



SevenBridges

---

# Next Generation Sequencing Technologies

Boris Majić - [boris.majic@sbgenomics.com](mailto:boris.majic@sbgenomics.com)

08.03.2022.

# Why Sequence the Genome?

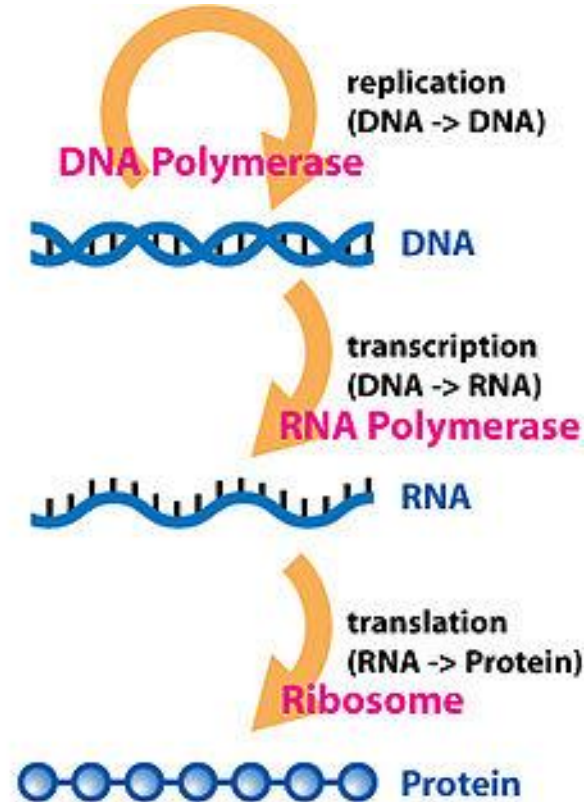
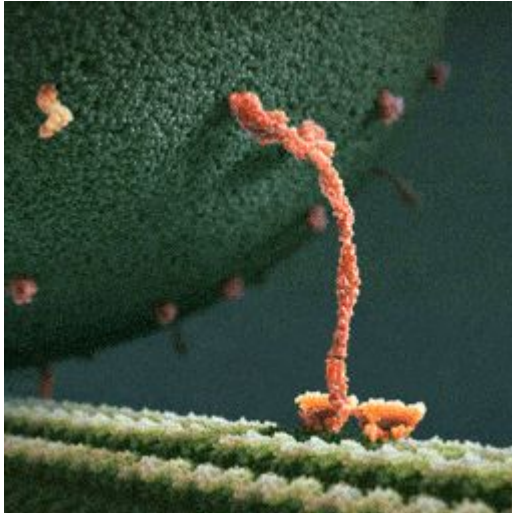
---

What are the benefits of genome sequencing?



# Refresher

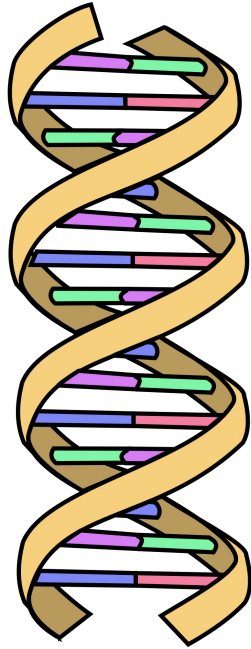
## Central dogma



# Sequencing

- What does “**sequencing**” mean?
- Why sequence?
- Why sequence DNA and RNA?

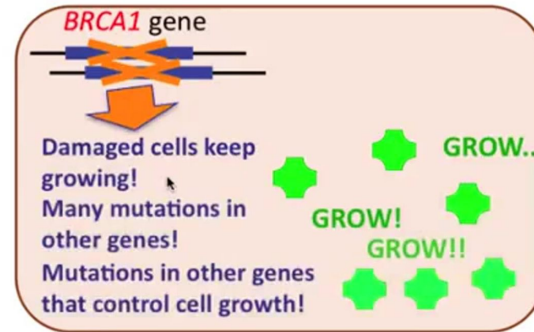
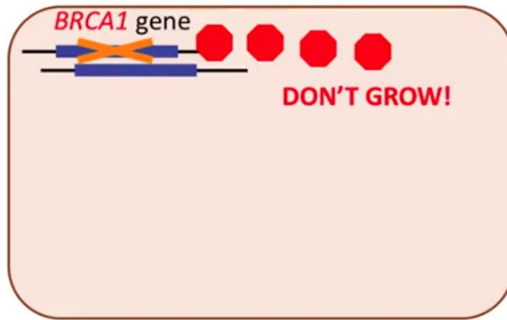
# What does it mean?



ACCATCATAGTCGTGAT  
CGTAGCAGTGCCATGG  
GGTCATATATAGCAGTA  
CAGATCGATGCATCGA  
TGAATTTTCAACAGTGC  
C

# Why sequence DNA?

BRCA1 and BRCA2 genes (BReast CAncer susceptibility genes) code for proteins that block cell growth when cell is damaged

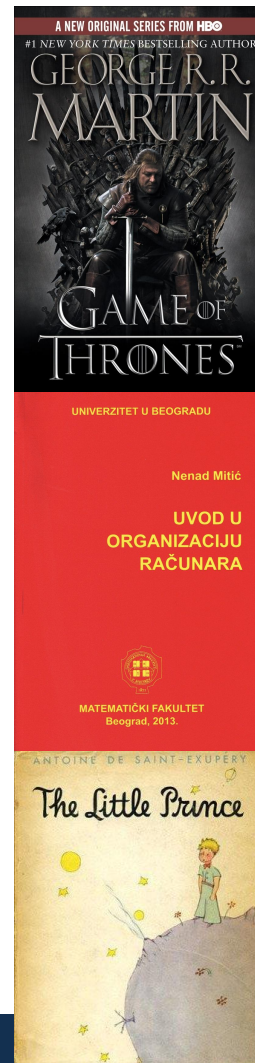


# Do we need computers?

- Human genome is 3 billion bps long
- It would take 1M pages (stack 50m high) to write down a **single** human genome sequence
- Impossible to analyze manually



Yes, we need computers

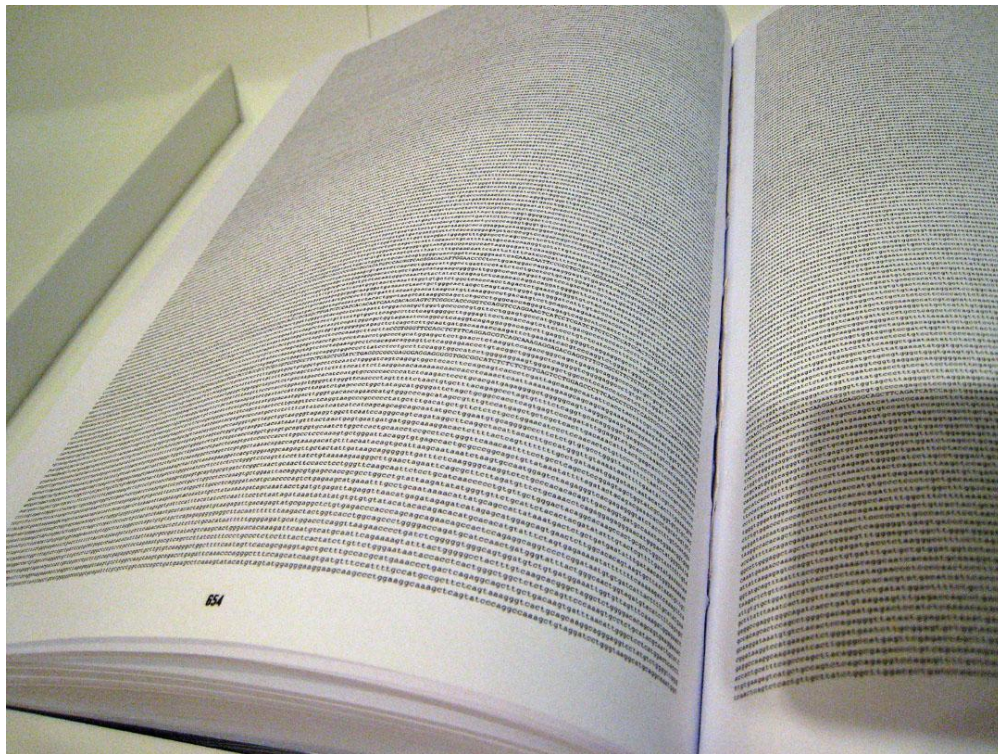


x3000

x10000

x30000





'The Human Genome printed out into books'  
 Title: [Genome](#). Image by [Dmitry Diouze](#), via [flickr](#) (CC BY 2.0)

Page from The Human Genome. Taken from [everydaybird](#)  
[blog](#)



# The Birth of Sequencing Technologies

The discovery of DNA polymerase, modified nucleotides  
and the Chain termination method

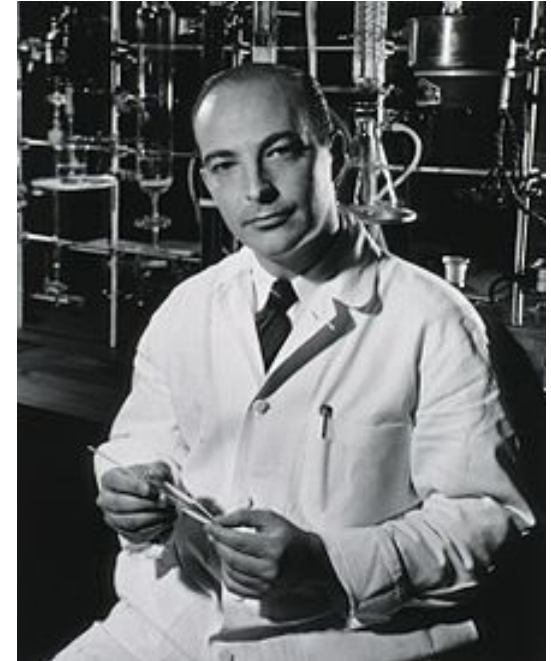


# Important concepts and ideas for understanding sequencing

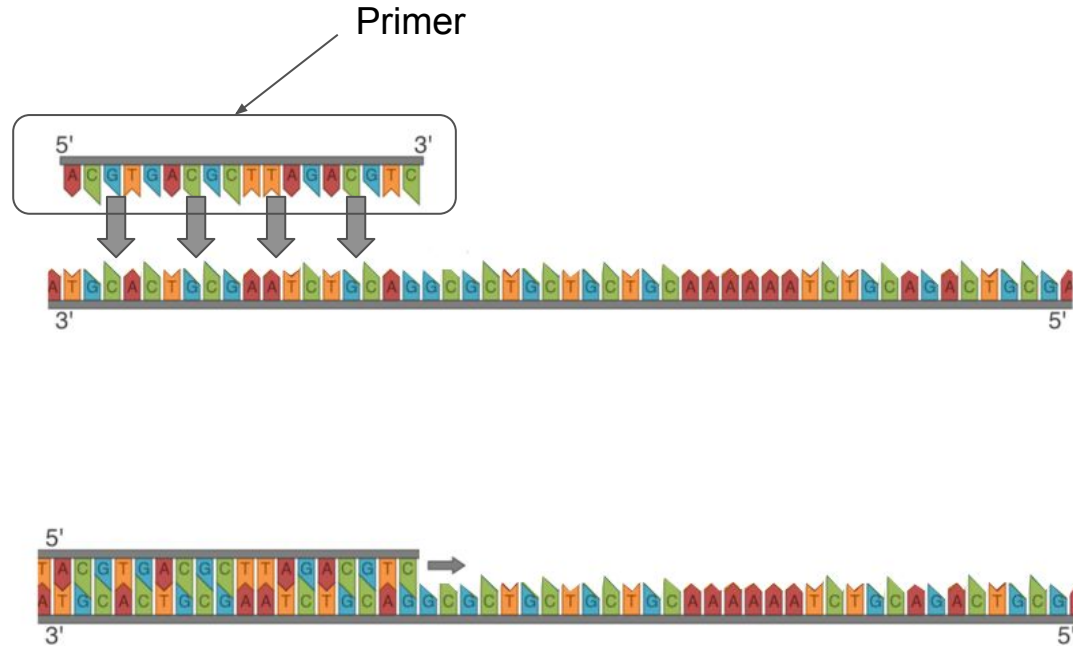
- DNA replication
- DNA ligation
- DNA fragmentation and gel electrophoresis
- Nucleotide labeling
- Synthesis termination

# Crucial Concept: DNA Replication

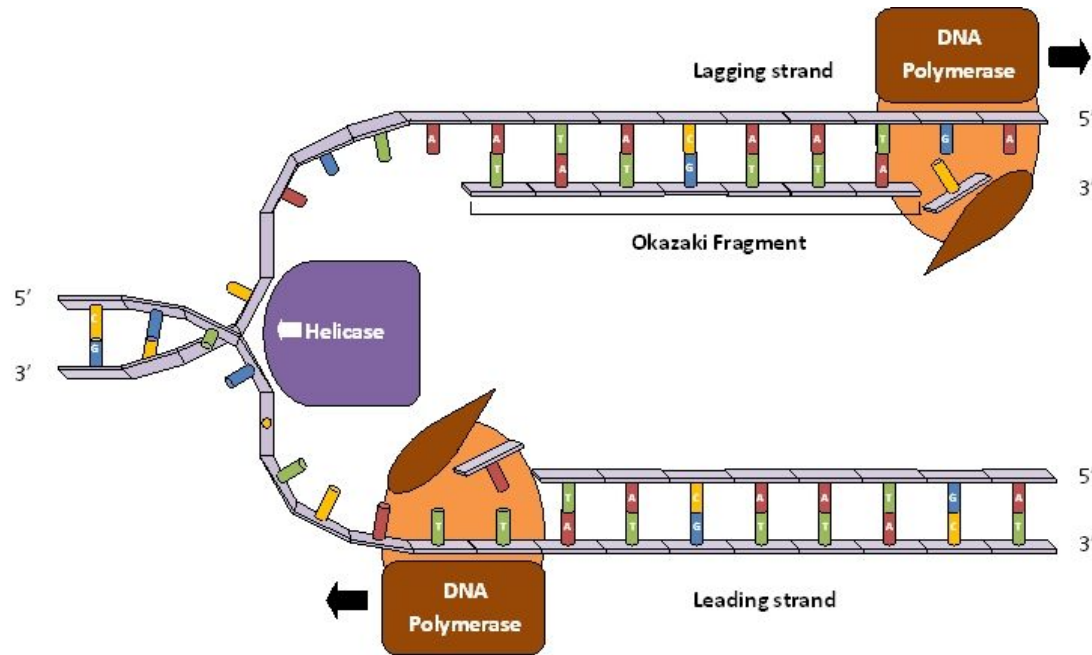
- **1956** - Arthur Kornberg et. al. (New York University)
  - discovered the **DNA Polymerase**
  - And that polymerase facilitates **DNA synthesis**
    - zipper molecule



# PCR Primers



# DNA Replication

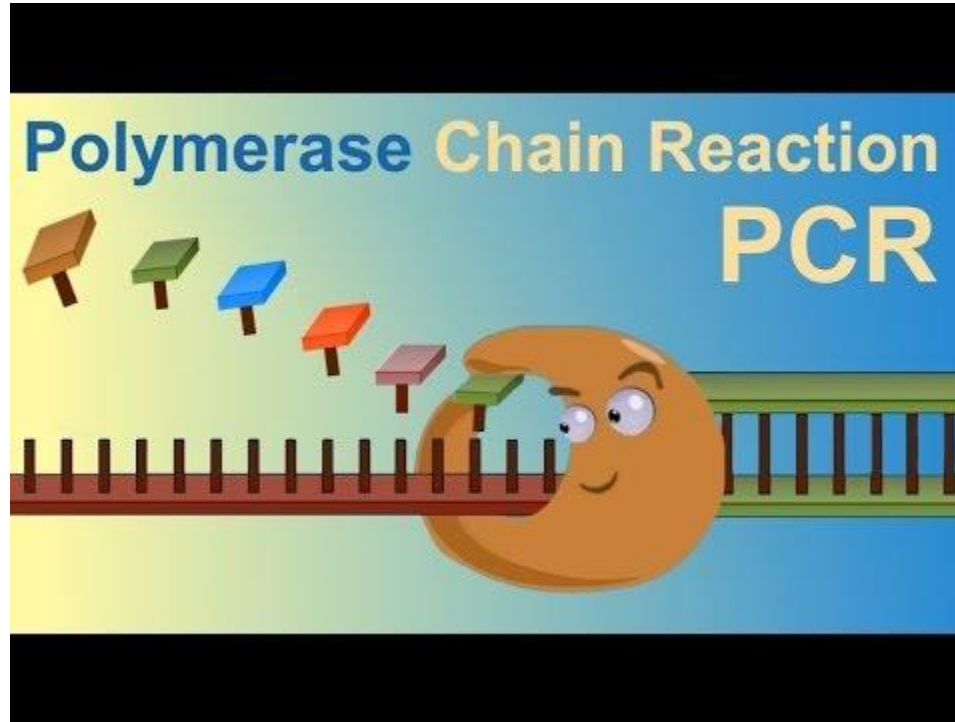


# PCR: Polymerase Chain Reaction

- DNA Polymerase Chain Reaction (PCR) chemistry based on **thermo cycling**; acute and repeated i.e. cyclic temperature fluctuations for denaturation and hybridization of complementary strands of DNA
- **Taq Polymerase** - *Thermus Aquaticus*
  - Not human polymerase
  - Taq is more temperature resistant to high denaturation temperatures of thermocycler

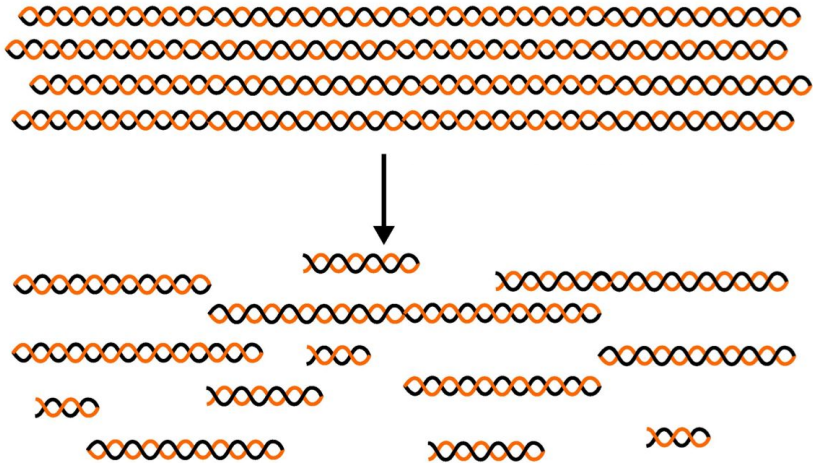


# PCR: Polymerase Chain Reaction





# DNA fragmentation and ligation



Fragmentacija

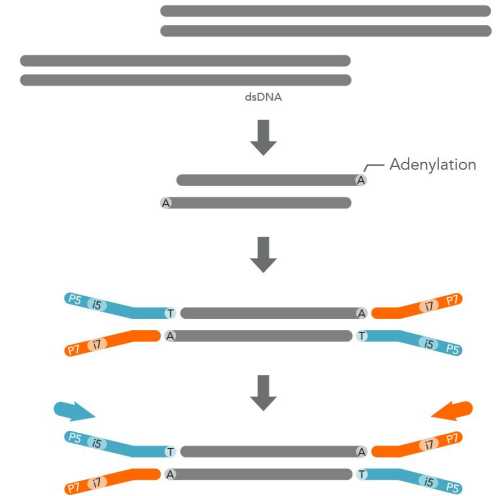
Preuzeto sa [Labster theory](#)

Fragmentation

End repair and A-tailing

Ligation

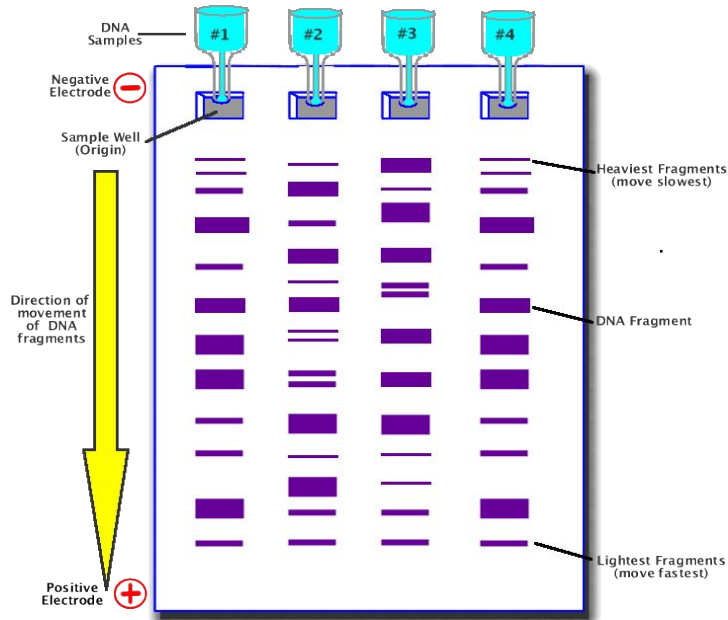
PCR amplification



Ligacija

Preuzeto sa [IDTDNA](#)

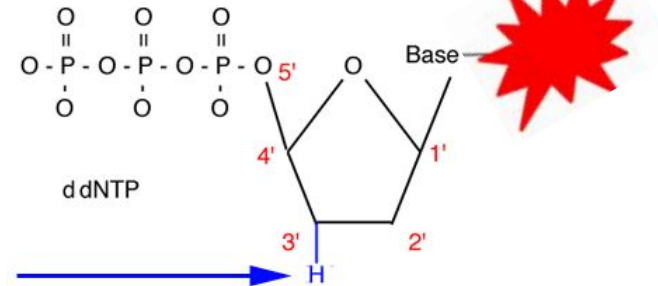
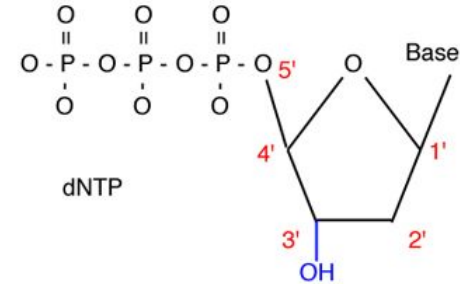
# Gel Electrophoresis



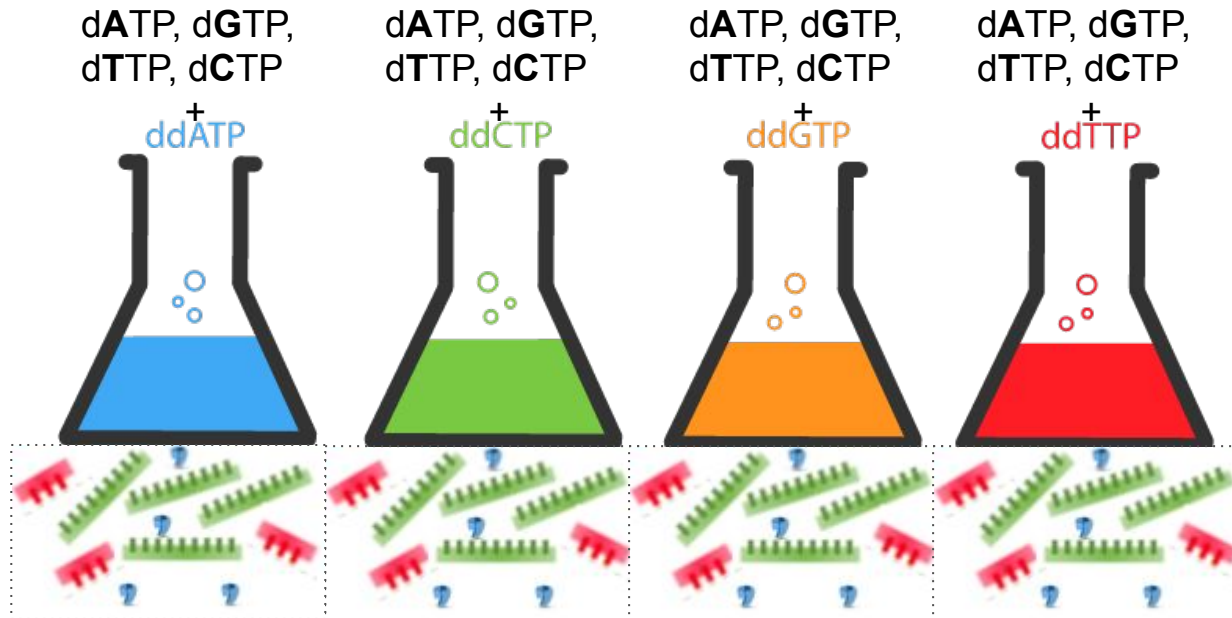
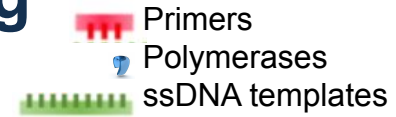
**Gel Electrophoresis**  
(Creating a DNA Profile)

# Terminating and Labeling Nucleotides

- **Sanger**
  - Chain **Termination** Method
  - **dNTPs** + labeled **ddNTPs**
  - Southern Blotting on Electrophoresis gel
- **NGS**
  - Reversible termination
  - No termination
  - Slides



# Sanger sequencing

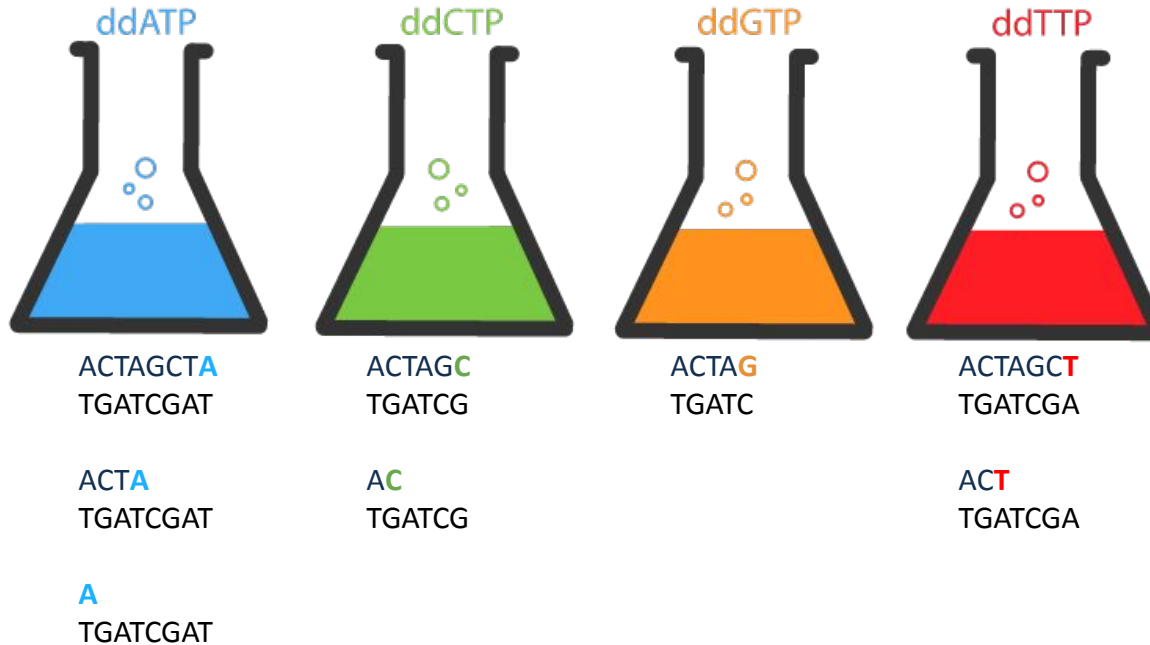


## Sanger sequencing example

**ACTAGCTA**

TGATCGAT

# Sanger sequencing



# Sanger sequencing

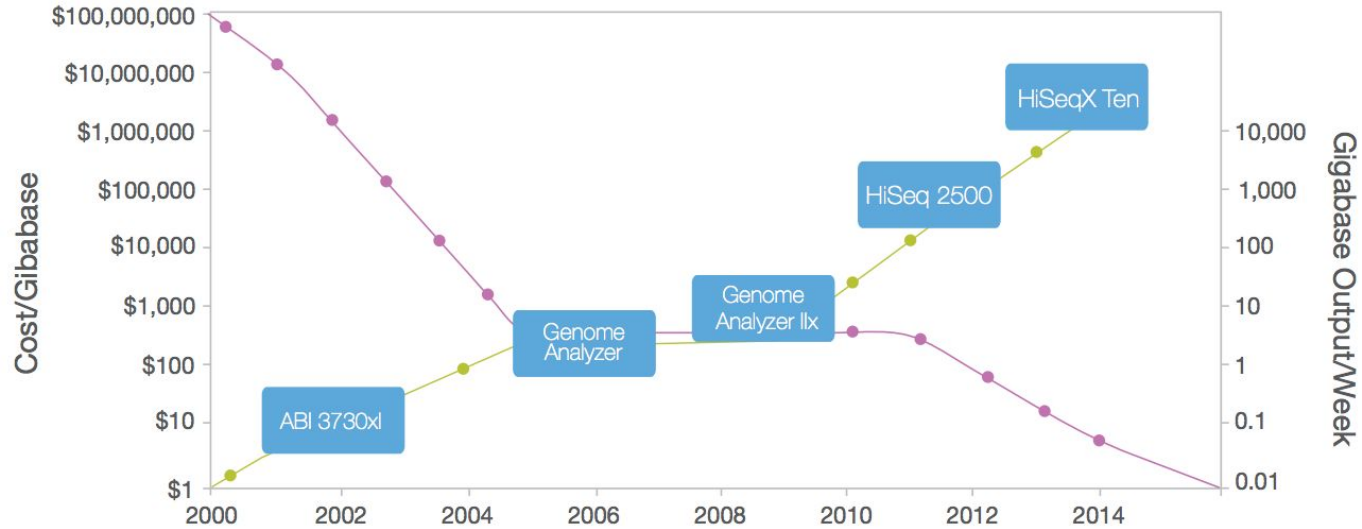




# Sanger sequencing



# Sequencing Trends



**Figure 1: Sequencing Cost and Data Output Since 2000**—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

# The Birth of Next Generation Sequencing Technologies

---

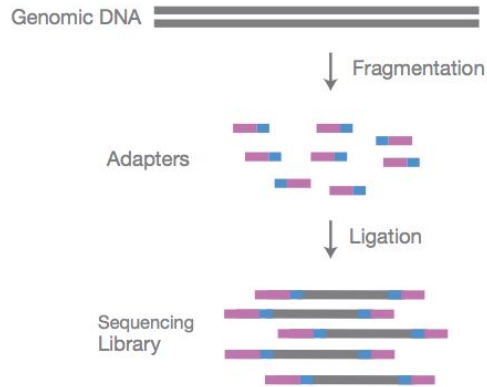
DNA polymerase in the sequencing by synthesis approach



illumina®

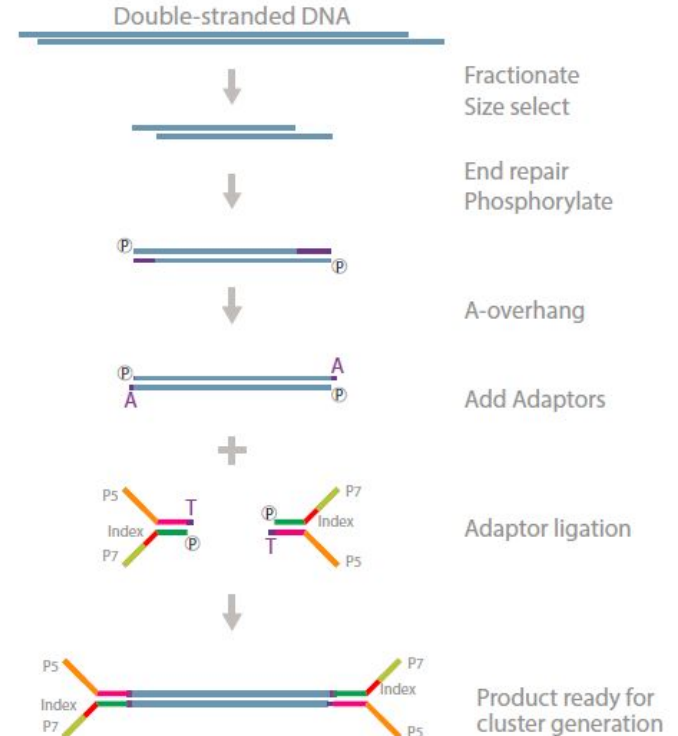
# Illumina Library Prep

## A. Library Preparation

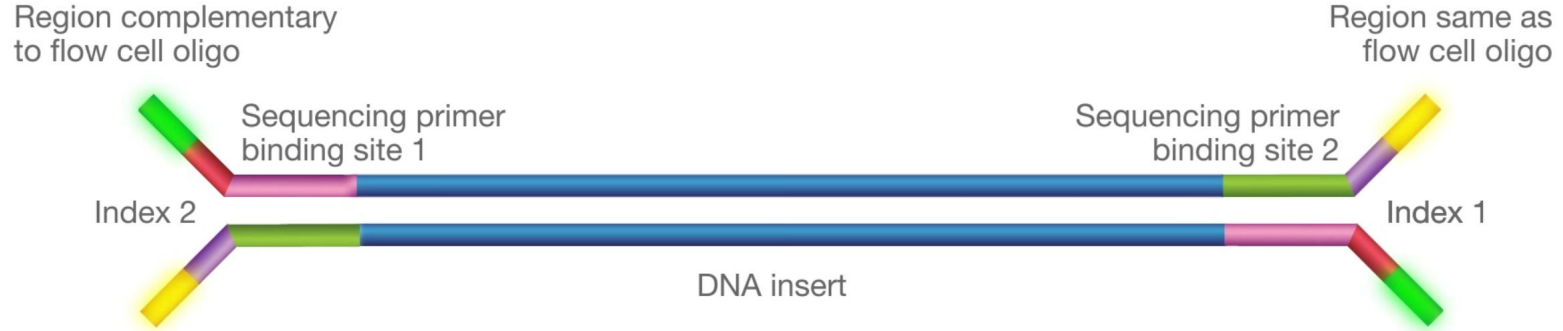


NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

## TruSeq PCR Free



# Illumina library prep



# Illumina flow cell

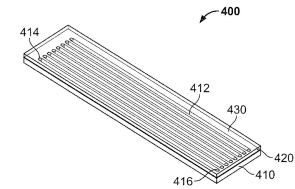
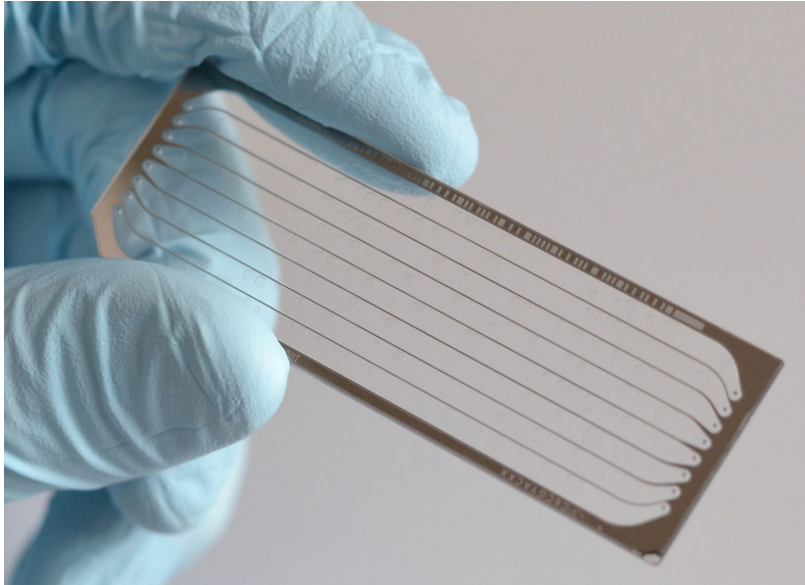


FIG. 3A

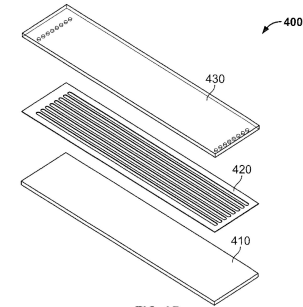
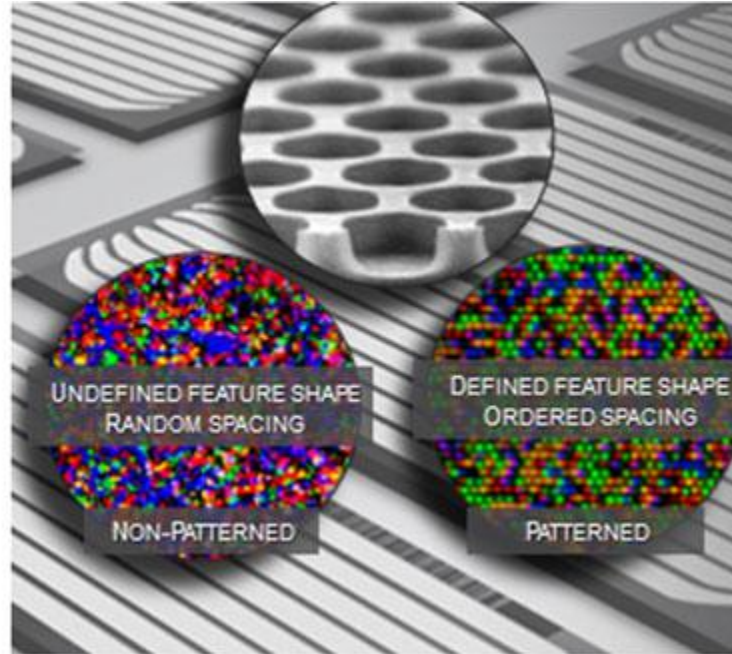
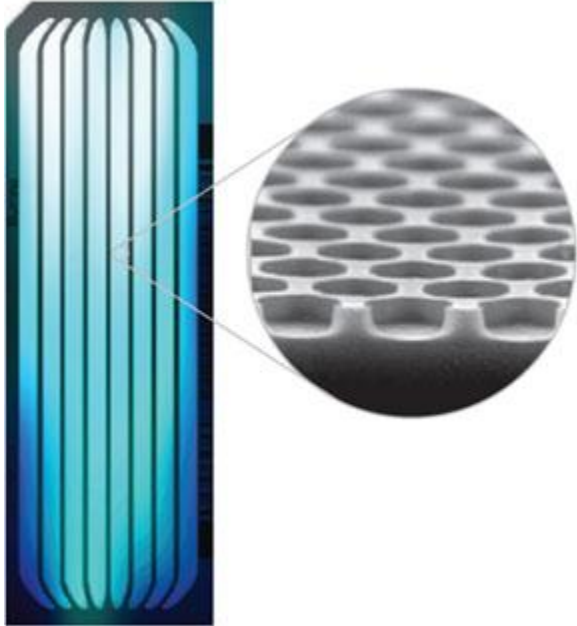


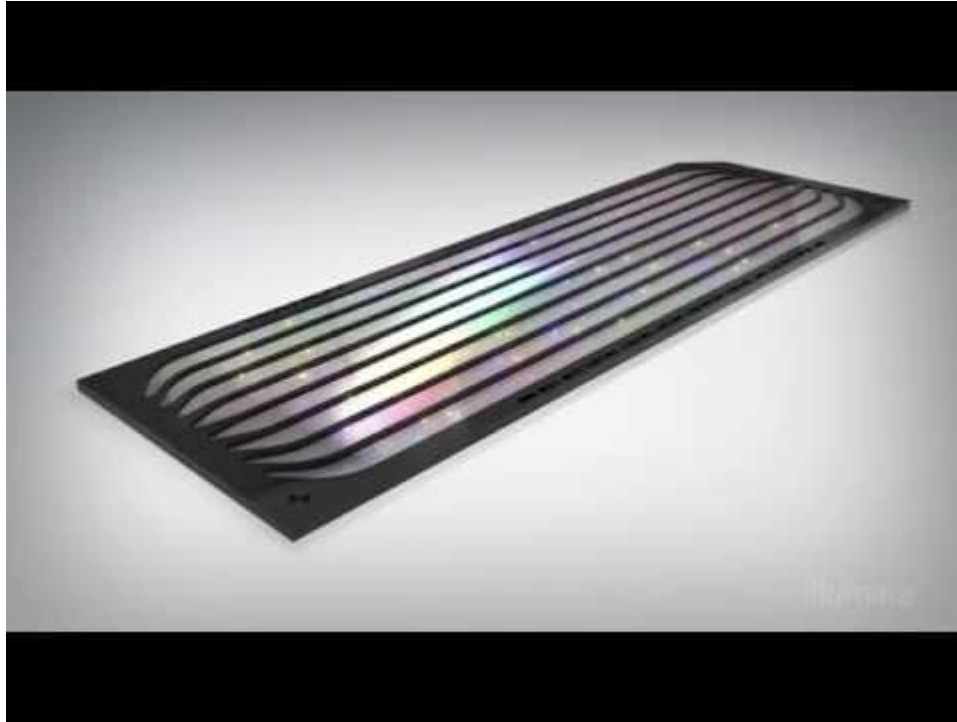
FIG. 3B



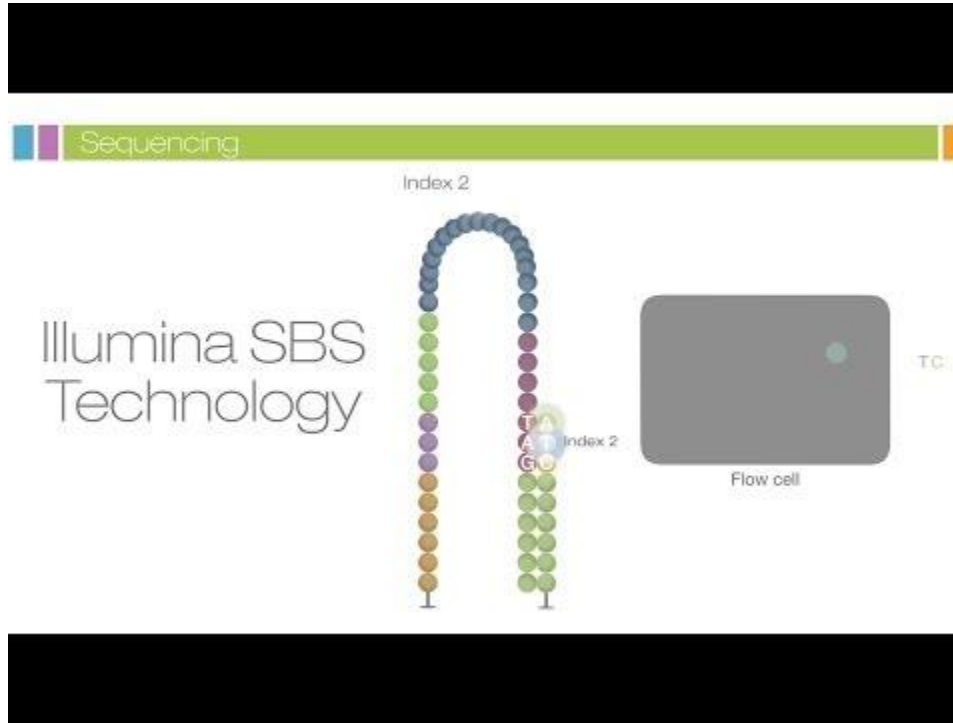
# Illumina flow cell



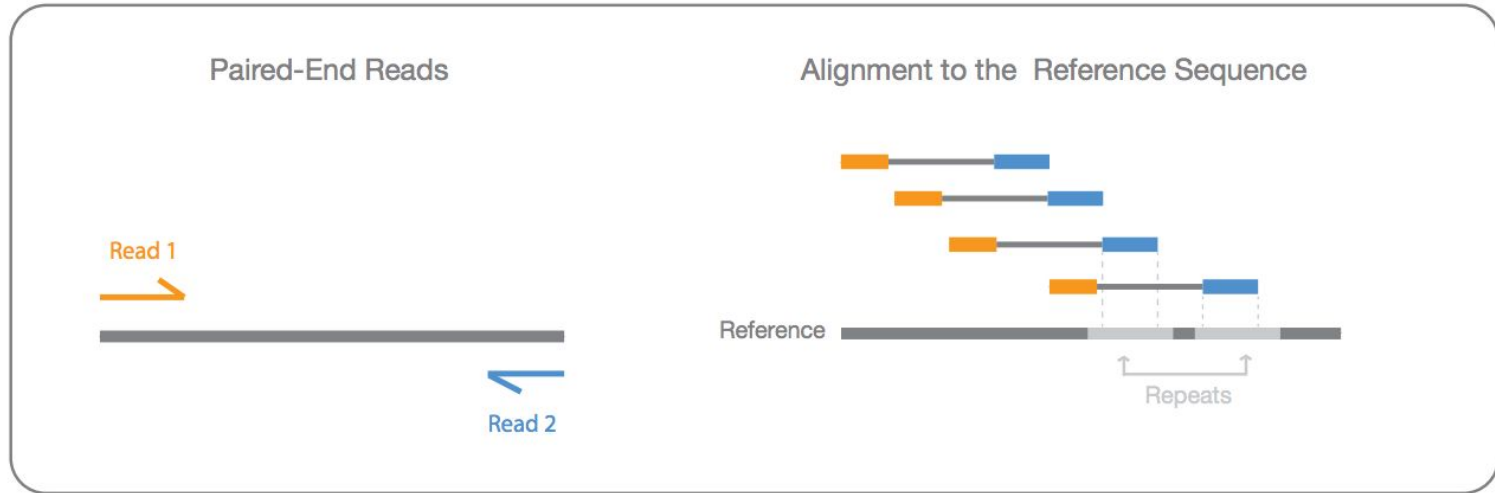
# Cluster Amplification



# Illumina sequencing

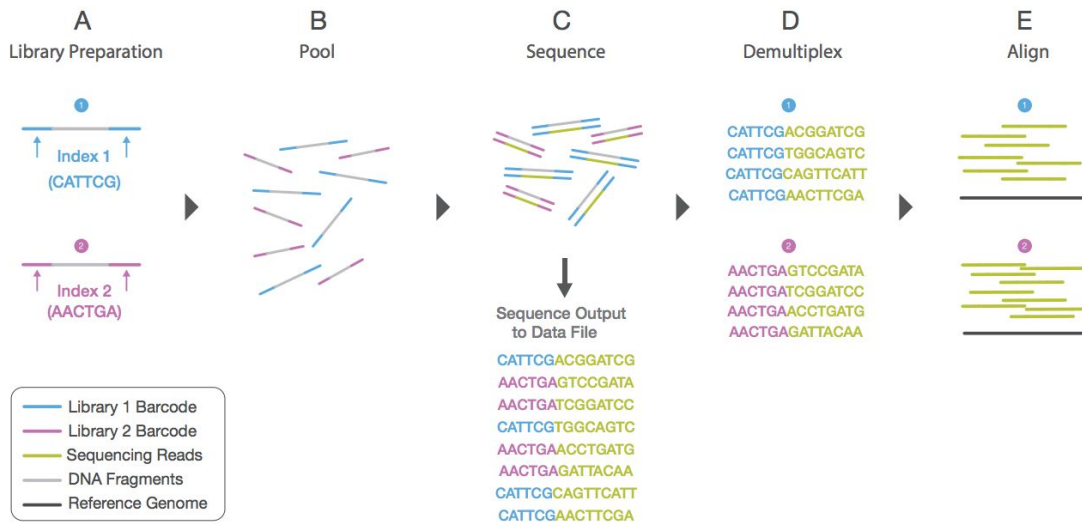


# Illumina sequencing: Paired ends



**Figure 4: Paired-End Sequencing and Alignment**—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

# Illumina sequencing: Multiplexing



**Figure 5: Library Multiplexing Overview.**

- Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.
- Libraries are pooled together and loaded into the same flow cell lane.
- Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.
- A demultiplexing algorithm sorts the reads into different files according to their indexes.
- Each set of reads is aligned to the appropriate reference sequence.

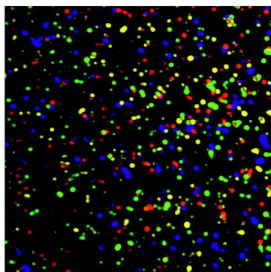
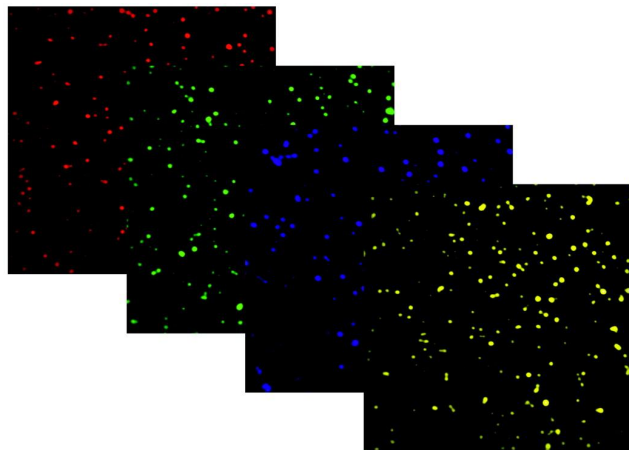
# Recording called bases or reads

---

The FASTQ file format



# Image processing



- Sequencing produces high-resolution TIFF images
- 100 tiles per lane, 8 lanes per flow cell, 100 cycles
- 4 images (A,G,C,T) per tile per cycle = 320,000 images
- Each TIFF image ~ 7Mb = 2,240,000 Mb of data (2.24TB)

Base calling - uses cluster intensities and noise estimate to output the sequence of bases read from each cluster, along with a confidence level for each base.



# Phred Scoring

## Quality score interpretation

$$Q = -10 \log_{10} P \quad \longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

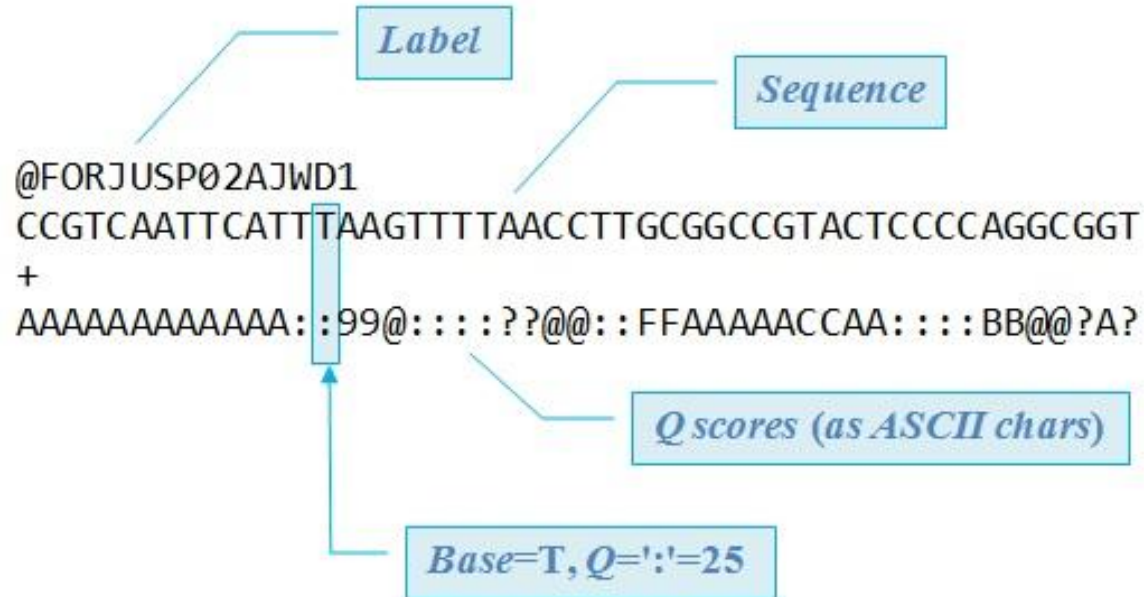
Materials from Wikipedia

# ASCII Base + 33 encoding

ASCII\_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

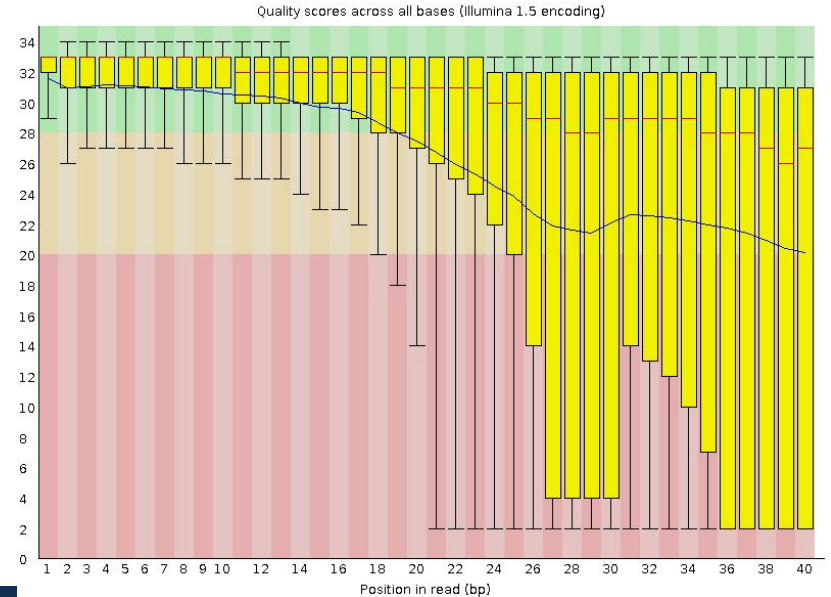
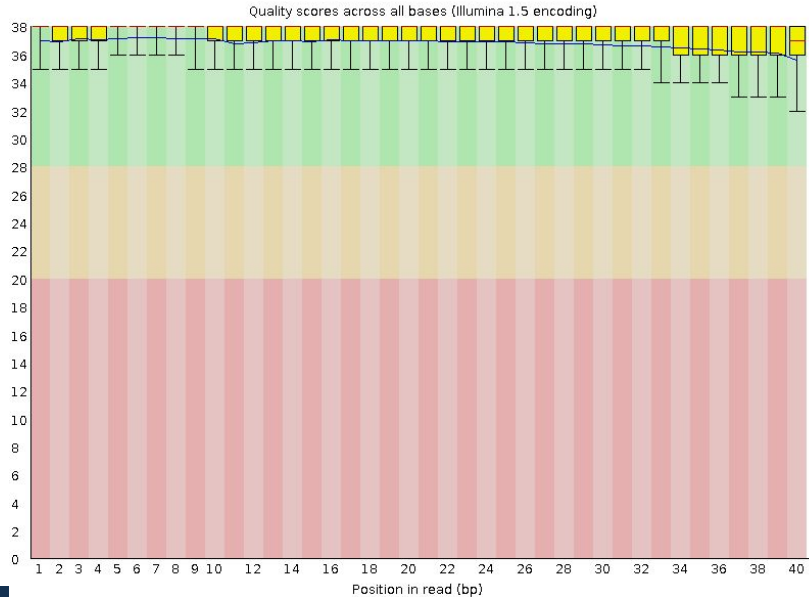
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :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

# FASTQ file format



# FASTQ - Analysis

## Per Base Sequence Quality



# Other NGS Technologies

---

Nanopore sequencing



# SMRT/ZMV Sequencing



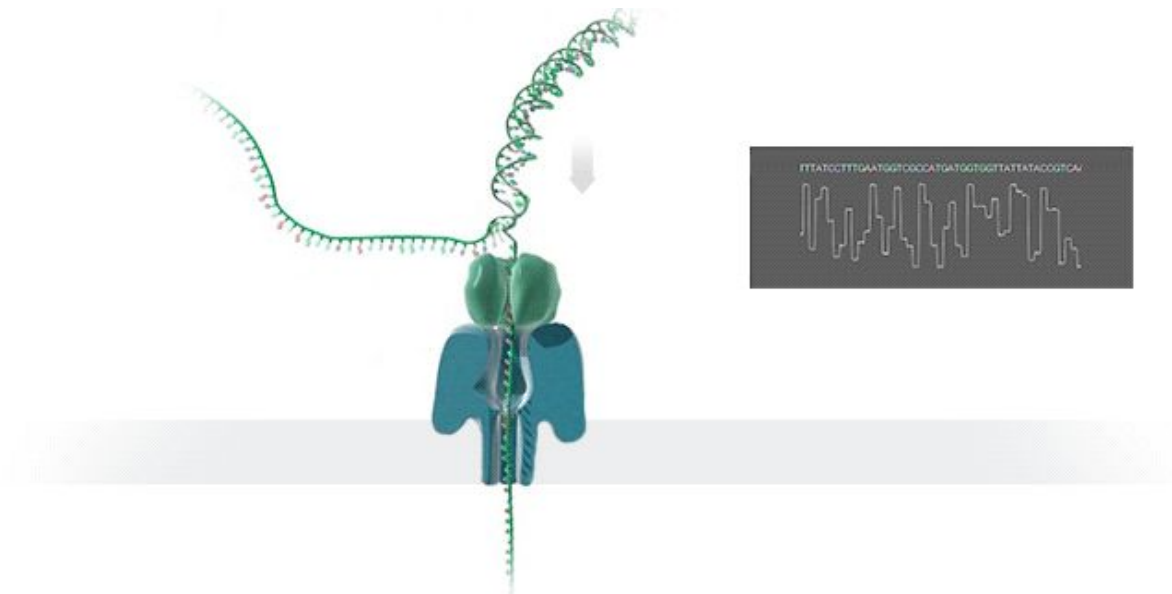
PACIFIC  
**BIOSCIENCES®**




Oxford

The logo for Oxford Nanopore Technologies, featuring a stylized circular icon composed of teal and dark grey segments, followed by the text "Oxford" in teal, "NANOPORE" in large bold dark grey letters, and "Technologies" in teal.

Technologies



# Oxford Nanopore MinION



Topics | Missions | Galleries | NASA TV | Follow NASA | Downloads | About | NASA Audiences

Search

Latest

Related

Space Station Science Highlights: Week of Feb 26, 2018  
4 days ago

Space Station Science Highlights: Week of Feb 19, 2018  
11 days ago

Space Station Science Highlights: Week of Feb 12, 2018  
18 days ago

Space Station Science Highlights: Week of Feb 5, 2018  
22 days ago

3-D Printable Tools May Help Study Astronaut Health  
a month ago

Bursting with Excitement – A Look at Bubbles and Fluids in Space  
a month ago

Aug. 29, 2016

## First DNA Sequencing in Space a Game Changer






For the first time ever, DNA was successfully sequenced in microgravity as part of the [Biomolecule Sequencer](#) experiment performed by NASA astronaut Kate Rubins this weekend aboard the [International Space Station](#). The ability to sequence the DNA of living organisms in space opens a whole new world of scientific and medical possibilities. Scientists consider it a game changer.


DNA, or deoxyribonucleic acid, contains the instructions each cell in an organism on Earth needs to live. These instructions are represented by the letters A, G, C and T, which stand for the four chemical bases of DNA, adenine, guanine, cytosine, and thymine. Both the number and arrangement of these bases differ among organisms, so their order, or sequence, can be used to identify a specific organism.

The [Biomolecule Sequencer](#) investigation moved us closer to this ability to sequence DNA in space by demonstrating, for the first time, that DNA sequencing is possible in an orbiting spacecraft.


With a way to sequence DNA in space, astronauts could diagnose an illness, or identify microbes growing in the [International Space Station](#) and determine whether or not they represent a health threat. A space-based DNA sequencer would be an important tool to help protect astronaut health during long duration missions on the journey to Mars, and future explorers could also potentially use the technology to identify DNA-based life forms beyond Earth.

The Biomolecule Sequencer investigation sent samples of mouse, virus and bacteria DNA to the space station to test a commercially available DNA sequencing device called MinION, developed by Oxford Nanopore Technologies. The MinION works by sending a positive current through pores embedded in membranes inside the device, called nanopores. At the same time, fluid containing a DNA sample passes through the device. Individual DNA molecules partially block the nanopores and





NASA Astronaut Kate Rubins sequenced DNA in space for the first time ever for the Biomolecule Sequencer investigation, using the MinION sequencing device.  
**Credits: NASA**



## [MinION in Space Video](#)





# Rosalind

[Counting DNA Nucleotides](#)

[Transcribing DNA into RNA](#)

[Complementing a Strand of DNA](#)

Parsing FASTQ file assignment:

- Count the number of reads
- Calculate average read quality
- Calculate per base average quality
- Create histogram of average read quality
- Create histogram of read quality per base

# Resources

- Useful Genetics, <https://www.youtube.com/channel/UCtXCrx28msMBQ-vFUIOIReA>
- Codecademy, <http://www.codecademy.com>
- Illumina inc., <https://www.youtube.com/channel/UCxWMU29FF4kIG8YmQf6Zv0g>
- AWS <https://aws.amazon.com/documentation/ec2/>
- Ion Torrent sequencing, Thermo Fisher Sci., Life Tech., <https://www.youtube.com/user/LifeTechnologiesCorp>
- Oxford Nanopore, <https://www.youtube.com/channel/UC5yMIYjHSgFfZ37LYq-dzig>
- PacBio, <https://www.youtube.com/user/PacificBiosciences>
- IPython, virtualenv, virtualenvwrapper
- Jupyter Notebook
- nbviewer, <http://nbviewer.ipython.org/>
- Sequencing, [http://mol-biol4masters.masters.grkraj.org/html/Genetic\\_Engineering5A-DNA\\_Synthesis\\_&Sequencing.htm](http://mol-biol4masters.masters.grkraj.org/html/Genetic_Engineering5A-DNA_Synthesis_&Sequencing.htm)