Bioinformatics Algorithms
COS-BIOL-530/630
Lecture11

Days & Times	**Room**	**Meeting Dates**
Tu 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025
Th 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

Sequencing Technologies
- Lecture11-

Announcements

Lecture11
Lab11

- Activity 11
- Discussion 11

Quiz 9: Open Friday April 4th 5 pm (**this week!**)

- Lecture/Lab 10 (Gene Prediction)
- Lecture/Lab 11 (High Throughput Sequencing)

Exam 2: Thursday April 17th 2pm (GOS)-2178

- Lecture/Lab 7 to Lecture/Lab 11

Genome Alignment and Assembly
- Lecture12-

RNA-seq
- Lecture13-

Quiz 10: Due on April 22nd

- Lectures/Lab 12 and Lecture 13

Final Project: Due on May 1st

Sequencing Technologies
- Lecture11-

Topics:

- Genome sequencing

- Sanger sequencing

- High Throughput Sequencing (HTS) technologies

Genome Sequencing
Why?

structure

- *De novo* whole genome sequencing
  - Requires *de novo* whole genome assembly

- Polymorphism discovery
  - Target enrichment (exome)
  - Whole genome (resequencing)
  - Single nucleotide polymorphism (SNPs), copy number variations, insertions, deletions…

- Expressed sequence discovery
  - Expressed Sequence Tags (ESTs)
  - cDNAs
  - miRNAs, piRNAs (non-coding RNAs)

- Functional genomics
  - Chromatin ImmunoPrecipitation (ChIP)
  - Expression profiling
  - Nucleosome positioning
  - Accessible Chromatin (ATAC-seq)

function

Genome Sequencing

- The goal in genomics is to obtain the complete genomes and genetic variants of **all living organisms**.

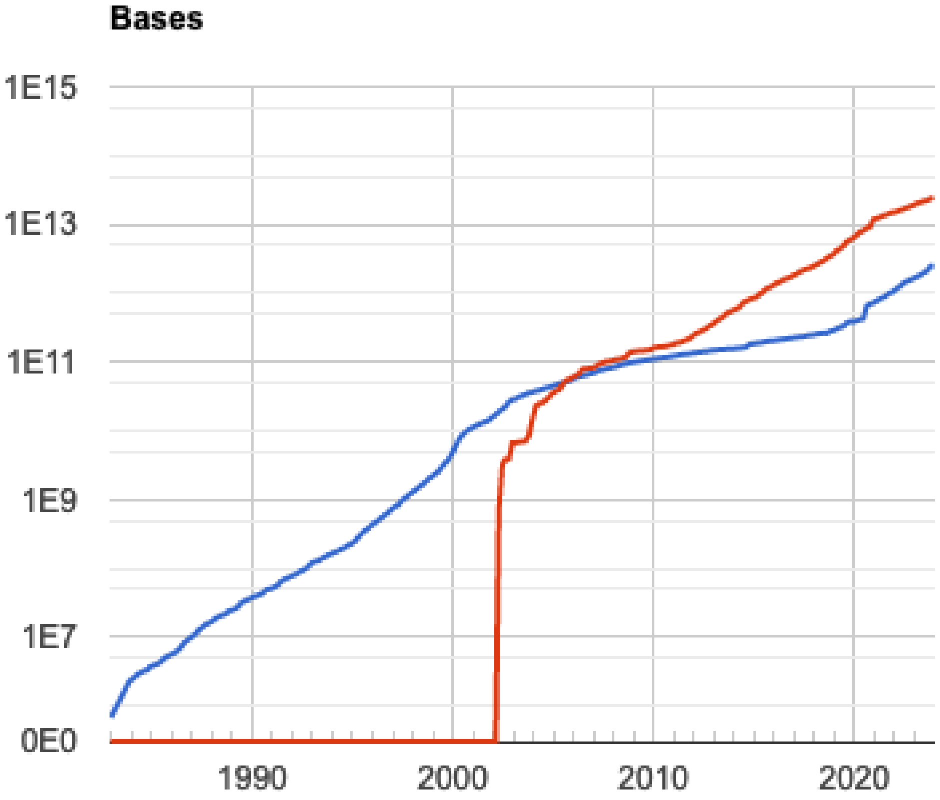
- **Next-generation sequencing** (NGS - HTS) technology has made an enormous contribution to our understanding of the relationship between single-nucleotide polymorphisms (SNPs) and various biological phenomena, including diseases and evolution.

- However, variant calling/gene prediction/genome structure is highly dependent on the **quality of the reference** genome, as it begins with mapping HTS reads onto the reference.

- Short-read sequencing technology (∼250 bp) makes difficult to analyze **large structural variants** and repetitive genomic regions, and it remains a challenge to assemble high-quality *de novo* genome.

- Advances in **long-read sequencing** technology have solved these problems by providing highly accurate and long (∼1 Mb) reads at reasonable costs.

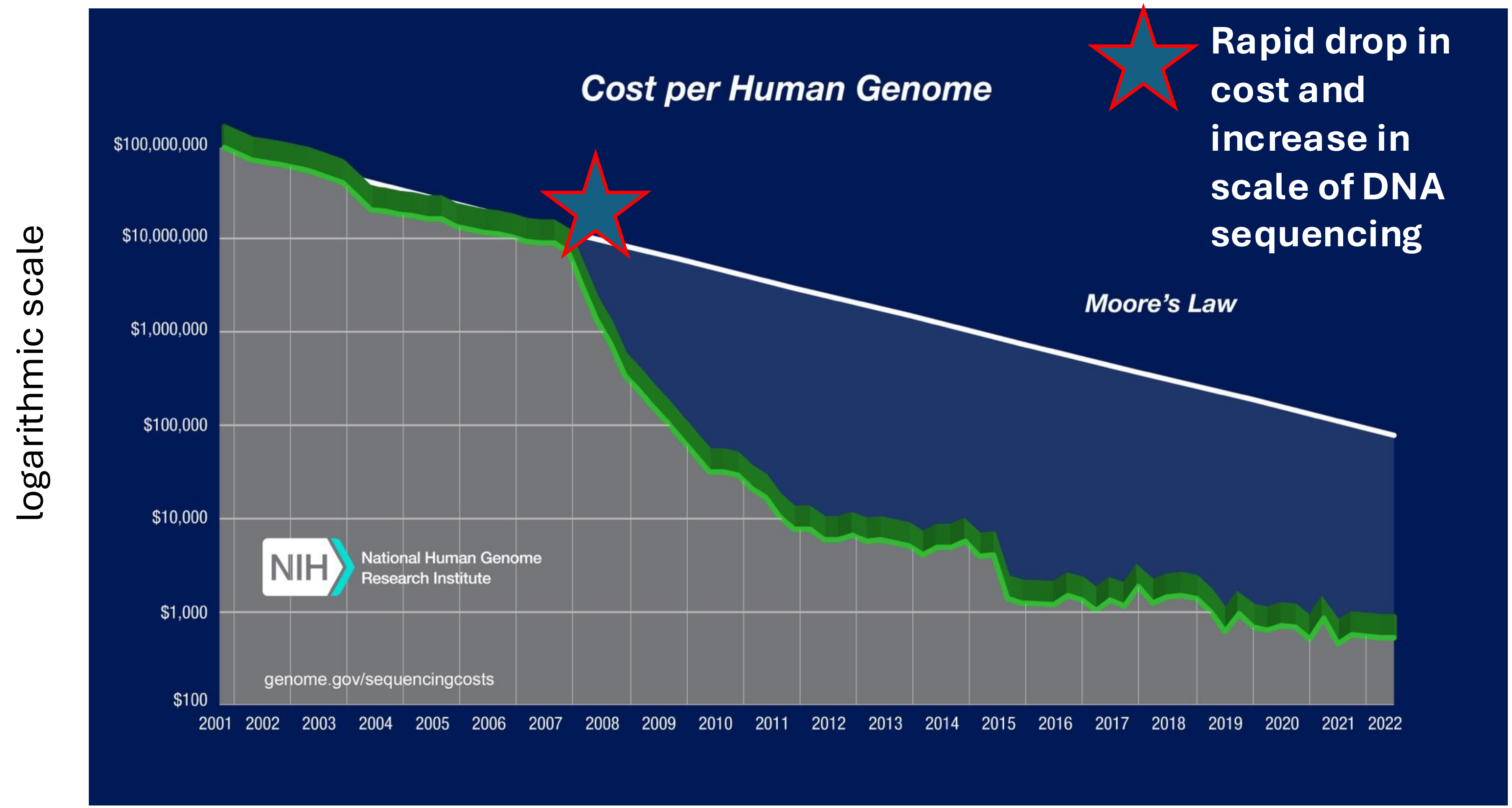
Genome analysis: The beginning



- Phage Phi-x174: 5kb, 1977.
- Bacteriophage lambda: 50kb, 1982.
- *Haemophilus influenzae*: 1.8Mb, 1995.
- Budding yeast *Saccharomyces cerevisiae*: 12Mb, 1996
- The roundworm *C. elegans*: 100Mb, 1998
- *Drosophila melanogaster* : 180 Mb 2000
- Human: 3.1Gb, 2003

https://www.ncbi.nlm.nih.gov/genbank/statistics/

Genome analysis: Big data



Rapid drop in cost and increase in scale of DNA sequencing

Moore's Law states that the number of components on a single chip *doubles every* two years at minimal cost

DNA Sequencing Approaches

- Chemical degradation of DNA
  - Maxam-Gilbert (1977) – obsolete

- Sequencing by synthesis (SBS)
  - Uses DNA polymerase in a primer extension reaction
  - Most common approach
  - First developed by Sanger (Sanger sequencing)
  - Illumina (aka Solexa), Pacific Biosciences (PacBio), Ion Torrent, 454

- Ligation bases
  - Sequencing using short probes that hybridize with template
  - SOLiD (Life Technologies)

- Other
  - Oxford Nanopore

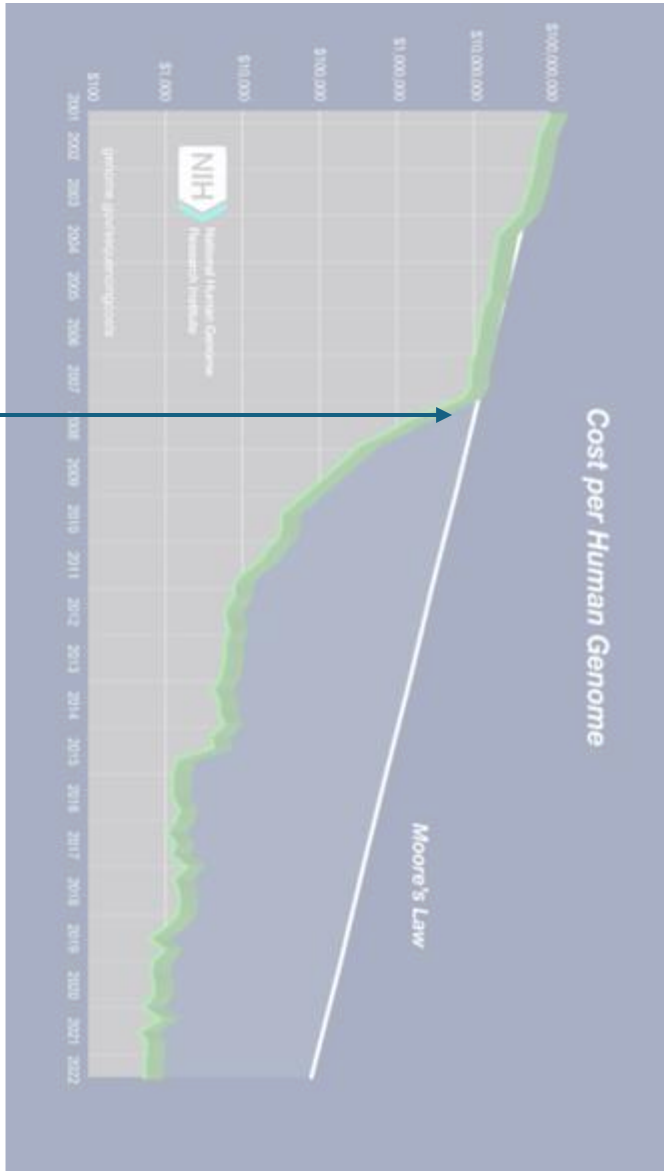
DNA Sequencing Approaches

- Chemical degradation of DNA
  - Maxam-Gilbert (1977) – obsolete

- Sequencing by synthesis (SBS)
  - Uses DNA polymerase in a primer extension reaction
  - Most common approach
  - First developed by Sanger (Sanger sequencing)
  - Illumina (aka Solexa), Pacific Biosciences (PacBio), Ion Torrent, 454

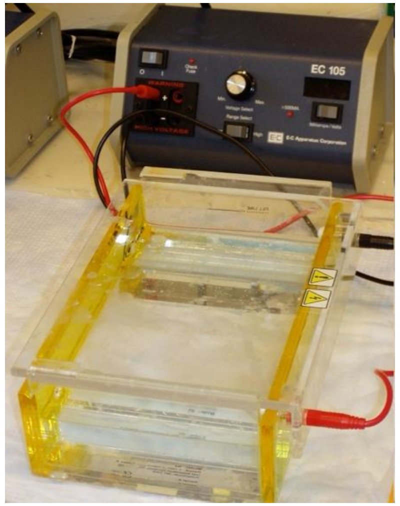
- Ligation bases
  - Sequencing using short probes that hybridize with template
  - SOLiD (Life Technologies)

- Other
  - Oxford Nanopore

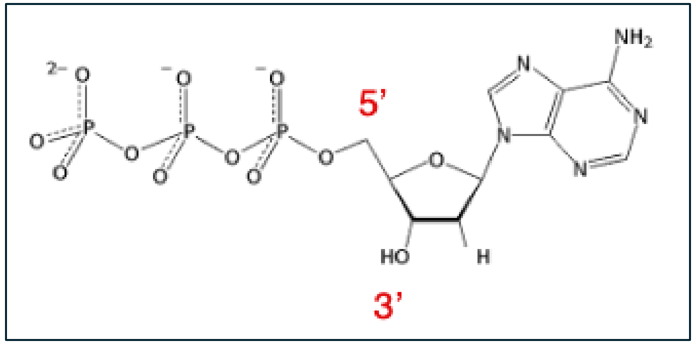


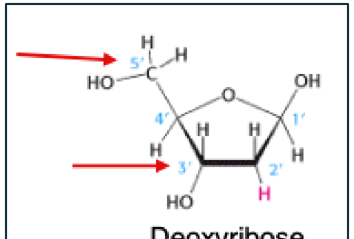
DNA Sequencing Machines

Although the sequencing principle is the same, there was/is tremendous progress in automating the collection of sequences.



Sequencing by synthesis (SBS)



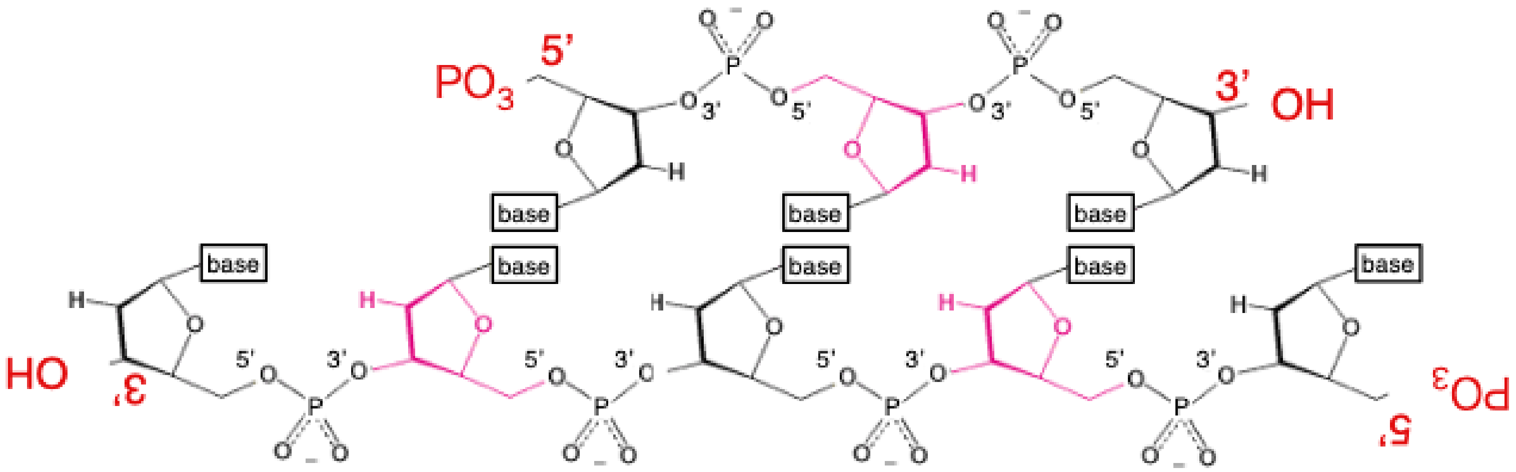


Deoxyribose

Base	plus sugar “nucleoside”
Adenine	Adenosine
Guanine	Guanosine
Cytosine	Cytidine
Thymine	Thymidine

in DNA: “deoxyadenosine”

plus triphosphate
“deoxynucleotide”
“2’-deoxyadenosine 5’-triphosphate” = dATP



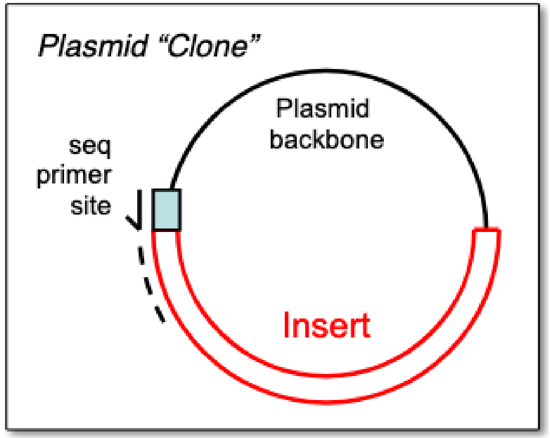
If I throw in DNA polymerase and free nucleotide, **which end gets extended**?"

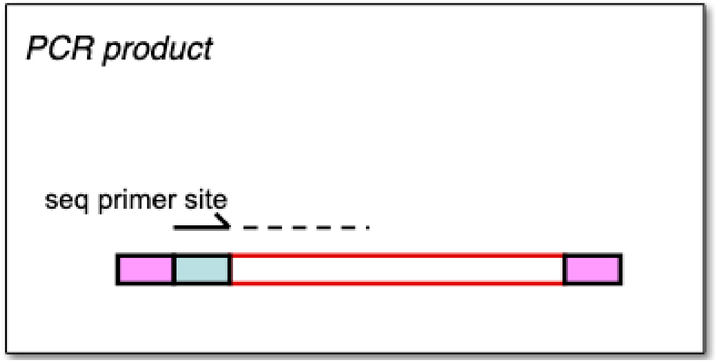
Source: *Biochemistry*, Berg et al.

Sanger Sequencing

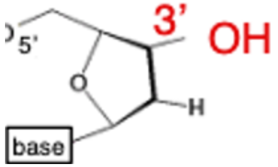
- Based on dideoxy terminator technique (Sanger 1977).
- Limits to stretches of approximately 1000 bases (or less).
- It uses specialized DNA polymerase enzymes to extend oligonucleotide primers bound to single-stranded DNA template.
- Primer walk approach: custom primers designed to be complimentary to extend nucleotide sequence data iteratively from the distal end of known regions into unknown regions.
- But serial generation of data in this manner is far too time consuming for large-scale sequencing.

Templates:



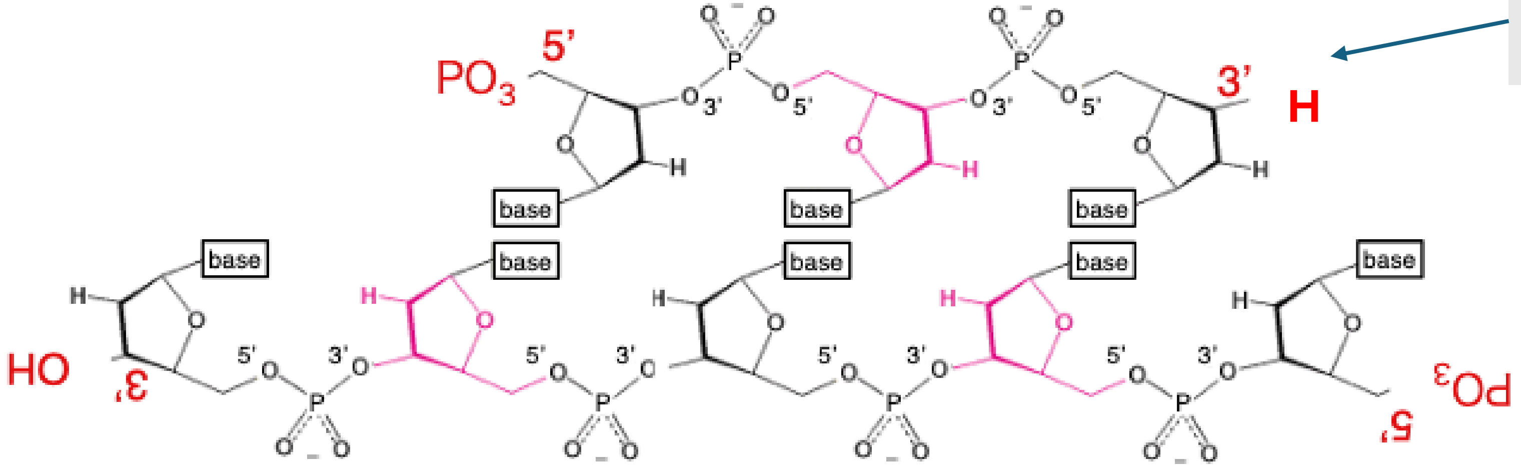


Sanger Sequencing



Deoxy
dNTP

Dideoxy nucleotides cannot be further extended, and so terminate the sequence chain



Dideoxy
ddNTP

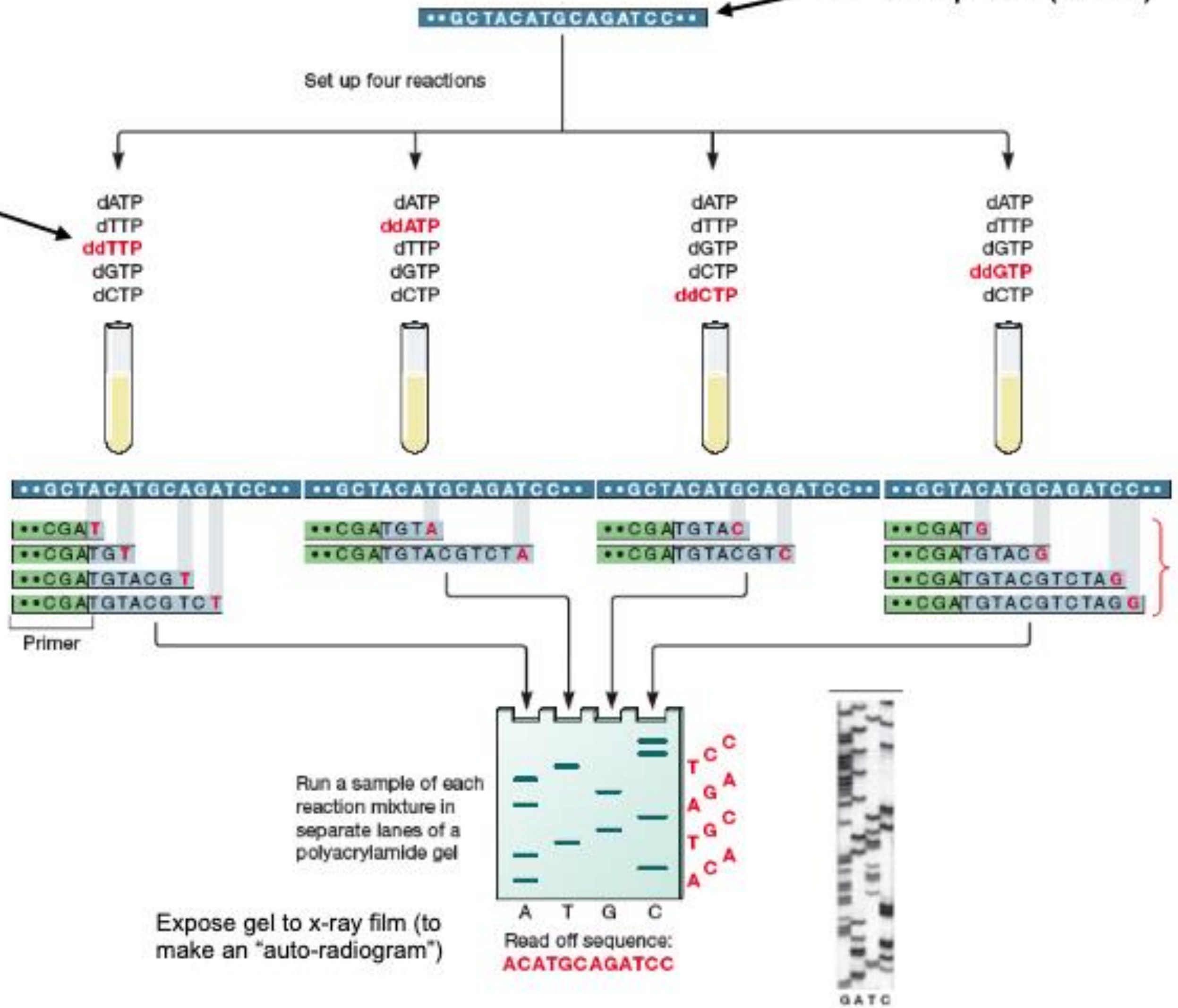
Source: *Biochemistry*, Berg et al.

Sanger Sequencing: Radioactive signal

Template (Crick)

very low concentration of ddNTPs compared to dNTPs

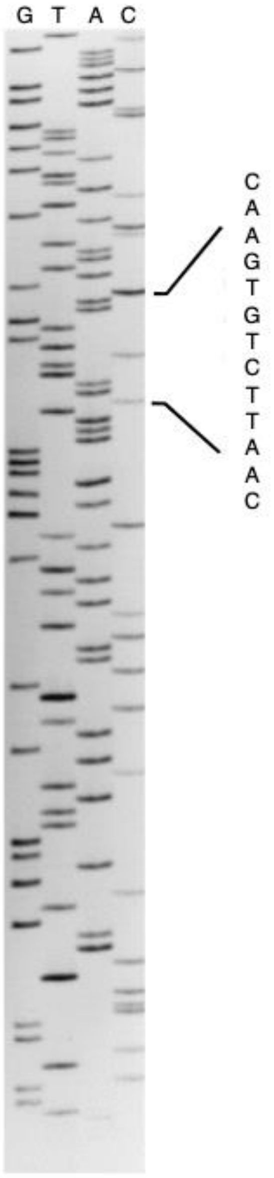
Watsons



A nested series of DNA fragments ending in the based specified by the terminator-ddNTP

Source: *Gene and Genomes*, Singer.

Sanger Sequencing



Nod bad, but:

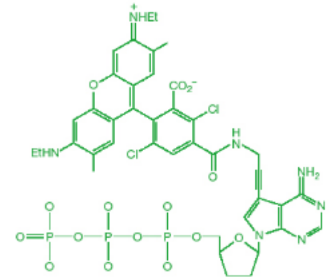
- Wouldn’t it be great to run everything in **one lane**?
    - Save space and time
    - more efficient

- Unable to read sequence near the top, as the bands get closer and closer together.

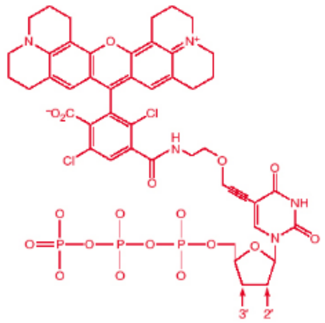
- **Fluorescently label** the ddNTPs so that they each appear a different color, and can be read by a laser at a fixed point.

A T C G

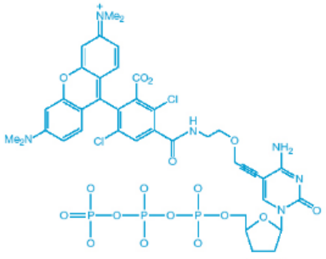
Sanger Sequencing



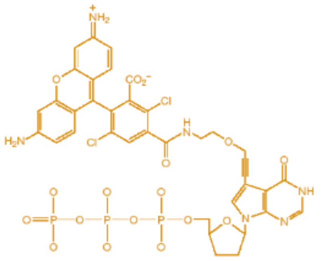
ddATP



ddTTP



ddCTP



ddGTP

Nod bad, but:

- Wouldn’t it be great to run everything in **one lane**?
  - Save space and time
  - more efficient

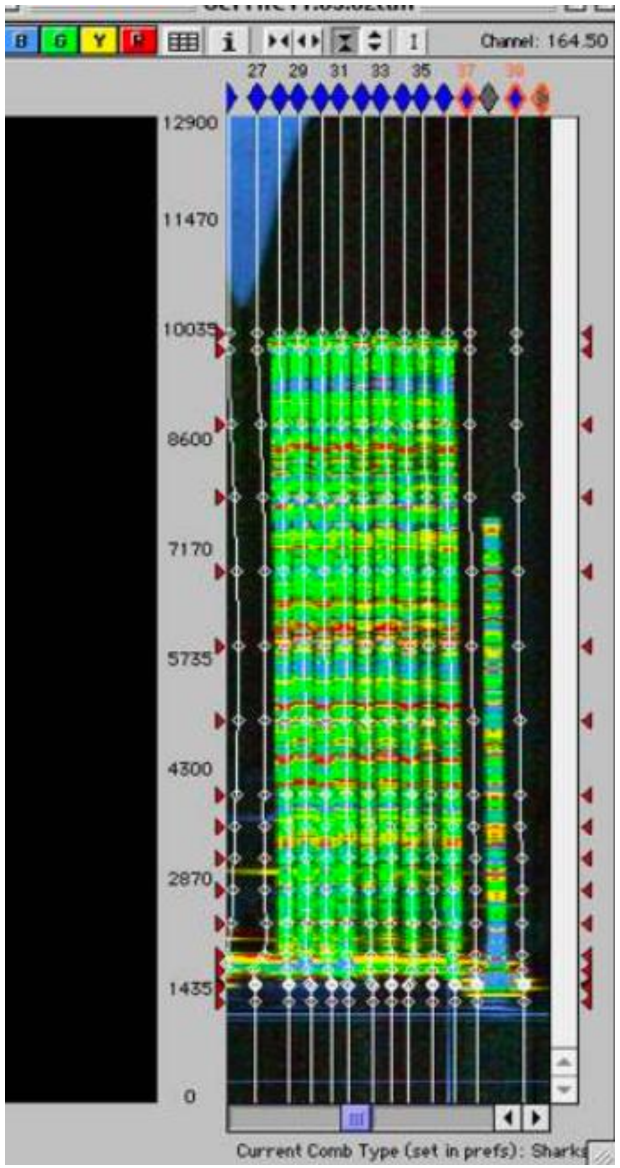
- Unable to read sequence near the top, as the bands get closer and closer together.

- **Fluorescently label** the ddNTPs so that they each appear a different color, and can be read by a laser at a fixed point.

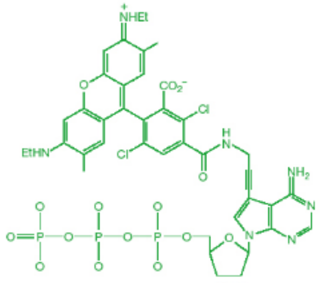
A T C G

But the principle is the same: shorter fragments run faster.

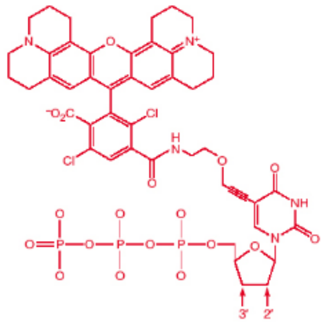
Sanger gel



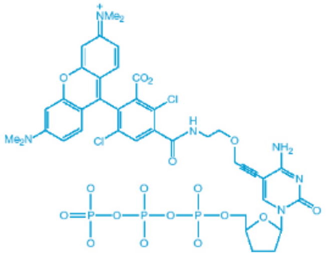
Sanger Sequencing



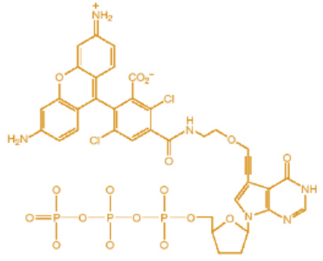
ddATP



ddTTP



ddCTP



ddGTP

Fluorescent Sanger Sequencing:

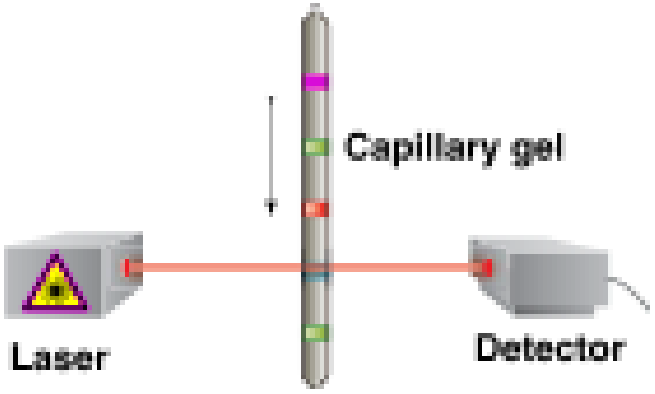
- Lane/capillary signal: real fluorescent signal

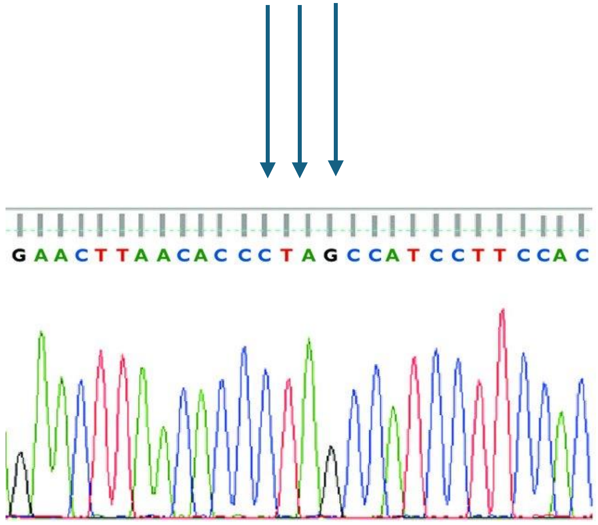
Various **algorithms**:
- boost signal/noise
- correct for dye-effects,
- Mobility differences
- Generates the 'final' trace

- Trace: shows DNA sequence

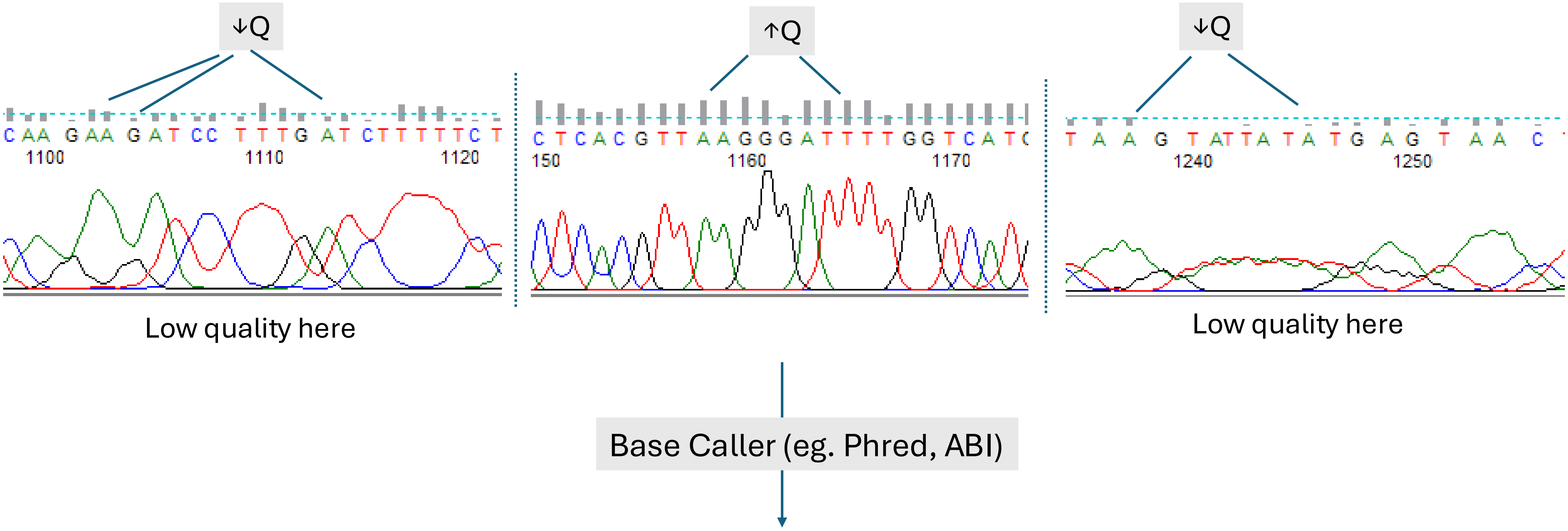
A T C G

But the principle is the same: shorter fragments run faster.





Sanger Base Calling



Low quality here

Low quality here

Base Caller (eg. Phred, ABI)

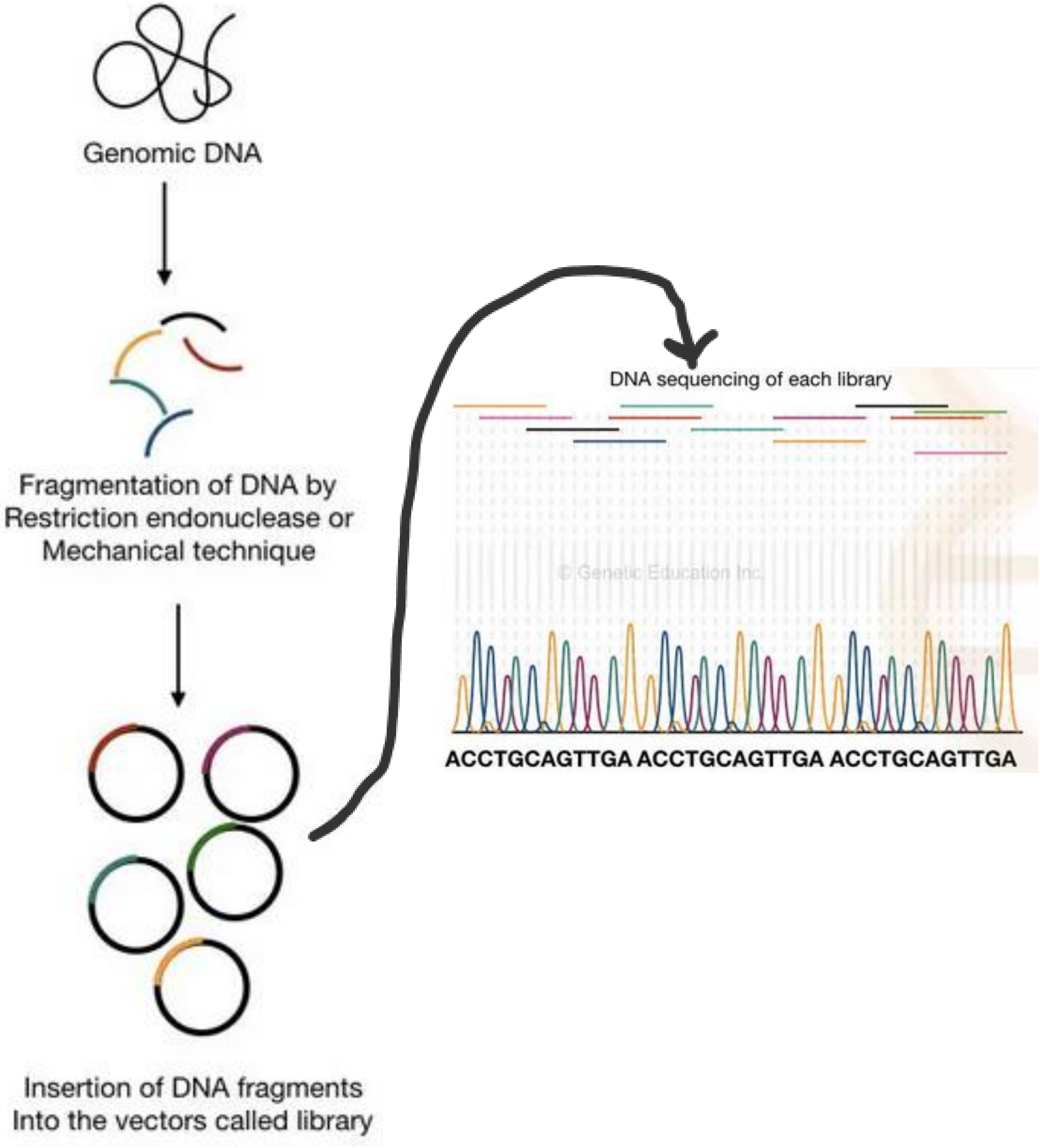
Quality score = -10log$_{10}$(probability of error)

For **Q20**, probability of error = 0.01 or 1 in 100. Base call accuracy 99%.

For **Q40**, probability of error = 0.0001 or 1 in 10,000. Base call accuracy 99.99%.

For Q99, probability of error ~10^{-10}

Whole Genome Shotgun


Genomic DNA
Fragmentation of DNA by Restriction endonuclease or Mechanical technique
Insertion of DNA fragments Into the vectors called library
DNA sequencing of each library
© Genetic Education Inc.
ACCTGCAGTTGA ACCTGCAGTTGA ACCTGCAGTTGA

Some day in the future, technology would be able to determine end-to-end nucleotide sequences. But during the Sanger sequencing early days:

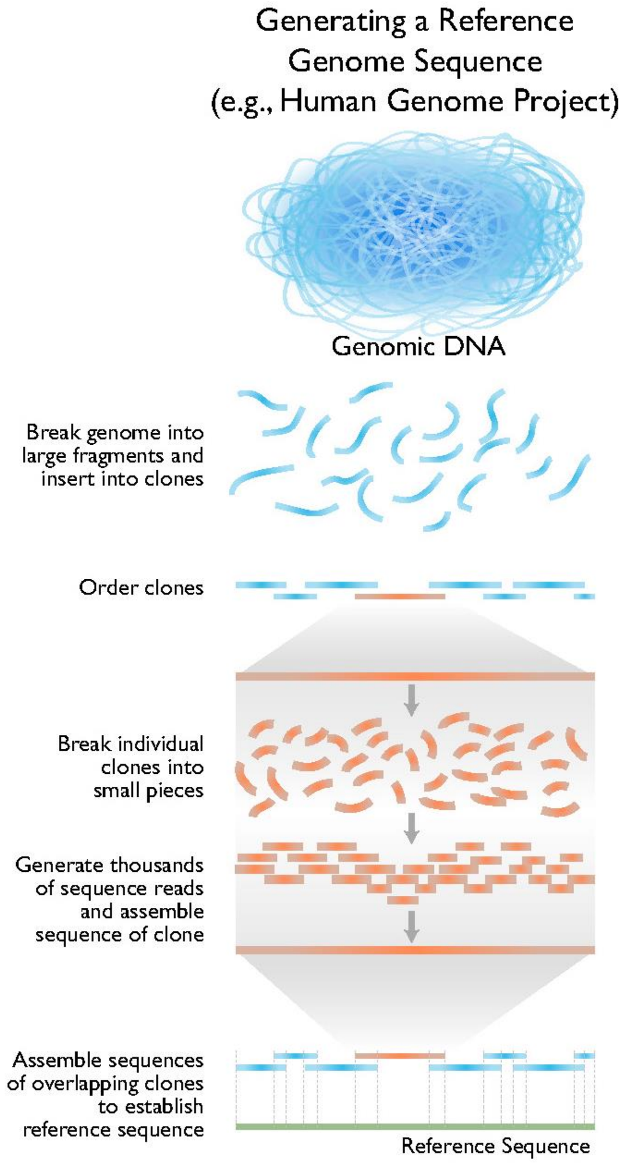
- In the shotgun sequencing method, DNA to be sequenced is sheared randomly into size fragments that are ligated into specially designed sequencing vectors.

- The vectors provide a mechanism for clonal amplification of each fragment by allowing for replication in a suitable host (bacteria).

- The generation of many of these cloned fragments in parallel lends itself to a highly efficient DNA sequence production.

The Human Genome Project

- Done by shotgun sequencing
- $3 billion project was formally founded in 1990
- Declared complete in April 2003
- Genome size 3.1 Gb


Generating a Reference Genome Sequence (e.g., Human Genome Project)
Genomic DNA
Break genome into large fragments and insert into clones
Order clones
Break individual clones into small pieces
Generate thousands of sequence reads and assemble sequence of clone
Assemble sequences of overlapping clones to establish reference sequence
Reference Sequence

Source: https://www.genome.gov/

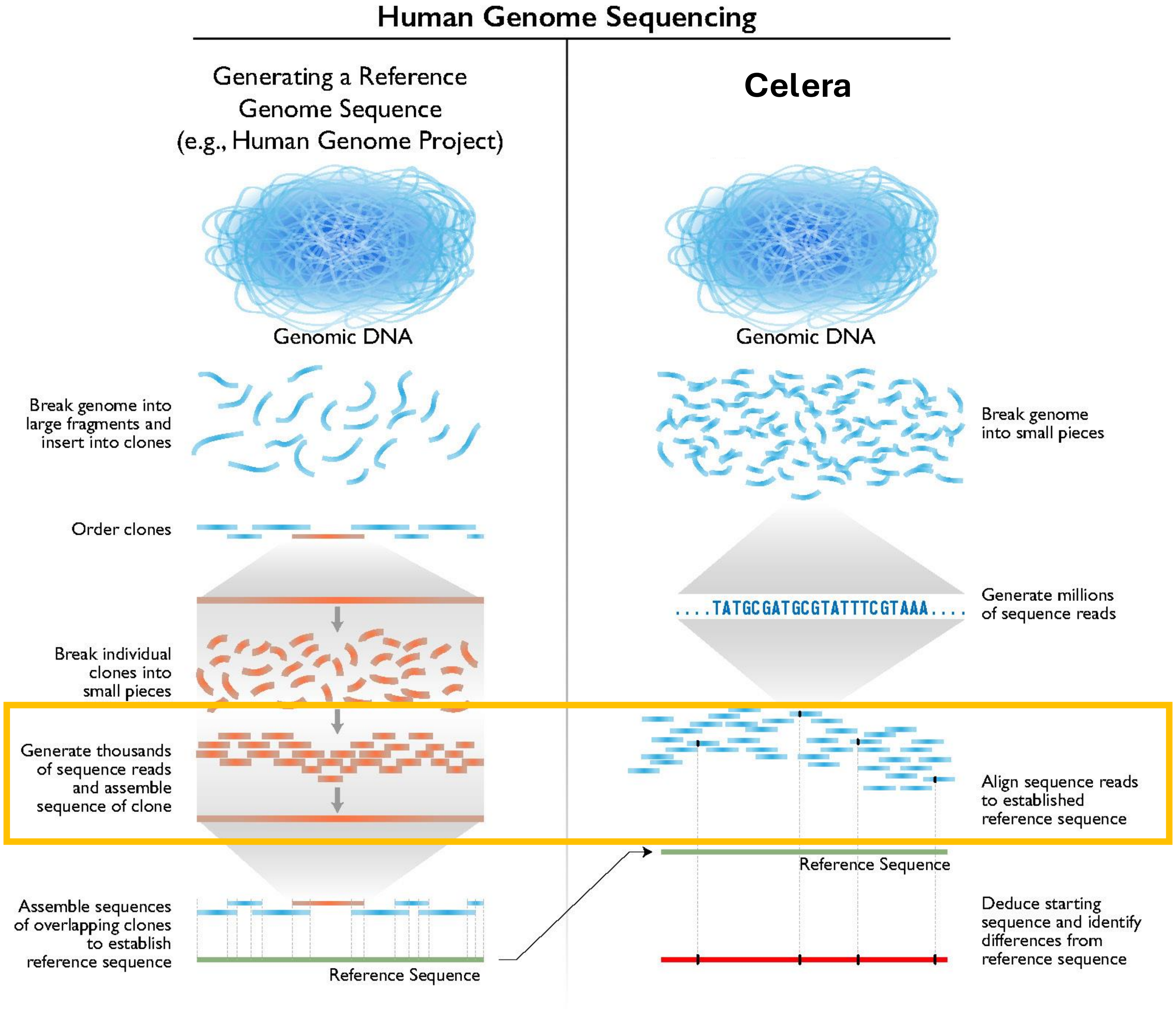
The Human Genome ~~Project~~ Battle

Public consortium:

- $3 billion
- Hierarchical approach (Lander et al. 2001)
- Initial step of clone-based physical mapping to generate a list of overlapping BAC clones to be sequenced ($$$$)
- Followed by shotgun sequencing of the individually mapped BAC clones (100 kb – 200 kb)

Celera:

- $300 million
- It was a proof-of-concept
- Whole genome shotgun approach
- No clones: generates data more quickly
- Direct sequencing of randomly sheared genomic DNA ($$).
- Libraries insert size: 2, 10 and 50 Kb
- Computationally expensive: Relies on sophisticated assembly algorithms.


Human Genome Sequencing
Generating a Reference Genome Sequence (e.g., Human Genome Project)
Celera
Genomic DNA
Genomic DNA
Break genome into large fragments and insert into clones
Break genome into small pieces
Order clones
Generate millions of sequence reads
....TATGCGATGCGTATTTCGTAAA....
Break individual clones into small pieces
Generate thousands of sequence reads and assemble sequence of clone
Align sequence reads to established reference sequence
Reference Sequence
Assemble sequences of overlapping clones to establish reference sequence
Deduce starting sequence and identify differences from reference sequence
Reference Sequence

Celera used more computational power and more sophisticated (proprietary) algorithms, but it was based on the same algorithm the Human Genome Project was using.

Computational issues arise from alignment-based assembly.

The Human Genome Project

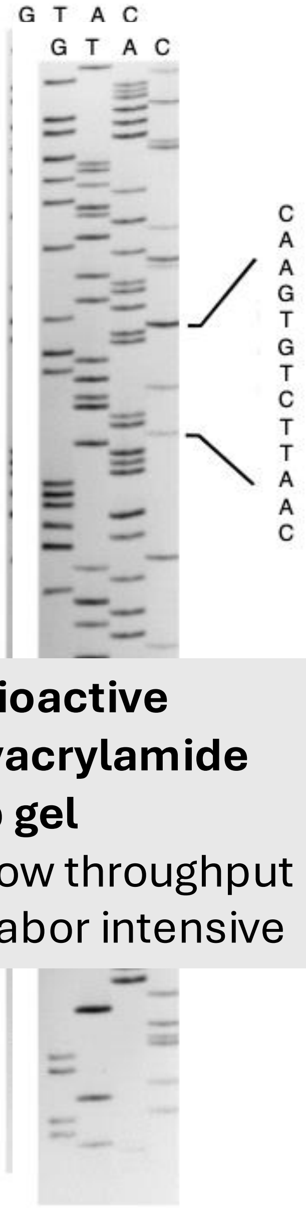
- However, both approaches have advantages.

- Genome sequencing of mouse, rat, and zebrafish was done with a hybrid approach of both clone-by-clone and whole genome shotgun sequencing.

- Shotgun sequencing was successful at determining the sequence of smaller segments of DNA.

- The assembly of shotgun reads from larger genomes is still a challenge.

Progress of Sanger Sequencing Technology



Radioactive polyacrylamide slab gel
- Low throughput
- Labor intensive

AB slab gel sequencers
(370, 373, 377)
Fluorescent sequencing
1990-1999
6 runs/day
96 reads/run
500 bp/read
288,000 bp/day

AB capillary sequencers
(3700, 3730)
1998-now
24 runs/day
96 reads/run
550 – 1,000 bp/read
1-2 million bp/day

Progress of Sequencing Technology

Sanger

~1,000-fold increase in throughput since 1985 accomplished by incremental improvements of the same underlying technology

THE NEXT GENERATION®

2nd Generation Sequencing Technologies

~500 - 30,000x more throughput than Sanger:
Illumina, SOLiD, 454 Pyrosequencing
+ PacBio, Ion Torrent, Nanopore

Illumina

- Illumina short read technology is currently the most popular sequencing technology (~ 90% market share).

- Uses **sequencing by synthesis** (SBS) technology relying on four fluorescently-labelled nucleotides to sequence the tens of millions of clusters on the flow cell surface in parallel.

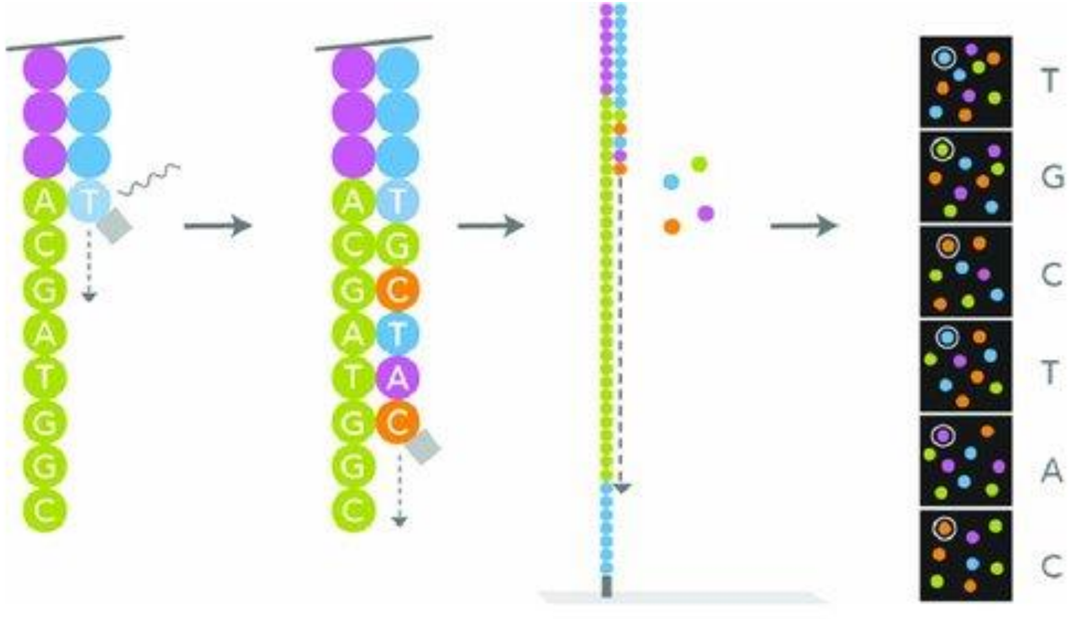
- During each sequencing cycle, a single labeled nucleotide is added to the nucleic acid chain.

- The nucleotide label serves as a terminator for polymerization. After each nucleotide incorporation:
  - the fluorescent dye is imaged to identify the base.
  - then enzymatically cleaved to allow incorporation of the next nucleotide.

- This allows controlling the addition of nucleotides one at a time, which minimizes incorporation bias.

- Base calls are made directly from **signal intensity measurements** during each cycle, which significantly reduces raw error rates compared to other technologies.

Reversible terminator technology



Illumina uses a sequencing concept that is similar to the chain termination procedure used in Sanger sequencing:

- The strand elongation is stopped after the incorporation of a fluorescently labelled base that prevents further strand elongation and the label of the incorporated base is read out to reveal sequence information.

- However, contrary to Sanger (irreversible), in the **reversible terminator technology**, termination is reversible, and the sequence is determined in real-time at the moment of incorporation of the fluorescently labeled bases.

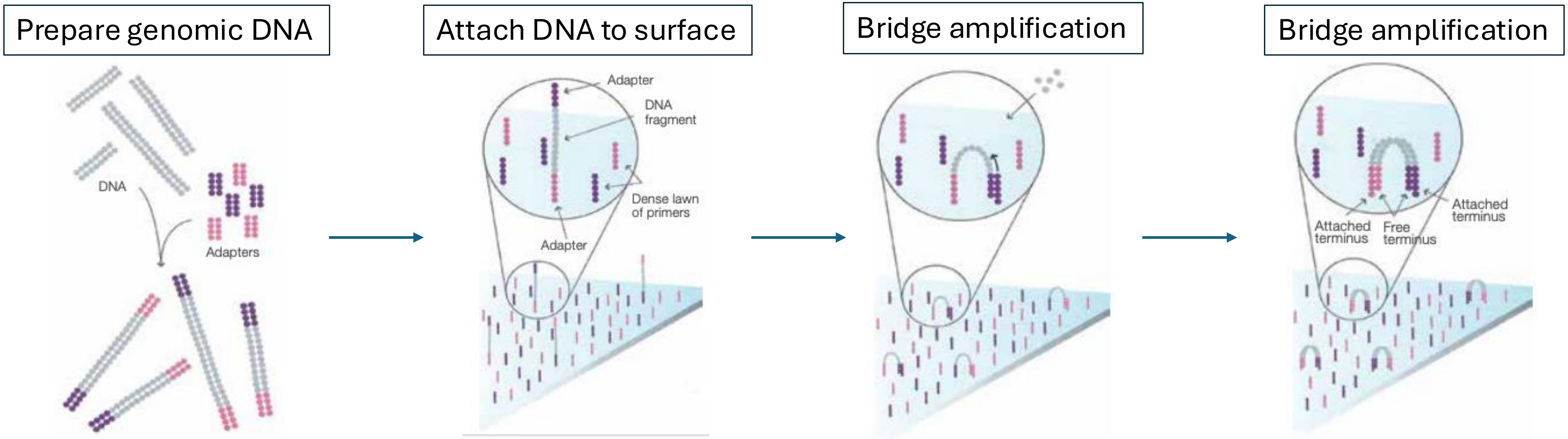
Illumina Cluster Generation and Bridge Amplification Process

- **DNA Fragment Preparation**: DNA is sheared and tagged with adapters.

- **Bridge Amplification**: DNA fragments undergo cycles of amplification, where each strand is complemented and folded over to attach to another receptor, creating two strands. This is repeated to form large DNA clusters.

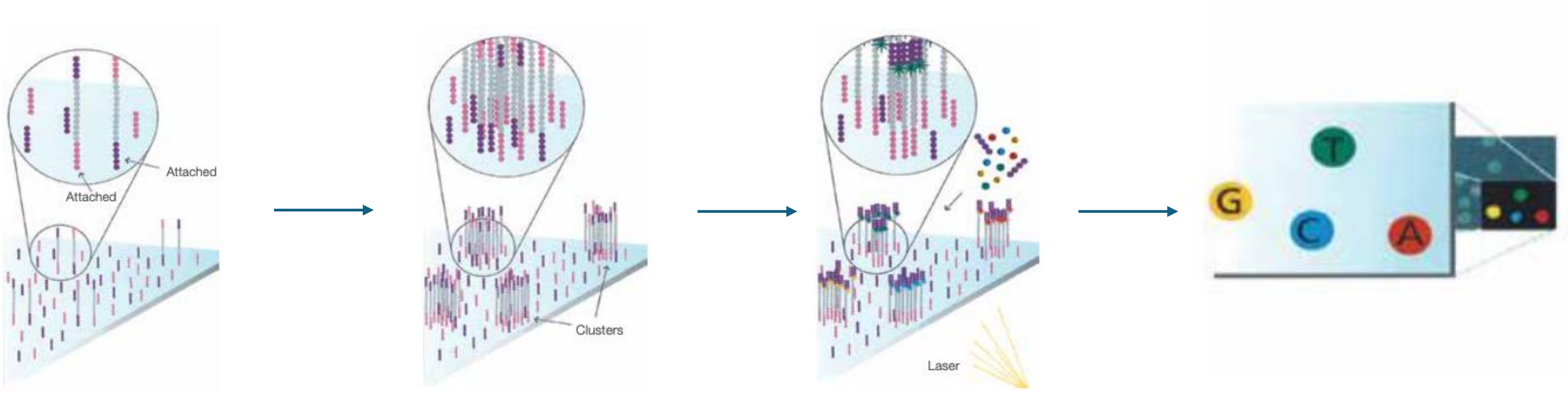
- **Sequencing by Synthesis**: Complementary base pairs are added and recorded over multiple cycles.

- **Read Error Rate**: Technology achieves a low read error rate of less than 1%.



Prepare genomic DNA → Attach DNA to surface → Bridge amplification → Bridge amplification

Illumina Cluster Generation and Bridge Amplification Process



Denature dsDNA

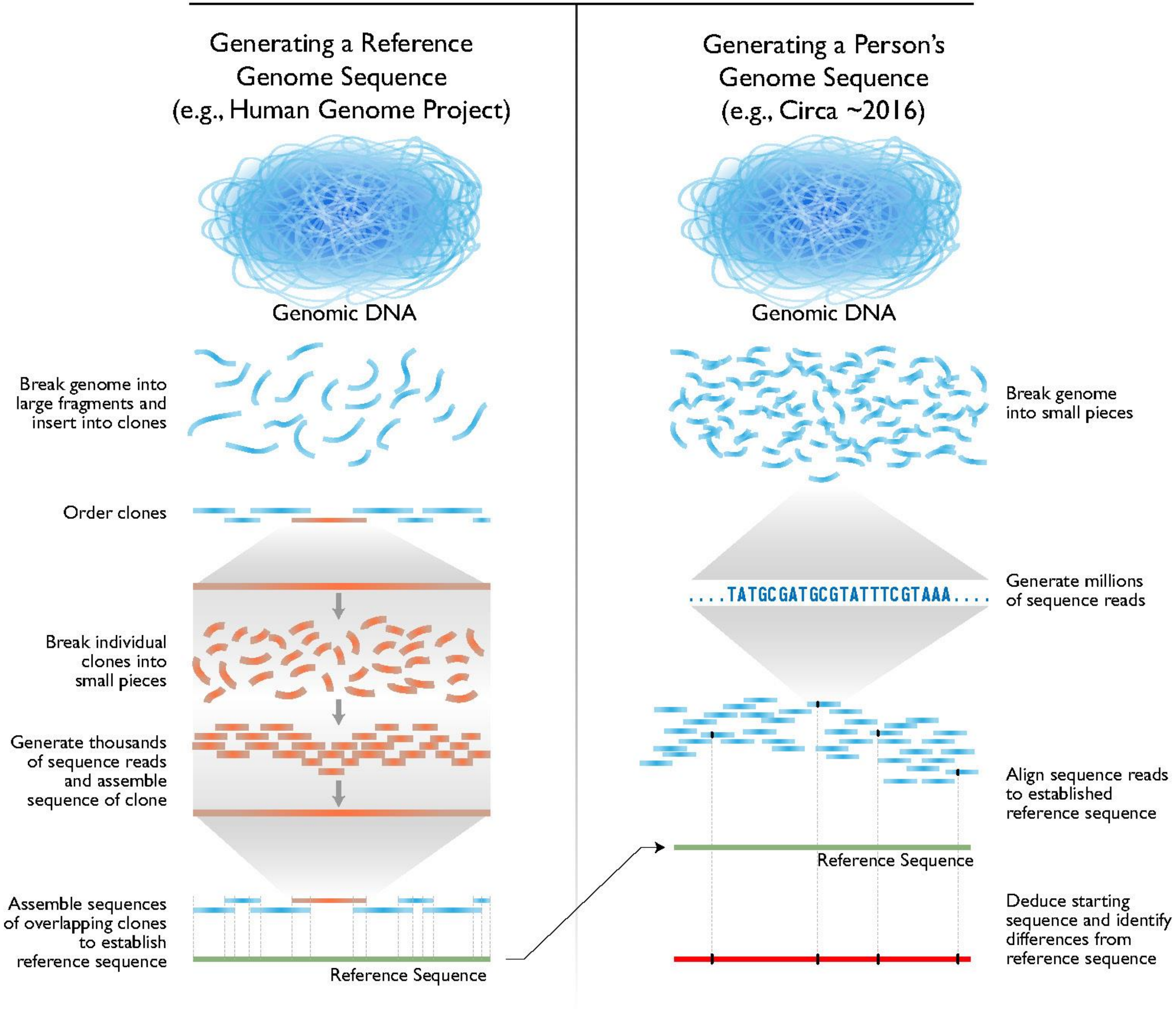
Complete amplification

Sequencing cycles

Image collection

https://youtu.be/fCd6B5HRaZ8

Human Genome Sequencing



Generating a Reference Genome Sequence (e.g., Human Genome Project)

Genomic DNA

Break genome into large fragments and insert into clones

Order clones

Break individual clones into small pieces

Generate thousands of sequence reads and assemble sequence of clone

Assemble sequences of overlapping clones to establish reference sequence

Reference Sequence

Generating a Person's Genome Sequence (e.g., Circa ~2016)

Genomic DNA

Break genome into small pieces

Generate millions of sequence reads

....TATGCGATGCGTATTTCGTAAA....

Align sequence reads to established reference sequence

Reference Sequence

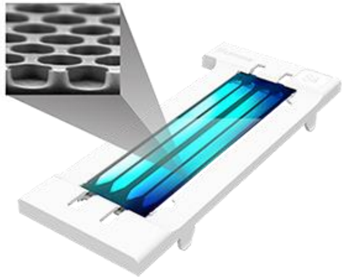
Deduce starting sequence and identify differences from reference sequence

Advantages and limitations of Illumina Technology

- **Advantages**:
  - Affordable and efficient for high-throughput sequencing.
  - It is fast (now): NovaSeq 2x (Paired End) 250 bp can generate 400 Gb in 38 hours!
  - Low read error rate of less than 1%. This means the polymerase has a lower substitution rate while amplifying the DNA.

Genomics Enterprise Center @RIT

- **Limitations**:
  - Short Read Length: the size of fragments that can be sequenced is approximately 250 bp.
  - Sequence Bias: Some genome regions are poorly sequenced, even with high coverage.
  - Homopolymers errors: Low complexity regions and repeats. Especially bad in AT-rich regions.



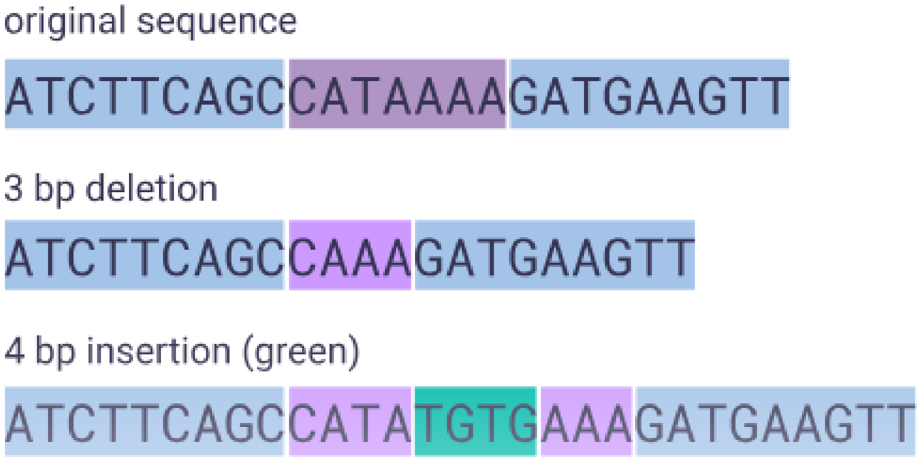
NovaSeq flow cell

Advantages and limitations of Illumina Technology

Homopolymers errors:

Homopolymers are arrays of identical nucleotides. "ATTTTTTGC", for example, has a homopolymer of length 6 (composed of "T" base).

These type of sequences will create insertions or deletions (**indels)**.



original sequence

ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (green)

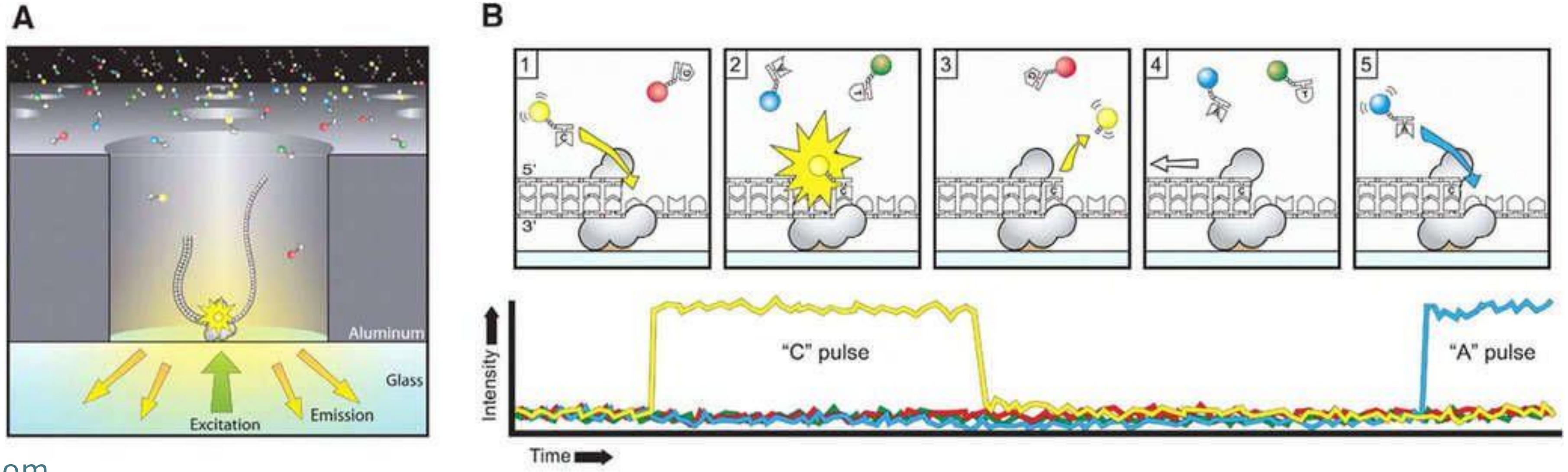
ATCTTCAGCCATATGTGAAAGATGAAGTT

PacBio SMRT Technology

- Single Molecule Real-Time (SMRT) DNA Sequencing performs real-time reading of the entire DNA molecule as it passes through a specialized sequencing apparatus.

- It sequences via light pulses. Compared to Illumina, this system is more sensitive optically, allowing for real-time sequencing.

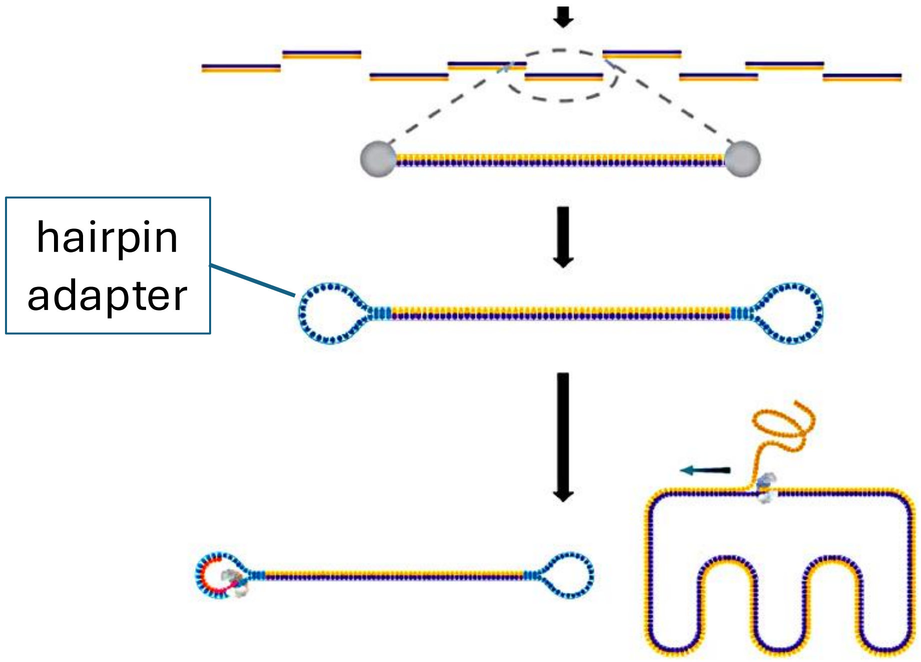
- With the **zero mode waveguide** *(*ZMW) technology, when a base is integrated, the pulse of light generated is about 10000 times as long as the baseline. The timings between pulses (corresponding to bases) are not regular, however, due to variable kinetics as a function of local DNA chemistry. This allows one to see chemical modifications of DNA without any preparation ahead of time.

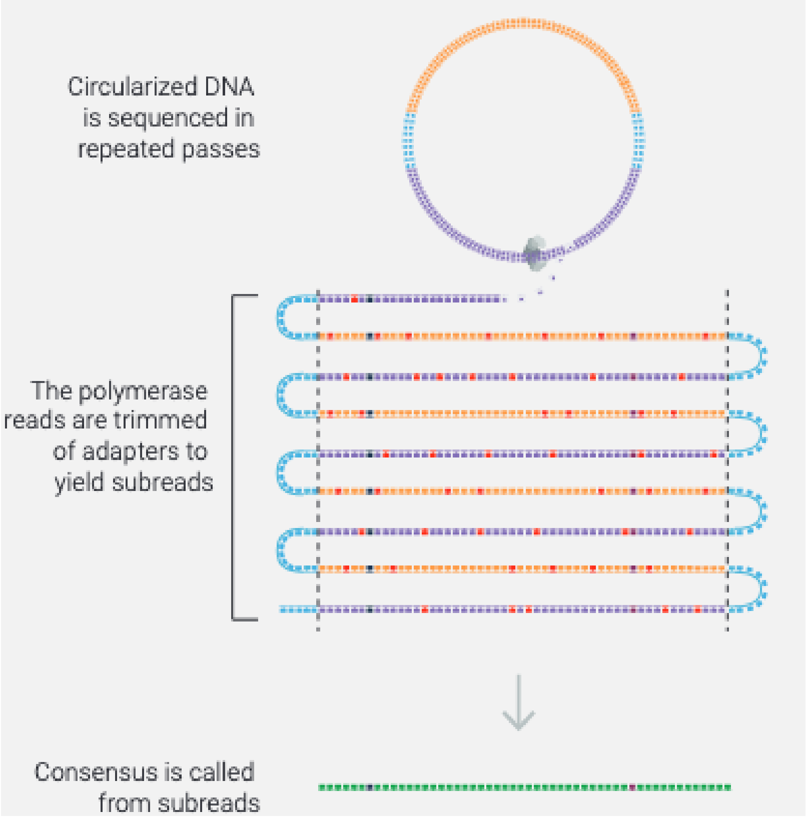


Source:
https://www.pacb.com

PacBio SMRT Technology

- PacBio uses **hairpin adapters**. The polymerase loops around during synthesis, displacing strands they encounter.


hairpin adapter


Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads

- Reads are generated by combining multiple consecutive observations of a DNA molecule (subreads), increasing accuracy.

Source: https://www.pacb.com

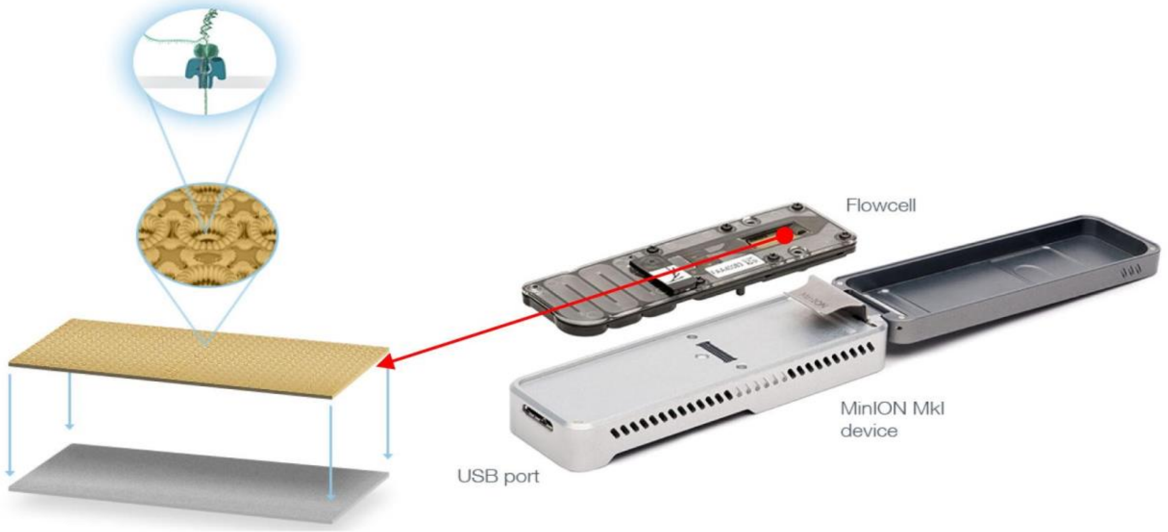
PacBio SMRT Technology

- **Advantages**:
  - PacBio SMRT technology is used to generate reads with a mean size of 10 - 15 kilobases, with a maximum size of 30 kilobases.
  - It Reads a single DNA molecule at a time.
  - Longer reads allow us to detect more structural variants (indels, repeats).
  - The errors are more stochastic and less biased (eg. Illumina). Solution: generate consensus sequences with subreads.
- **Limitations**: Affordable
  - Reads have a high error rate.
  - Not at affordable as Illumina:
    - Library preparation
    - Need a bigger amount of (high quality) DNA input
    - Average cost per Gb sequenced

PacBio cell

Oxford Nanopore Technology

- Similar to PacBio, Oxford Nanopore technology is able to read a single base pair at a time.
- DNA sequencing is performed by adding the sample to the flow cell.
- Identifies DNA bases by measuring the changes in electrical conductivity generated as DNA strands pass through a biological pore.
- The data streams are passed to the software that generates the signal-level data. This technology can probe very long lengths of DNA.



Oxford Nanopore Technology

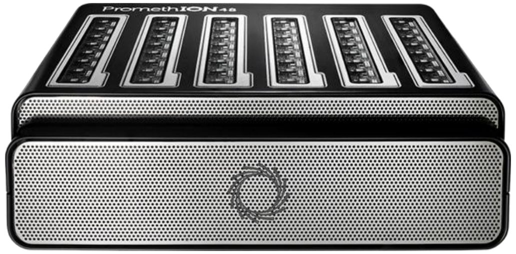
- Entered the market with the MinION, capable of sequencing one sample.

- Error rate (2-10%) is currently even higher than PacBio

- The MinION has portability, affordability, and speed in data production, which makes it suitable for real-time applications (it is portable!).

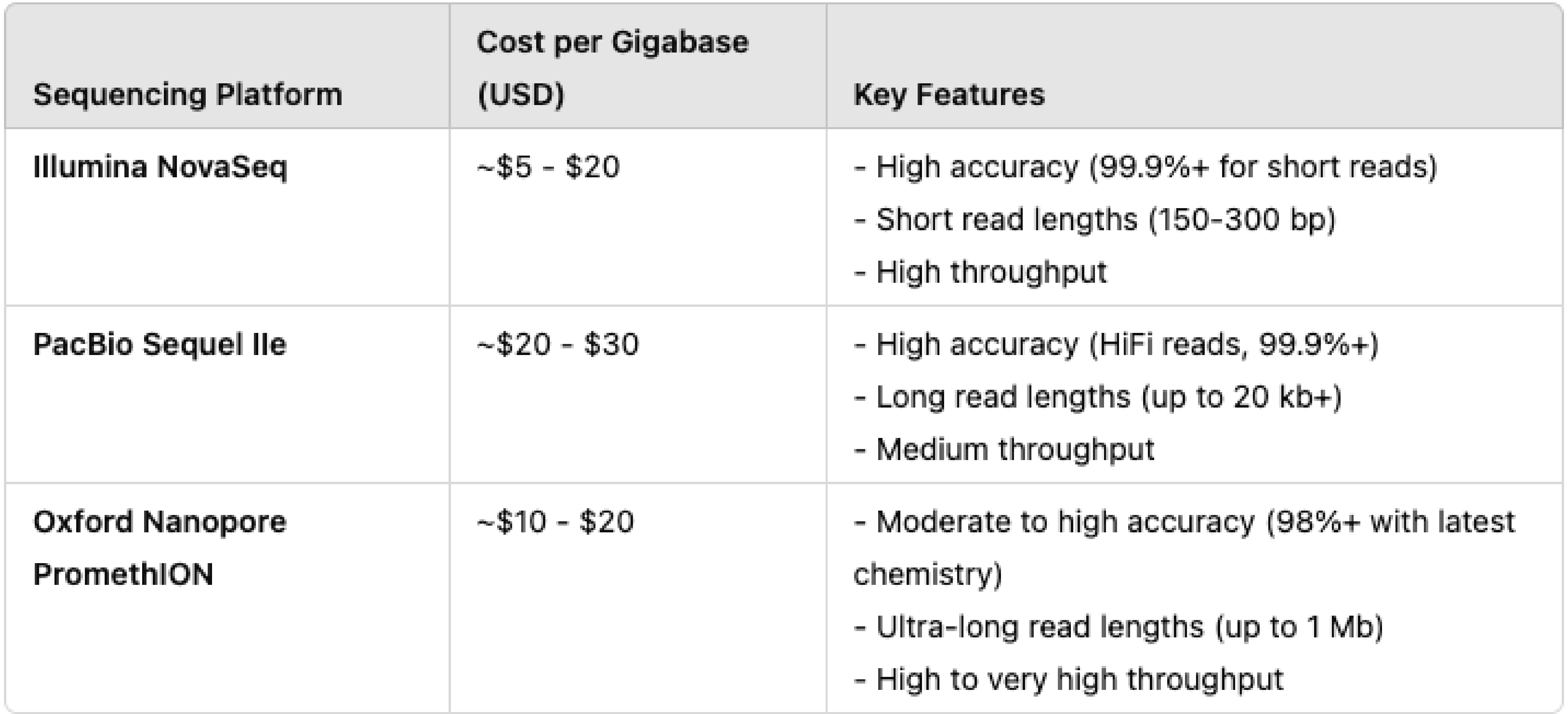
- A Nanopore MinION typically generates around 10-50 Gigabases (Gb).

- The new Nanopre device: PromethION
  - PromethION can generate up to 100 Gb of sequence data per flow cell and can run up to 24 flow cells simultaneously.
  - Sequencing of long DNA fragments (> 50 kb).

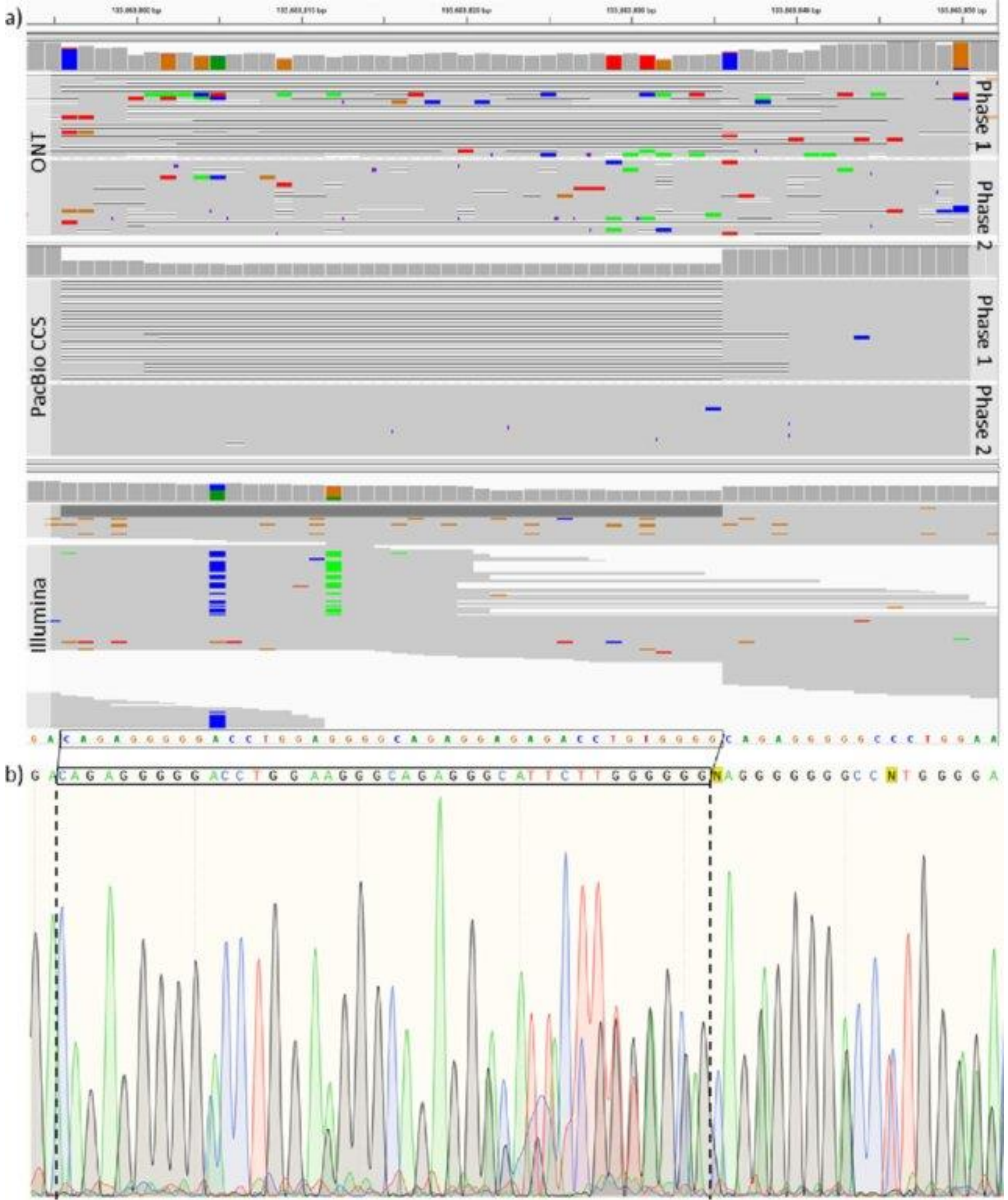


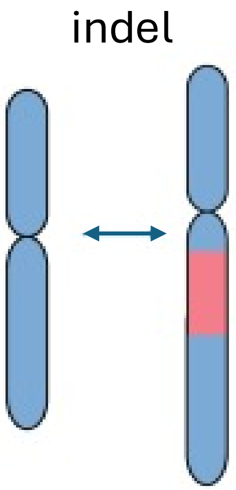
Choosing technologies

Choosing technologies



Sequencing Platform	Cost per Gigabase (USD)	Key Features
Illumina NovaSeq	~$5 - $20	- High accuracy (99.9%+ for short reads) - Short read lengths (150-300 bp) - High throughput
PacBio Sequel IIe	~$20 - $30	- High accuracy (HiFi reads, 99.9%+) - Long read lengths (up to 20 kb+) - Medium throughput
Oxford Nanopore PromethION	~$10 - $20	- Moderate to high accuracy (98%+ with latest chemistry) - Ultra-long read lengths (up to 1 Mb) - High to very high throughput

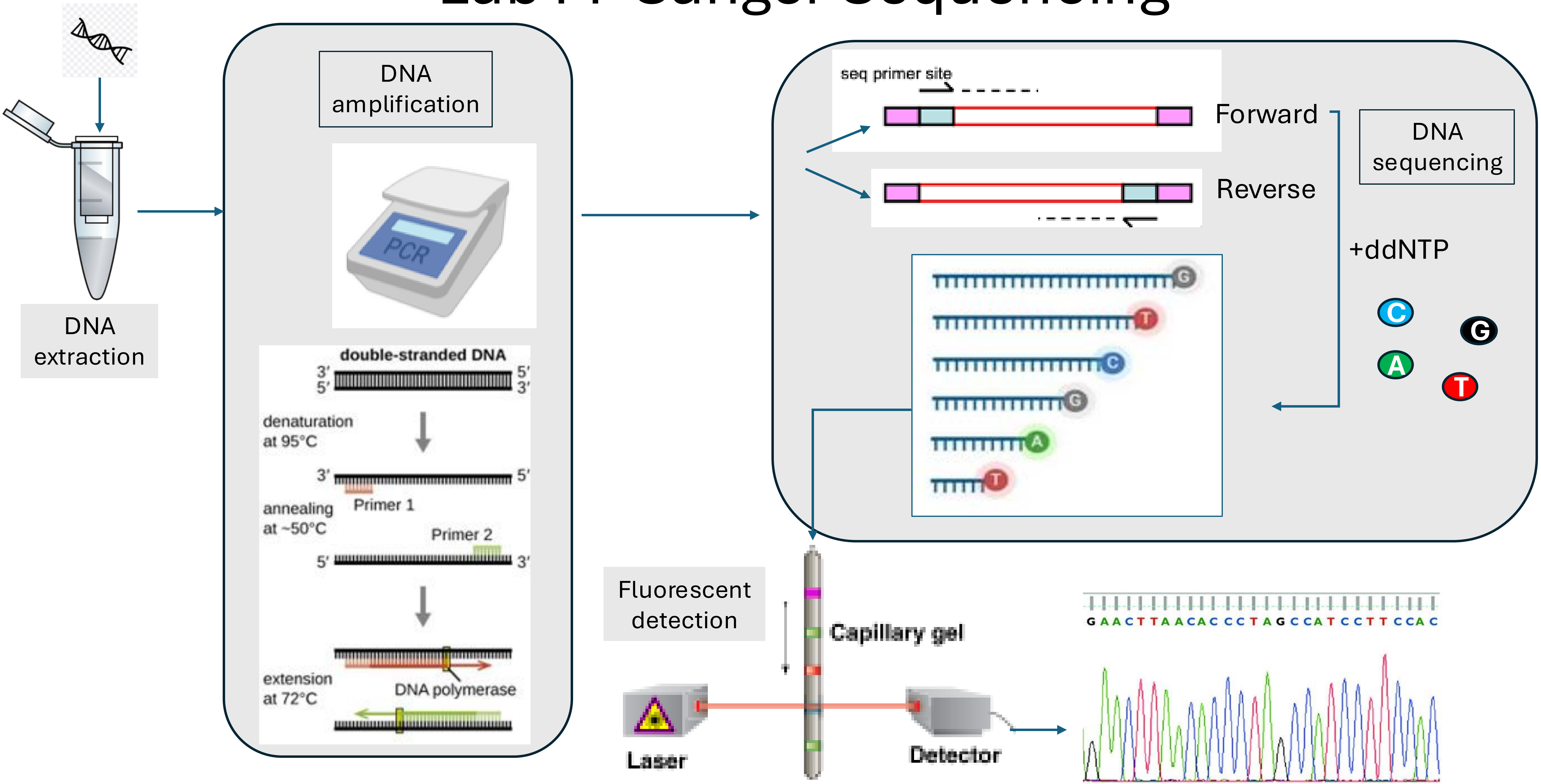




Evidence for novel deletions:
chr9:135663780-135663850
40-bp-long deletion (black box)

Source: 10.1186/s13059-021-02472-2

Lab11-Sanger Sequencing



DNA extraction

DNA amplification

DNA sequencing

Forward

Reverse

+ddNTP

Fluorescent detection