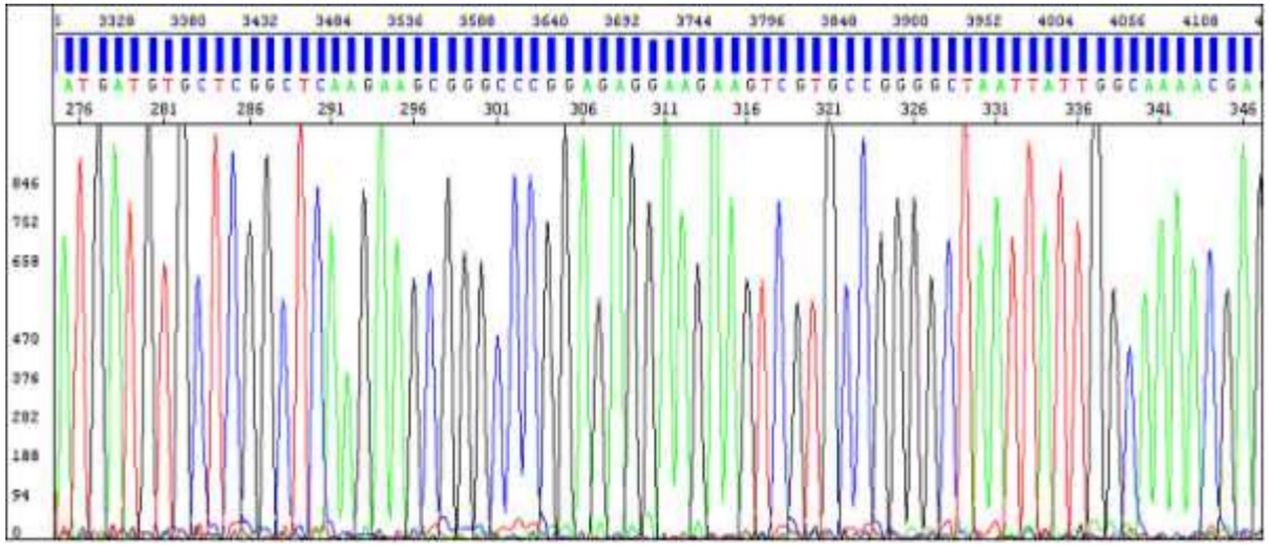


Source: Phil McClean Lab
https://www.ndsu.edu/pubweb/~mcclean/plsc411/index.htm

Raw data (data before analysis by the base caller algorithm) are data as they are recorded by the sequencer:



Electropherogram (data after analysis) shows a sequence of peaks in four colors, each color represents the base called for that peak and there is a textual version of recorded sequence visible:



DATA (SANGER SEQUENCING OUTPUT)

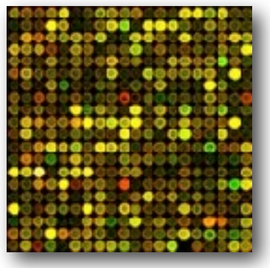
Results of DNA sequencing are provided in three data files – .ab1 file, .seq file and .phd.1 file.

- *.ab1 file contains the DNA sequence electropherogram as well as raw data and some other information.
- *.seq file is a simple sequence text file in FASTA format.
- *.phd.1 file (Phred file) is a simple text file containing bases with quality values for each base.

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

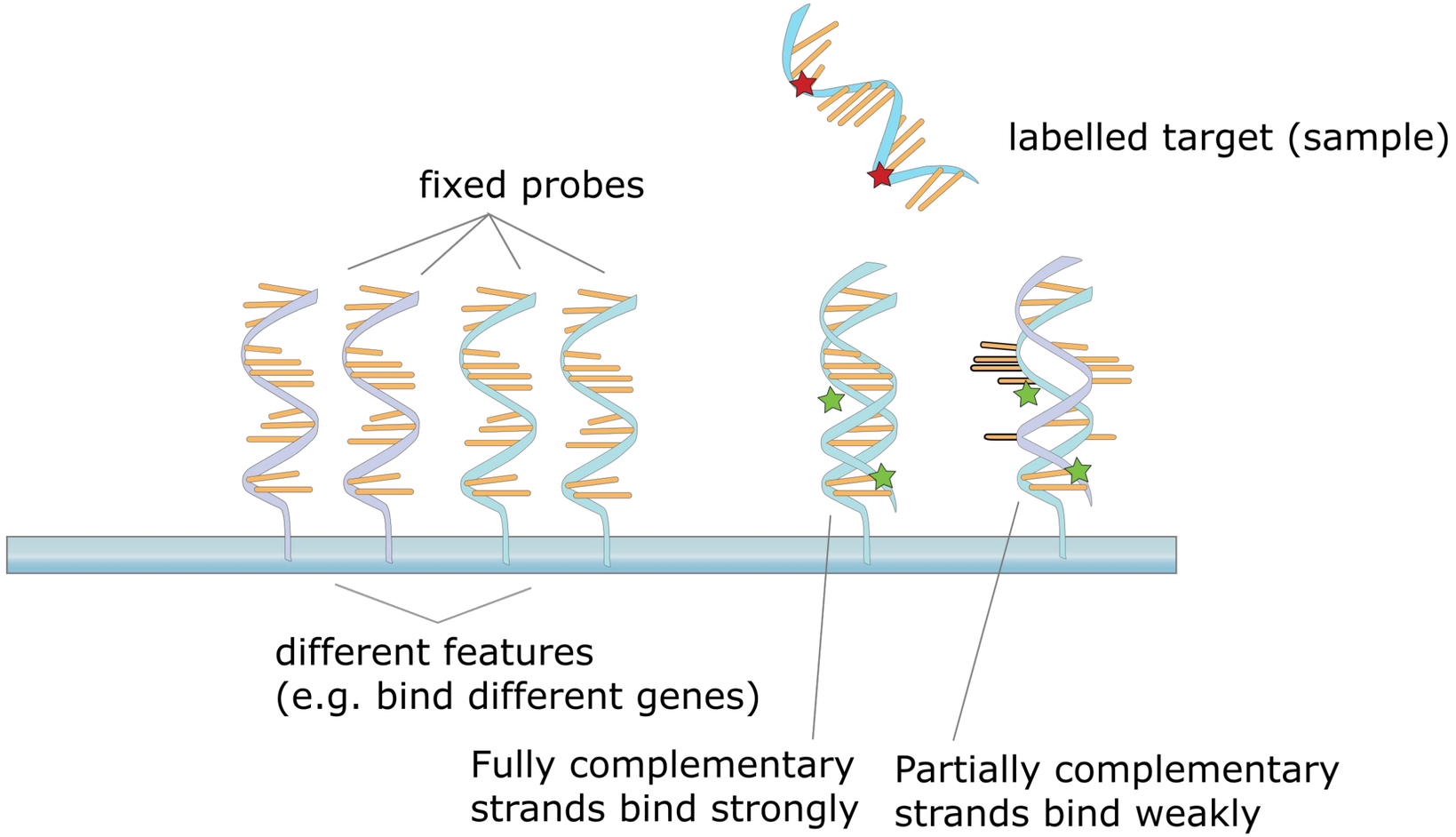
Since ~2010

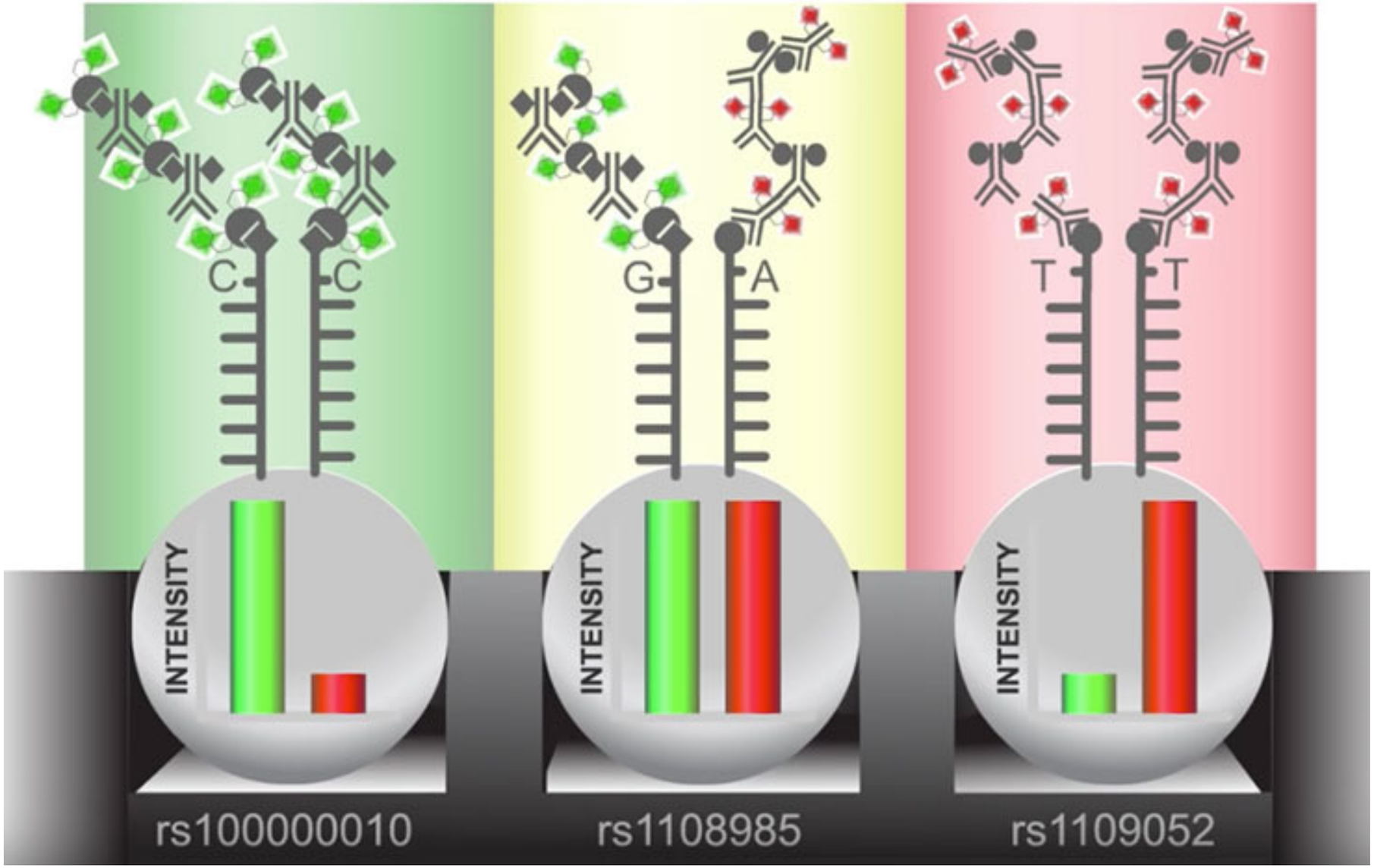
Solexa

illumina®

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

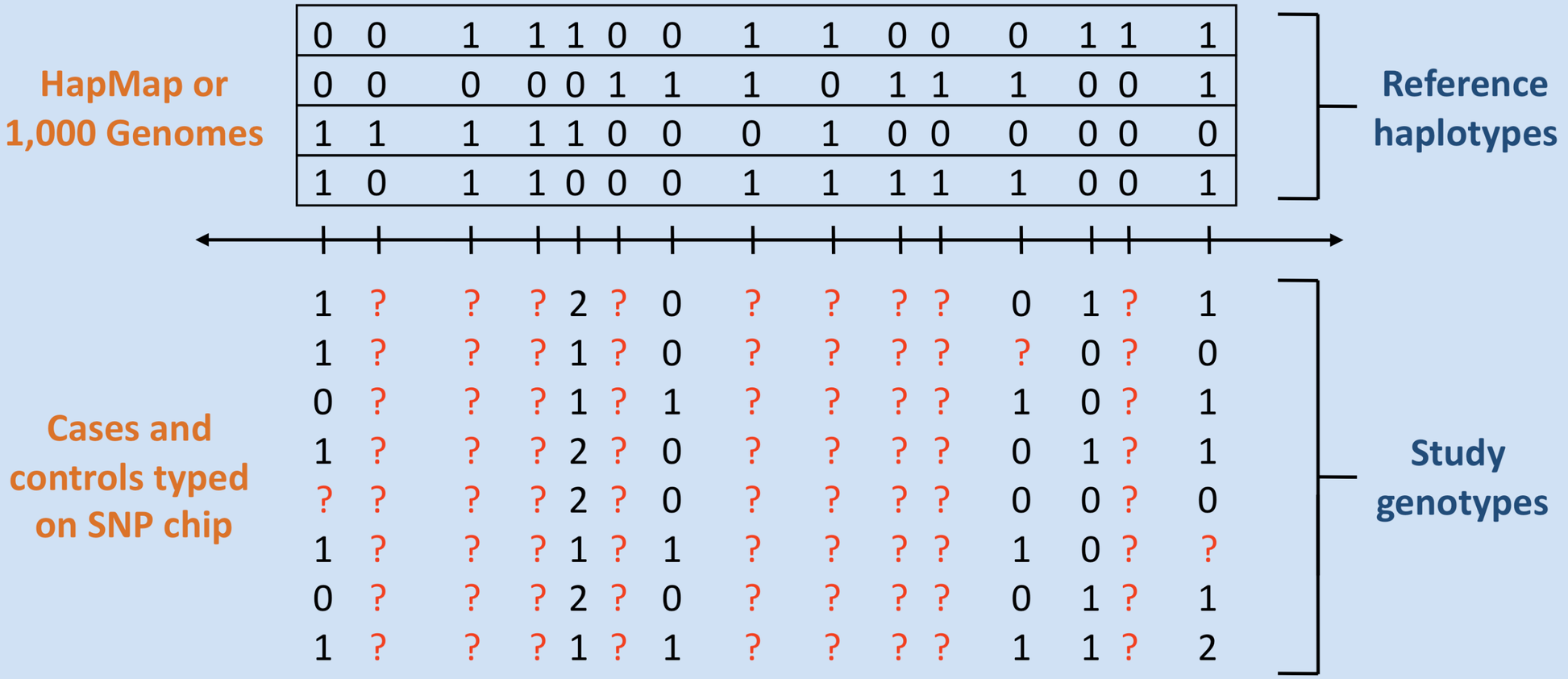

AFFYMETRIX
GeneChip®
Human Genome
U133 Plus 2.0
P/N: 520019
Lot #: 4010008
Exp. Date: 05/07/06
For Research Use Only
AFFYMETRIX
GeneChip®
Mouse Genome
430 2.0 Array
P/N: 520029
Lot #: 4015112
Exp. Date: 11/24/06
For Research Use Only


labelled target (sample)
fixed probes
different features
(e.g. bind different genes)
Fully complementary
strands bind strongly
Partially complementary
strands bind weakly



https://www.illumina.com/science/technology/microarray.html

Typical imputation scenario

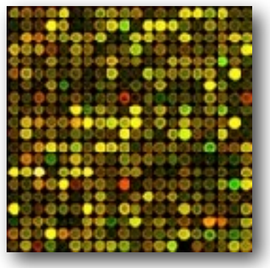


https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

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

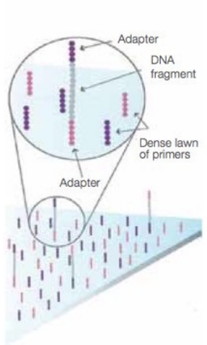
Solexa

illumina®

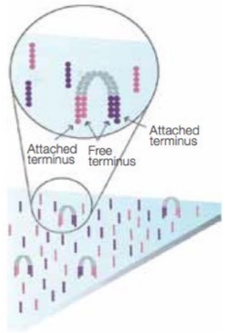
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Illumina NovaSeq 6000
Sequencing by Synthesis

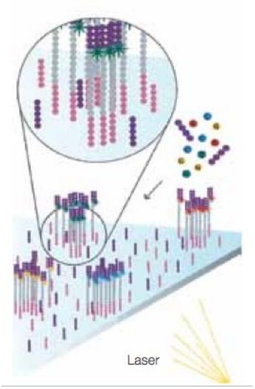
>3Tbp / day



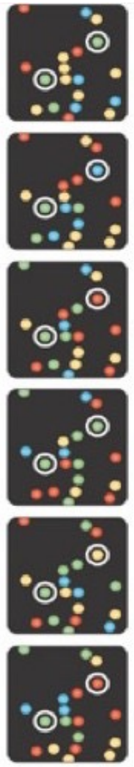
1. Attach



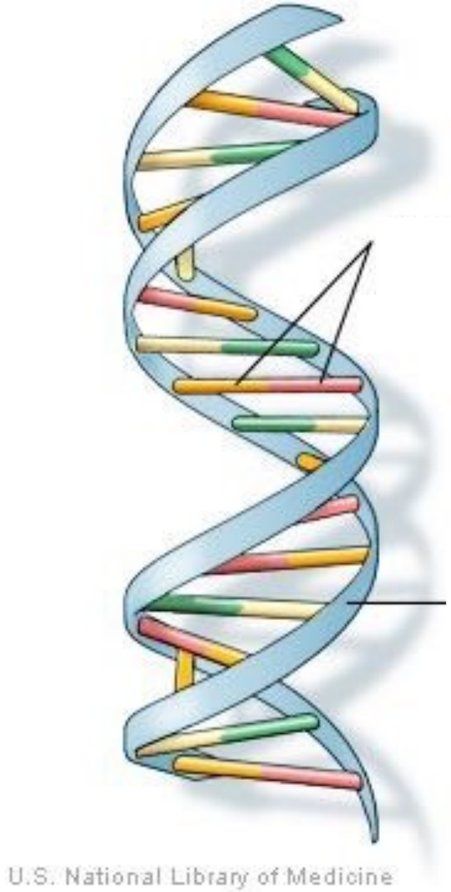
2. Amplify



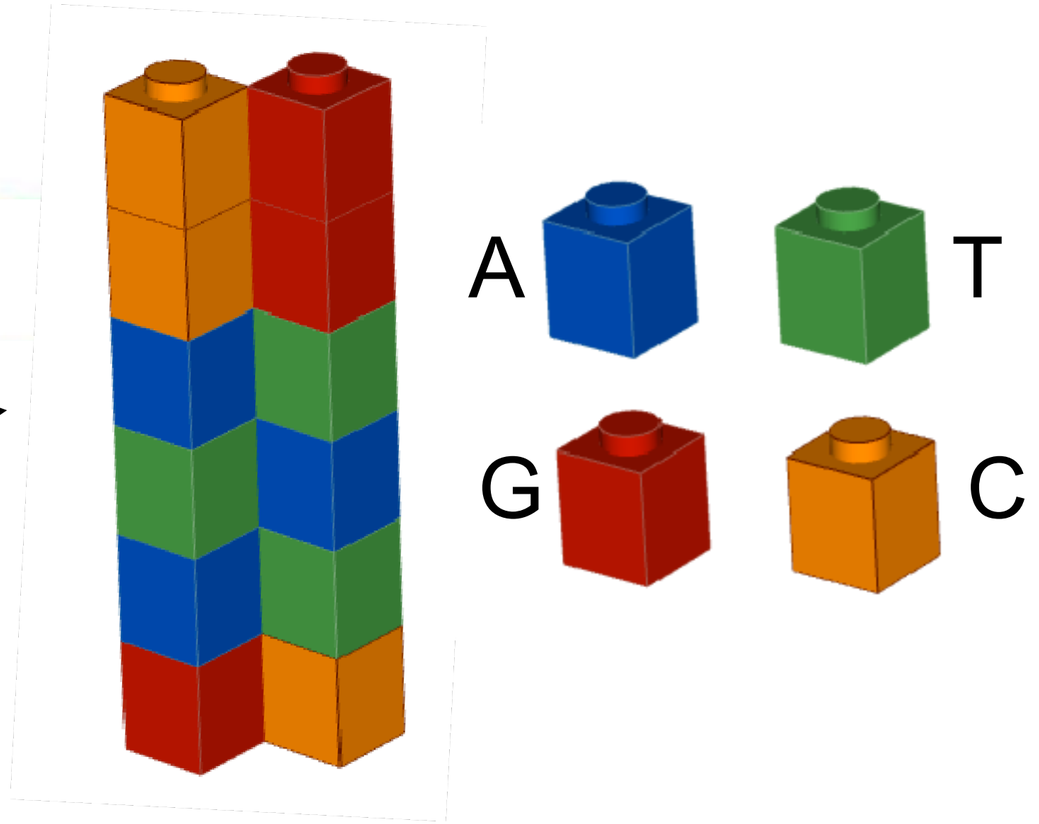
3. Image



Metzker (2010) Nature Reviews Genetics 11:31-46
https://www.youtube.com/watch?v=fCd6B5HRaZ8



Double stranded
DNA (double helix)



Double stranded
DNA (lego version)

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

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

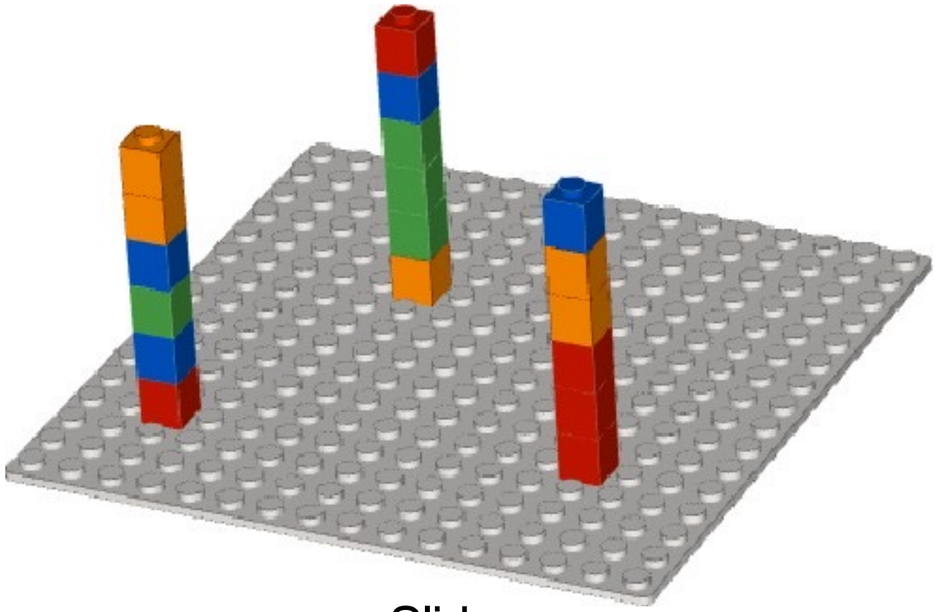
Deposit on slide

C
C
A
T
A
G

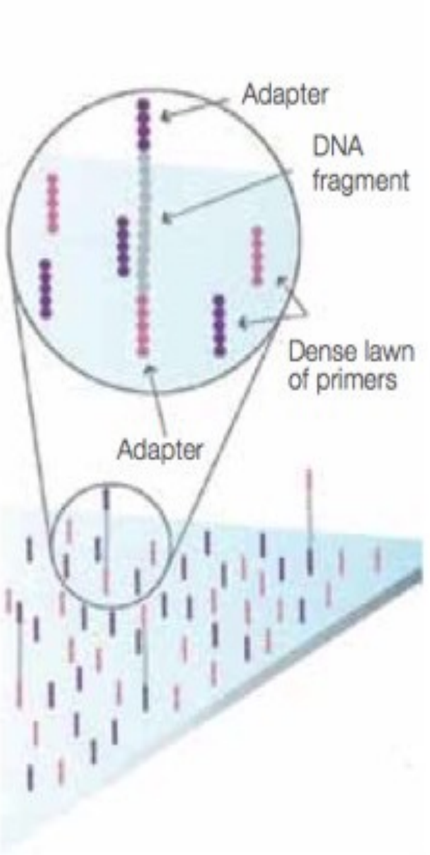
More details: Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6;456(7218):53-9

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

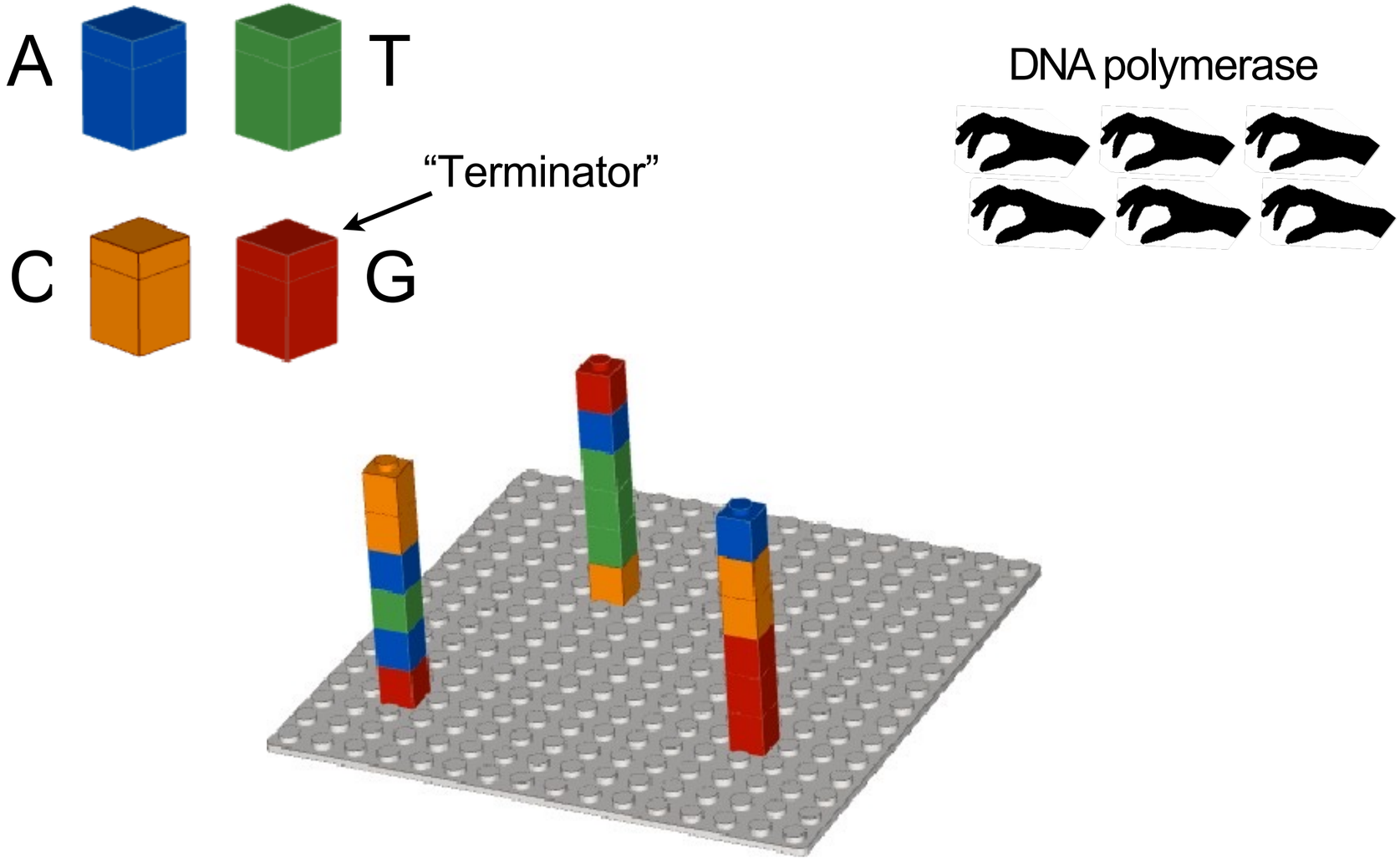
Template
(billions of them!)



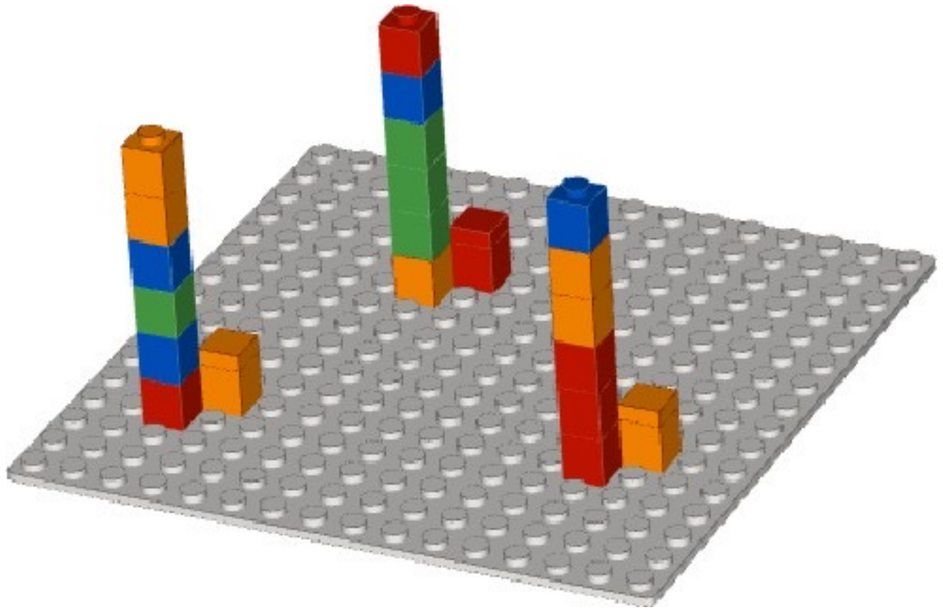
Slide



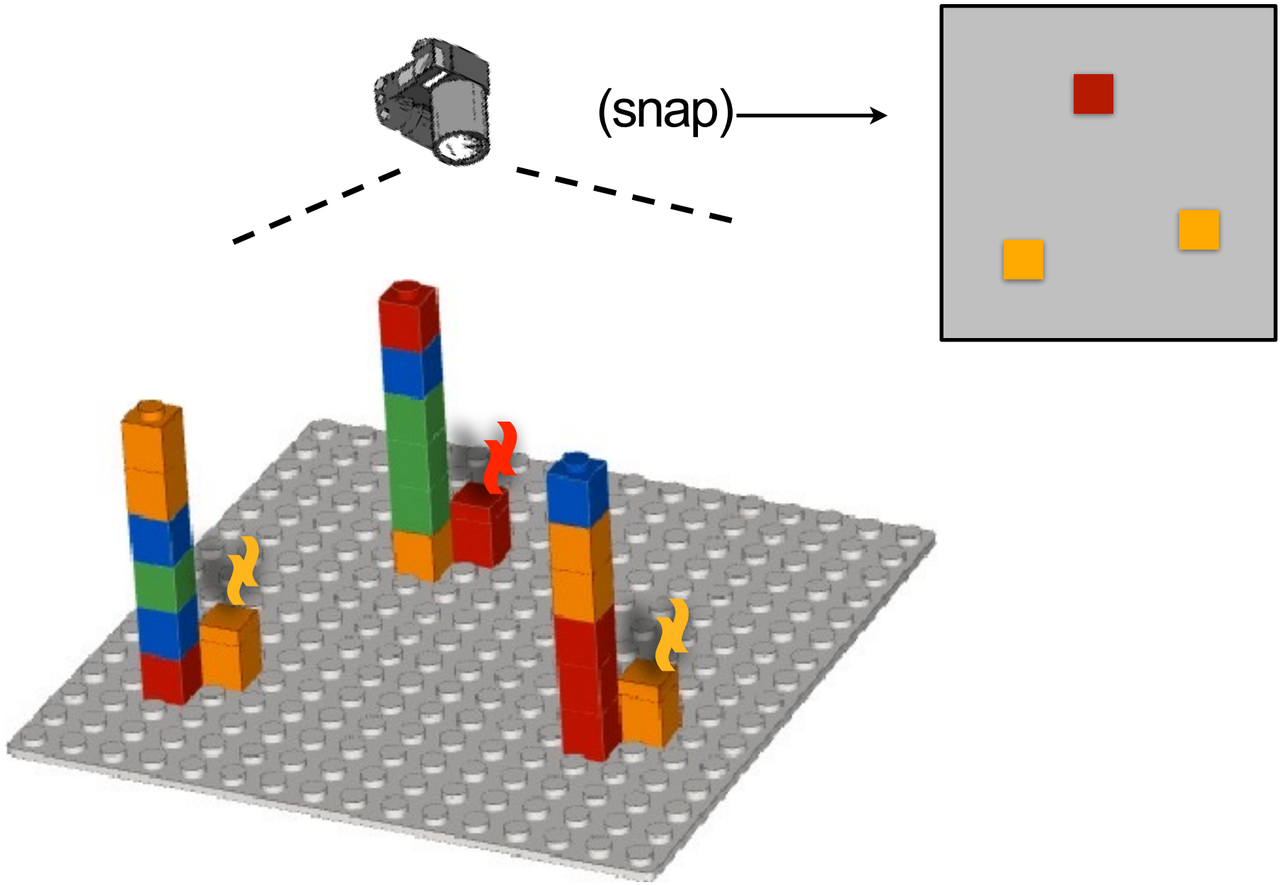
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



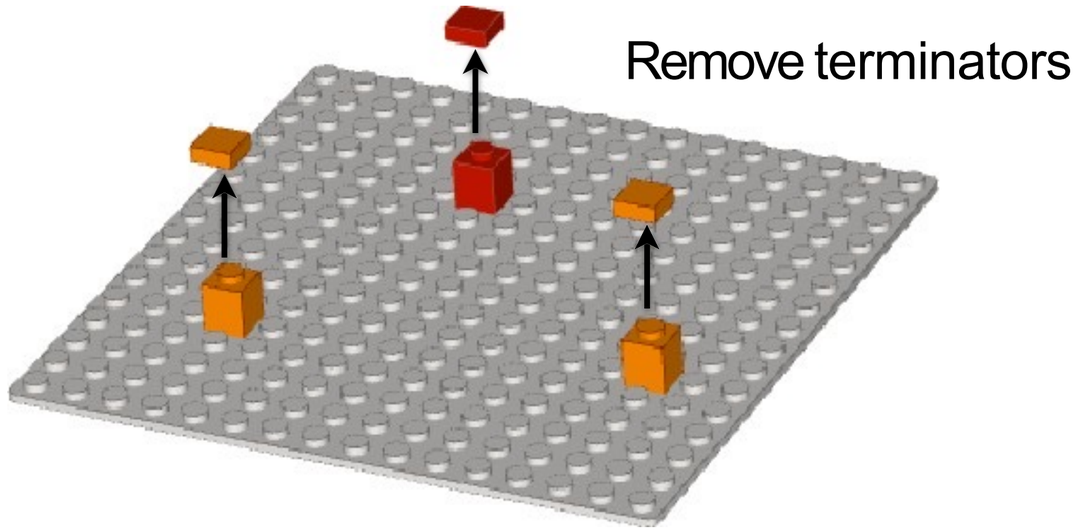
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

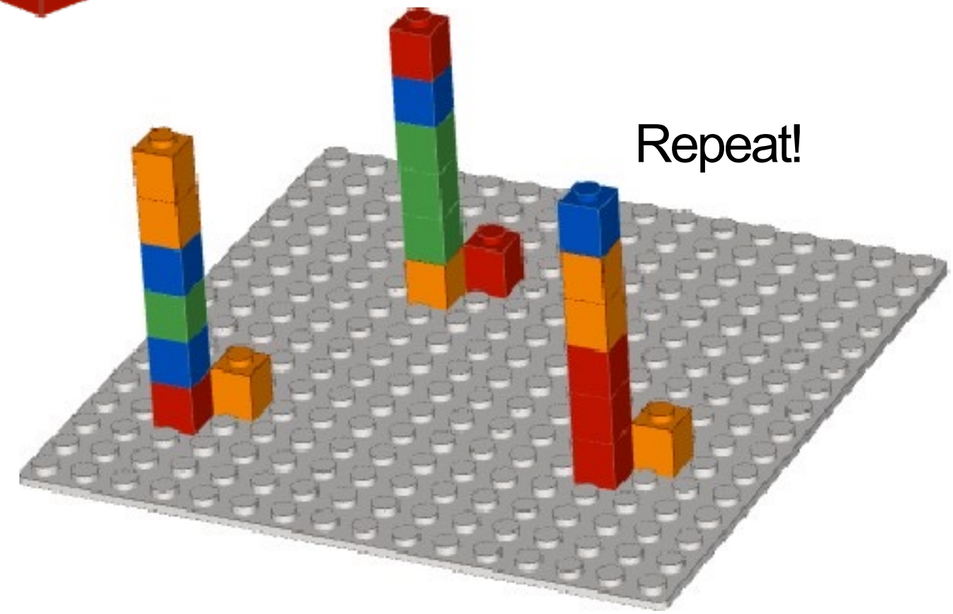


Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

A   T

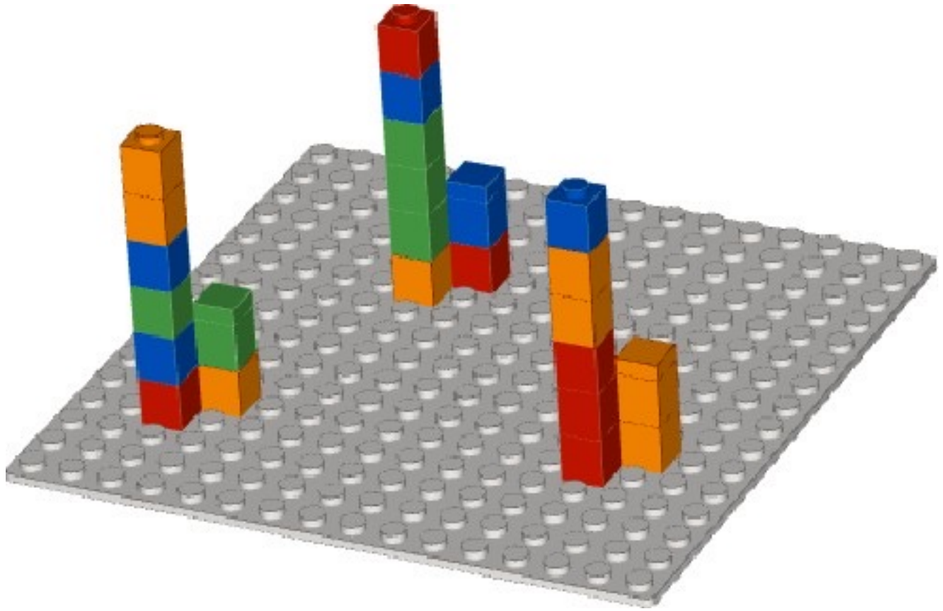
C   G

DNA polymerase

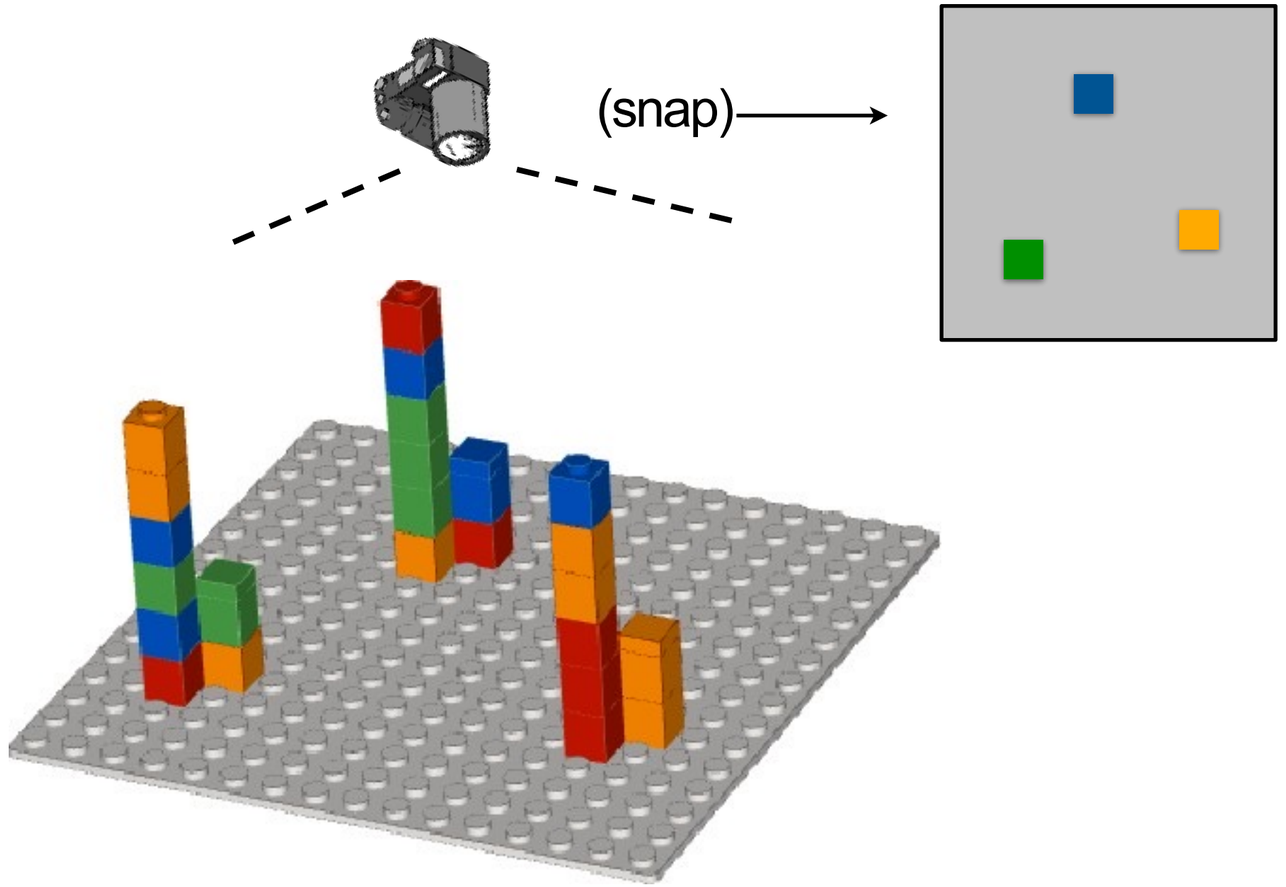


Repeat!

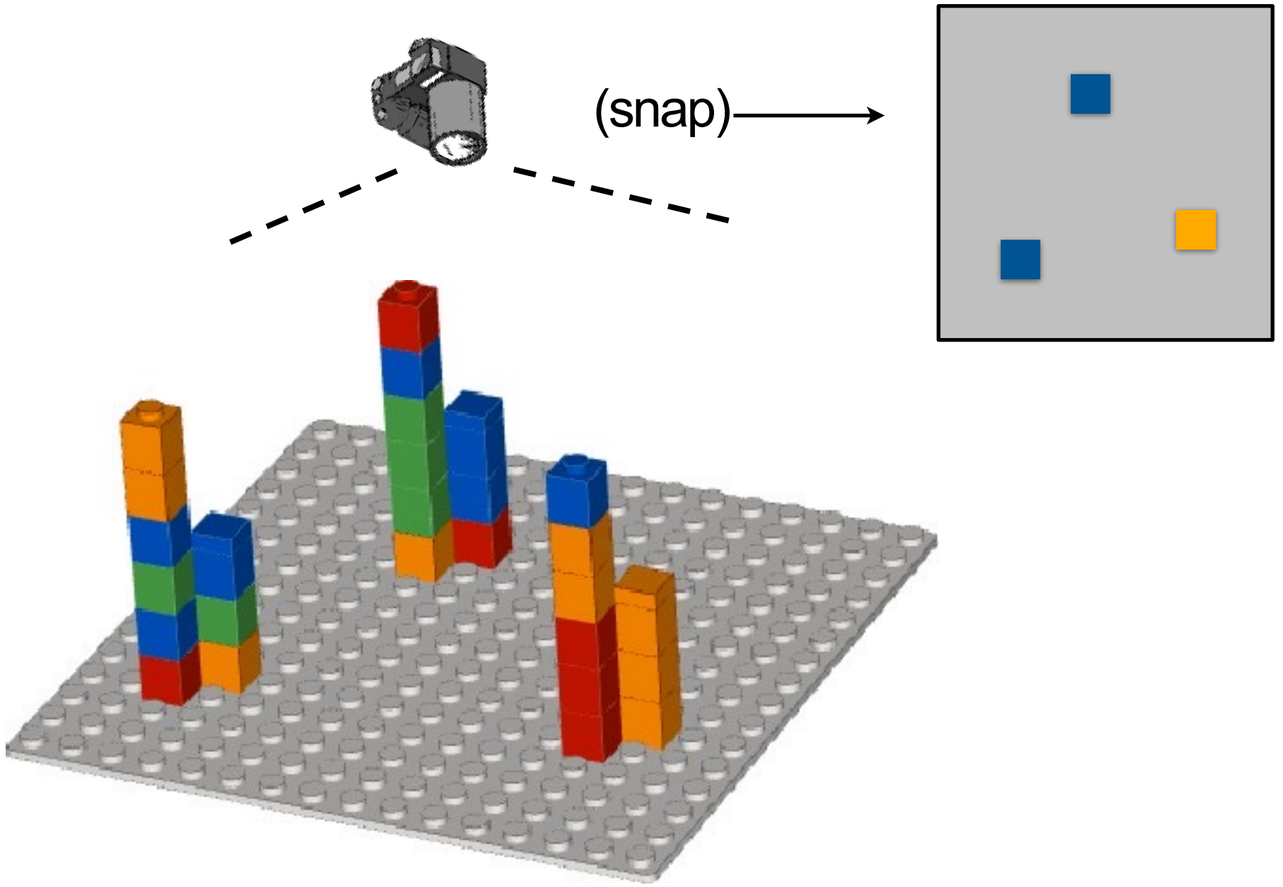
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



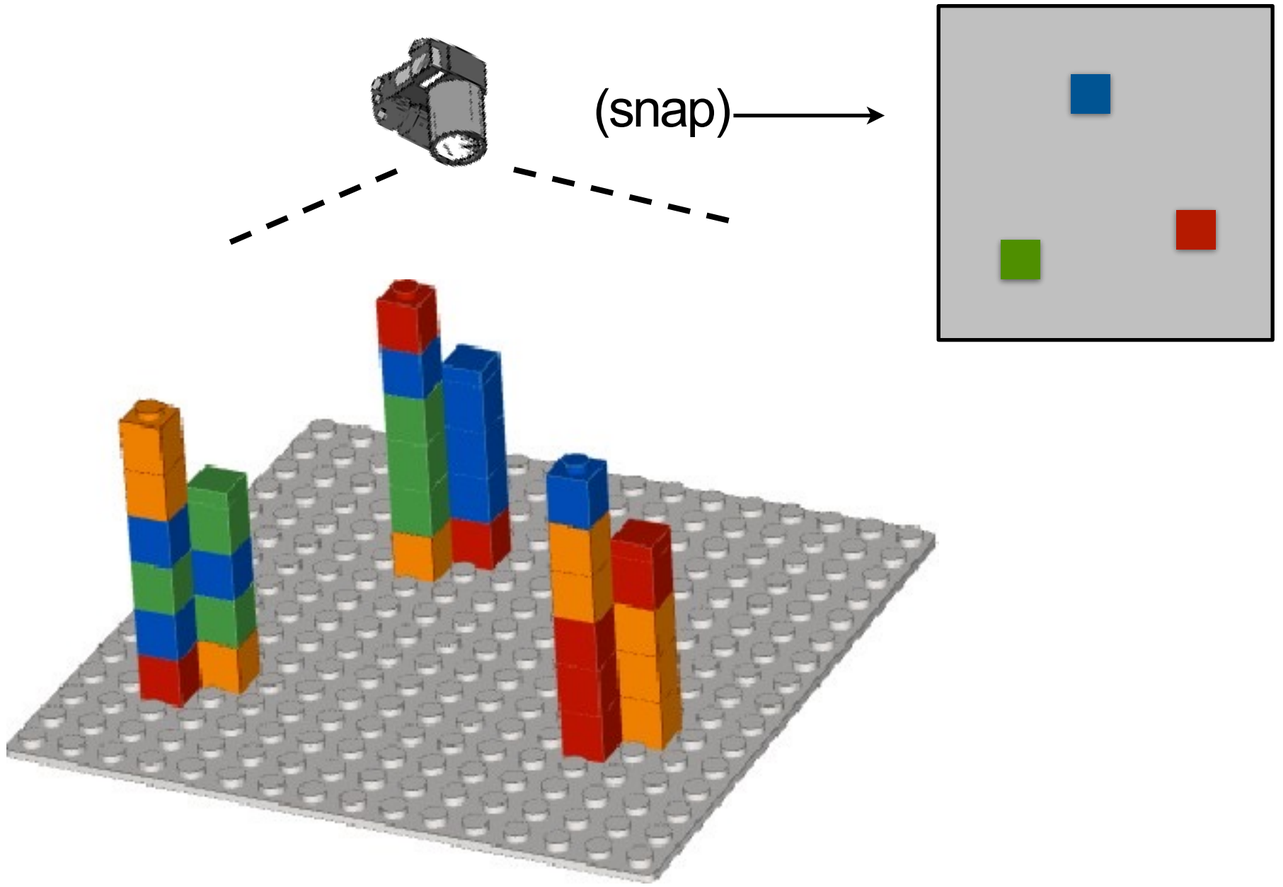
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



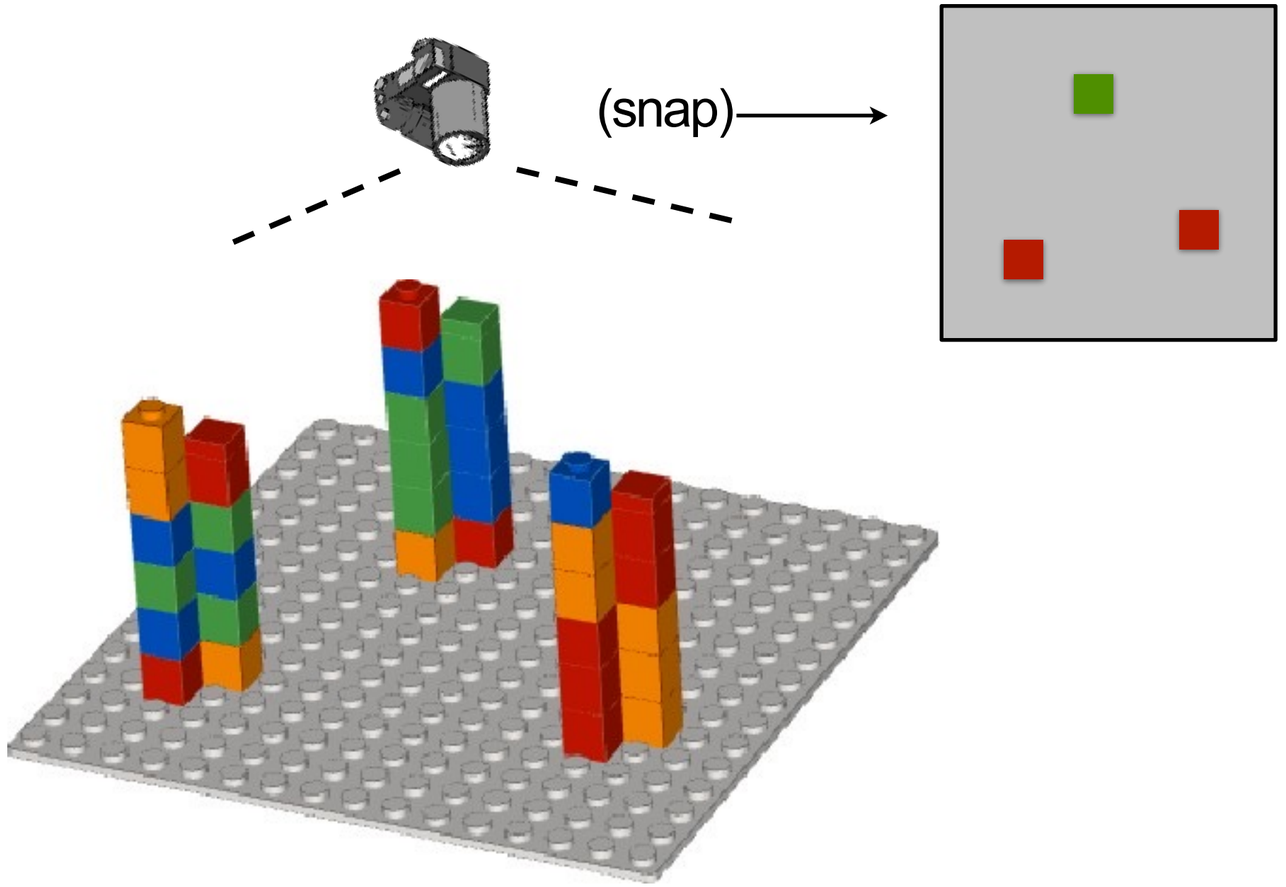
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



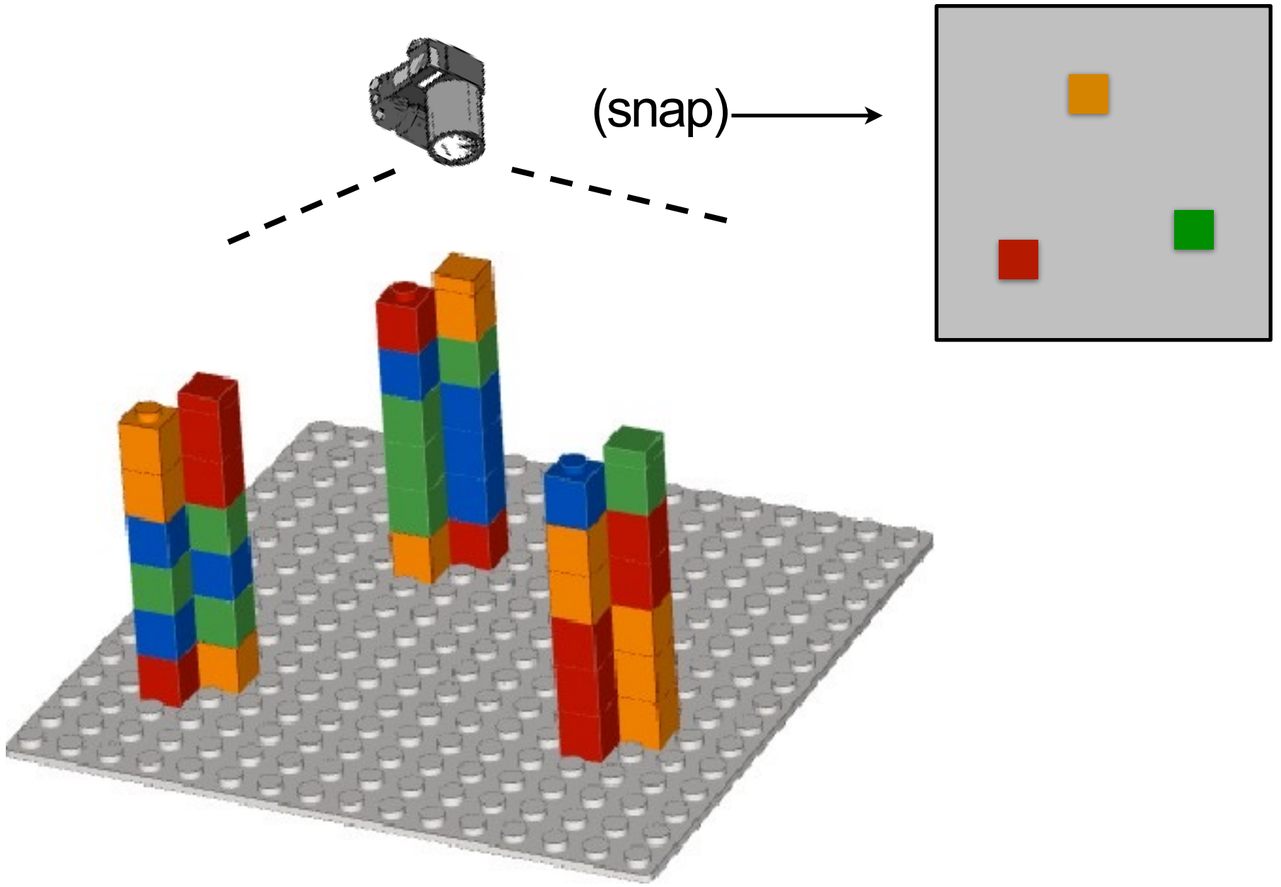
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

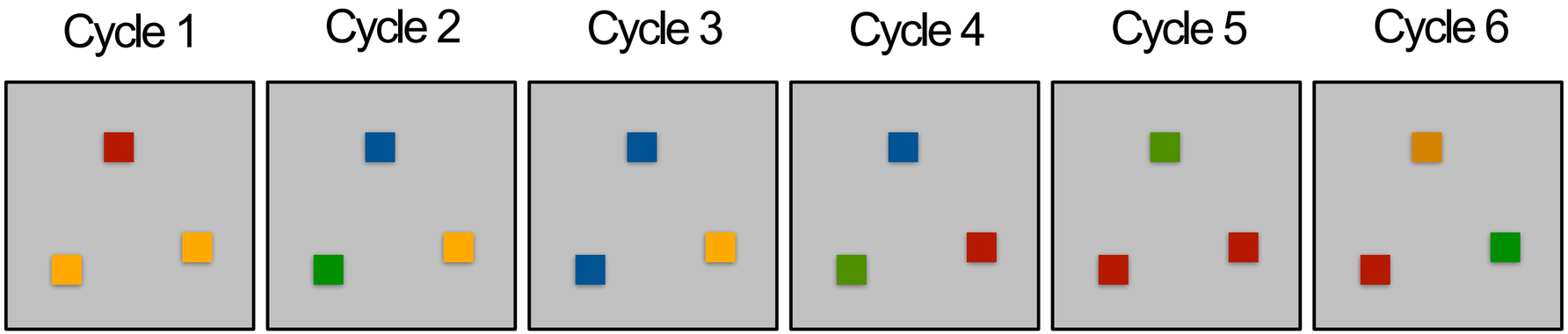


Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



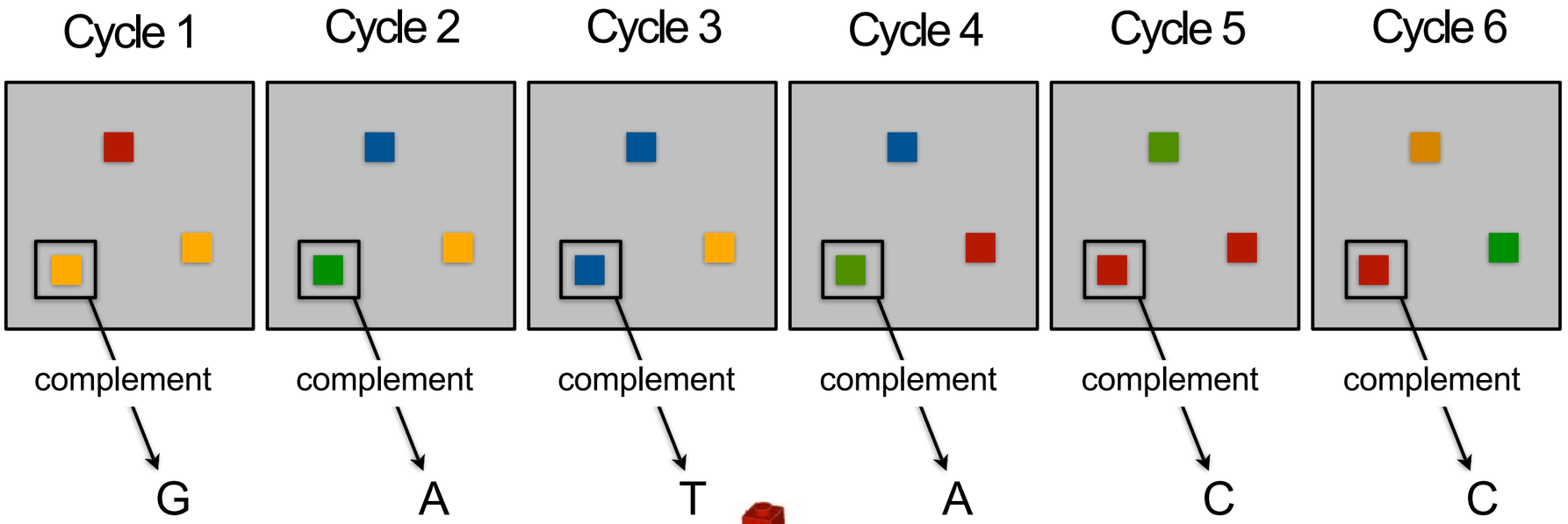
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

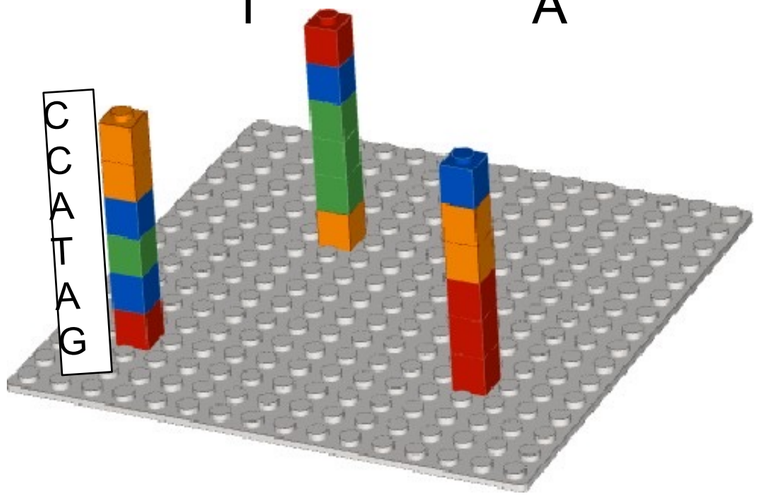
Sequencing by synthesis



Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

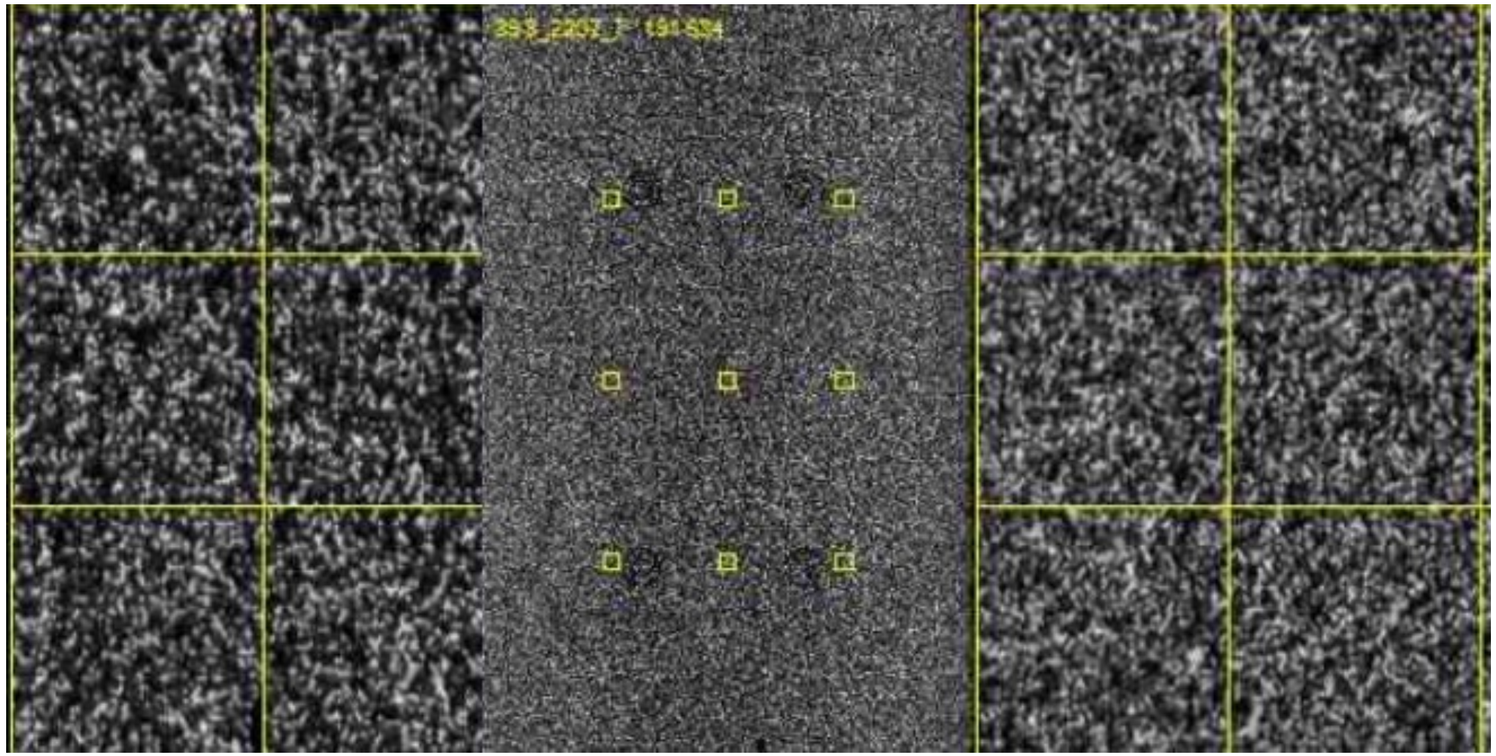
Sequencing by synthesis





Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Sequencing by synthesis



Actual Illumina HiSeq 3000 image

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

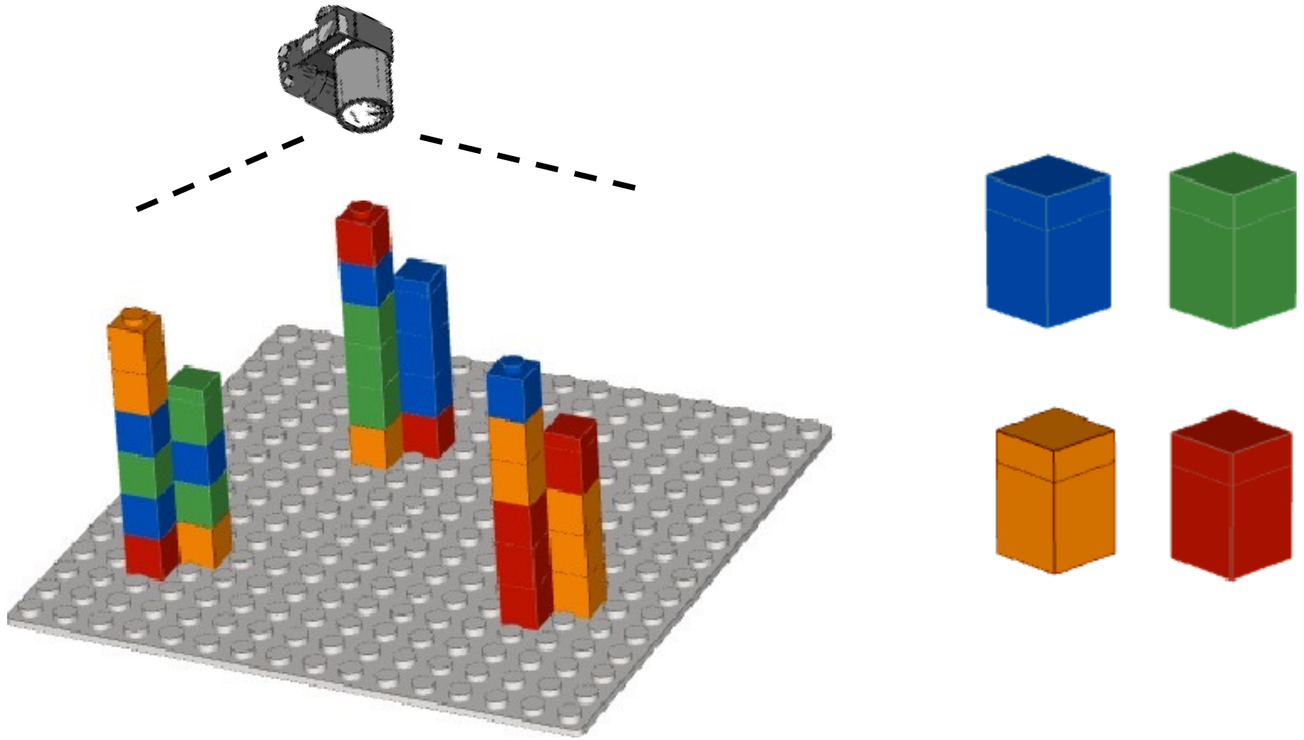
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Sequencing by synthesis

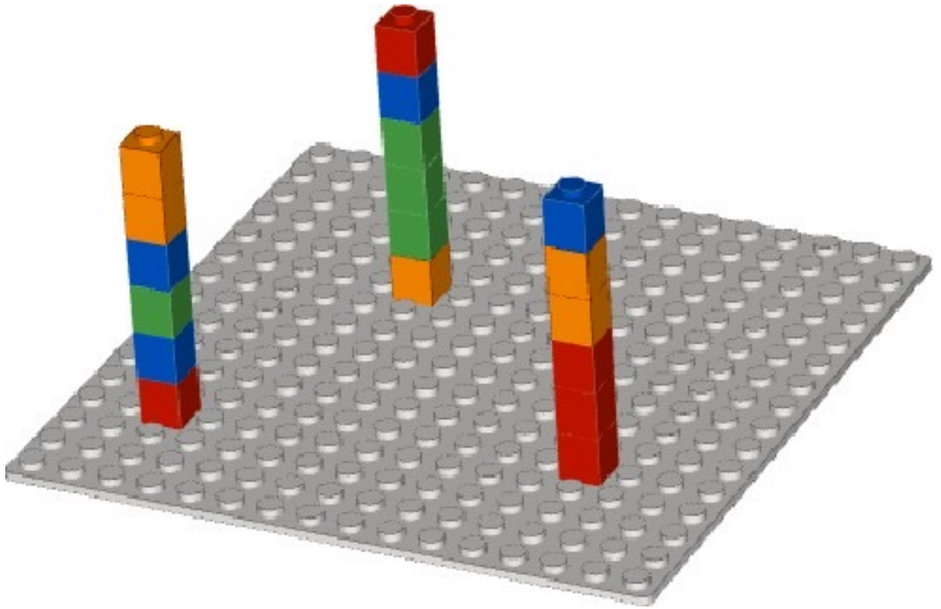
Billions of templates on a slide

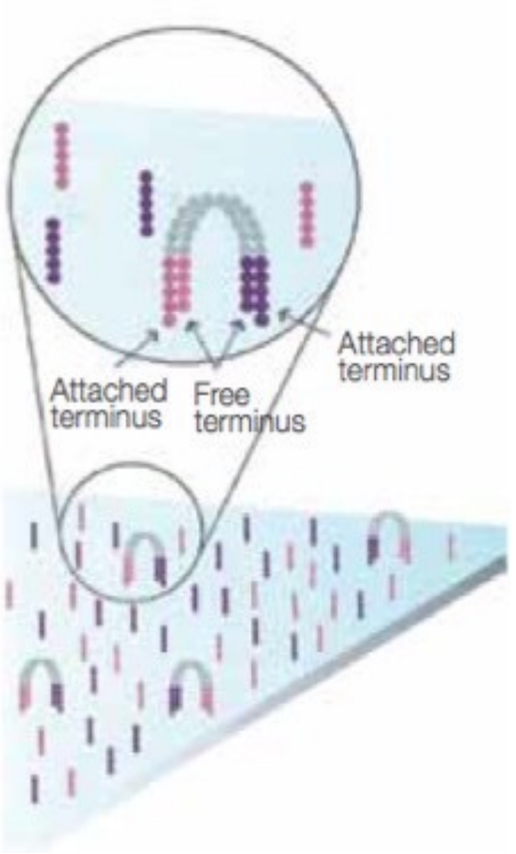
Massively parallel: photograph captures all templates simultaneously

Terminators are "speed bumps," keeping reactions in sync



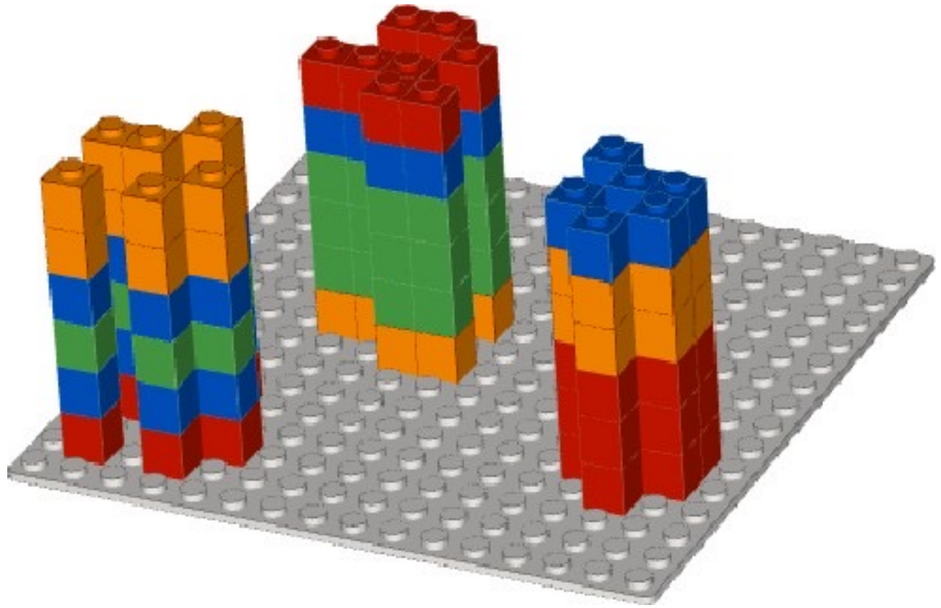
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



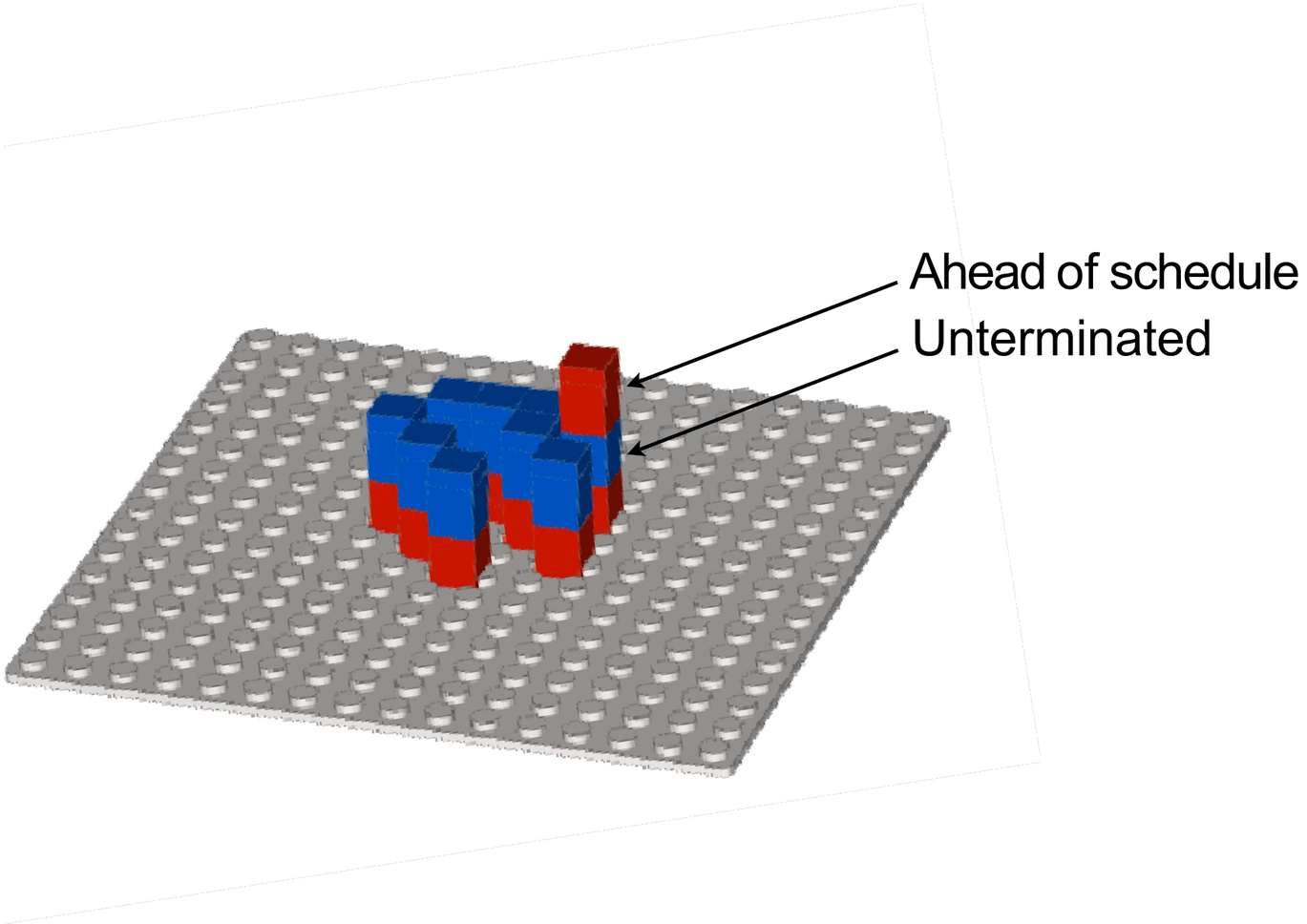


Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

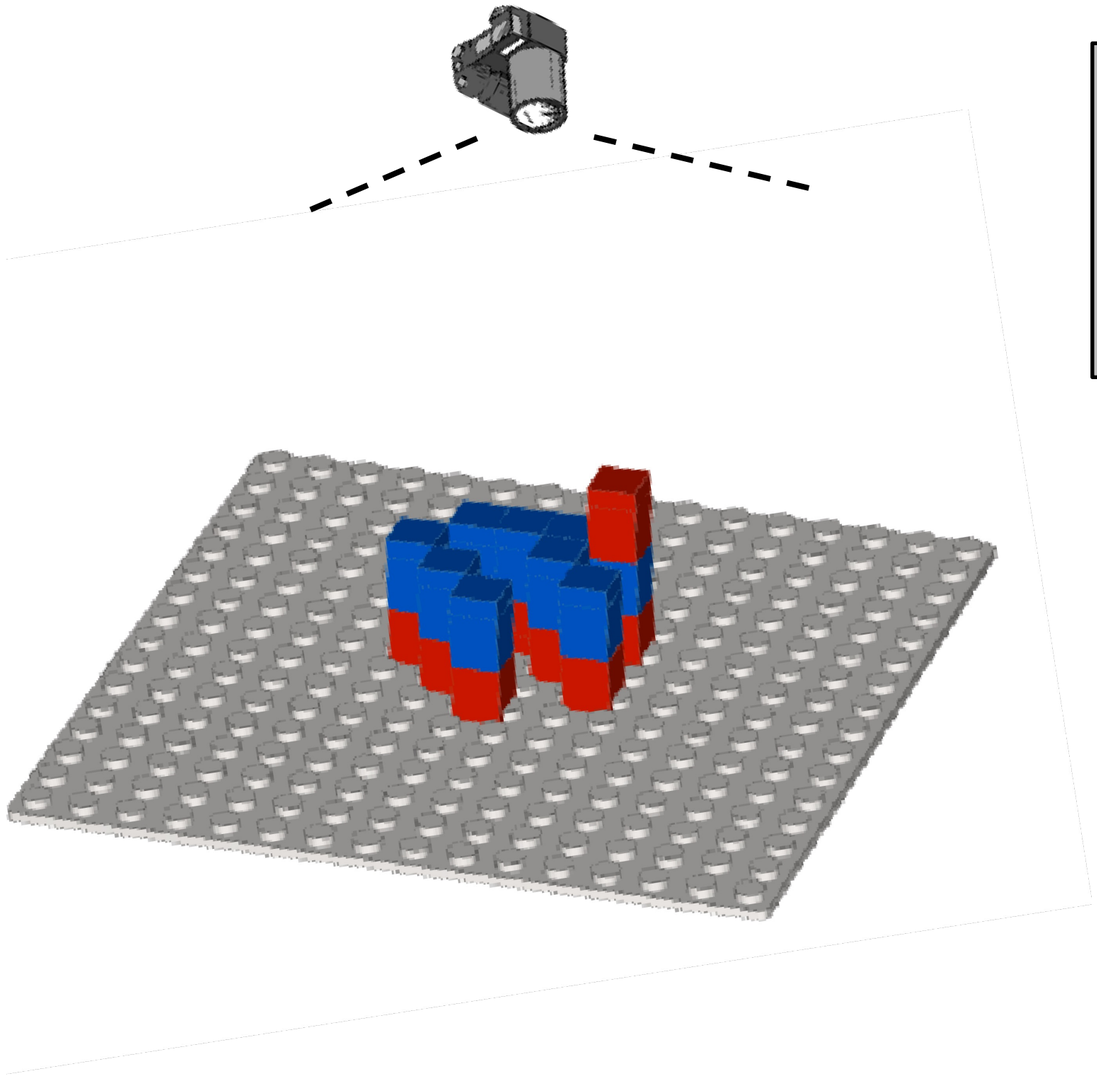
Cluster of clones

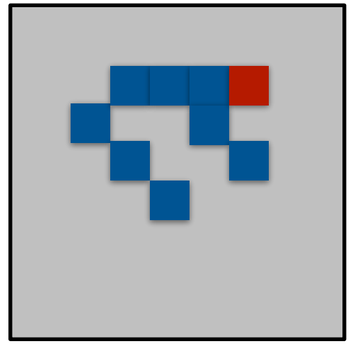


Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

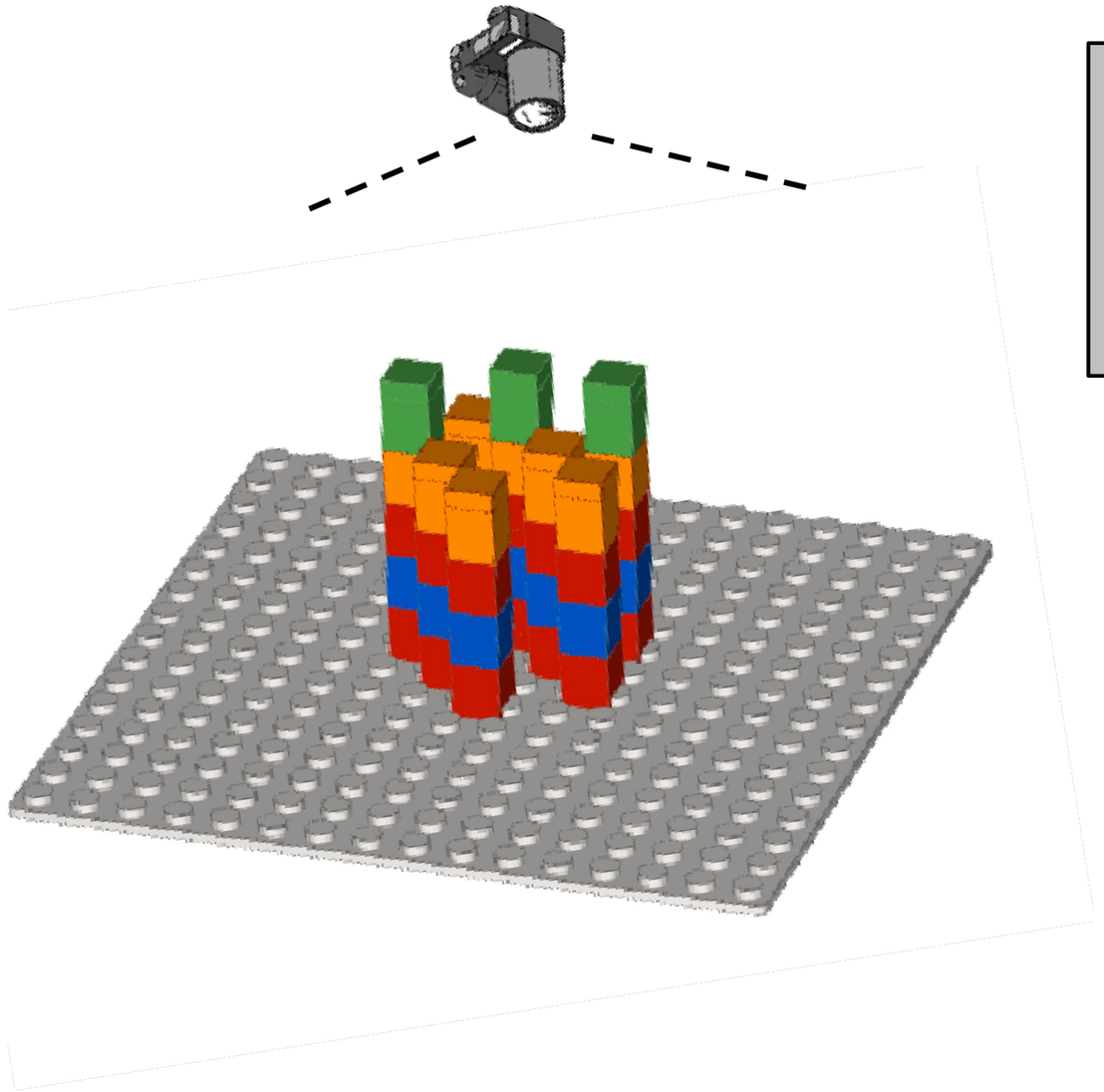


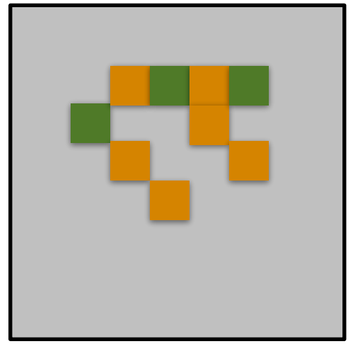
Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf





Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf





Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

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

Base quality

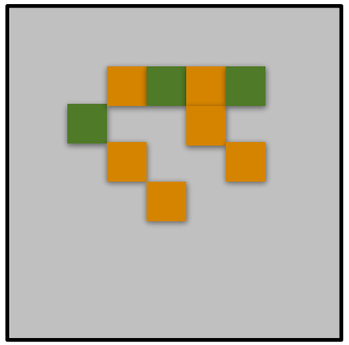
Probability that base call is incorrect

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

$Q = 20 \rightarrow$ 1 in 100

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

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



Call: orange (C)

Estimate *p*, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

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

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Base qualities

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
||||||||||||||||||||||||||||||||
IHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

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

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	*[NULL]*
1	1	1	1	*[START OF HEADING]*
2	2	10	2	*[START OF TEXT]*
3	3	11	3	*[END OF TEXT]*
4	4	100	4	*[END OF TRANSMISSION]*
5	5	101	5	*[ENQUIRY]*
6	6	110	6	*[ACKNOWLEDGE]*
7	7	111	7	*[BELL]*
8	8	1000	10	*[BACKSPACE]*
9	9	1001	11	*[HORIZONTAL TAB]*
10	A	1010	12	*[LINE FEED]*
11	B	1011	13	*[VERTICAL TAB]*
12	C	1100	14	*[FORM FEED]*
13	D	1101	15	*[CARRIAGE RETURN]*
14	E	1110	16	*[SHIFT OUT]*
15	F	1111	17	*[SHIFT IN]*
16	10	10000	20	*[DATA LINK ESCAPE]*
17	11	10001	21	*[DEVICE CONTROL 1]*
18	12	10010	22	*[DEVICE CONTROL 2]*
19	13	10011	23	*[DEVICE CONTROL 3]*
20	14	10100	24	*[DEVICE CONTROL 4]*
21	15	10101	25	*[NEGATIVE ACKNOWLEDGE]*
22	16	10110	26	*[SYNCHRONOUS IDLE]*
23	17	10111	27	*[ENG OF TRANS. BLOCK]*
24	18	11000	30	*[CANCEL]*
25	19	11001	31	*[END OF MEDIUM]*
26	1A	11010	32	*[SUBSTITUTE]*
27	1B	11011	33	*[ESCAPE]*
28	1C	11100	34	*[FILE SEPARATOR]*
29	1D	11101	35	*[GROUP SEPARATOR]*
30	1E	11110	36	*[RECORD SEPARATOR]*
31	1F	11111	37	*[UNIT SEPARATOR]*
32	20	100000	40	*[SPACE]*
33	21	100001	41	**!**
34	22	100010	42	**"**
35	23	100011	43	**#**
36	24	100100	44	**$**
37	25	100101	45	**%**
38	26	100110	46	**&**
39	27	100111	47	**'**
40	28	101000	50	**(**
41	29	101001	51	**)**
42	2A	101010	52	*****
43	2B	101011	53	**+**
44	2C	101100	54	**,**
45	2D	101101	55	**-**
46	2E	101110	56	**.**
47	2F	101111	57	**/**
48	30	110000	60	**0**
49	31	110001	61	**1**
50	32	110010	62	**2**
51	33	110011	63	**3**
52	34	110100	64	**4**
53	35	110101	65	**5**
54	36	110110	66	**6**
55	37	110111	67	**7**
56	38	111000	70	**8**
57	39	111001	71	**9**
58	3A	111010	72	**:**
59	3B	111011	73	**;**
60	3C	111100	74	**<**
61	3D	111101	75	**=**
62	3E	111110	76	**>**
63	3F	111111	77	**?**
64	40	1000000	100	**@**
65	41	1000001	101	**A**
66	42	1000010	102	**B**
67	43	1000011	103	**C**
68	44	1000100	104	**D**
69	45	1000101	105	**E**
70	46	1000110	106	**F**
71	47	1000111	107	**G**
72	48	1001000	110	**H**
73	49	1001001	111	**I**
74	4A	1001010	112	**J**
75	4B	1001011	113	**K**
76	4C	1001100	114	**L**
77	4D	1001101	115	**M**
78	4E	1001110	116	**N**
79	4F	1001111	117	**O**
80	50	1010000	120	**P**
81	51	1010001	121	**Q**
82	52	1010010	122	**R**
83	53	1010011	123	**S**
84	54	1010100	124	**T**
85	55	1010101	125	**U**
86	56	1010110	126	**V**
87	57	1010111	127	**W**
88	58	1011000	130	**X**
89	59	1011001	131	**Y**
90	5A	1011010	132	**Z**
91	5B	1011011	133	**[**
92	5C	1011100	134	****
93	5D	1011101	135	**]**
94	5E	1011110	136	**^**
95	5F	1011111	137	**_**
96	60	1100000	140	**`**
97	61	1100001	141	**a**
98	62	1100010	142	**b**
99	63	1100011	143	**c**
100	64	1100100	144	**d**
101	65	1100101	145	**e**
102	66	1100110	146	**f**
103	67	1100111	147	**g**
104	68	1101000	150	**h**
105	69	1101001	151	**i**
106	6A	1101010	152	**j**
107	6B	1101011	153	**k**
108	6C	1101100	154	**l**
109	6D	1101101	155	**m**
110	6E	1101110	156	**n**
111	6F	1101111	157	**o**
112	70	1110000	160	**p**
113	71	1110001	161	**q**
114	72	1110010	162	**r**
115	73	1110011	163	**s**
116	74	1110100	164	**t**
117	75	1110101	165	**u**
118	76	1110110	166	**v**
119	77	1110111	167	**w**
120	78	1111000	170	**x**
121	79	1111001	171	**y**
122	7A	1111010	172	**z**
123	7B	1111011	173	**{**
124	7C	1111100	174	**|**
125	7D	1111101	175	**}**
126	7E	1111110	176	**~**
127	7F	1111111	177	*[DEL]*

ASCII TABLE

These character don’t print

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	*[NULL]*
1	1	1	1	*[START OF HEADING]*
2	2	10	2	*[START OF TEXT]*
3	3	11	3	*[END OF TEXT]*
4	4	100	4	*[END OF TRANSMISSION]*
5	5	101	5	*[ENQUIRY]*
6	6	110	6	*[ACKNOWLEDGE]*
7	7	111	7	*[BELL]*
8	8	1000	10	*[BACKSPACE]*
9	9	1001	11	*[HORIZONTAL TAB]*
10	A	1010	12	*[LINE FEED]*
11	B	1011	13	*[VERTICAL TAB]*
12	C	1100	14	*[FORM FEED]*
13	D	1101	15	*[CARRIAGE RETURN]*
14	E	1110	16	*[SHIFT OUT]*
15	F	1111	17	*[SHIFT IN]*
16	10	10000	20	*[DATA LINK ESCAPE]*
17	11	10001	21	*[DEVICE CONTROL 1]*
18	12	10010	22	*[DEVICE CONTROL 2]*
19	13	10011	23	*[DEVICE CONTROL 3]*
20	14	10100	24	*[DEVICE CONTROL 4]*
21	15	10101	25	*[NEGATIVE ACKNOWLEDGE]*
22	16	10110	26	*[SYNCHRONOUS IDLE]*
23	17	10111	27	*[ENG OF TRANS. BLOCK]*
24	18	11000	30	*[CANCEL]*
25	19	11001	31	*[END OF MEDIUM]*
26	1A	11010	32	*[SUBSTITUTE]*
27	1B	11011	33	*[ESCAPE]*
28	1C	11100	34	*[FILE SEPARATOR]*
29	1D	11101	35	*[GROUP SEPARATOR]*
30	1E	11110	36	*[RECORD SEPARATOR]*
31	1F	11111	37	*[UNIT SEPARATOR]*
32	20	100000	40	*[SPACE]*
33	21	100001	41	**!**
34	22	100010	42	**"**
35	23	100011	43	**#**
36	24	100100	44	**$**
37	25	100101	45	**%**
38	26	100110	46	**&**
39	27	100111	47	**'**
40	28	101000	50	**(**
41	29	101001	51	**)**
42	2A	101010	52	*****
43	2B	101011	53	**+**
44	2C	101100	54	**,**
45	2D	101101	55	**-**
46	2E	101110	56	**.**
47	2F	101111	57	**/**
48	30	110000	60	**0**
49	31	110001	61	**1**
50	32	110010	62	**2**
51	33	110011	63	**3**
52	34	110100	64	**4**
53	35	110101	65	**5**
54	36	110110	66	**6**
55	37	110111	67	**7**
56	38	111000	70	**8**
57	39	111001	71	**9**
58	3A	111010	72	**:**
59	3B	111011	73	**;**
60	3C	111100	74	**<**
61	3D	111101	75	**=**
62	3E	111110	76	**>**
63	3F	111111	77	**?**
64	40	1000000	100	**@**
65	41	1000001	101	**A**
66	42	1000010	102	**B**
67	43	1000011	103	**C**
68	44	1000100	104	**D**
69	45	1000101	105	**E**
70	46	1000110	106	**F**
71	47	1000111	107	**G**
72	48	1001000	110	**H**
73	49	1001001	111	**I**
74	4A	1001010	112	**J**
75	4B	1001011	113	**K**
76	4C	1001100	114	**L**
77	4D	1001101	115	**M**
78	4E	1001110	116	**N**
79	4F	1001111	117	**O**
80	50	1010000	120	**P**
81	51	1010001	121	**Q**
82	52	1010010	122	**R**
83	53	1010011	123	**S**
84	54	1010100	124	**T**
85	55	1010101	125	**U**
86	56	1010110	126	**V**
87	57	1010111	127	**W**
88	58	1011000	130	**X**
89	59	1011001	131	**Y**
90	5A	1011010	132	**Z**
91	5B	1011011	133	**[**
92	5C	1011100	134	****
93	5D	1011101	135	**]**
94	5E	1011110	136	**^**
95	5F	1011111	137	**_**
96	60	1100000	140	**`**
97	61	1100001	141	**a**
98	62	1100010	142	**b**
99	63	1100011	143	**c**
100	64	1100100	144	**d**
101	65	1100101	145	**e**
102	66	1100110	146	**f**
103	67	1100111	147	**g**
104	68	1101000	150	**h**
105	69	1101001	151	**i**
106	6A	1101010	152	**j**
107	6B	1101011	153	**k**
108	6C	1101100	154	**l**
109	6D	1101101	155	**m**
110	6E	1101110	156	**n**
111	6F	1101111	157	**o**
112	70	1110000	160	**p**
113	71	1110001	161	**q**
114	72	1110010	162	**r**
115	73	1110011	163	**s**
116	74	1110100	164	**t**
117	75	1110101	165	**u**
118	76	1110110	166	**v**
119	77	1110111	167	**w**
120	78	1111000	170	**x**
121	79	1111001	171	**y**
122	7A	1111010	172	**z**
123	7B	1111011	173	**{**
124	7C	1111100	174	**|**
125	7D	1111101	175	**}**
126	7E	1111110	176	**~**
127	7F	1111111	177	*[DEL]*

Value	Character	Value	Character	Value	Character	Value	Character	Value	Character	Value	Character
0	Null	22	Synchronous Idle	44	,	66	B	88	X	110	n
1	Start of Heading	23	End of Transmission Block	45	-	67	C	89	Y	111	o
2	Start of Text	24	Cancel	46	.	68	D	90	Z	112	p
3	End of Text	25	End of Medium	47	/	69	E	91	[	113	q
4	End of Transmission	26	Substitute	48	0	70	F	92	\	114	r
5	Enquiry	27	Escape	49	1	71	G	93	]	115	s
6	Acknowledgement	28	File Separator	50	2	72	H	94	^	116	t
7	Bell (Causes an alert sound)	29	Group Separator	51	3	73	I	95	_	117	u
8	Backspace	30	Record Separator	52	4	74	J	96	@	118	v
9	Horizontal Tab	31	Unit Separator	53	5	75	K	97	a	119	w
10	Line Feed	32	space	54	6	76	L	98	b	120	x
11	Vertical Tab	33	!	55	7	77	M	99	c	121	y
12	Form Feed	34	"	56	8	78	N	100	d	122	z
13	Carriage Return	35	#	57	9	79	O	101	e	123	{
14	Shift Out	36	$	58	:	80	P	102	f	124	|
15	Shift In	37	%	59	;	81	Q	103	g	125	}
16	Data Link Escape	38	&	60	<	82	R	104	h	126	~
17	Device Control 1 (often XON)	39	'	61	=	83	S	105	i	127	Delete
18	Device Control 2	40	(	62	>	84	T	106	j		
19	Device Control 3 (often XOFF)	41	)	63	?	85	U	107	k		
20	Device Control 4	42	*	64	@	86	V	108	l		
21	Negative Acknowledgement	43	+	65	A	87	W	109	m		

Base 33 (typical)

Base 64 (old/rare)

Value	Character	Value	Character	Value	Character	Value	Character	Value	Character	Value	Character
0	Null	22	Synchronous Idle	44	,	66	B	88	X	110	n
1	Start of Heading	23	End of Transmission Block	45	-	67	C	89	Y	111	o
2	Start of Text	24	Cancel	46	.	68	D	90	Z	112	p
3	End of Text	25	End of Medium	47	/	69	E	91	[	113	q
4	End of Transmission	26	Substitute	48	0	70	F	92	\	114	r
5	Enquiry	27	Escape	49	1	71	G	93	]	115	s
6	Acknowledgement	28	File Separator	50	2	72	H	94	^	116	t
7	Bell (Causes an alert sound)	29	Group Separator	51	3	73	I	95	_	117	u
8	Backspace	30	Record Separator	52	4	74	J	96	@	118	v
9	Horizontal Tab	31	Unit Separator	53	5	75	K	97	a	119	w
10	Line Feed	32	space	54	6	76	L	98	b	120	x
11	Vertical Tab	33	!	55	7	77	M	99	c	121	y
12	Form Feed	34	"	56	8	78	N	100	d	122	z
13	Carriage Return	35	#	57	9	79	O	101	e	123	{

Base 33 (typical)

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
|||||||||||||||||||||||||||||||||
IHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

```
Value - base = Phred
  73  -  33  = 40
```

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

Phred	Error Probability	Confidence
0	1 / 1	0%
10	1 / 10	90%
20	1 / 100	99%
30	1 / 1000	99.9%
40	1 / 10000	99.99%
50	1 / 100000	99.999%
60	1 / 1000000	99.9999%

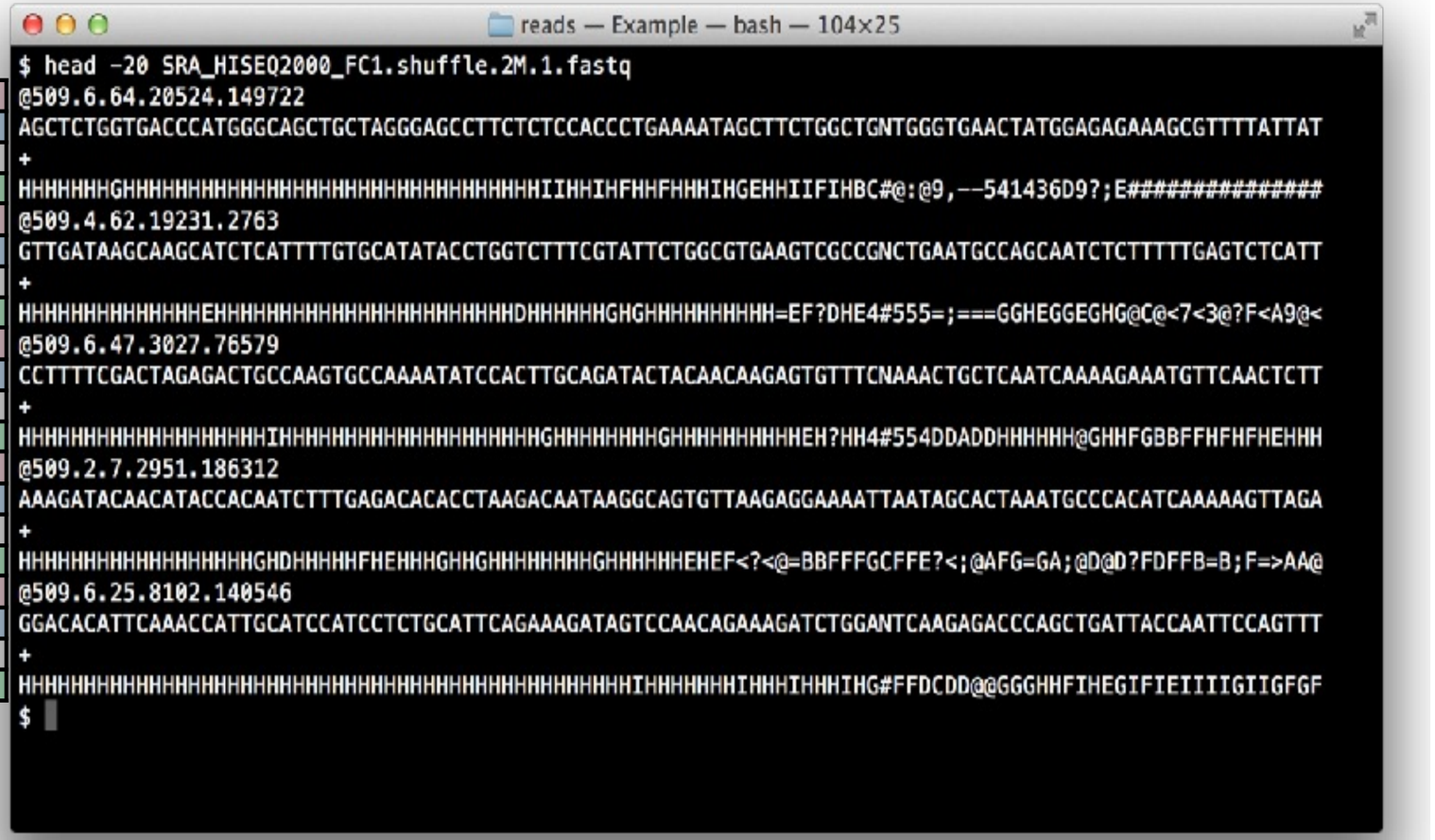
A read in FASTQ format

Name	@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:3607 2:Y:18:ATCACG
Sequence	ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
(ignore)	+
Base qualities	?@@FFBFFDDHHBCEAFGEGIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEGG

Always starts with “@”
ERR194146.1HSQ1008:141:D0CC8ACXX:3 – Machine, Run, Flowcell, Lane
1308:20201:3607 – Tile, X-pos, Y-pos
2:Y:18:ATCACG – Direction, Filtered?, Control bits, Index/Sample

FASTQ

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



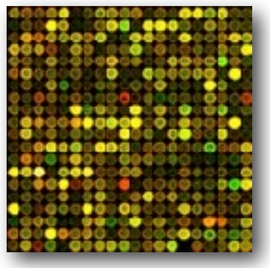
Sample_S1_L001_R1_001.fastq.gz

Sample: Sample name
S1: Sample number
L001: Lane number
R1: Read direction (If starts with “I”, Index read direction)
001: File number (always 001 on modern systems)

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

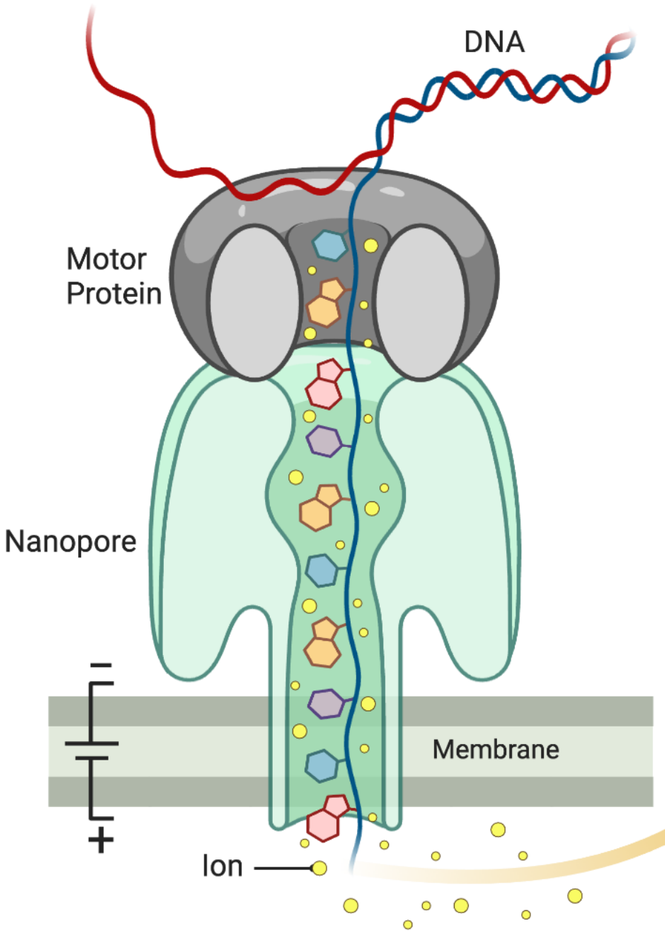
Solexa

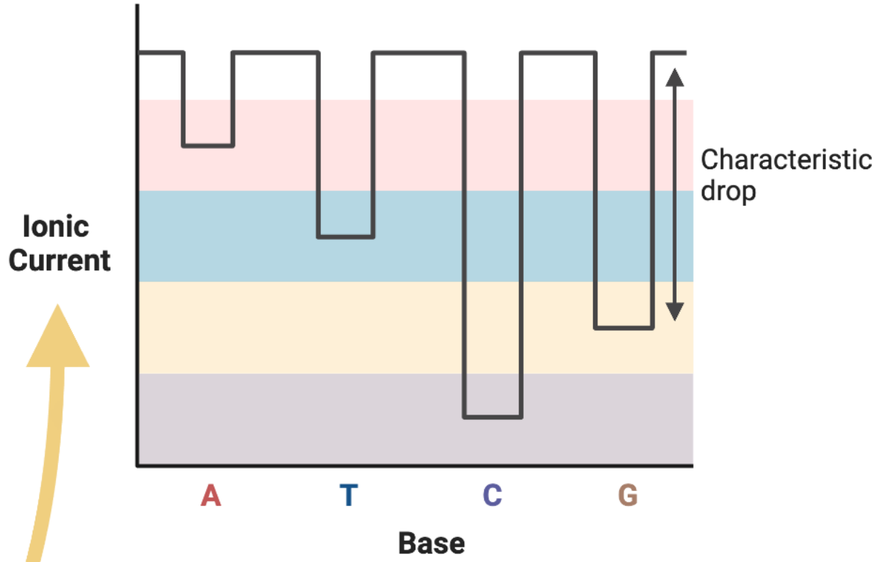
illumina®

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

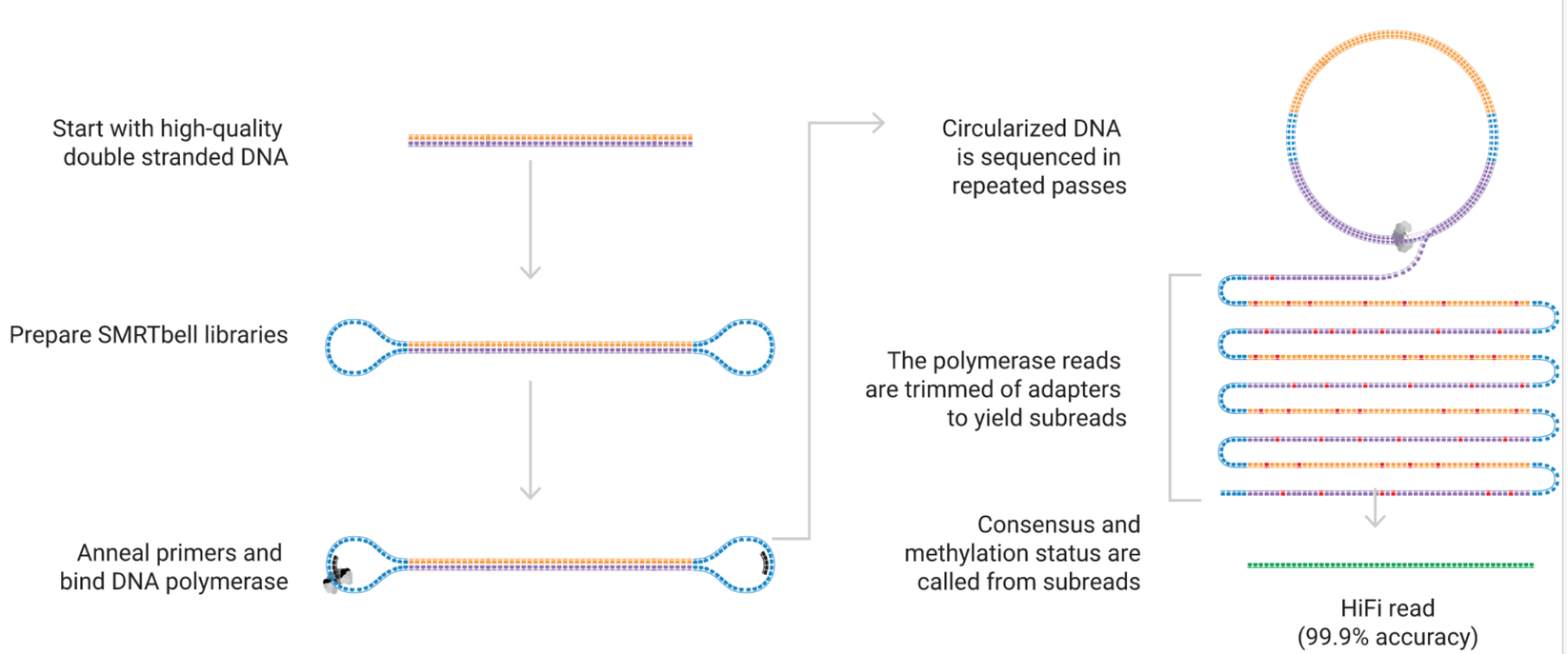
Nanopore Sequencing

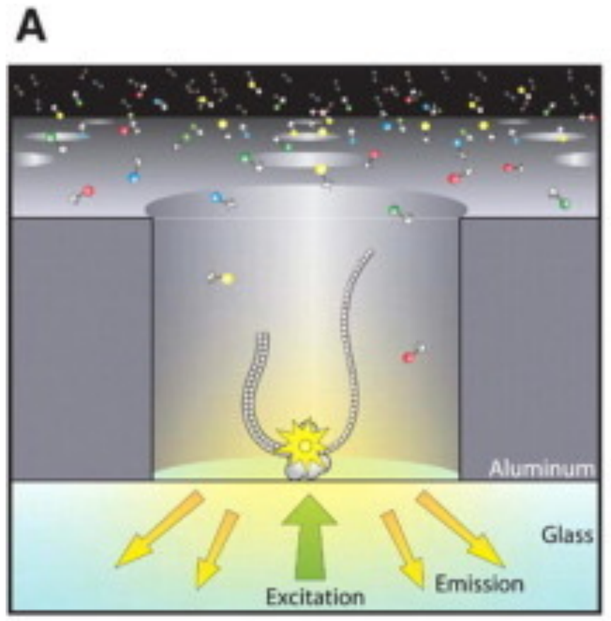
1 DNA is unwound by the motor protein and one strand is translocated through the pore to the +ve side of membrane



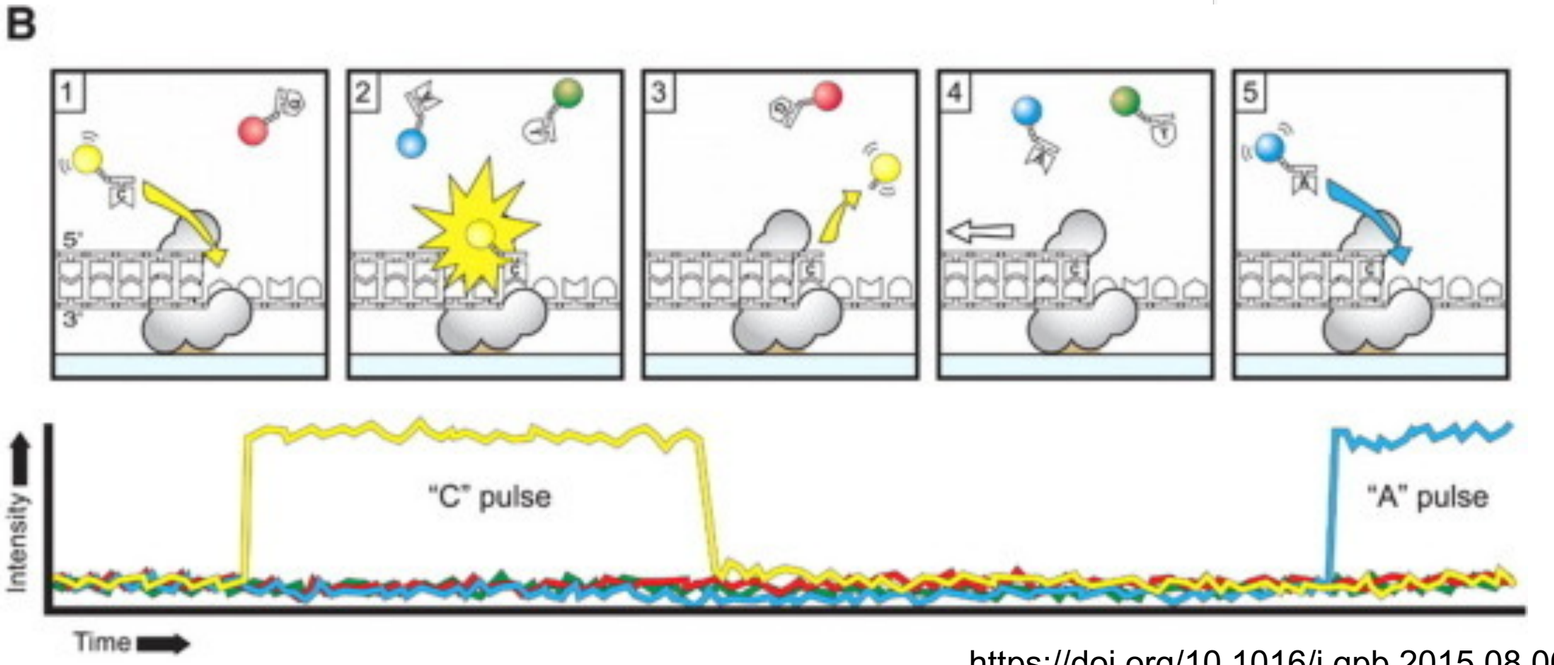


2 Each base gives a characteristic reduction in the ionic current, allowing the DNA to be sequenced





ZMW (**z**ero **m**ode **w**aveguide)



https://doi.org/10.1016/j.gpb.2015.08.002

		Illumina						
	Sanger	**MiSeq**	**NextSeq**	**HiSeq**	**NovaSeq**	**Ion Torrent**	**Pacific Biosciences**	**Oxford Nanopore**
Throughput range per run (Gb)	*c.* 0.0005	10–15	10–120	1000–1800	2000–6000	1–15	0.5–10	0.1–1
Read length	Up to 1 kb	300	150	150	250	200–400	up to 60 kb	up to 100 kb
Read type	SR	SR, PE	SR, PE	SR, PE	SR, PE	SR	SR	SR
Error rate (%)	0.001–1	0.8	0.8	0.2	0.2–0.8	1–2	13	5–40
Error type	Substitutions	Substitutions	Substitutions	Substitutions	Substitutions	Indels, homopolymers	Indels	Indels, deletions
Advantages	Read accuracy and length	Read length	Throughput	Throughput, low error rate	High throughput	Speed, read length	Speed, read length	Read length, portability

Proxima aula…

18-Oct Bioinformática - Linux - Processamento de dados de sequenciamento (teórico-prática)

Google Cloud Shell