Lecture 2

Introduction to bioinformatics (MVE510)

Autumn, 2020

Read *Ten years of next-generation sequencing technology,* van Dijk EL, Auger H, Jaszczyszyn Y, and Thermes C, Trends in Genetics, 30(10) 2014.

Repetition

- R is a statistical programming language that you will use throughout this course
- R is the most used tool for analysis of complex biological datasets
- Computer exercise 1 aims to introduce you to R
- Rstudio/R installed on the computer systems. Can also be downloaded to your own computer for free

Todays agenda

- The history of DNA sequencing
- From serial to ‘massively parallel’
- Next generation sequencing platforms
    - Massively parallel pyrosequencing
    - Illumina
    - Pacific bioscience (PacBio)
    - Oxford Nanopore
- Applications of next generation sequencing

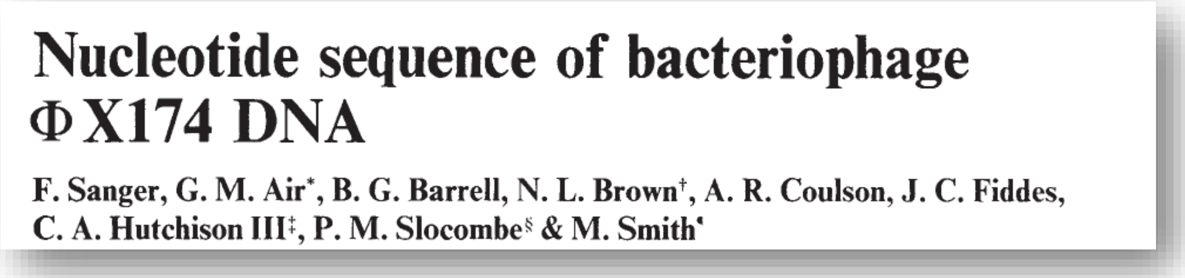
History of DNA sequencing

Watson & Crick
Nobel price 1962

- Structure of DNA was discovered in 1953.
- First sequences in 1965.

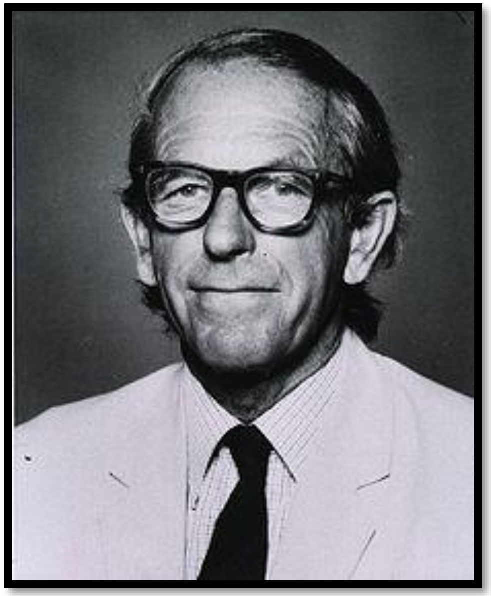
History of DNA sequencing

- ‘Rapid’ DNA sequencing developed by Sanger 1977.



Nucleotide sequence of bacteriophage ΦX174 DNA

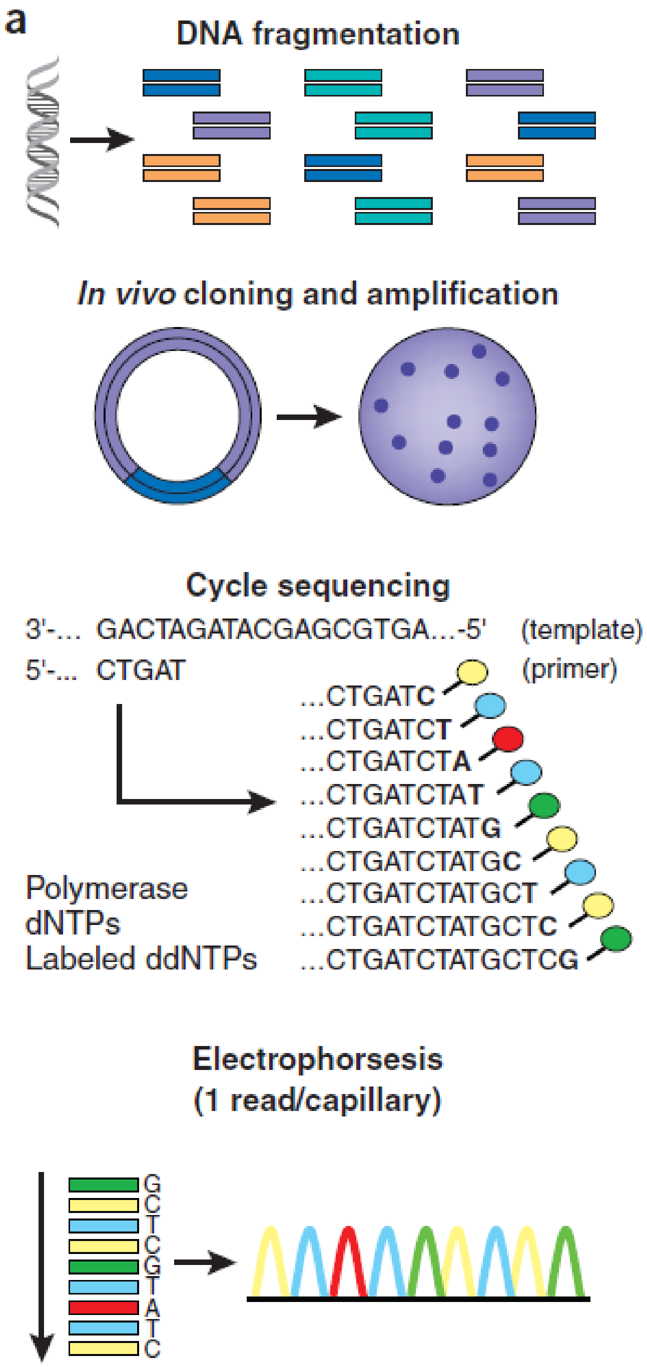
F. Sanger, G. M. Air[*], B. G. Barrell, N. L. Brown[†], A. R. Coulson, J. C. Fiddes, C. A. Hutchison III[‡], P. M. Slocombe[§] & M. Smith[¶]



Fred Sanger
Nobel price 1980

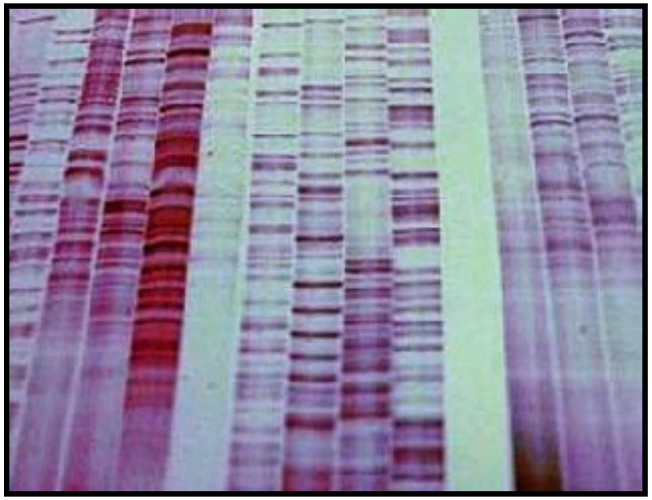
Sanger sequencing

- "First generation" sequencing
- Based on
  - Fragmentation of DNA
  - Cloning in bacteria
  - Chain-termination using labeled ddNTPs
  - Electrophoresis to separate labeled fragments
- Originally only short fragments (80 bases fragments). Can today sequence >1kb fragments.



Sanger sequencing

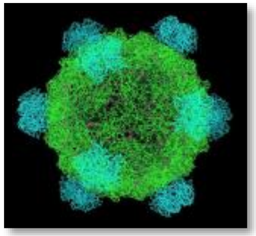
- Serial process – only one fragment is sequenced in one reaction
- Has a limited output and expensive
- Has a high accuracy – sequencing errors are relatively uncommon
- Today still often used as a 'golden standard'



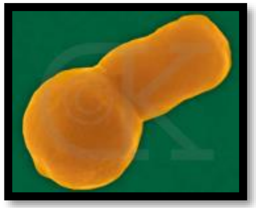
Gel used for base-calling

Multiple sequencing machines at the Sanger institute

History of genome sequencing



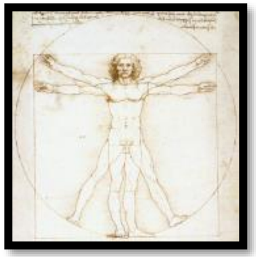
- **Bacteriophage Phi X 174**
  - First sequenced genome. Done by Fred Sanger.
  - 11 genes, 5,386 bases
  - Published 1977



- **Haemophilus influenzae**
  - First sequenced free living organism
  - 1800 genes, 1.8 million bases
  - Published 1995

History of DNA sequencing

- **Saccharomyces cerevisiae**
  - First sequenced eukaryote
  - Genome consists of 6000 genes and 12 million bases
  - Published 1997 – the project took 7 years



- **Homo sapiens**
  - The Human Genome Project
  - Genome consists of approximately 21.000 genes and 3.25 billion bases

The Human Genome Project

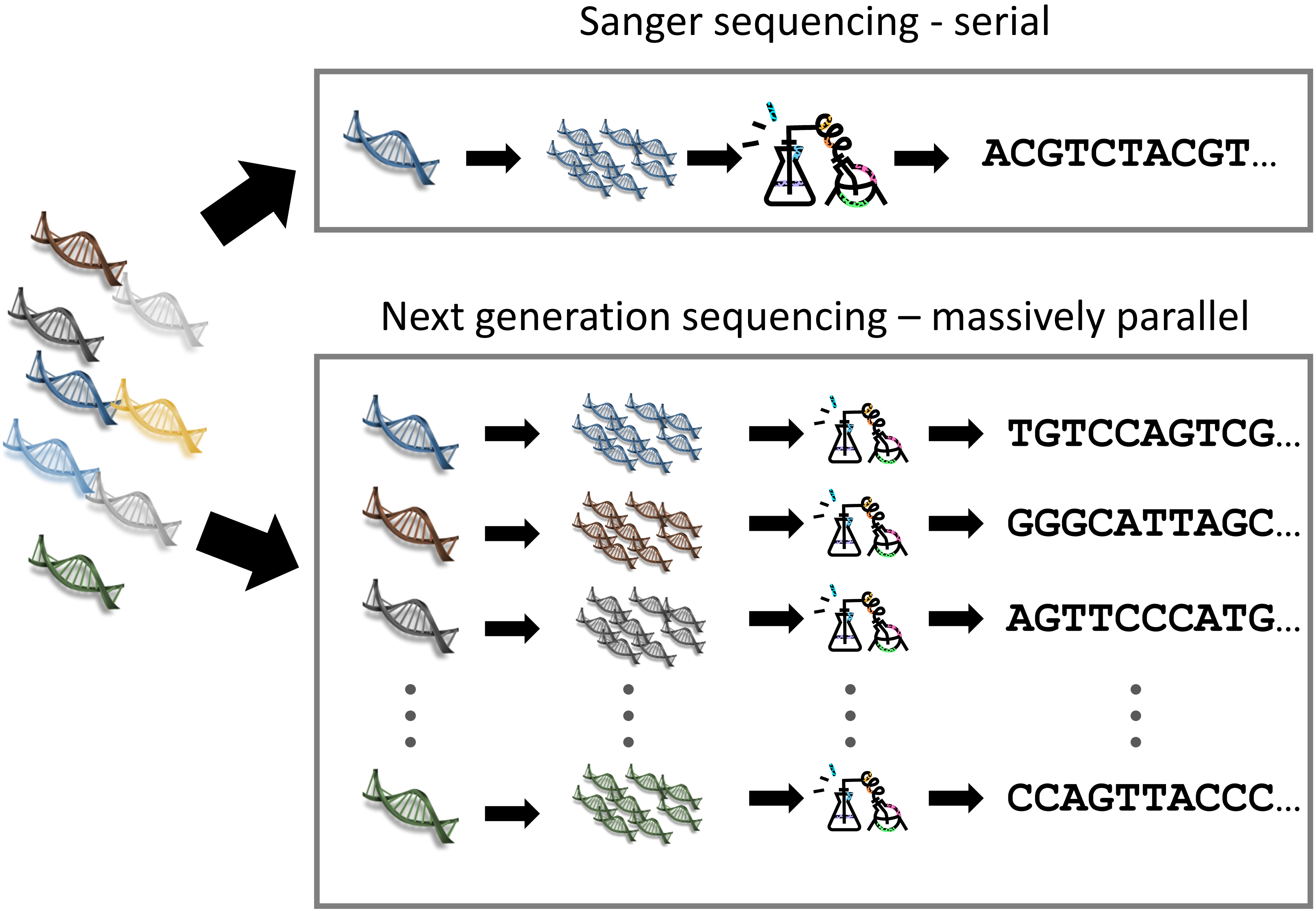
- Initiated 1990 – finished 13 year later
- Massive research effort
  - 200 research groups worldwide
  - Total cost estimated to $3 billion
- Sequence still not 100% complete (parts remain that can not be sequenced with todays technology)

nature
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities

Next generation sequencing

- Introduced 10 years ago (2006)
- From serial to parallel. Multiple DNA sequences are ‘read’ in one chemical reaction.
- Several platforms on the market – each have their own chemistry

- The output DNA sequences are referred to as ‘reads’.



Sanger sequencing - serial

ACGTCTACGT…

Next generation sequencing – massively parallel

TGTCCAGTCG…

GGGCATTAGC…

AGTTCCCATG…

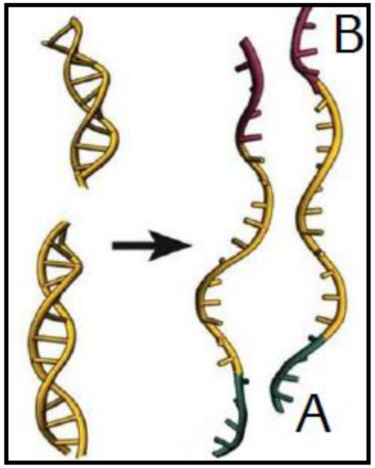
CCAGTTACCC…

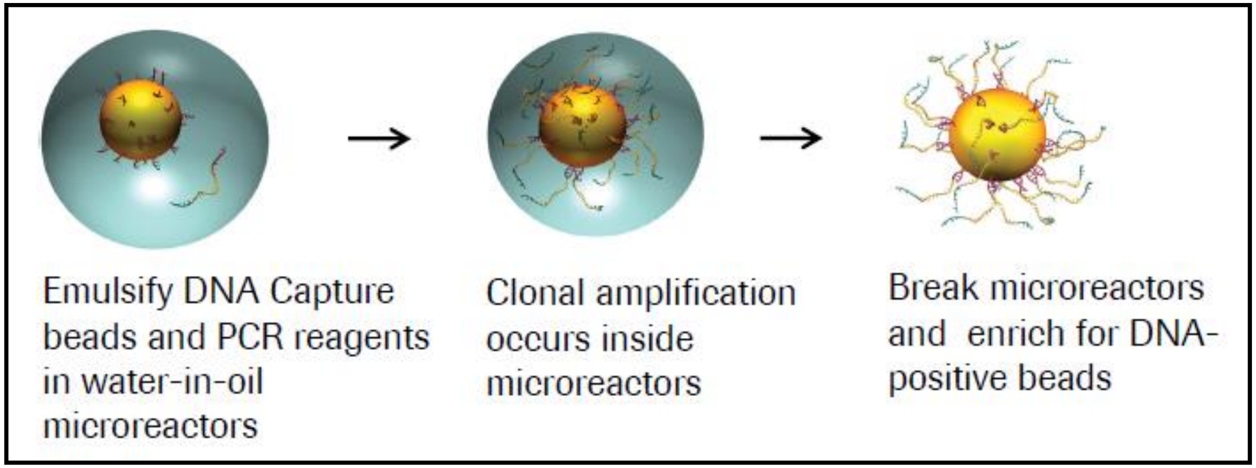
Massively parallel pyrosequencing

- Early NGS technique available between 2005-2015.
- Sequence 1 million of DNA fragment simultaneously.
- Read lengths around 400 bases

454 LIFE SCIENCES

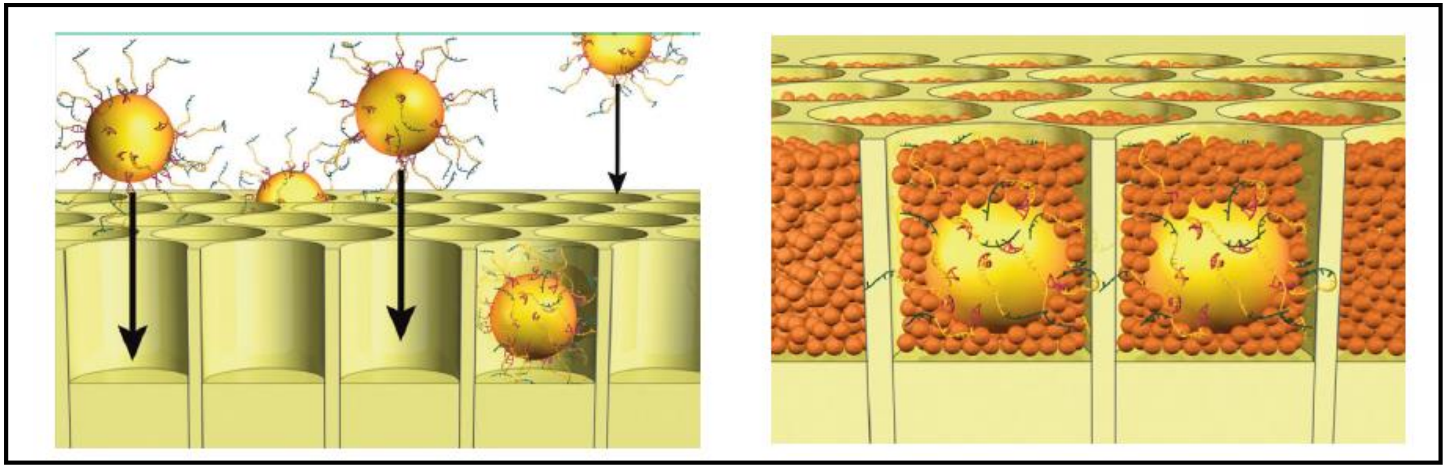
Massively parallel pyrosequencing



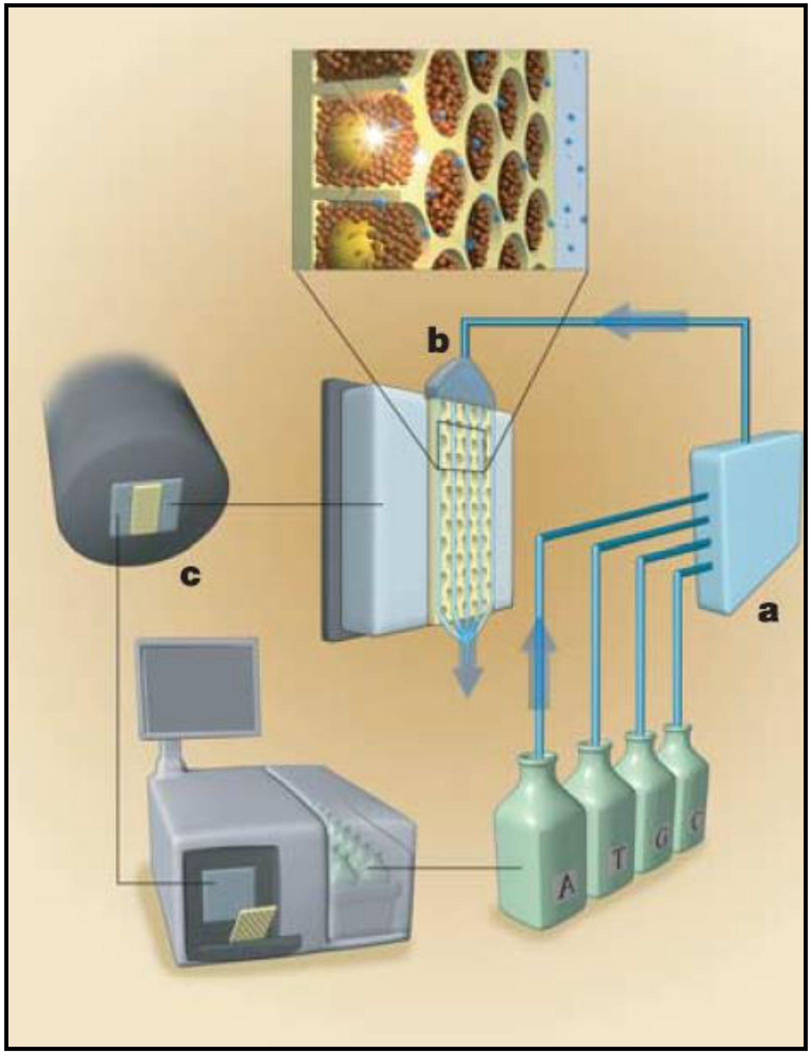

Emulsify DNA Capture beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors and enrich for DNA-positive beads

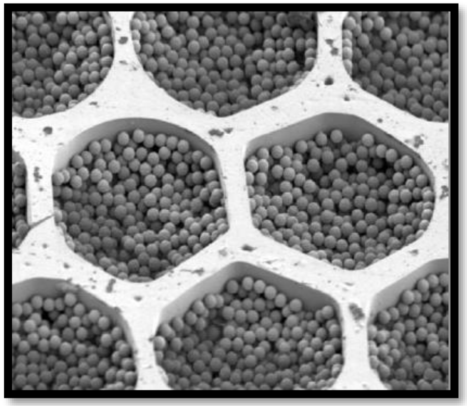


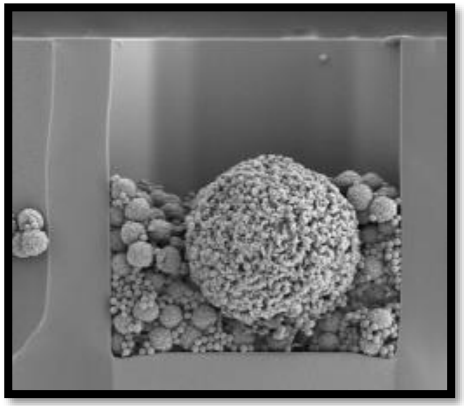
454 sequencing

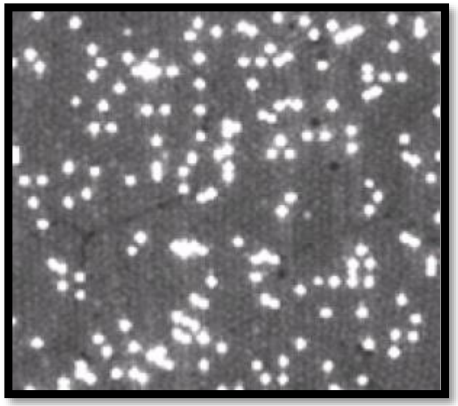


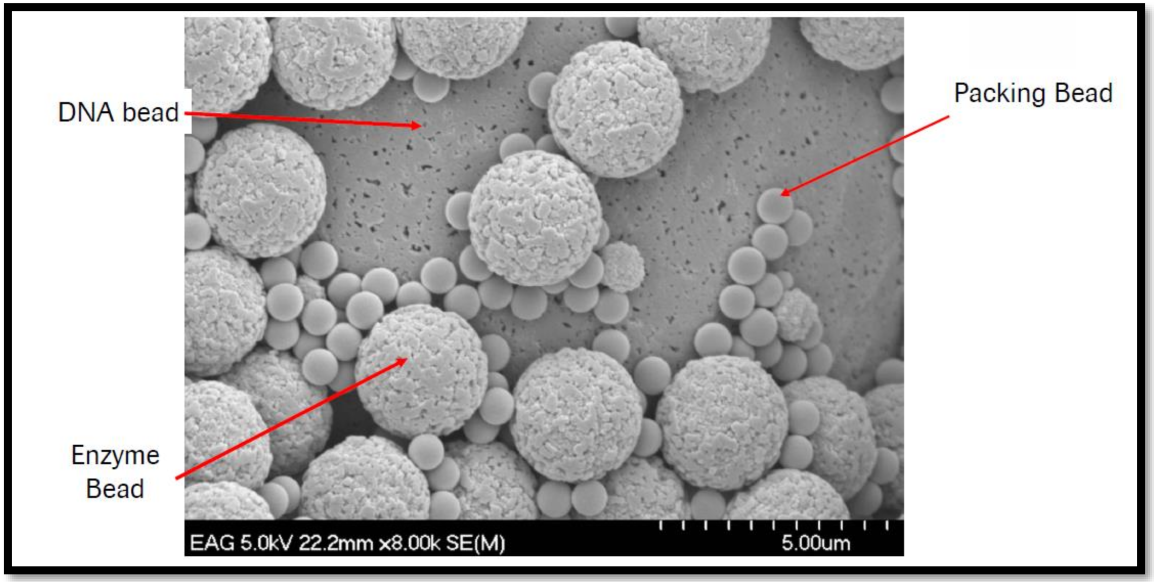
- Nucleotides are flowed sequentially (**a**)
- When a nucleotide is incorporated on the opposite strand, energy is released.
- Luciferase converts the energy to light (**b**)
- A CCD camera is generating an image after each flow (**c**)
- The signal strength is proportional to the number of incorporated nucleotides.

454 sequencing

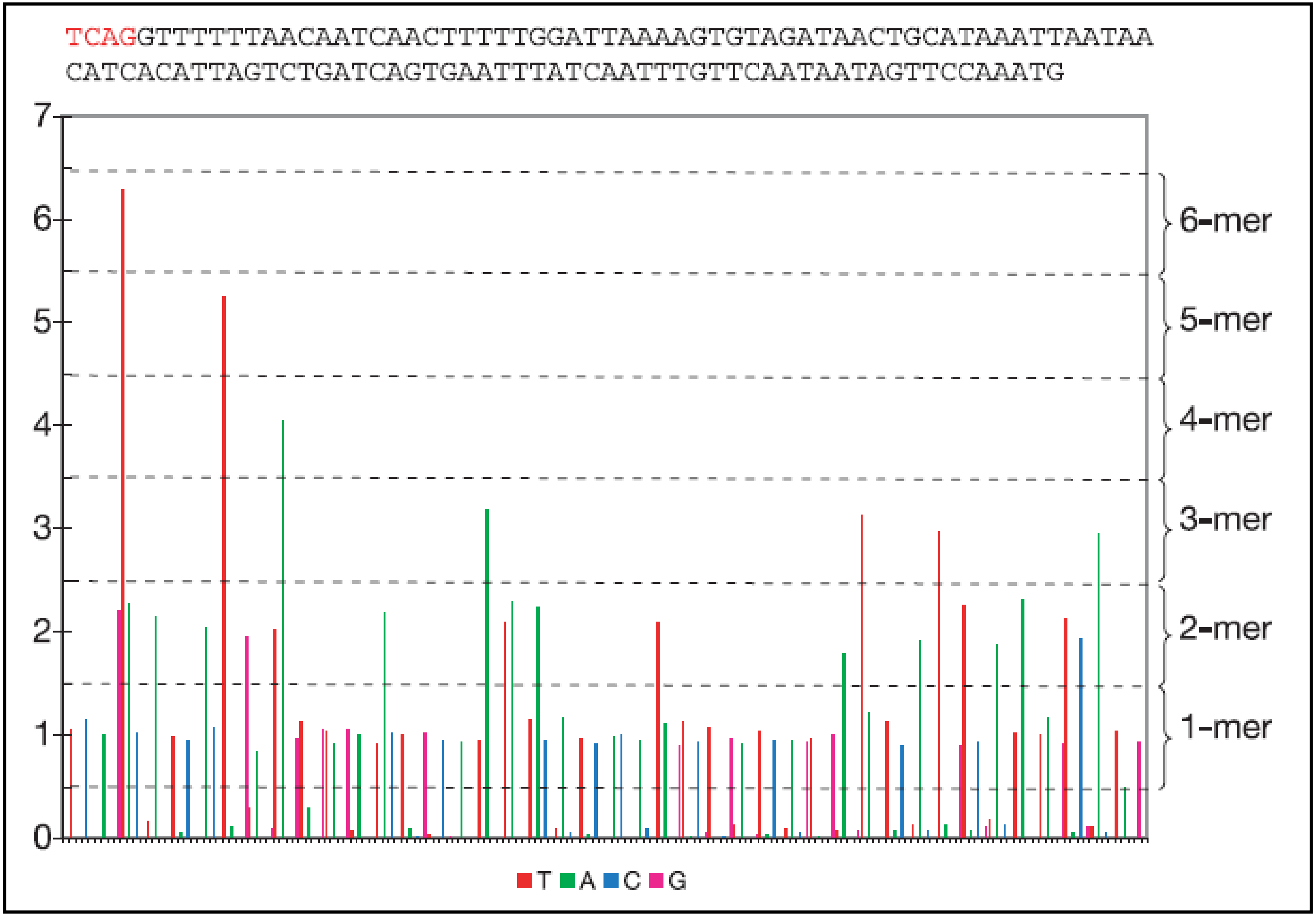








Flowgram from massively parallel pyrosequencing



Massively parallel pyrosequencing

Advantages

- Early technique – was a driving force of the NGS development
- Significantly higher throughput than Sanger

Disadvantages

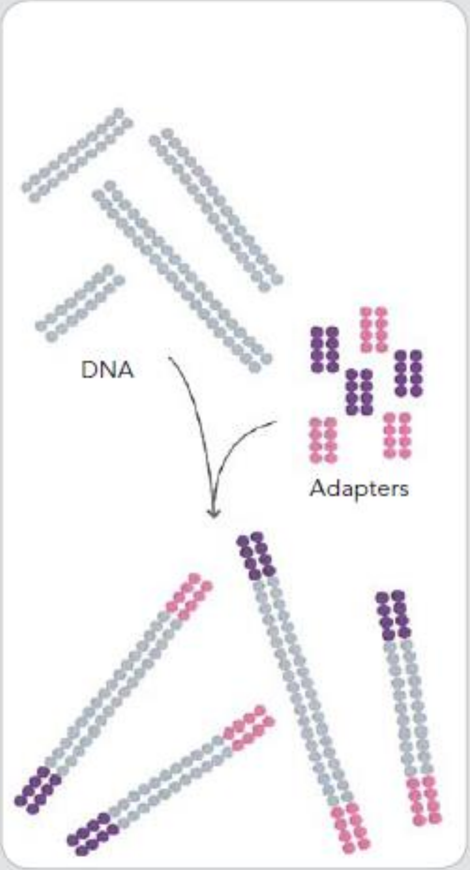
- Did not scale well – throughput maxed at 1 million fragments/sequencing run
- Complex and error-prone sample preparation
- Large error rate

Illumina sequencing

- Introduced 2006
- High throughput, can sequence billions of fragments in a single run
- Shot read lengths (up to 350 base pairs)

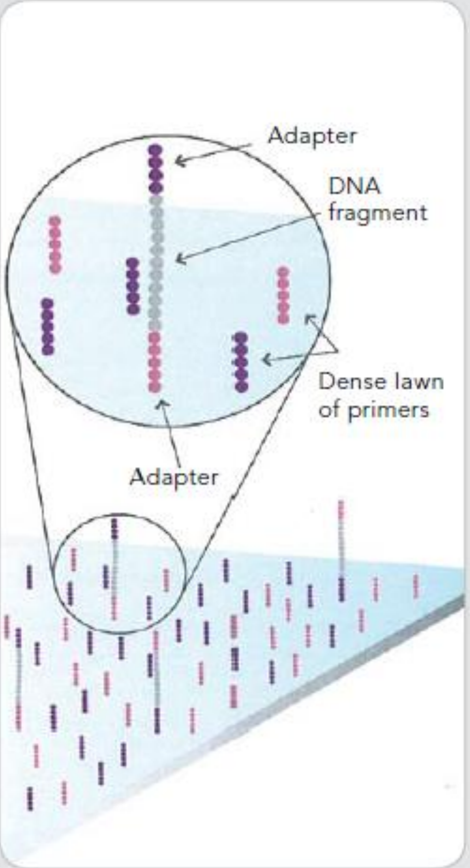
Illumina sequencing

1. PREPARE GENOMIC DNA SAMPLE



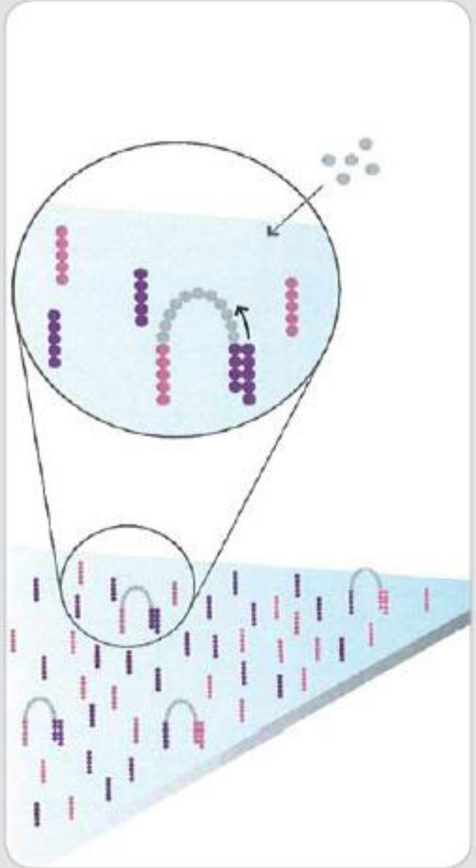
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

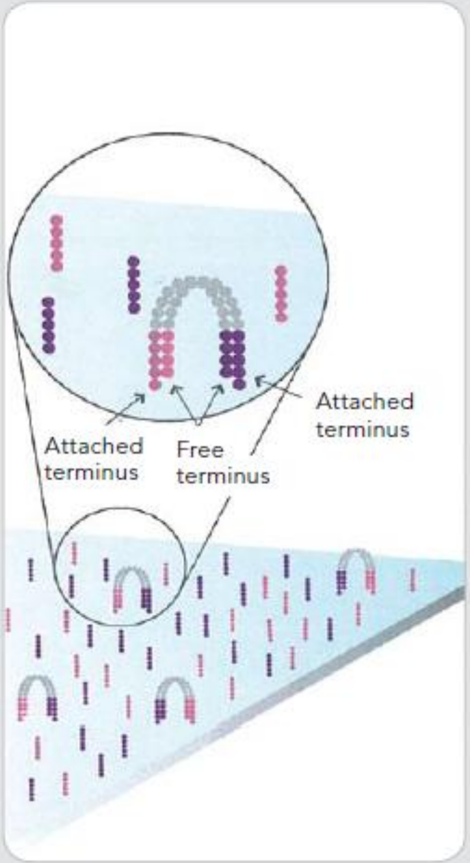
3. BRIDGE AMPLIFICATION



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

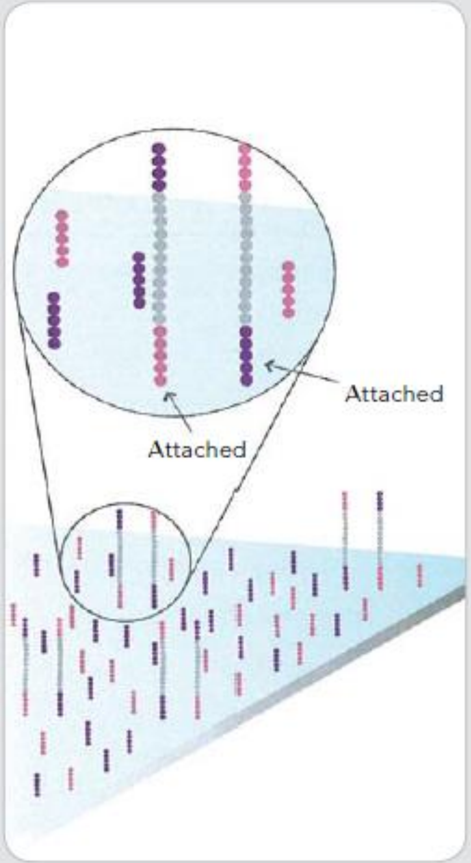
Illumina sequencing

4. FRAGMENTS BECOME DOUBLE-STRANDED



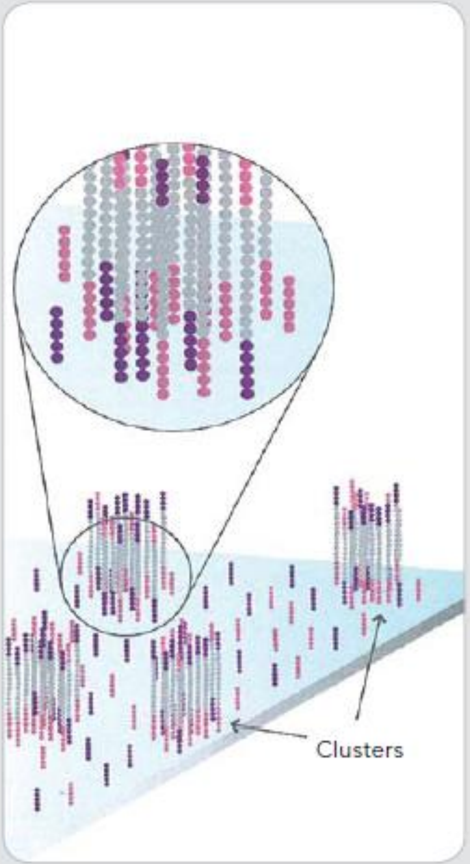
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

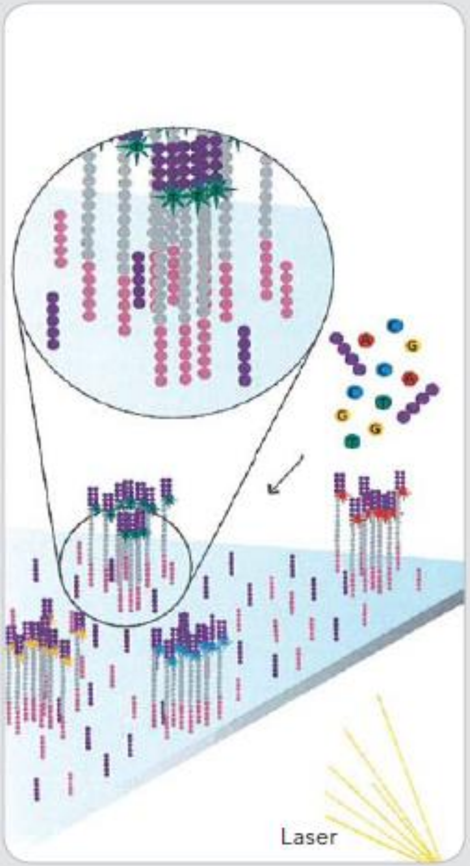
6. COMPLETE AMPLIFICATION



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

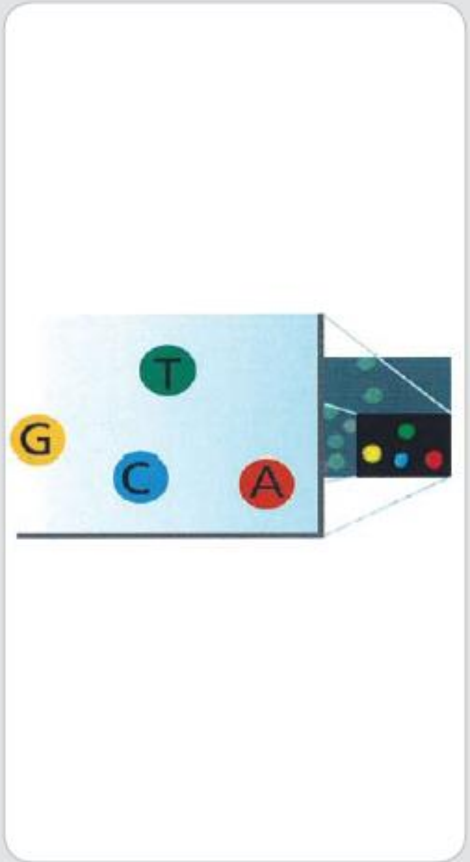
Illumina sequencing

7. DETERMINE FIRST BASE



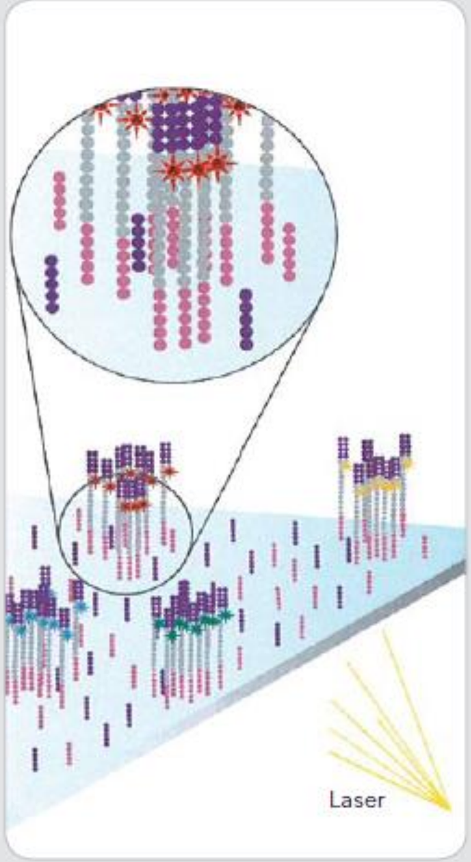
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. IMAGE FIRST BASE

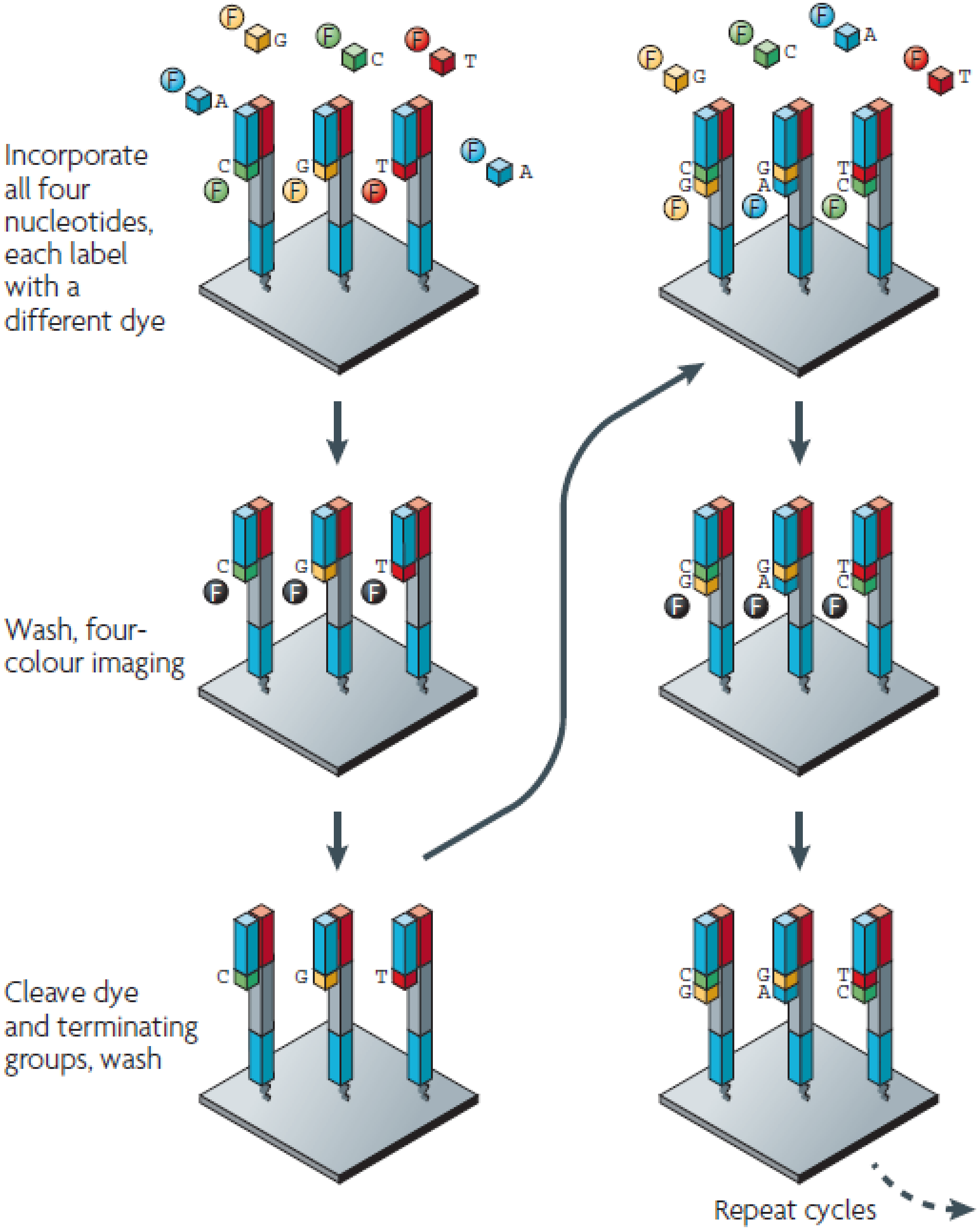


After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

9. DETERMINE SECOND BASE



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

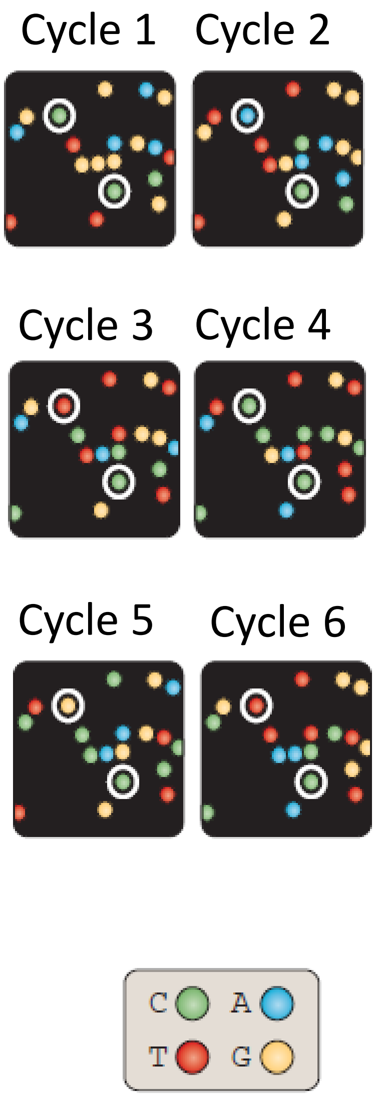


Incorporate all four nucleotides, each label with a different dye

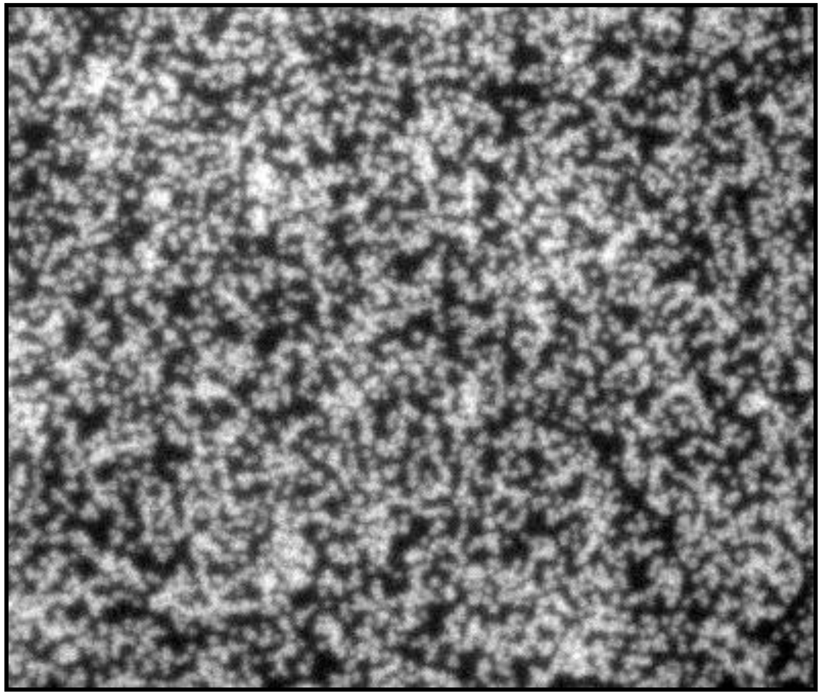
Wash, four-colour imaging

Cleave dye and terminating groups, wash

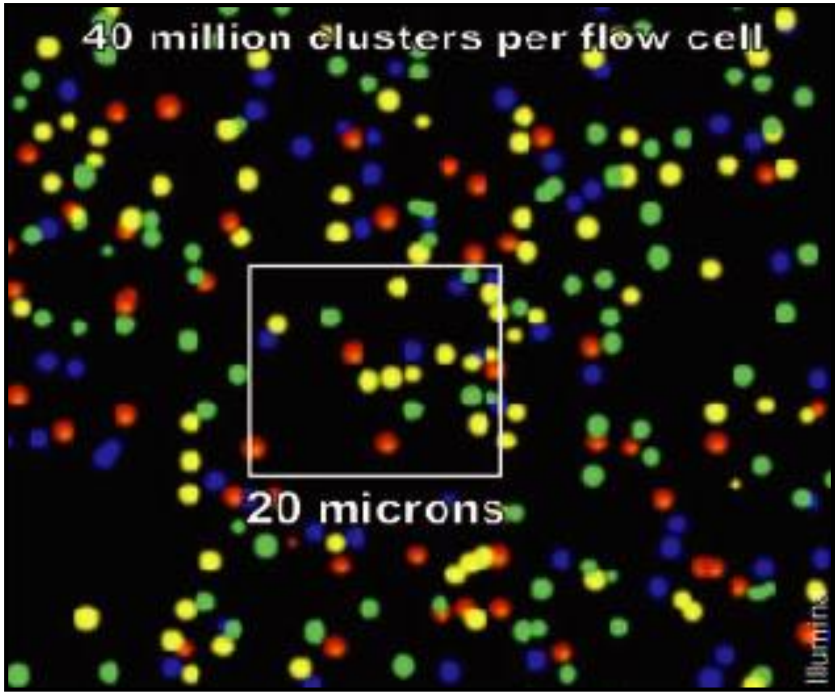
Repeat cycles



From Metzker, M. (2010). Sequencing Technologies – the next generation. Nature Reviews.



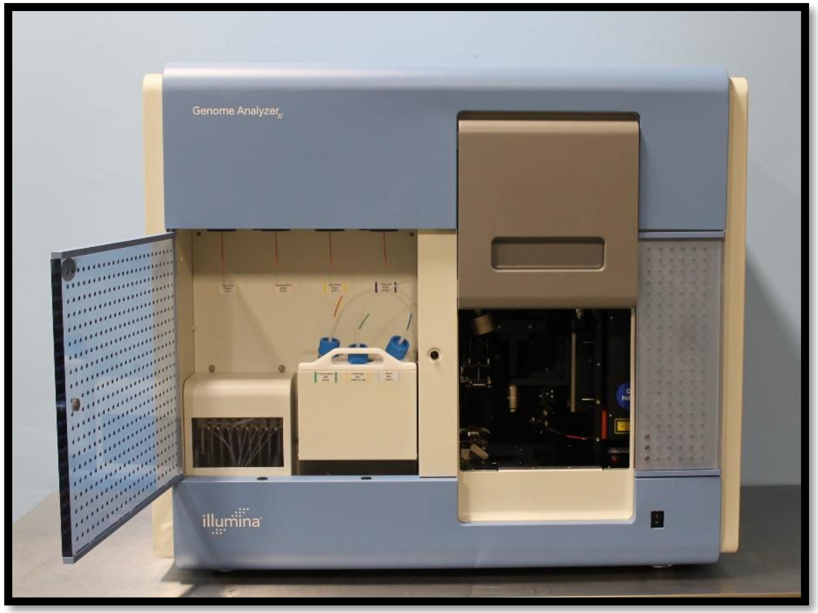
Raw image



Pseudo-colored image

- Hundreds of thousands high-resolution images are analyzed during a sequencing run
- Processing of these images (terabytes of data) requires a small computer cluster.

Genome Analyzer II



First version sequenced around 30 million reads corresponding to 1 gigabase (1×10^9). One run took two days.

NovaSeq 6000

Sequences 20 billion reads at once corresponding to 6 terabases (1.8×10^{12}). One run takes two days

Illumina sequencing

Advantages

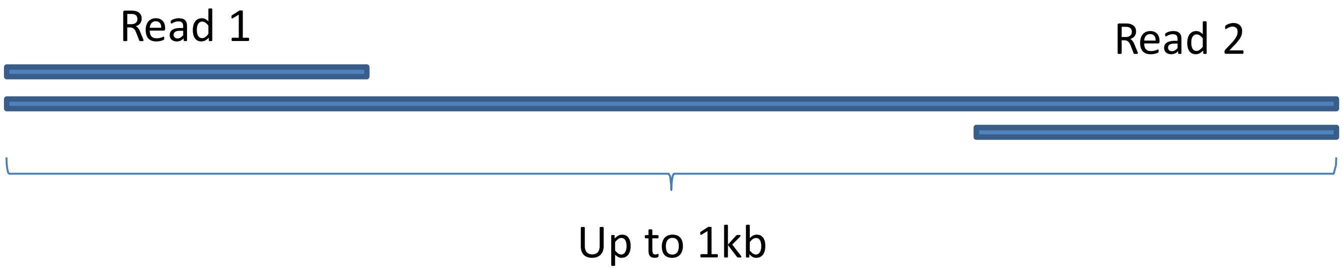
- Very high throughput. Cost-efficient when sequencing larges volumes of DNA.
- Low error rate (compared to other NGS platforms)
- Paired-end reads.

Disadvantages

- Short sequence reads (up to 350 bases)
- High error rates in certain genomic contexts (e.g. high GC-content)

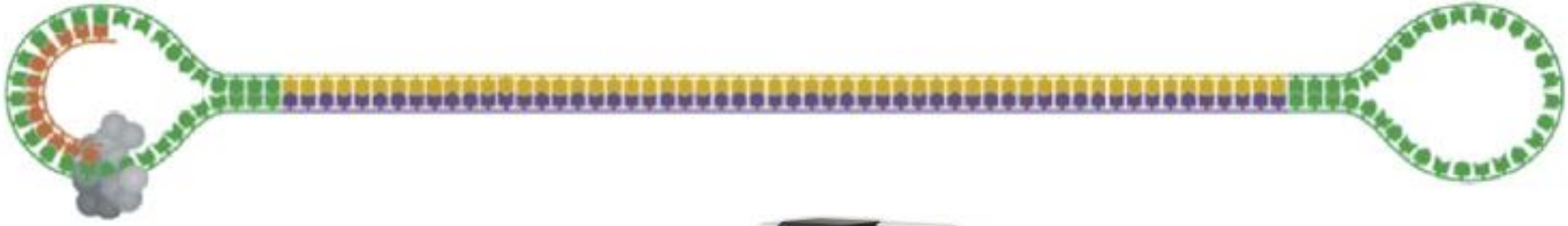
Paired-end sequencing

- Illumina sequencing can be done ‘paired-end’ where both ends of the same DNA fragment is sequenced.
- Simplifies the data analysis, for example the reconstruction of DNA regions.

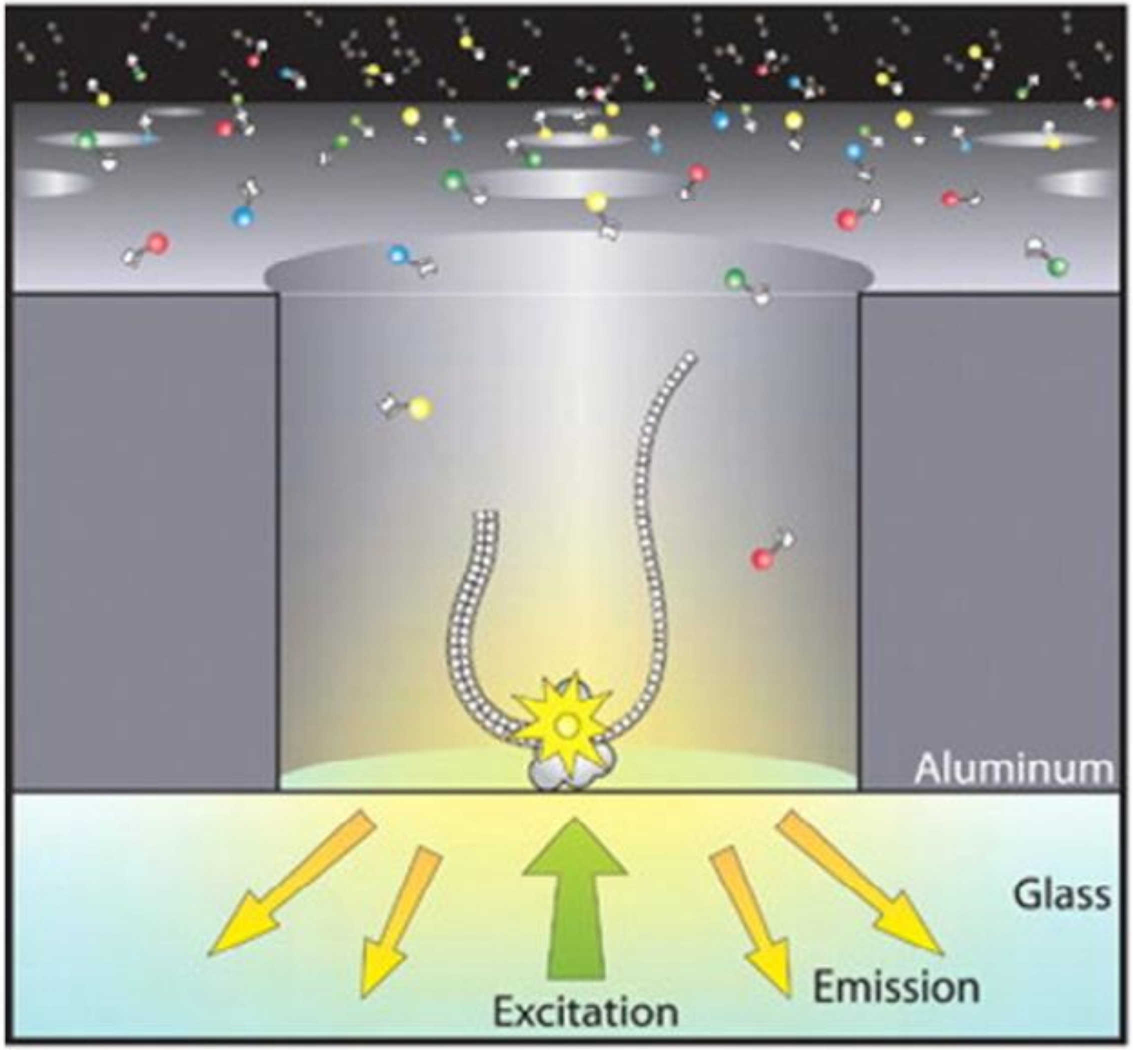


Pacific Bioscience – SMRT sequencing

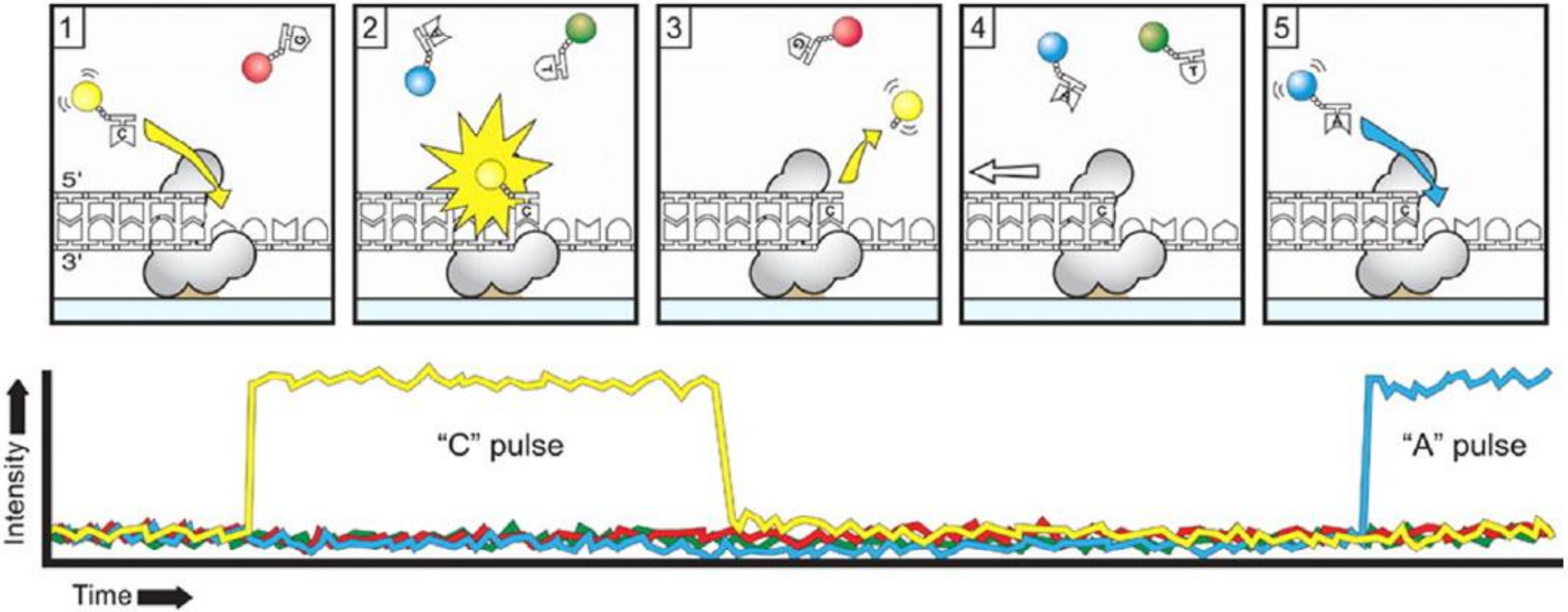
- Introduced early 2011
- Single-molecule sequencing in ‘real time’
- Has a relatively low throughput but can generate very long reads. Currently up to 100,000 bases



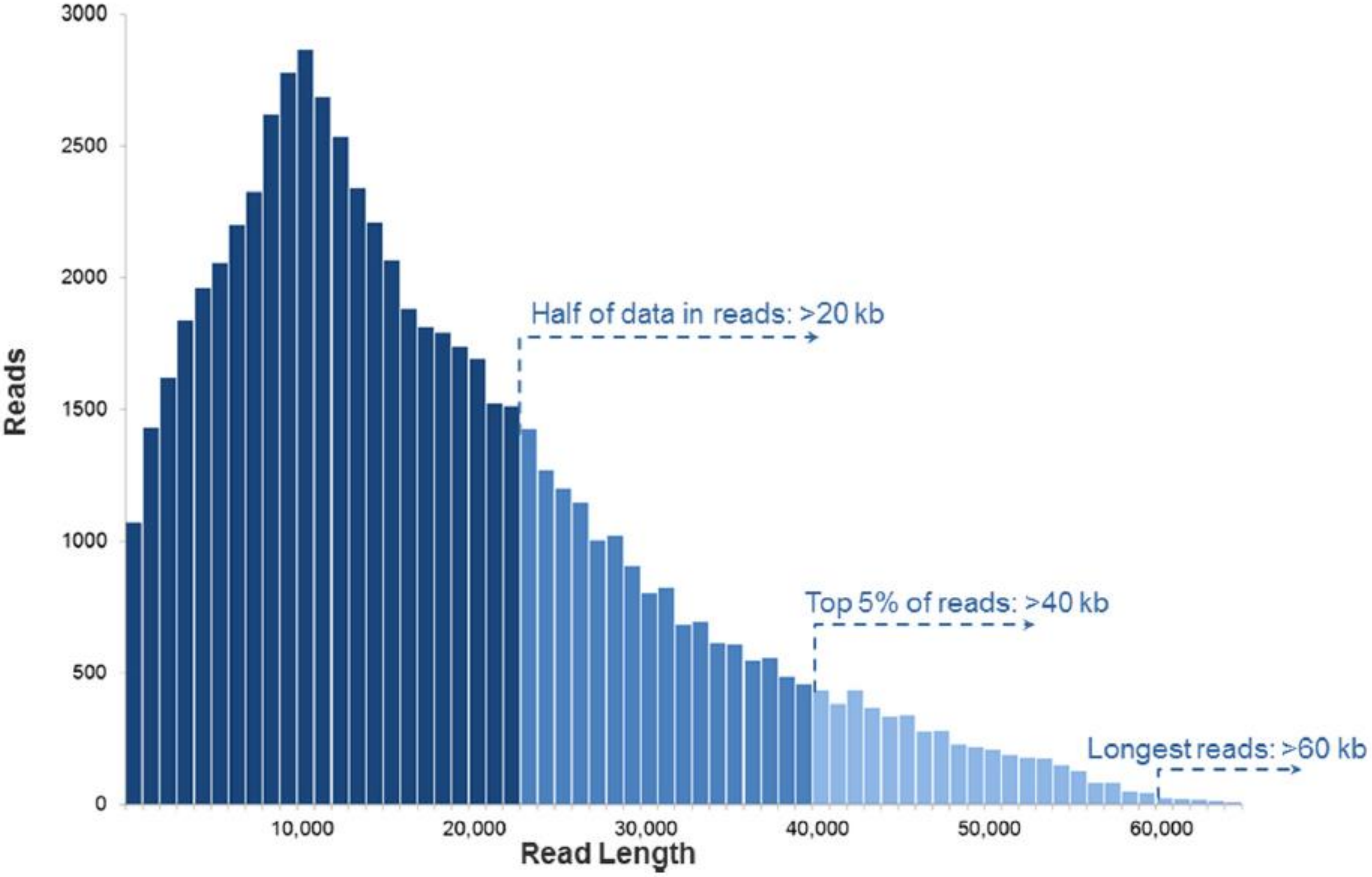
Pacific Bioscience – SMRT sequencing



Pacific Bioscience – SMRT sequencing

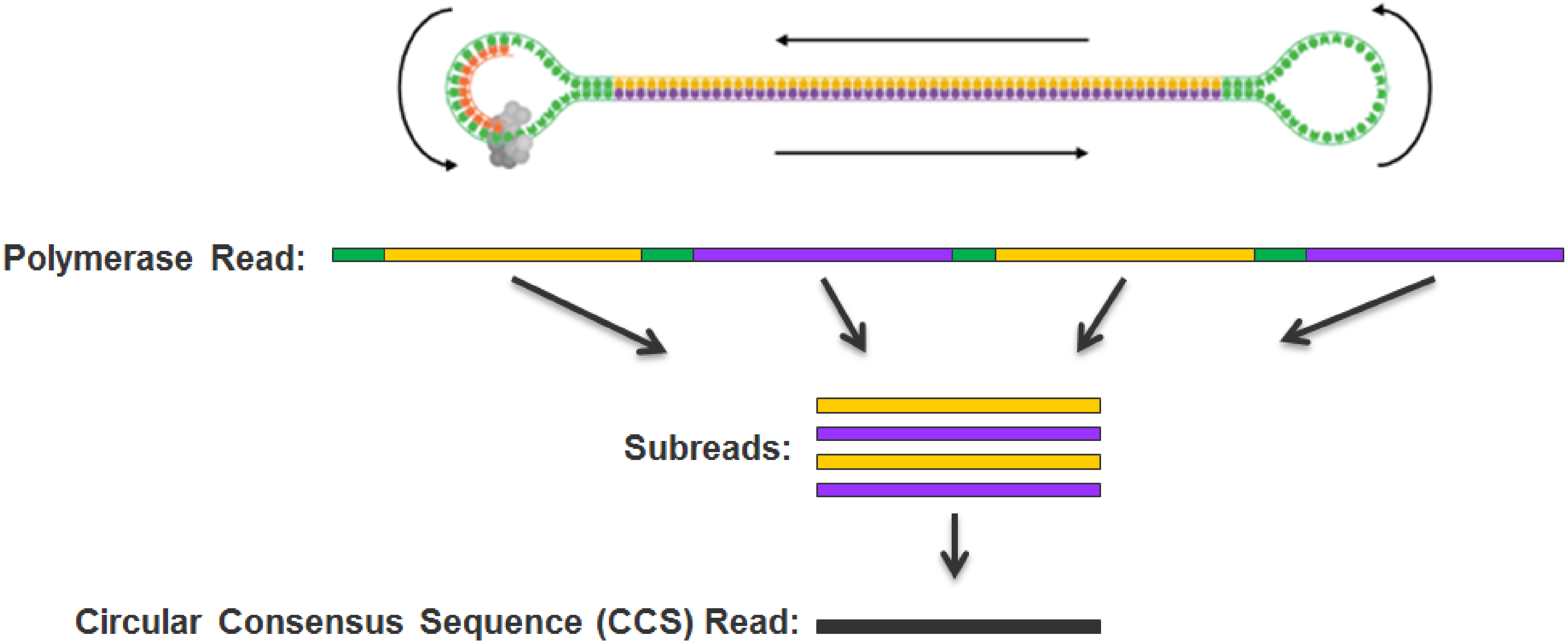


Pacific Bioscience – SMRT sequencing



Pacific Bioscience - circular consensus sequencing

- Short fragments can be sequenced multiple times. The reads (‘subreads’) can then be combined into a consensus.
- This is called circular consensus sequencing (CCS).



Polymerase Read:

Subreads:

Circular Consensus Sequence (CCS) Read:

Pacific Bioscience – SMRT sequencing

Advantages

- Long sequence reads
- Single molecule – no PCR amplification necessary
- Error patterns are more random

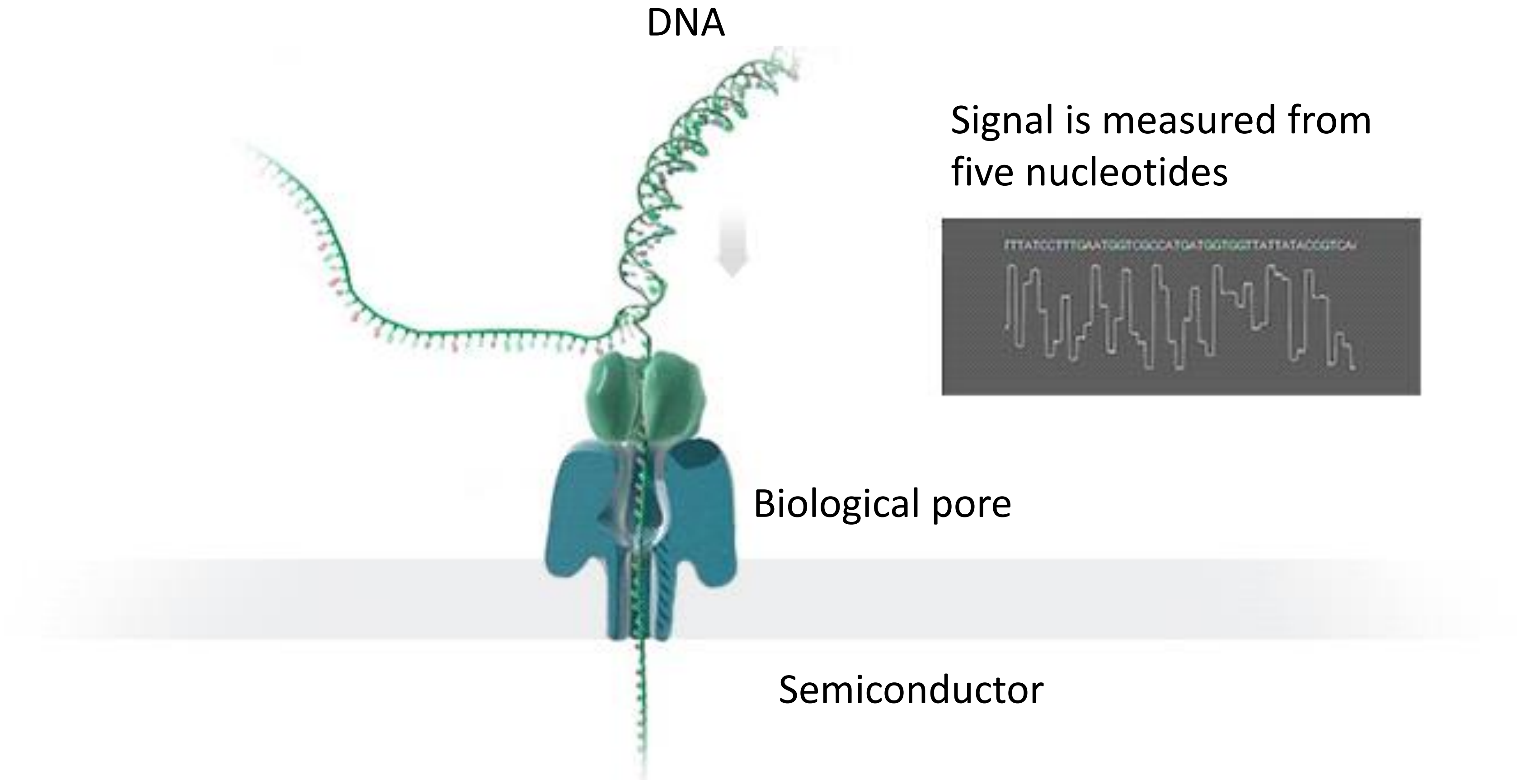
Disadvantages

- Lower throughput, relatively expensive
- Error prone (>10% sequencing errors)
- Requires DNA of very high quality

Oxford Nanopore

- Introduced May 2015
- Single molecule ’real-time’ sequencing by pulling the DNA strand through a biological pore
- The MiniION sequencers weights 90 grams and is connected to the USB port of a computer.

Oxford Nanopore

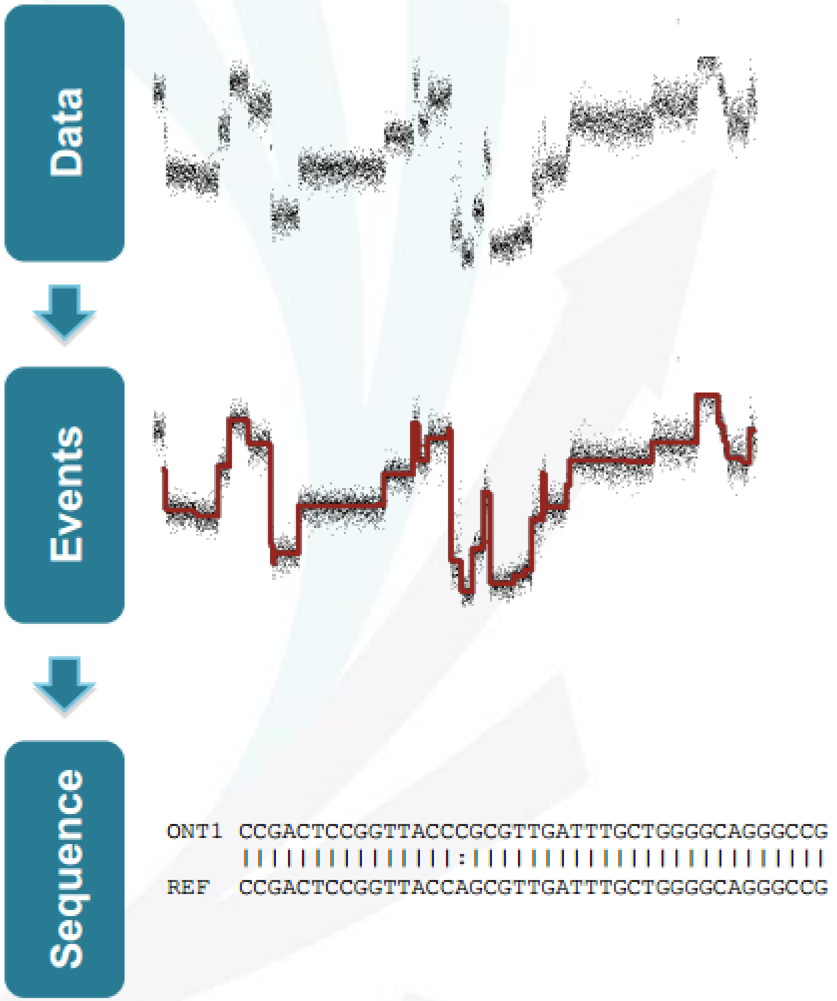


DNA

Signal is measured from five nucleotides

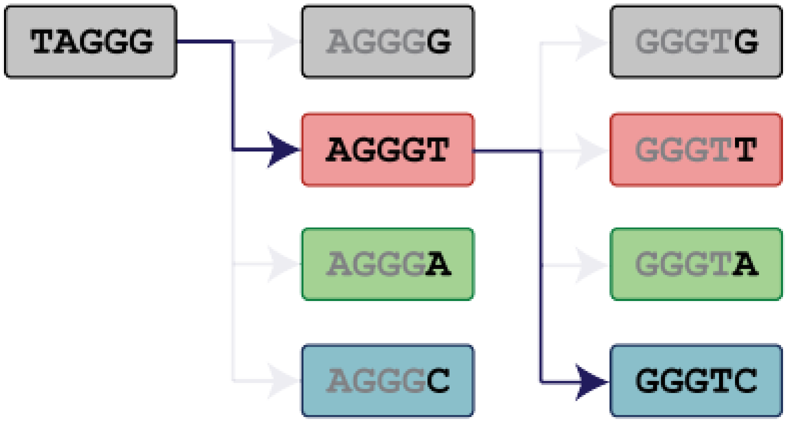
Biological pore

Semiconductor



```
ONT1 CCGACTCCGGTTACCCGCGTTGATTTGCTGGGGCAGGGCCG
     |||||||||||||||:|||||||||||||||||||||||||
REF  CCGACTCCGGTTACCAGCGTTGATTTGCTGGGGCAGGGCCG
```

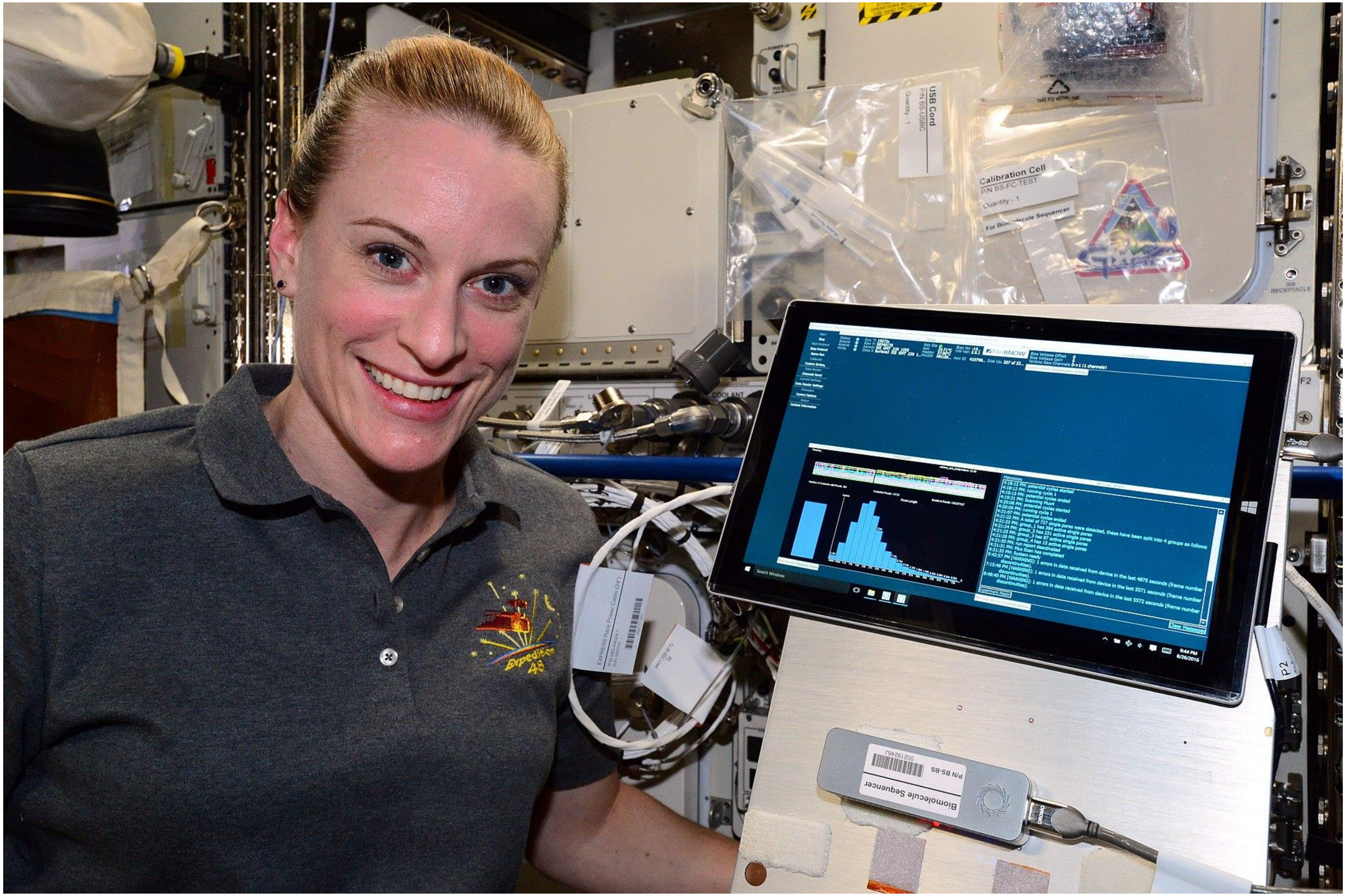
- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length



- Form probabilistic path through measured states currents and transitions
    - e.g. Viterbi algorithm

From https://konradpaszkiewicz.wordpress.com/2014/04/10/nanopore/

Oxford Nanopore MiniION in space



Astronaut Kate Rubins with a MinION sequencer on the International Space Station.

Oxford Nanopore

Output estimated to correspond to up to 512 MiniIONs
(up to 15 terabases in 48 hours).

Nanopore sequencing

Advantages

- Long sequence reads (up to 100kb)
- Single molecule – no PCR amplification necessary
- Easy sample preparation
- Portable – sequence DNA anywhere?

Disadvantages

- Lower throughput, relatively expensive
- Error prone (>10% sequencing errors). Problems with homopolymers
- Requires DNA of high quality

Major NGS platforms on the market

Massively parallel pyrosequencing (454)

IonTorrent (ThermoFisher)

Illumina

SMRT sequencing (Pacific Bioscience)

MinION (Oxford Nanopore)

Second generation sequencers

Third generation sequencers

Summary of NGS platforms

Second generation DNA sequence platforms

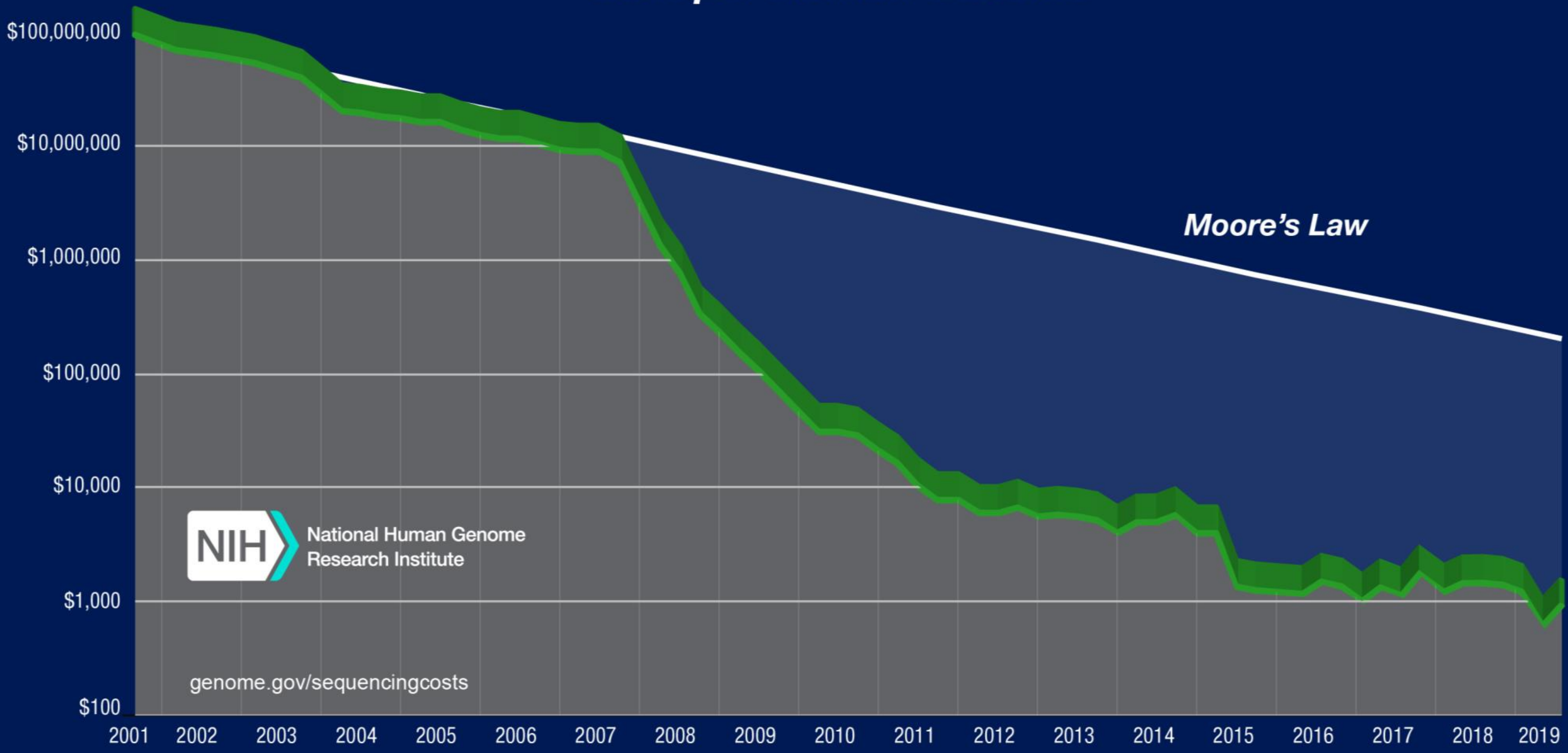
- High throughput
- Short reads
- Low error rate (still higher than Sanger sequencing)
- Requires amplification

Third generation DNA sequence platforms

- Lower throughput
- Long reads
- High error rate
- Real time and single molecule

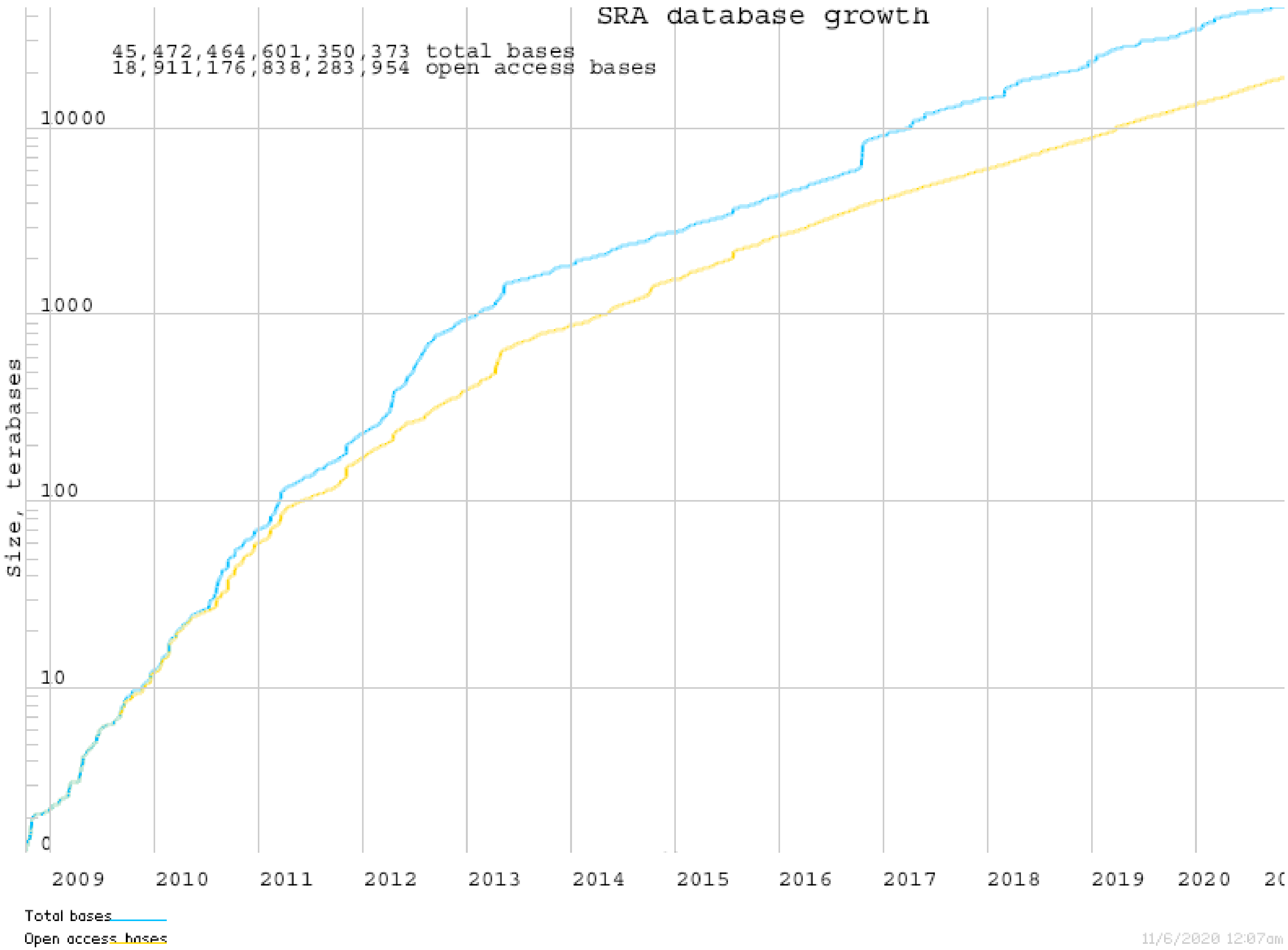
What are the implications of next generation DNA sequencing?

Cost per Human Genome



NIH National Human Genome Research Institute

genome.gov/sequencingcosts


SRA database growth
45,472,464,601,350,373 total bases
18,911,176,838,283,954 open access bases
Size, terabases
10000
1000
100
10
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
20
Total bases
Open access bases
11/6/2020 12:07am

Implications of NGS

- Output from DNA sequencing has increased more than **1 million times** in less than 10 years.
- The cost of sequencing a human genome is today **less than $1000**
- Public databases has accumulated **larges volumes** of DNA sequence data (petabytes of data!)

Current focus is not on analysis and interpretation of DNA sequence data.

Major applications of NGS

In this course we will focus on three major applications

- Genome sequencing
- Transcriptome sequencing (RNA-seq)
- Metagenomics

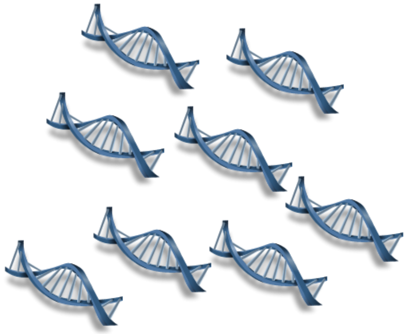
There are many other applications including

- *De novo* genome/transcriptome sequencing
- Chromatin Immunoprecipitation Sequencing (ChIPSeq)

These will not be covered in this course!

Major applications of NGS

Genome sequencing: characterization of the DNA present in eukaryotic and prokaryotic cells



DNA

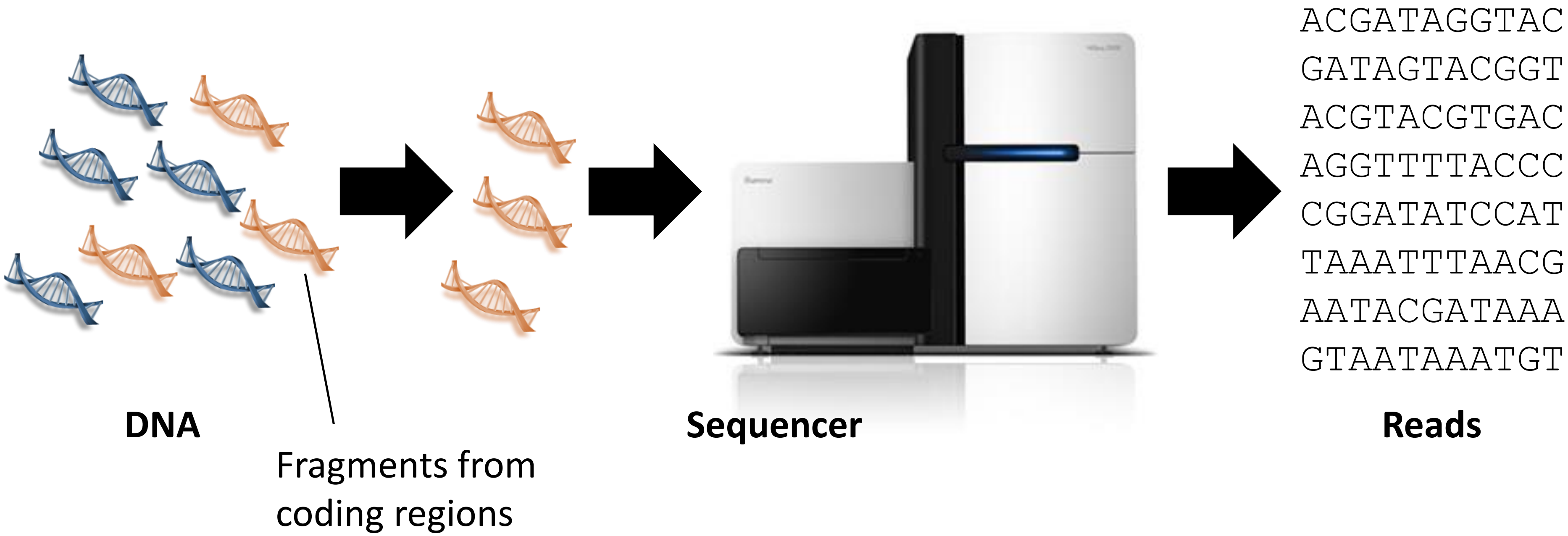
Sequencer

```
ACGATAGGTAC
GATAGTACGGT
ACGTACGTGAC
AGGTTTTACCC
CGGATATCCAT
TAAATTTAACG
AATACGATAAA
GTAATAAATGT
```

Reads

Major applications of NGS

Exome sequencing: characterization of the DNA present in coding regions of eukaryotic and prokaryotic cells



Applications of NGS

Genome sequencing

- The sequence reads are compared to a reference genome
- Differences are used to identify mutations

CTCTACAGCAGATGTGGAA

CTCTACATCTGATGTGGAA

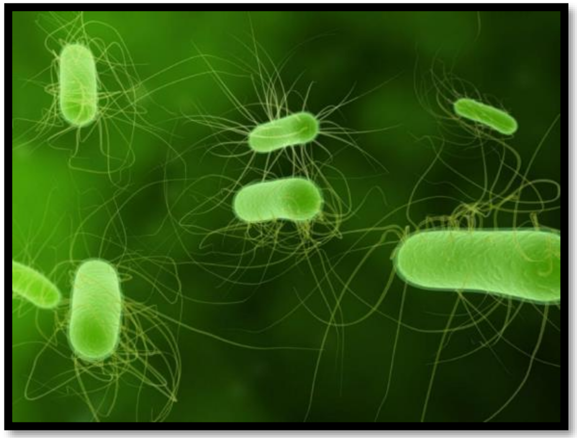
Mutation

Example: Cancer genomics

- Cancer is caused by genetic mutations that results in cells with abnormal growth and the ability to invade other parts of the body.
- Genome sequencing is used to identify mutations carried by a tumor
- The mutations can be used in diagnostics, targeted treatments and to understanding how tumors develop and spread.

Example: Antibiotic resistance

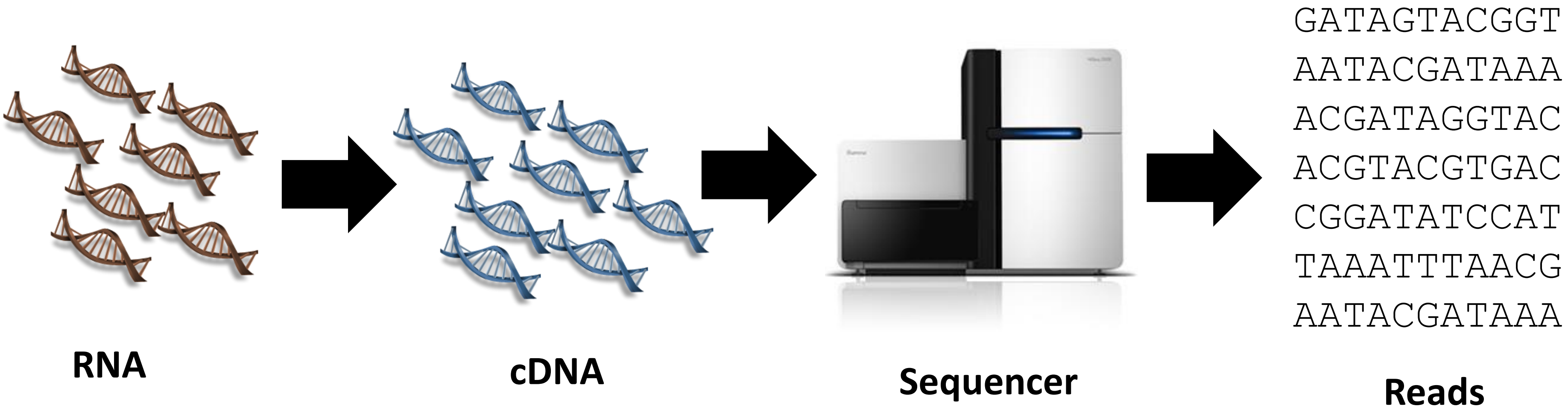
- Antibiotics are used to treat bacterial infections.
- Antibiotic resistance is caused by mutations in the bacterial genome.
- Genome sequencing of bacteria
  - Enables improved identification of resistant bacteria
  - Can be used to guide treatment by suggesting more efficient antibiotics
  - Can be used to monitor the spread of resistance over the world.



Applications of NGS

Transcriptome sequencing: Sequencing of RNA expressed in a cell.

- Measures gene expression through mRNA abundance of different genes (mRNA abundance)
- Describes the sequence of transcribed RNA.

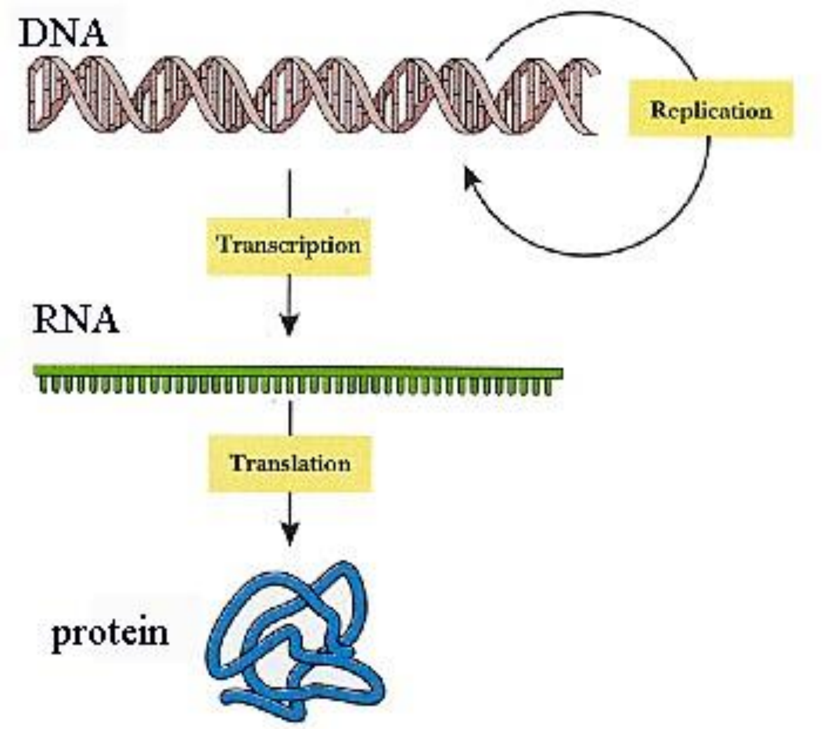


GATAGTACGGT
AATACGATAAA
ACGATAGGTAC
ACGTACGTGAC
CGGATATCCAT
TAAATTTAACG
AATACGATAAA

Applications of NGS

Transcriptome sequencing

- The sequence reads are matched to genes in the genome
- Used to analyze which genes that are expressed



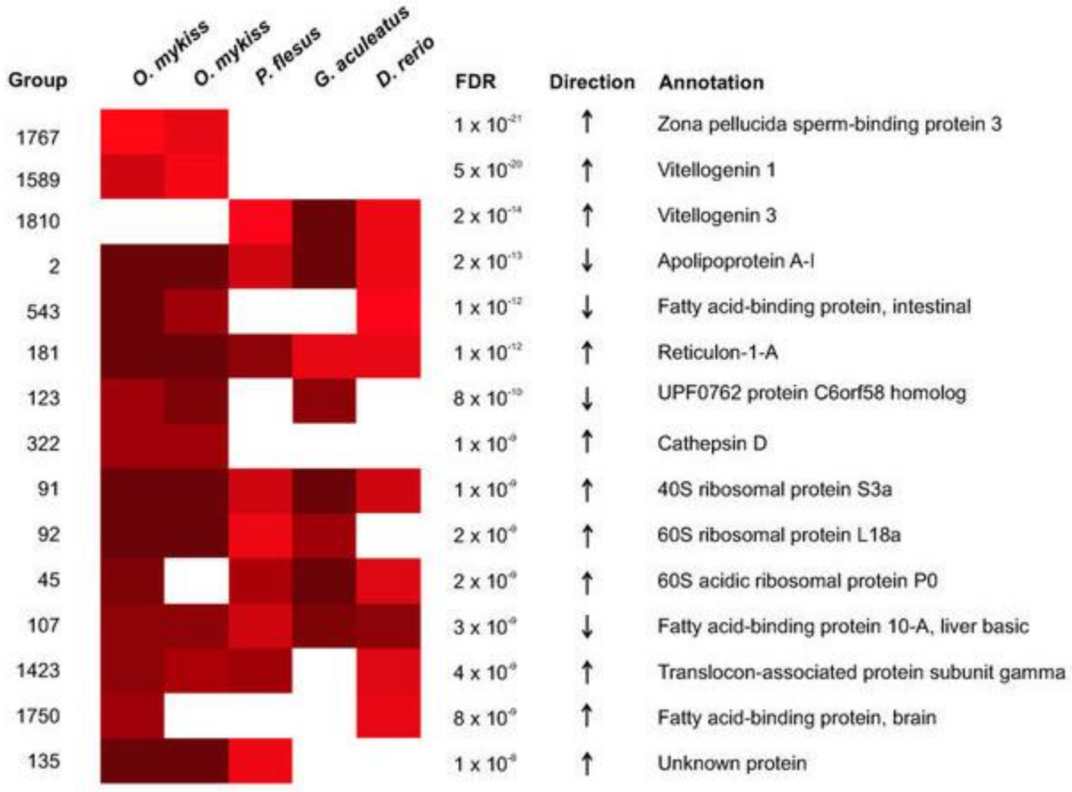
Example: Gene expression changes in human disease

- Many human diseases, including cancer, are associated with changes in genes expression.
- Identification of these changes provides information the molecular mechanics of the disease.
- Genes regulated during disease can constitute targets for novel drugs.
- Gene expression patterns can be used for target diagnostics and treatment.

Example: Pollution in the enviroment

- Toxic chemicals are emitted into aquatic environments (lakes and the sea) and can have a negative impact on wild life.

- Analysis of changes in gene expression can provide information about the mode of toxicity and adverse effects.



Group	FDR	Direction	Annotation
1767	1 x 10^{-21}	↑	Zona pellucida sperm-binding protein 3
1589	5 x 10^{-20}	↑	Vitellogenin 1
1810	2 x 10^{-14}	↑	Vitellogenin 3
2	2 x 10^{-13}	↓	Apolipoprotein A-I
543	1 x 10^{-12}	↓	Fatty acid-binding protein, intestinal
181	1 x 10^{-12}	↑	Reticulon-1-A
123	8 x 10^{-10}	↓	UPF0762 protein C6orf58 homolog
322	1 x 10^{-9}	↑	Cathepsin D
91	1 x 10^{-9}	↑	40S ribosomal protein S3a
92	2 x 10^{-9}	↑	60S ribosomal protein L18a
45	2 x 10^{-9}	↑	60S acidic ribosomal protein P0
107	3 x 10^{-9}	↓	Fatty acid-binding protein 10-A, liver basic
1423	4 x 10^{-9}	↑	Translocon-associated protein subunit gamma
1750	8 x 10^{-9}	↑	Fatty acid-binding protein, brain
135	1 x 10^{-8}	↑	Unknown protein

Applications of NGS

Metagenomics: Sequencing of DNA from a complex mixture of organisms.

DNA from mixture of organisms

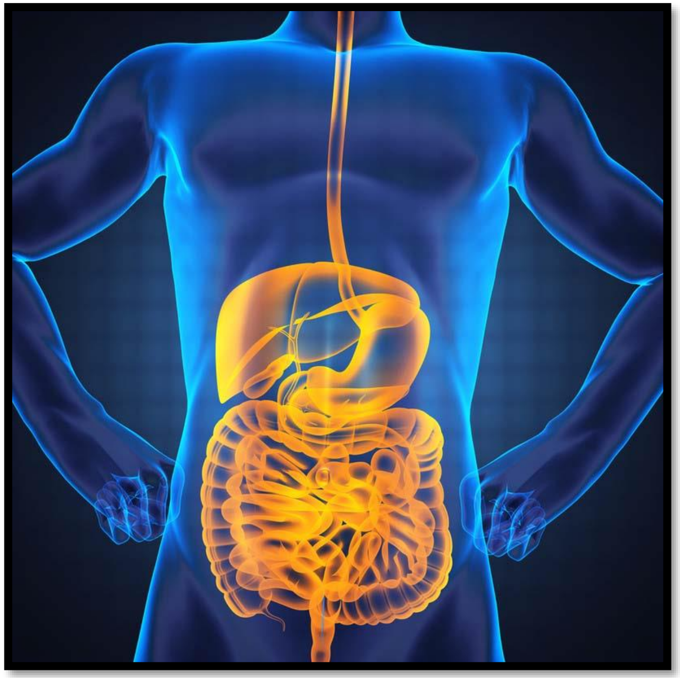
Sequencer

ACGATAGGTAC
GATAGTACGGT
ACGTACGTGAC
AGGTTTTACCC
CGGATATCCAT
TAAATTTAACG
AATACGATAAA
GTAATAAATGT

Reads

Example: The human gut

- The human gut consists of a complex ecosystem containing thousands of bacterial species.
- Changes in the composition is connected to several diseases, such as obesity and diabetes.
- Identification of these changes using metagenomics can be used for diagnostics.



Summary of todays lecture

- During the last 10 year, next generation sequencing has transformed life sciences
- Compared to Sanger sequencing, the next generation sequencing platforms has higher output but the data is generally more prone to errors.
- Common short read platform is Illumina, which also has the highest throughput and thus the lowest cost
- Common long read platforms are PacBio and Oxford Nanopore which both offers read lengths > 100kb