أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Instructor: Sakhaa Alsaedi
TA: Ebtihal Hani
Sakhaa.Alsaedi@kaust.edu.sa

Day 1: Sequence Analysis
9th January 2024

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Course Schedule and Content Overview

أكاديمية كاوست
KAUST ACADEMY

Day	Content Overview
Day 1	• Introduction to Bioinformatics • Sequence analysis
Day 2	• Genetic variant analysis
Day 3	• Diagnosing human genetic disease
Day 4	• Genome Assembly
Day 5	• Transcriptomics + Exam

Course Timing and Duration

- **Training Sessions:** Total of 6 hours,
  - Morning (9:00 AM - 12:00 PM)
  - Afternoon (2:00 PM - 5:00 PM)
- **Breaks:** 12:00 PM to 2:00 PM **(**Prayer and Lunch)
- **Operational Days**: 5 consecutive days, (9th - 13th of Jan.)

GitHub: https://github.com/Sakhaa-Alsaedi/Bioinformatics-/blob/main/README.md

أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Understanding the Digital Frontier in Biomedicine

Part 1: Introduction

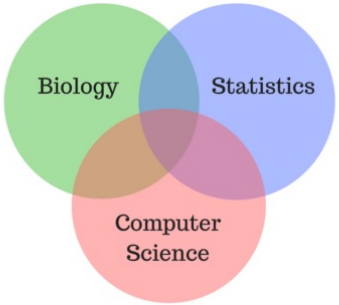
جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Bioinformatics

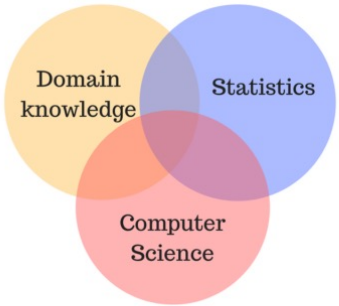
The science of collecting and analysing **complex** and **digital** **biological data**

→ **Understanding** the biological significance behind vast amounts of data in different applications

Bioinformatics



Data Science



Biological Data and Complexity

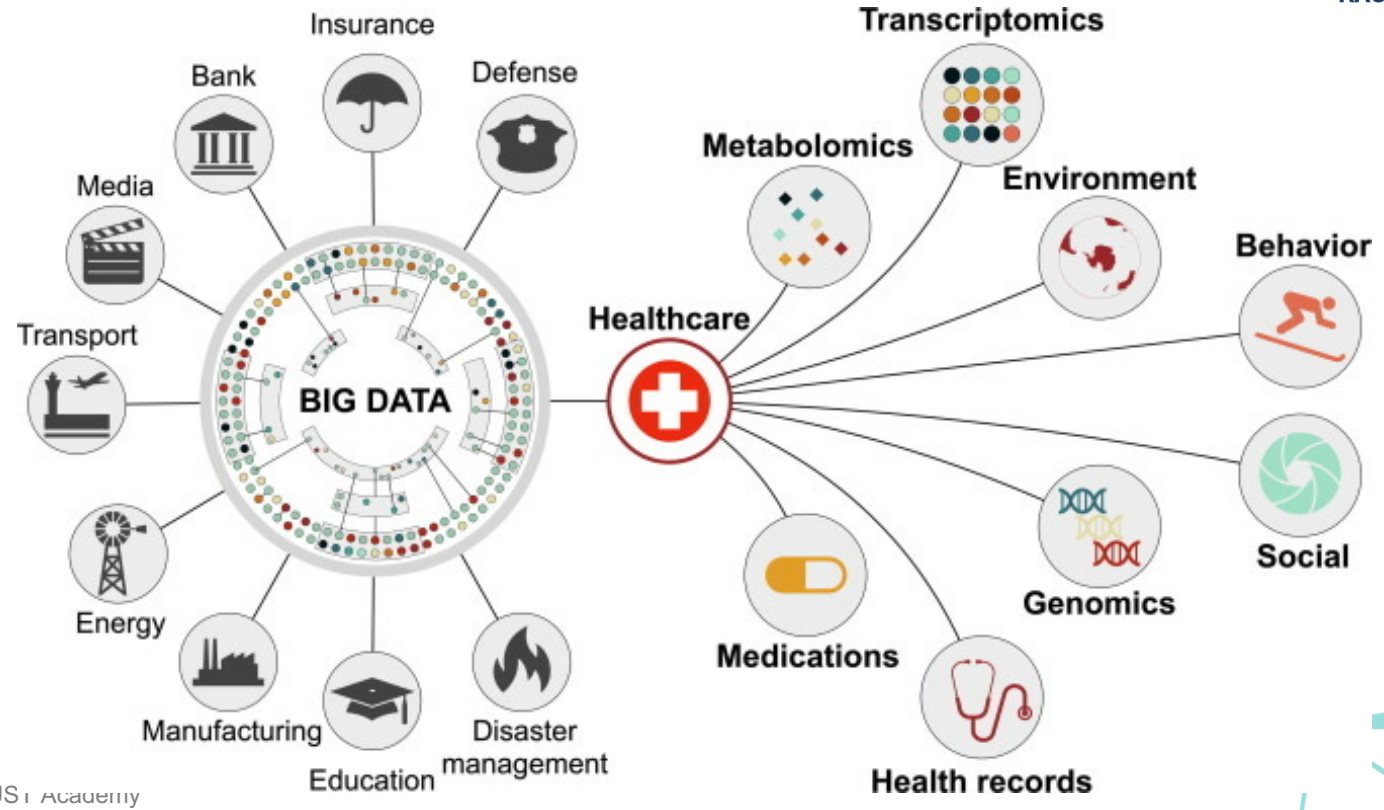
أكاديمية كاوست
KAUST ACADEMY

- Complex biological data refers to the intricately detailed, highly interconnected, and voluminous datasets that are characteristic of biological research and studies.

Dimensions of Biological Data Complexity

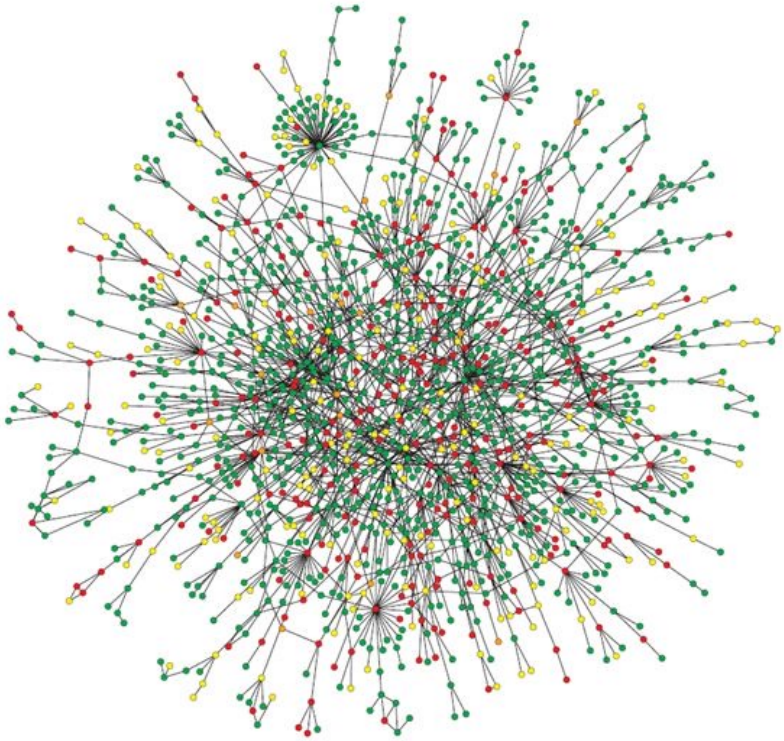
- Diversity of data Types, high dimensionality, and noise and uncertainty.
- Heterogeneity, Computational Intensity, and Rapid Data Generation.

Example: Biomedical Data → Big Data



Example: Complex Molecular Networks

أكاديمية كاوست
KAUST ACADEMY



- Protein-Protein interactions
- Protein-DNA interactions
- Genetic interactions
- Metabolic reactions
- Co-expression interactions
- Text mining interactions
- Association Networks
- Etc.

Barabasi & Oltvai, Nature Reviews, 2004

Digitizing Biomedical Data

أكاديمية كاوست
KAUST ACADEMY

- Converting traditional biological and medical data into digital formats for computational analysis.

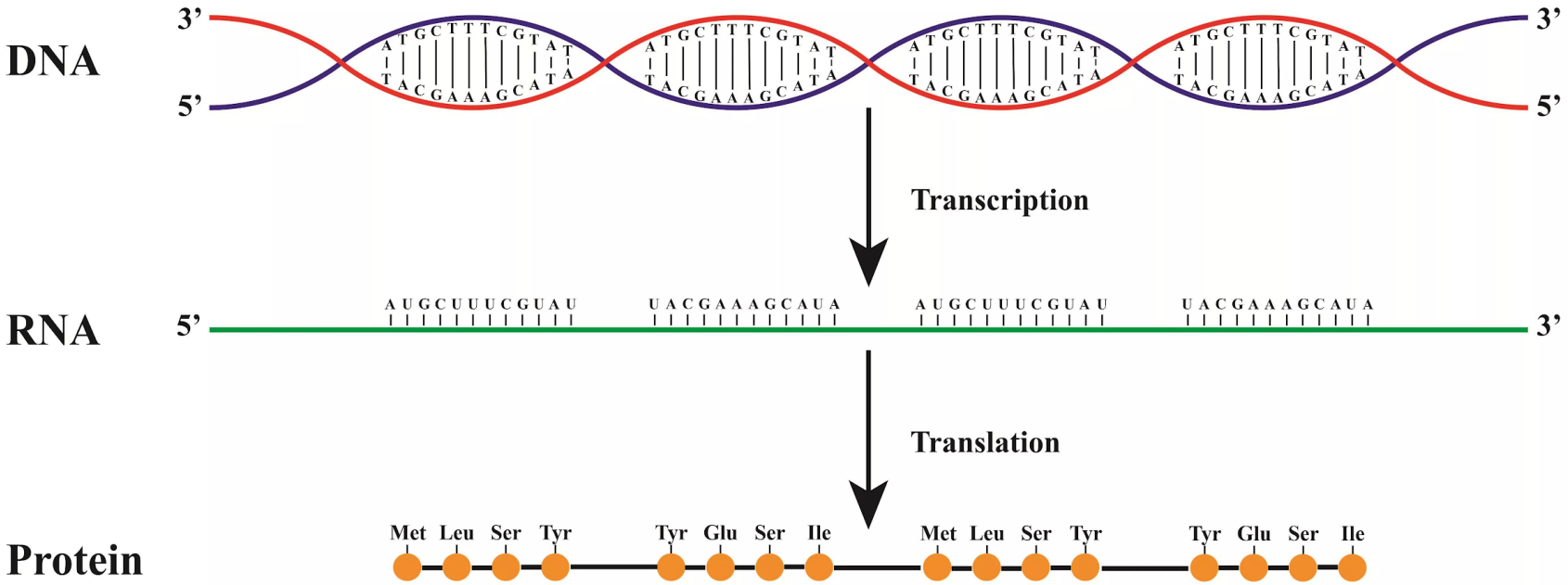
Why study computational algorithms?

- What is possible?
- What is practical?
- Where will the next contribution be?

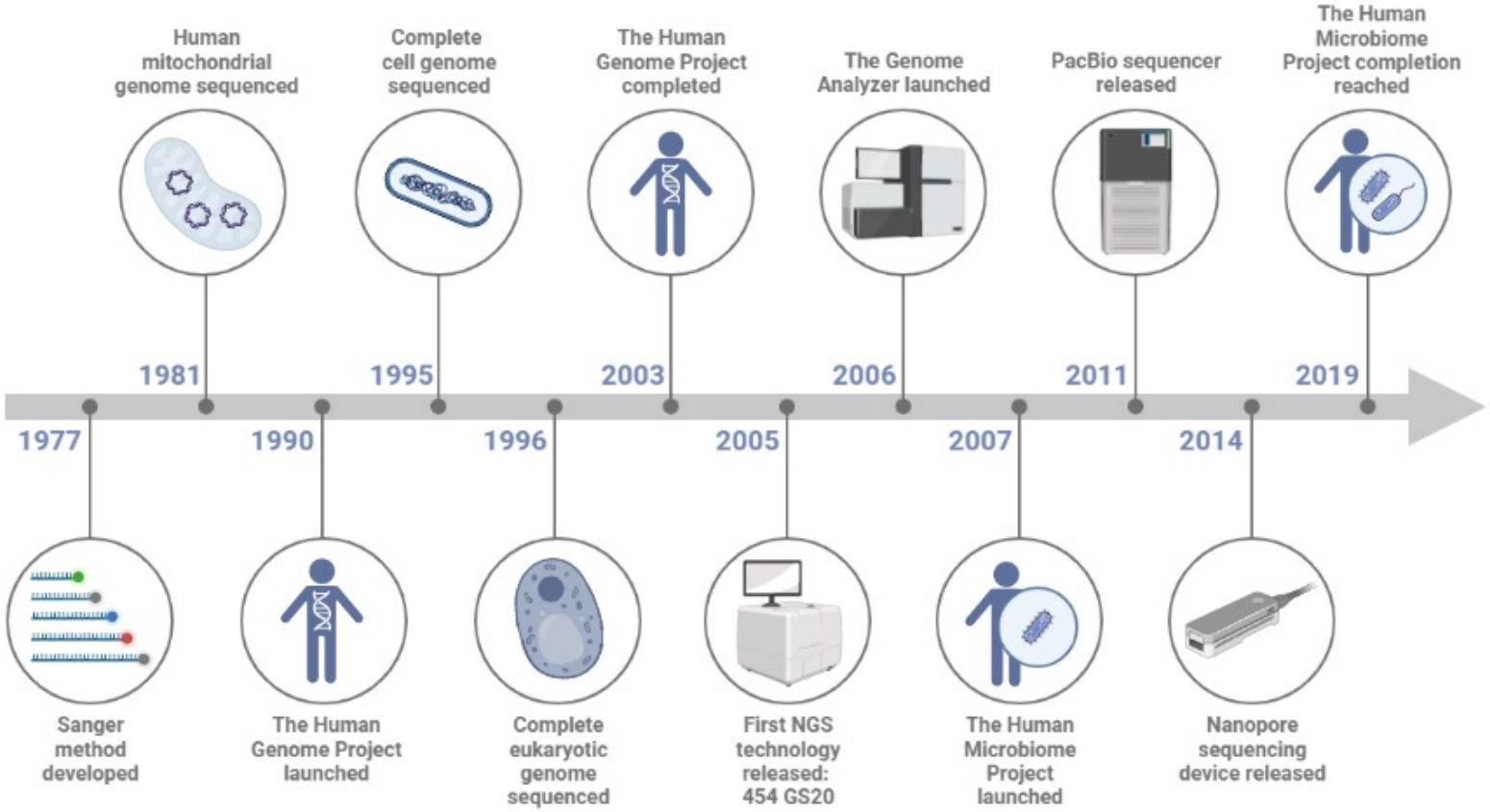
The Need for Digitizing Medical Data

- Enhances accessibility, storage, and analysis capabilities.
- Improved diagnosis, treatment planning, and predictive medicine.

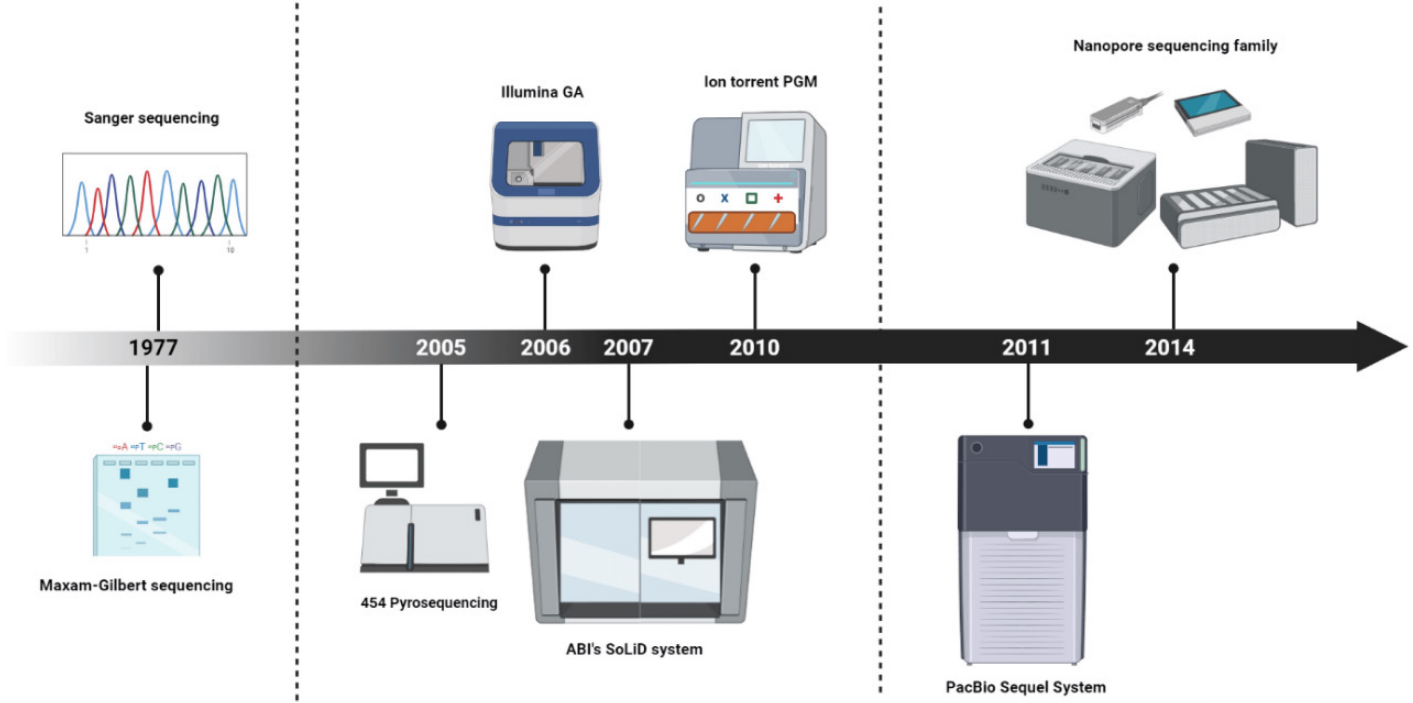
DNA Sequencing



History of Sequencing Technology



History of Sequencing Technology

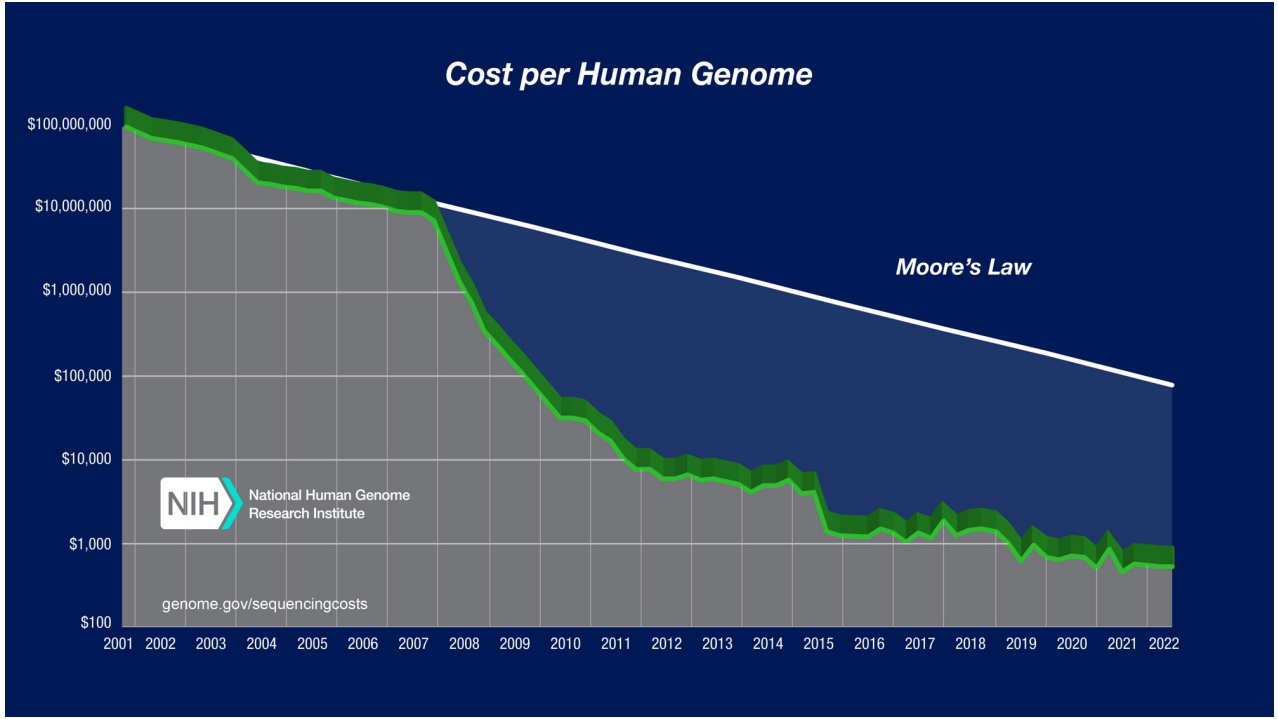


First-generation sequencing

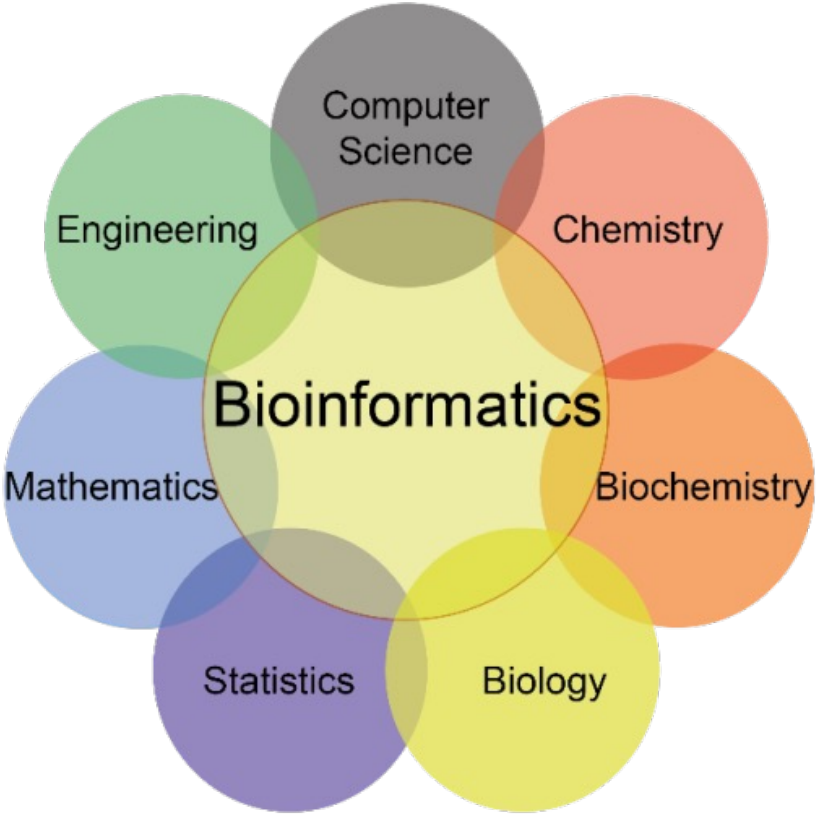
Next-generation sequencing (NGS)

Third-generation sequencing

Sequencing Technology Cost



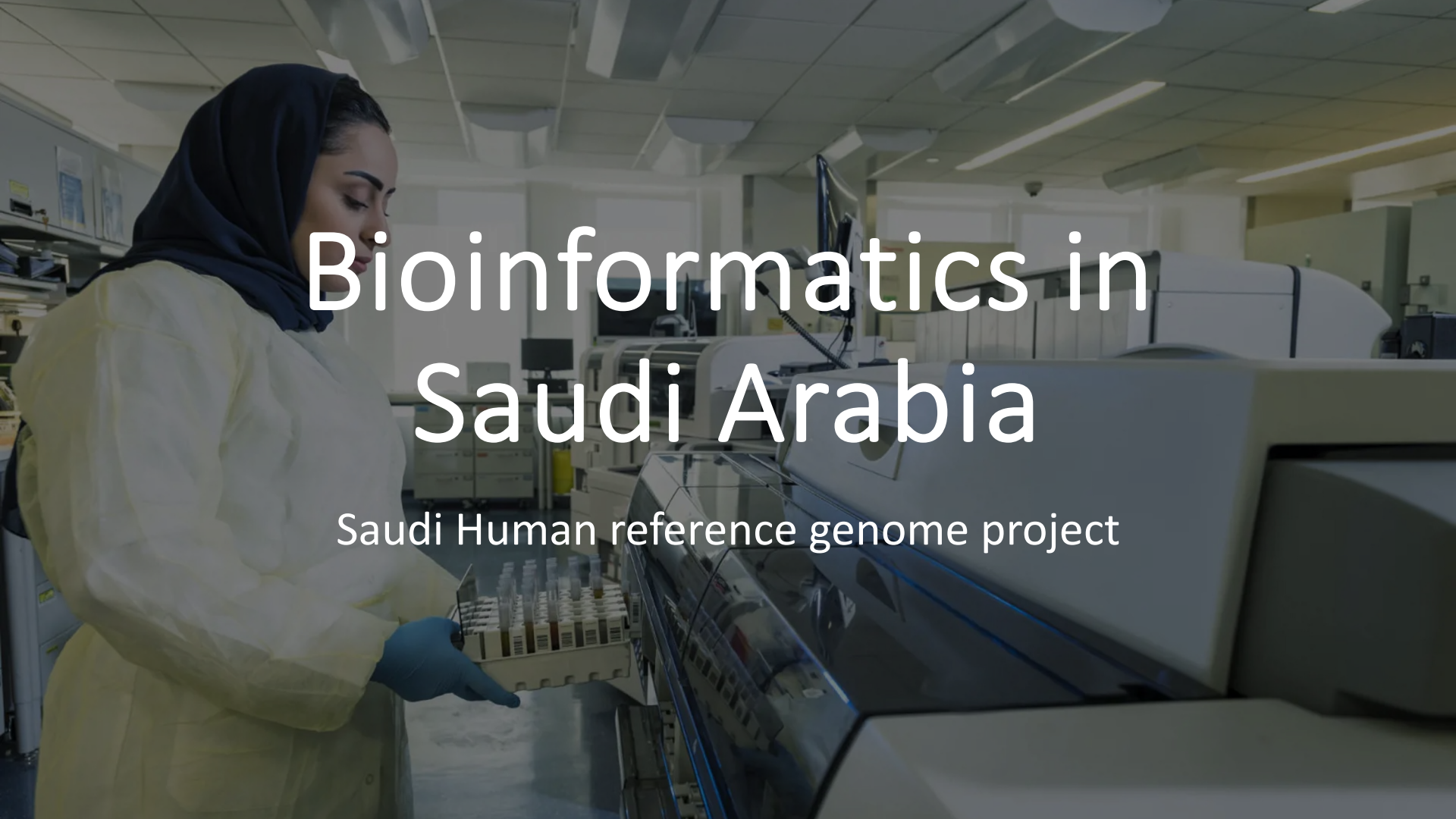
Interdisciplinary Foundations of Bioinformatics



Saudi gene hunters comb country's DNA to prevent rare diseases

Research could help prevent disorders that result from marriages between relatives

8 DEC 2016 · BY JOCELYN KAISER



Bioinformatics in Saudi Arabia

Saudi Human reference genome project

Bioinformatics in Saudi Arabia

- **SARS-COV2 comparative analysis**

NoorDX Startup

- Rapid COVID-19 PCR test
- Metagenomics analysis
- KAUST

KMAP Platform

أكاديمية كاوست
KAUST ACADEMY

KAUST Metagenomic Analysis Platform (KMAP), enabling access to massive analytics of re-annotated metagenomic data

Intikhab Alam [1], Allan Anthony Kamau [2], David Kamanda Ngugi [3], Takashi Gojobori [2], Carlos M Duarte [2 4], Vladimir B Bajic [2]

Affiliations + expand

PMID: 34075103 PMCID: PMC8169707 DOI: 10.1038/s41598-021-90799-y

Free PMC article

Microbial Habitats

1 Samples

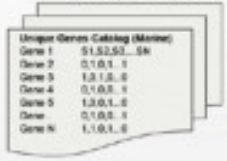
- Metadata [Temp., Salinity]
- Shotgun metagenomic sequencing
- 40,000-60,000 free living or host associated shotgun metagenomic samples available at EBI .

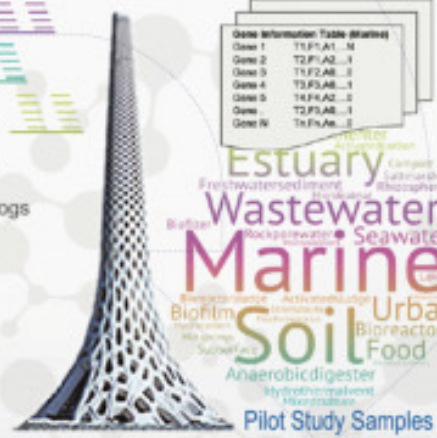
Microbes

Microbes are everywhere, an estimated number is ~1 trillion species.

2 Assemble Metagenomes

- Get full-length genes
- Make unique genes catalogs & map sample reads





Pilot Study Samples Habitats

4 Data Sharing

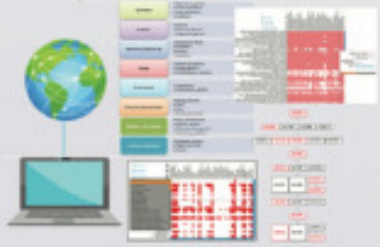
Shaheen and KMAP were used to produce Pilot Study: 40 Gene Catalogs, 275 million genes.

Research Groups

Gene information tables (AAMG TSV) available for research organizations with advanced computational resources and skills.

Individuals

Graphic User Interface access to indexed gene information tables for browsing and comparisons. Available for individuals with less computational resources.



THE FUTURE OF HEALTH

01

Provide a comprehensive experience throughout the entire care process

02

Support individual empowerment and proactive prevention

03

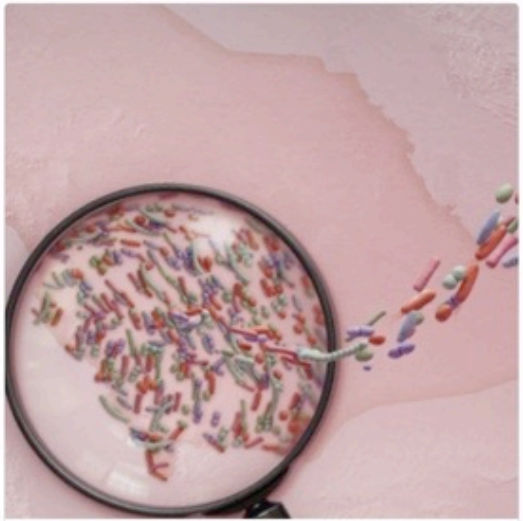
Design a digital-first ecosystem that's available 24/7

04

Spearhead genetics as the future of personalized care

KAUST Smart-Health Initiative

Health and Wellness: Accelerating Impact in KSA



04 December, 2023

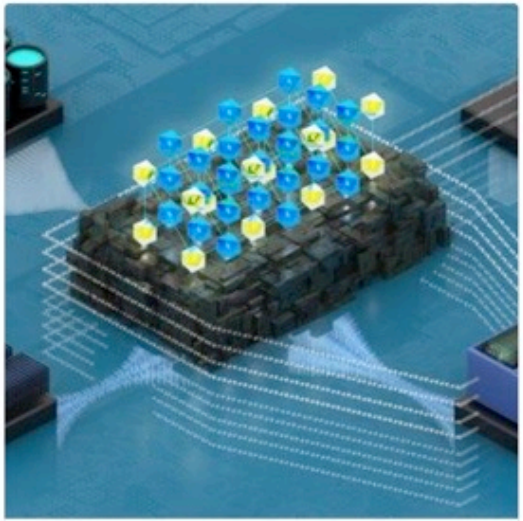
The weird and wonderful world of Saudi Arabia's microbiomes

Saudi Arabia is home to microbial communities that can boost coral reef health, sequester carbon in

03 December, 2023

The first ride across Saudi Arabia on a hand bike

To showcase the remarkable potential of people with physical disabilities, KAUST Professor Matteo Parsani will travel from the east to the west of Saudi Arabia by



22 November, 2023

Safeguarding the right to be forgotten

An open-source software can help align artificial intelligence applications in healthcare with data privacy regulations.

أكاديمية كاوست
KAUST ACADEMY

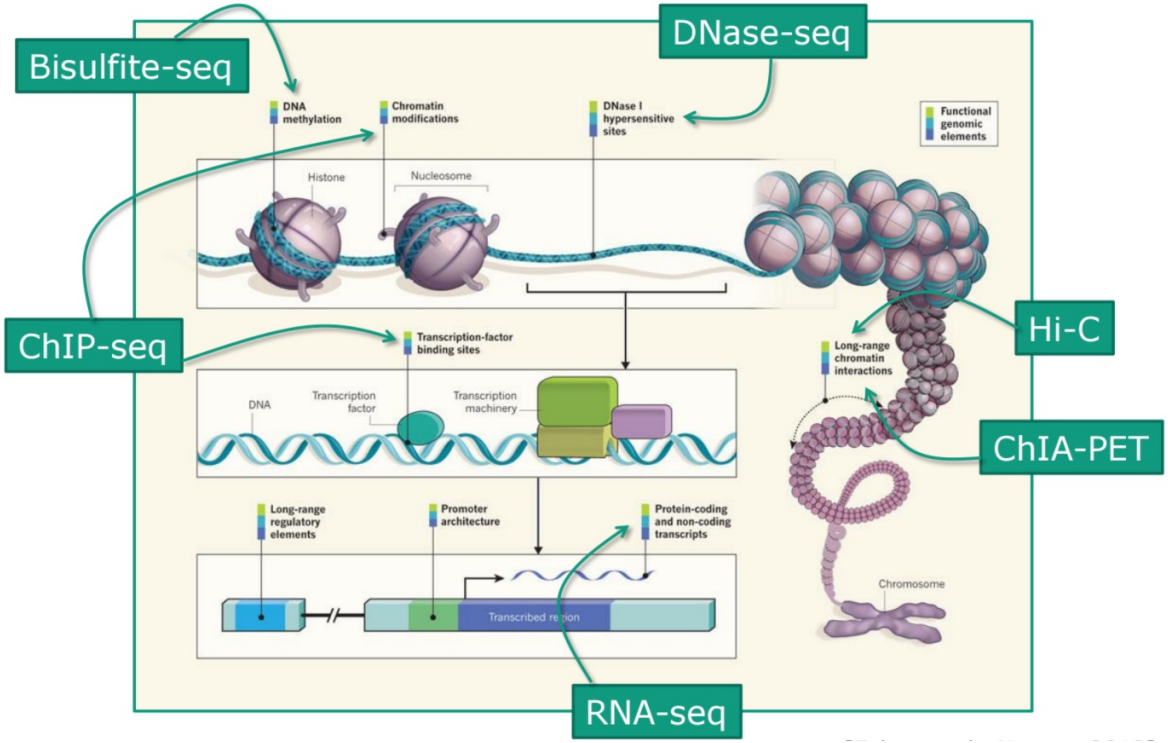
Introduction to Bioinformatics

How to Get Started in Bioinformatics

Part 2: Sequence analysis

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Where is Data Coming From?



Ecker et al, Nature, 2012

Sequencing Technologies

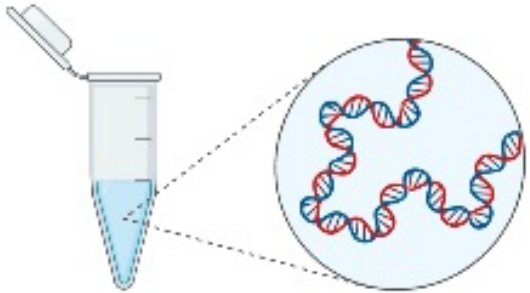
NGS Technology	Description
ChIP-seq	Analyzes protein-DNA interactions to identify binding sites of DNA-associated proteins
RNA-seq	Sequences RNA to examine transcriptome profiles, including gene expression and splicing patterns
Exome-seq	Focuses on sequencing exons, the coding regions of the genome, to find genetic variants
DNA-seq	Involves sequencing the whole DNA to identify genetic variations and mutations.
Metagenomic Sequencing	Studies genetic material directly from environmental samples.
Single-Cell Sequencing	Analyzes genomes or transcriptomes of individual cells for detailed cellular variation
Methyl-Seq	Studies DNA methylation, an important epigenetic modification
ChIP-exo	An enhanced version of ChIP-seq offering higher resolution.
ATAC-seq	Assesses chromatin accessibility, important for understanding transcriptional regulation.

Wet Lab

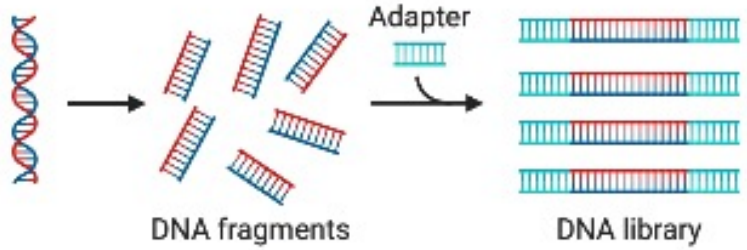
Dry Lab

Next Generation Sequencing Workflow

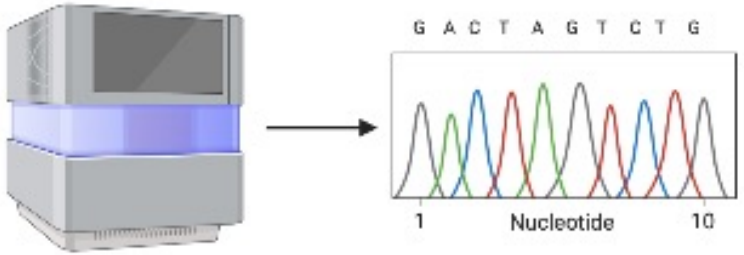
Step 1:
DNA extraction



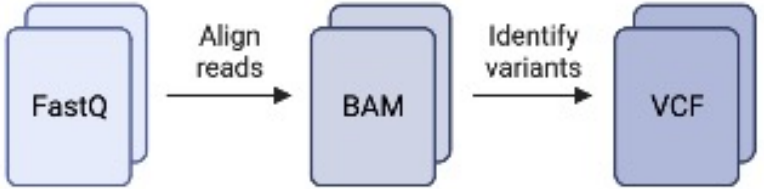
Step 2:
Library preparation



Step 3:
Sequencing



Step 4:
Analysis

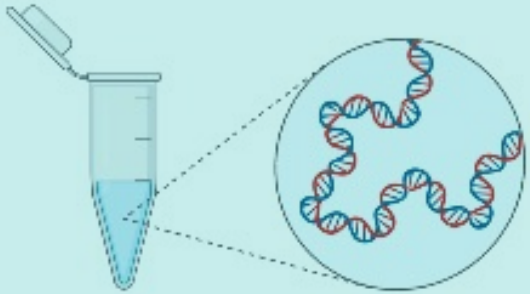


Next Generation Sequencing Workflow

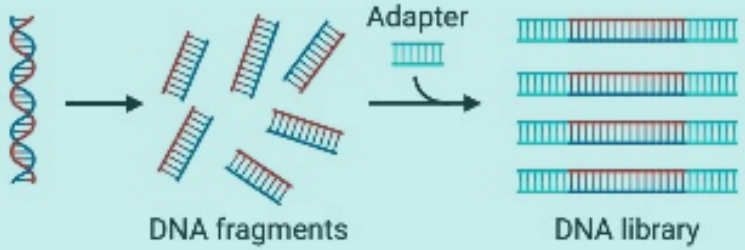
Wet Lab

Dry Lab

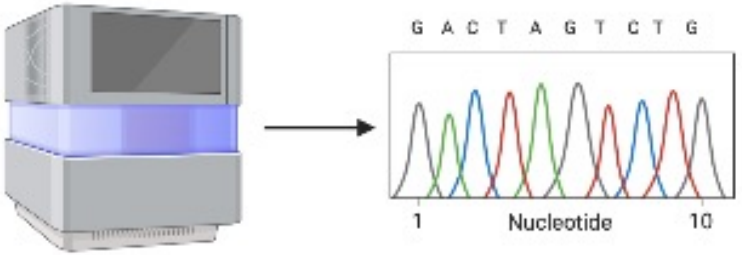
Step 1:
DNA extraction



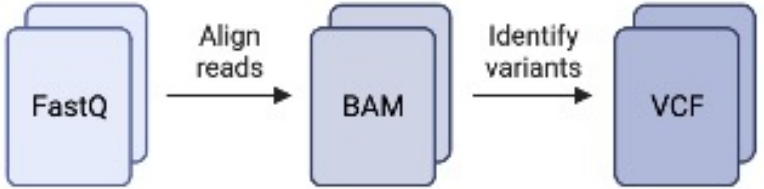
Step 2:
Library preparation



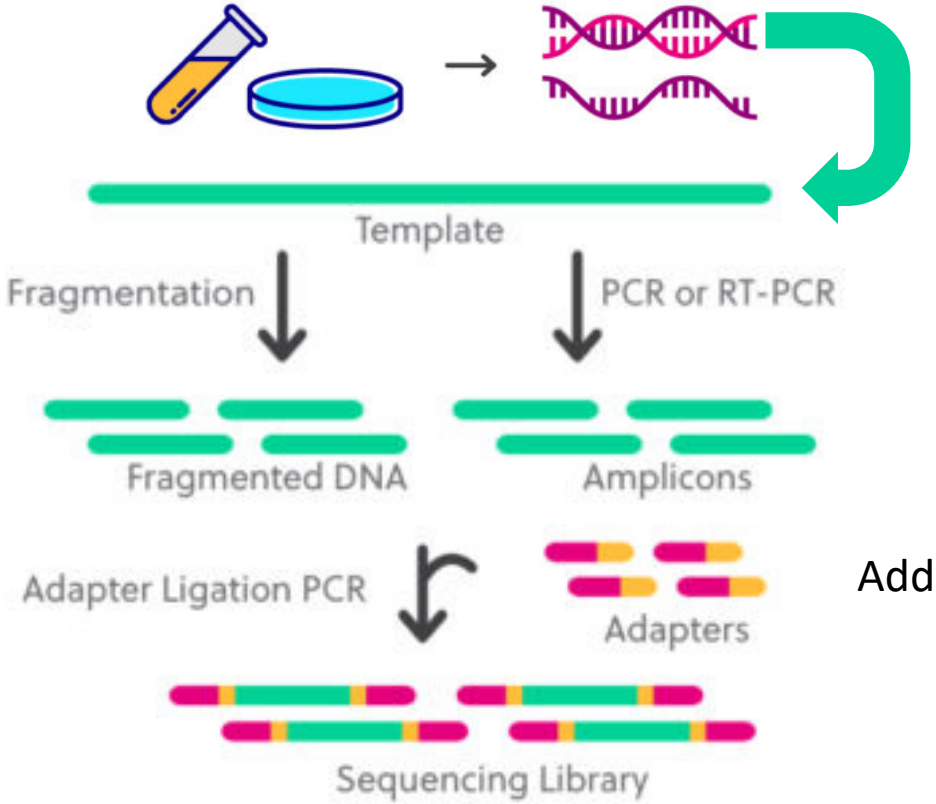
Step 3:
Sequencing



Step 4:
Analysis



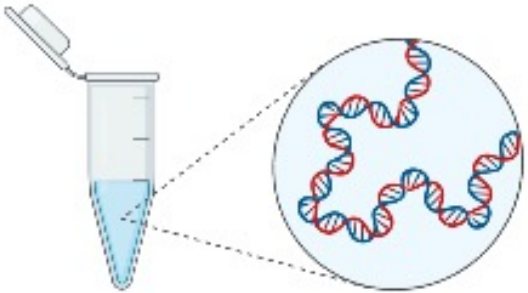
Wet-Lab: Library Preparation



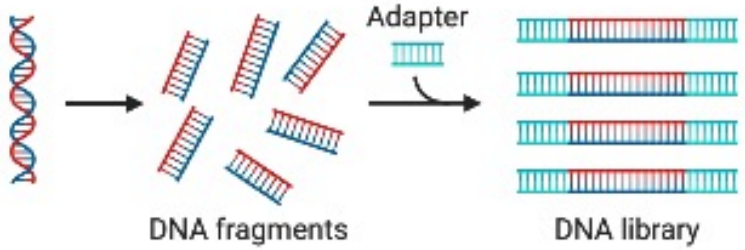
Add your ingredents ☺

Wet Lab

Step 1:
DNA extraction



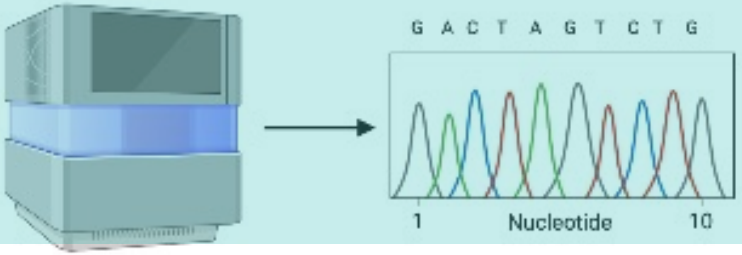
Step 2:
Library preparation



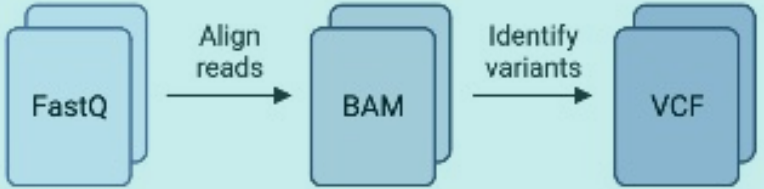
Next Generation Sequencing Workflow

Dry Lab

Step 3:
Sequencing



Step 4:
Analysis



DNA Sequncer

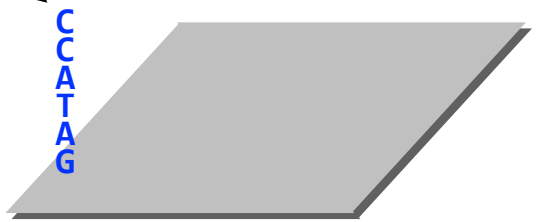
Input DNA

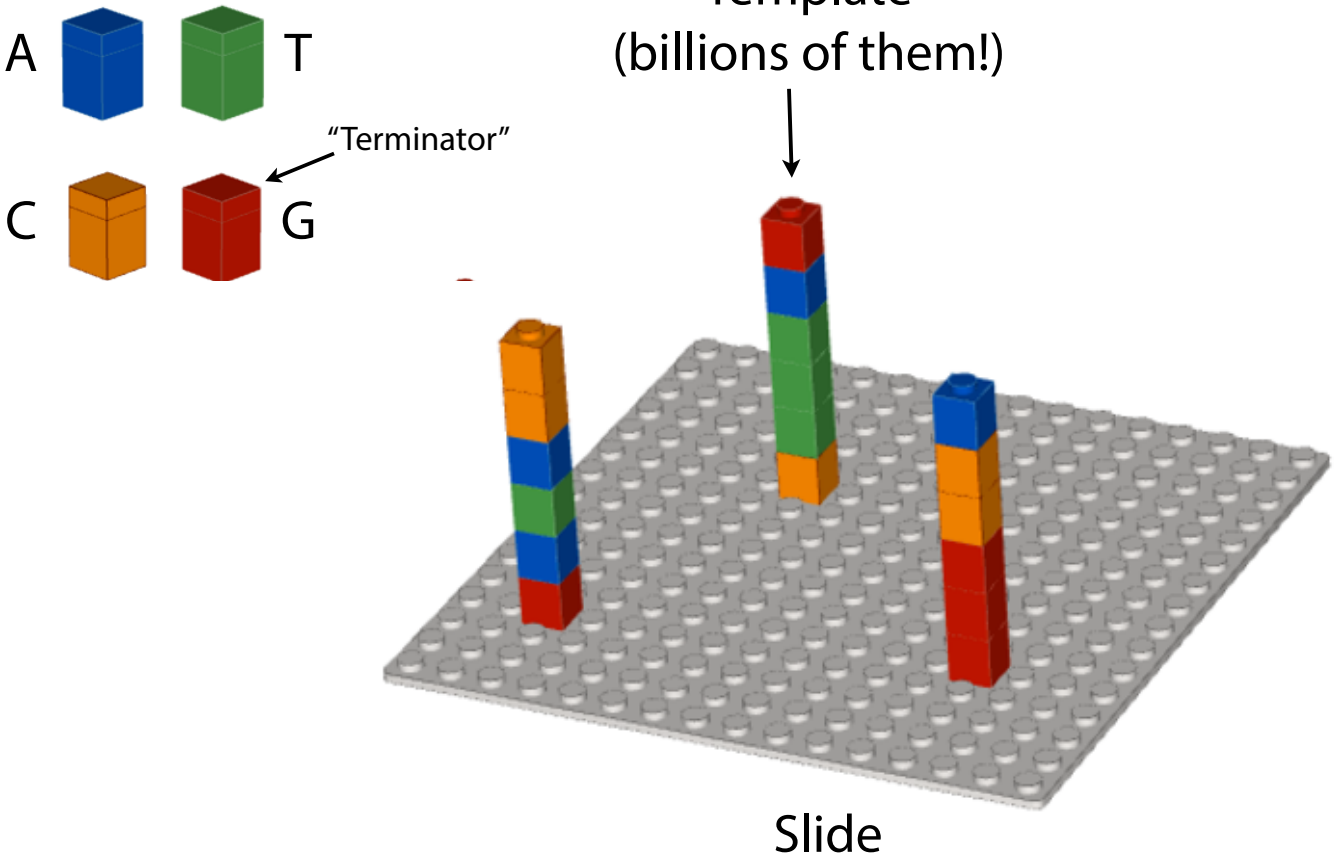
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT
CCATAGTATATCTCGGCTCTAGGCCCTCATTTTTT

Cut into snippets

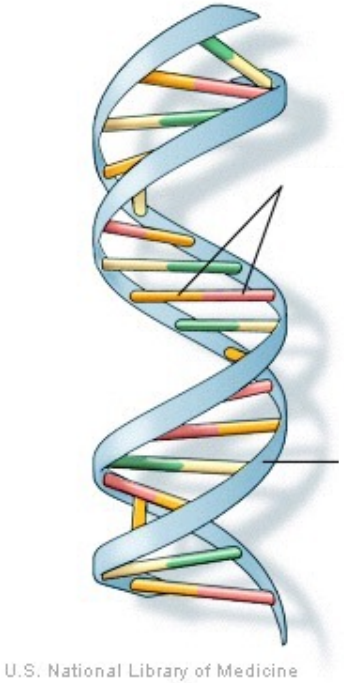
CCATAGTA TATCTCGG CTCTAGGCCCTC ATTTTTT
CCA TAGTATAT CTCGGCTCTAGGCCCTCA TTTTTT
CCATAGTAT ATCTCGGCTCTAG GCCCTCA TTTTTT
CCATAG TATATCT CGGCTCTAGGCCCT CATTTTTT

Deposit on slide

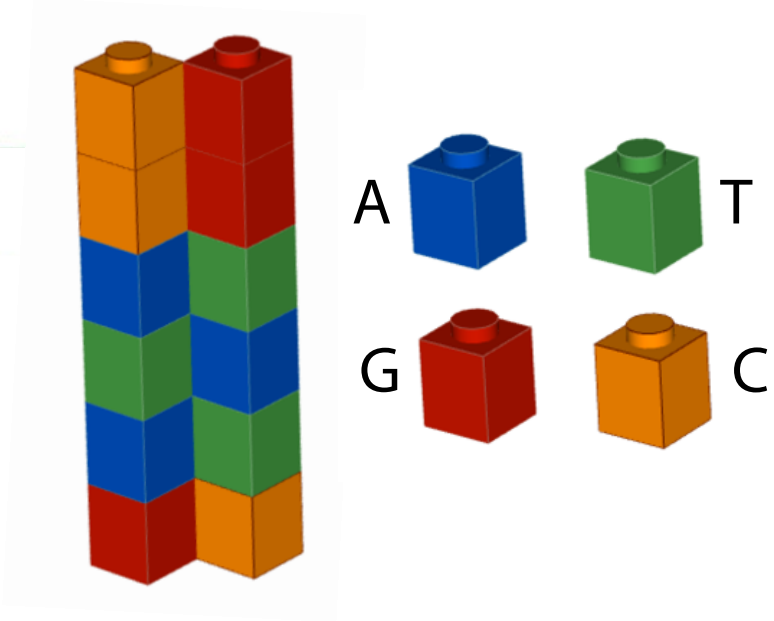



A
T
C
G
"Terminator"
Template
(billions of them!)
Slide

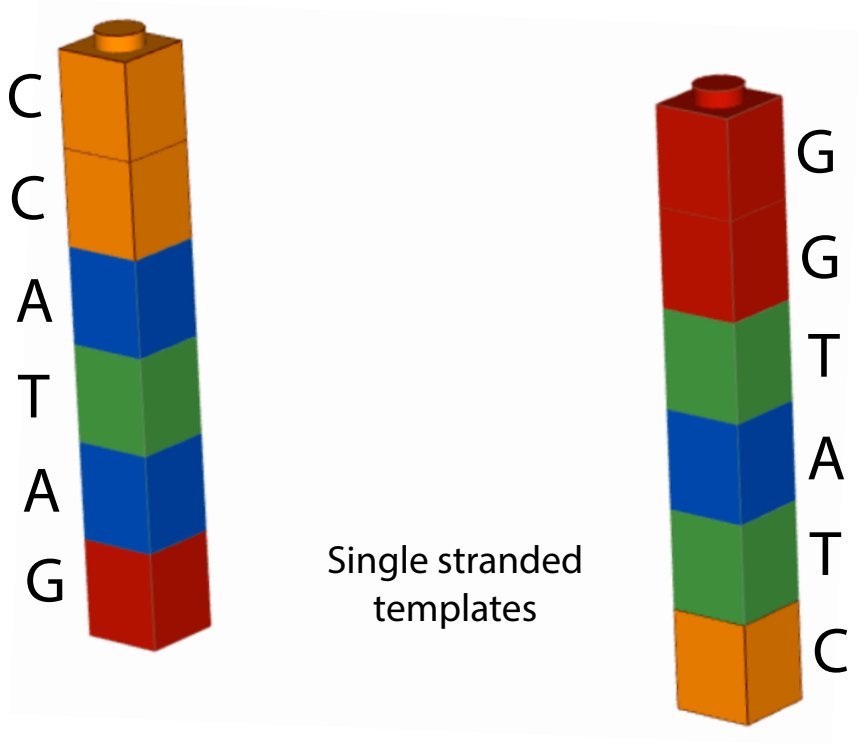
How DNA is copied

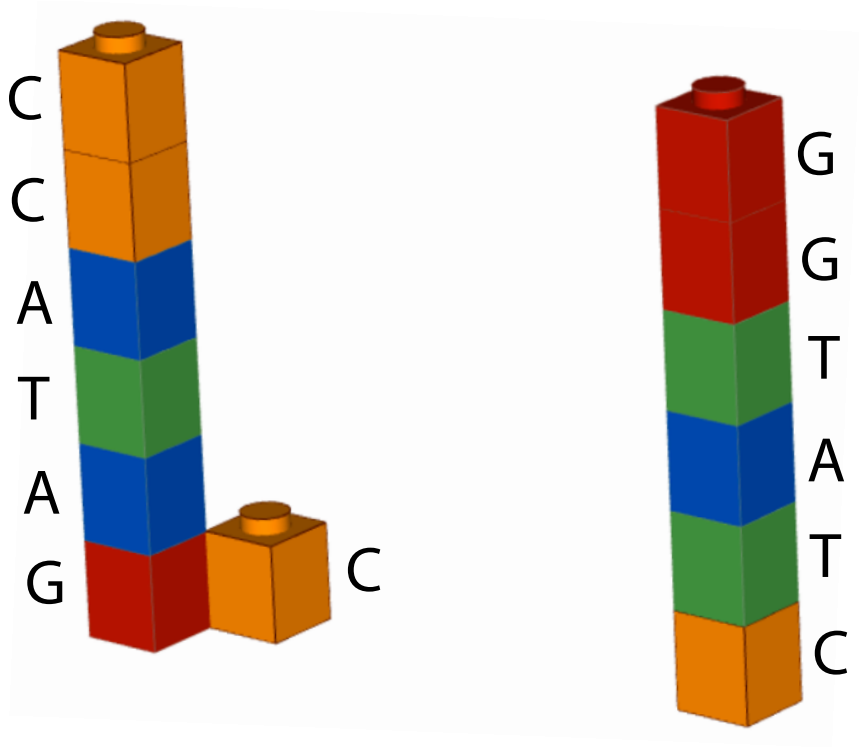


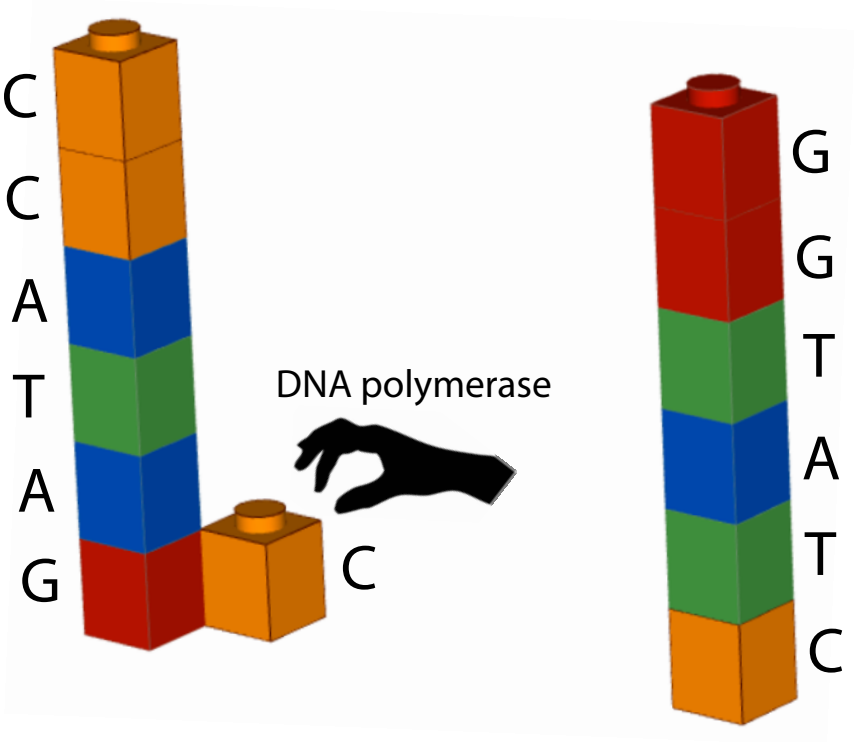
Double stranded DNA (double helix)

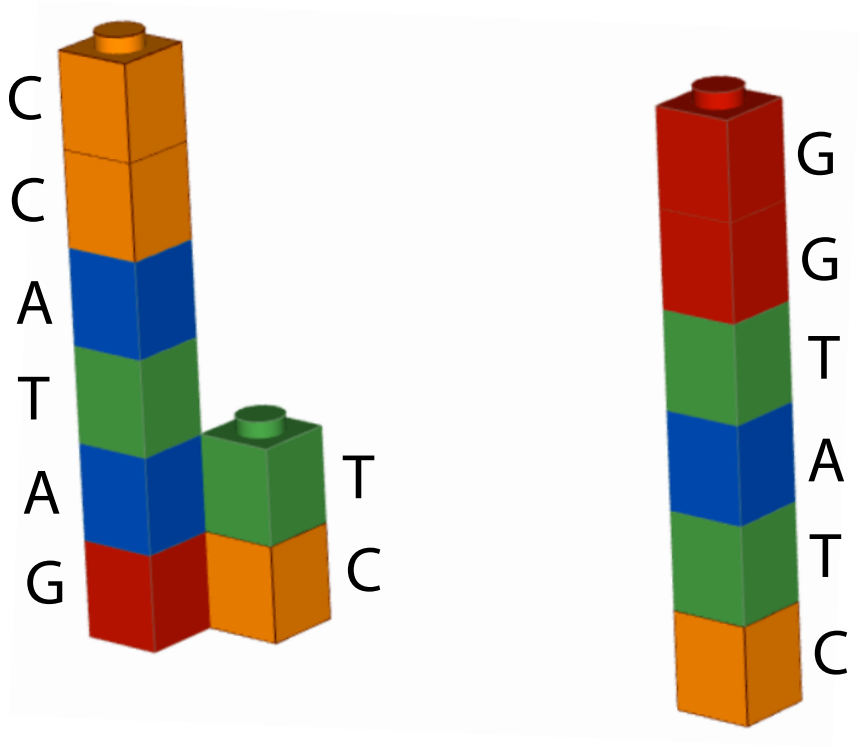


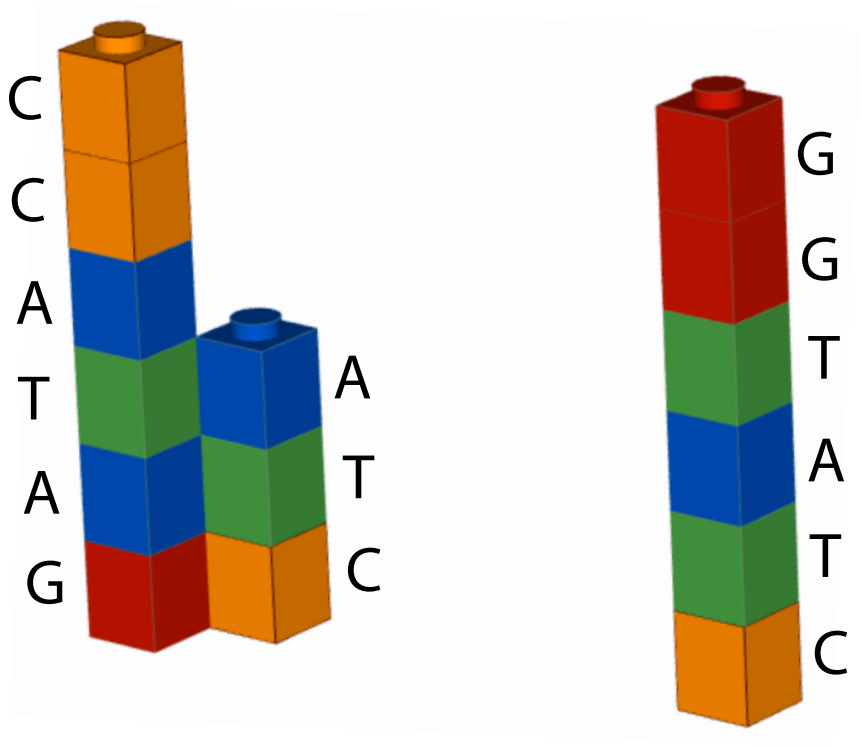
Double stranded DNA (lego version)

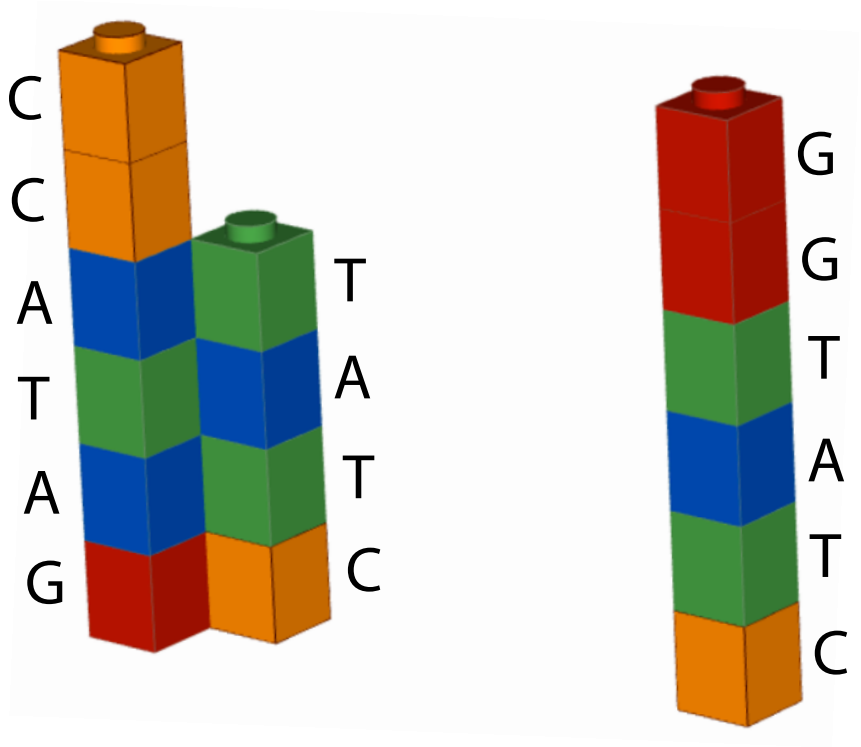

C
C
A
T
A
G
Single stranded templates
G
G
T
A
T
C

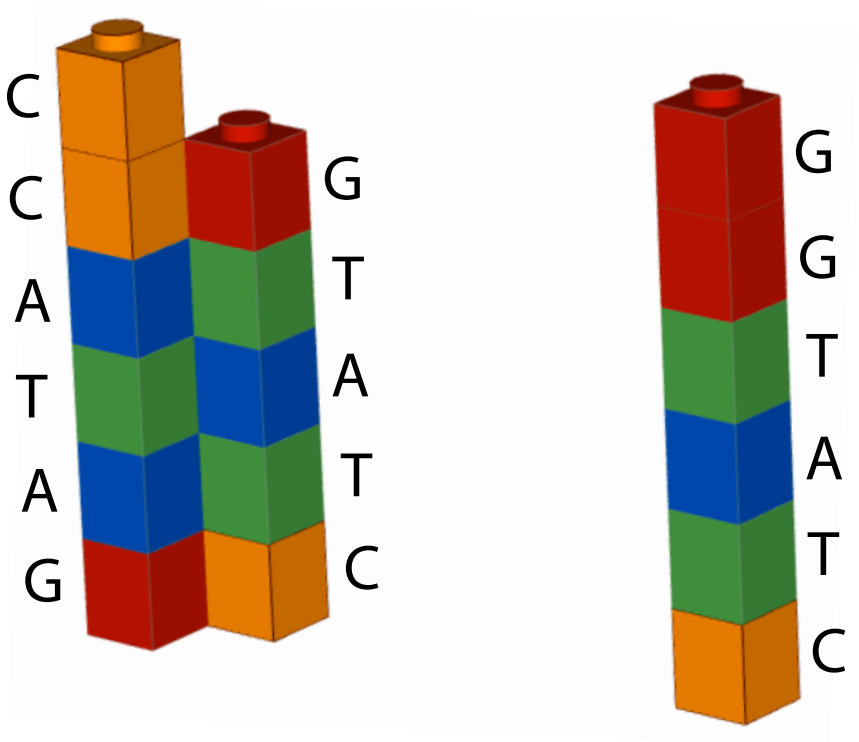

C
C
A
T
A
G
C
G
G
T
A
T
C

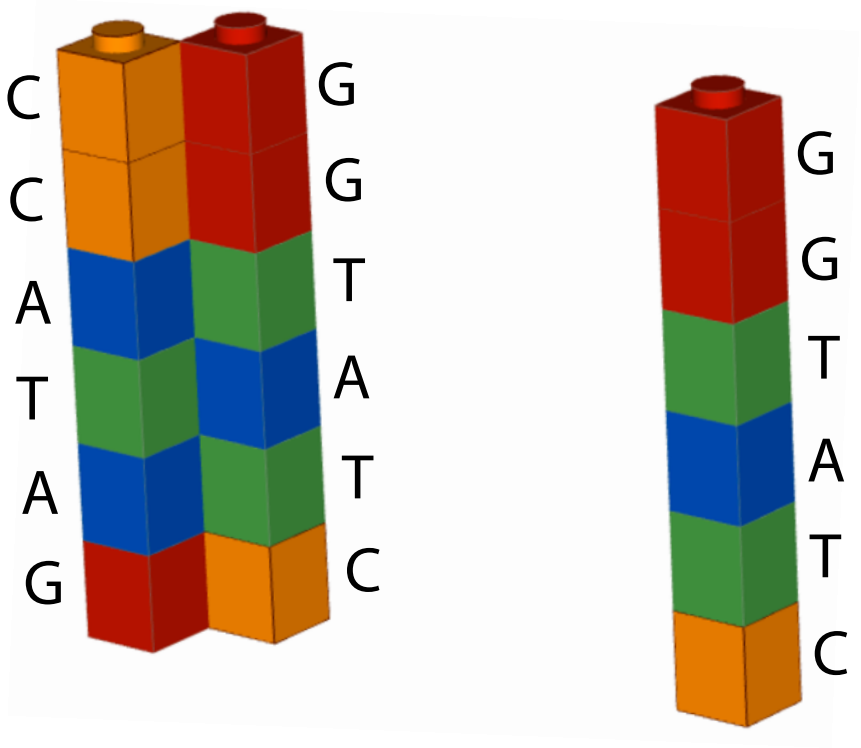

C
C
A
T
A
G
DNA polymerase
C
G
G
T
A
T
C

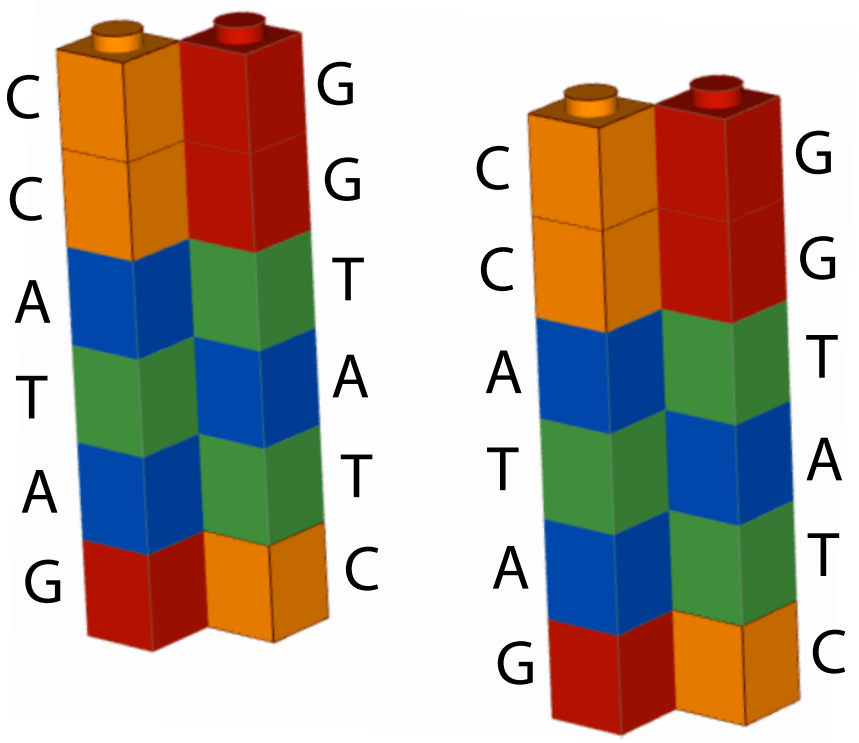

C
C
A
T
A
G
T
C
G
G
T
A
T
C

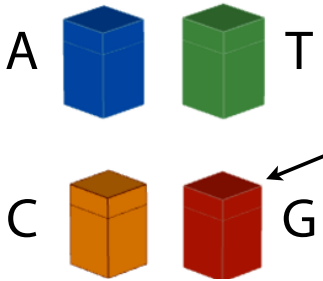

C
C
A
T
A
G
A
T
C
G
G
T
A
T
C


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

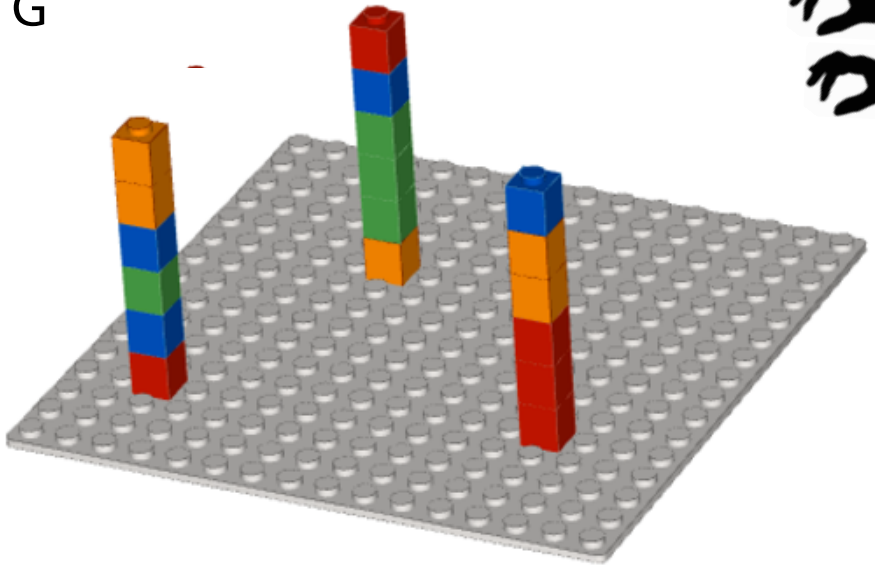

C
C
A
T
A
G
G
G
T
A
T
C
G
G
T
A
T
C

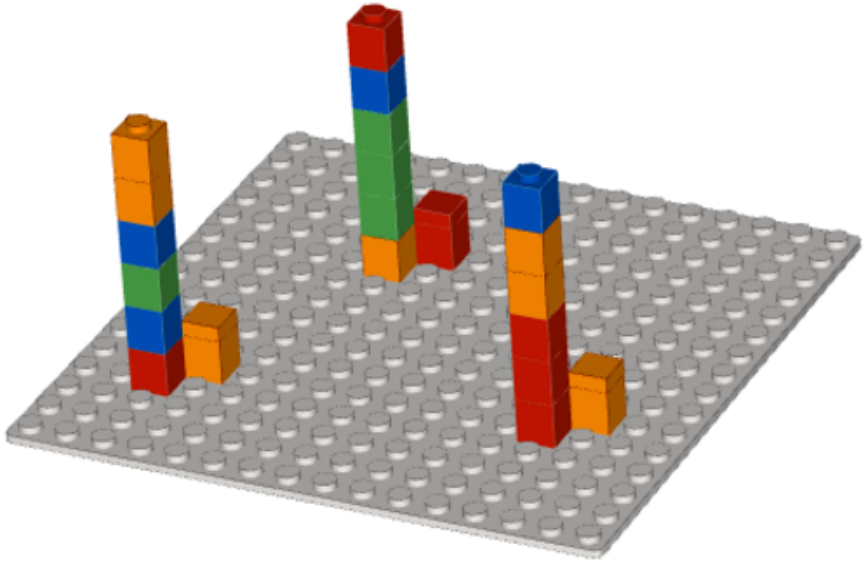

C
C
A
T
A
G
G
G
T
A
T
C
C
C
A
T
A
G
G
G
T
A
T
C

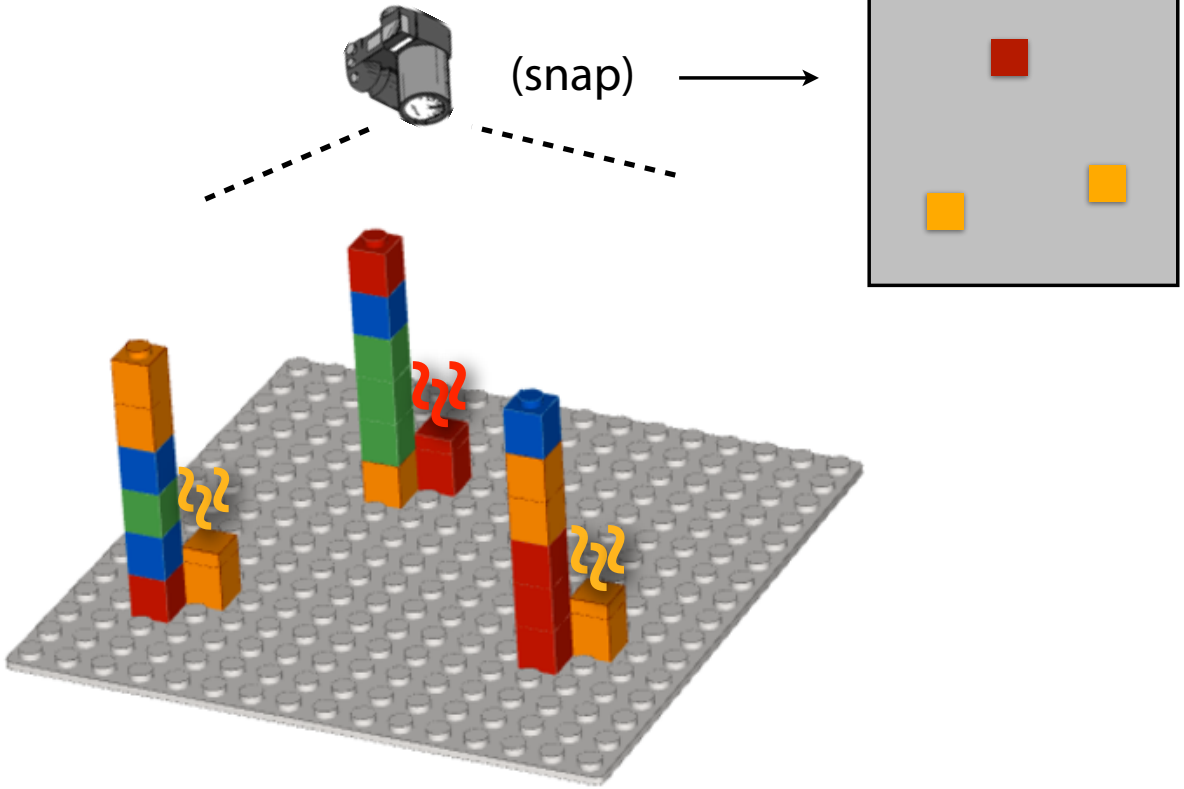


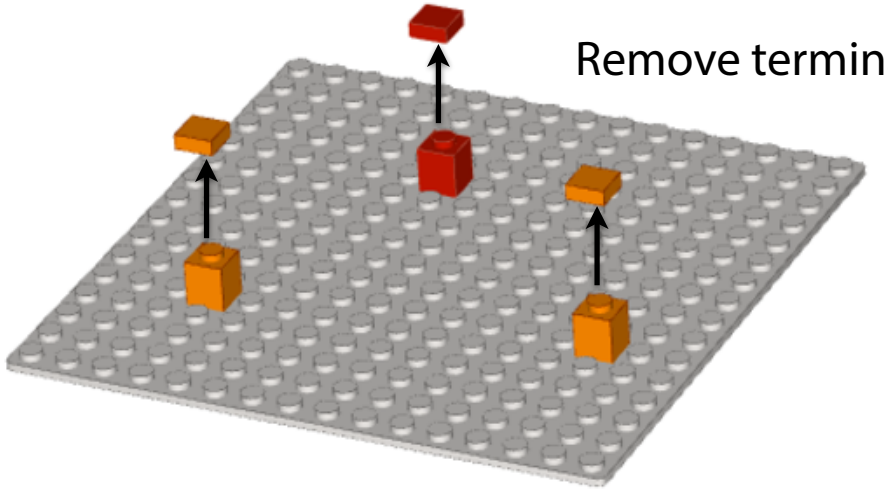
"Terminator"

DNA polymerase






(snap)

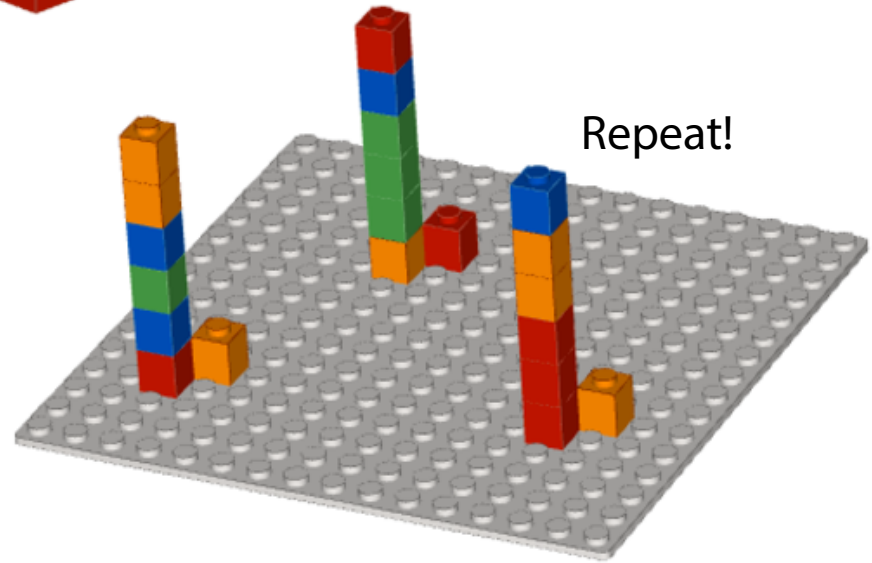


Remove terminators

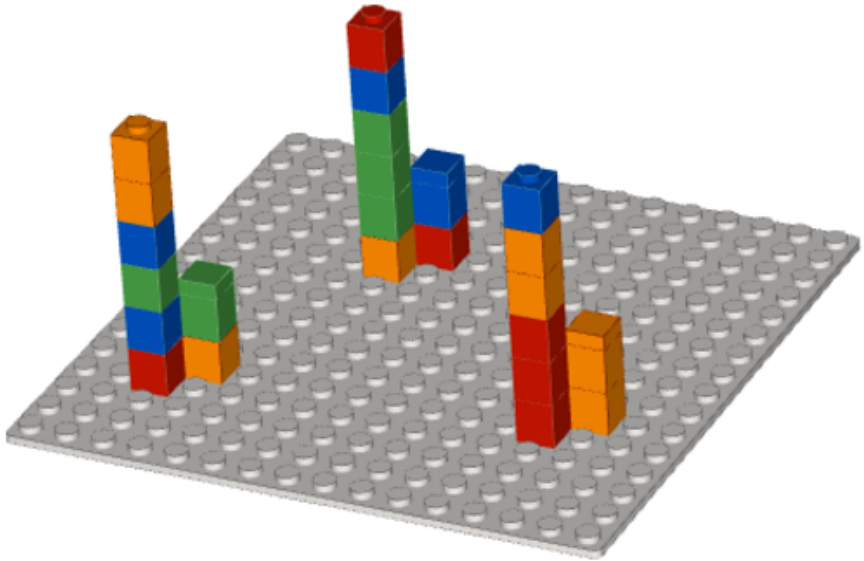
A   T

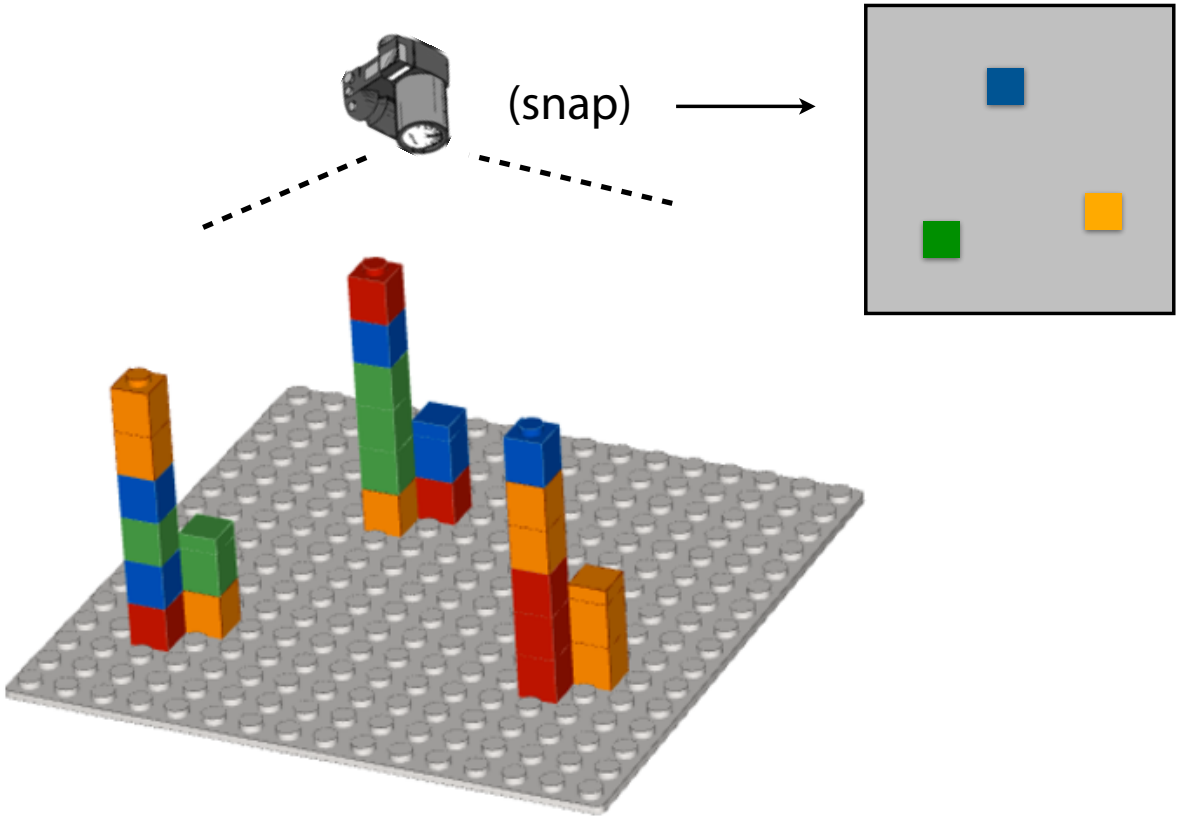
C   G

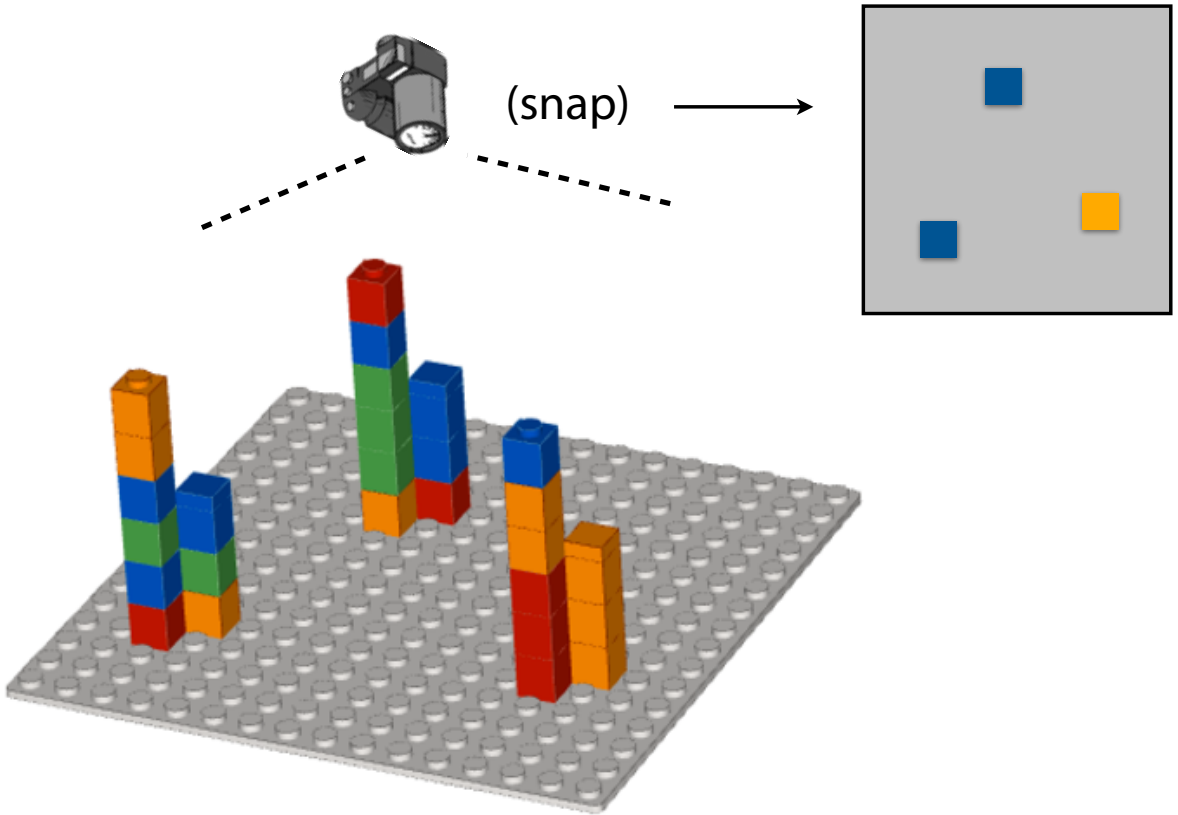
DNA polymerase

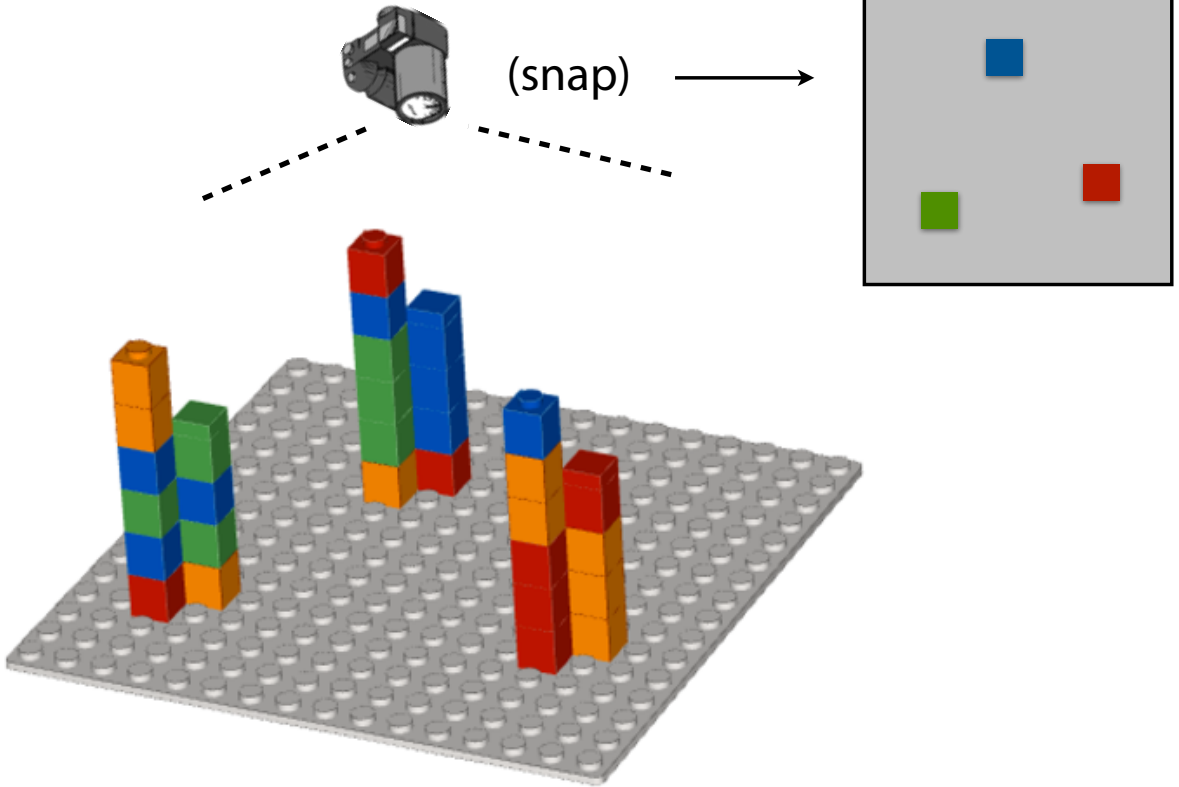


Repeat!

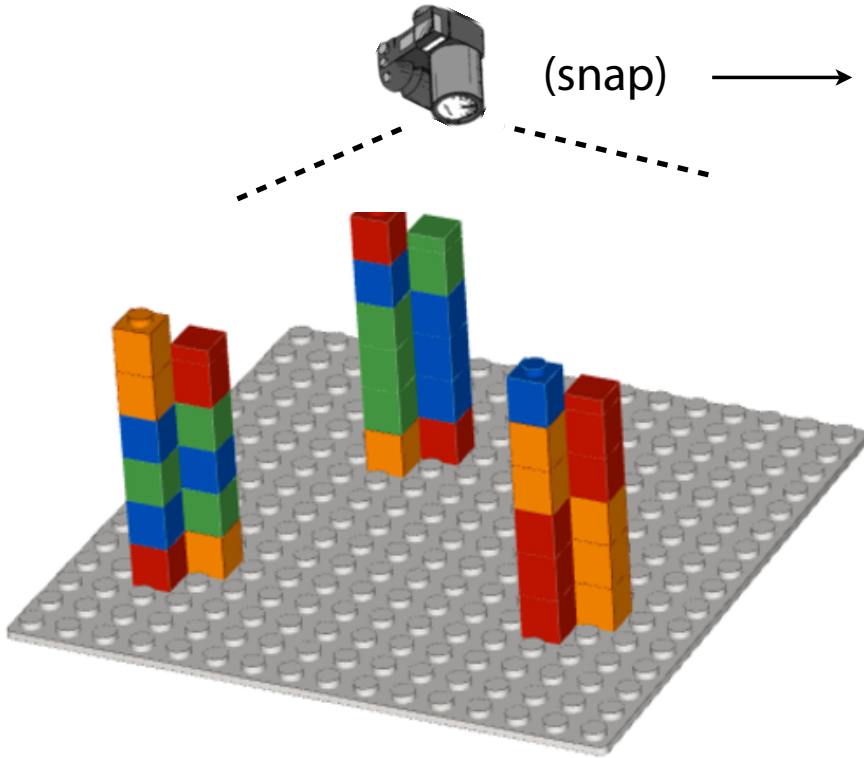



(snap)

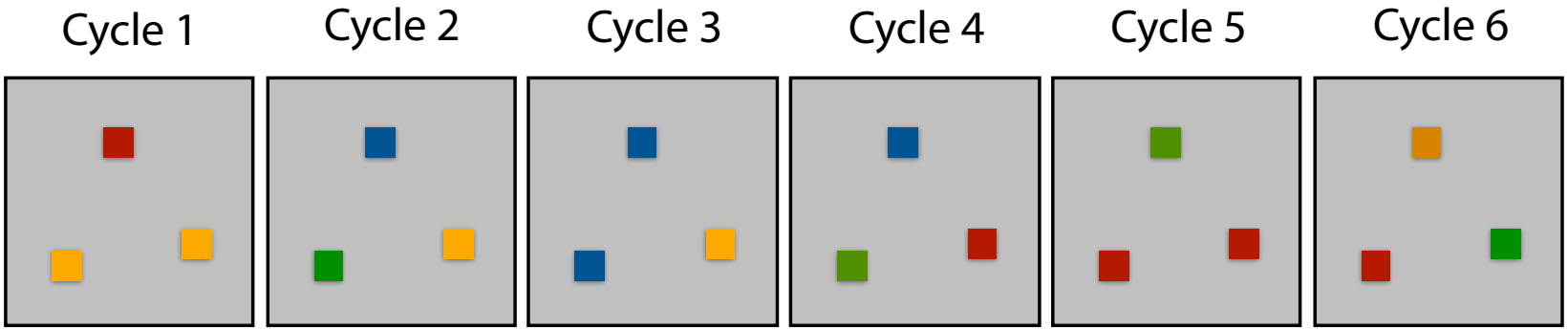

(snap)


(snap)

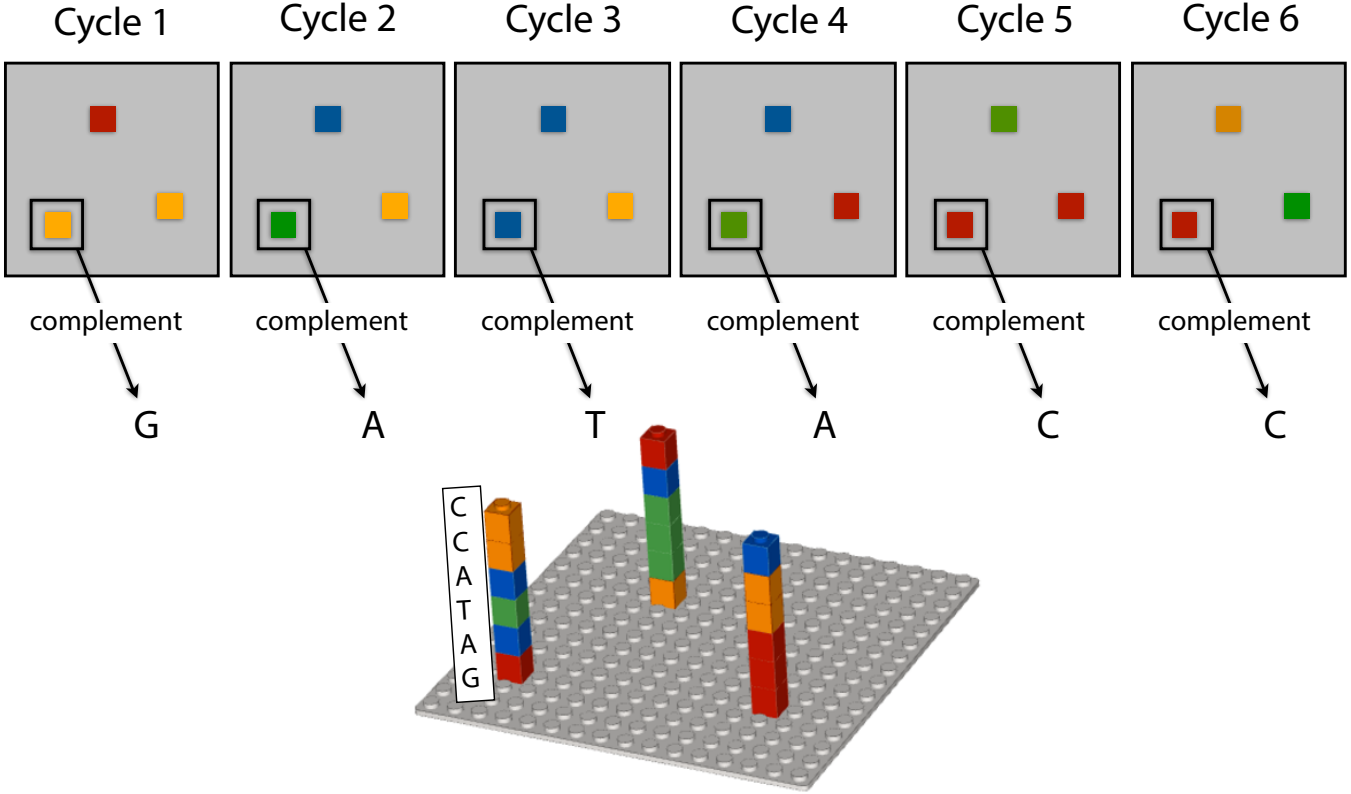
Example



Cycles



Sequencing by Synthesis

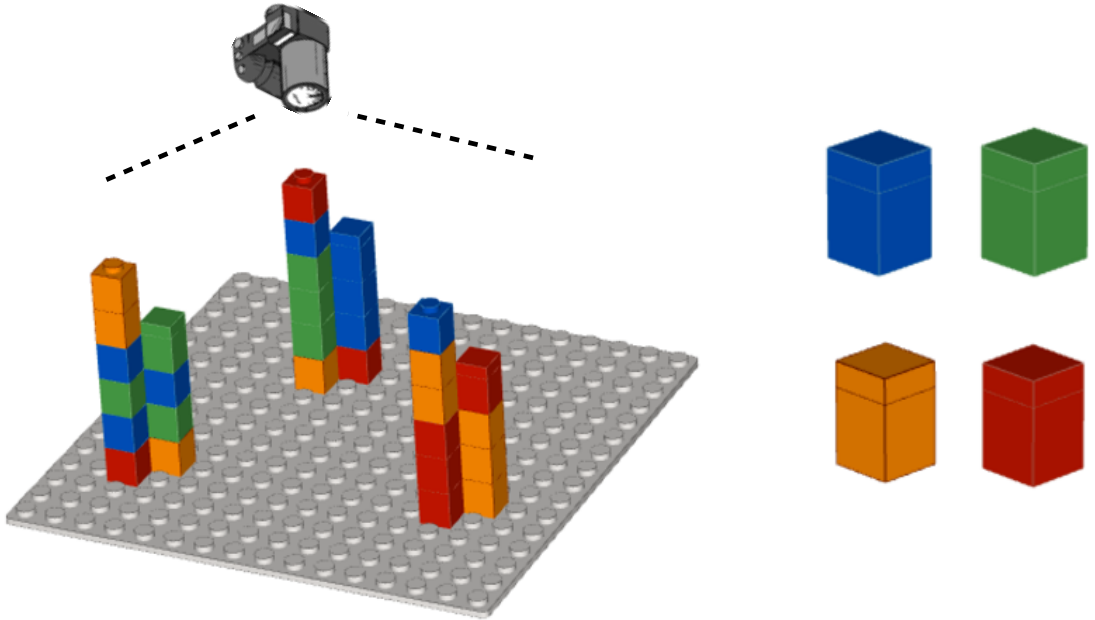


Exan Sequencing by synthesis

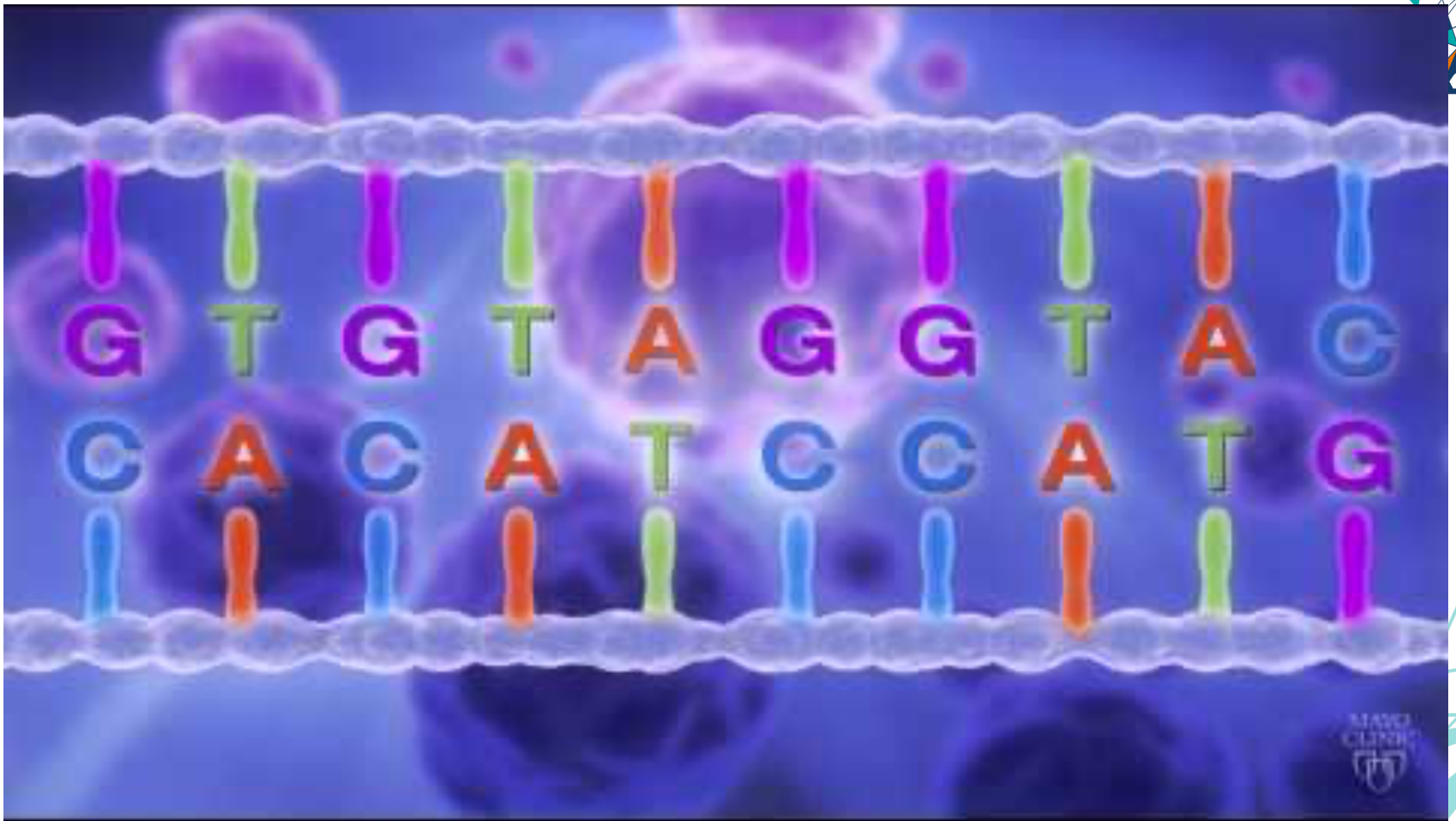
Billions of templates on a slide

Massively parallel: photograph captures all templates simultaneously

Terminators are “speed bumps,” keeping reactions in sync



From experiments to data

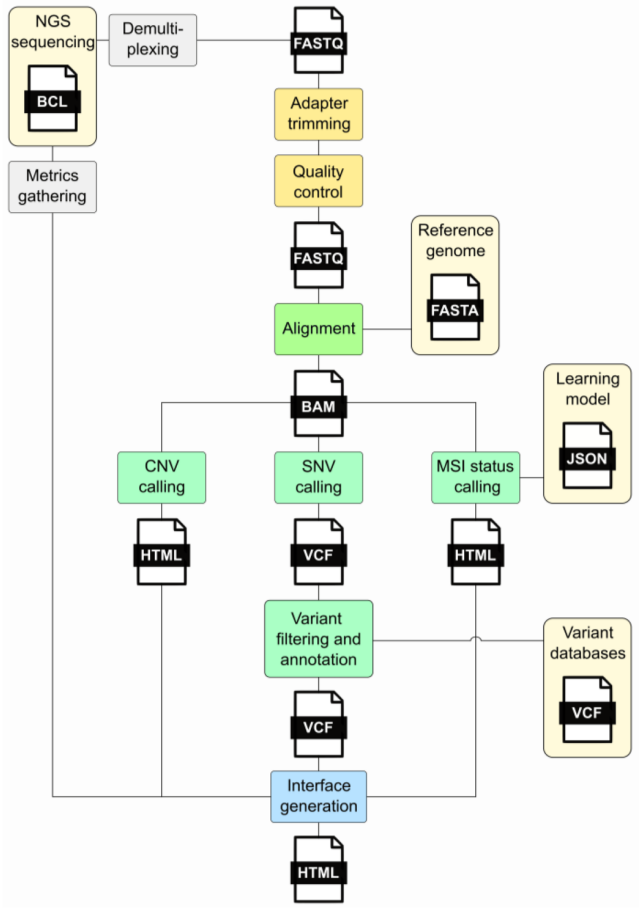


Sequncing Data File Formats

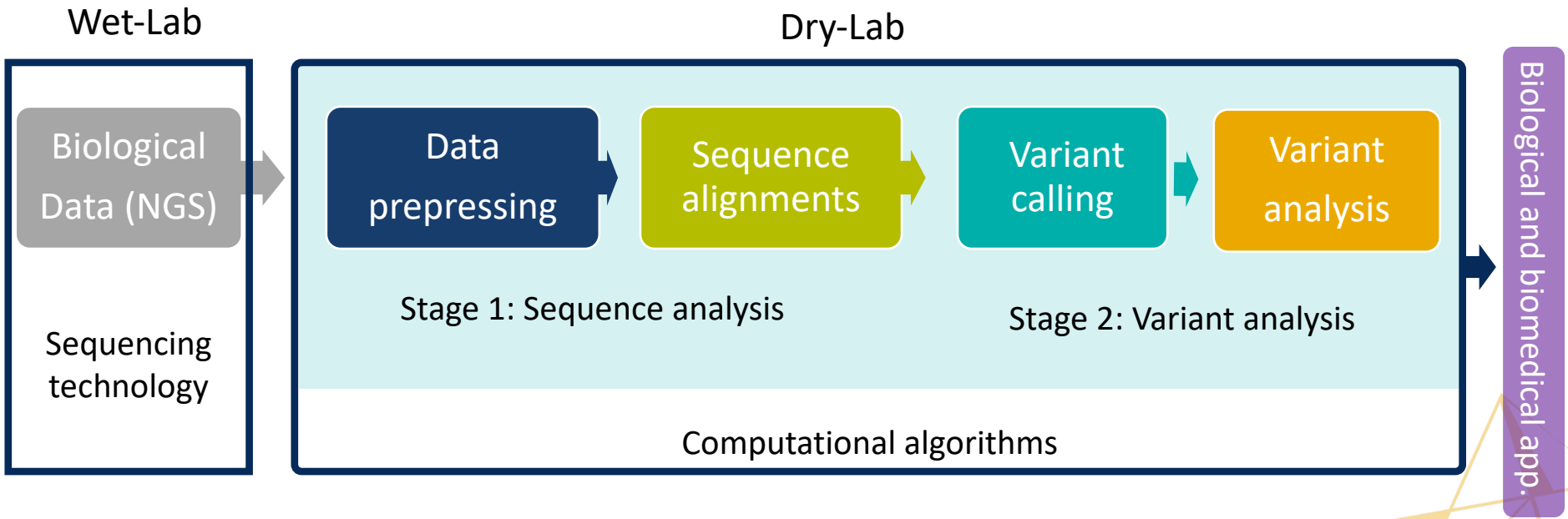
أكاديمية كاوست
KAUST ACADEMY

File Format	Description
FASTA	Represents nucleotide or peptide sequences without quality scores
FASTQ	Stores raw sequence data and quality scores for each nucleotide
BCL	Binary Base Call format; raw base call data from Illumina sequencing platforms
SAM	Verbose text-based format for storing sequence alignment data
BAM	Binary version of SAM, compressed and indexed for efficient data processing
CRAM	Compressed file format for storing sequencing reads, optimized for space
VCF	Describes gene sequence variations like SNPs and indels
BED	Specifies genomic regions, such as those identified in sequencing analyses
GFF/ GTF	Detailed annotation of genes and other genomic features
JSON	JavaScript Object Notation; a lightweight data-interchange format, often used for APIs and config files.
HTML	HyperText Markup Language; used for creating web pages and web applications

Sequncing Data Format Workflow



How to Get Started in Bioinformatics



NGS Bioinformatics workflow

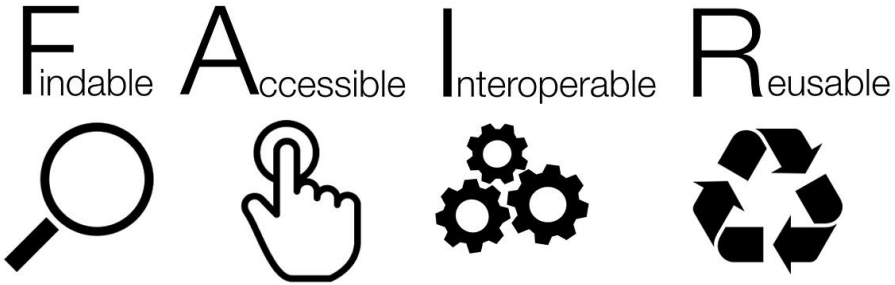
Galaxy PROJECT Bioinformatics Tool

أكاديمية كاوست
KAUST ACADEMY

Introduction of Galaxy Platform

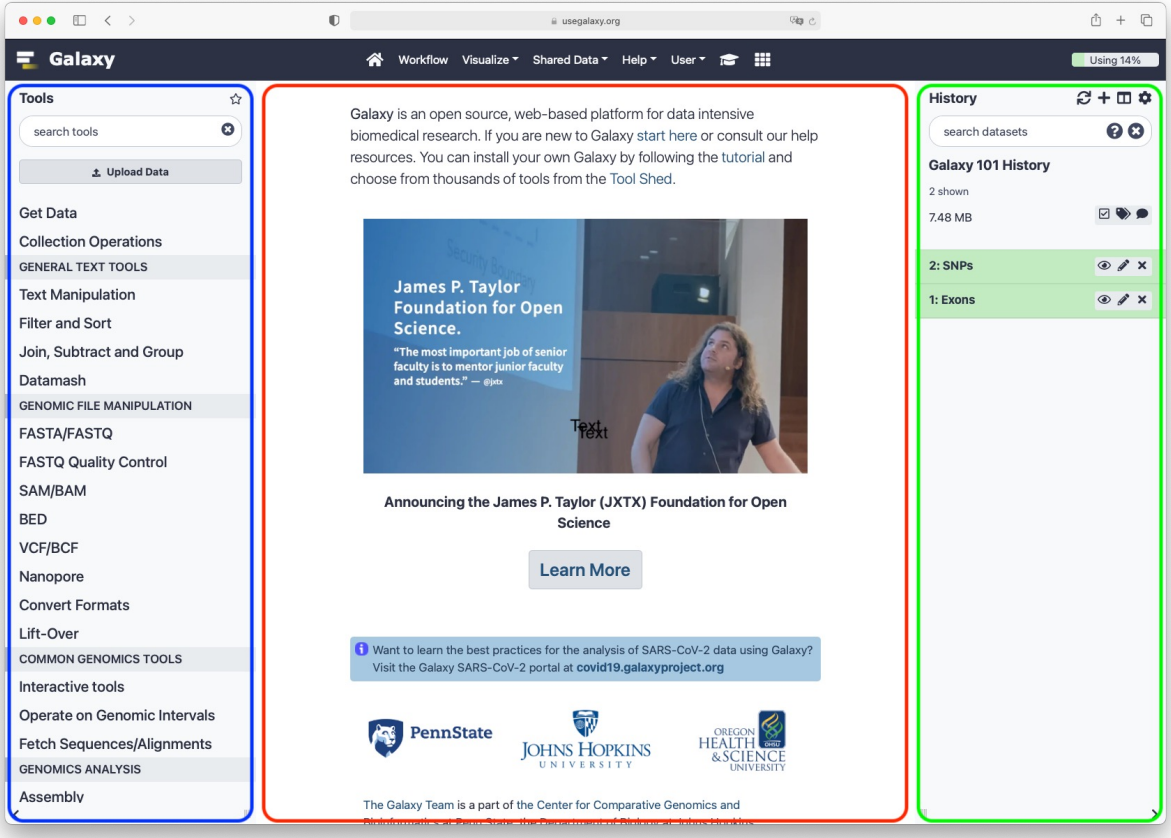
Galaxy is an **open-source** platform for **FAIR** **data analysis** that enables users to:

- Use **tools** from various domains through its graphical web interface.
- Run code in **interactive environments** along with other tools or workflows.
- **Manage data** by sharing and publishing results, workflows, and visualizations.
- **Ensure reproducibility** by capturing the necessary information to repeat and understand data analyses.



https://galaxyproject.org/

Discover galaxy platform!

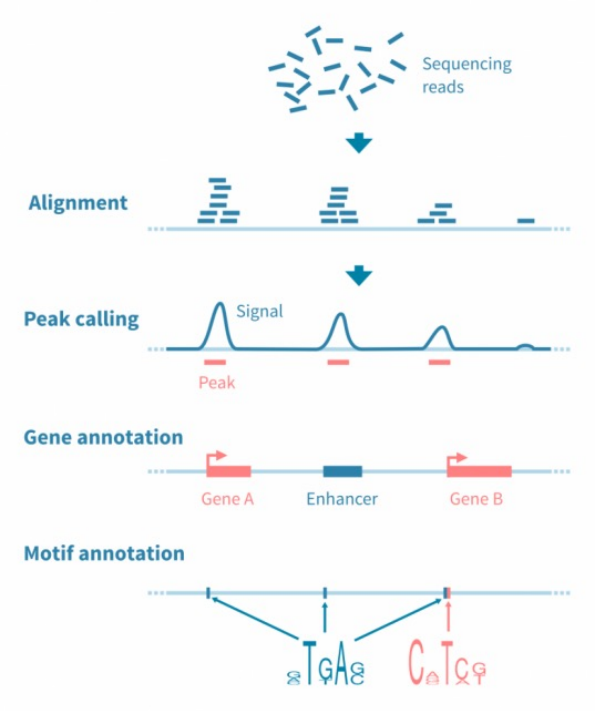


From peaks to genes

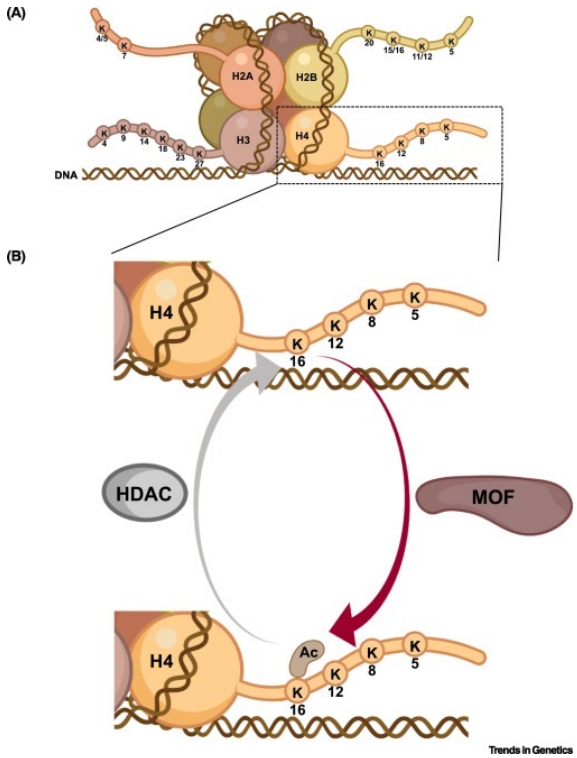
أكاديمية كاوست
KAUST ACADEMY

How to get from peak regions to a list of gene names?

- What is a peak?
- Peak regions in DNA, identified via sequencing techniques like ChIP-seq, are areas with high read enrichment, suggesting where proteins bind to the genome. These peaks are mapped to nearby genes to infer which genes may be regulated by these proteins.



GSE37268 Dataset for MOF Gene



- MOF(MOF histone acetyltransferase) is a protein that plays a crucial role in regulating the core transcriptional network of embryonic stem cells.
- Li et al. (2012) conducted a study to identify the target genes of MOF using ChIP-seq in mice.
- The raw data from the study is available through GEO, but the list of MOF target genes is not included in the paper's supplement or the GEO submission.

GSE37268 Dataset for MOF Gene

ChIP-seq analysis result

Chromosome Number	Start	End	Peak Length	Signal Strength	Score	Statistical Data
1	3660676	3661050	375	210	62.08762504	2.003293867

- **Chromosome Number:** The first column (e.g., '1') indicates the chromosome on which the region is located.
- **Start:** The second column (e.g., '3660676') shows the starting position of the peak region on the chromosome.
- **End:** The third column (e.g., '3661050') indicates the ending position of the peak region.
- **Peak Length:** The fourth column (e.g., '375') might represent the length of the peak region.
- **Signal Strength/Score:** The next columns, like '210' and '62.0876250438913', could represent the signal location or a score that quantifies the enrichment.
- **Statistical Data:** The last column (e.g., '-2.003') is statistical data related to the peak, such as a p-value or fold-change.

Hands-on and Practical Part

Galaxy PROJECT

GitHub

Part 1: Introduction of galaxy platform

- Collecting data
- Gene Identification

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Understanding the Digital Frontier in Biomedicine

Part 3: Quality Control

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Expected Errors

- Samples collection
- Nucleic acid extraction
- Quality Control

- Adaptor ligation/barcoding
- Size selection
- Amplification/purification
- Quality control

- Sequencing
- Data analysis
  - Base calling
  - Read alignment
  - Variant calling
  - Variant annotation

Expected Errors

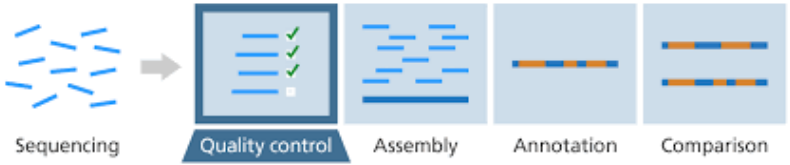
أكاديمية كاوست
KAUST ACADEMY

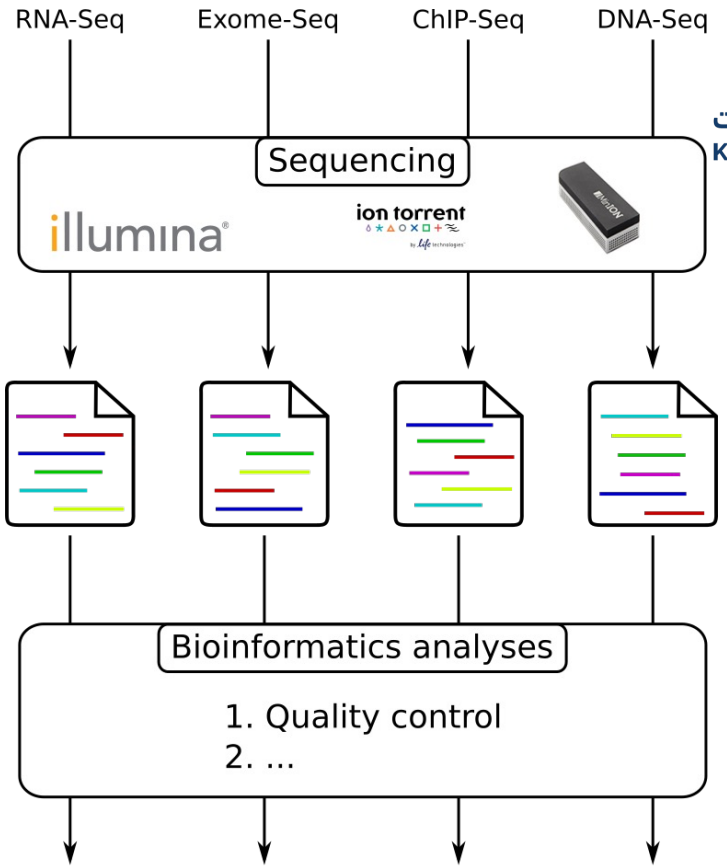
- **Sample Preparation:** Contamination with other DNA, RNA degradation, or insufficient quantity of starting material can lead to errors.
- **Library Construction:** Inaccurate size selection, adapter ligation problems, or PCR amplification errors can introduce biases.
- **Sequencing Process:** Issues like sequencing instrument calibration, cluster generation failures on flow cells, or phasing/pre-phasing problems can cause errors.
- **Base Calling:** Incorrect interpretation of fluorescent signals can lead to incorrect nucleotide identification.
- **Read Quality:** Low-quality reads can result from sequencing reaction issues or signal decay over time, especially at the end of reads.
- **Alignment:** Incorrectly mapped reads due to repetitive regions or sequence similarity can misrepresent the genomic location.
- **Variant Calling:** Misidentification of true genetic variants versus sequencing errors.

Quality Control

The process of evaluating the quantity, purity, and intactness of the genomic DNA

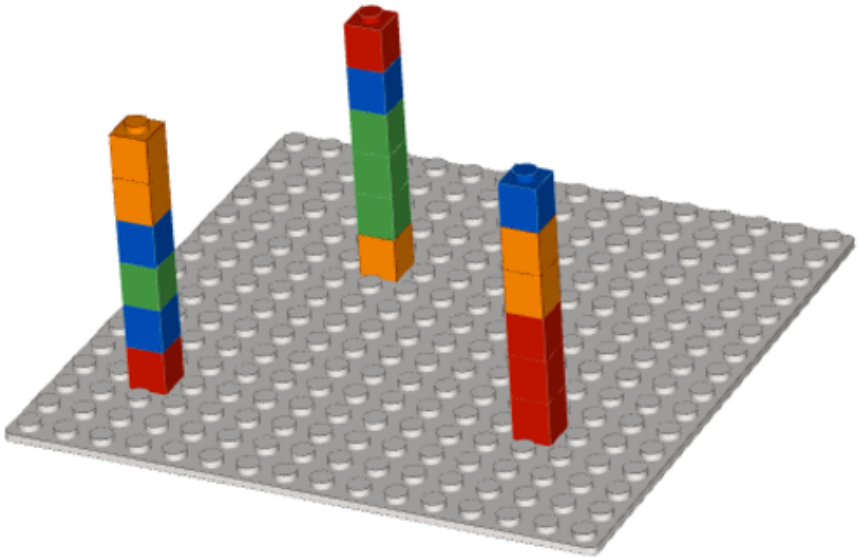
- Pre-sequencing Errors
- Sequencing Errors
- Post-sequencing Errors



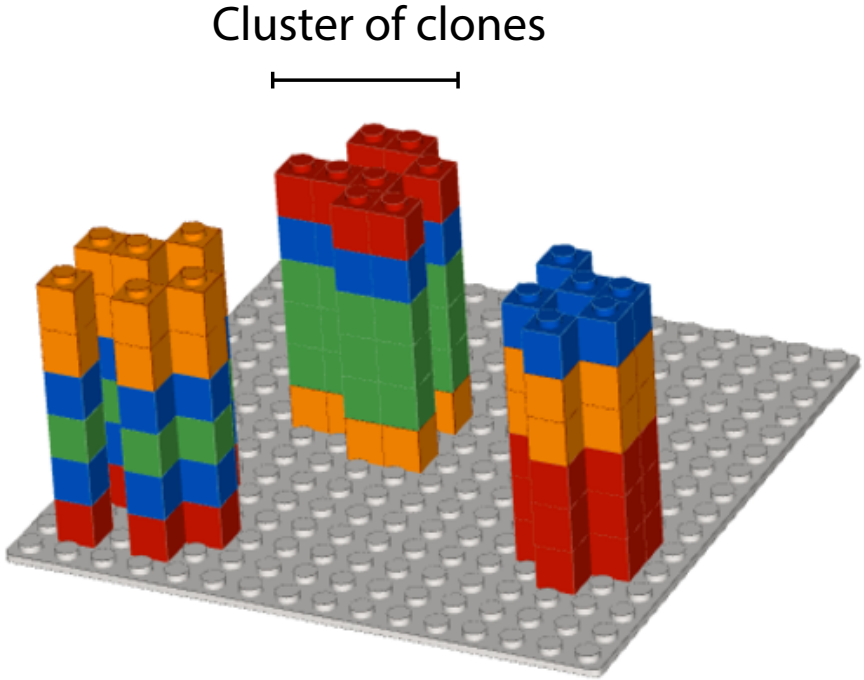


Quality control = First step of the bioinformatics analyses

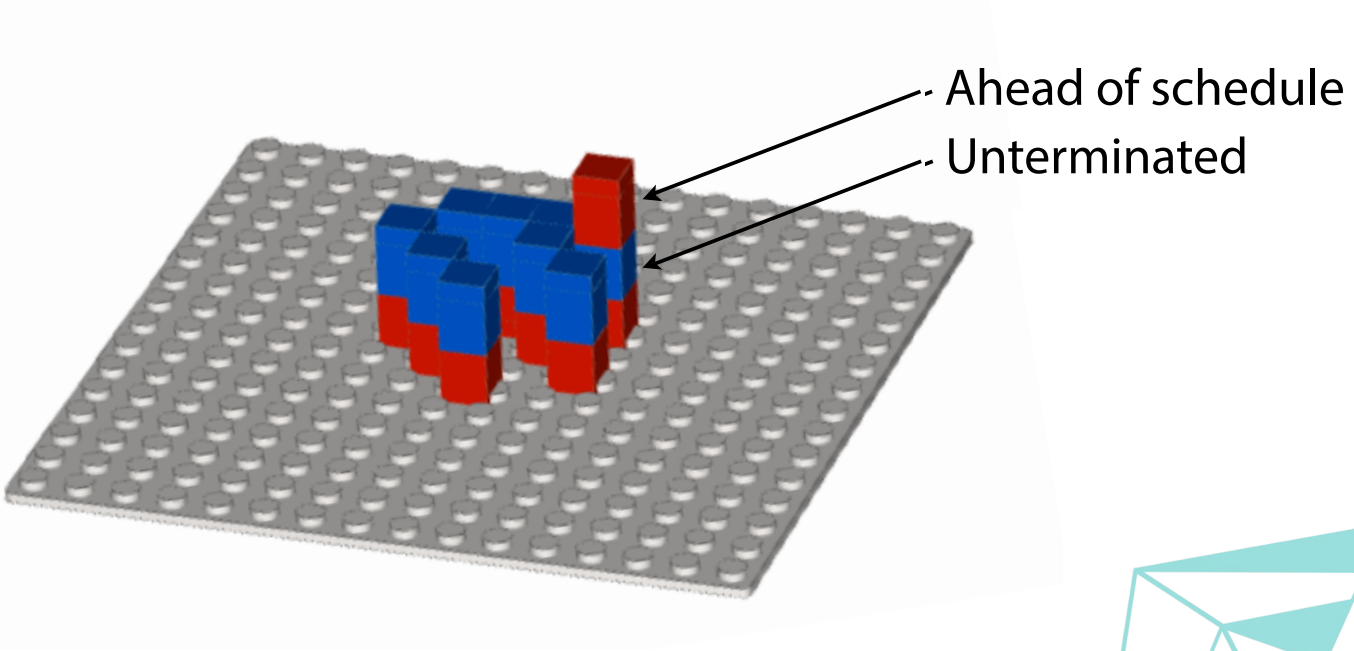
Example: Sequencing Errors



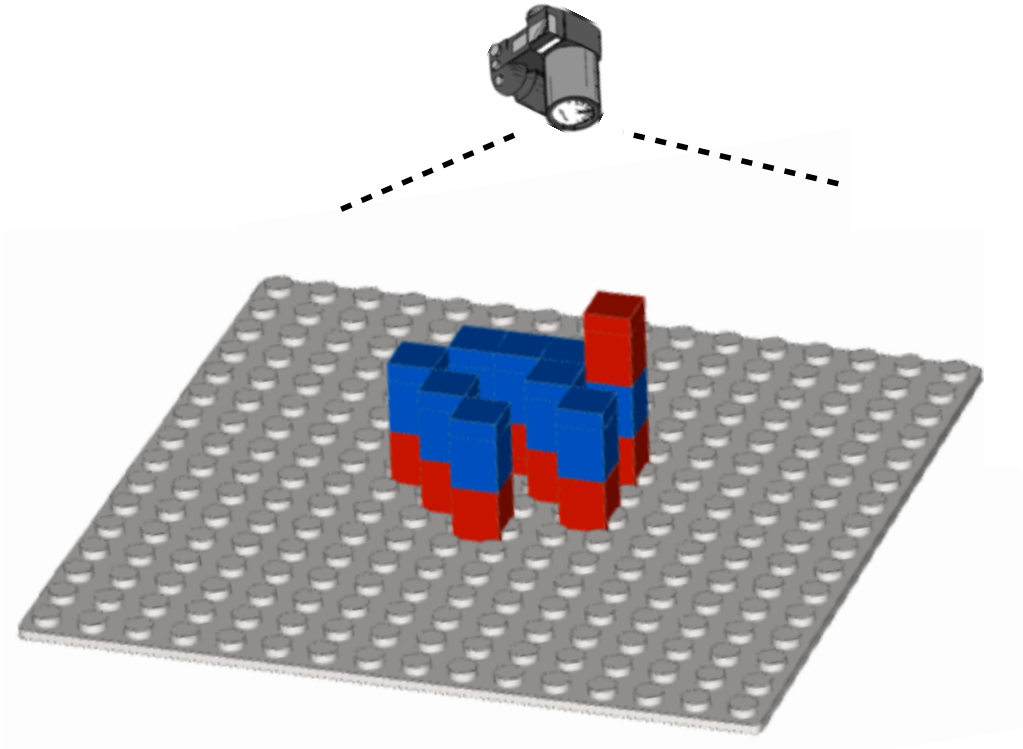
Example: Sequencing Errors

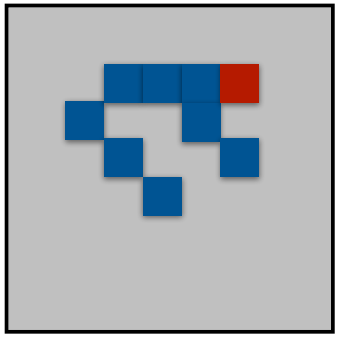


Example: Sequencing Errors

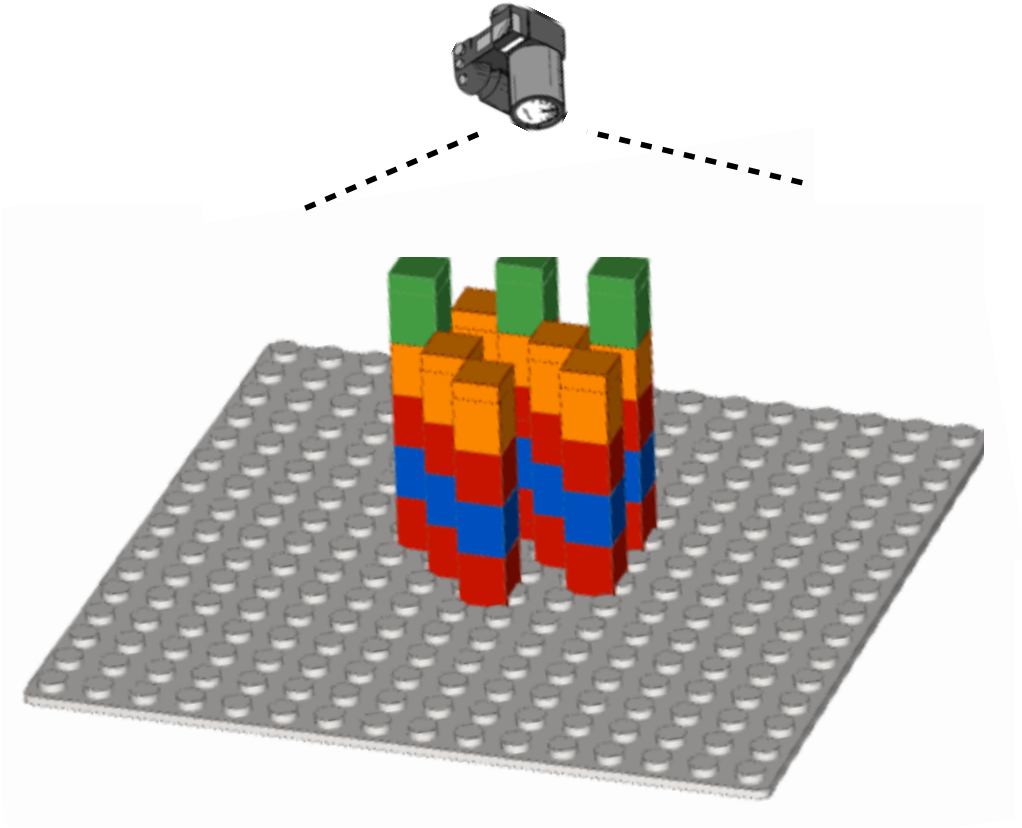


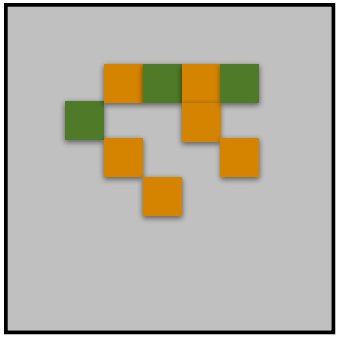
Example: Sequencing Errors



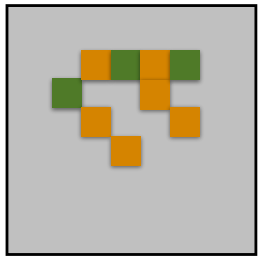


Example: Sequencing Errors





Example: Sequencing Errors



Call: orange (C)

Estimate *p*, probability incorrect:

non-orange light / total light

p = 3 green / 9 total = 1/3

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

Quality Score Calculation

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

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

Explor FASTA File

أكاديمية كاوست
KAUST ACADEMY

- Sequences: FASTA

>Copy>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX

Explor FASTQ File

أكاديمية كاوست
KAUST ACADEMY

- Sequences: FASTAQ

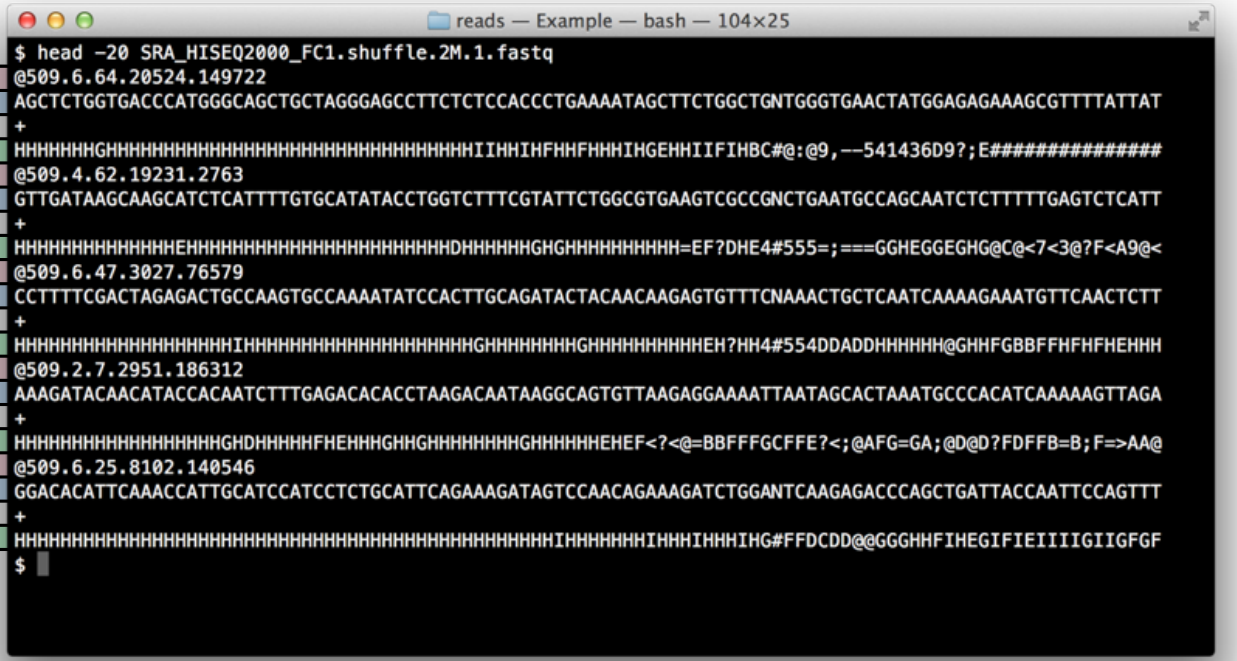
>Copy>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ

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



Base qualities

Bases and qualities line up:

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

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

ASCII Code

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	ˇ
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	fl	255	

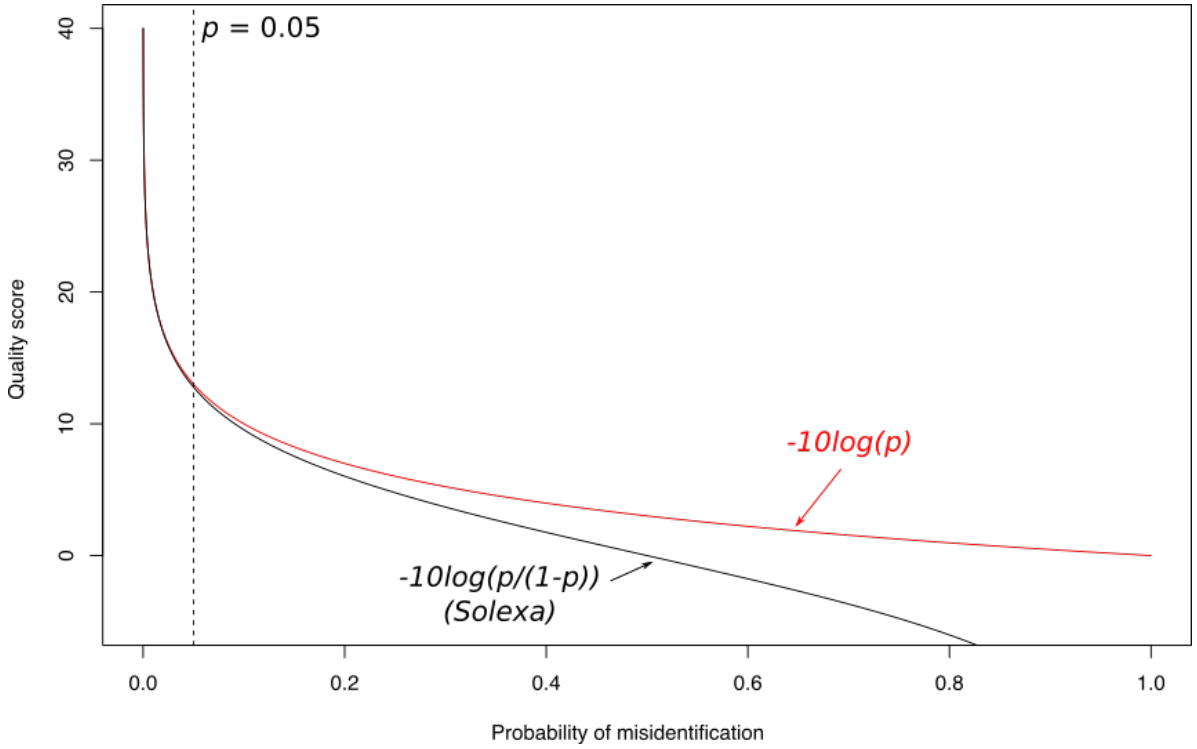
Quality Scores

أكاديمية كاوست
KAUST ACADEMY

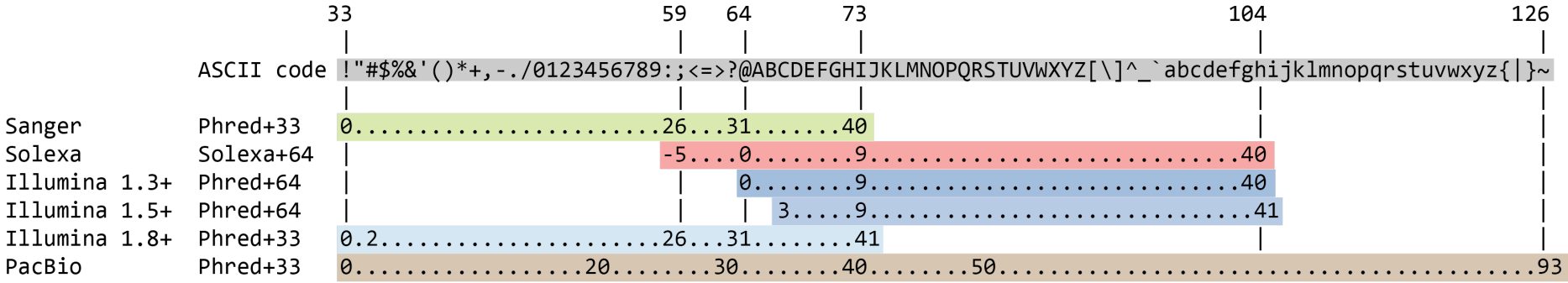
Measure of the quality of the identification of the nucleobases generated by automated DNA sequencing

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality Score



Quality score encoding



Identifying Potential Quality Issues

أكاديمية كاوست
KAUST ACADEMY

- Per-base quality
- Pre-sequence quality
- Per-tile quality
- Per-base sequence content
- Per-sequence GC content

- FastQC tool : a versatile tool for short and long reads quality control

FastQC Report

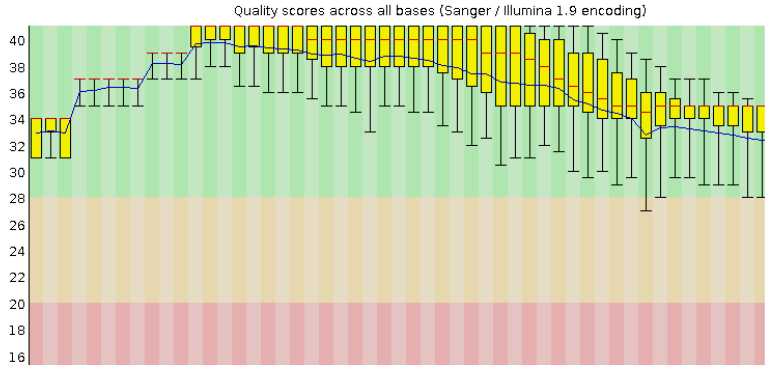
Mon 6 May 2019
MCL1-DK.gz

Summary

Basic Statistics

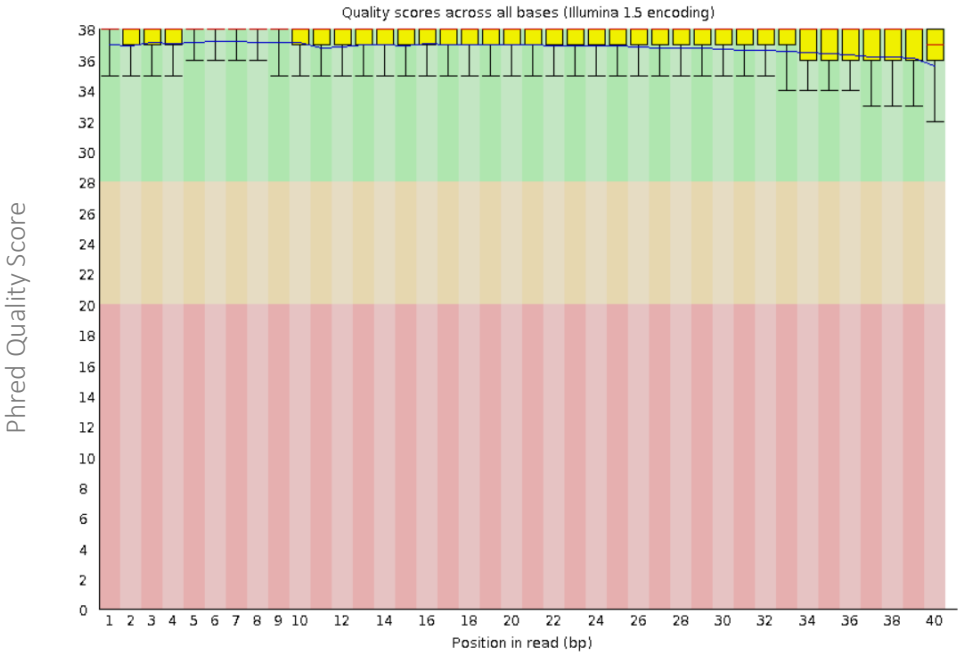
Measure	Value
Filename	MCL1-DK.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	100
%GC	50

Per base sequence quality

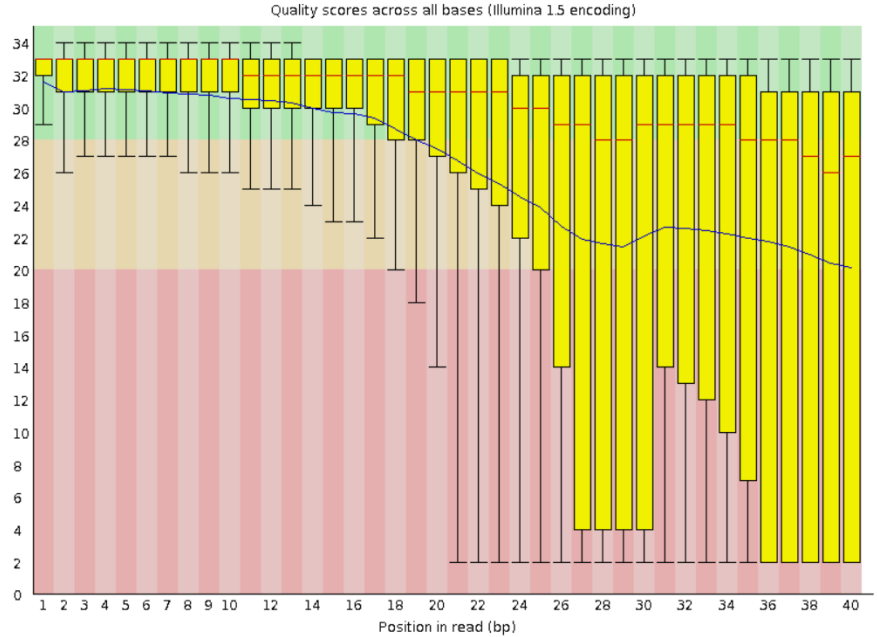


Per-base Quality

Per base sequence quality

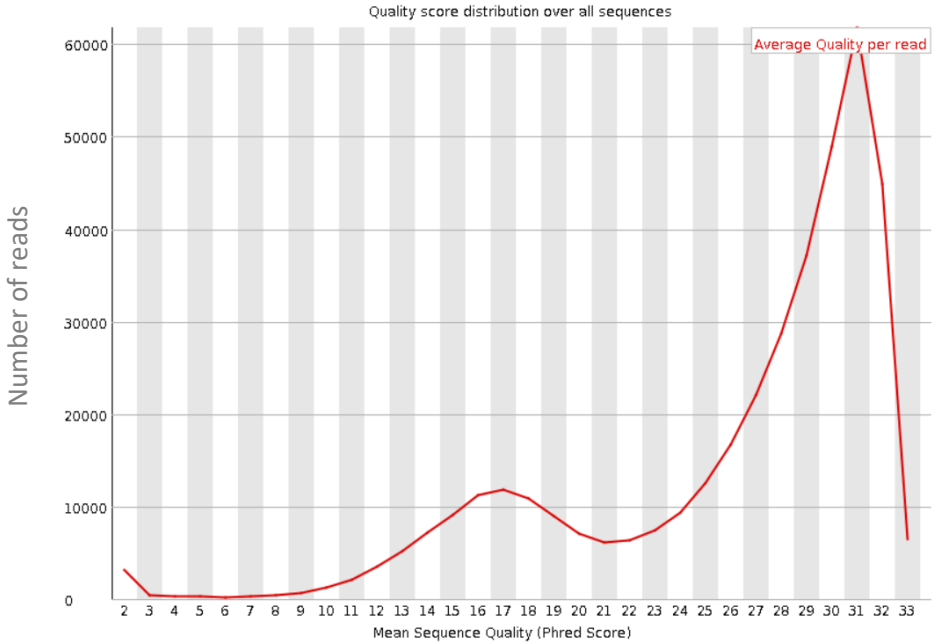


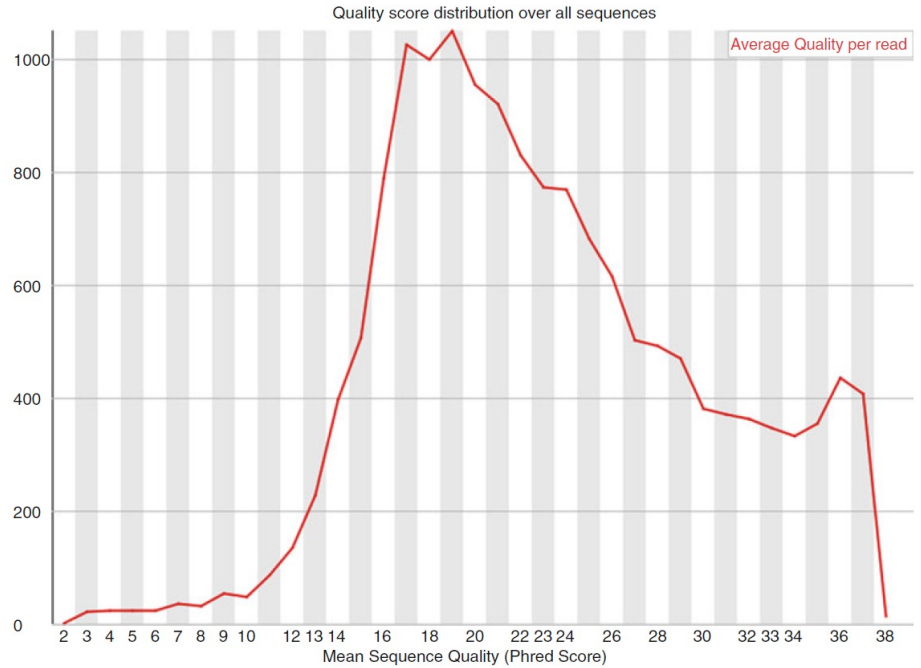
Per base sequence quality



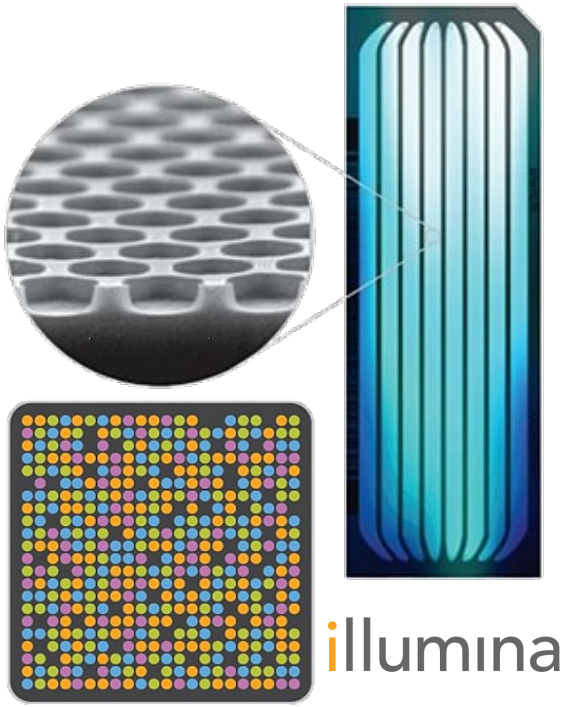
Per-sequence Quality

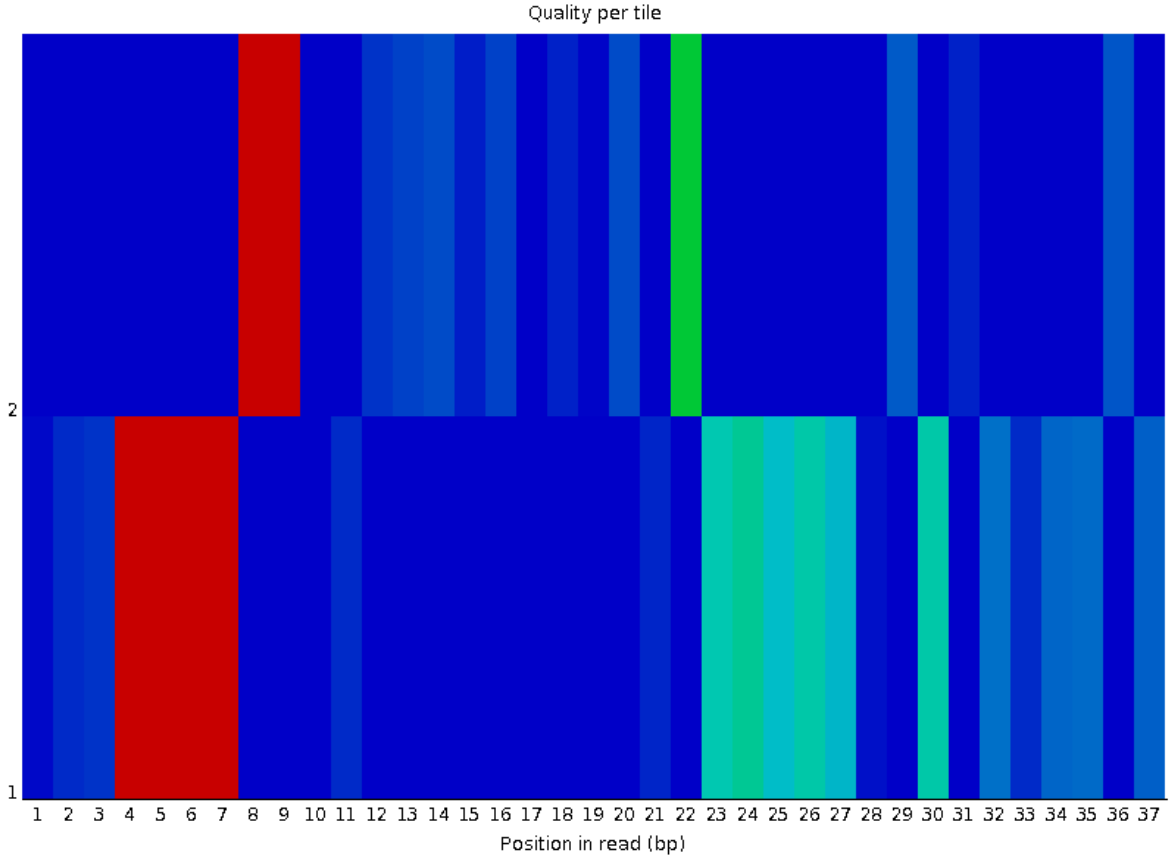
Per sequence quality scores





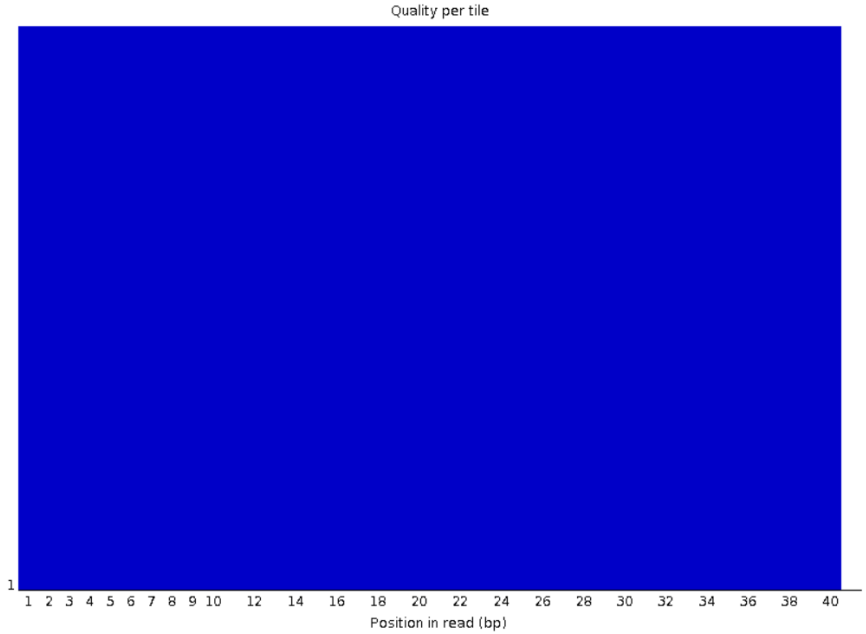
Per-tile Quality



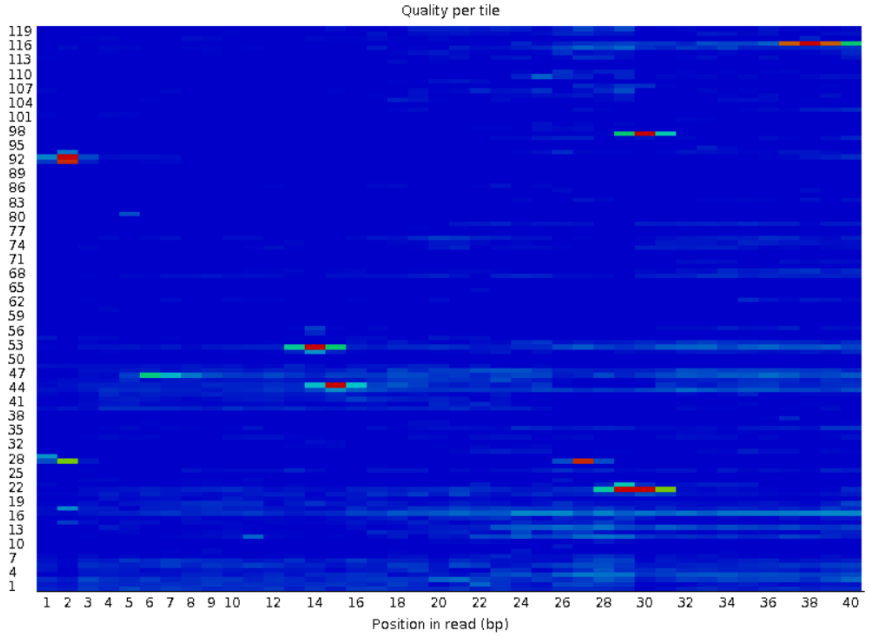


Per-tile Quality

Per tile sequence quality

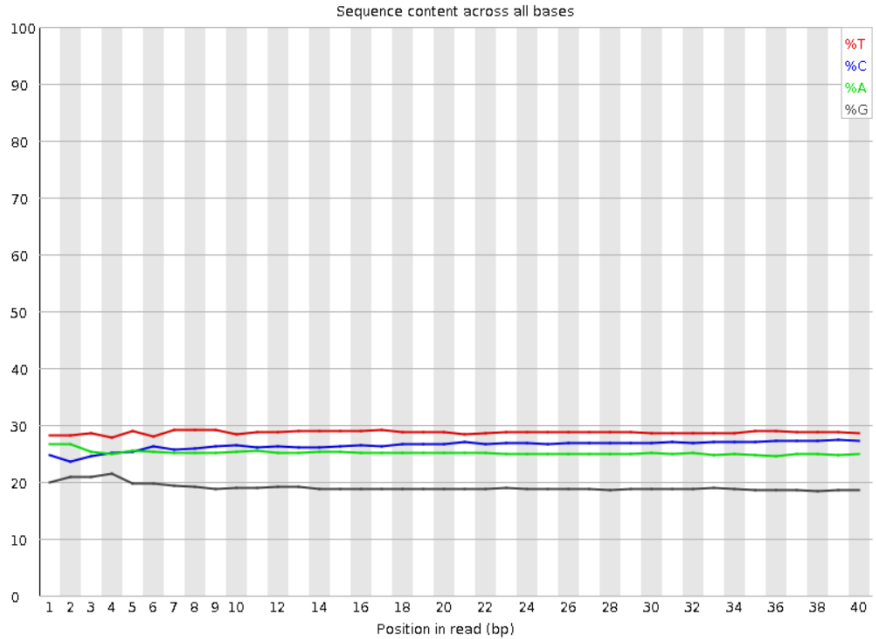


Per tile sequence quality

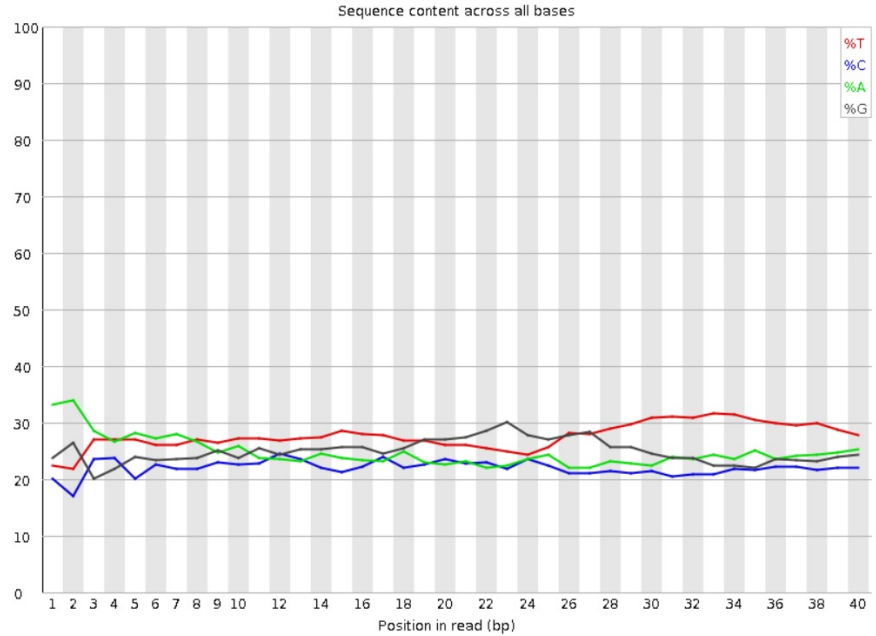


Per-base Sequence Content

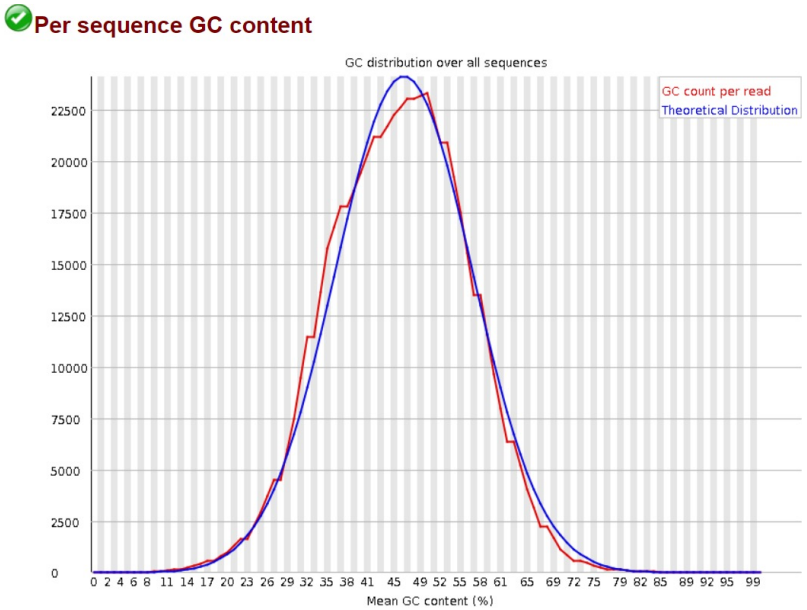
Per base sequence content

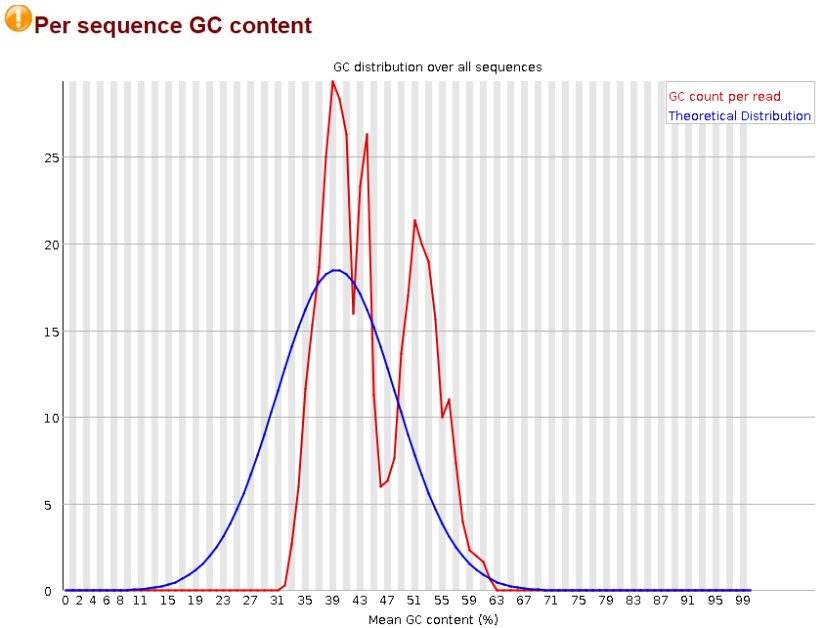


Per base sequence content

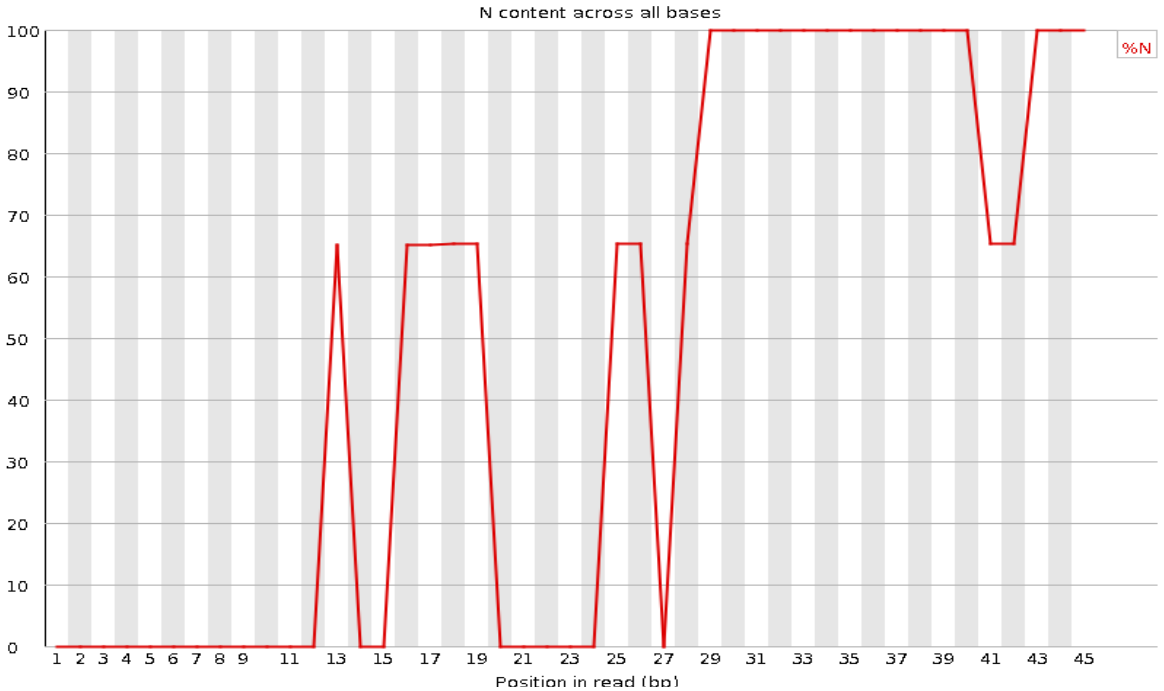


Per-sequence GC Content





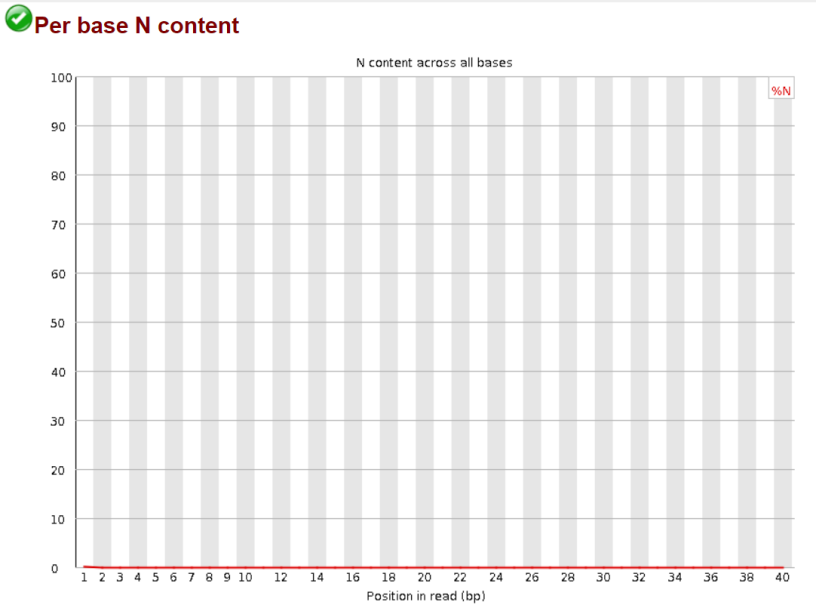
Per-base N Content

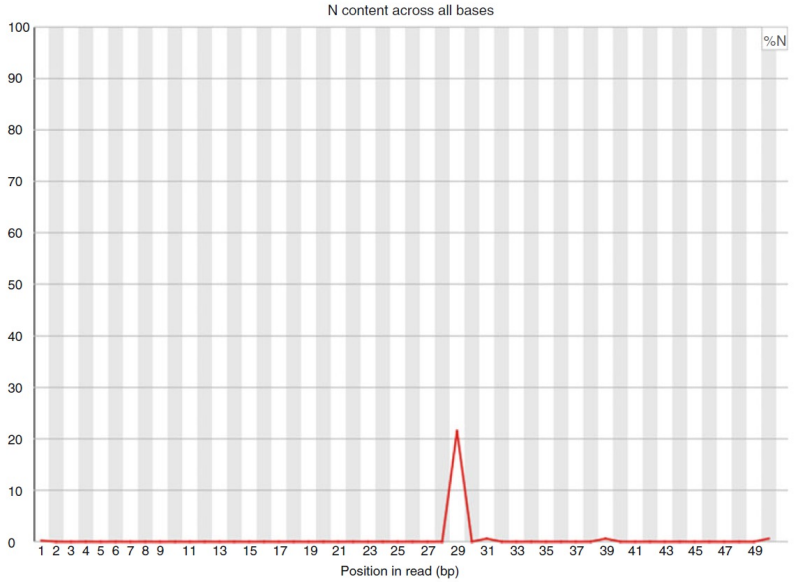


Name @ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1

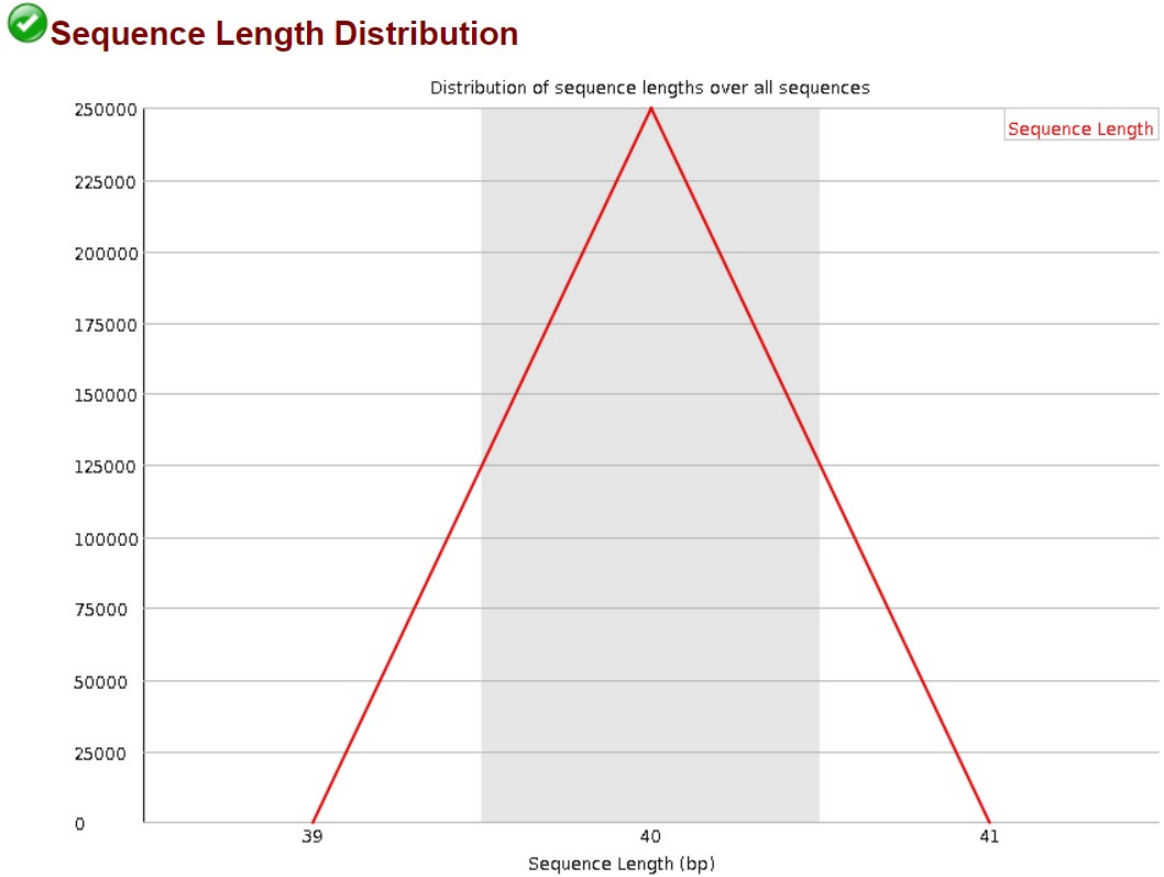
Sequence ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT

Per-base N content



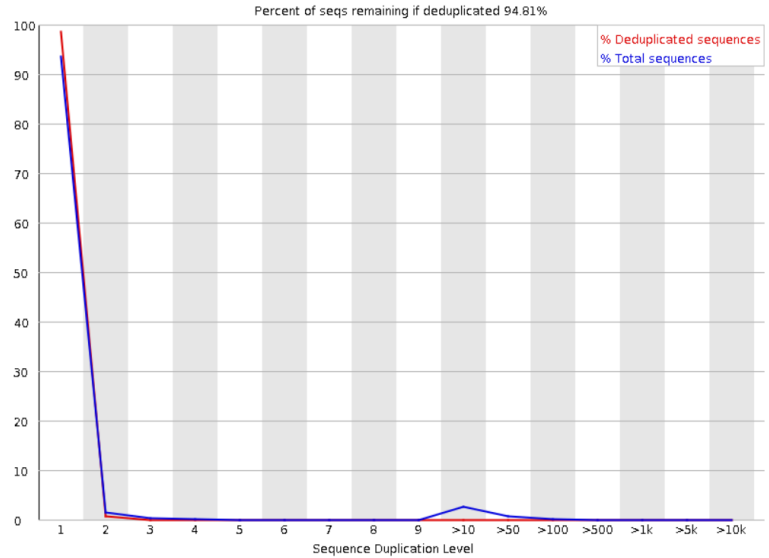


Sequence Length Distribution

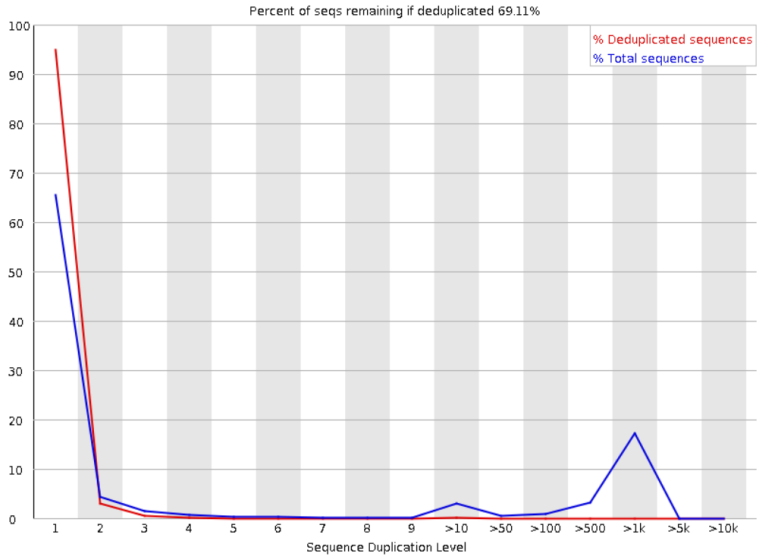


Per-sequence Duplication Quality

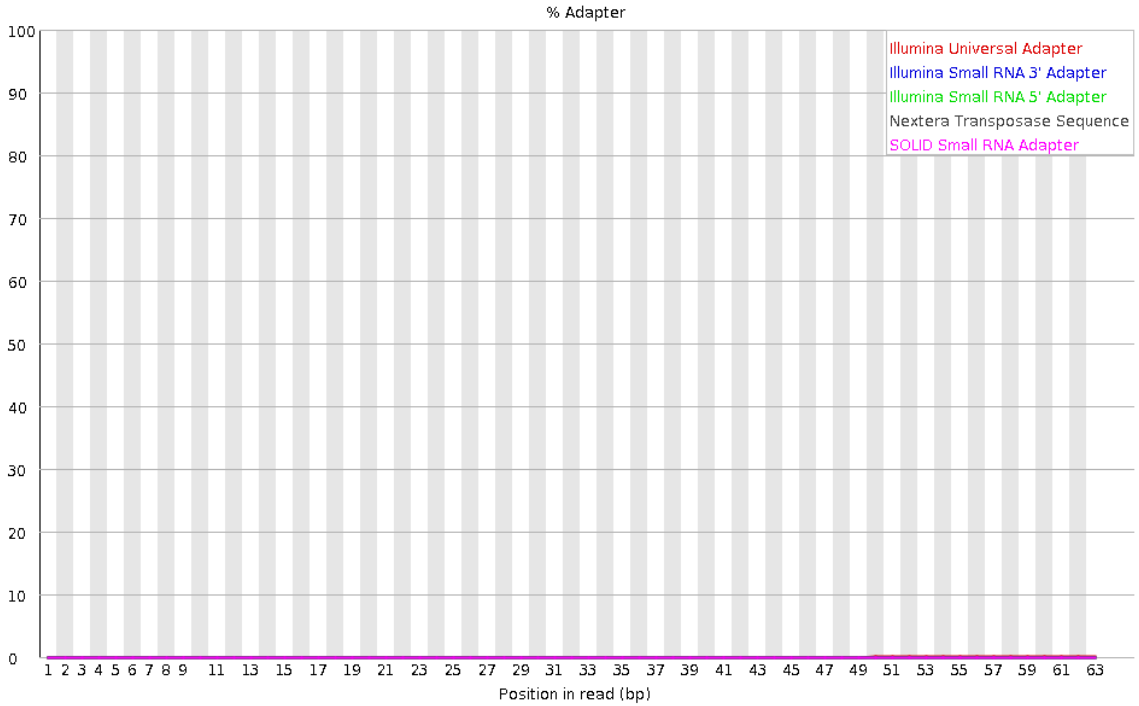
Sequence Duplication Levels



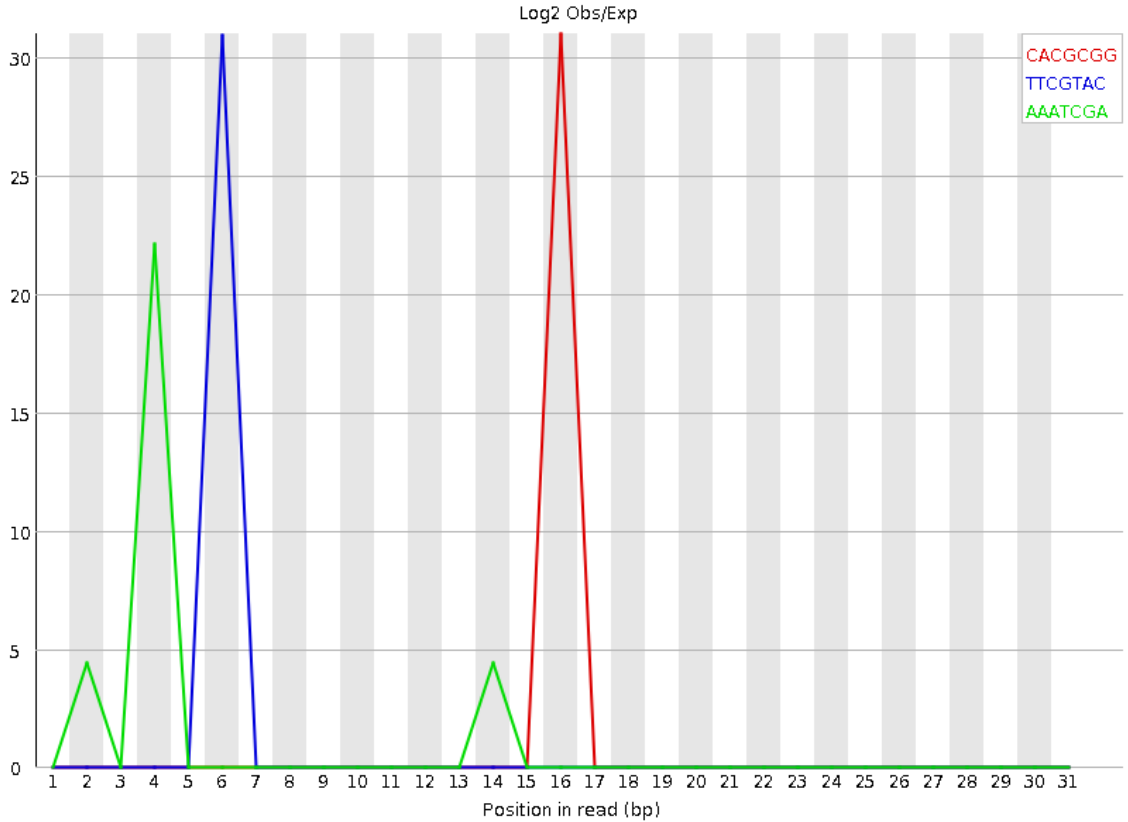
Sequence Duplication Levels



Tag sequences: Adapter Contamination



Tag Sequences: K-mer Content



Key Points

أكاديمية كاوست
KAUST ACADEMY

- Run quality control on every sequencing dataset before any other analyses
- Choose QC parameters carefully
- Re-run FastQC to check the impact of the quality control

Hands-on and Practical Part

Galaxy PROJECT

GitHub

Part 2: QC

- Inspect a raw sequence file
- Assess quality with FASTQE
- Assess quality with FastQC

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

أكاديمية كاوست
KAUST ACADEMY

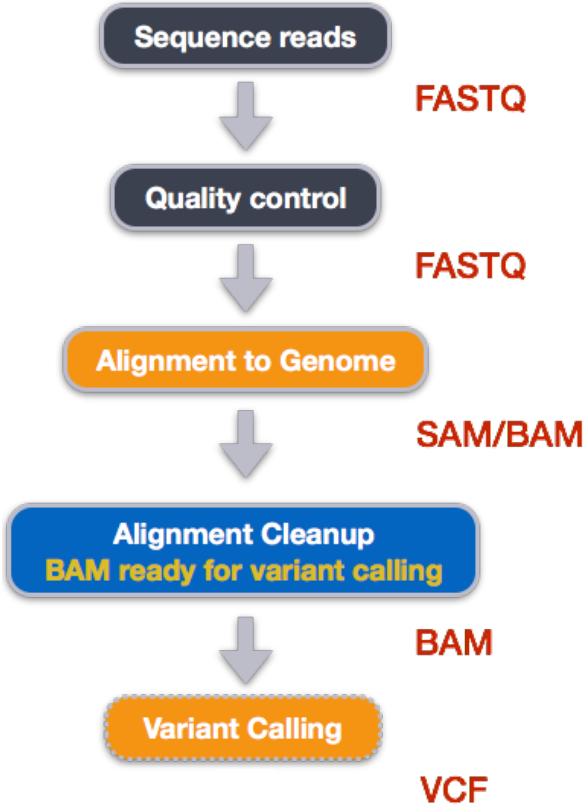
Introduction to Bioinformatics

Understanding the Digital Frontier in Biomedicine

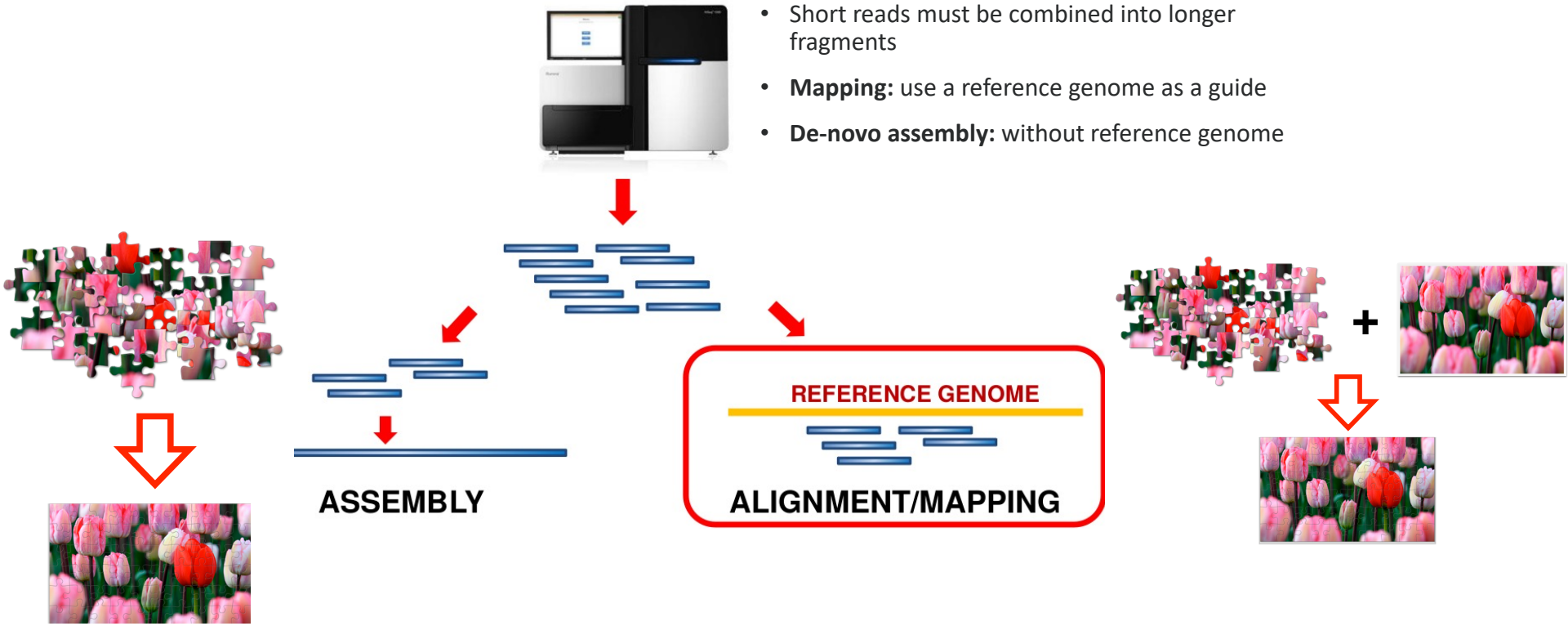
Part 3: Reference mapping

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Example of NGS Workflow



What is Mapping or Alignment in NGS?



- Short reads must be combined into longer fragments
- **Mapping:** use a reference genome as a guide
- **De-novo assembly:** without reference genome

Sequence Alignment

أكاديمية كاوست
KAUST ACADEMY

- Determine position of short read on the reference genome

```
Reference: . . . A A C G C C T T . . .

Read:              A G G G G C C T T
```

Sequence Alignment

أكاديمية كاوست
KAUST ACADEMY

- Determine position of short read on the reference

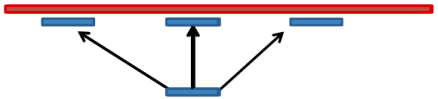
```
Reference: . . . A A - C G C C T T . . .          | = match
.                  | : - : | | | | |              : = mismatch
Read:              A G G G G C C T T              - = gap
```

Sequence Alignment

- Determine position of short read on the reference

```
Reference: . . . A A - C G C C T T . . .          | = match
.                  | : - : | | | | |             : = mismatch
Read:              A G G G G C C T T             - = gap
```

- Read could align to multiple places

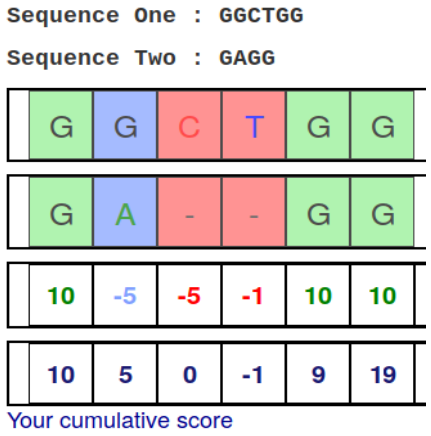


•How to handle multi-mapped reads? Depends on tool:
  - Map to best region (but what is "best"? And what about ties?)
  - Map to all regions
  - Map to one region randomly
  - Discard read

Alignment Scoring (basics)

أكاديمية كاوست
KAUST ACADEMY

- **Reward** for a match (e.g. +10), **penalty** for a mismatch (e.g. -5)
- **Penalty** for gaps
  - *Linear:* every gap same penalty (e.g. -5)
  - *Affine:* gap open vs gap extend (e.g. -5 and -1)
- Different tools use different scoring values (and give different results)



Sequence Alignment

```
Reference: AAA CAGTGA GAA
Observed:  AAA TCTCT  GAA
```

Alignment	
AAA-CAGTGAGAA |||-|--|::||| AAATC--TCTGAA	Maybe like this?
AAACAGTGAGAA |||-::|::||| AAA-TCTCTGAA	Or this?
AAACAGTGAGAA |||:-:|::||| AAAT-CTCTGAA	Or..?
AAACAGTCA-----GAA |||----------||| AAA------TCTCTGAA	What about this?

Sequence Alignment


```
Reference: AAA CAGTGA GAA
Observed:  AAA TCTCT  GAA
```

Alignment	Tool
AAA-CAGTGAGAA |||-|--|::||| AAATC--TCTGAA	Novoalign
AAACAGTGAGAA |||-::|::||| AAA-TCTCTGAA	Ssaha2
AAACAGTGAGAA |||:-:|::||| AAAT-CTCTGAA	BWA
AAACAGTCA-----GAA |||----------||| AAA------TCTCTGAA	Complete Genomics

Sequence Alignment

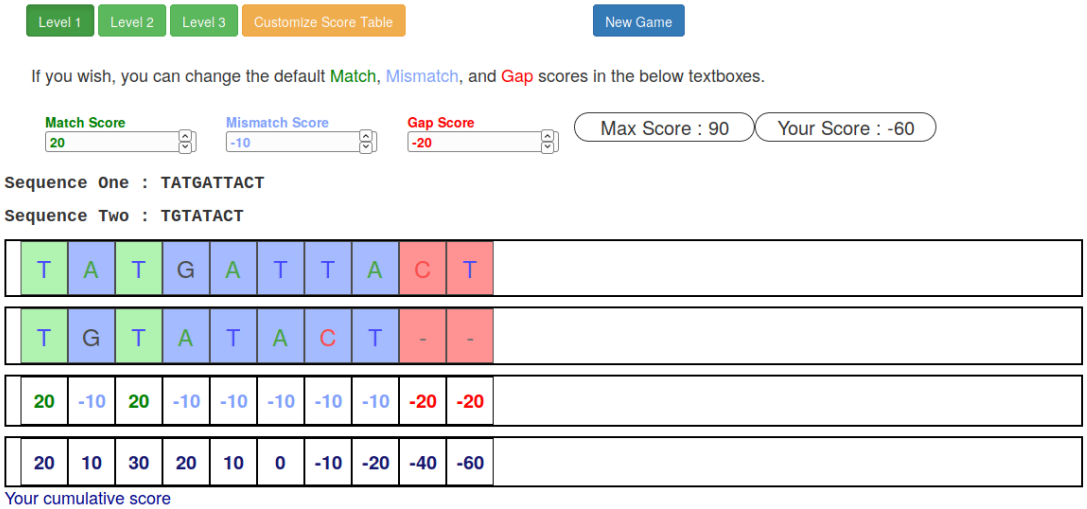
```
Reference: AAA CAGTGA GAA
Observed:  AAA TCTCT  GAA
```

Alignment	Variant calls
AAA-CAGTGAGAA |||-|--|::||| AAATC--TCTGAA	ins T del AG sub GA -> CT
AAACAGTGAGAA |||-::|::||| AAA-TCTCTGAA	del C sub AG -> TC sub GA -> CT
AAACAGTGAGAA |||:-:|::||| AAAT-CTCTGAA	snp C -> T del A snp G -> C sub GA -> CT
AAACAGTGA-----GAA |||----------||| AAA------TCTCTGAA	del CAGTGA ins TCTCT

Sequence Alignment

أكاديمية كاوست
KAUST ACADEMY

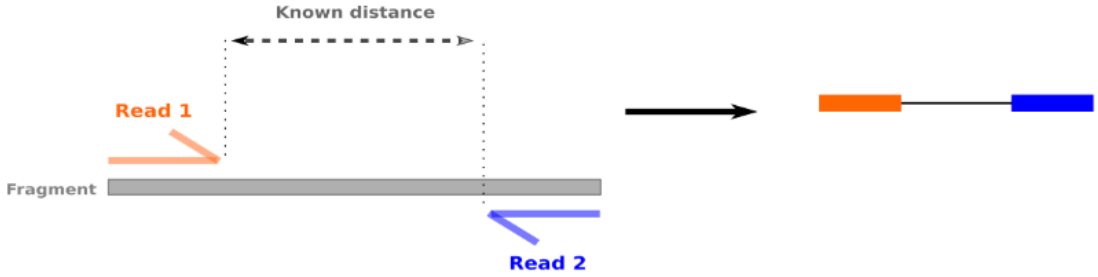
- Lego time! Who wants to volunteer?
- Or try this online sequence alignment game:



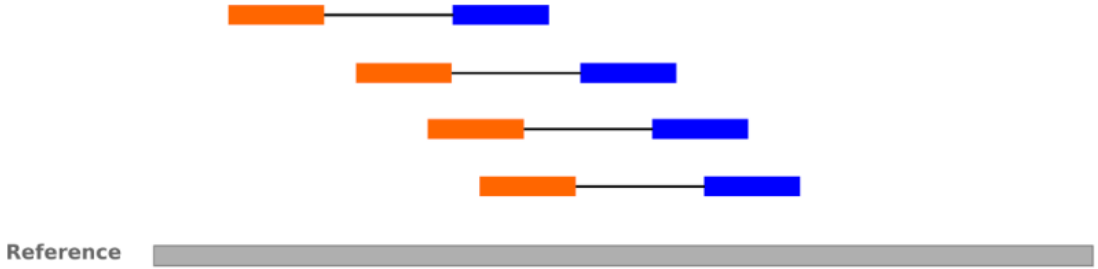
Paired-end sequencing

أكاديمية كاوست
KAUST ACADEMY

- **Sequencing:** Cut longer fragments of DNA, sequence only the ends

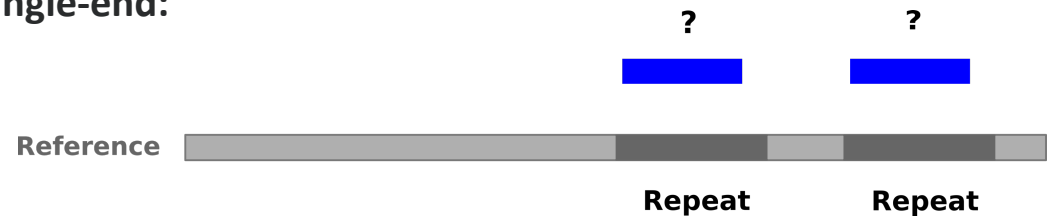


- **Mapping:** known distance between reads improves accuracy



Repeats

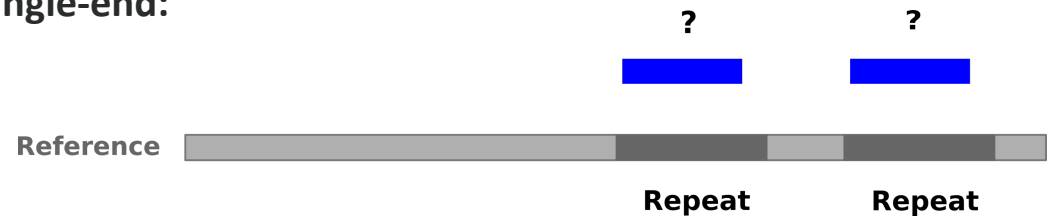
- Multi-mapped reads (e.g. because of repeats) may now be resolved
- **Single-end:**



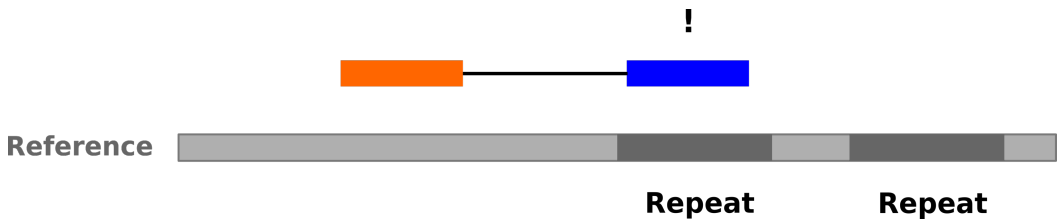
- **Paired-end:**

Repeats

- Multi-mapped reads (e.g. because of repeats) may now be resolved
- **Single-end:**

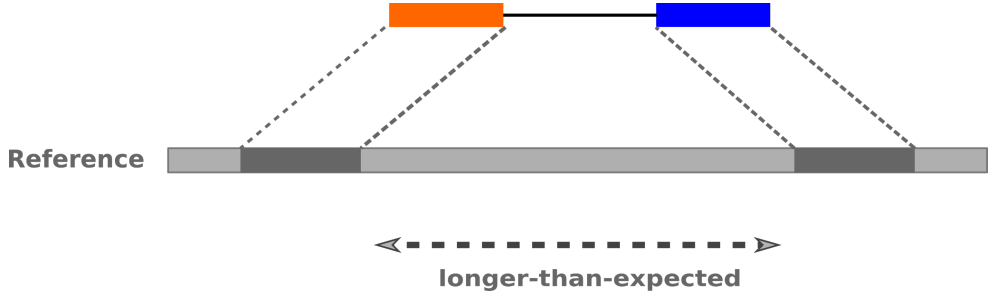


- **Paired-end:**

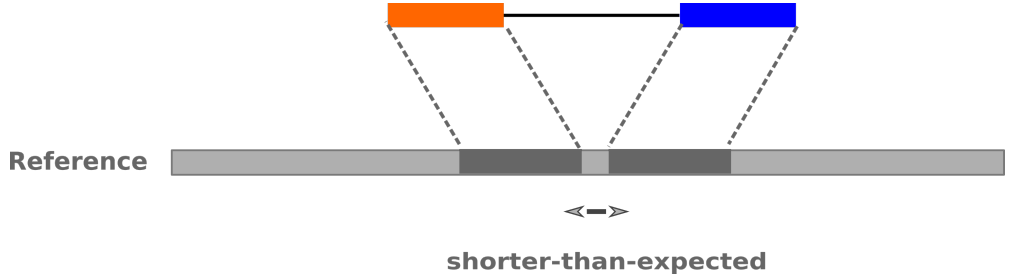


InDels (Insertions / Deletions)

- Discordant insert size may indicate insertion or deletion between reads
- **Deletions:** Longer mapping distance than expected

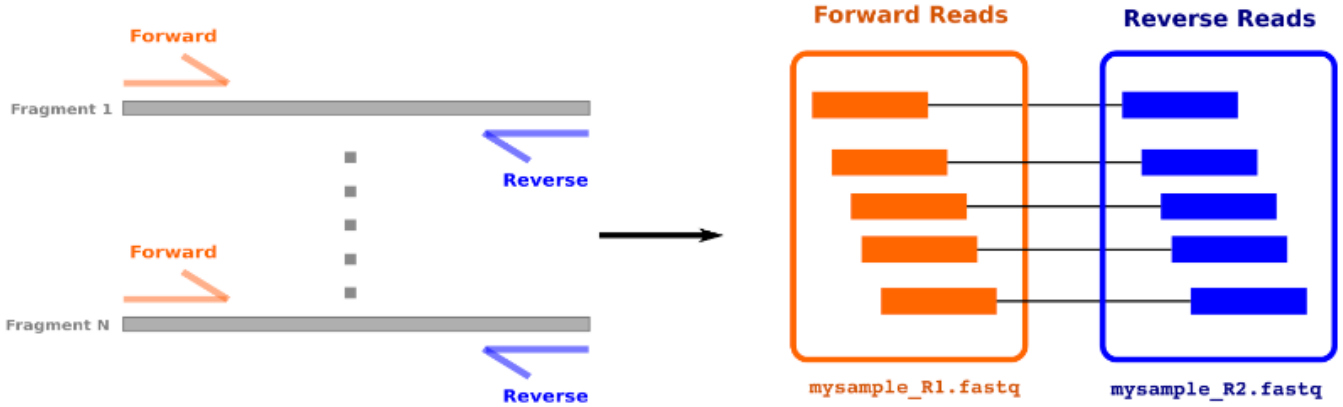


- **Insertions:** Shorter mapping distance than expected



Paired-end FASTQ Files

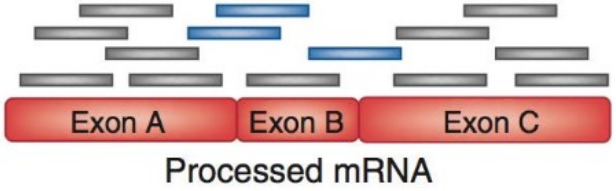
- Sequencer produces two FASTQ files:
  - **Forward** reads (usually `_1` or `_R1` in file name)
  - **Reverse** reads (usually `_2` or `_R2` in file name)



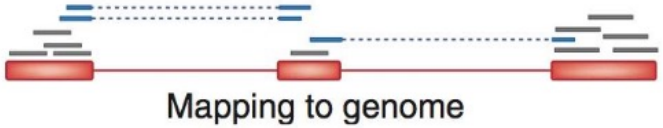
Choosing an Aligner

أكاديمية كاوست
KAUST ACADEMY

- Each tool makes **different choices** during alignment
  - Choice of aligner may **affect downstream results**
  - Default options may not be best for your data

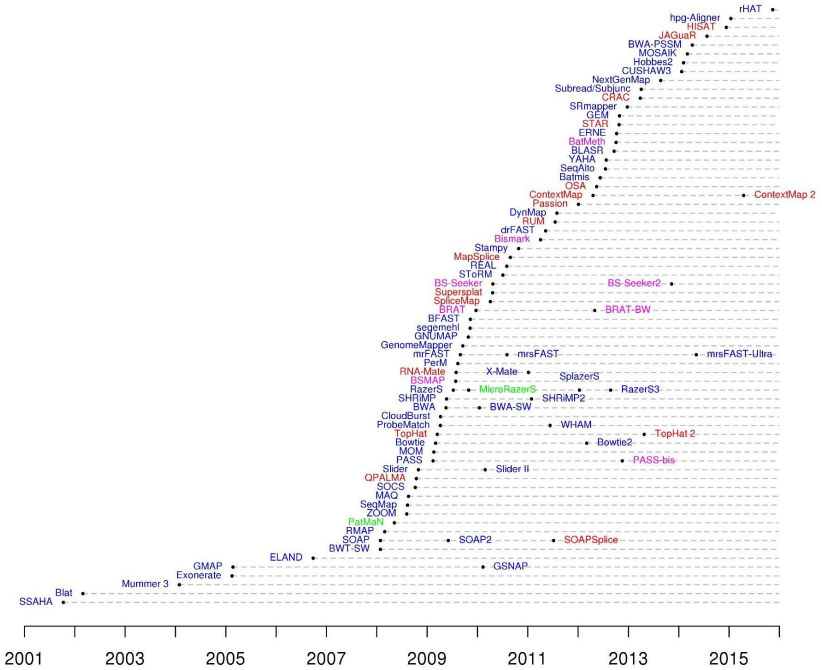


- Best tool for your data **depends on many factors**
  - Type of experiment (e.g. DNA, RNA, Bisulphite)
  - Sequencing platform
  - Compute resources vs sensitivity
  - Read characteristics (paired/single end, read length)



- Know your data!
- *“... there is no tool that outperforms all of the others in all the tests. Therefore, the end user should clearly specify [their] needs in order to choose the tool that provides the best results.”* - Hatem et al BMC Bioinformatics 2013, 14:184
- DOI: 10.1186/1471-2105-14-184

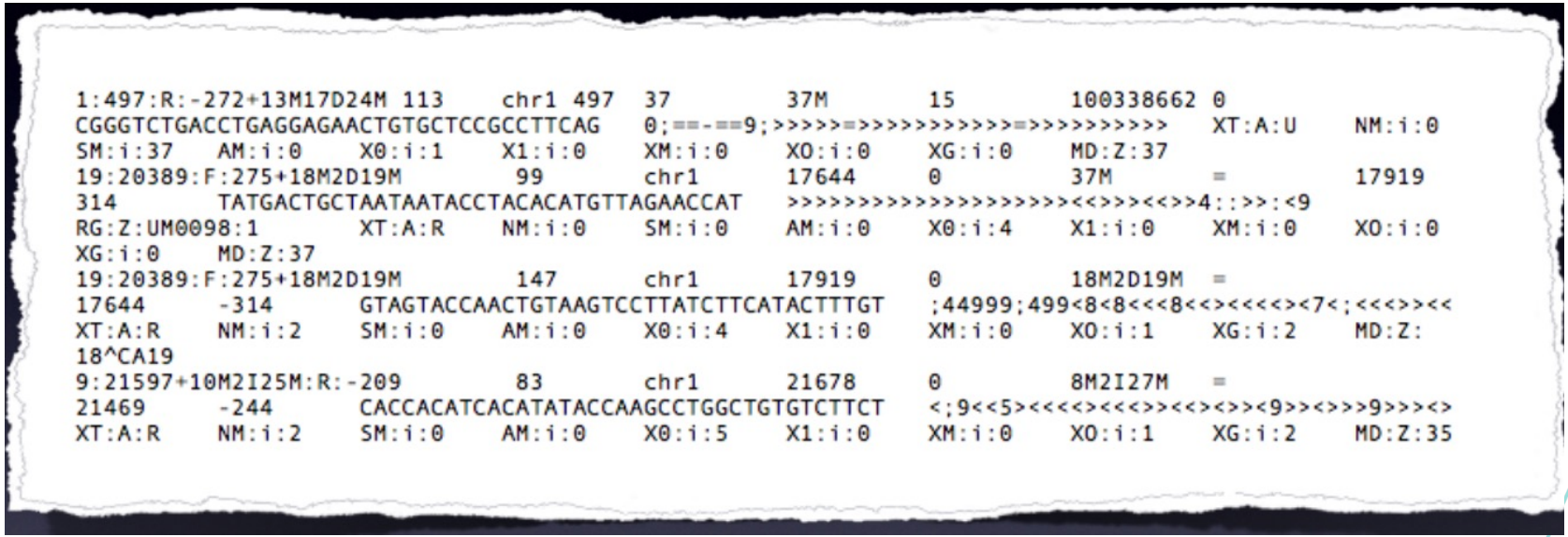
Mapping Tools



Mapping tool	Uses	Characteristics
HISAT2	DNA/RNA	Short reads. Based on GCSA. Reference.
RNASTAR	RNA	Short reads. Extremely fast. High sensitive and accuracy. Based on Maximal Mappable Prefixes (MMPs). Reference.
BWA-MEM2	DNA	Short reads. Twice as faster as BWA-MEM. Memory efficient. Based on Burrows-Wheeler. Reference.
Minimap2	DNA/RNA	Long reads (PacBio and ONT). Extremely fast. Based on DALIGN and MHAP. Reference.
Bismark	DNA/RNA	Short reads. Bisulfite treated sequencing. Based on GCSA. Reference.
BBMap	DNA/RNA	Short and long reads (PacBio and ONT). Memory demanding. Reference.
Whisper 2	DNA	Short reads. Indel sensitive. Variant-calling oriented. Reference.
S-conLSH	DNA	Long reads (ONT). High sensitivity and accuracy. Reference.

60+ different mappers, many comparison papers. Figure from 10.1093/bioinformatics/bts605

SAM/BAM File Format



```
1:497:R:-272+13M17D24M 113     chr1 497  37        37M       15        100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG    0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>    XT:A:U    NM:i:0
SM:i:37    AM:i:0    X0:i:1    X1:i:0    XM:i:0    XO:i:0    XG:i:0    MD:Z:37
19:20389:F:275+18M2D19M         99        chr1      17644     0         37M       =         17919
314        TATGACTGCTAATAATACCTACACATGTTAGAACCAT    >>>>>>>>>>>>>>>>>>>><<>>>><<>>4::>>:<9
RG:Z:UM0098:1        XT:A:R    NM:i:0    SM:i:0    AM:i:0    X0:i:4    X1:i:0    XM:i:0    XO:i:0
XG:i:0    MD:Z:37
19:20389:F:275+18M2D19M         147       chr1      17919     0         18M2D19M  =
17644      -314      GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT    ;44999;499<8<8<<<8<<><<<<<><7<;<<<>><<
XT:A:R     NM:i:2    SM:i:0    AM:i:0    X0:i:4    X1:i:0    XM:i:0    XO:i:1    XG:i:2    MD:Z:
18^CA19
9:21597+10M2I25M:R:-209         83        chr1      21678     0         8M2I27M   =
21469      -244      CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT    <;9<<5><<<<><<<><<><><><><9><><>>9><><>
XT:A:R     NM:i:2    SM:i:0    AM:i:0    X0:i:5    X1:i:0    XM:i:0    XO:i:1    XG:i:2    MD:Z:35
```

SAM: Sequence **A**lignment **M**ap

BAM: Binary (compressed) SAM; not human-readable Formats

SAM/BAM File Format

- Original read information (from FASTQ) plus mapping information
  - Position on reference, alignment, quality score, uniqueness, ..

Genome Browsers

- Visualise aligned reads (BAM files)

This is IGV (Integrative Genome Browser) DOI: 10.1038/nbt.1754

Key points

أكاديمية كاوست
KAUST ACADEMY

- Mapping is not trivial
- There are many mapping tools, best choice depends on your data
- Choice of mapper can affect downstream results
- Know your data!
- Genome browsers can be used to view aligned reads

Hands-on and Practical Part

Galaxy PROJECT

GitHub

Part 3: Sequence alignments

- Map reads on a reference genome
- Inspect a BAM/SAM file

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Done with Day 1, Heyyyyy!

Thank You !

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology