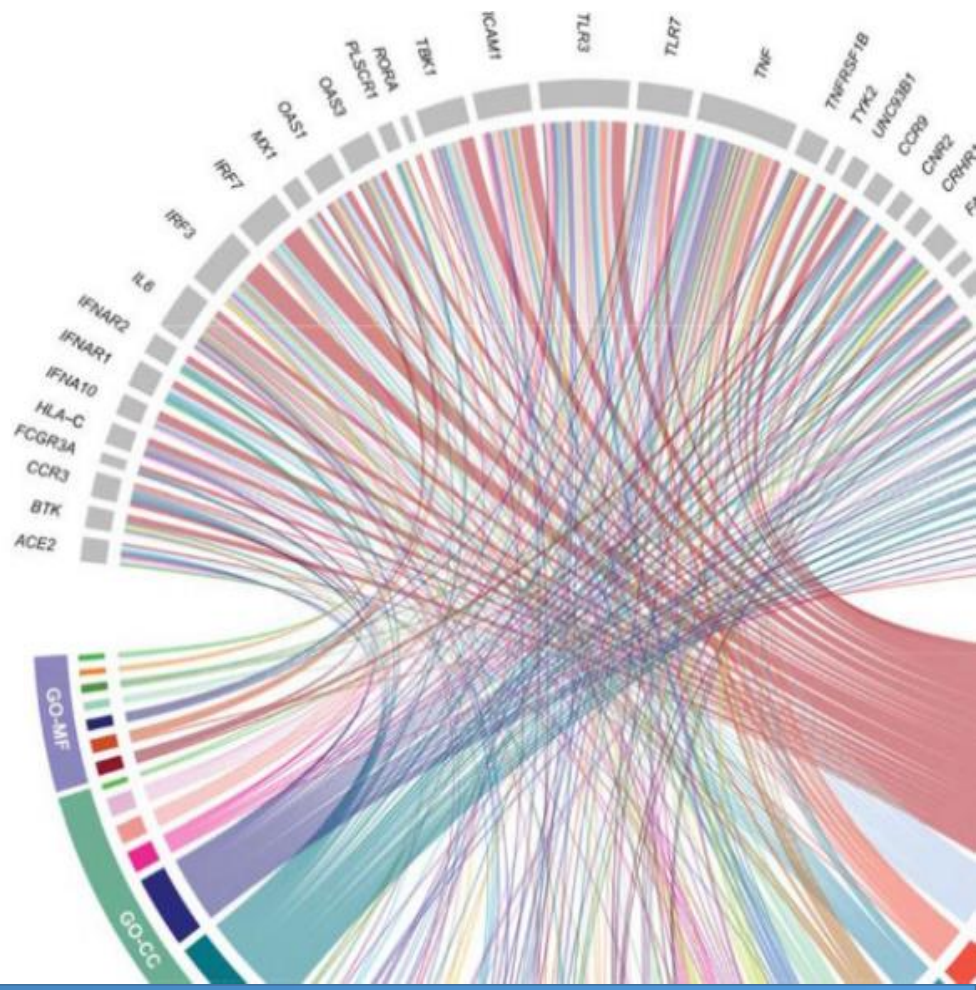
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Tècniques i Eines Bioinformàtiques

Sequence Data Analysis



Santiago Marco-Sola (santiago.marco@upc.edu)

Màster en Enginyeria Informàtica, UPC
Departament of Computer Science
Facultat d'Informàtica de Barcelona (FIB), UPC

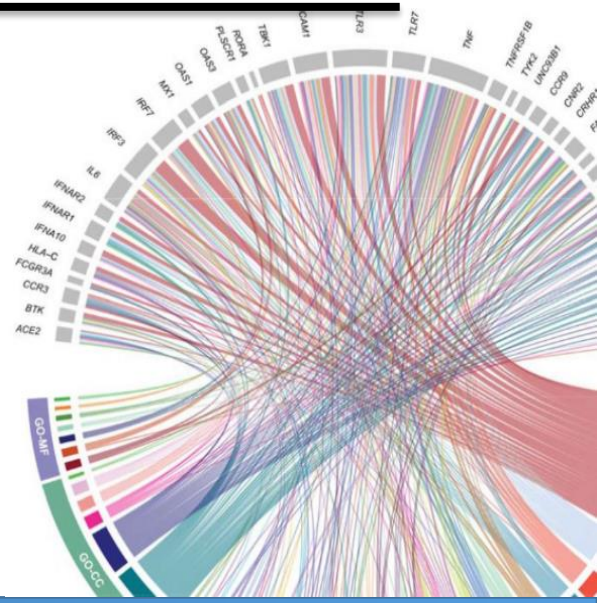
Acknowledgements

Many pictures and materials are taken from **Ben Langmead's course**.

Course heavily inspired in:

- **Genome-Scale Algorithm Design**. Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu. Cambridge University Press.
- **Algorithms on Strings, Trees, and Sequences**. Dan Gusfield. Cambridge University Press.
- **An Introduction to Bioinformatics Algorithms.** Neil C. Jones, Pavel A. Pevzner. MIT Press.

1 Brief Introduction to Genomics

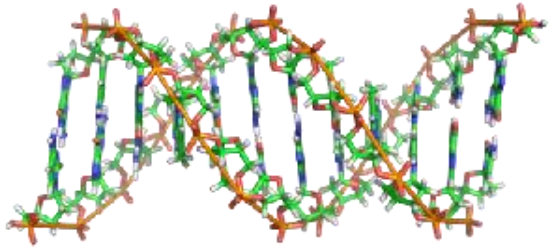


Genome

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

Oxford dictionaries

"Blueprint" or "recipe" of life



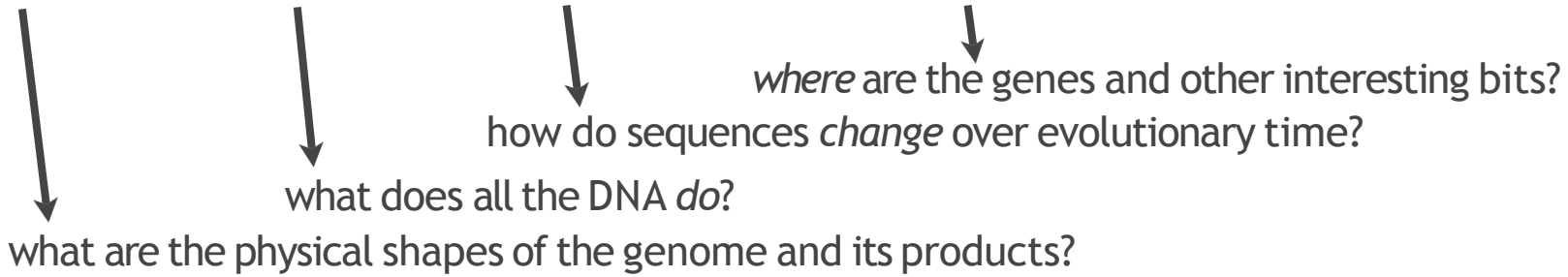
Self-copying store of read-only information about how to develop and maintain an organism

TAGCCCGACTTG

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."

Genomics: contrast with biology & genetics*

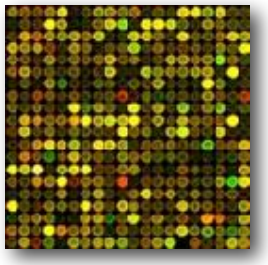
*This slide has gross generalizations

Biology & Genetics		Genomics
Targeted studies of one or a few genes	scope	Studies considering all genes in a genome
Targeted, low-throughput experiments	technology	Global, high-throughput experiments
Clever experimental design, painstaking experimentation	hard part	Tons of data, uncertainty, computation

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)

Genomics: tool 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?

Genomics: tool 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?

Computational Genomics

- Addresses crucial problems at the intersection of genomics and computer science

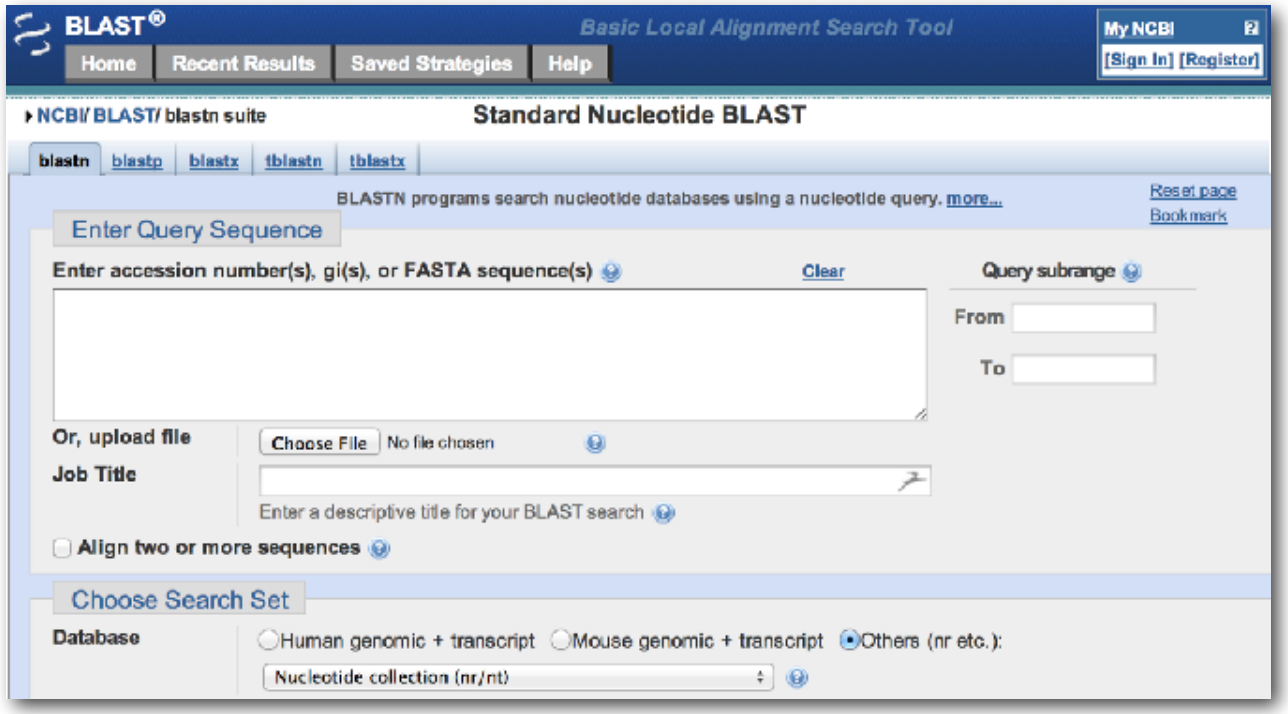
- The intersection:

- 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**

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, …

Computational Genomics: success stories



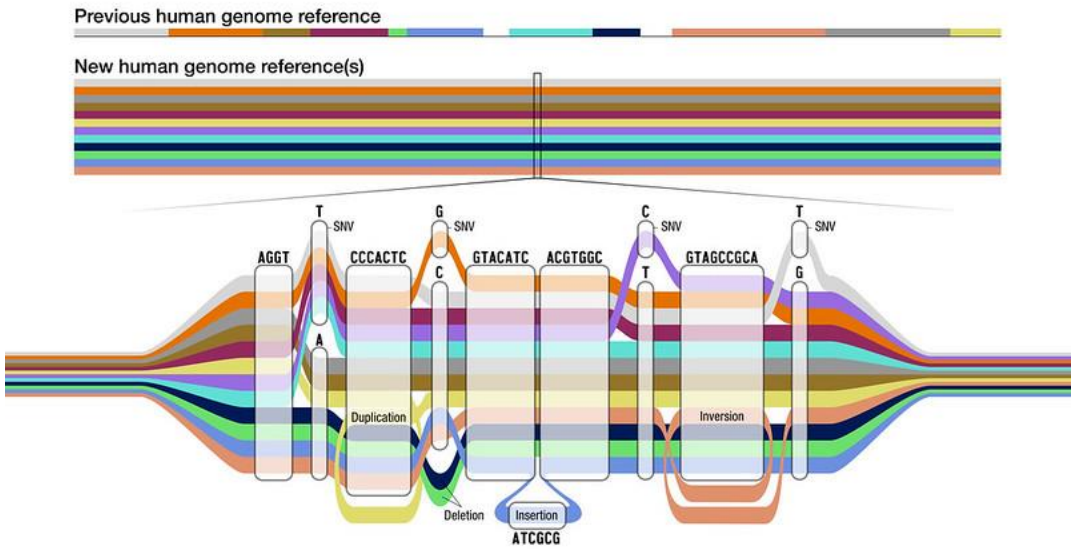
http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

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

Computational Genomics: success stories

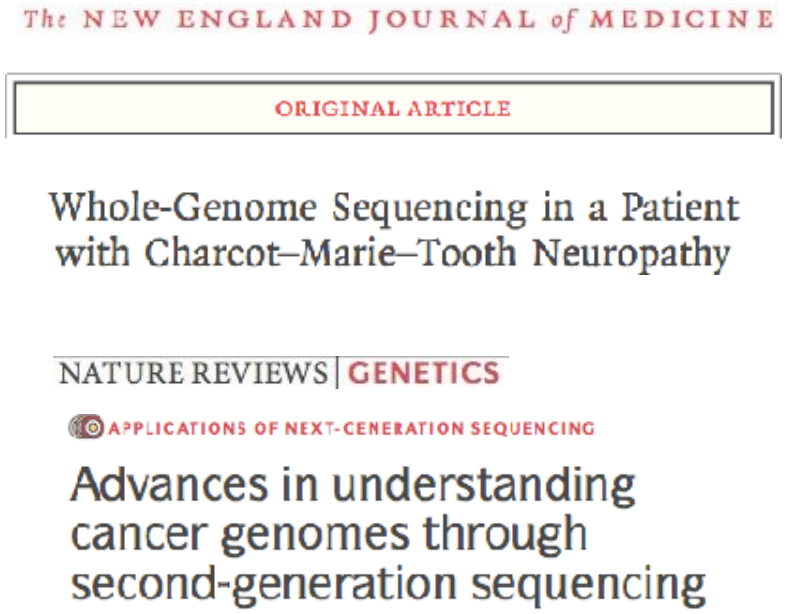
The Human Genome Project depended crucially on contributions by computer scientists, especially new methods for assembling DNA fragments into chromosomes.

Computational Genomics: success stories



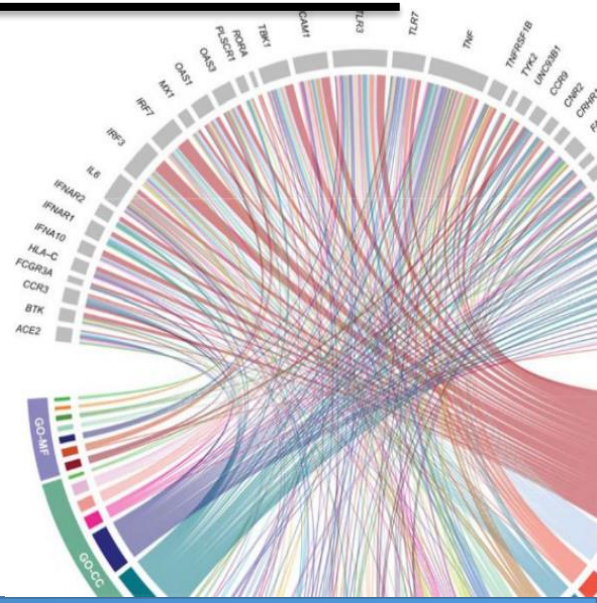
Human Pangenome Reference Consortium has released a high-quality collection of human genome reference sequences, which together comprise a human “pangenome” reference.
Encompassing **genome sequences from 47 people of diverse ancestries** (with the goal of increasing that number to 350 by mid-2024), the human pangenome reference captures significantly more population diversity than the previous reference sequence.

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.

2 Sequencing Technologies

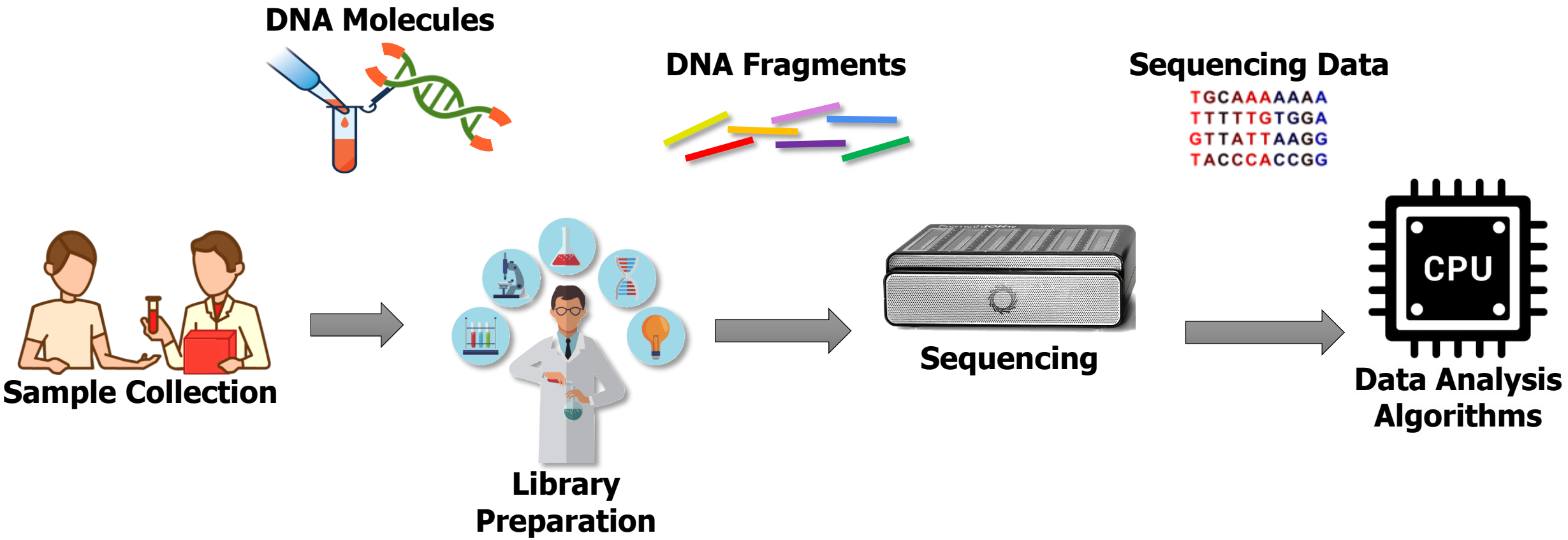


All the Information lies in the DNA

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.

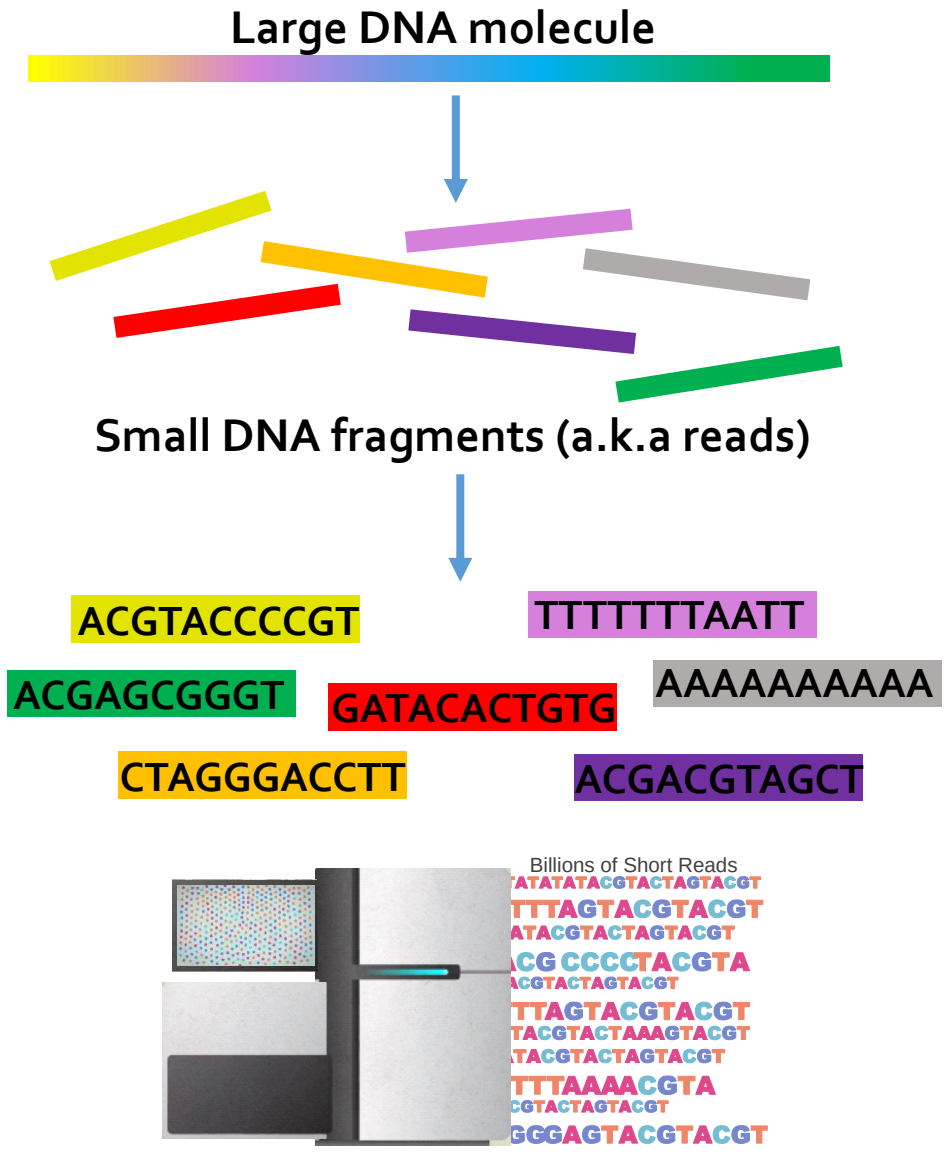
Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.

From Sequencing (Wet-Lab) to Analysis (Computation)



Sequencing Machines (Translating Molecules to Digital Data)

- **What Do Sequencing Machines Do?**
  - Convert biological molecules (DNA/RNA) into digital data
  - Allow computational analysis of genetic information
- **How Do They Work? Requirements?**
  - DNA Preparation: Sample processing and library preparation
  - Fragmentation: DNA is chopped into smaller pieces
  - Sequencing Reaction: Reads the DNA bases
  - Signal Conversion: Optical/electrical signals into ASCII DNA characters (A, T, C, G)
- **A DNA Sequencer is Like a Chopper Machine**
  - Breaks DNA into readable pieces
  - Repeated pieces (PCR) and sequencing errors introduced



Sequencing technologies

No sequencing technology yet invented can read much more than 10,000 nucleotides at a time with reasonable cost, throughput, accuracy

Instead, there's a vigorous race to see whose sequencer can read "short" fragments of DNA (around 100s of nucleotides) with best cost, throughput, accuracy

Decoding DNA With Semiconductors

By NICHOLAS WADE
Published: July 20, 2011

Company Unveils DNA Sequencing Device Meant to Be Portable, Disposable and Cheap

By ANDREW POLLACK
Published: February 17, 2012

Cost of Gene Sequencing Falls, Raising Hopes for Medical Advances

By JOHN MARKOFF
Published: March 7, 2012

Source: nytimes.com

Multiple Sequencing Technologies

Short-Read Technologies

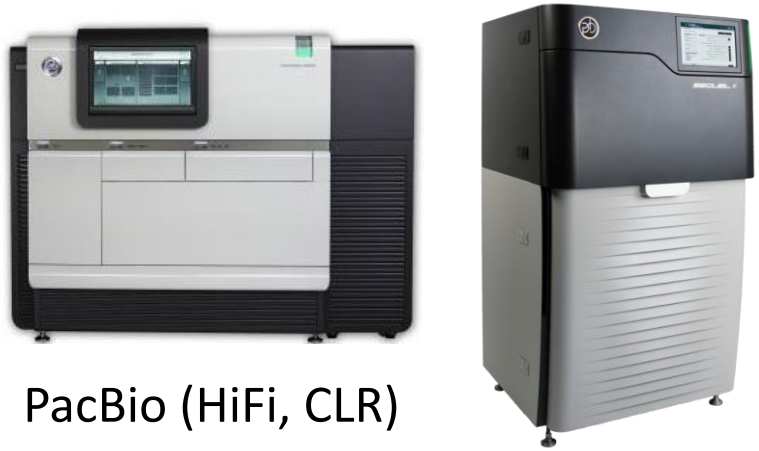
Long-Read Technologies

MiSeq NextSeq 500 AVITI NovaSeq 6000

Illumina Sequencing Technologies

Oxford Nanopore Technologies (ONT)

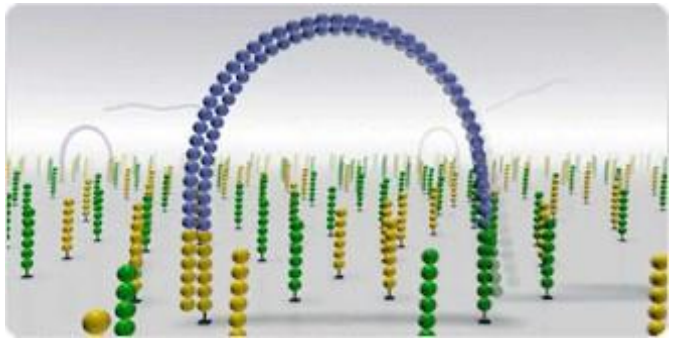
BGI/MGI Sequencing (DNBSEQ - DNA Nanoball)



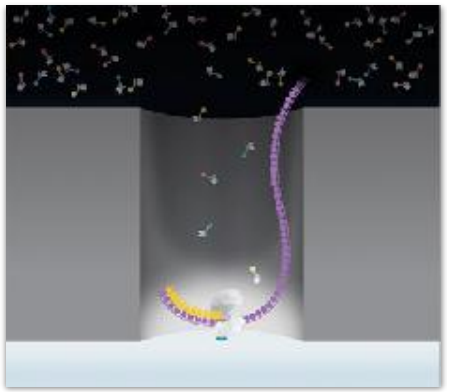
PacBio (HiFi, CLR)

Multiple Sequencing Technologies

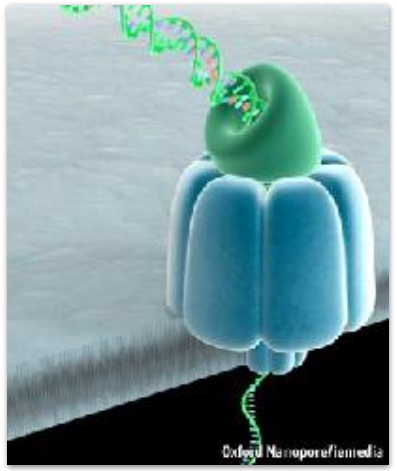
- Since 2005, many DNA sequencing instruments have been described and released. They are based on a few different principles
- Each technology produce data with different properties (e.g., throughput, sequence length, error-rate/quality).
  - Sequencing by synthesis ("massively parallel sequencing") provides greatest throughput, and is the most prevalent today
  - SMRT and Nanopore produce longer reads but higher error-rate (i.e., less quality)



Synthesis / ligation



SMRT cell

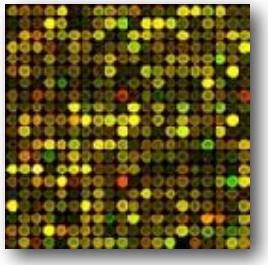


Nanopore

Pictures: http://www.illumina.com/systems/miseq/technology.ilmn, http://www.genengnews.com/gen-articles/third-generation-sequencing-debuts/3257/

Sequencing Generations

Sanger DNA sequencing
1977-1990s

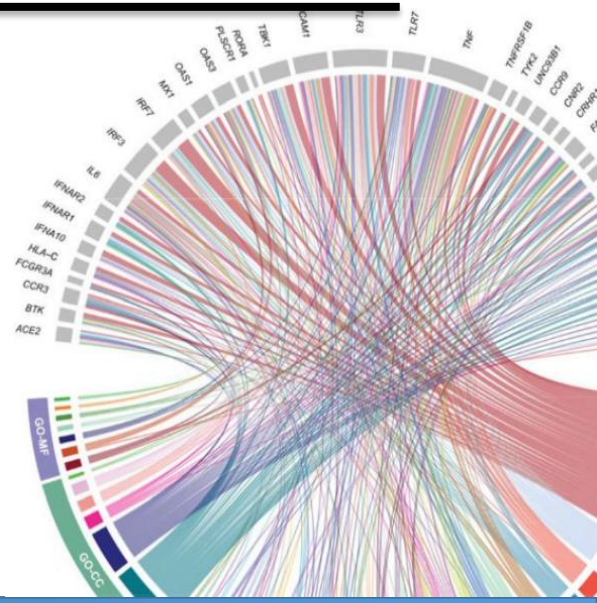


DNA Microarrays
Since mid-1990s

2nd-generation DNA sequencing
Since ~2007

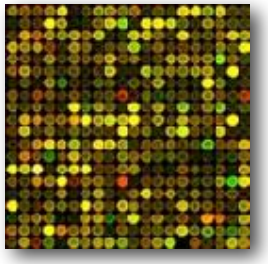
3rd-generation single-molecule DNA sequencing
Since ~2010

3 Sanger Sequencing



Genomics technology

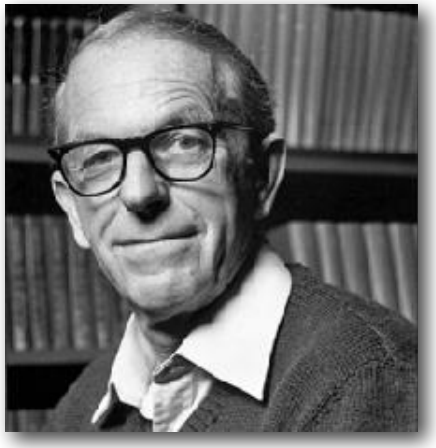
Sanger DNA sequencing
1977-1990s



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

https://www.youtube.com/watch?v=6ldtdWjDwes

https://www.youtube.com/watch?v=dVRB4CaLizc

From "DNA" documentary, episode 3

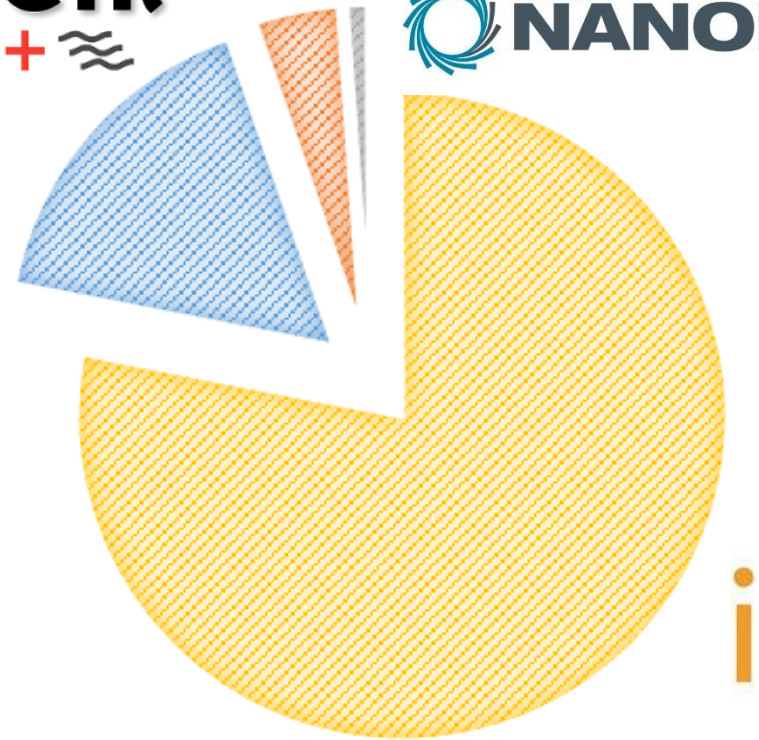
Sequencing Technologies by Market Share

ion torrent

PACBIO®

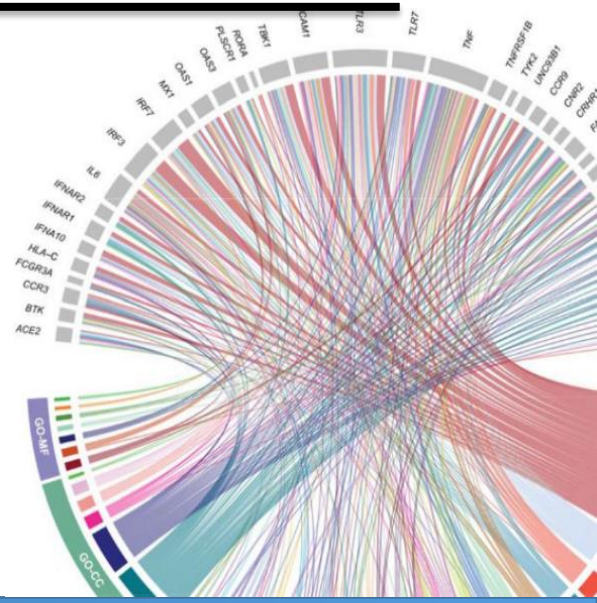
Oxford NANOPORE Technologies

BGI华大



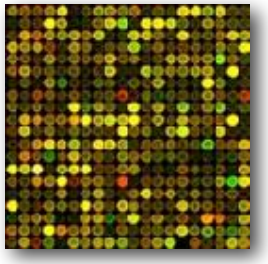
illumina®

4 Sequencing by Synthesis (Illumina)



Sequencing technologies

Sanger DNA sequencing
1977-1990s



DNA Microarrays
Since mid-1990s

2nd-generation DNA sequencing
Since ~2007

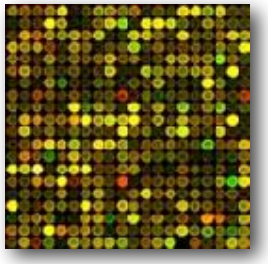
3rd-generation single-molecule DNA sequencing
Since ~2010

Solexa

illumina®

Sequencing technologies

Sanger DNA sequencing
1977-1990s



DNA Microarrays
Since mid-1990s

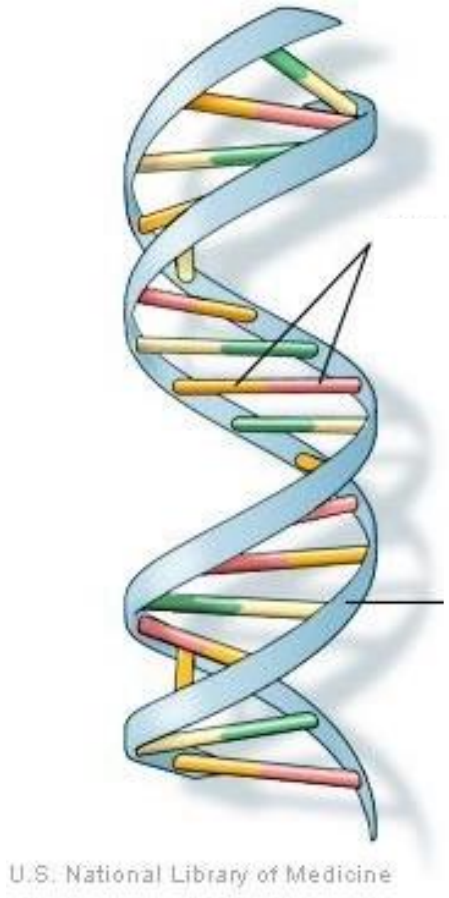
2nd-generation DNA sequencing
Since ~2007

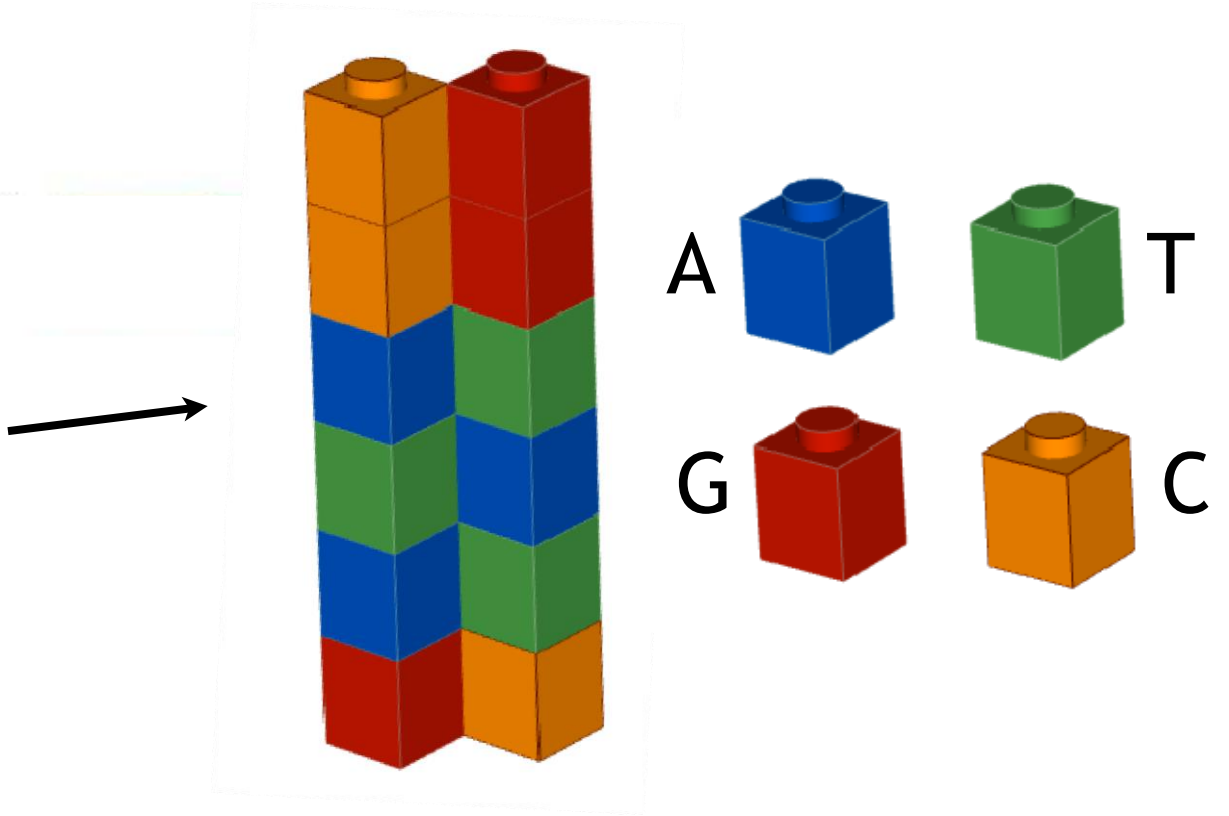
3rd-generation single-molecule DNA sequencing
Since ~2010

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Solexa

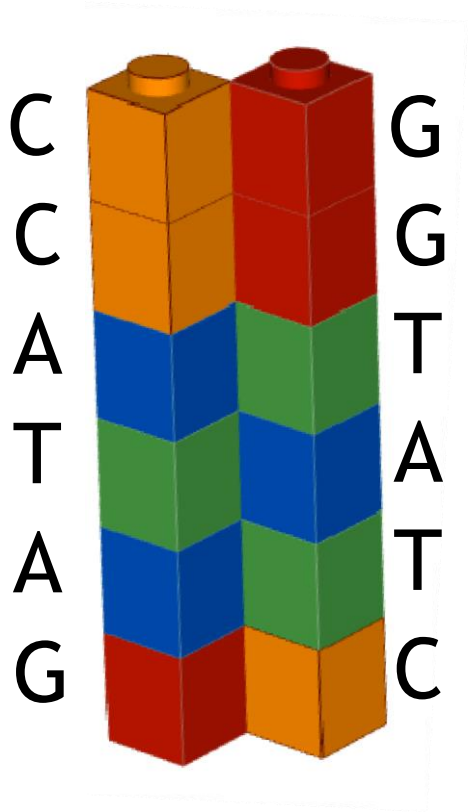
illumina®

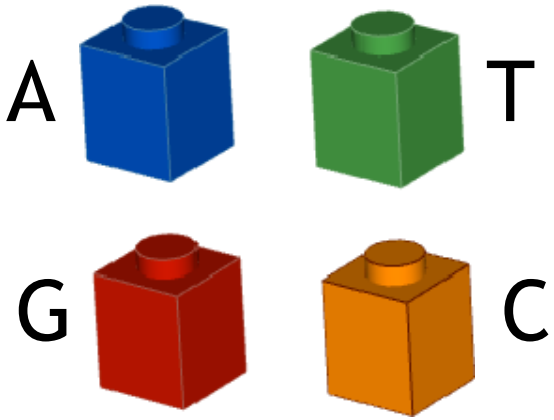


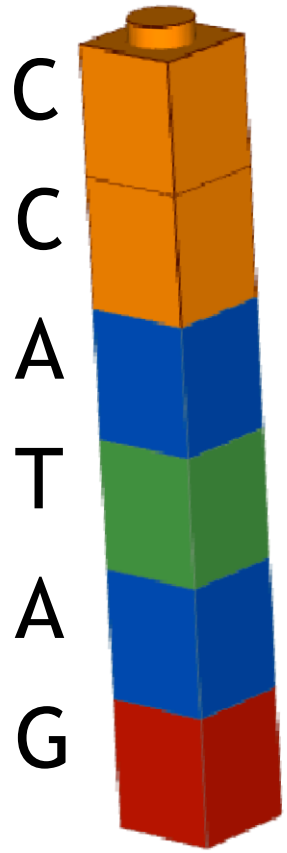


Double stranded DNA (double helix)

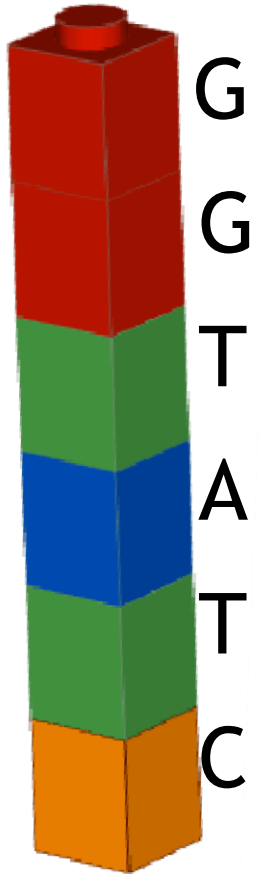
Double stranded DNA (lego version)

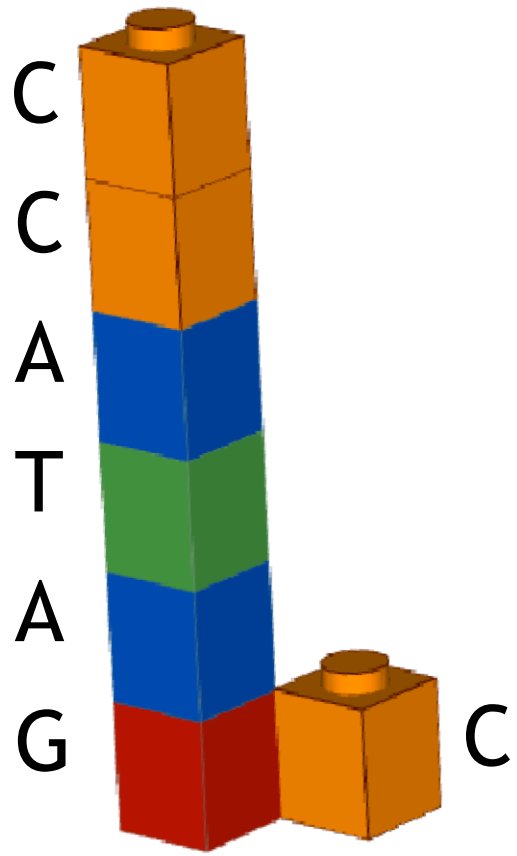

C G
C G
A T
T A
A T
G C

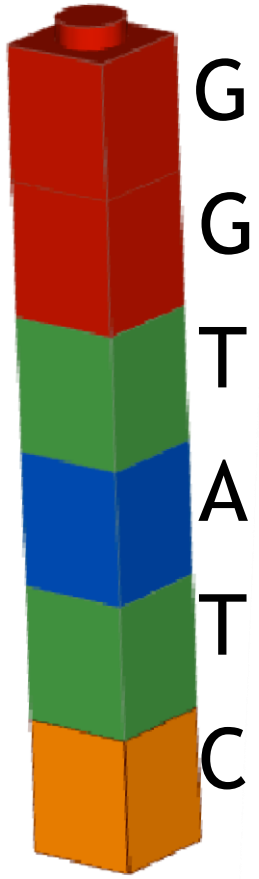

A
T
G
C

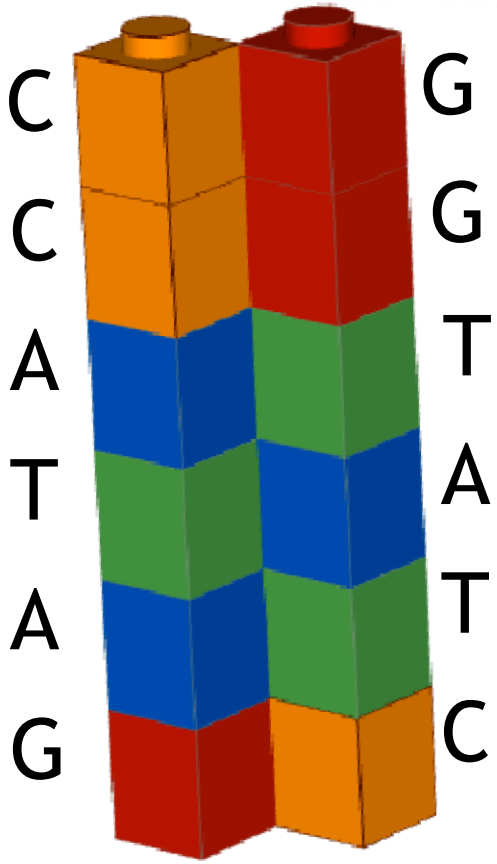


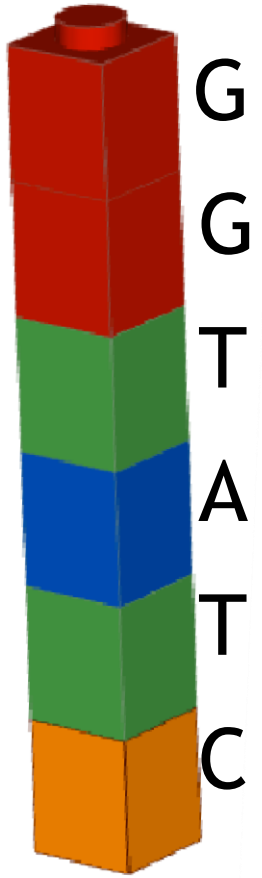
Single stranded templates

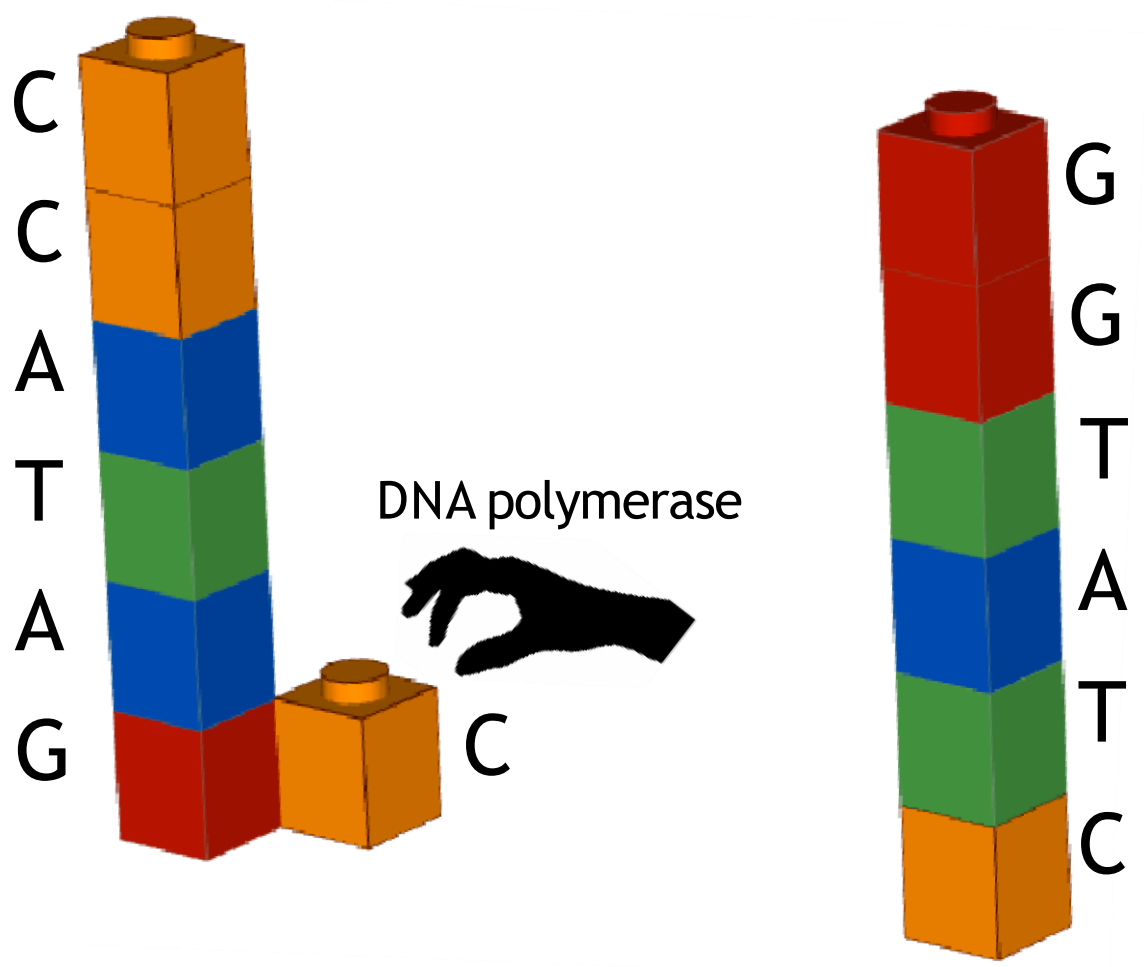


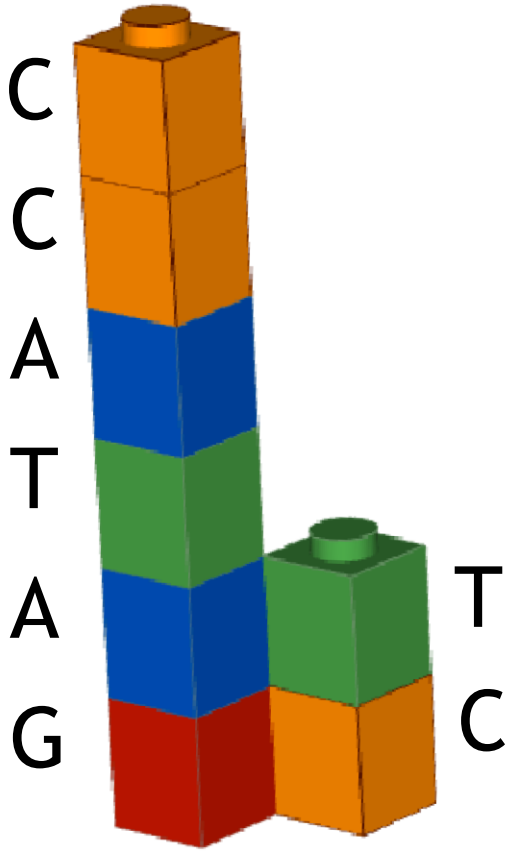

C
C
A
T
A
G
C

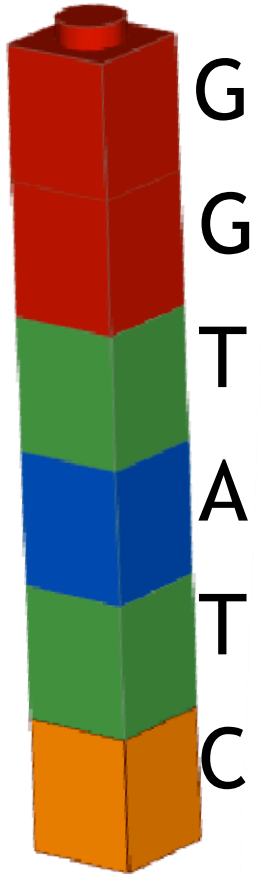

G
G
T
A
T
C

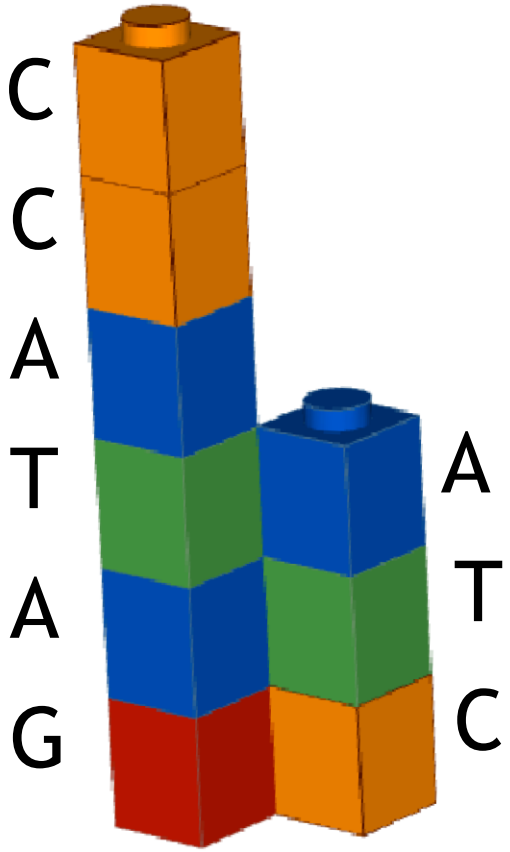

C
C
A
T
A
G
G
G
T
A
T
C

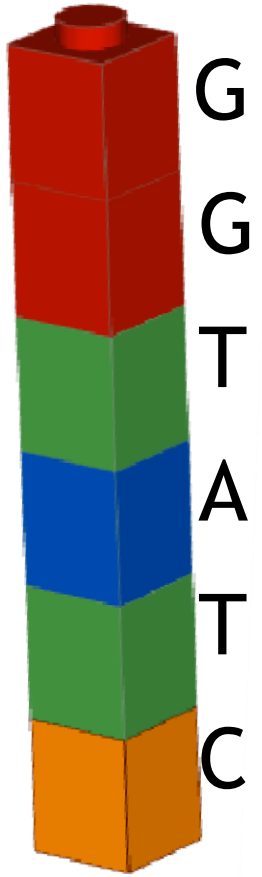

G
G
T
A
T
C

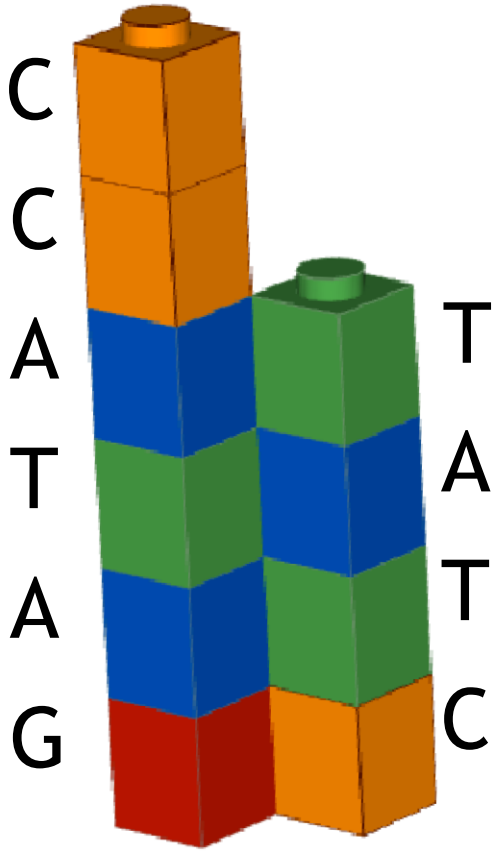

C
C
A
T
A
G
DNA polymerase
C
G
G
T
A
T
C

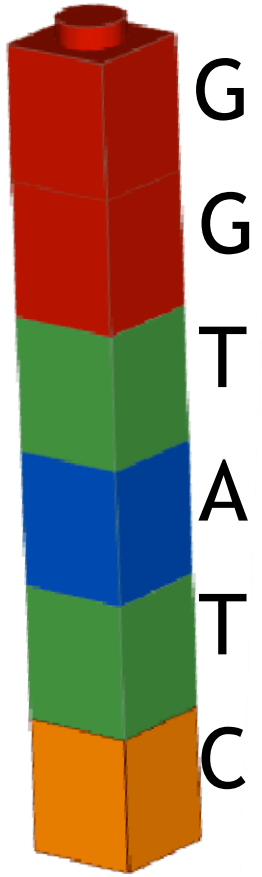

C
C
A
T
A
G
T
C

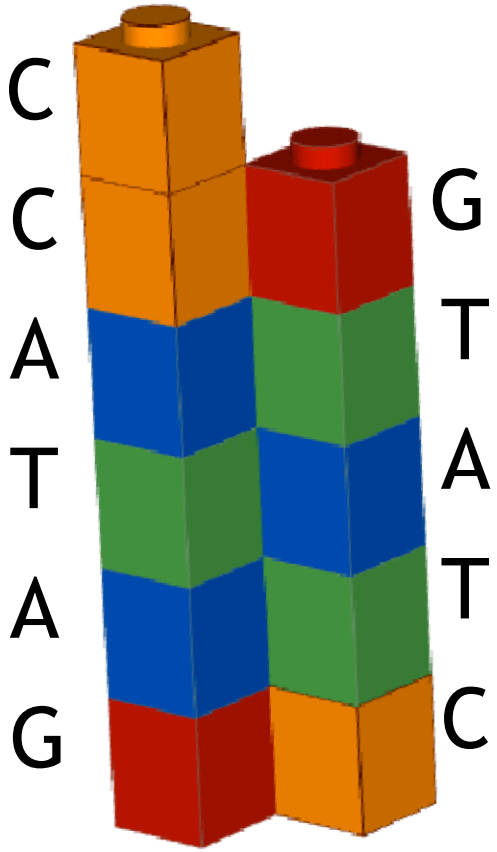

G
G
T
A
T
C

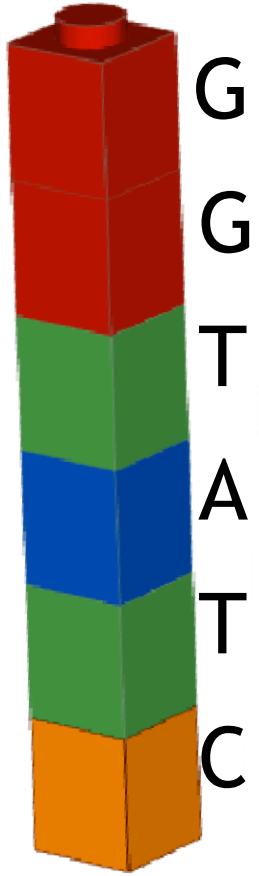

C
C
A
T
A
G
A
T
C

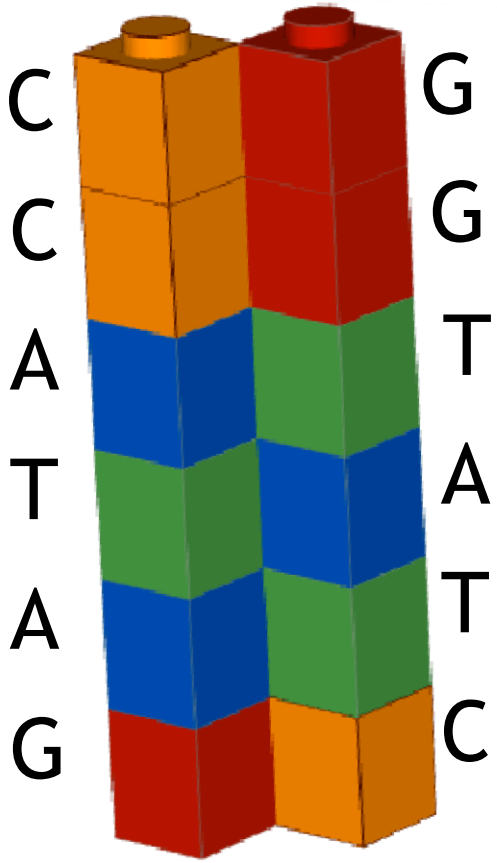

G
G
T
A
T
C

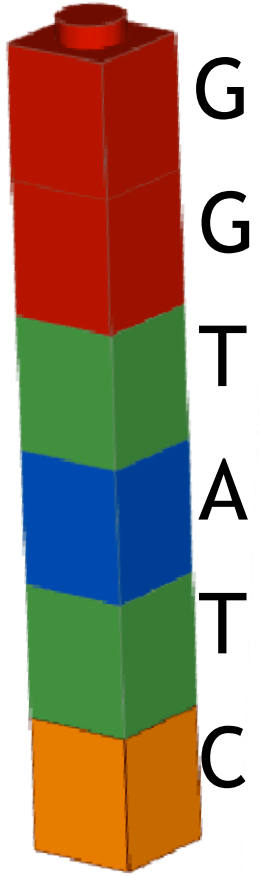

C
C
A
T
A
G
T
A
T
C

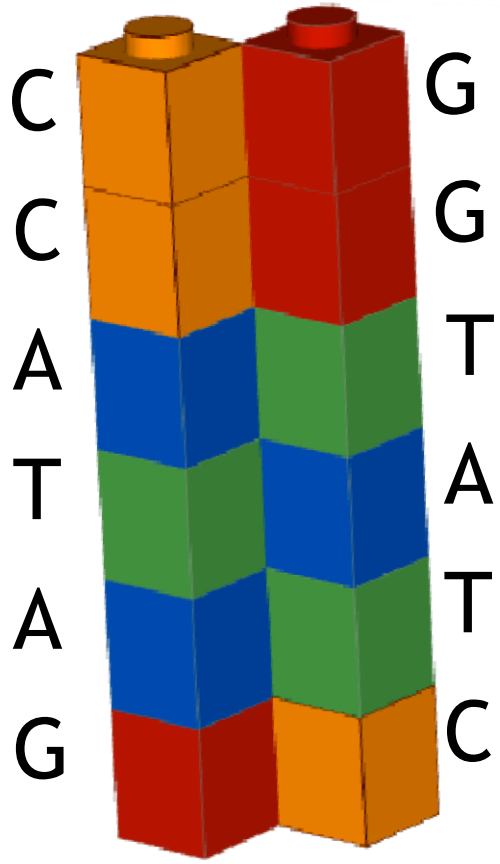

G
G
T
A
T
C

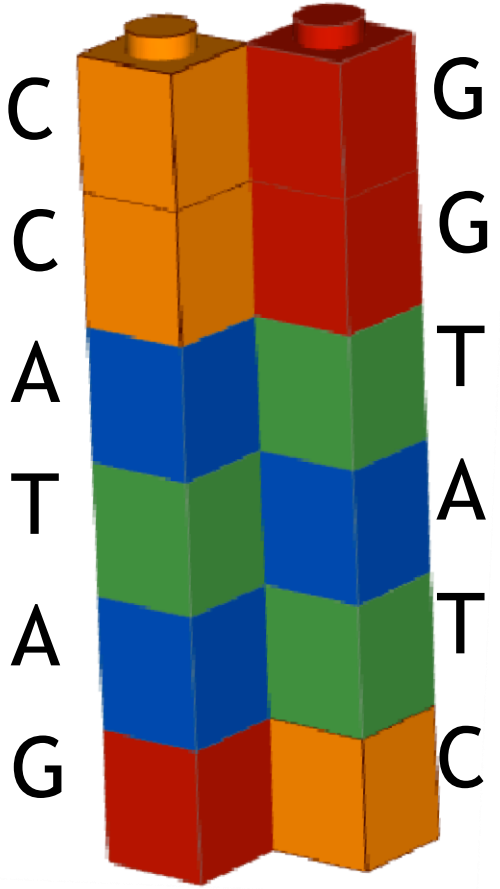

C
C
A
T
A
G
G
T
A
T
C


G
G
T
A
T
C


C
C
A
T
A
G
G
G
T
A
T
C


G
G
T
A
T
C


C
C
A
T
A
G
G
G
T
A
T
C


C
C
A
T
A
G
G
G
T
A
T
C

Input DNA

CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT

Cut into snippets

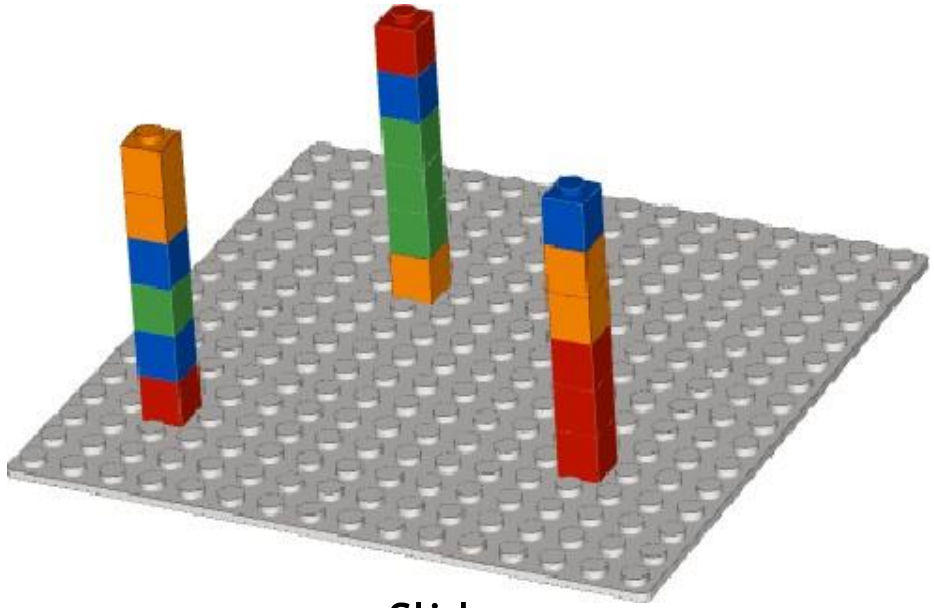
CCATAGTA TATCTCGG CTCTAGGCCCTC ATTTTTT
CCA TAGTATAT CTCGGCTCTAGGCCCTCA TTTTTT
CCATAGTAT ATCTCGGCTCTAG GCCCTCA TTTTTT
CCATAG TATATCT CGGCTCTAGGCCCT CATTTTTT

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

Template
(billions of them!)



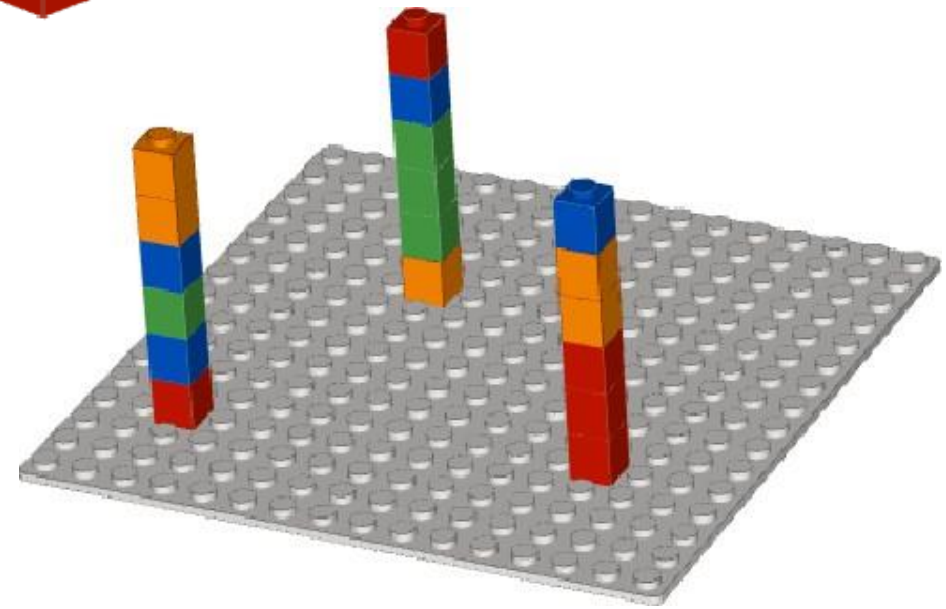
Slide

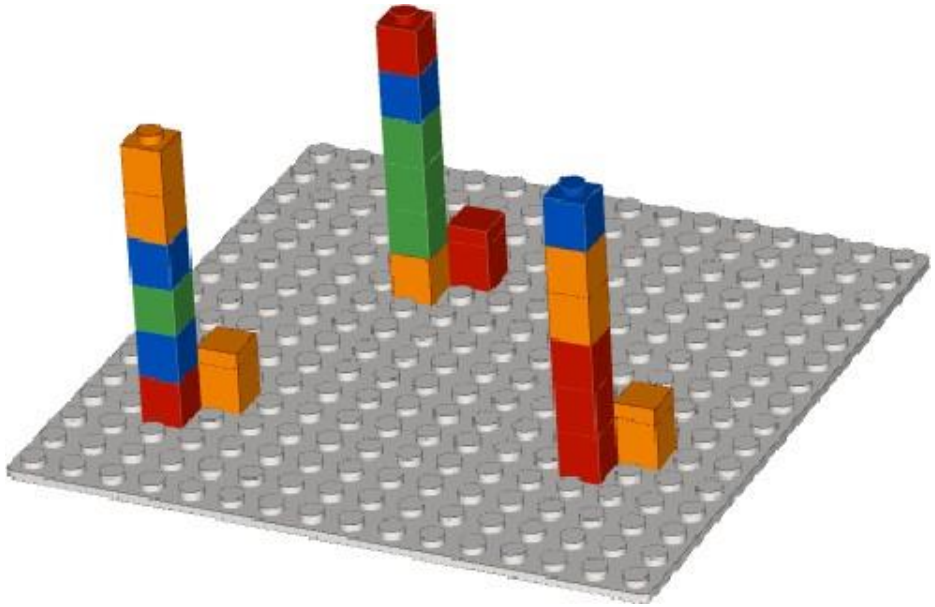
A   T

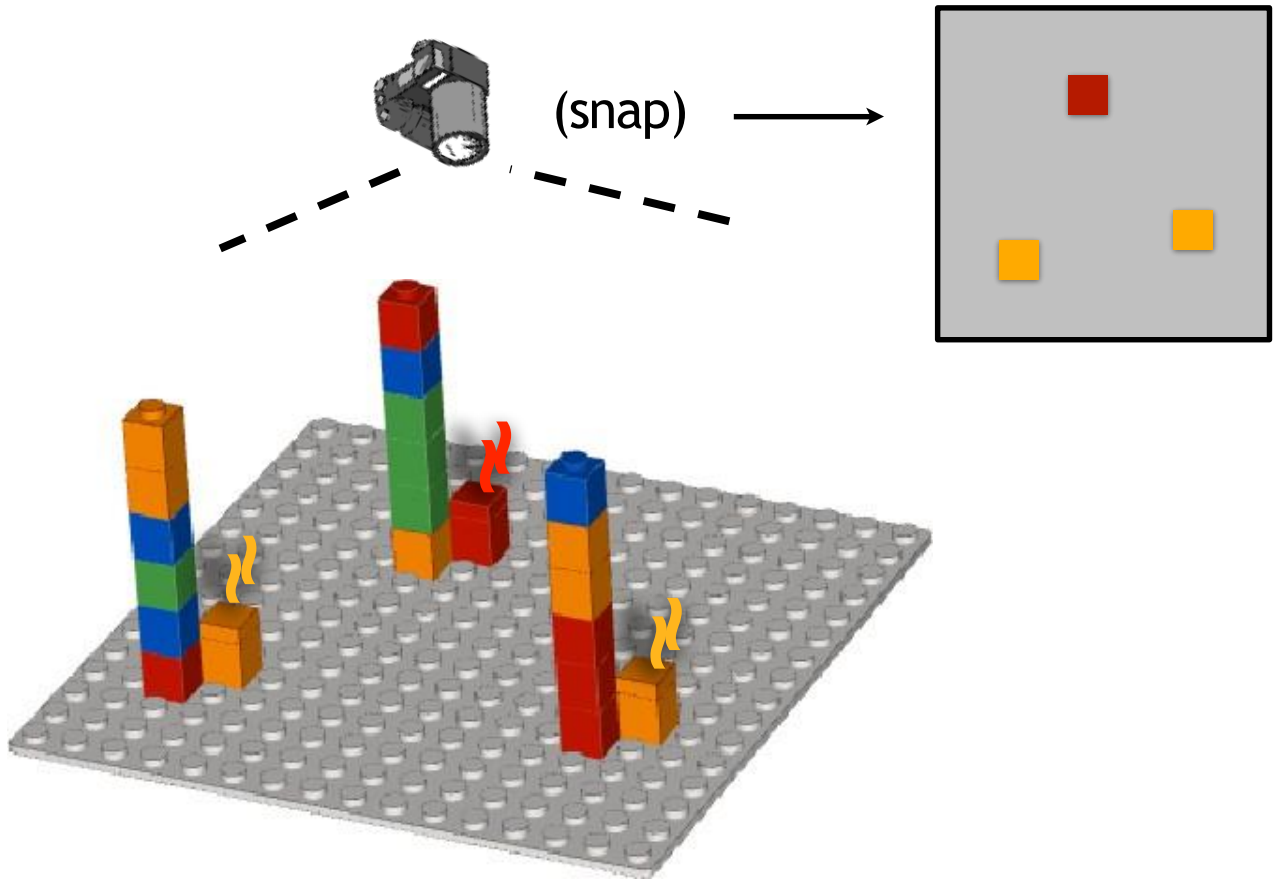
"Terminator"

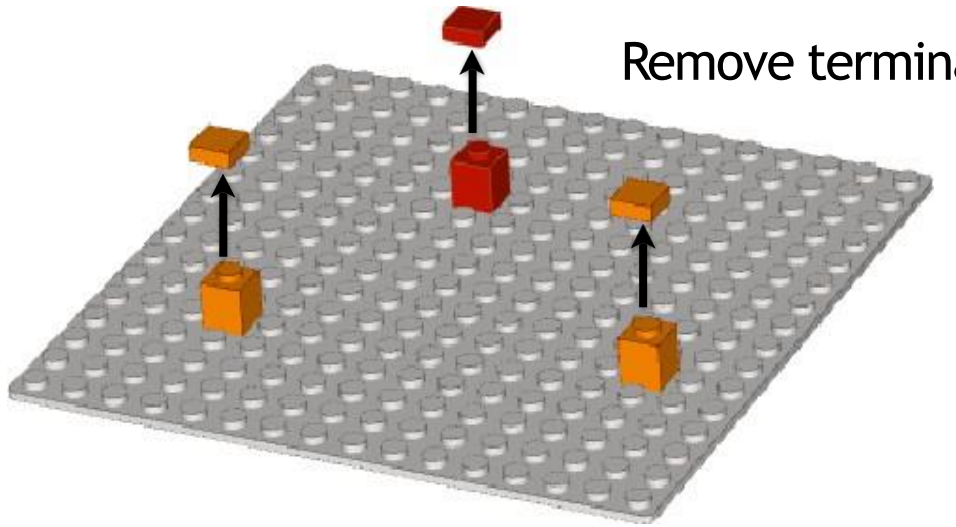
C   G

DNA polymerase






(snap)

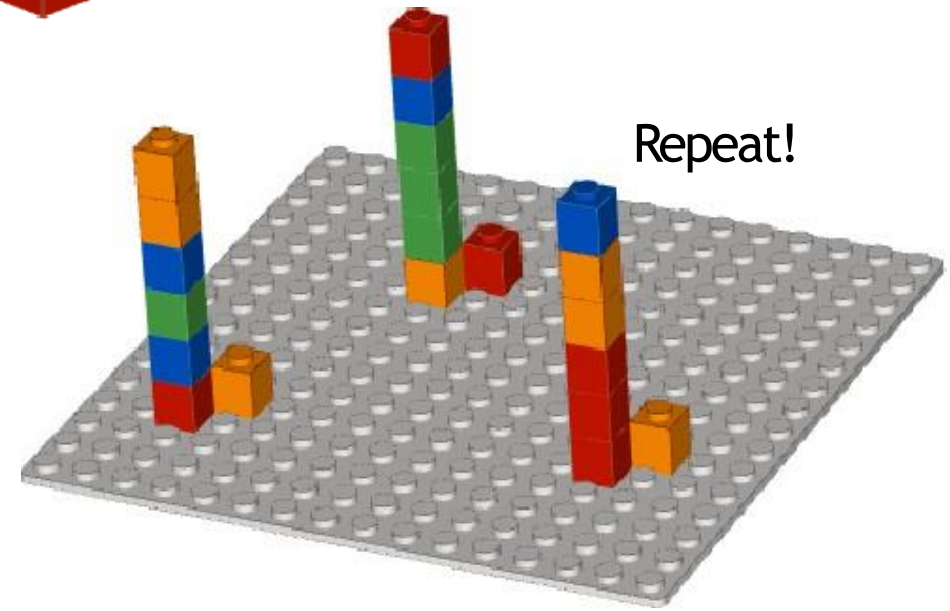


Remove terminators

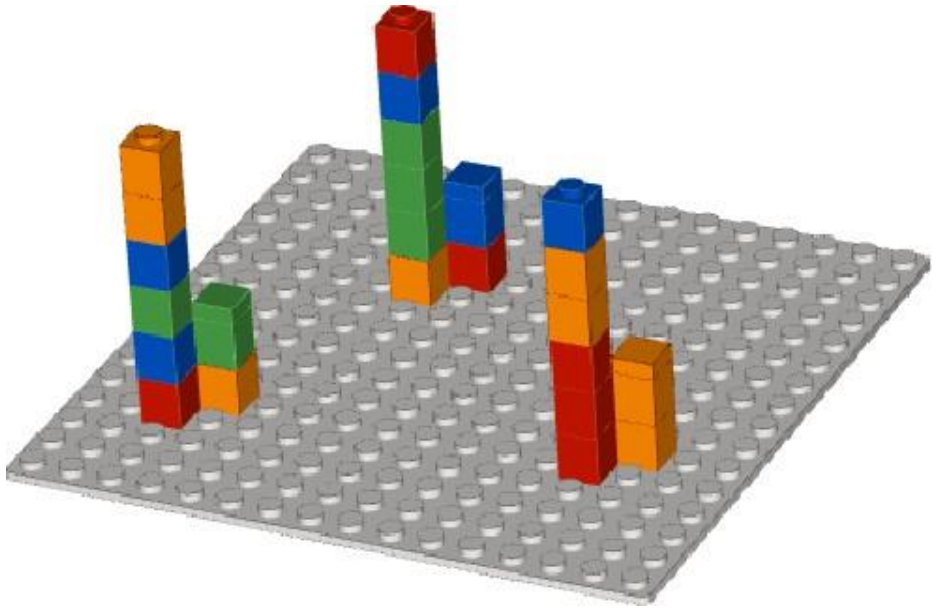
A   T

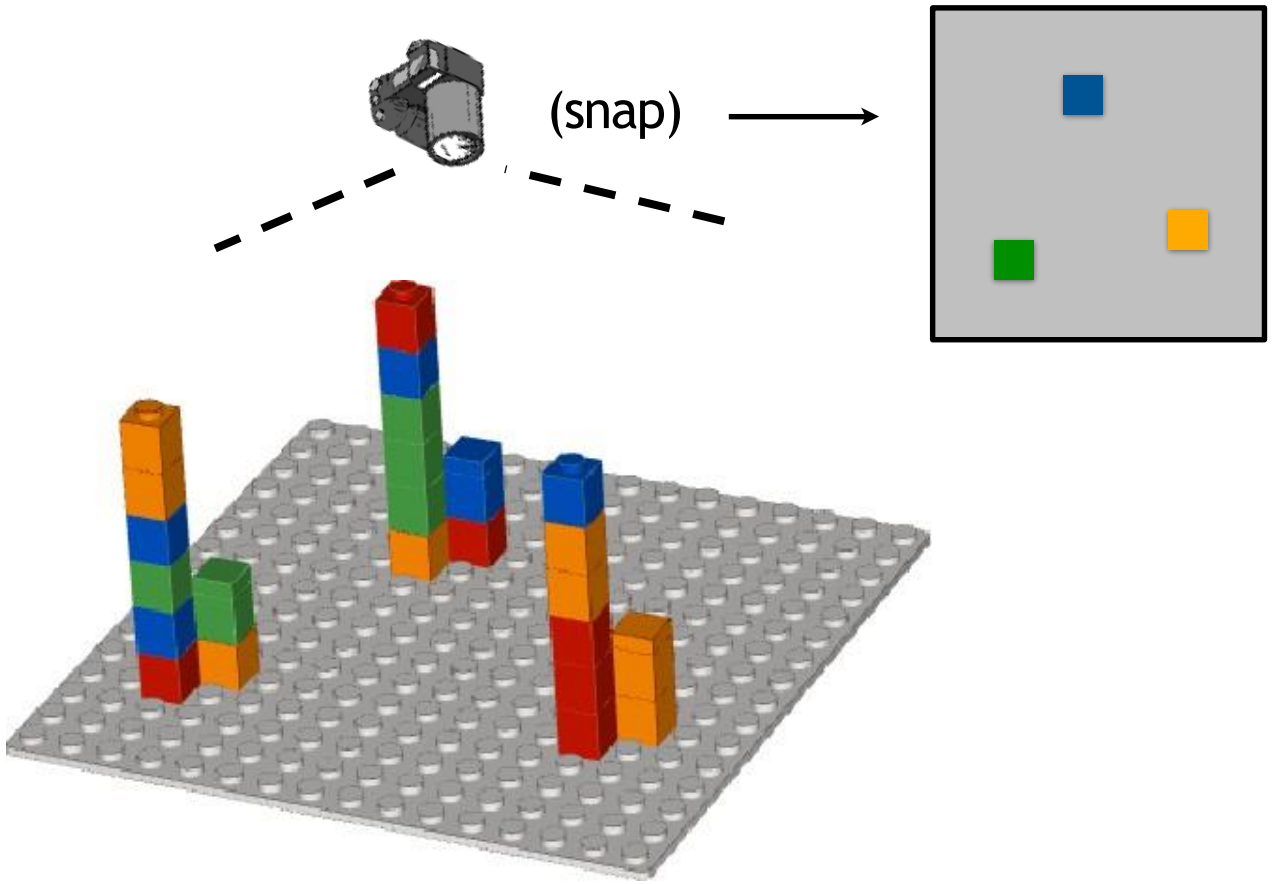
C   G

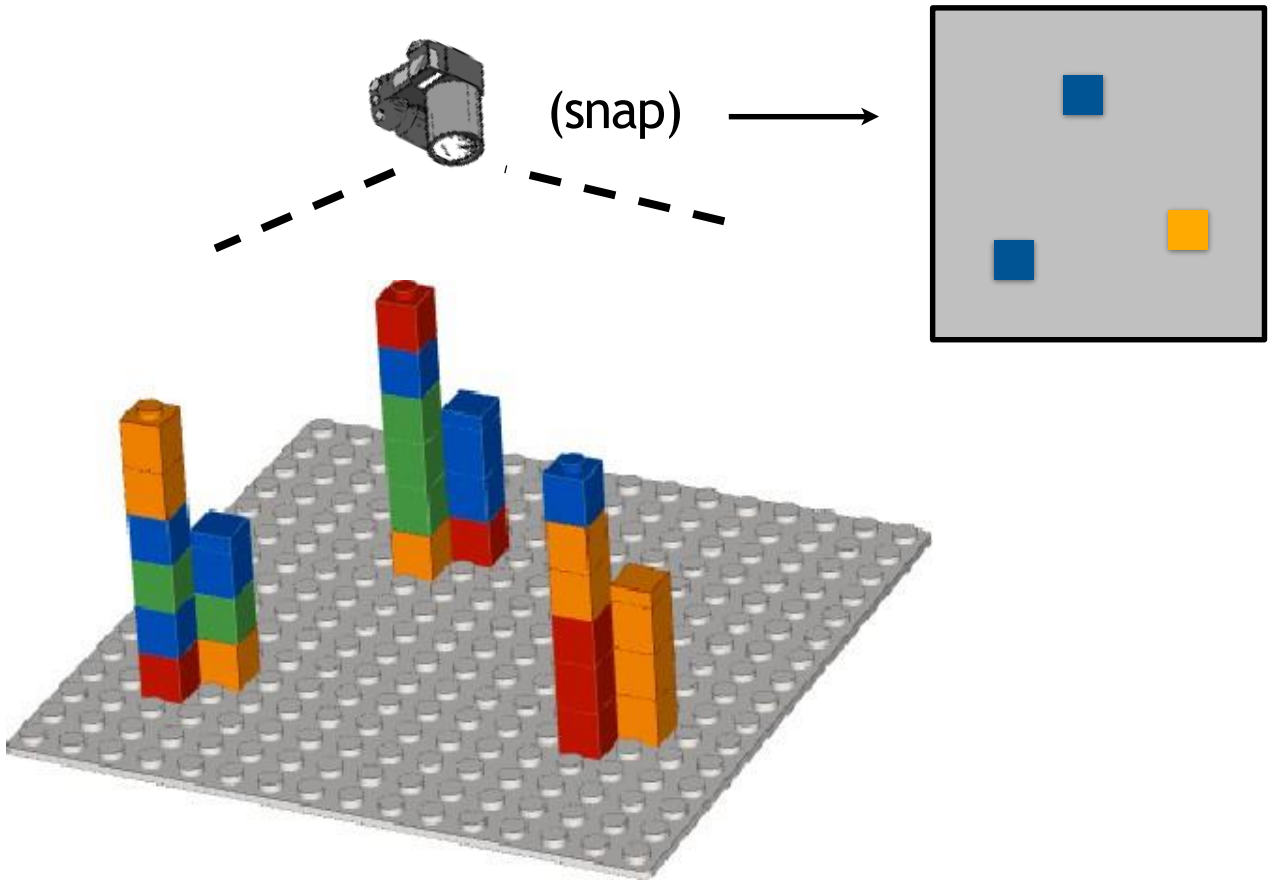
DNA polymerase

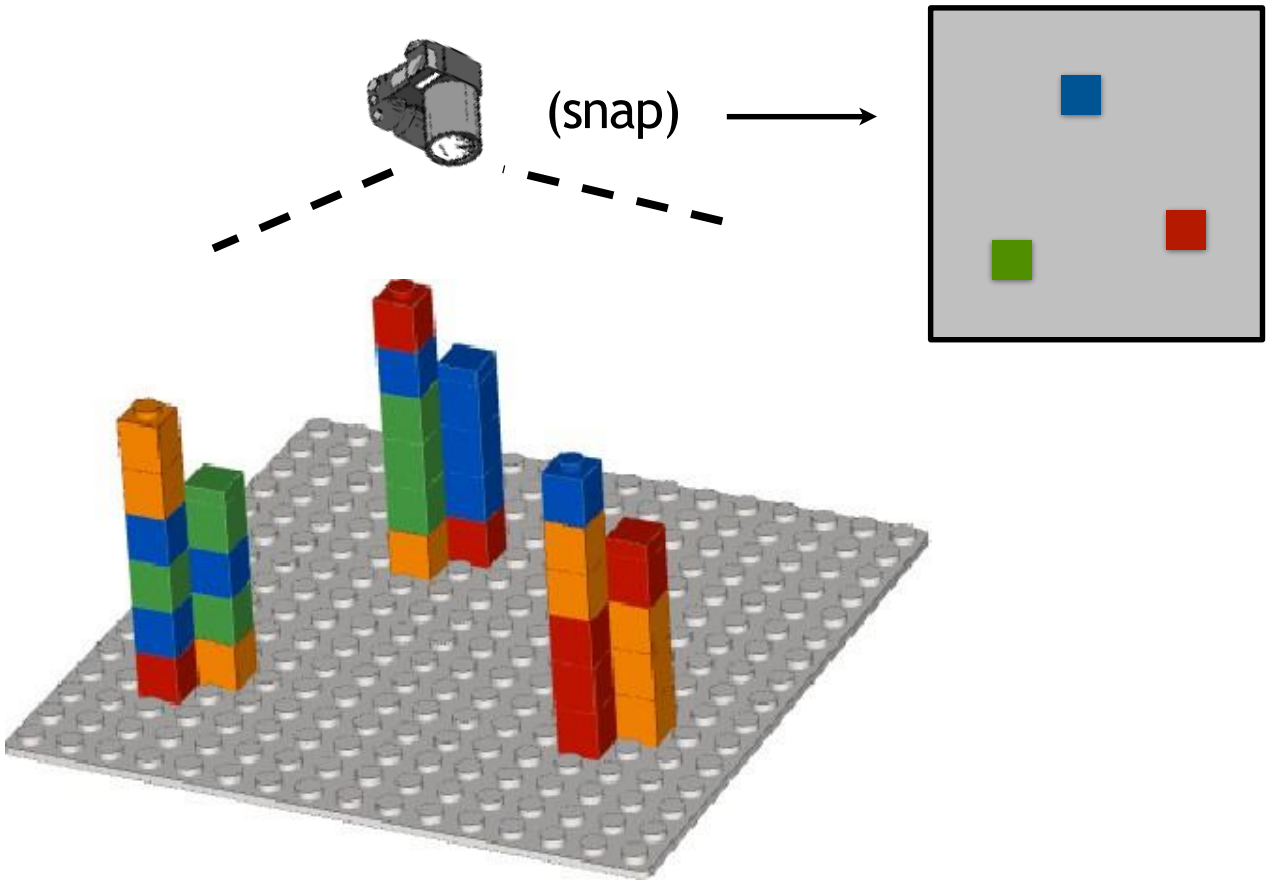


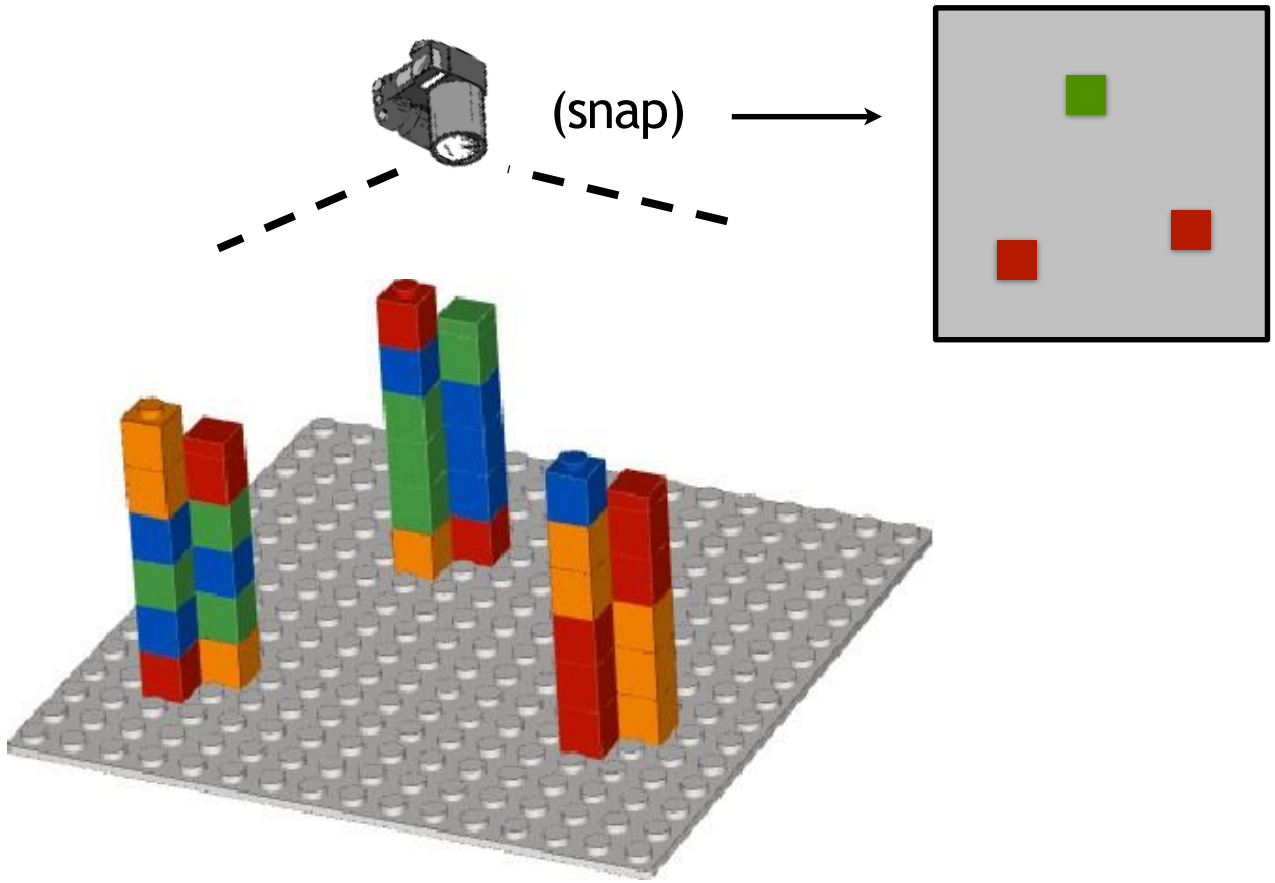
Repeat!

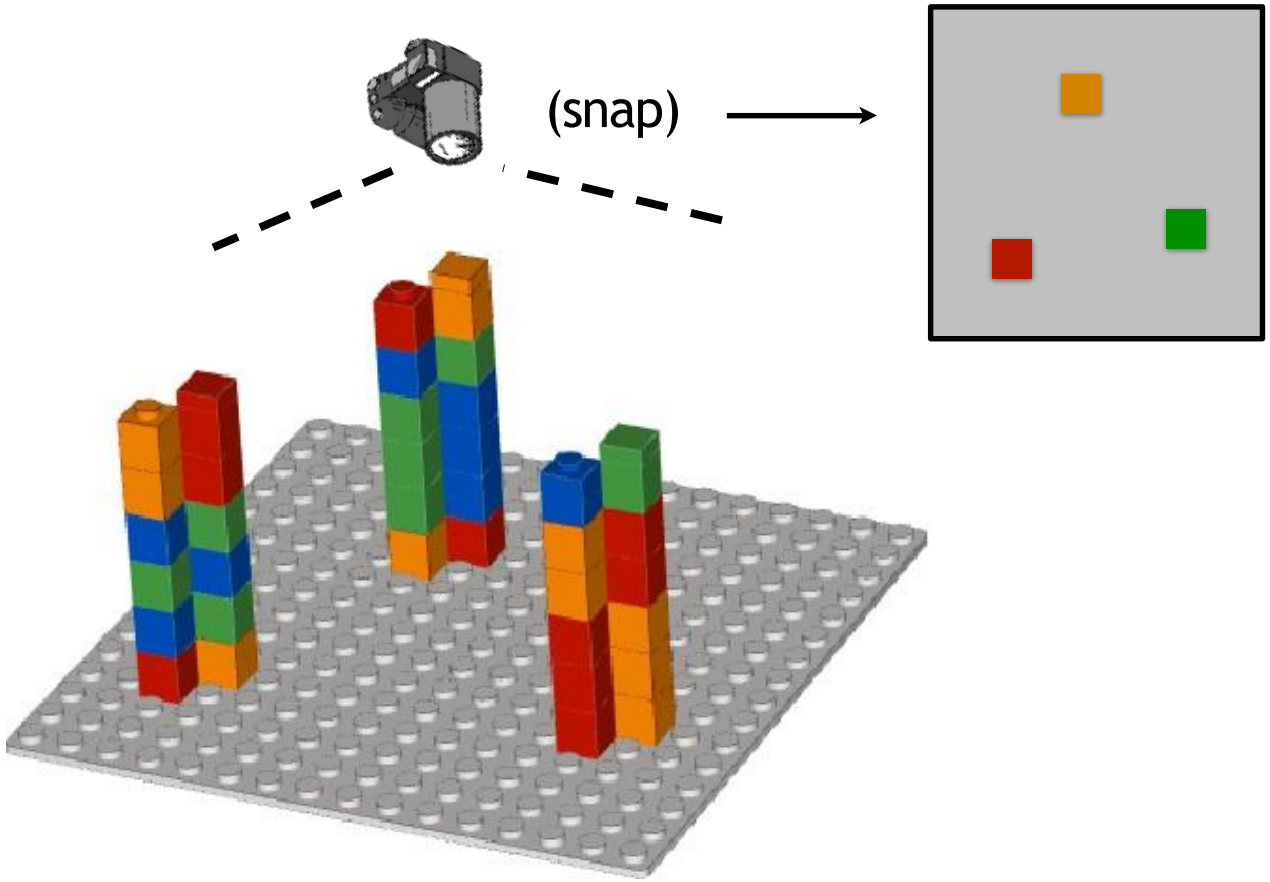



(snap)

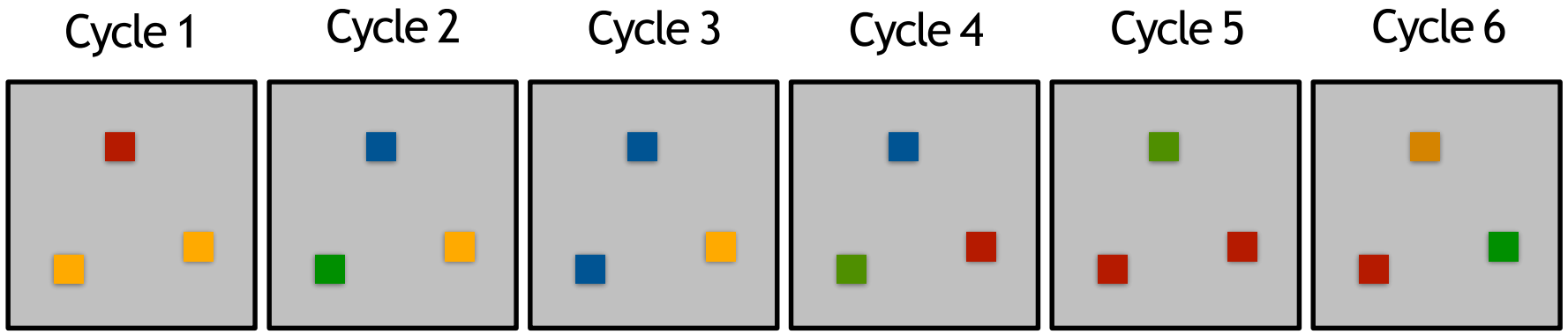

(snap)


(snap)

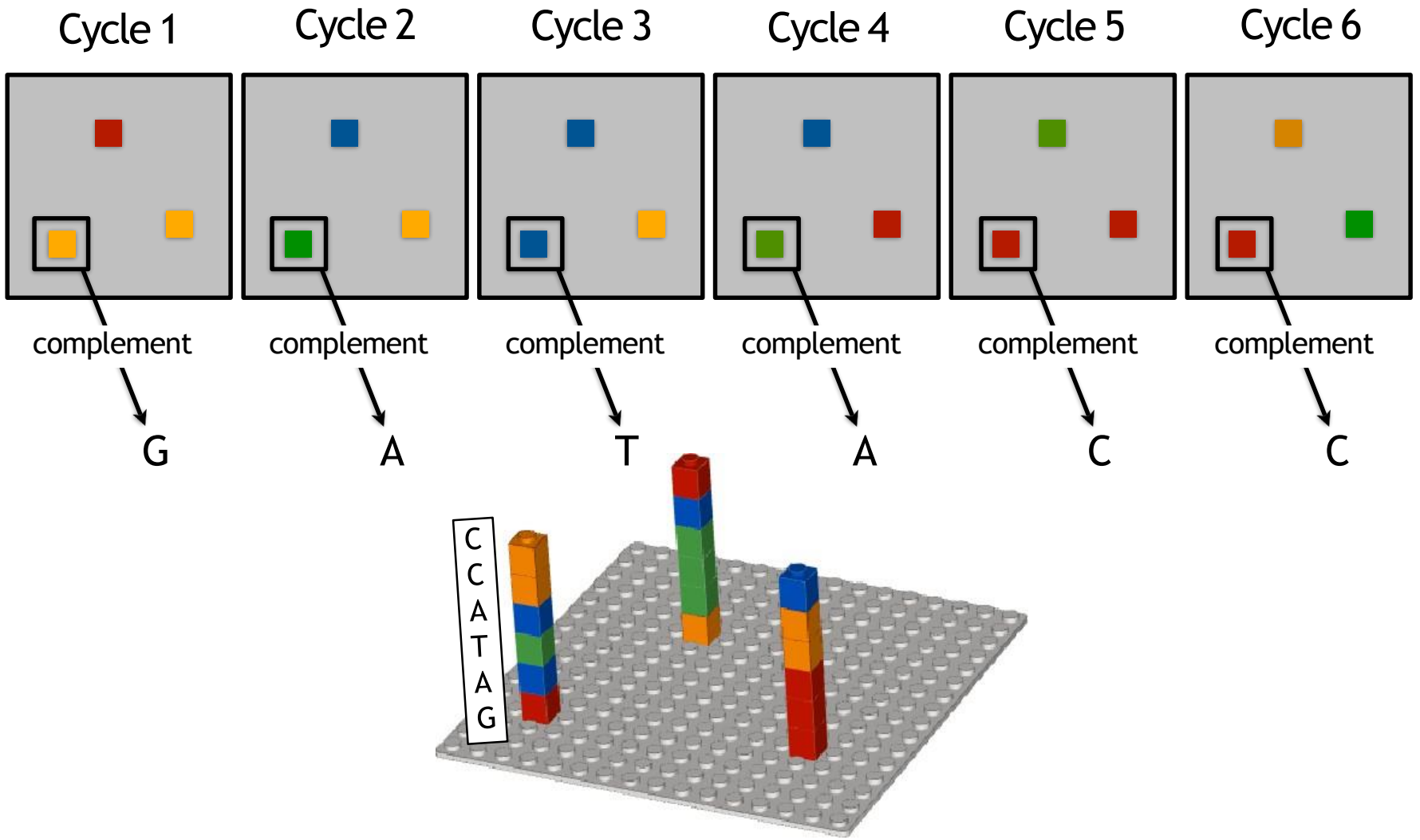

(snap)


(snap)

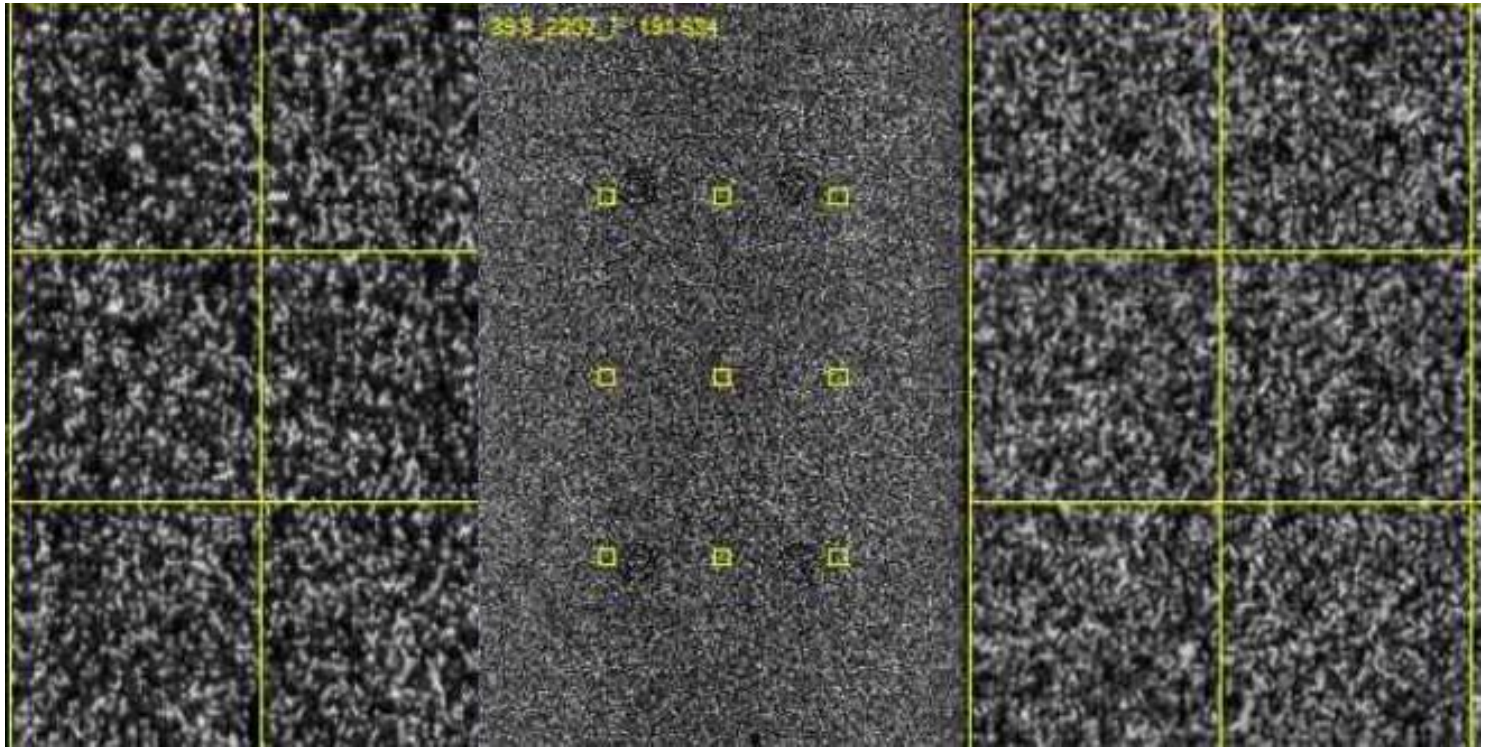
Sequencing by synthesis



Sequencing by synthesis



Sequencing by synthesis



Actual Illumina HiSeq 3000 image

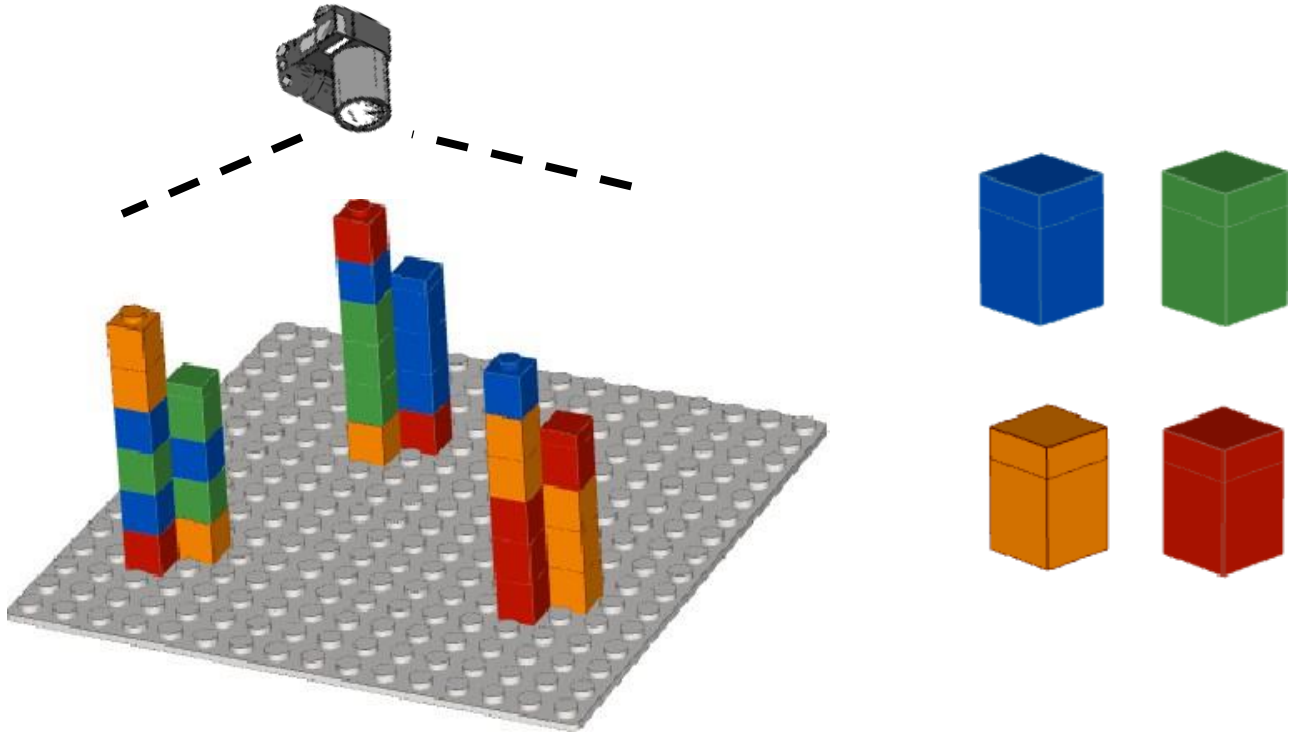
http://dnatech.genomecenter.ucdavis.edu/2015/05/07/first-hiseq-3000-data-download/

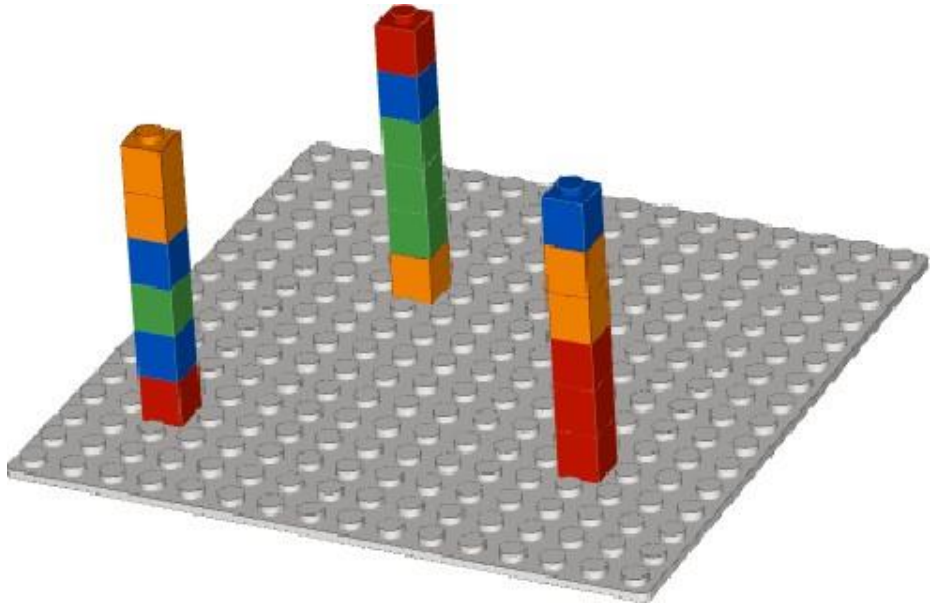
Sequencing by synthesis

Billions of templates on a slide

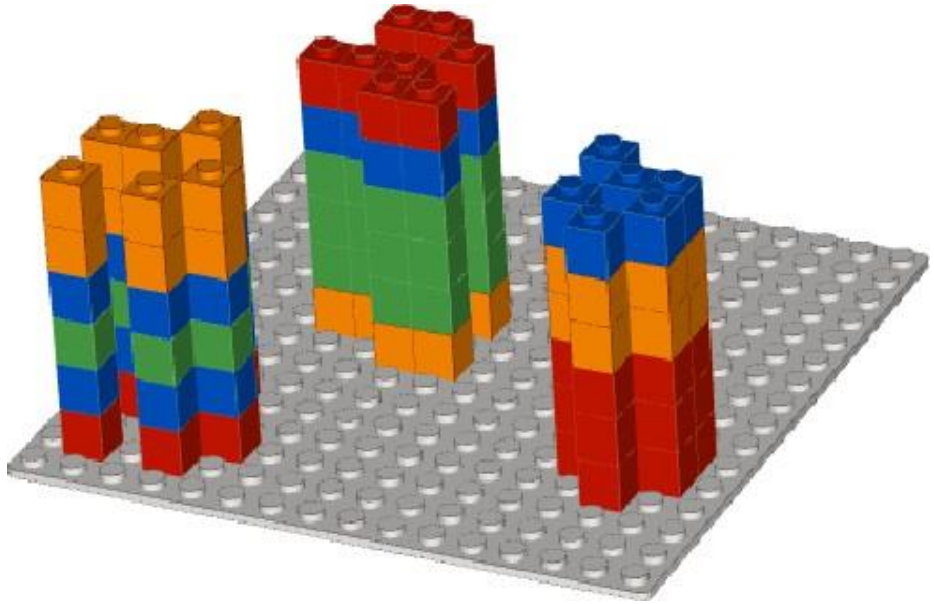
Massively parallel: photograph captures all templates simultaneously

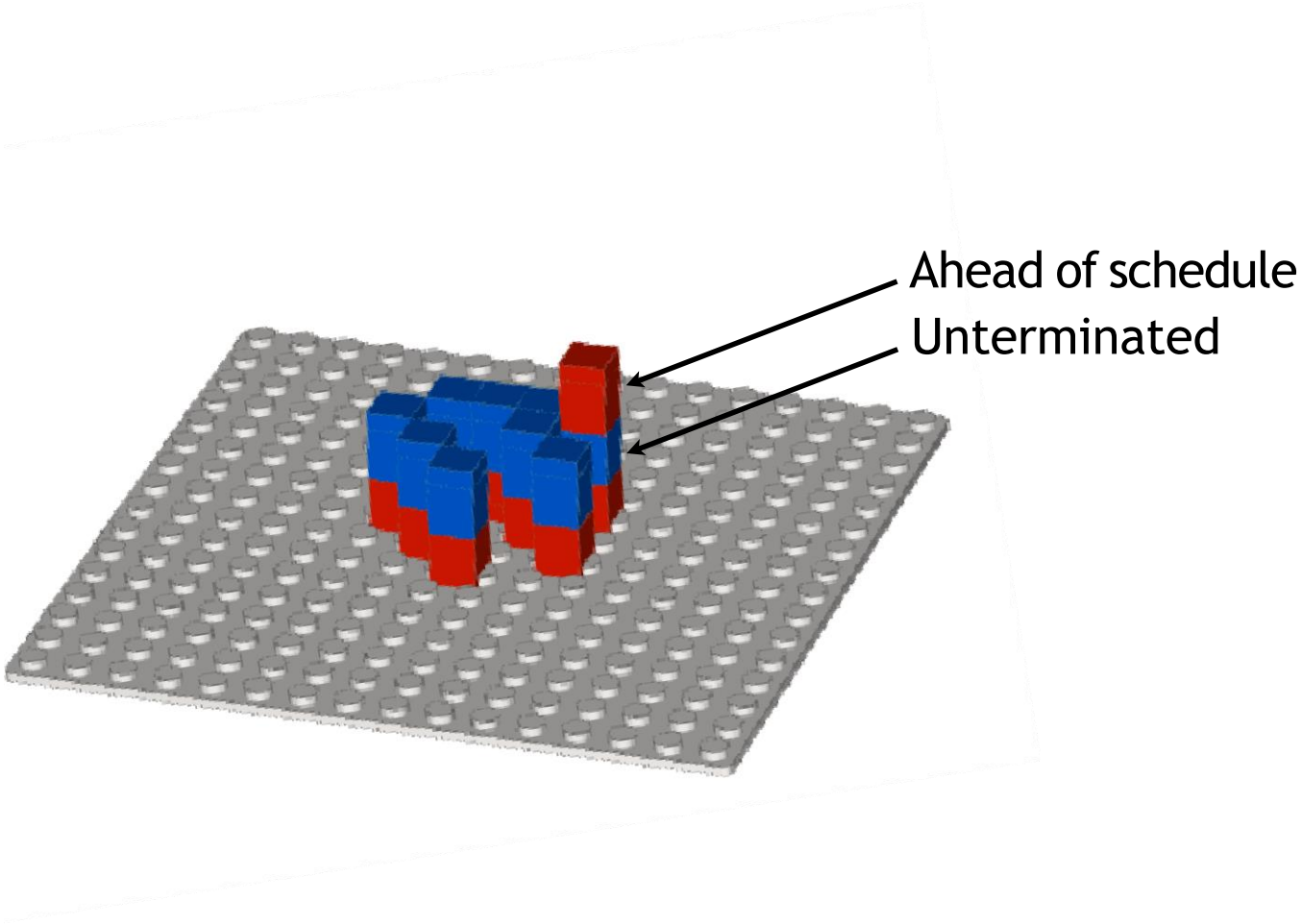
Terminators are "speed bumps," keeping reactions in sync

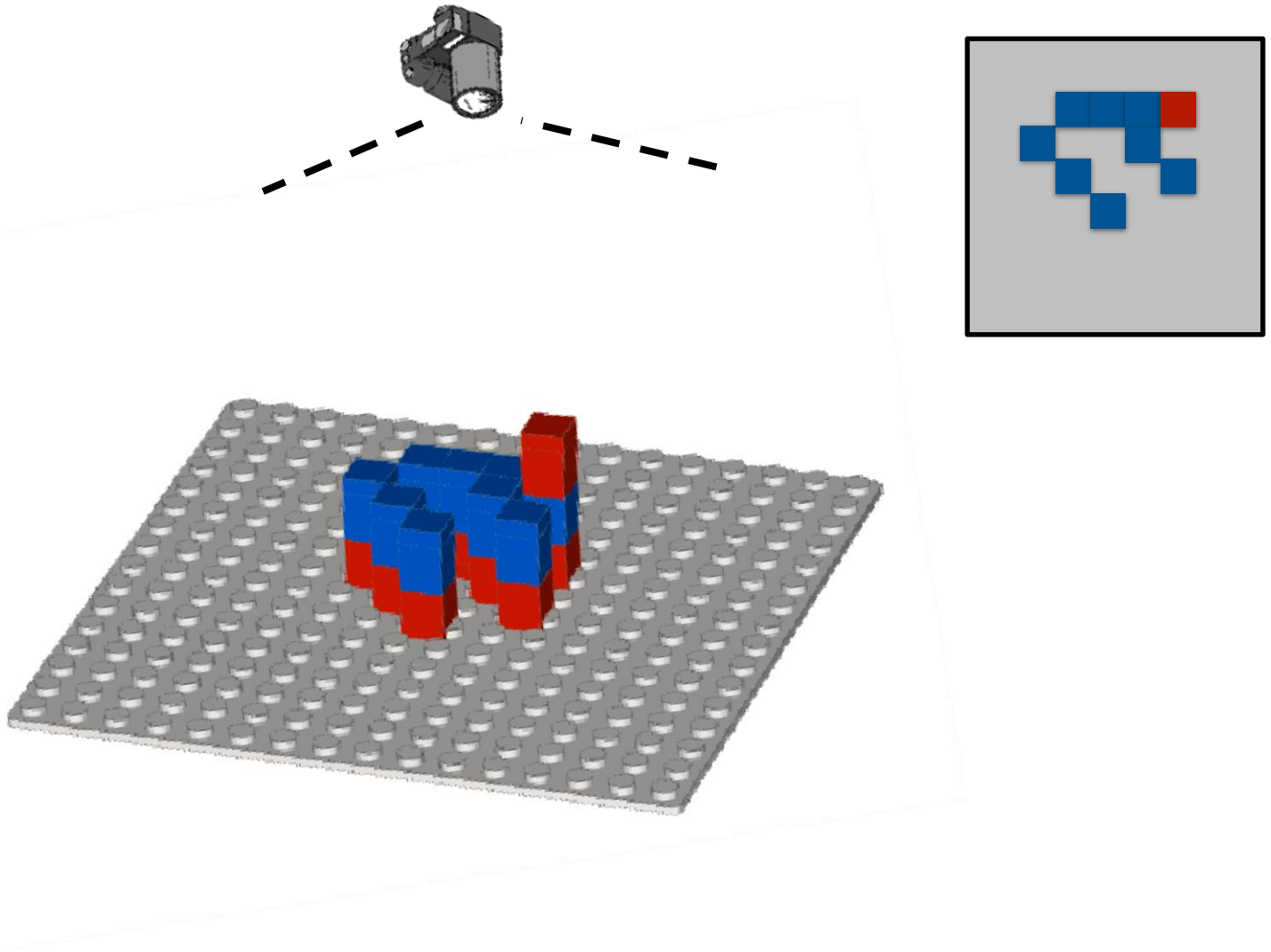


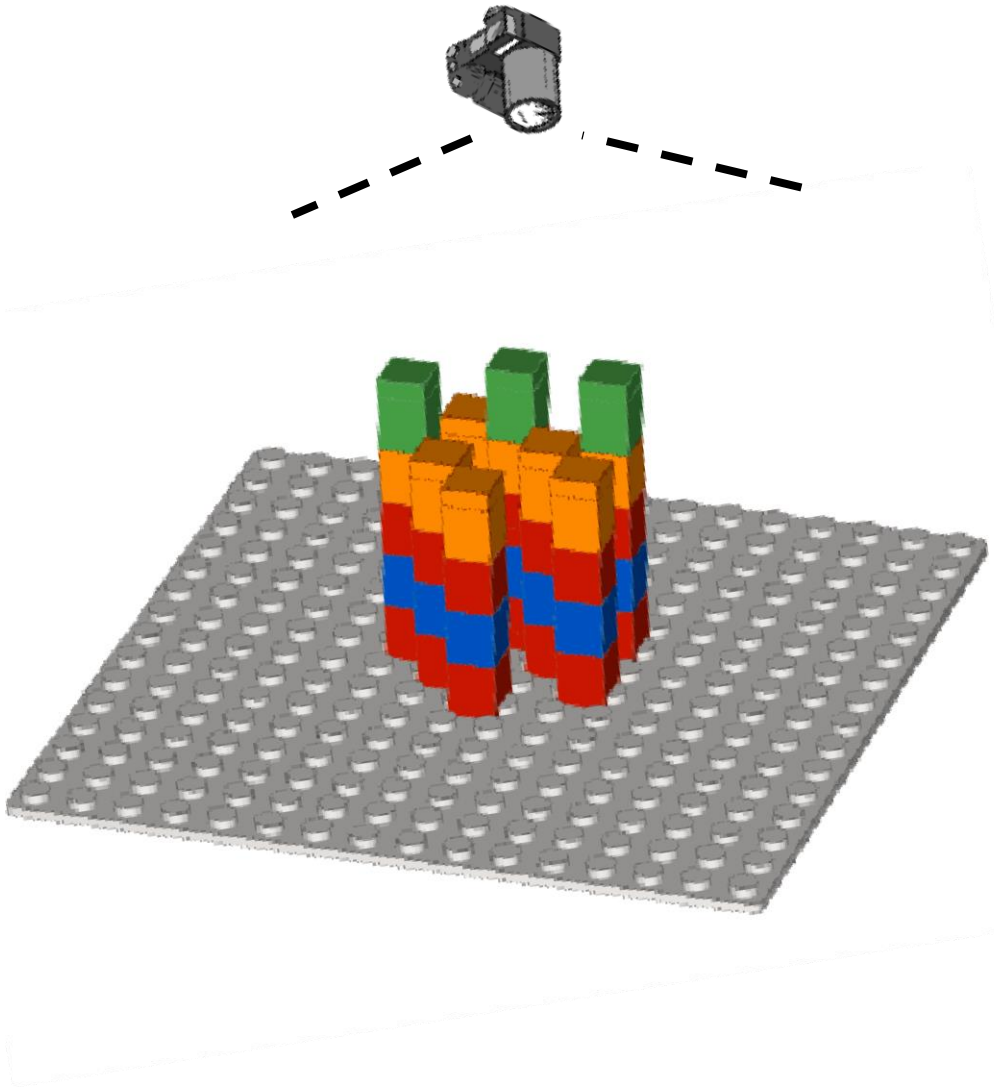


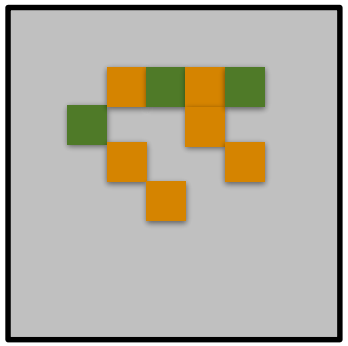
Cluster of clones




Ahead of schedule
Unterminated







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

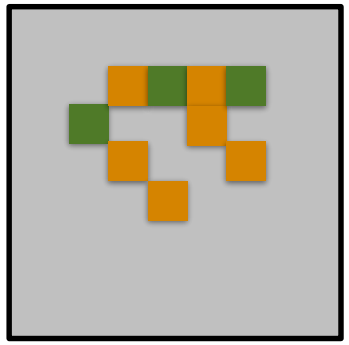
Base quality

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



Call: orange (C)

Estimate p, probability incorrect:

non-orange light / total light

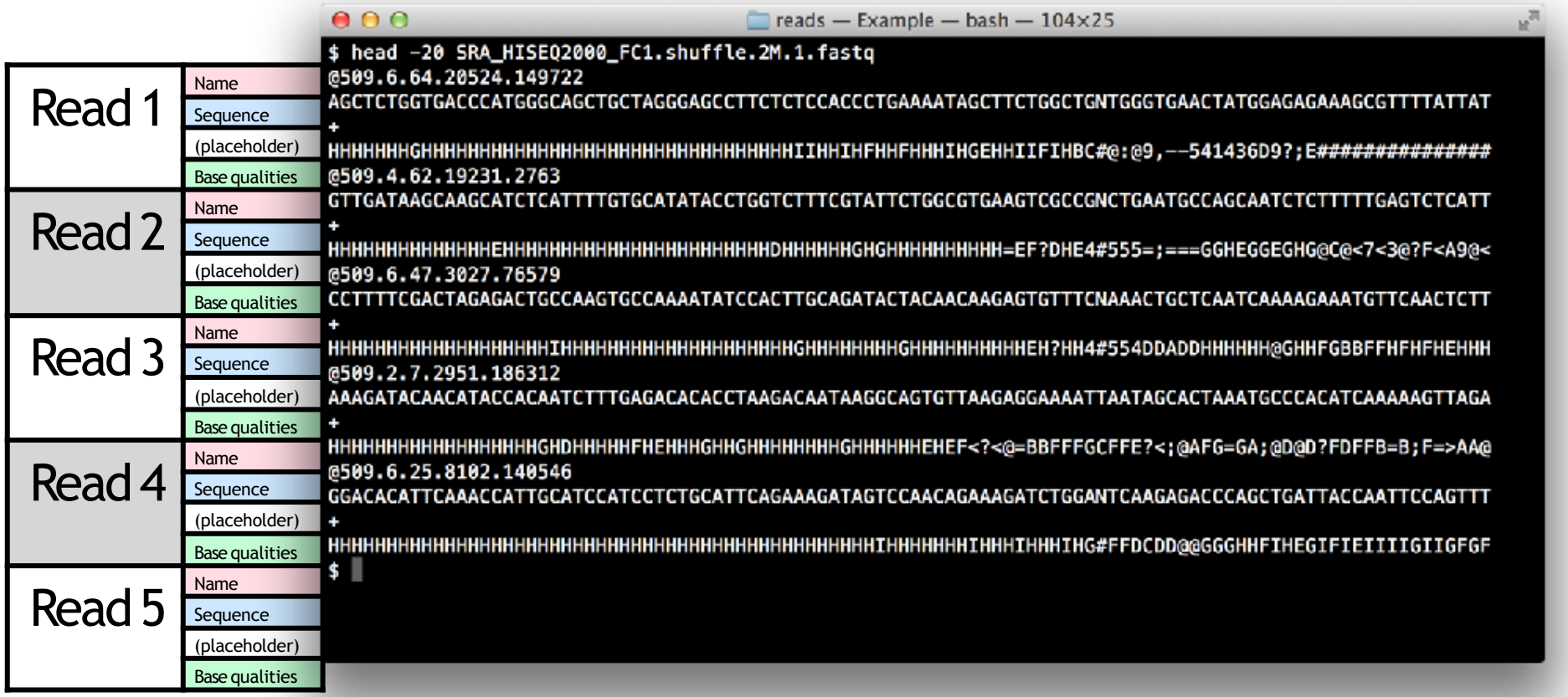
p = 3 green / 9 total = 1/3

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

A read in FASTQ format

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

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
reads — Example — bash — 104×25
$ head -20 SRA_HISEQ2000_FC1.shuffle.2M.1.fastq
@509.6.64.20524.149722
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGAGCCTTCTCTCCACCCTGAAAATAGCTTCTGGCTGNTGGGTGAACTATGGAGAGAAAGCGTTTTATTAT
+
HHHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIIHHIHFHHFHHHIHGEHHIIFIHBC#@:@9,--541436D9?;E###############
@509.4.62.19231.2763
GTTGATAAGCAAGCATCTCATTTTGTGCATATACCTGGTCTTTCGTATTCTGGCGTGAAGTCGCCGNCTGAATGCCAGCAATCTCTTTTTGAGTCTCATT
+
HHHHHHHHHHHHHHEHHHHHHHHHHHHHHHHHHHHHHHHHDHHHHHHGHGHHHHHHHHHH=EF?DHE4#555=;===GGHEGGEGHG@C@<7<3@?F<A9@<
@509.6.47.3027.76579
CCTTTTCGACTAGAGACTGCCAAGTGCCAAAATATCCACTTGCAGATACTACAACAAGAGTGTTTCNAAACTGCTCAATCAAAAGAAATGTTCAACTCTT
+
HHHHHHHHHHHHHHHHHHHIHHHHHHHHHHHHHHHHHHHGHHHHHHHHGHHHHHHHHHHEH?HH4#554DDADDHHHHHH@GHHFGBBFFHFHFHEHHH
@509.2.7.2951.186312
AAAGATACAACATACCACAATCTTTGAGACACACCTAAGACAATAAGGCAGTGTTAAGAGGAAAATTAATAGCACTAAATGCCCACATCAAAAAGTTAGA
+
HHHHHHHHHHHHHHHHHHHGHDHHHHHFHEHHHGHHGHHHHHHHHHGHHHHHHEHEF<?<@=BBFFFGCFFE?<;@AFG=GA;@D@D?FDFFB=B;F=>AA@
@509.6.25.8102.140546
GGACACATTCAAACCATTGCATCCATCCTCTGCATTCAGAAAGATAGTCCAACAGAAAGATCTGGANTCAAGAGACCCAGCTGATTACCAATTCCAGTTT
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHIHHHIHHHIHG#FFDCDD@@GGGHHFIHEGIFIEIIIIGIIGFGF
$

Base qualities

Bases and qualities line up:

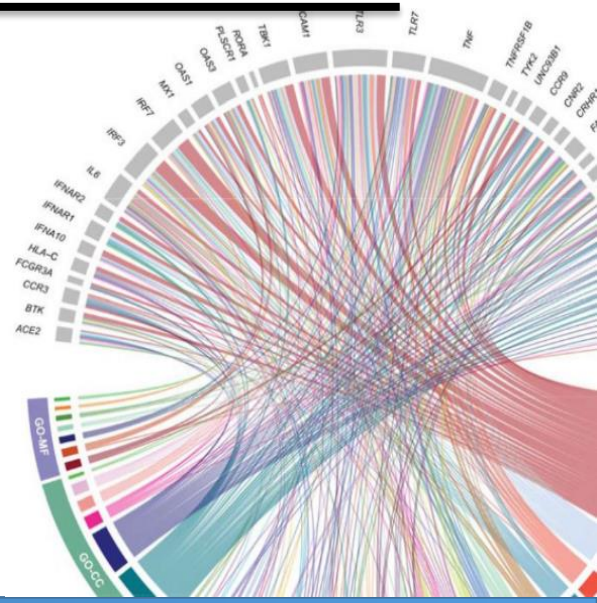
```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
||||||||||||||||||||||||||||||||
HHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

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

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

5 Oxford Nanopore Sequencing



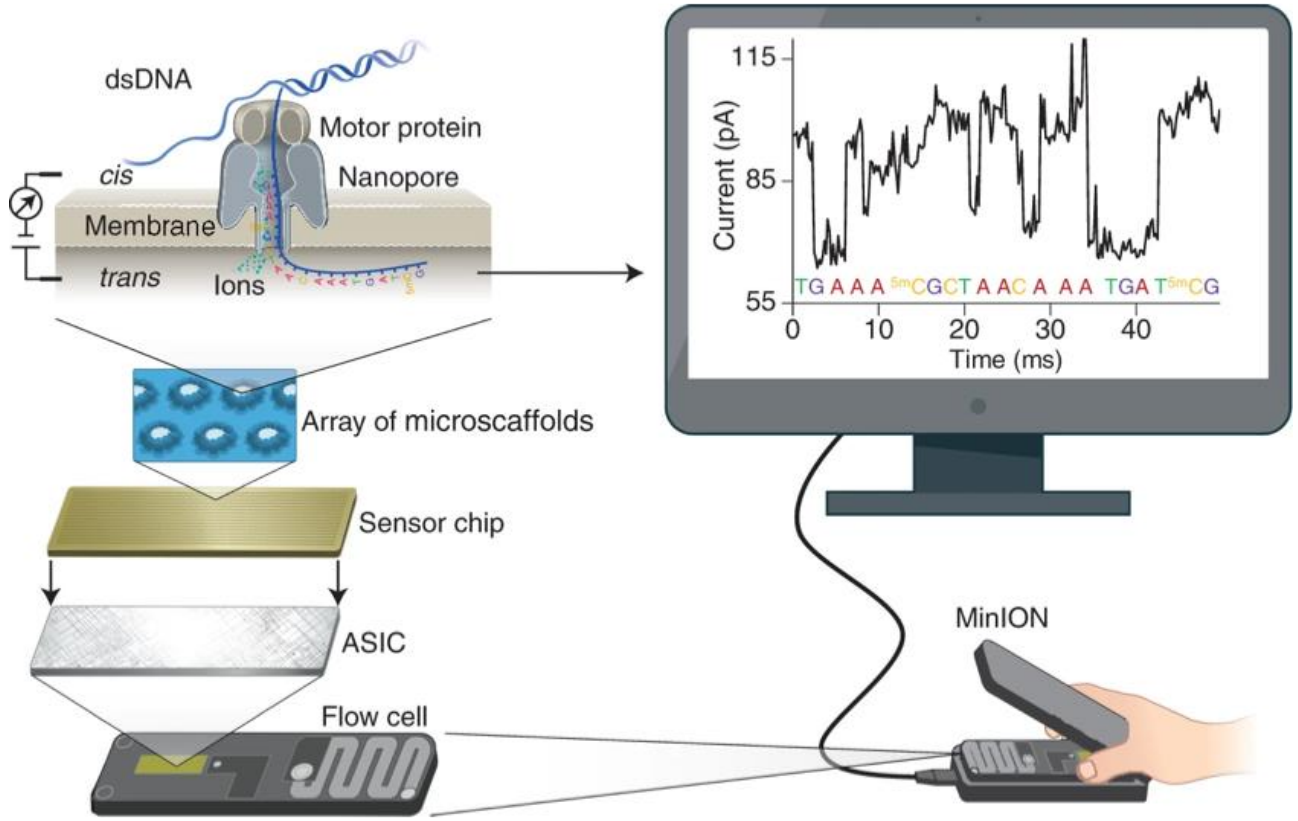
Oxford Nanopore Sequencing

- Nanopore Sequencing
  - Long reads (>1 million bases reads)
  - Portable sequencing
  - Cost-effective
  - Real-time analysis
- Raw Signals. Ionic current measurements generated as DNA passes through the nanopore.
- Real-Time Analysis. Translating to bases or directly analyzing raw signals.
- Real-Time Decisions. Stopping sequencing early based on real-time analysis.

https://www.youtube.com/watch?v=E9-Rm5AoZGw

Oxford Nanopore Sequencing

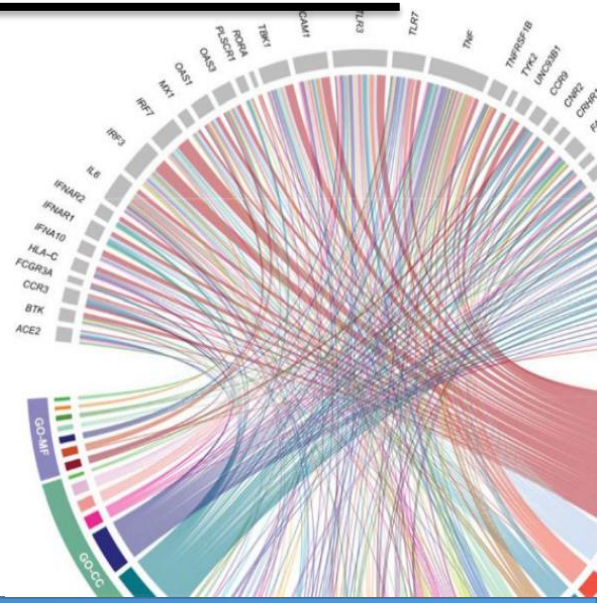
- Nanopore Sequencing
  - Long reads (>1 million bases reads)
  - Portable sequencing
  - Cost-effective
  - Real-time analysis
- Raw Signals. Ionic current measurements generated as DNA passes through the nanopore.
- Real-Time Analysis. Translating to bases or directly analyzing raw signals.
- Real-Time Decisions. Stopping sequencing early based on real-time analysis.



Wang+, "Nanopore sequencing technology, bioinformatics and applications," *Nature Biotechnology*, 2021.

https://www.youtube.com/watch?v=E9-Rm5AoZGw

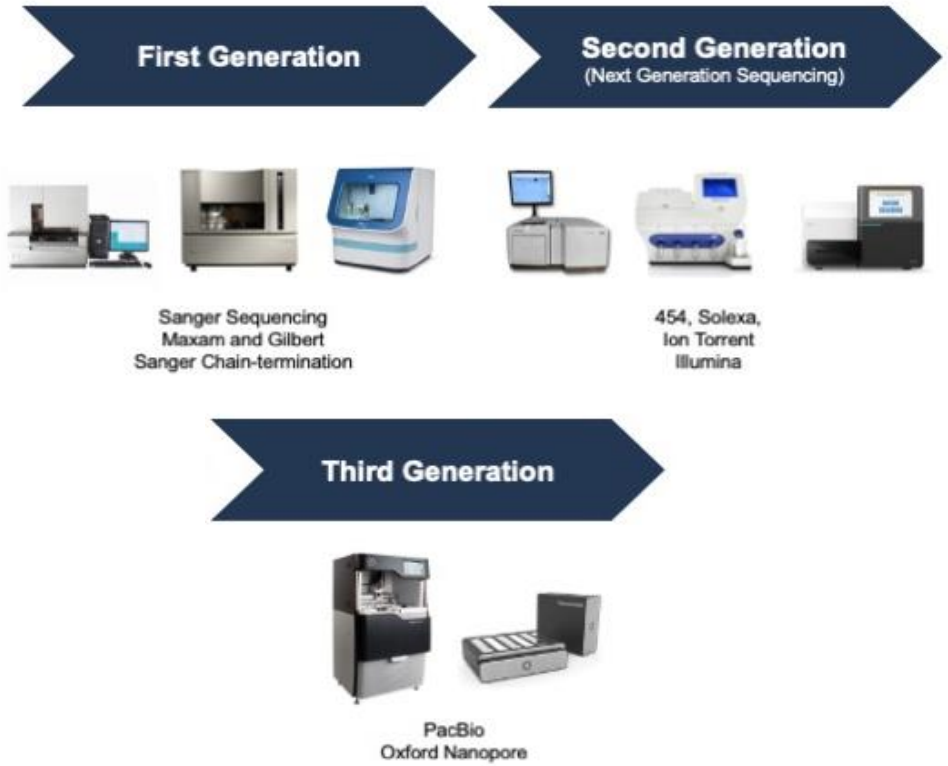
6 Genome Data Analysis

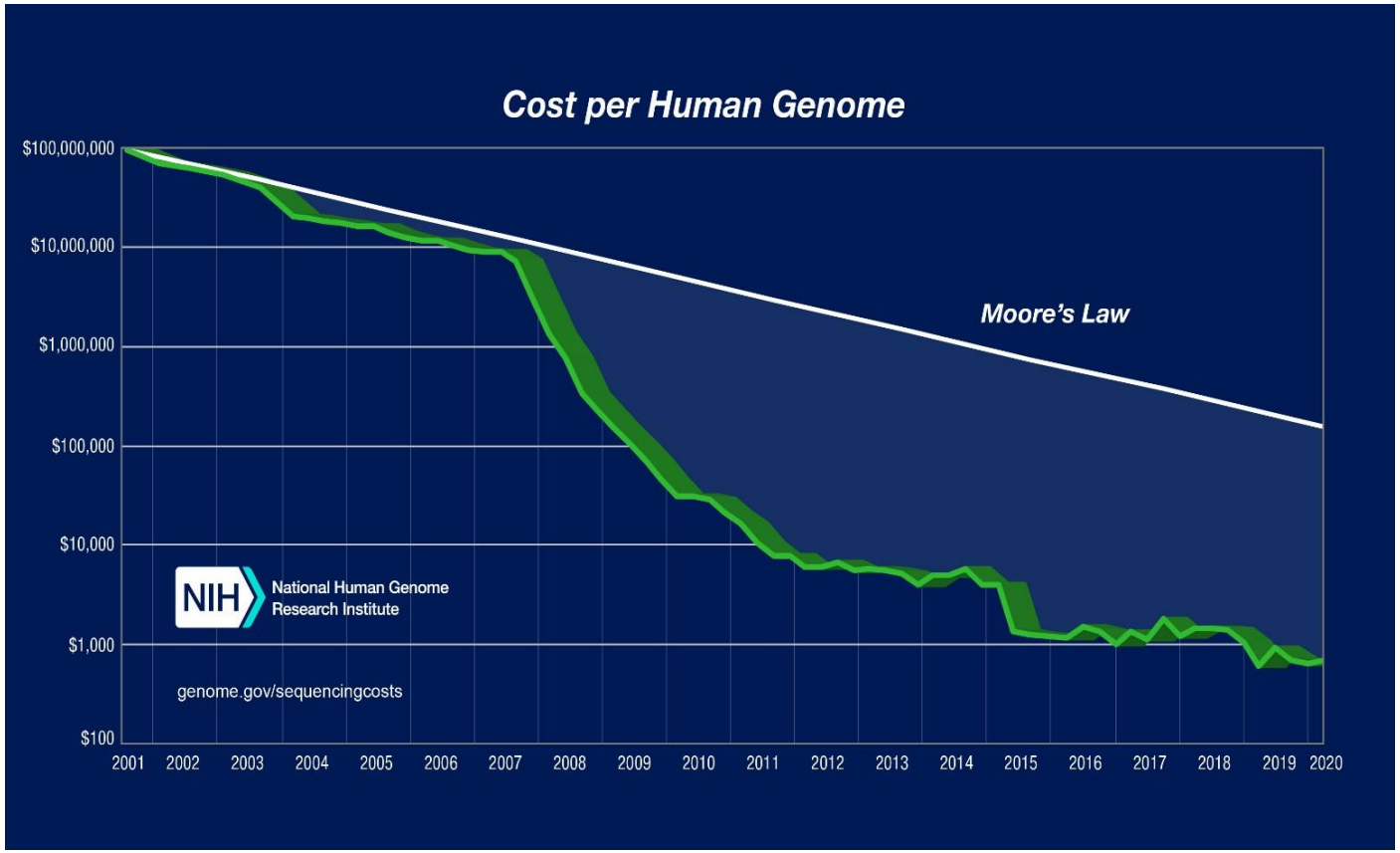


Sequencing Has Become Clinically Affordable

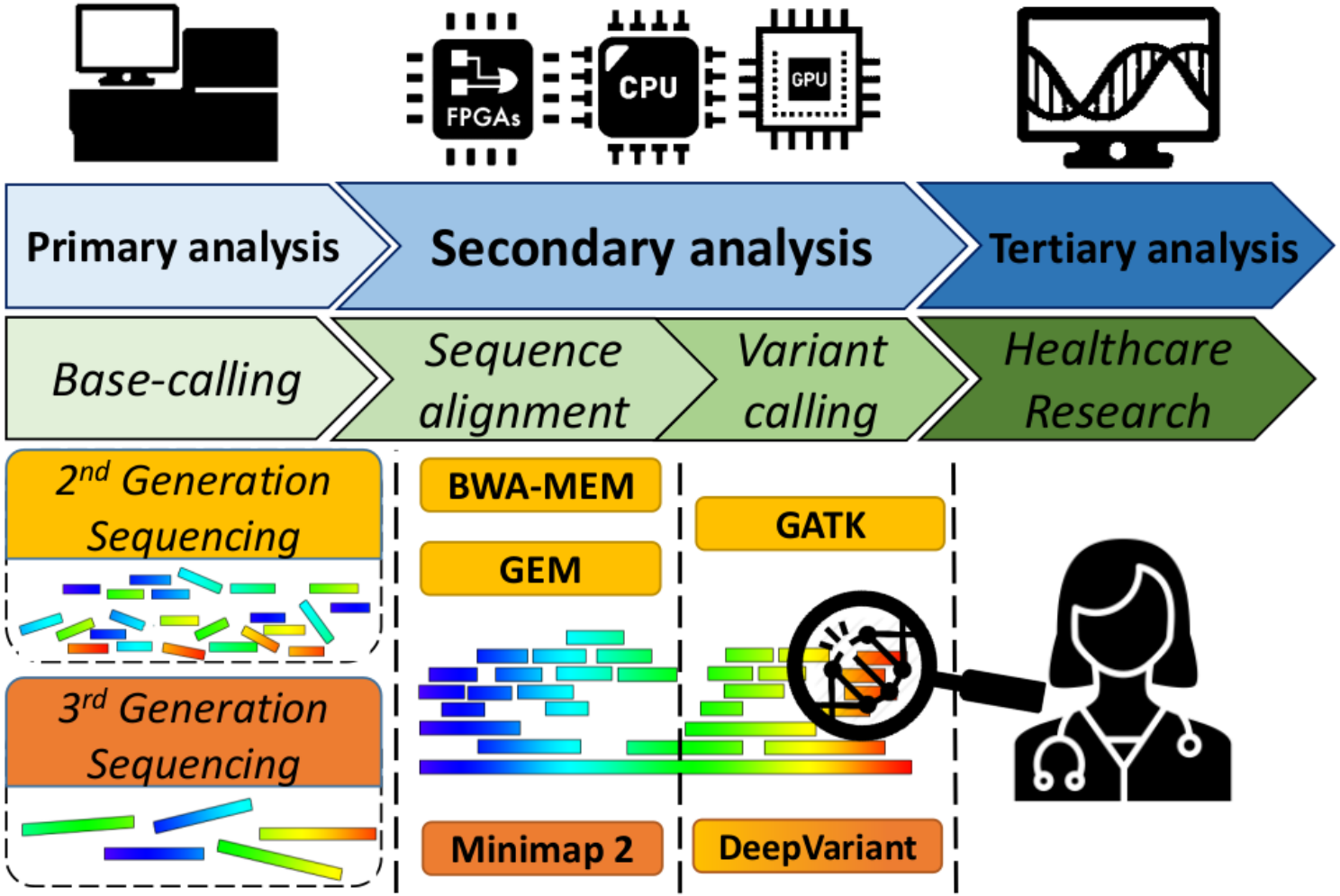
Nowadays, we can sequence a complete individual (i.e., **whole genome**) in less 48h for **less than $1000.**

Whole Exome for **less than $200.**

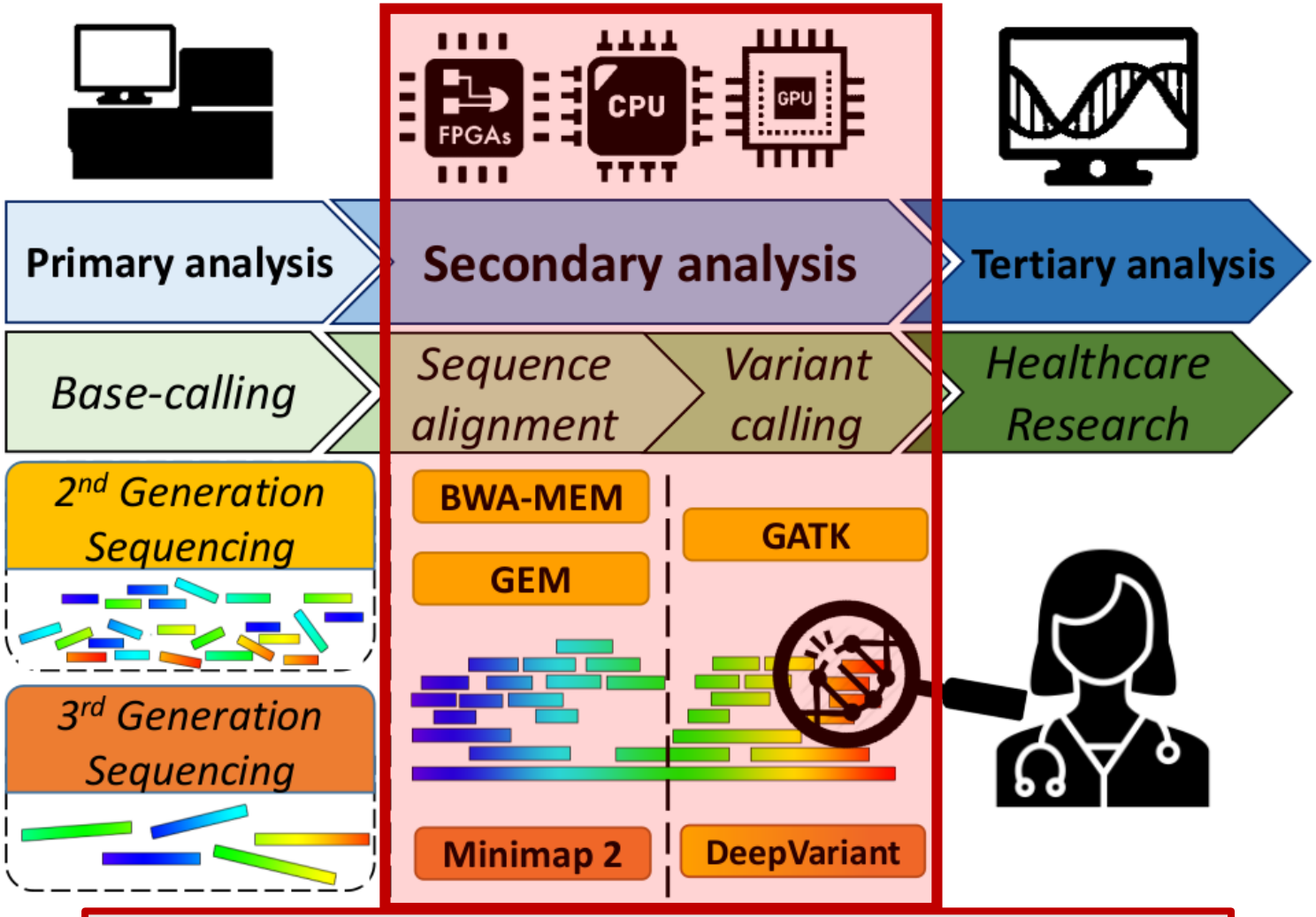




Sequencing Data Analysis

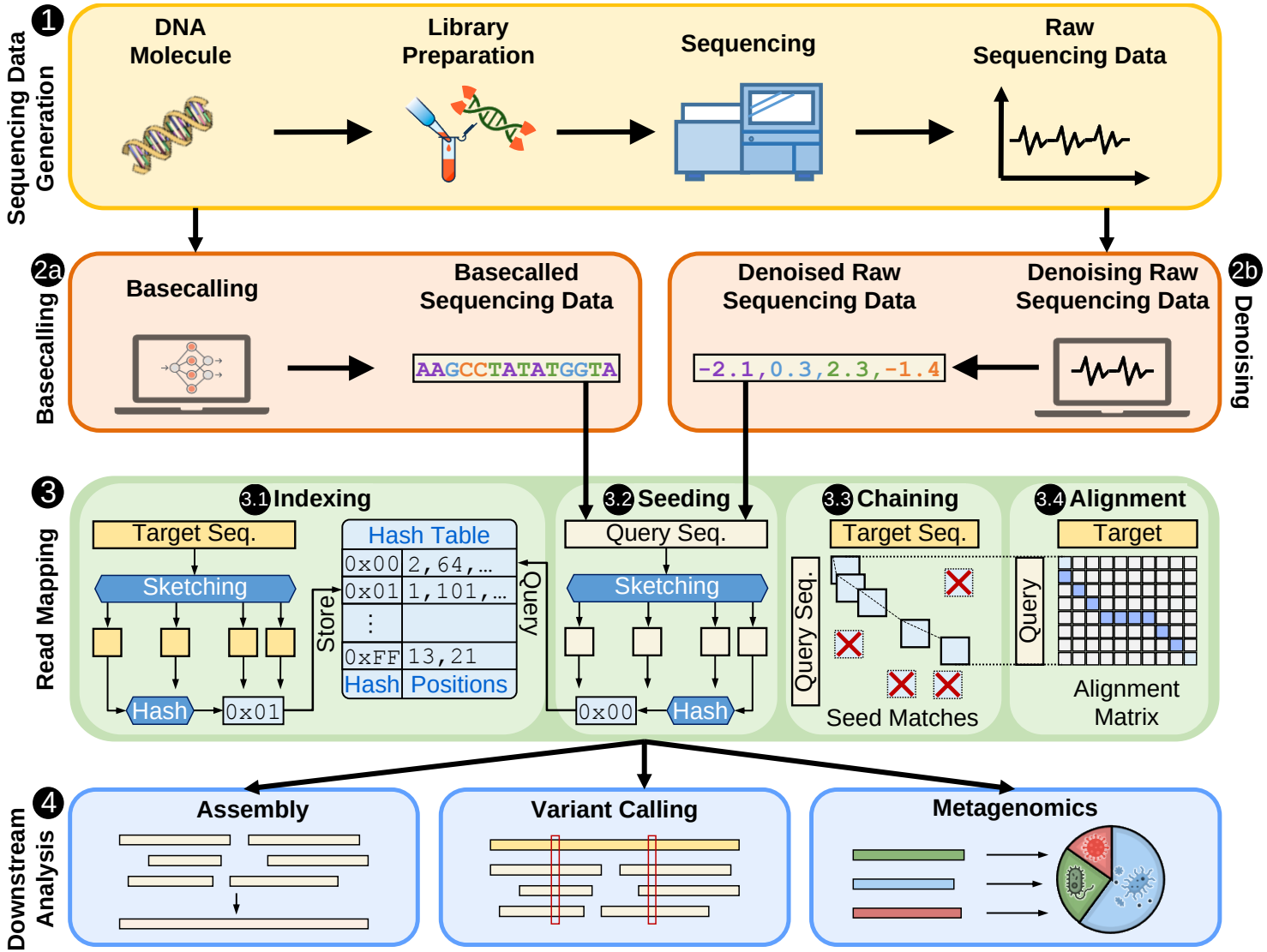


Sequencing Data Analysis Bottleneck

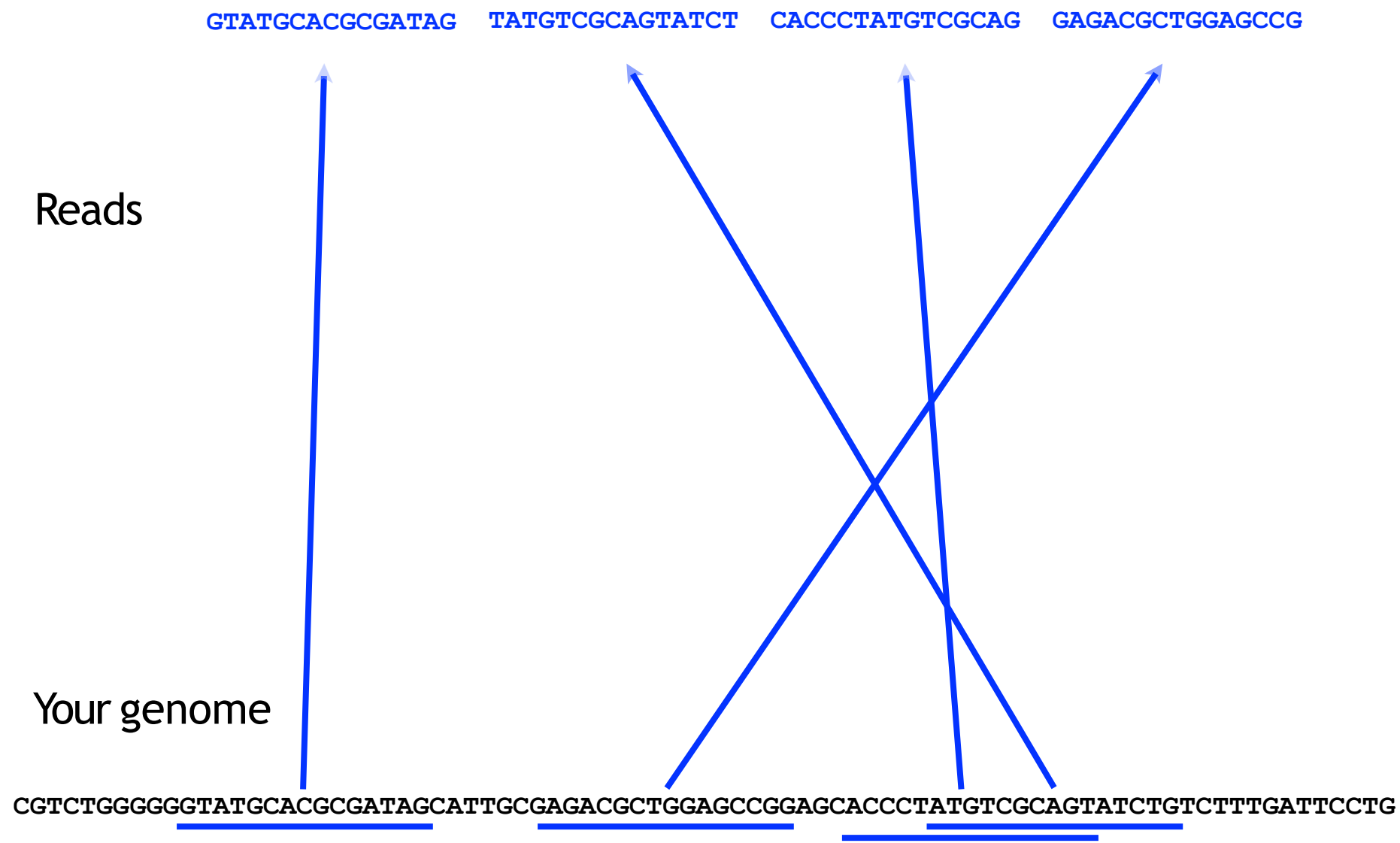


As sequencing becomes inexpensive, computational analyses become the bottleneck.

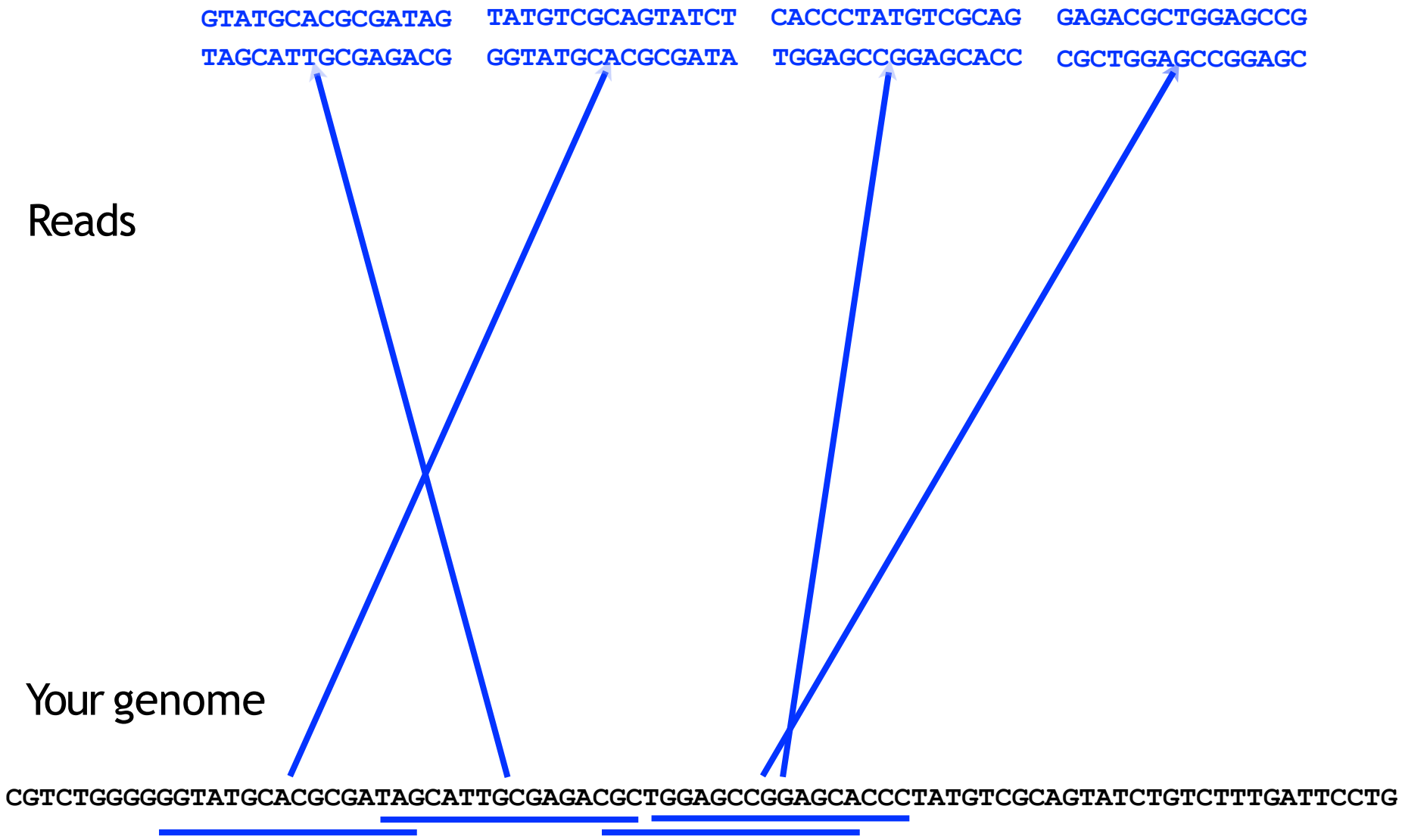
Sequencing Data Analysis (Pipelines & Tools)



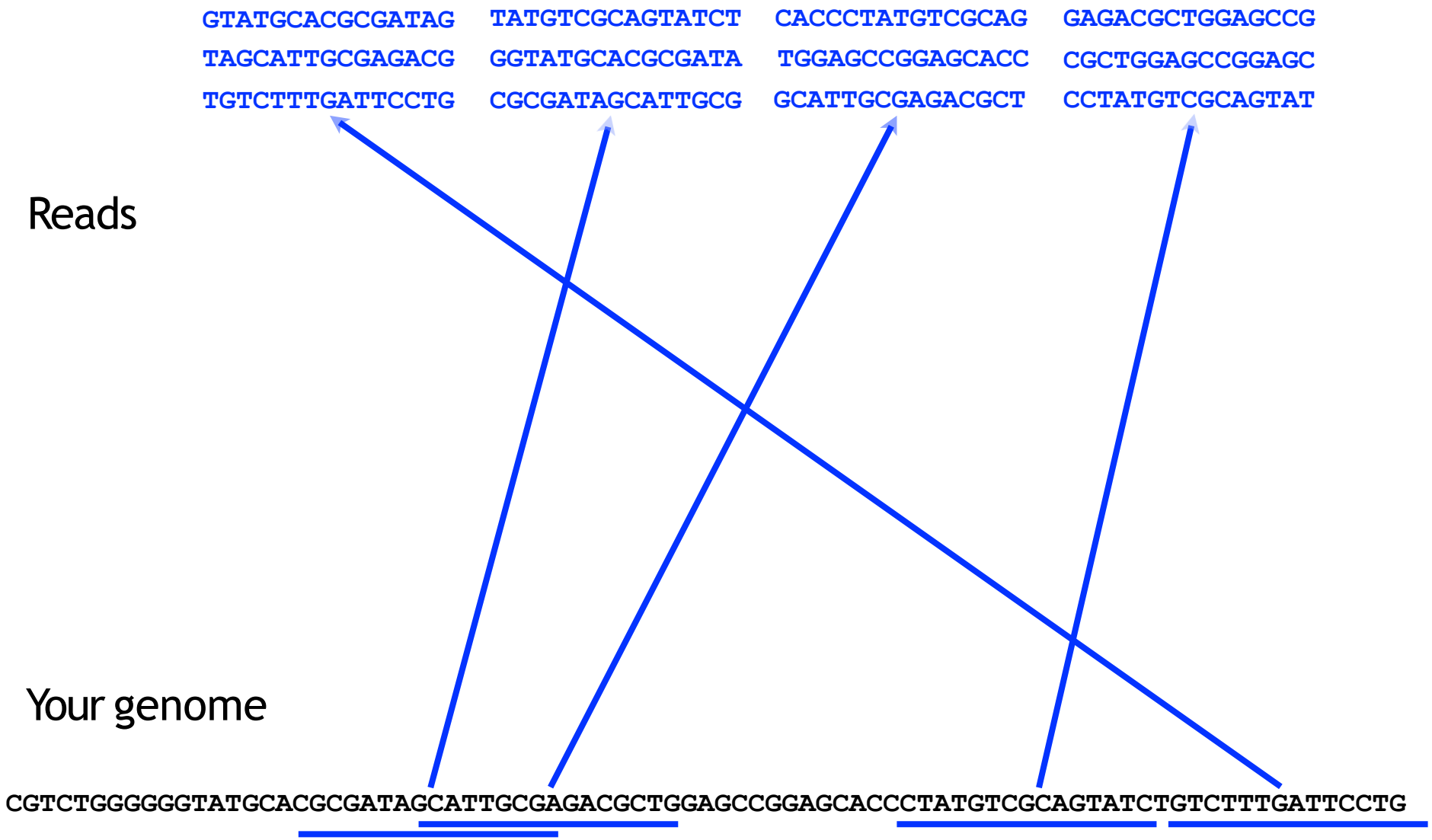
Reconstruction of a Genome



Reconstruction of a Genome



Reconstruction of a Genome



Reconstruction of a Genome

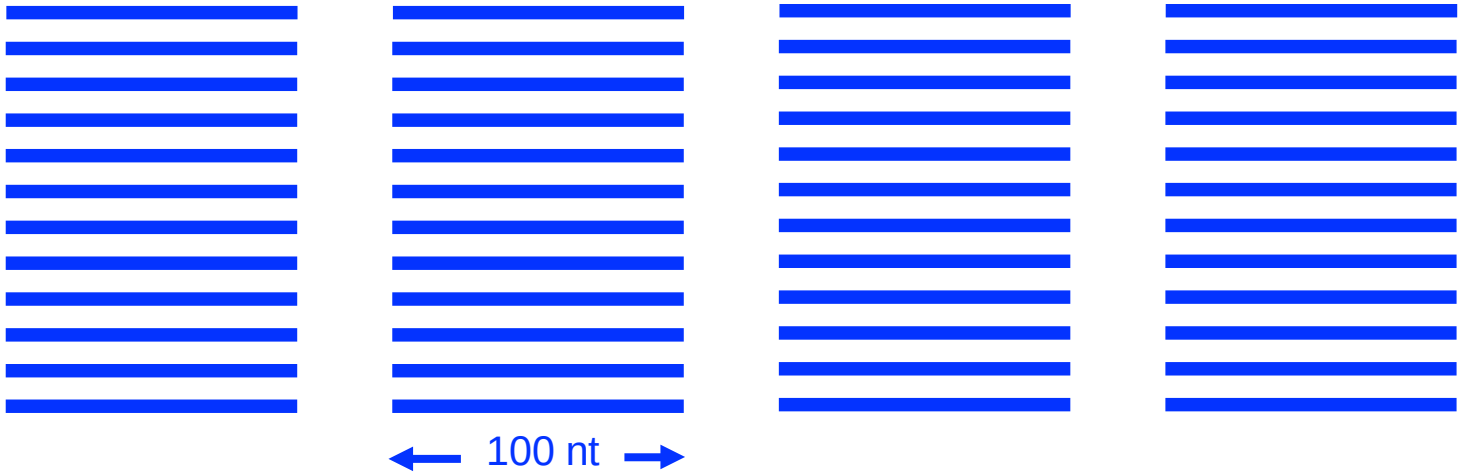
Reads

GTATGCACGCGATAG	TATGTCGCAGTATCT	CACCCTATGTCGCAG	GAGACGCTGGAGCCG
TAGCATTGCGAGACG	GGTATGCACGCGATA	TGGAGCCGGAGCACC	CGCTGGAGCCGGAGC
TGTCTTTGATTCCTG	CGCGATAGCATTGCG	GCATTGCGAGACGCT	CCTATGTCGCAGTAT
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG
GTATGCACGCGATAG	ACCTACGTTCAATAT	TATTTATCGCACCTA	CCACTCACGGGAGCT
GCGAGACGCTGGAGC	CTATCACCCTATTAA	CTGTCTTTGATTCCT	ACTCACGGGAGCTCT
CCTACGTTCAATATT	GCACCTACGTTCAAT	GTCTGGGGGGTATGC	AGCCGGAGCACCCTA
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG

Your genome

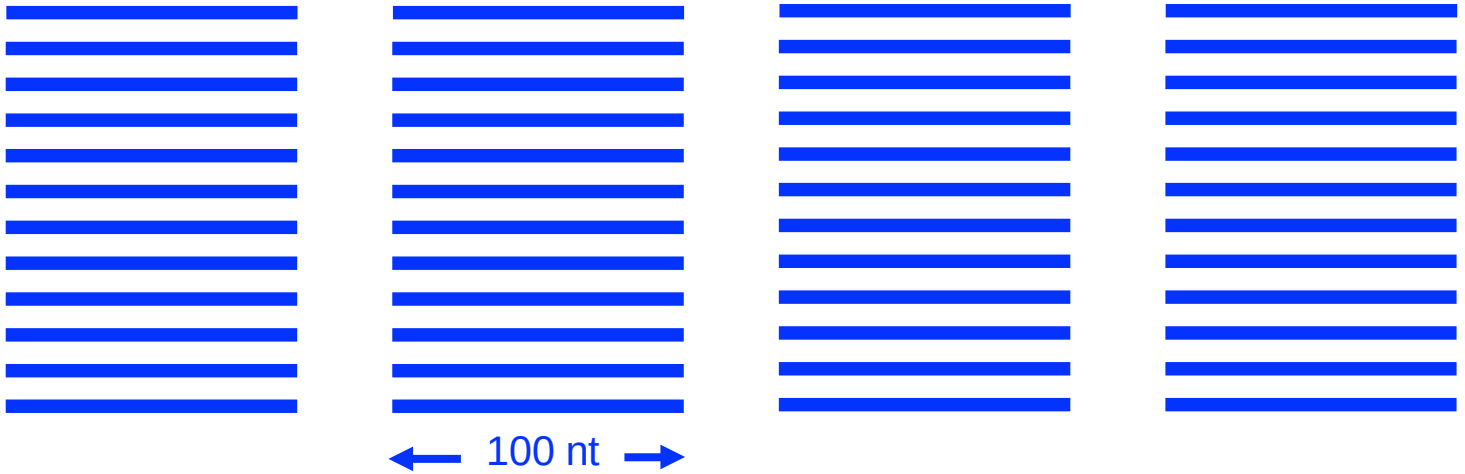
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

Reconstruction of a Genome



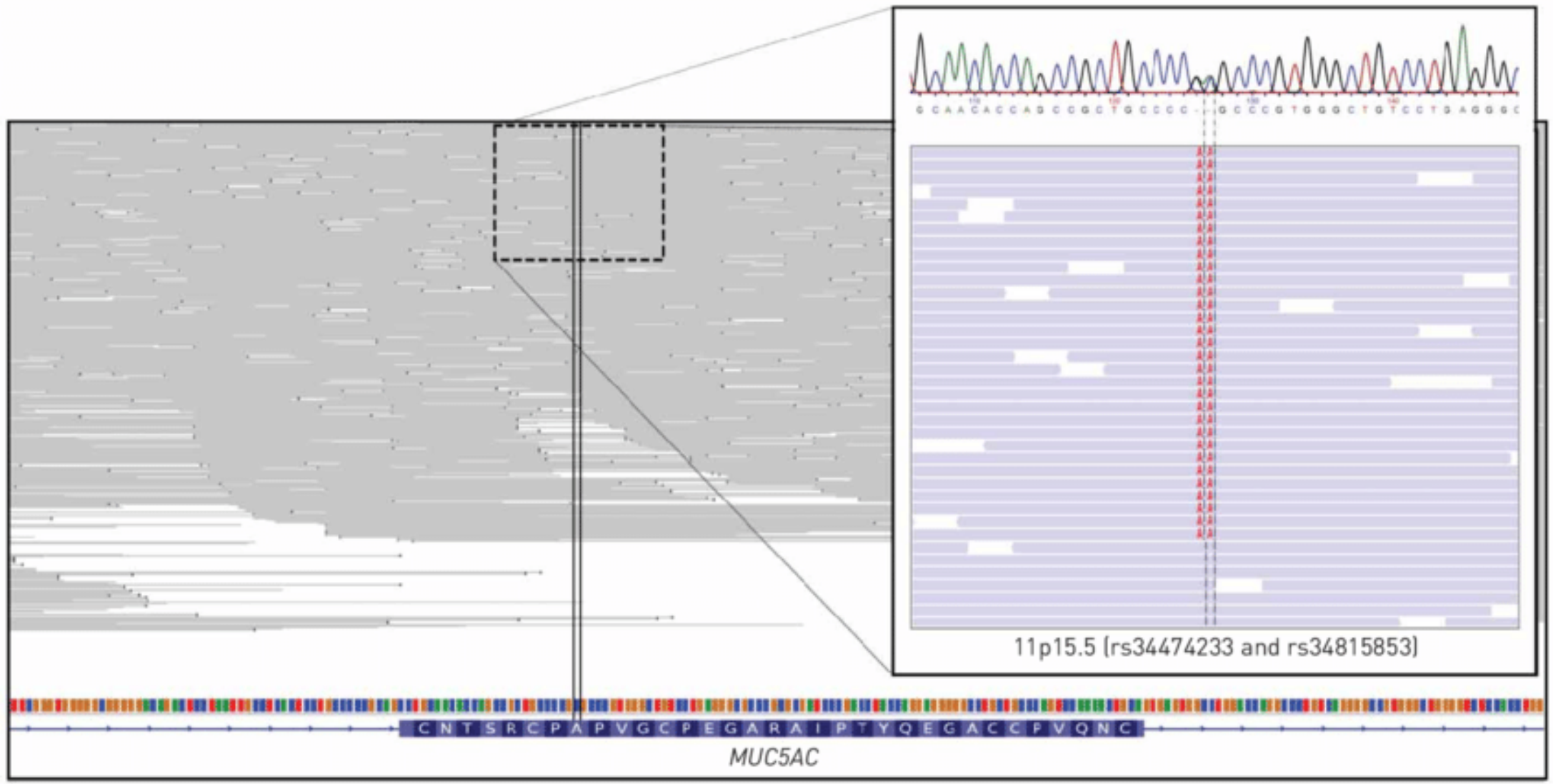
Your genome

Reconstruction of a Genome

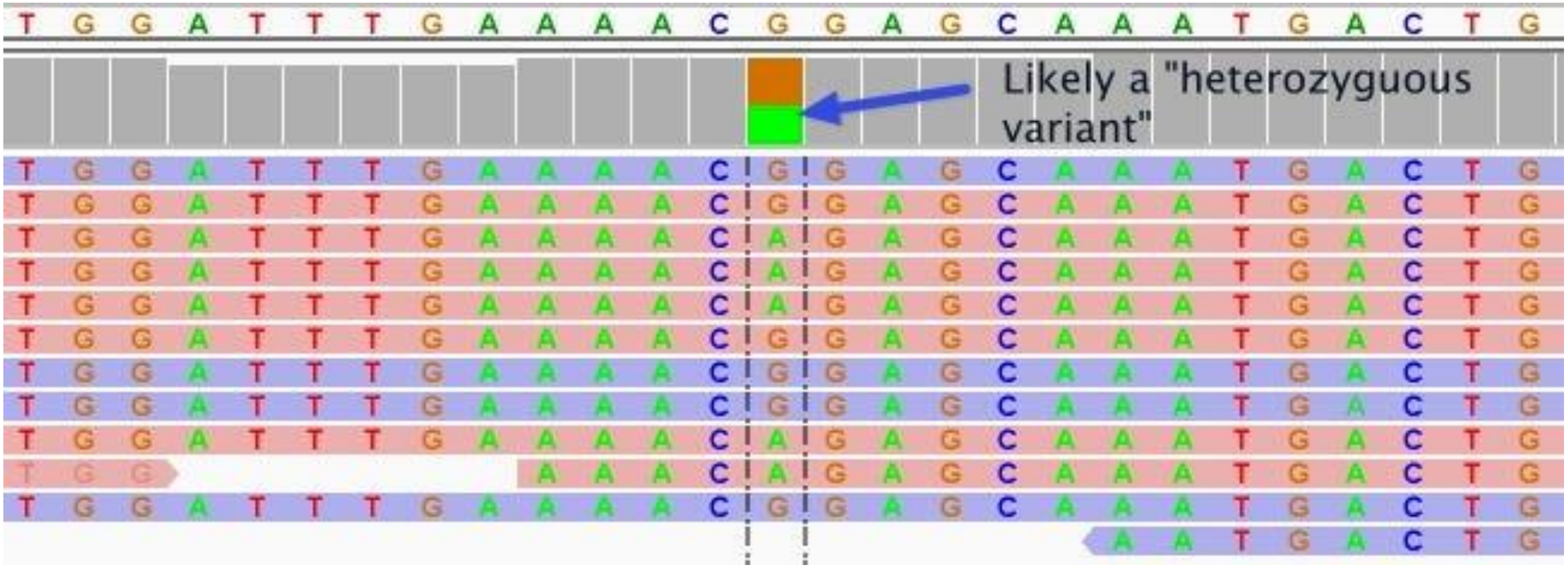


Your genome

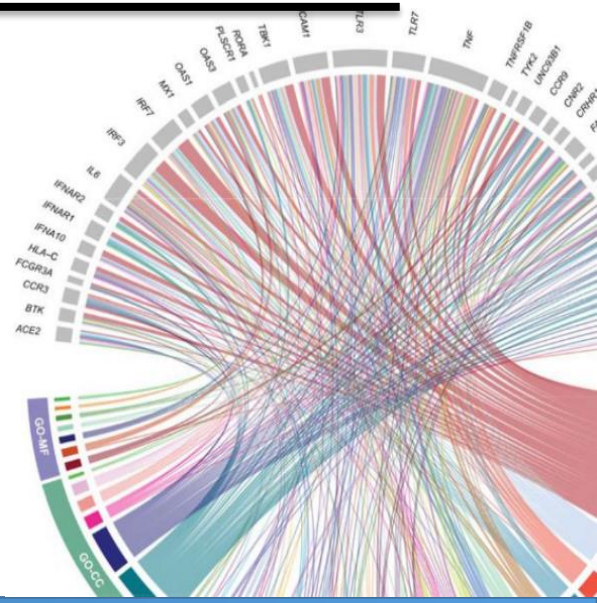
Reconstruction of a Genome



Reconstruction of a Genome



7 Genome Sequencing Data

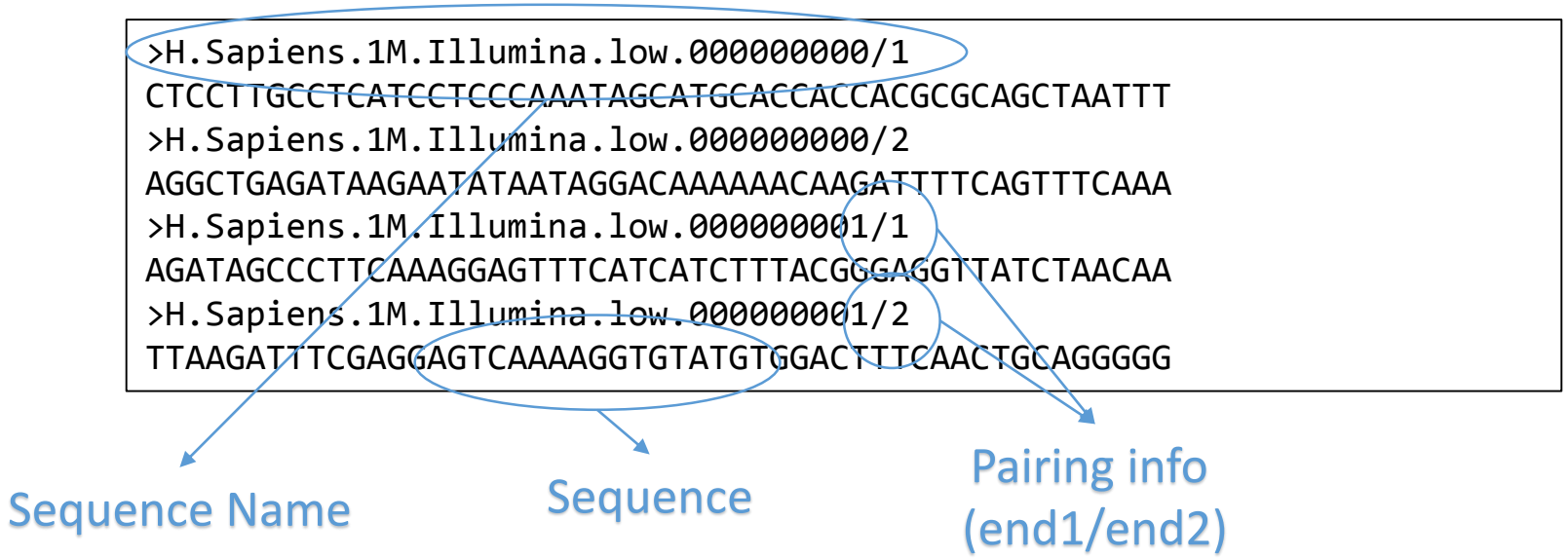


Bioinformatics File Formats

- Briefly introduce a few very relevant formats used in bioinformatics (sequence analysis).
  - **FASTQ/FASTA** – Stores sequenced reads with/without qualities
  - **MultiFASTA** – Stores collections of FASTA sequences (e.g. reference genomes)
  - **SAM** – Stores aligned sequenced reads against a reference genome
  - **VCF** – Stores variants from a sample against a reference
- These file formats are textual databases of genomic data
  - Stored in plain text format for simplicity and convenience
  - Data processing based on textual database handling

FASTA

- **FASTA** format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes
  - More info: https://en.wikipedia.org/wiki/FASTA_format



FASTQ

- **FASTQ** format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.
  - More info: https://en.wikipedia.org/wiki/FASTQ_format
  - Example: https://www.ebi.ac.uk/ena/data/view/SRX1615302

```
@H.Sapiens.1M.Illumina.l100.low.000000000/1
CTCCTTGCCTCATCCTCCCAAATAGCATGCACCACCACGCGCAGCTAATTT
+
KGOLHSAHTEFKMLGJPENFEGMIDAHKMIJFIEIHKCFCGHEGMJGHHPH
@H.Sapiens.1M.Illumina.l100.low.000000000/2
AGGCTGAGATAAGAATATAATAGGACAAAAAACAAGATTTTCAGTTTCAAA
+
IHHHIIIIHHIHHIHIIHHGFKGGIHGJGEIIFEEGIBHJKKJJIJJEJJF
@H.Sapiens.1M.Illumina.l100.low.000000001/1
AGATAGCCCTTCAAAGGAGTTTCATCATCTTTACGGGAGGTTATCTAACAA
+
FELPDDGC>PDFTDNF<MGCTG@IJ?HPGLCDGMCHIEMMIJDCJNIFBGF
@H.Sapiens.1M.Illumina.l100.low.000000001/2
TTAAGATTTCGAGGAGTCAAAAGGTGTATGTGGACTTTCAACTGCAGGGG
+
IIIHIHHIIHHHHHHIGHHHHGKGKJHHHHHHIJJJICGGFAHHIGLIKC
```

Quality String

Multi-FASTA

- A multi-FASTA file contains multiple FASTA formatted sequences.
  - Sequence-names can contain additional information (separated by space)
  - Sequences can span over multiple lines
  - Bases can appear masked (lowercase)

```
>seq001 description
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCACGAGCTAC
ACGACTAATCACGACTACAGCGACTCAGTCAGACTAGCTACACTGACTACGACTACGACT
>seq002 description
ACGACTAATCACGACTACAGCGACTCAGTCAGACTAGCTACACTGACTACGTCAGACTAC
ACGACTAATCACGACTACACTGACTACGACGACTAATCACGACTACGACTAATCACGACT
>seq003 description
AAAAACCATGCATCAGACAACACATGACTCACACAGTTGCATACCGACTCAGACTACTGT
GGGTCAAAGACTTTACGCGAAGGTACTCAGACATCGAGACTAGCTACATGCTAAGCGATC
CACACAAAAAGGGGCGATCAGCCACTGCCACGCTAGAATAAGCGATCTACGAGCTACGAT
AGCTATCACTACGATaaaaaaggggcaaaaACGACTCAGCGATCAGCTCATCAGCTACTA
```

Sequence Name

Description/Info (Optional)

Masked bases

Uncalled stretches

SAM

- Sequence Alignment Map (**SAM**) is a text-based format for storing biological sequences aligned to a reference sequence
  - Reference: https://samtools.github.io/hts-specs/SAMv1.pdf

```
@HD     VN:1.3
@SQ     SN:2micron      LN:6318
@SQ     SN:chrI LN:230208
@SQ     SN:chrII        LN:813178
...
@SQ     SN:chrXV        LN:1091289
@SQ     SN:chrXVI       LN:948062
@RG     ID:0    PG:GEM  PL:ILLUMINA     SM:0
@PG     ID:BWA  VN:1.875
Seq.001  83   chr17  29927 30 10M1D90M  =  2906   -422   CA..GG  BH..IC  NM:i:0  MD:Z:10T90
Seq.002  163  chr17  2906  57 100M      =  29927  422    AT..GT  HH..HH  NM:i:0  MD:Z:100
Seq.003  99   chrX   12466 56 36M1D64M  =  1242   323    CT..TA  IH..II  NM:i:0  MD:Z:36A64
Seq.004  147  chrX   1242  60 100M      =  12466  -323   CC..GC  II..IH  NM:i:0  MD:Z:73T26
```

Sequence Name
Flags
Chromosome
Position
MAPQ
CIGAR
Sequence
Qualities

Variant Call Format (VCF)

- The Variant Call Format (**VCF**) is a file format used to store gene sequence variations.
  - More info:

https://www.ebi.ac.uk/training/online/course/human-genetic-variation-introduction/exercise-title/want-know-how-we-did-it

```
chr1    47703379        .       C       T
chr1    48010488        .       G       A
chr1    48030838        .       A       T
chr1    48032875        .       CTAT    -
chr2    48032937        .       T       C
chr2    48033273        .       TTTTTGTTTTAATTCCT       -
chr2    48033551        .       C       G
chr3    48033910        .       A       T
chr4    5632048         .       G       T
```

CHROM: Chromosome where the variance is observed

POS: Position of the chromosome where the variance is observed

REF: Reference base (or bases in the case of an indel) at this locus

ALT: Alternative allele at this locus

Example. FASTQ to FASTA

- The following script reads a FASTQ file and converts it to FASTA

```
# Open the file
fastq_file = open("mbio.sample.fastq", "rt")
fasta_file = open("mbio.sample.fasta", "wt")
# If the file is not empty (line is not null) keep reading
while True:
    # Tag
    tag = fastq_file.readline()
    if not tag: break # End of file
    # Sequence
    sequence = fastq_file.readline()
    # '+'
    fastq_file.readline() # Ignore it
    # Qualities
    fastq_file.readline() # Ignore them
    # Write FASTA
    fasta_file.write(">%s\n%s\n" % (tag[1:-1:],sequence[0:-1:]))
# Close files
fastq_file.close()
fasta_file.close()
```

Example. FASTQ to FASTA

- The following script reads a FASTQ file and converts it to FASTA

```
# Open the file
fastq_file = open("mbio.sample.fastq", "rt")
fasta_file = open("mbio.sample.fasta", "wt")
# If the file is not empty (line is not null) keep reading
line_no = 0
for line in fastq_file:
    # Select line
    if line_no % 4 == 0:
        tag = line # Tag
    elif line_no % 4 == 1:
        sequence = line # Sequence
        # Write FASTA
        fasta_file.write(">%s\n%s\n" % (tag[1:-1:],sequence[0:-1:]))
    # Increment line number
    line_no += 1
# Close files
fastq_file.close()
fasta_file.close()
```

Questions