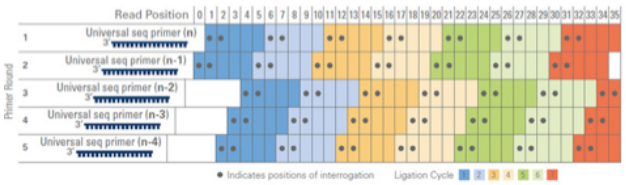
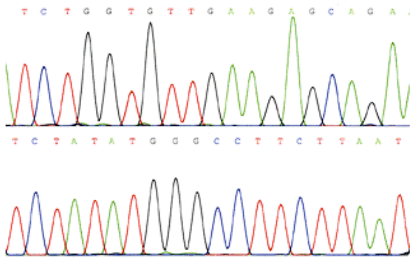
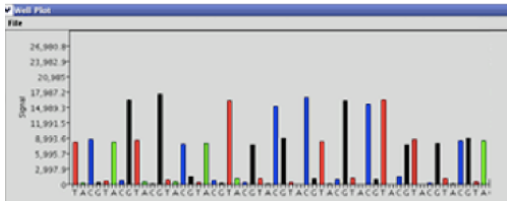


Overview of different types of next generation sequencing

Dr. Anne Jores-Fischer

24th November 2014

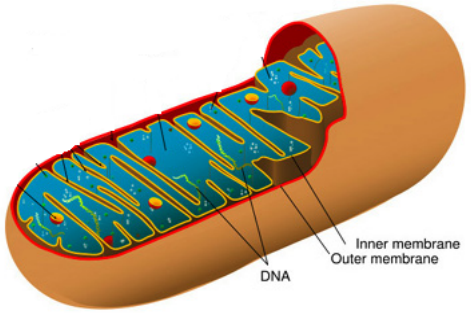




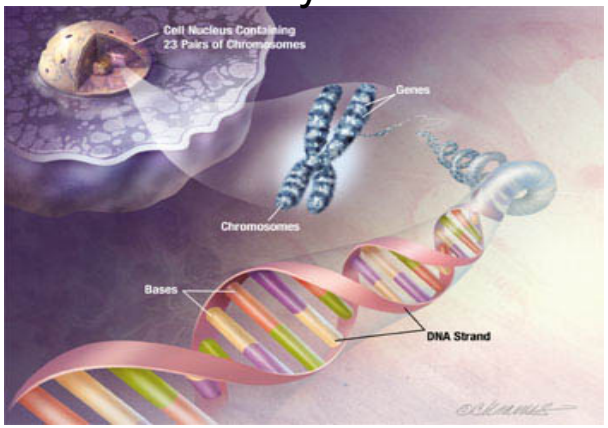
What is sequencing?

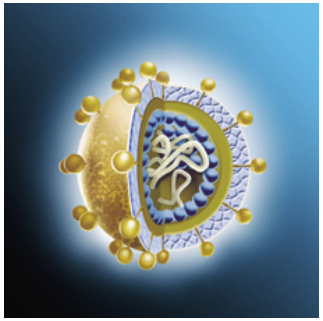
- Find the order of the nucleotide along a piece of DNA/RNA

Prokaryotes



Eukaryotes



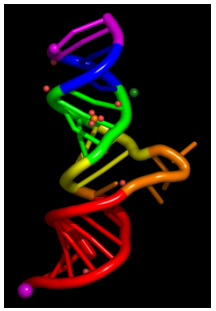


Virus

Why sequence?

- Nucleotide order determines Amino acids order and therefore protein sequence, structure and function

DNA → RNA → Protein → Phenotype

What is NGS?

- Sequence both DNA and RNA (with modified protocols)
- Short reads
- High-throughput

Outline

- Overview of sequencing technologies
  - First generation: Sanger
  - Second generation:
    - 454
    - Illumina
    - ABI solid
    - Ion torrent
  - Third generation:
    - Pacific Bioscience SMRT
    - Oxford nanopore
- Biological applications

Sequencing technologies

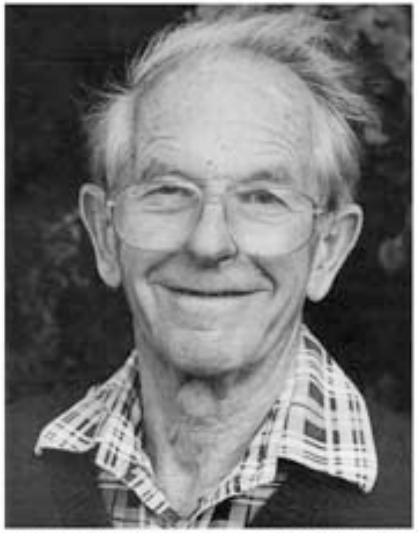
- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

1st generation sequencing

- Sanger sequencing or "chain termination" method

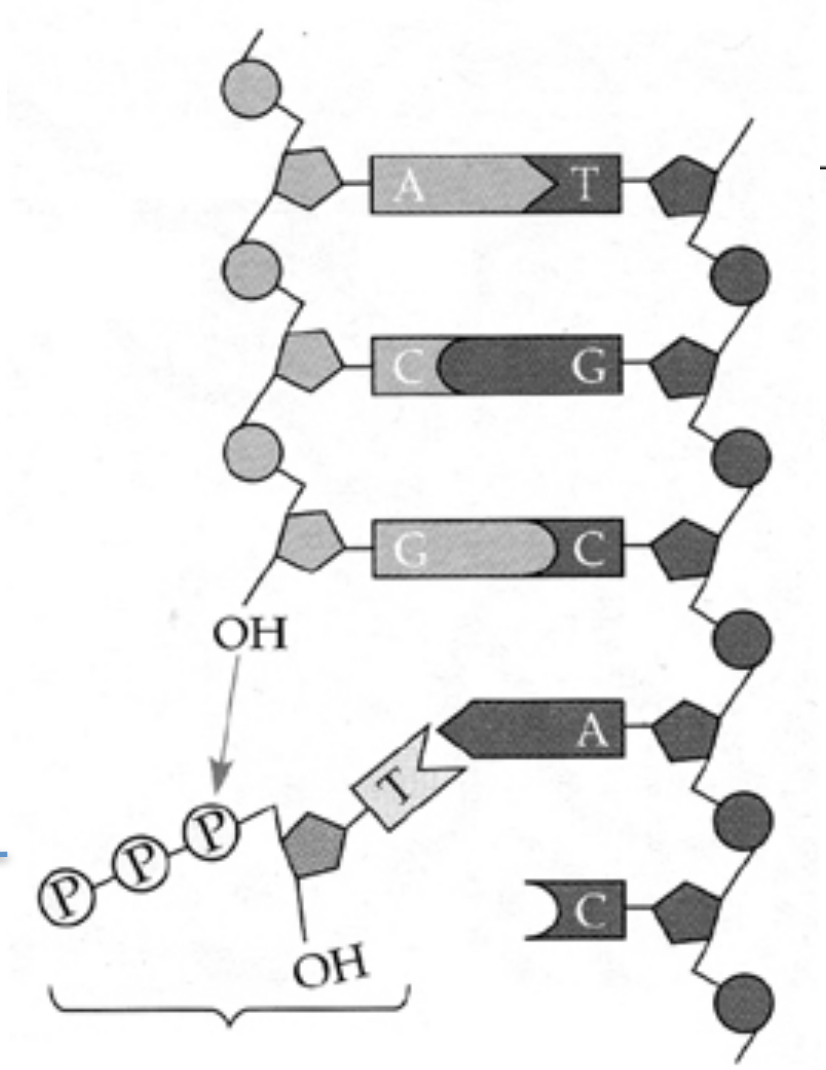
- Developed by Fred Sanger in 1977

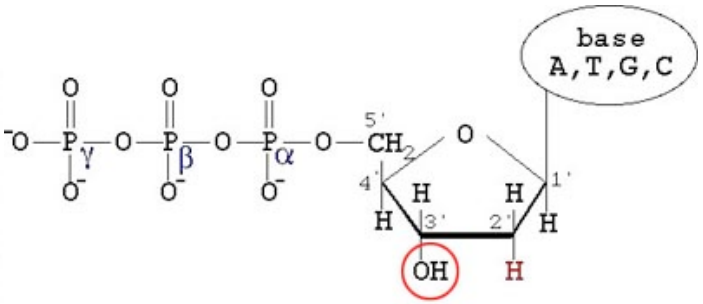
(Nobel prize in 1980)

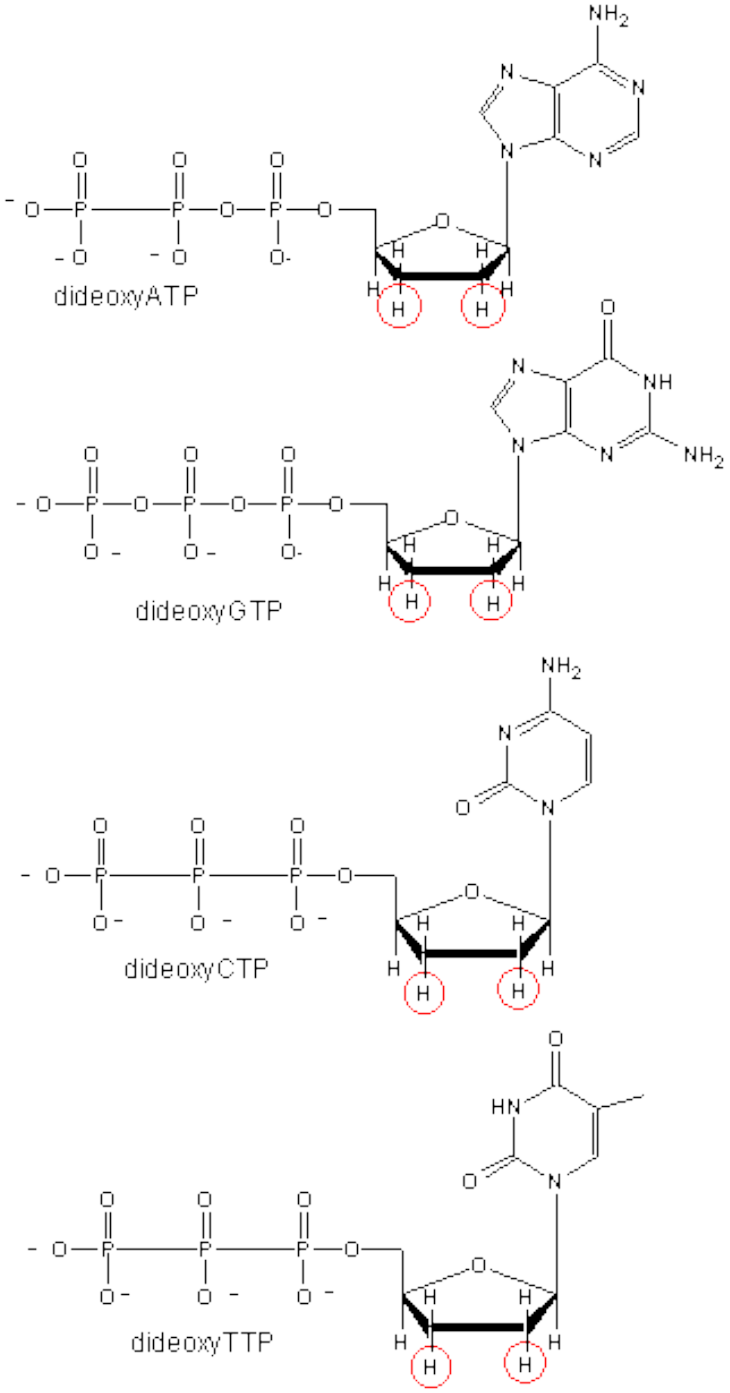


Courtesy of Dr. F. Sanger, MRC, Cambridge.
Noncommercial, educational use only.

Deoxyribonucleic acid bonding





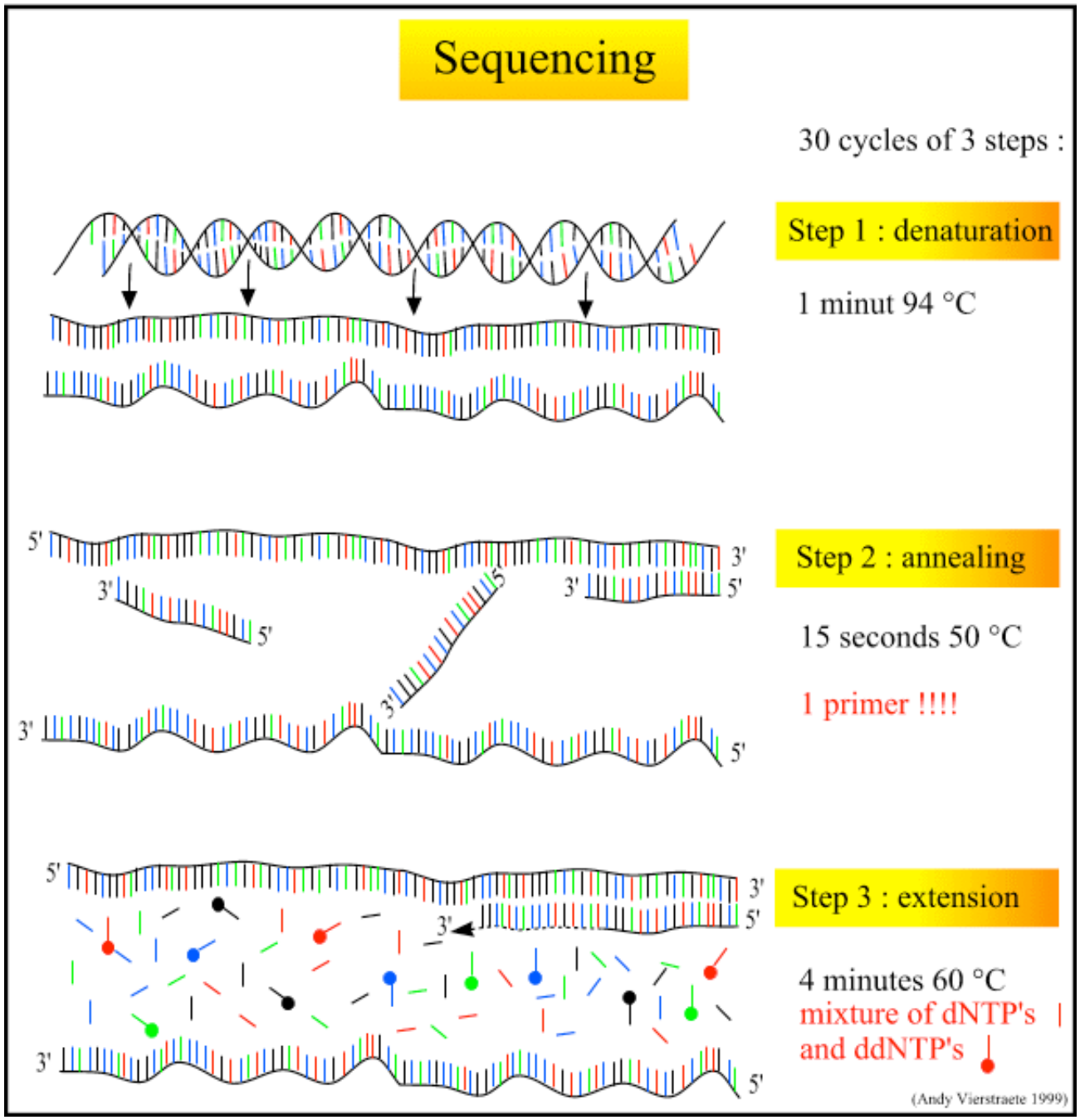


Mixture of Deoxyribonucleic acids (dNTPs) and **Dideoxyribonucleic acids (or ddNTPs)**

Replication proceeds 5’ ->3’ :
Polymerase can add ddNTPs to a strand, but after that the strand cannot be extended – replication is **halted**!

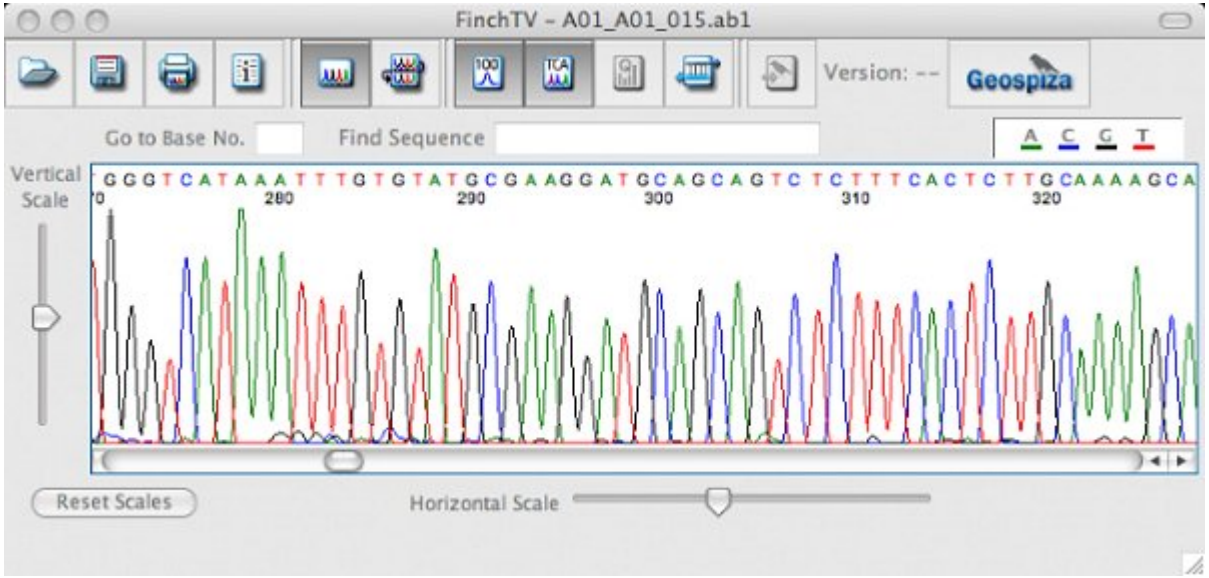
Requires:

Template DNA
Primer
DNA polymerase
A pool of normal nucleotides
A fraction of labelled ddNTPs



1st generation sequencing

- Reading the sequence: chromatogram



- Sequence length 800-1200bp
  - 1 sequencing plate 96 wells

Time frame Sanger sequencing

- First genome sequenced: Haemophilus influenzae (1995): 1.8Mb of sequence in 7 months
- Human genome (2001): 3Gb of data (started in 1990, cost: 3 billions USD)
- Very accurate, but long and tedious

Sequencing technologies

- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

2nd generation sequencing

- Pål Nyrén (Royal Tech Institute, Stockholm 1986): pyrosequencing

- First high throughput pyrosequencing: 454 (Roche 2004)

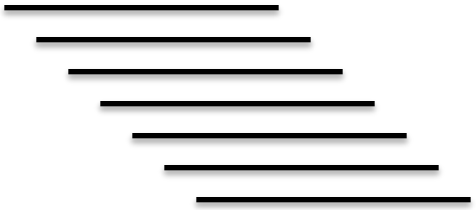
2nd generation sequencing

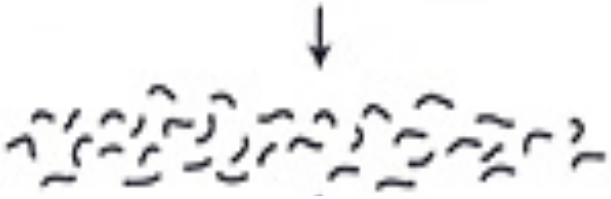
- Requires amplification of the material to be sequenced, ligation of adapters and PCR
- Sequence hundreds of millions of sequences at the same time in a single run
- Multiplex possible (barcode or primer)
- Single read, paired-end and mate pair

Library preparation

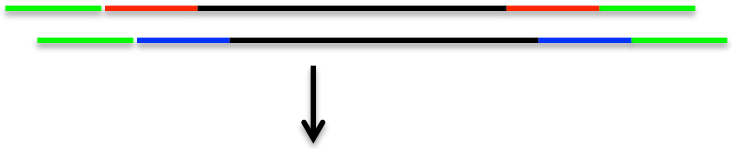
DNA/cDNA Shearing
(sonication or enzymatic

PCR products
Depending on the size will need to be sheared also



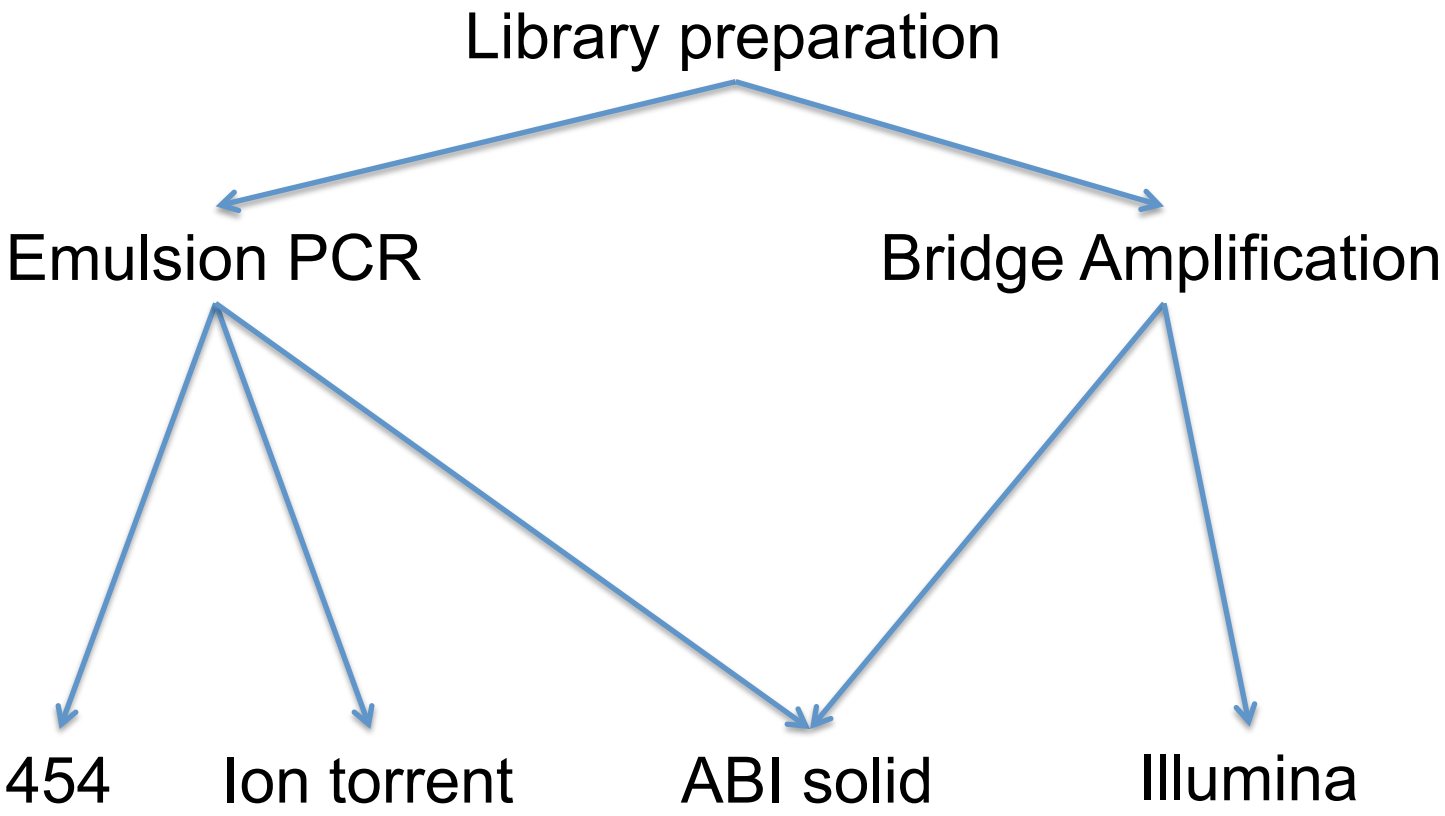


Adapter (and barcode) ligation

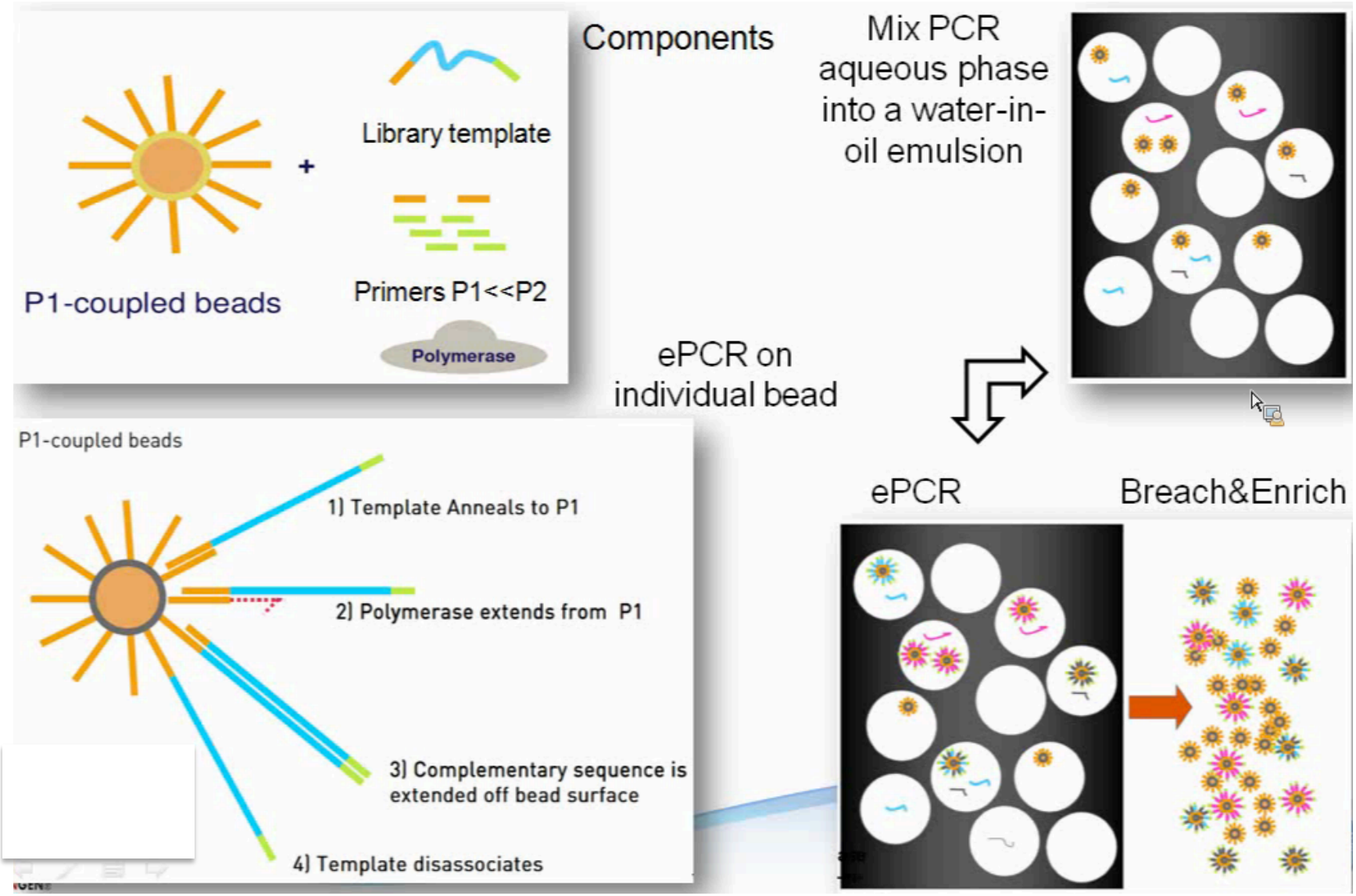


Size selection of fragments to be sequenced

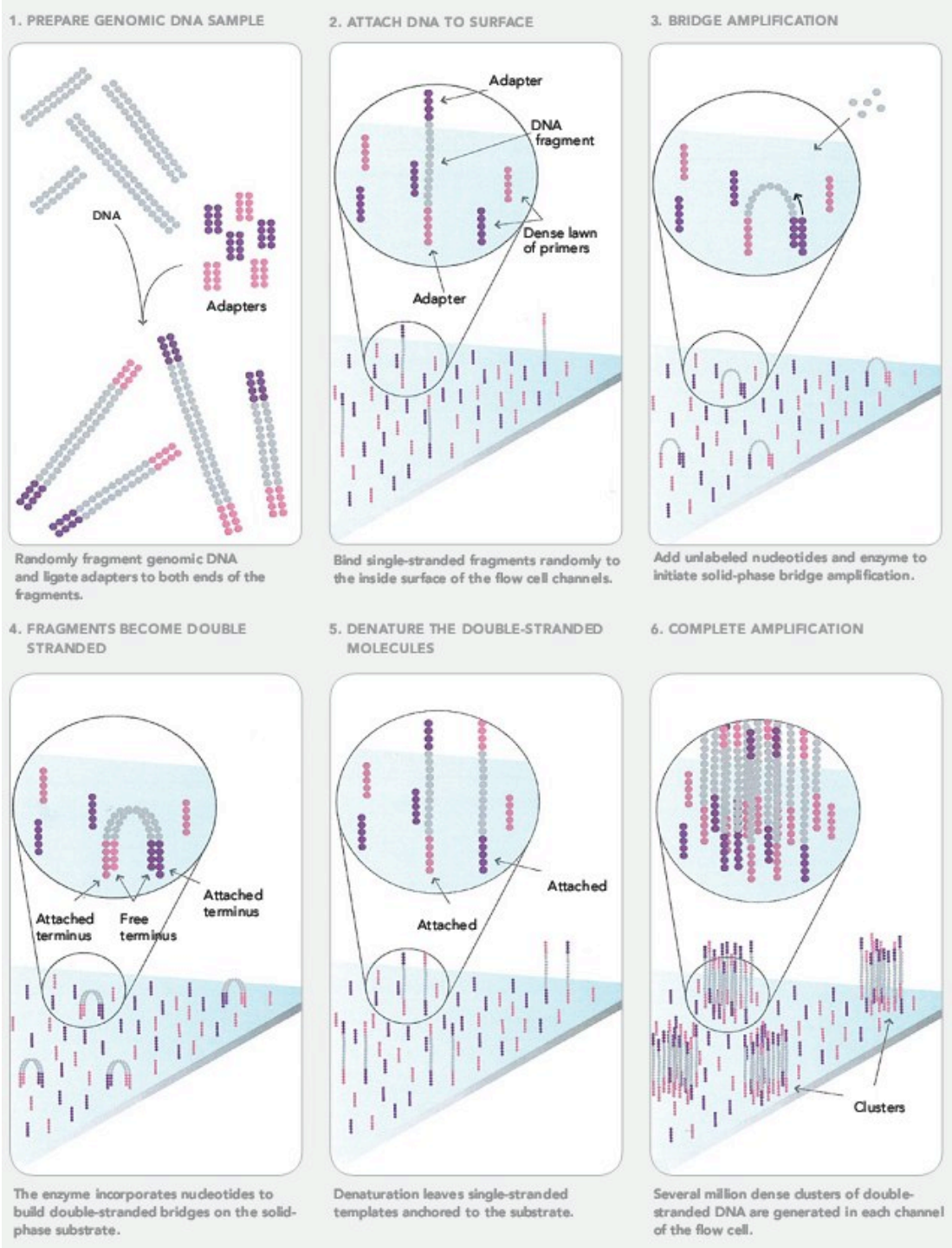
Single molecule amplification



Emulsion PCR (emPCR)



Bridge amplification

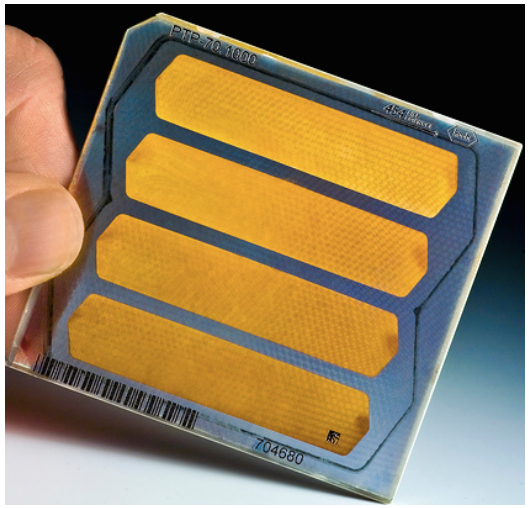


Sequencing technologies

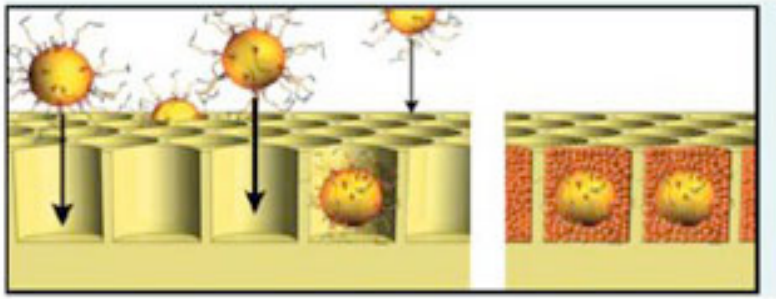
- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

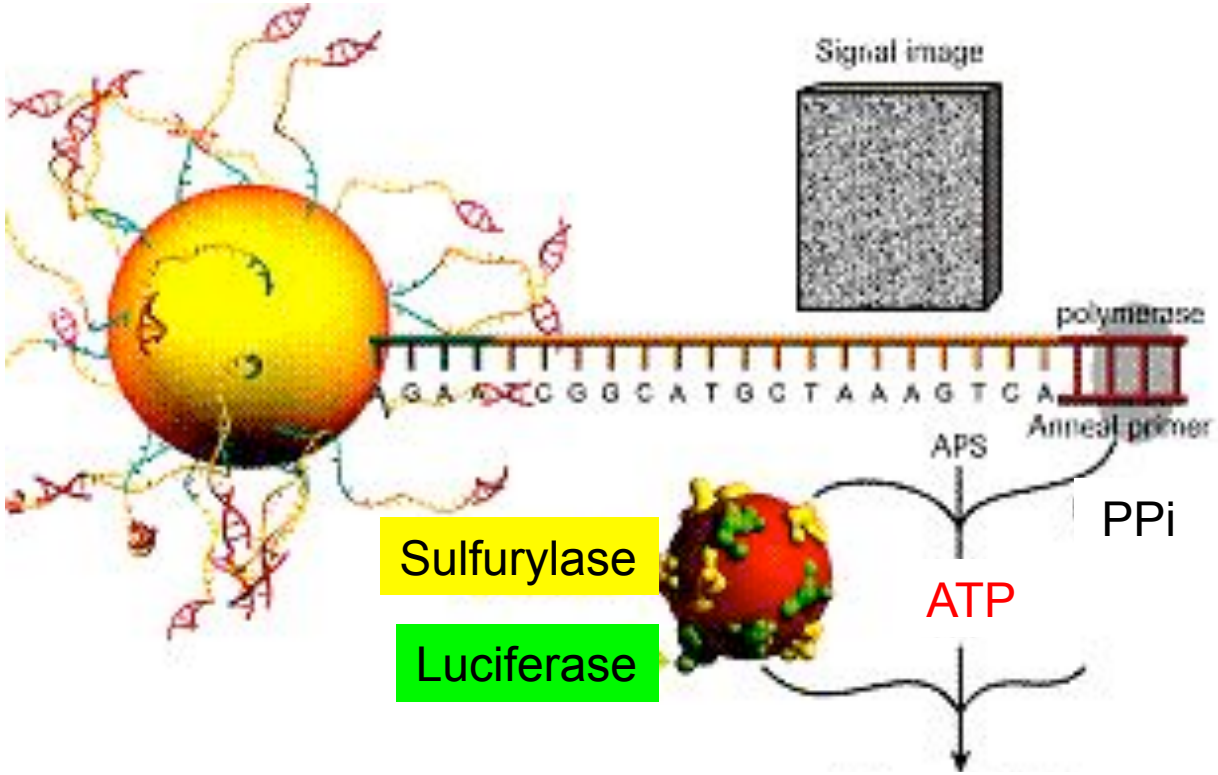
454: Pyrosequencing

- 1 PTP: 1M reads
- 700bp long reads
- 1 run in 24h
- Sequential flow of nucleotides
- Multiplexing possible

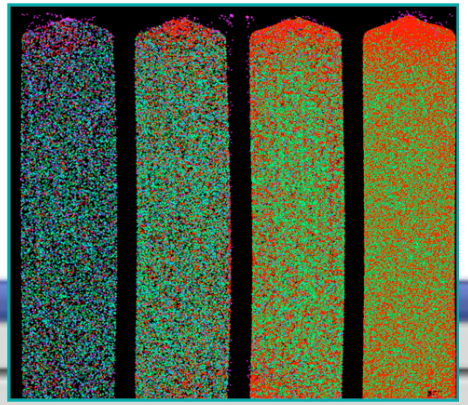


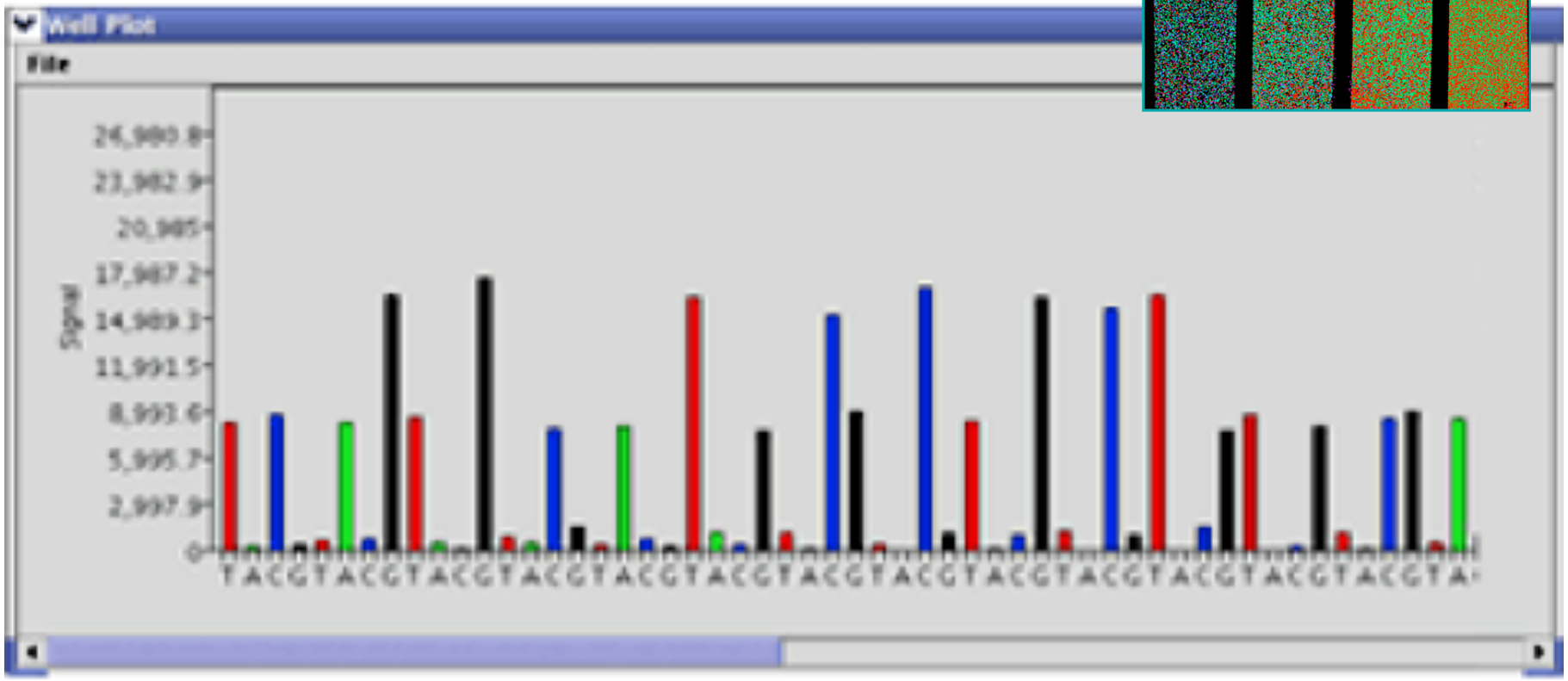
454: Pyrosequencing





454 Flowgram





Signal strength proportional to the number of nucleotides incorporated

454

- Longer reads
- Much lower throughput than other technologies
- 10$ per million bases
- Not so much used anymore

- First human genome (James Watson) in 2007: 2 million USD

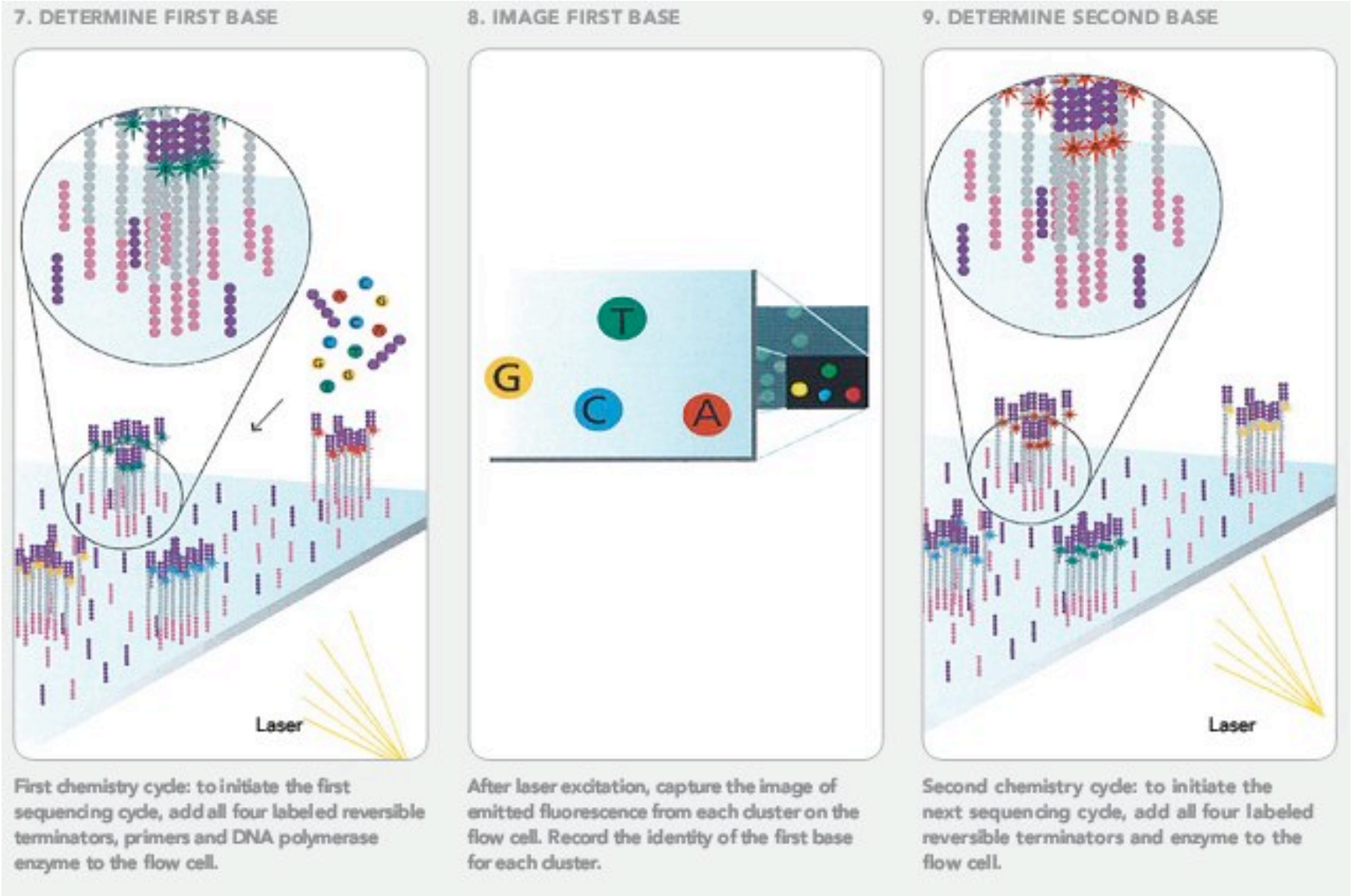
Sequencing technologies

- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

Illumina Hiseq-NextSeq

- 1 flow cell has 8 lanes
- 36, 75, 150, 250, 300 bp
- Single read or paired end
- Multiplexing possible
- 1 run 11days

Illumina sequencing by synthesis



Illumina HiSeq-Nextseq

- Up to 3 billion clusters
- 30x coverage of 2 human genomes on a single run
- 1 lane costs $1000
- 4k$ per sample

Sequencing technologies

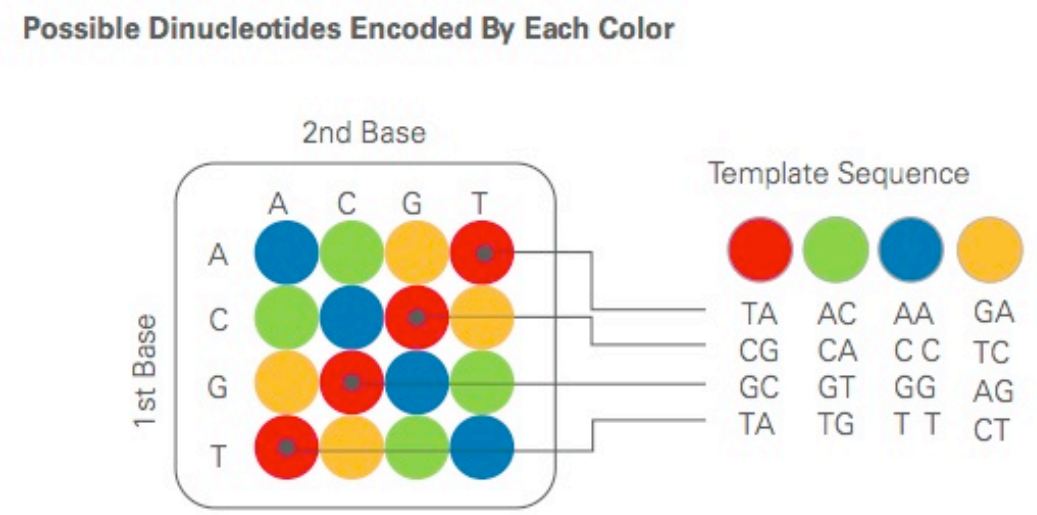
- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

ABI solid

- Short reads 100bp
- Run 7 to 14 days
- 2 flowchips: 1.2 to 1.4M reads
- 1 run 15,000$

ABI Solid: sequencing by ligation

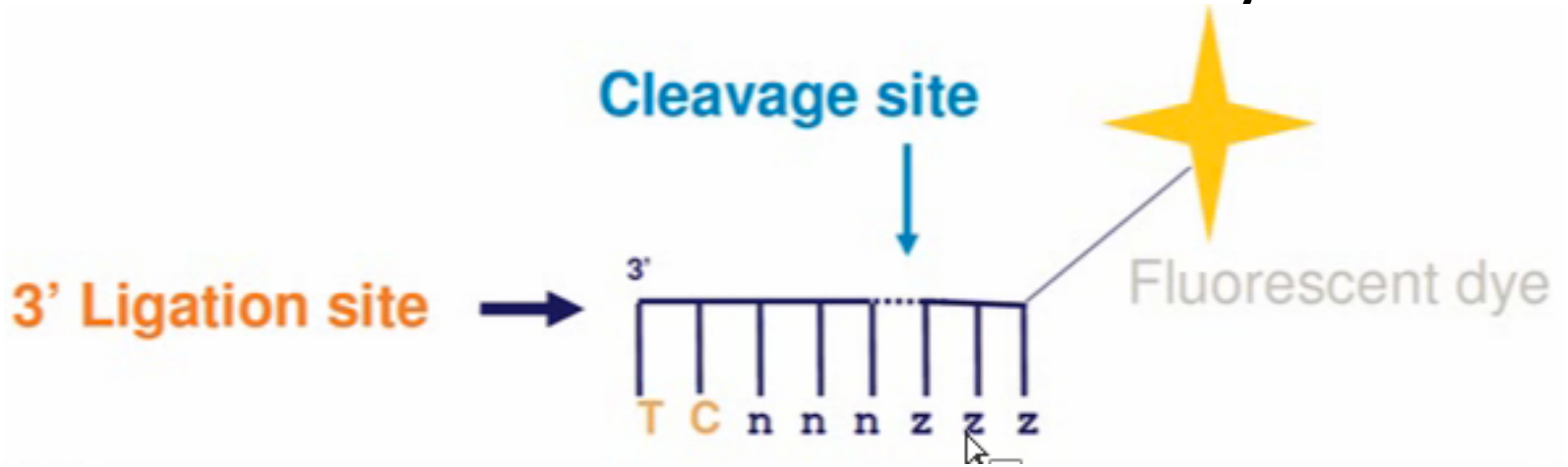
- Amplification by emulsion PCR
- Sequencing with labeled nucleotide ligated to one another



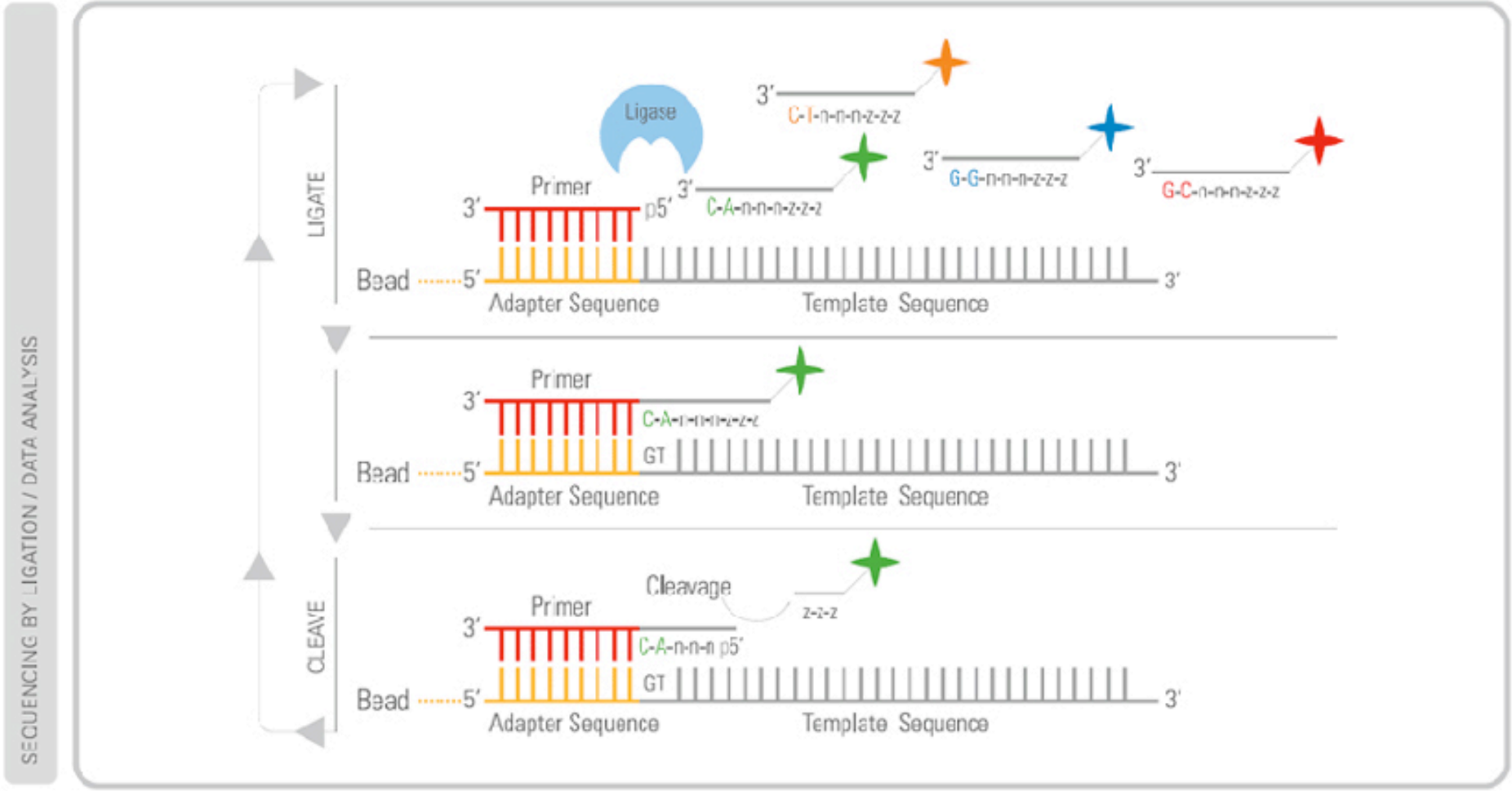
ABI Solid: sequencing by ligation

- 1024 octamer probes, 4 dyes
- Each dinucleotide has a color

- n degenerate bases
- Z universal base which binds to any nt

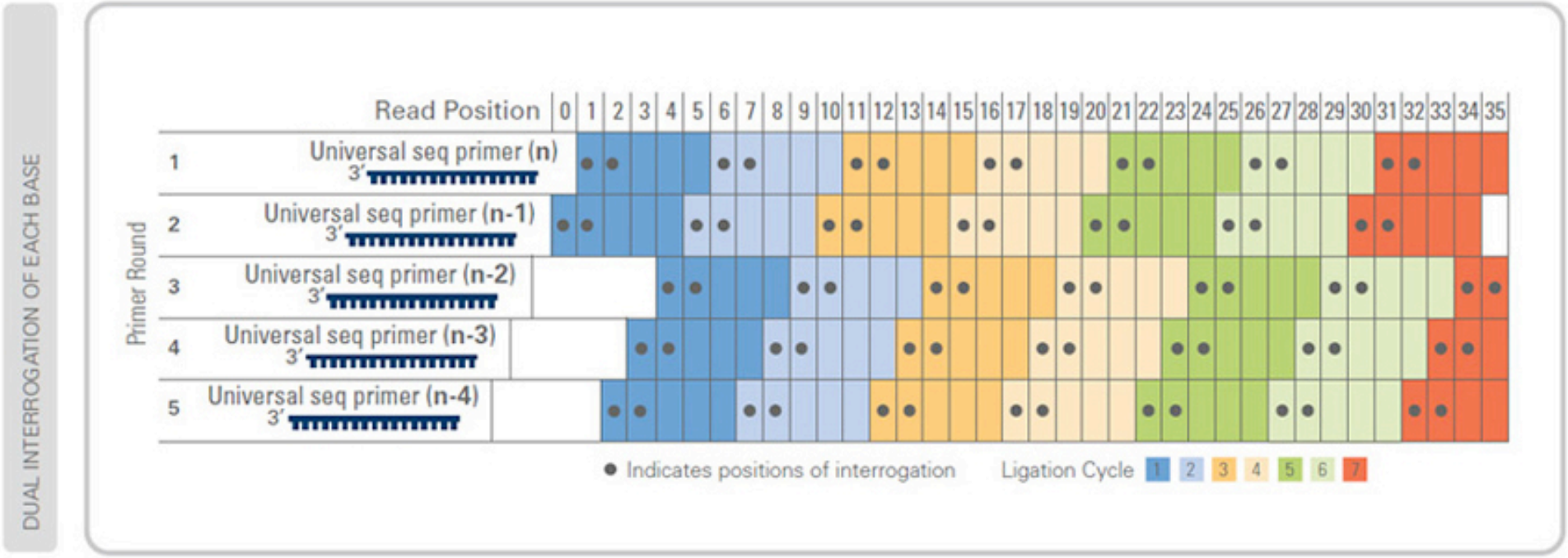


ABI Solid: sequencing by ligation



ABI Solid: sequencing by ligation

- Every single nt is sequenced twice:
  - Increased quality



Sequencing technologies

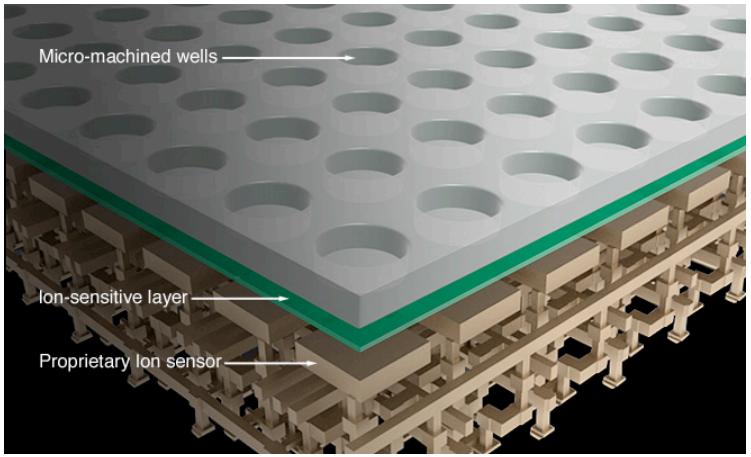
- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

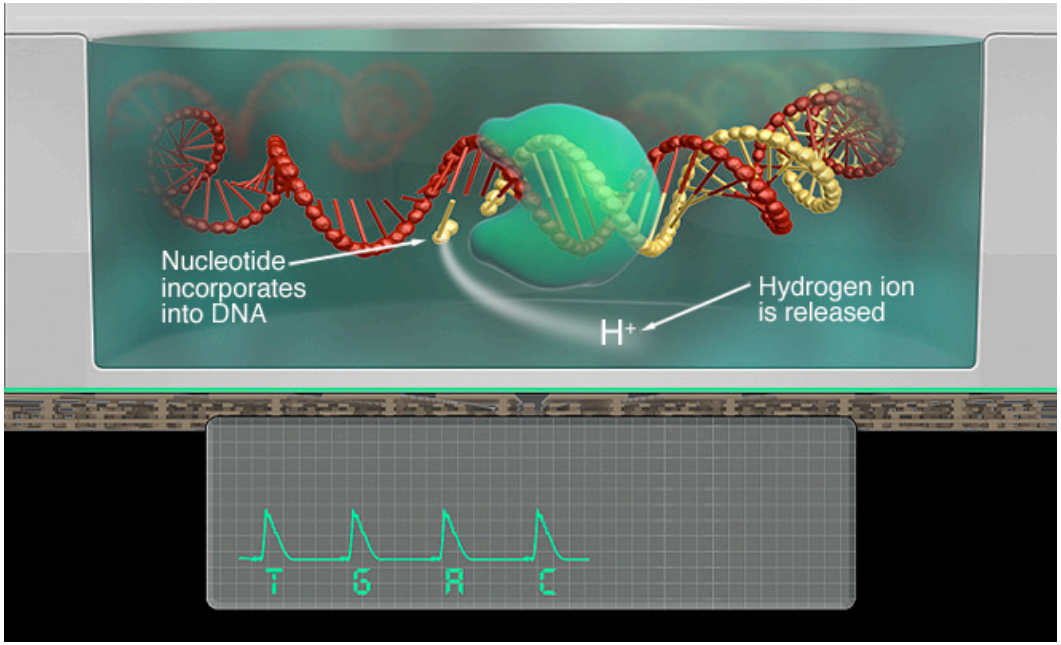
Ion torrent

- Machine is much cheaper: no camera needed to detect fluorescence
- Up to 80M reads
- Reads up to 400bp
- Runtime: 2-4 hours

Ion torrent

- Emulsion PCR
- Release of hydrogen ion when a base is incorporated, resulting in a change in pH
- Nucleotides flow sequentially





Sequencing technologies

- First generation: Sanger
- Second generation:
  - 454
  - Illumina
  - ABI solid
  - Ion torrent
- Third generation:
  - Pacific Bioscience SMRT
  - Oxford nanopore

Third generation sequencing

- Single molecule sequencing
- No PCR amplification and therefore no errors incorporated

Pacific Bioscience: PacBio SMRT

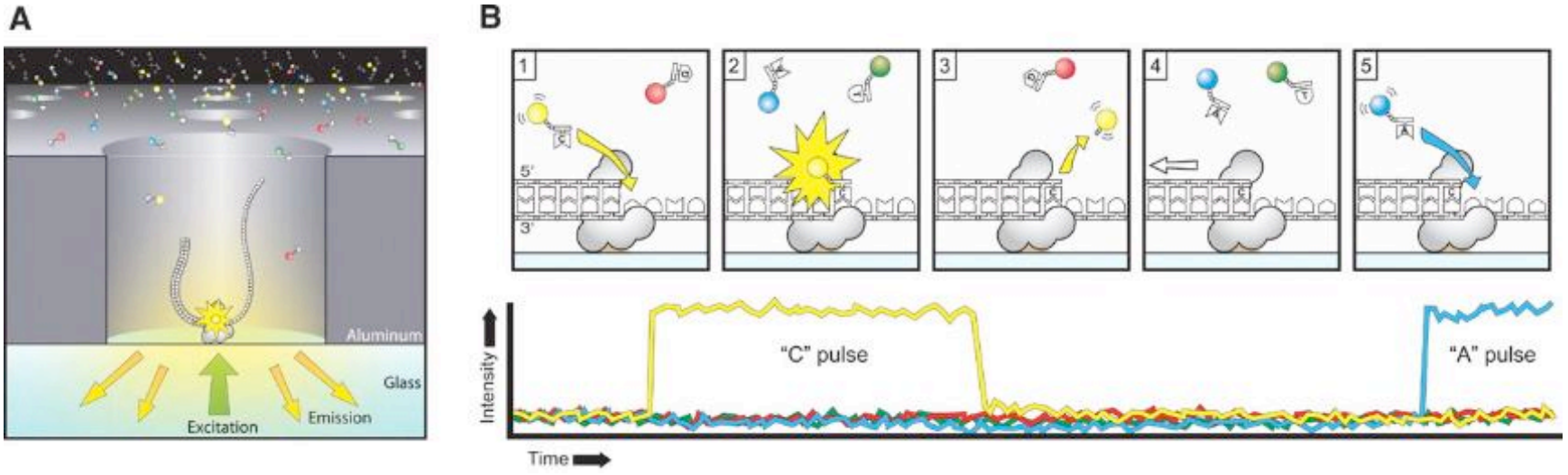
- Reads 3,000 to 15,000 bases
- Output 1GB
- Short runtime 30 minutes
- 1000$ per run
- Can detect DNA modifications: methylation

- Less accurate, higher error rate

Pacific Bioscience SMRT

SMRT: Single Molecule Real Time

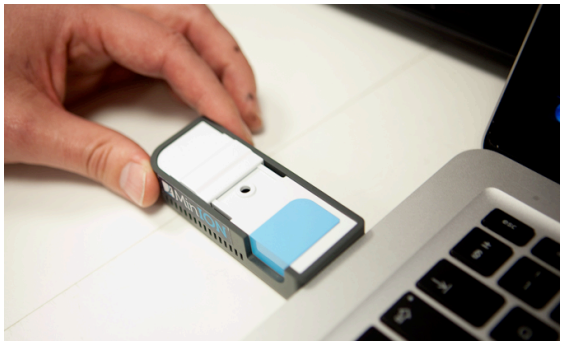
By synthesis: nucleotide incorporation and signal detection occur continuously
as fast as the polymerase incorporates nucleotides (750 nt/sec)

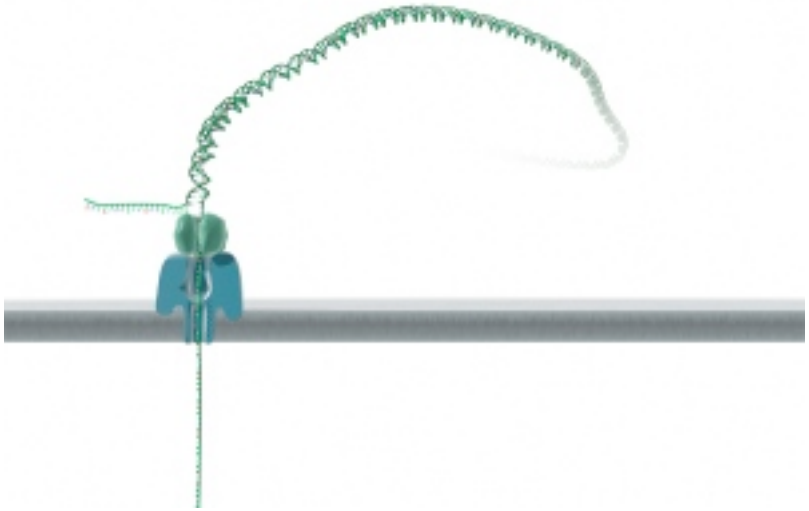


Oxford nanopore MInION

- Very long read??
- High error rate

http://vimeo.com/81212781





Summary

Method	Read length	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadvantages
Chain termination (Sanger sequencing)	400 to 900 bp	N/A	20 minutes to 3 hours	$2400	Long individual reads. Useful for many applications.	More expensive and impractical for larger sequencing projects.
Pyrosequencing (454)	700 bp	1 million	24 hours	$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	50 to 300 bp	up to 3 billion	1 to 10 days	$0.05 to $0.15	Potential for high sequence yield	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower than other methods. Have issue sequencing palindromic sequence
Single-molecule real-time sequencing (Pacific Bio)	10,000 bp to 15,000 bp avg	50,000 per SMRT cell, or 500-1000 megabases	30 minutes to 4 hours	$0.13–$0.60	Longest read length. Fast. Detects 4mC, 5mC, 6mA	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 400 bp	up to 80 million	2 hours	$1	Less expensive equipment. Fast.	Homopolymer errors.

Summary

- The rate of sequence data generation is more than doubling every year
- Data analysis is now the bottleneck

- That's why you are here!

Biological applications

- DNA
  - *de novo* sequencing: new genomes
  - Resequencing
  - Amplicon sequencing
  - SNP detection
  - Metagenomics
- RNA
  - RNAseq
  - miRNAseq
  - ncRNA
- DNA/RNA-Protein Interaction (transcription factor binding)
  - CHIP-seq
- Epigenetics:
  - Methylation
  - Histone modification

THANKS

Questions?