3000788 Intro to Comp Molec Biol

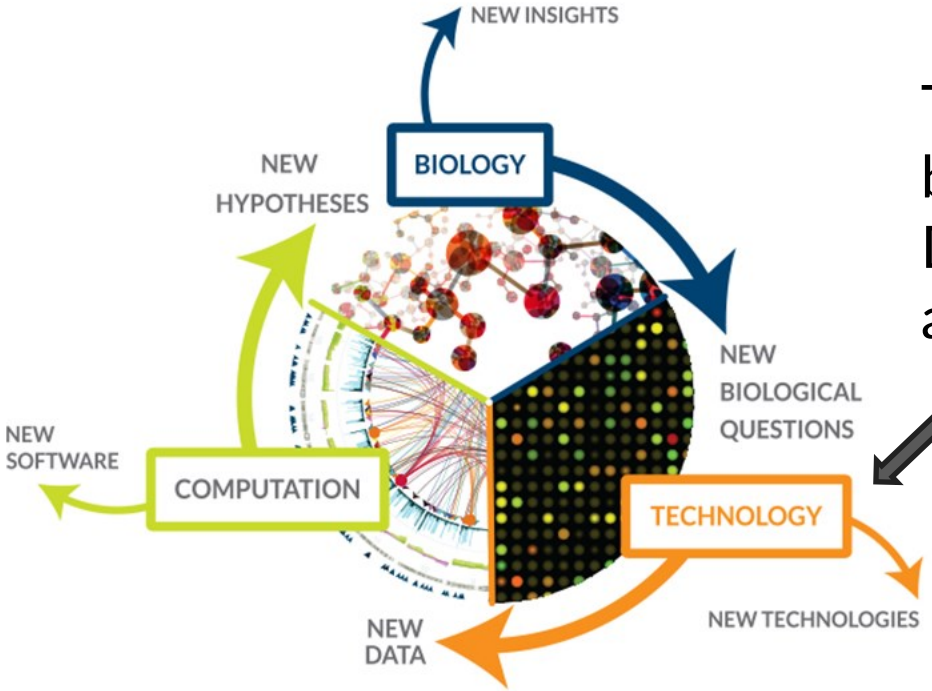
Lecture 4: DNA sequencing applications

August 28, 2023

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

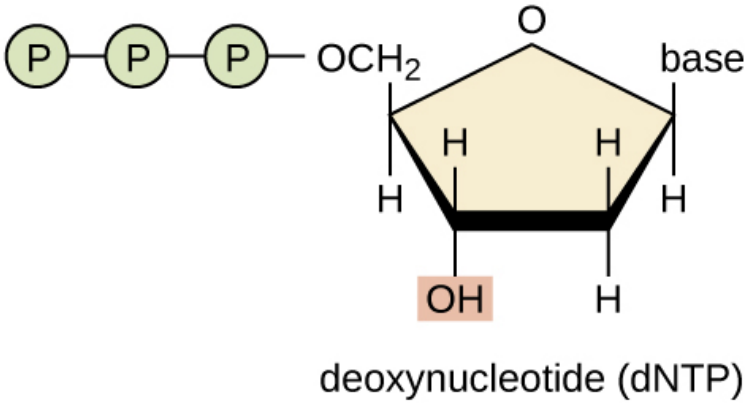
Technology for addressing biological questions



Today's content is about biological questions and how DNA sequencing help us address them

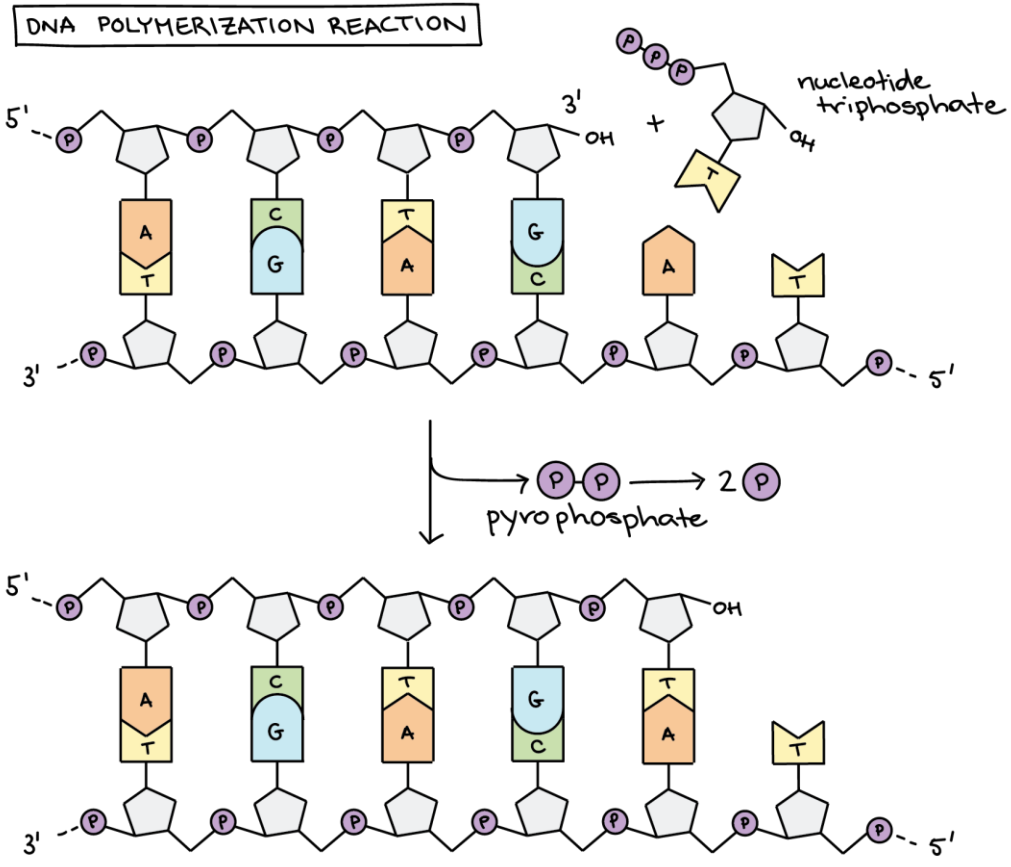
www.systemsbiology.org/about/what-is-systems-biology

DNA polymerization



deoxynucleotide (dNTP)

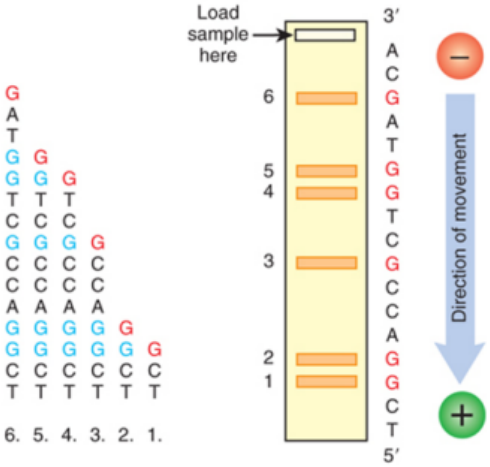
http://www.onlinebiologynotes.com/sangers-method-gene-sequencing/

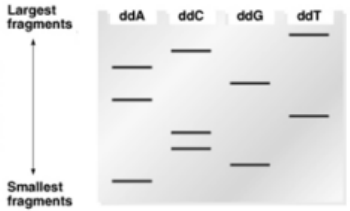


https://www.khanacademy.org

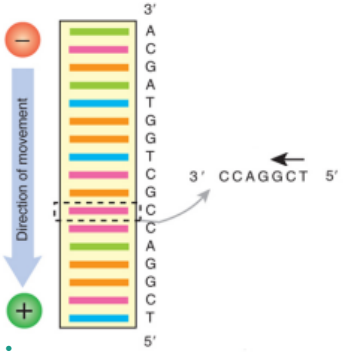
Sanger sequencing

Know only the last nucleotide





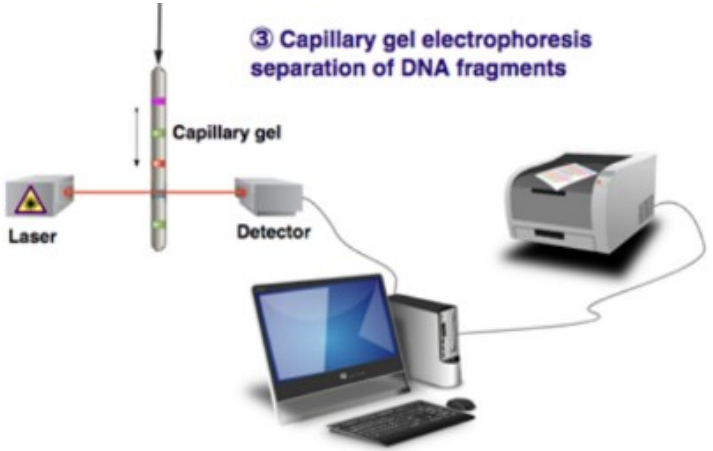
What is the sequence?

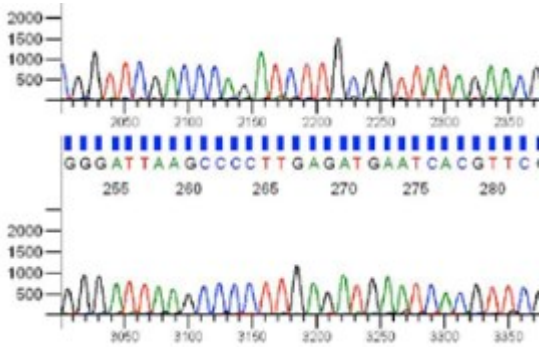


Generate all possible products, each with different length

Fluorescence-labeled ddNTP

Product length = bp position

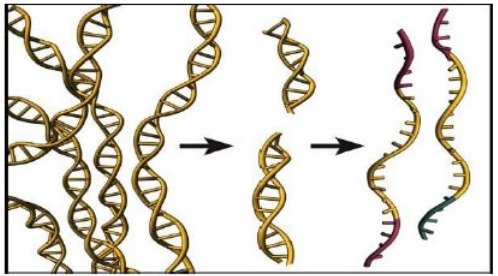




Images from https://en.wikipedia.org/wiki/Sanger_sequencing

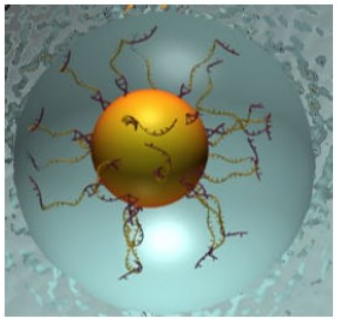
Next-Generation Sequencing (NGS)

High throughput from parallel reactions

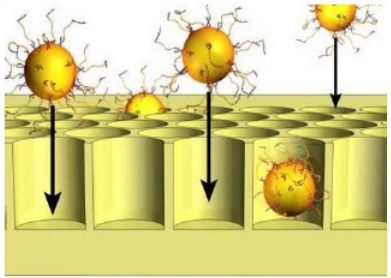


1) Adapter-ligated ssDNA library

Roche & Ion Torrent wells

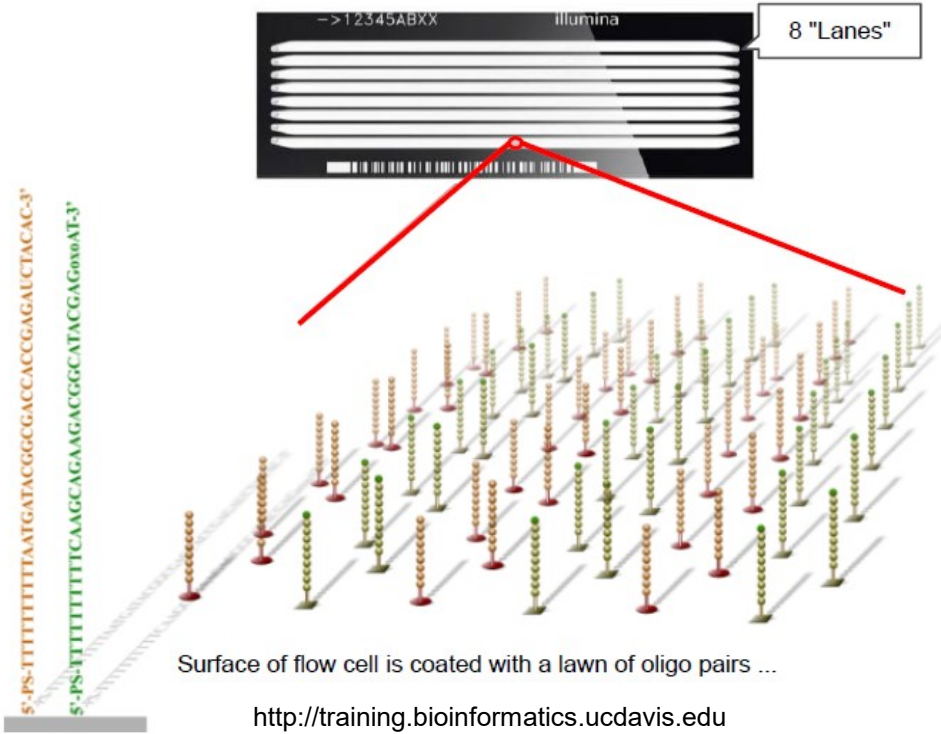


2) Clonal amplification on 28 micron beads ... emulsion PCR



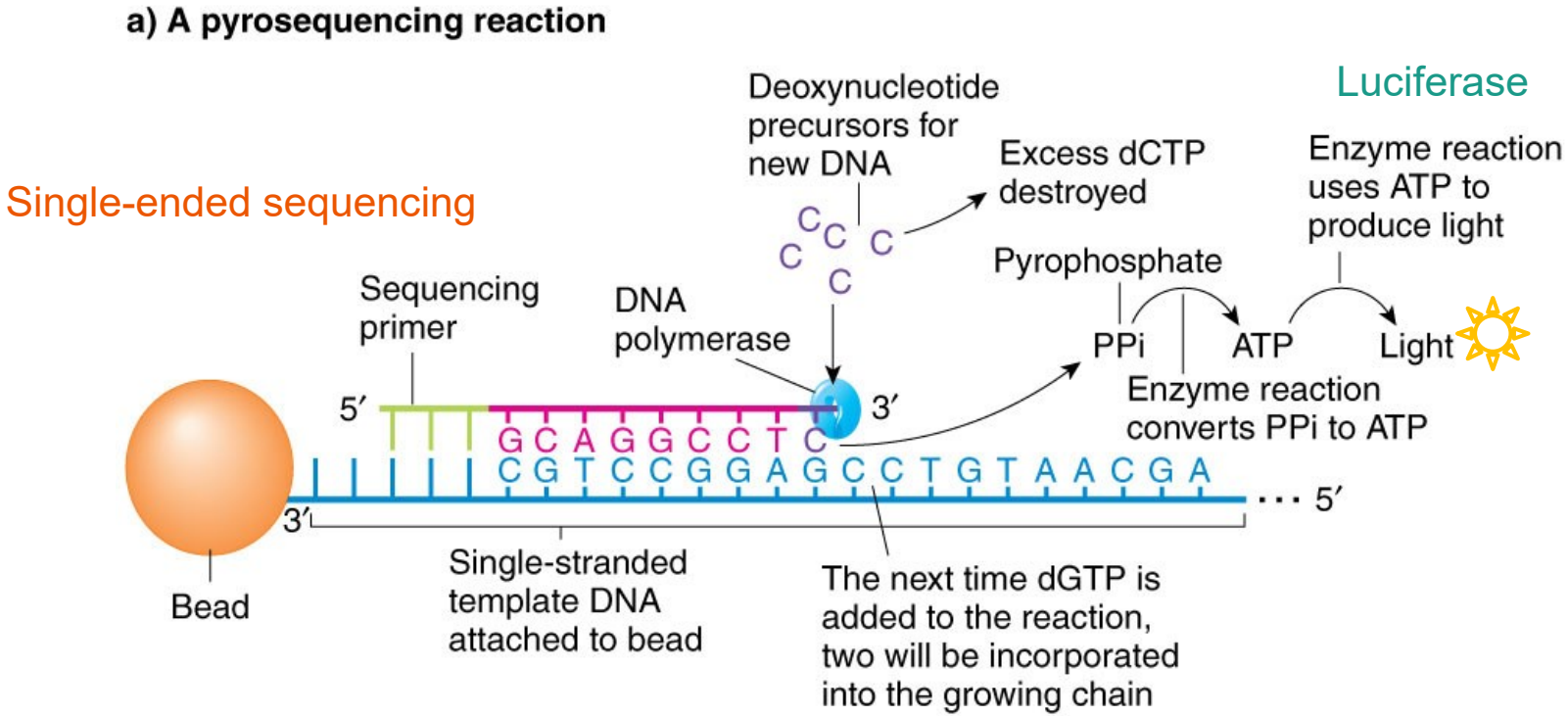
3) Beads deposited on PicoTiterPlate wells

Illumina's flow cell

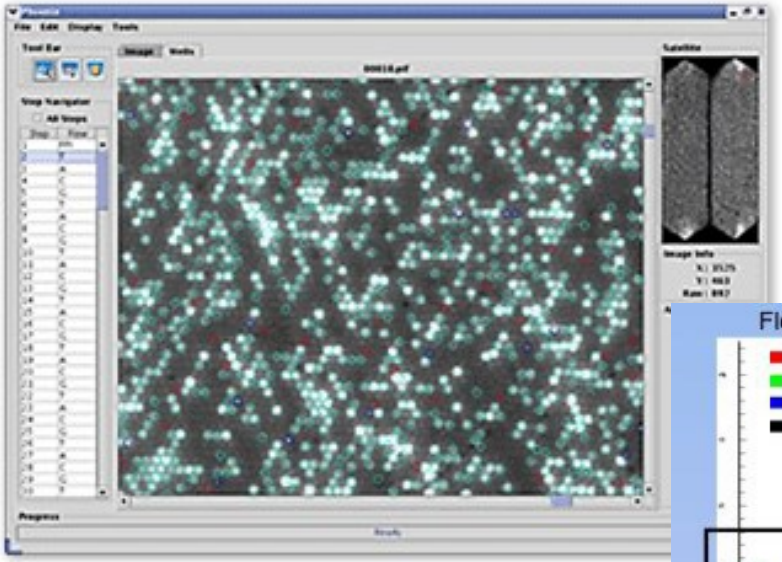


Pyrosequencing

Single-ended sequencing

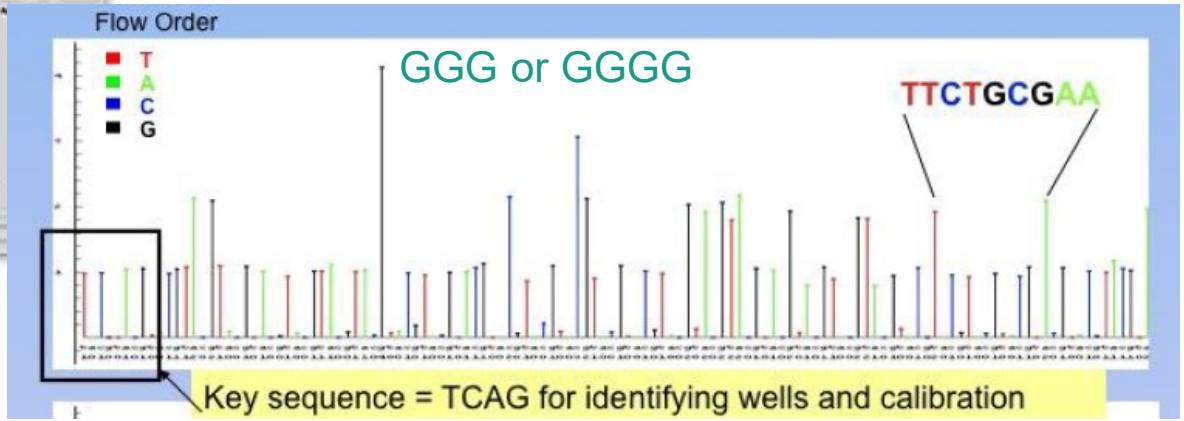


Limitation of pyrosequencing

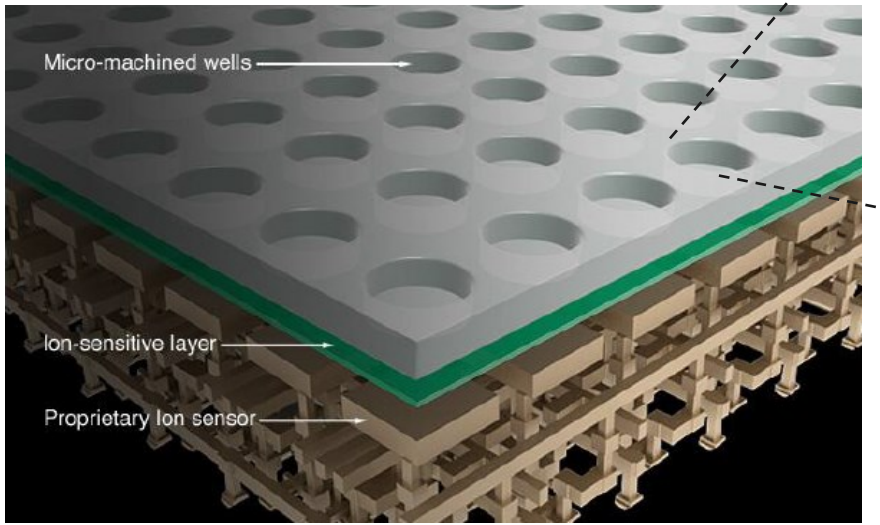


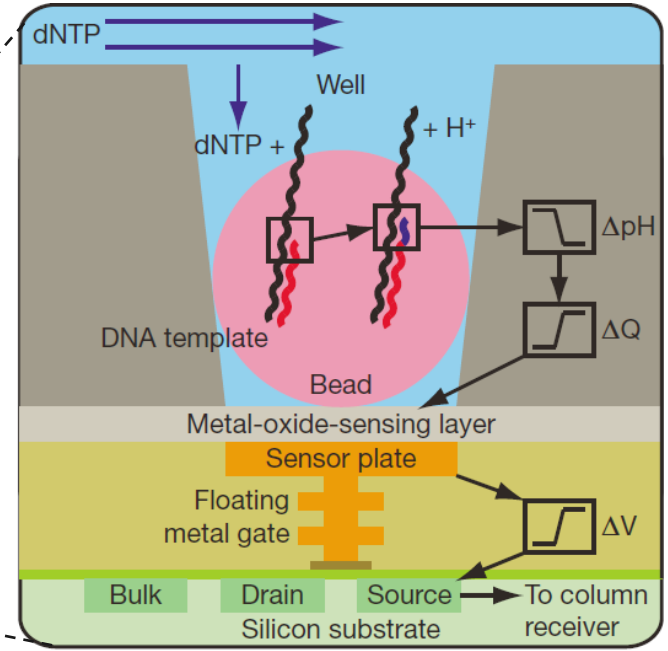
DATA ANALYSIS: OUTPUT PACKAGE

- Difficult to distinguish **homopolymer** of various lengths



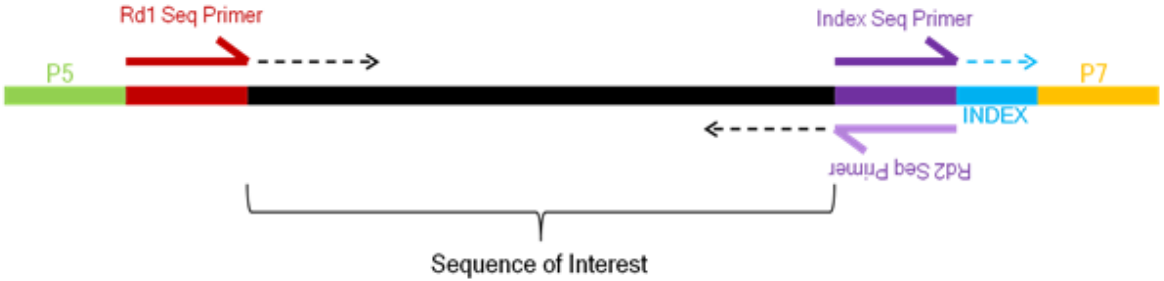
Ion Torrent

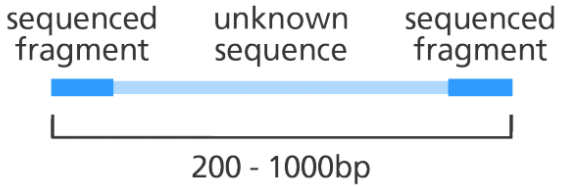




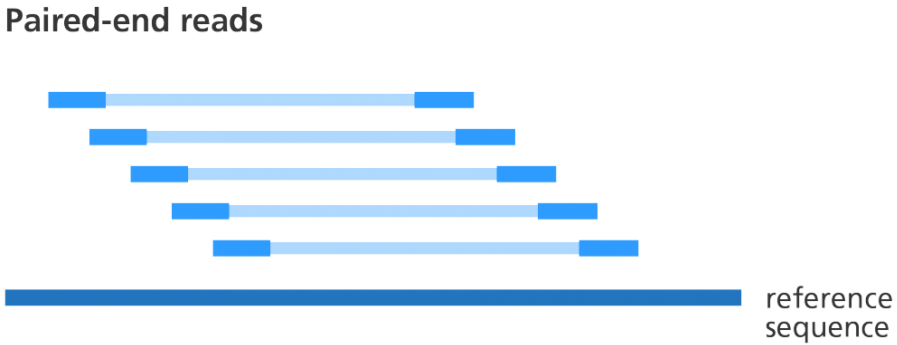
- Measure changes in pH
- Also has **homopolymer** limitation

Illumina / Solexa





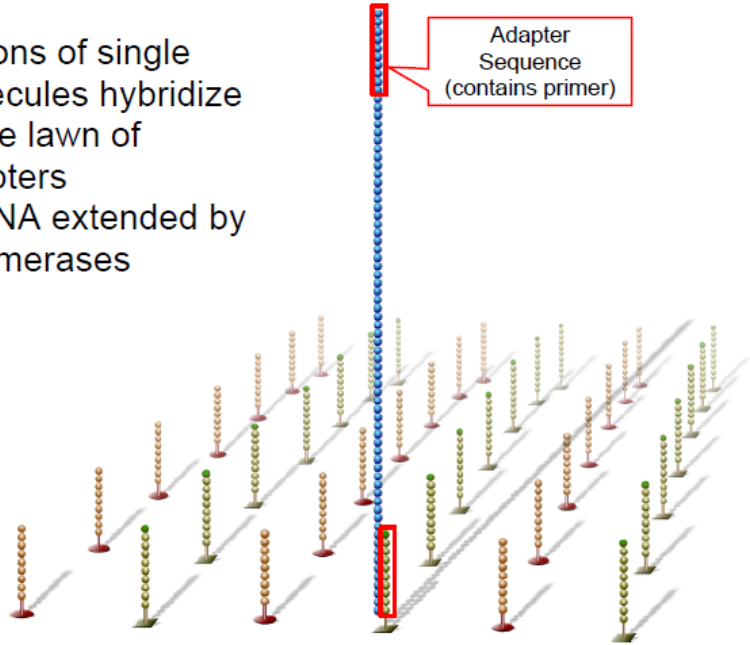
- Enable paired-end sequencing
- Improve mappability
- Identify splice junction
- Identify gene fusion
- Identify translocation



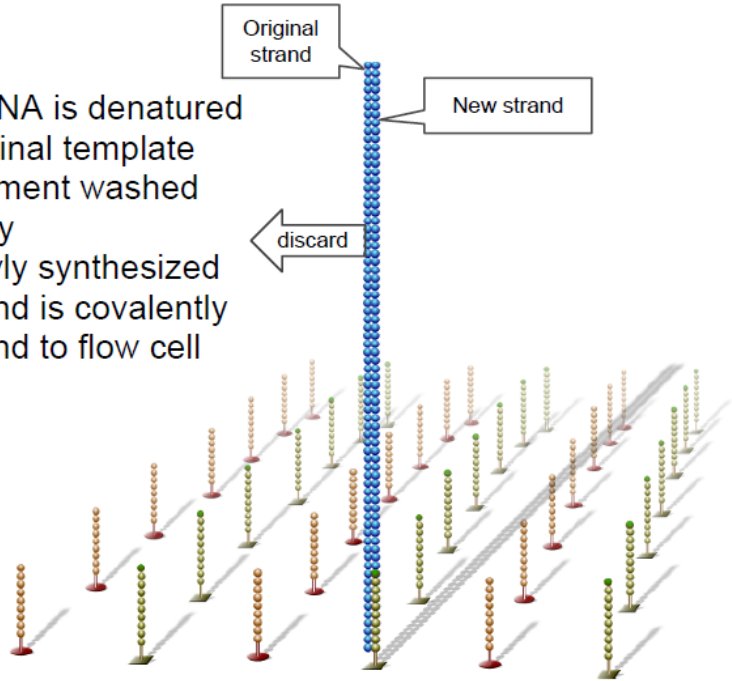
https://www.biostars.org/p/267167/

Amplification step 1

- Millions of single molecules hybridize to the lawn of adapters
- dsDNA extended by polymerases

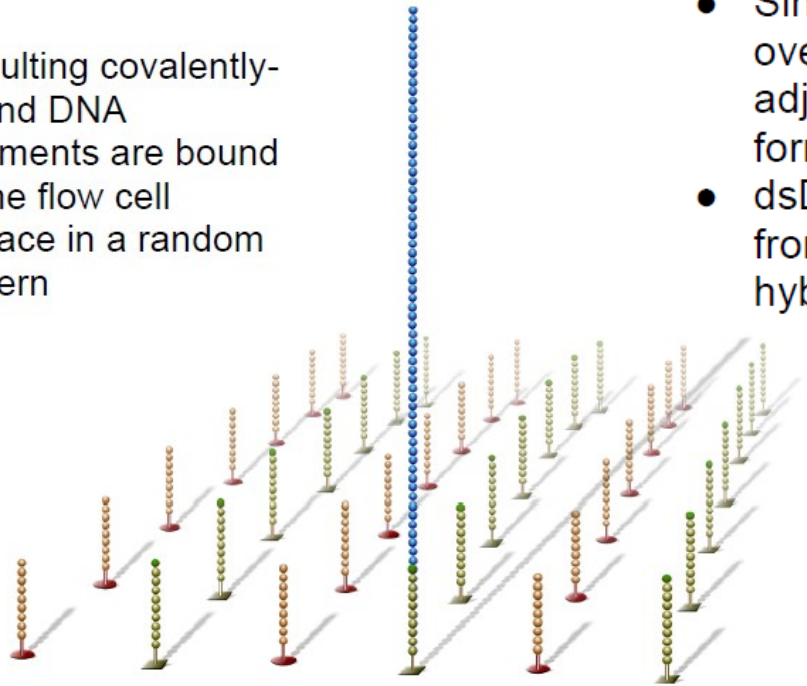


- dsDNA is denatured
- Original template fragment washed away
- Newly synthesized strand is covalently bound to flow cell

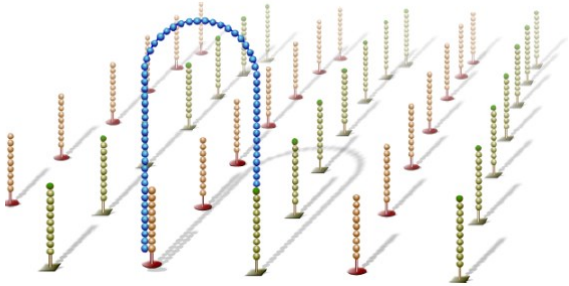


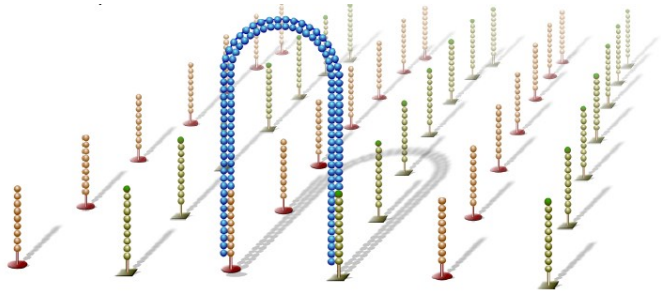
Amplification step 2

- Resulting covalently-bound DNA fragments are bound to the flow cell surface in a random pattern

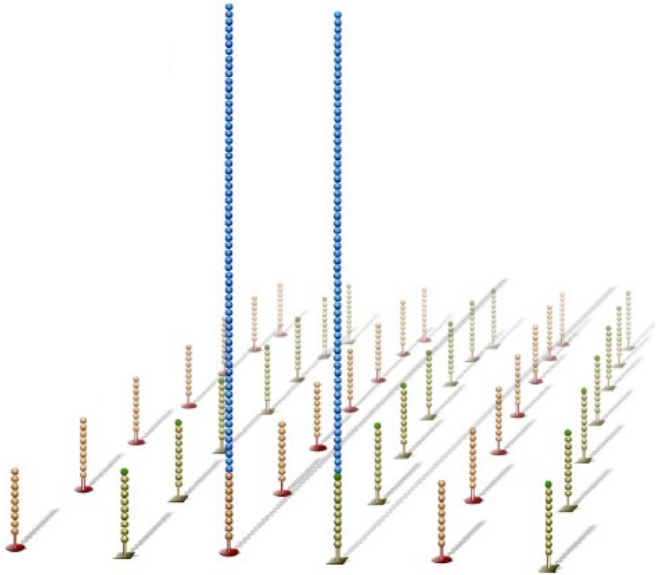


- Single-strand flops over to hybridize to adjacent adapter, forming a bridge
- dsDNA synthesized from primer in hybridized adapter

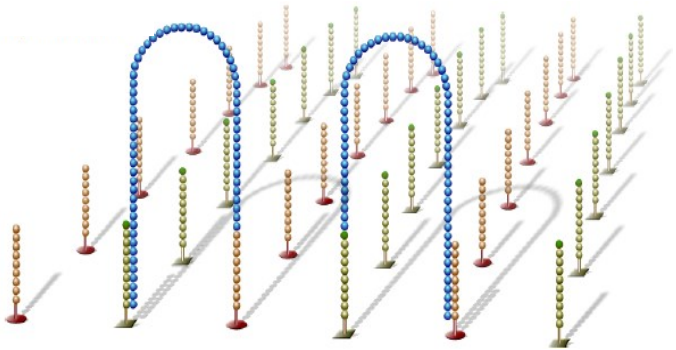




Amplification step 3



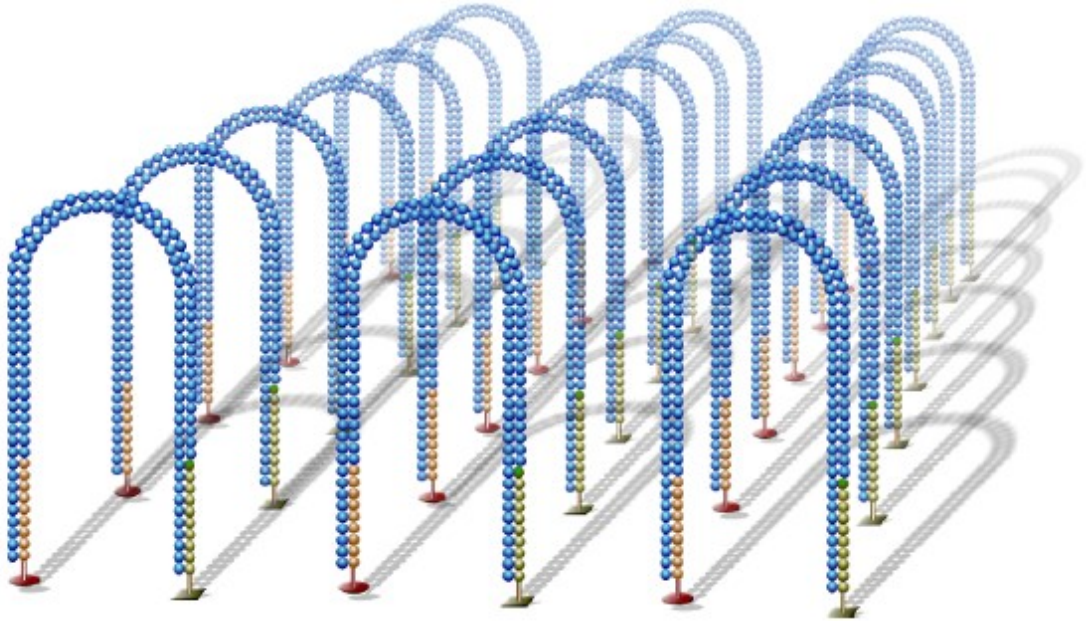
- dsDNA bridge is denatured



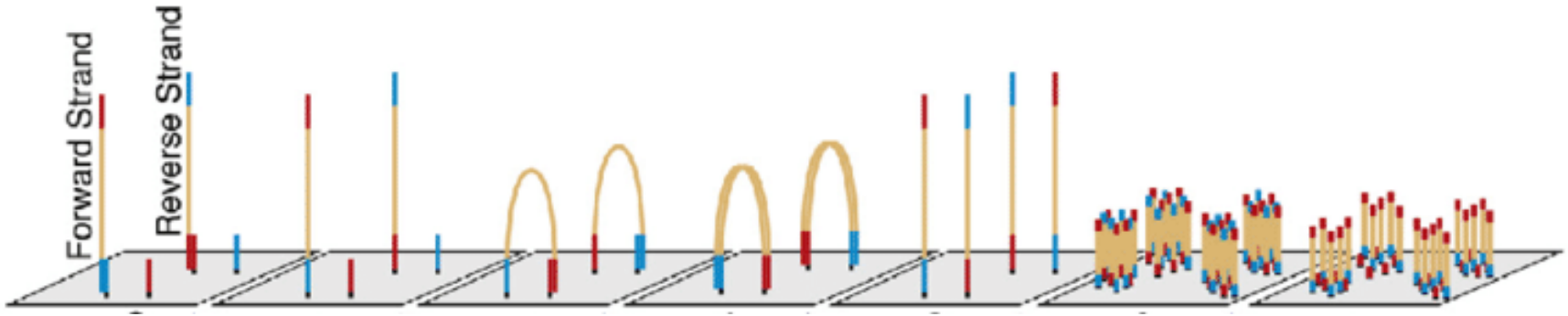
- Single strands flop over to hybridize to adjacent adapters, forming bridges
- dsDNA synthesized by polymerases

Amplification step 4

- Bridge amplification cycles repeated many times



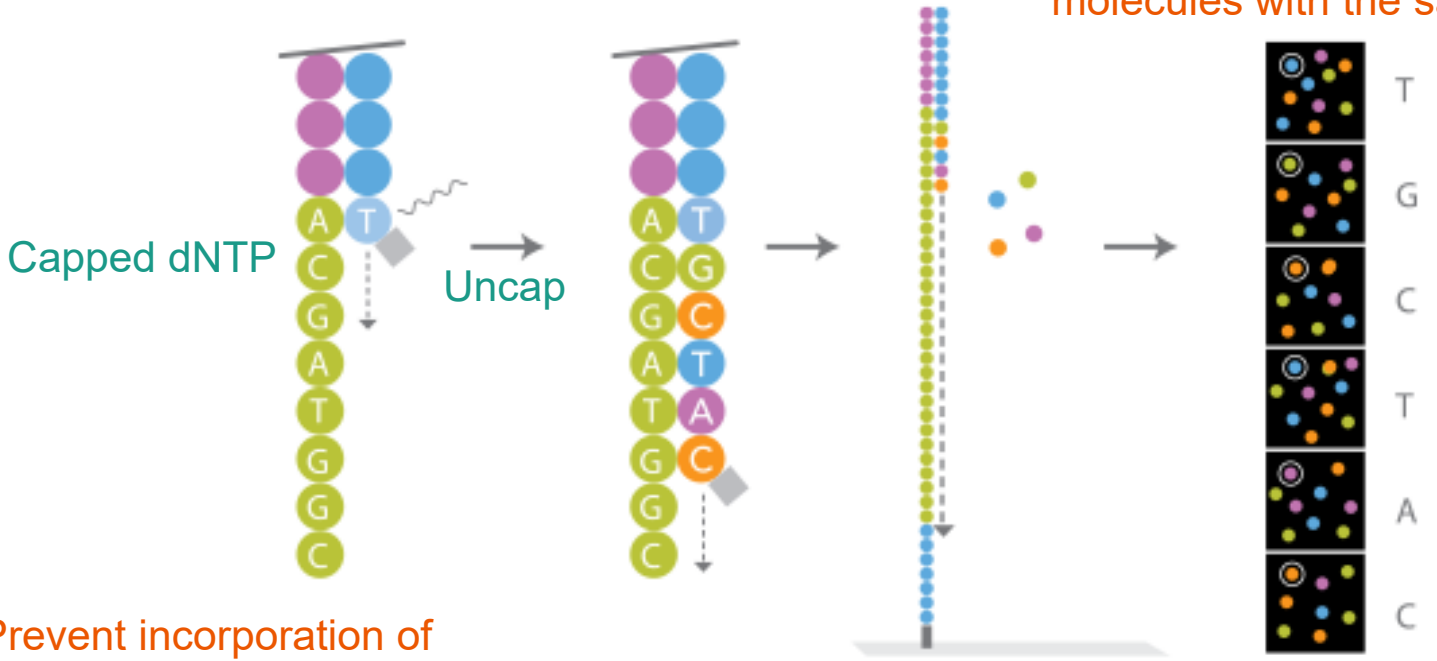
Illumina / Solexa DNA amplification



- Improve sensitivity by sequencing clusters of the amplified DNA molecules deriving from the same original DNA

Multi-step DNA polymerization

Each spot = cluster of multiple DNA molecules with the same sequence



Prevent incorporation of multiple dNTP

http://www.bioinformatics.nl/

Pros and cons

Platform	Read Length	Run Time	Gb/ Run	Advantage	Disadvantage
454 (Pyrosequencing)	400+	1 day	0.7	Long read length	Homopolymer error Single-ended only
Illumina	50-300	10 days	600	Low cost per base	Short reads Long run time
Ion Torrent	200-400	2 hrs	100	Fast run times	Homopolymer error

Tradeoffs

Sanger

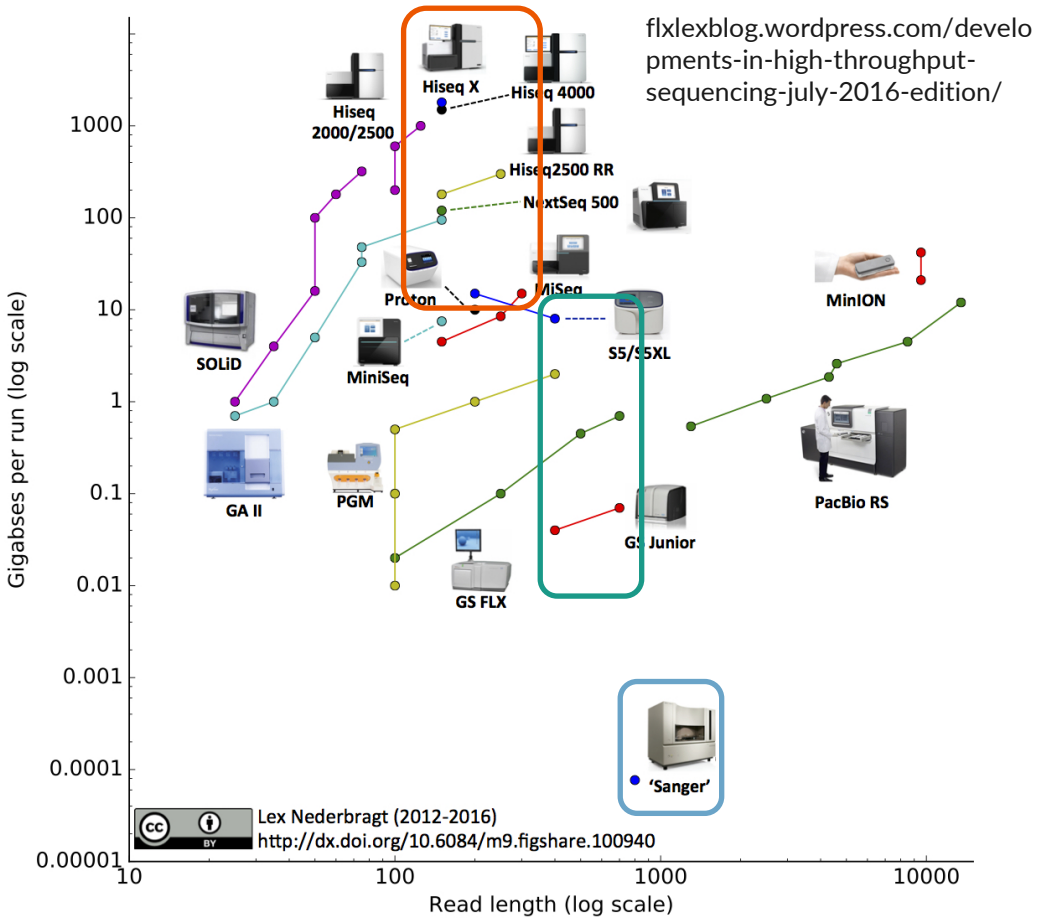
- 1000 bp, low throughput

454 and Ion Torrent

- 400+ bp, medium throughput

Illumina

- <300 bp, high throughput



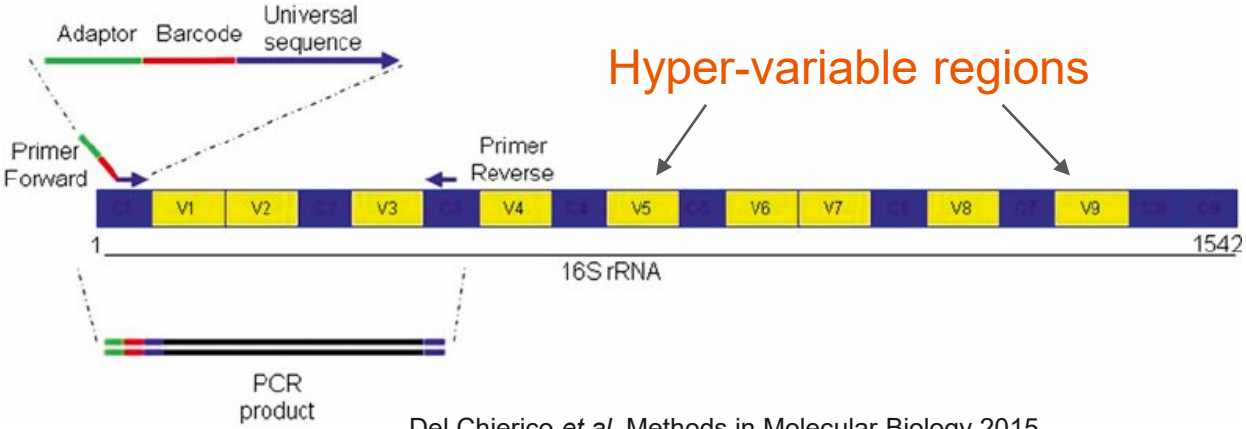
Use cases

Sanger
- Validate sequences

454 Pyrosequencing
- Metagenomics

Ion Torrent
- Fast turn-around situation

Illumina
- Whole-genome
- Practically anything…



Del Chierico *et al*. Methods in Molecular Biology 2015

3rd Generation Sequencing (Long-Read)

Single-Molecule Real-Time (SMRT) sequencing

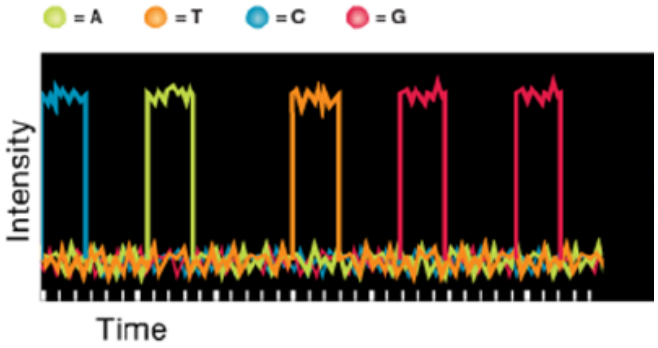
Zero-mode waveguide (ZMW)

Phospholinked nucleotide

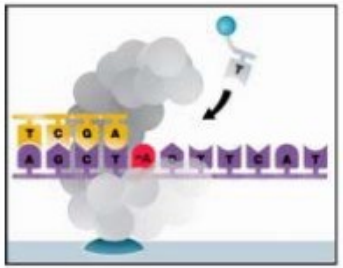
Images from Pacific Biosciences

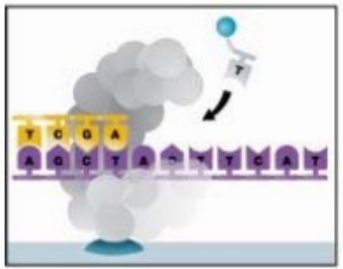
- Faster, more durable DNA polymerase
- Small wells with single DNA molecule
  - Zero-mode waveguide = nanophotonic confinement structure
  - Allow monitoring of fluorescence signal from individual reaction
- No amplification = direct quantification of DNA/RNA abundance

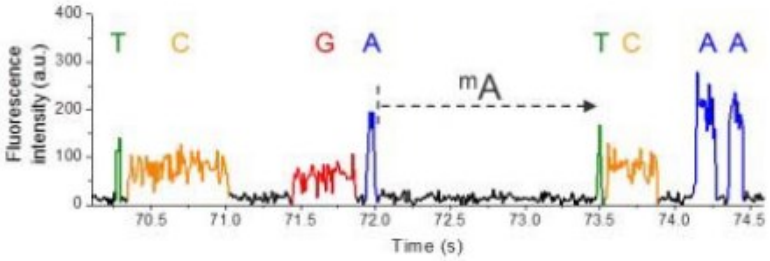
Video data

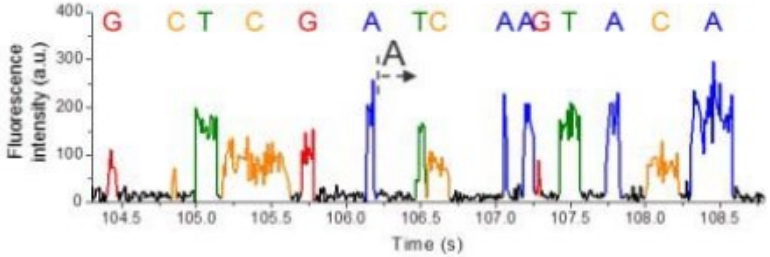


Images from Pacific Biosciences



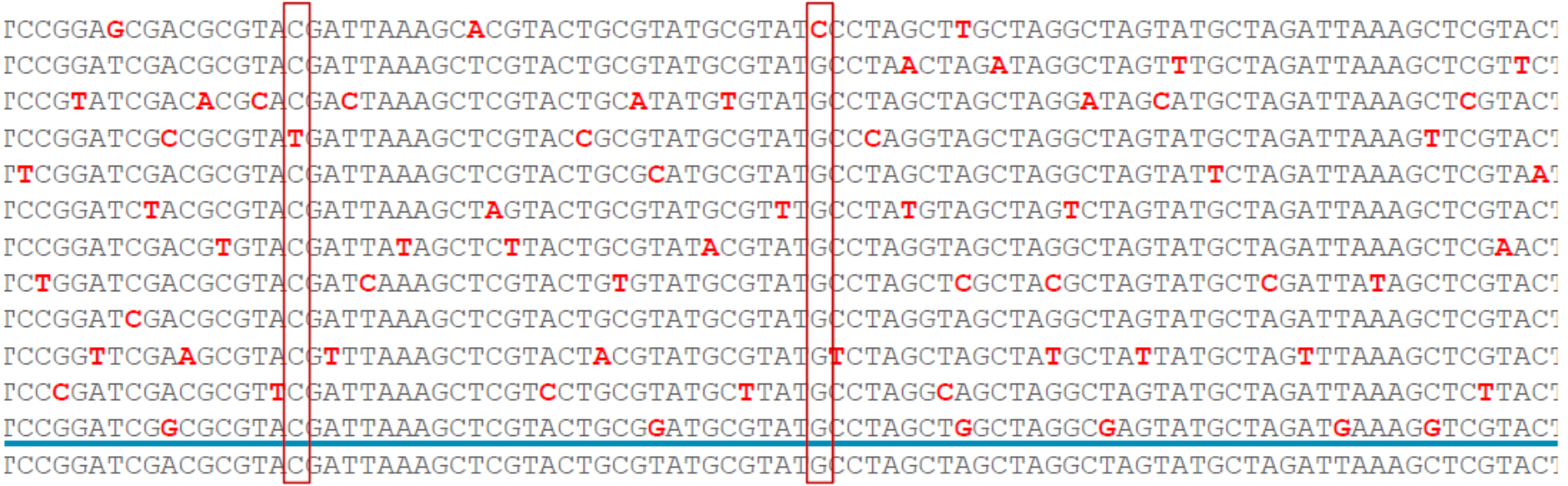






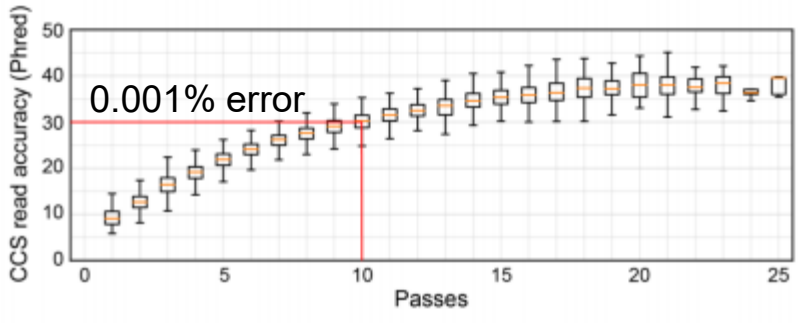
- Compared to image data from Illumina platform
- Video gives time information → identification of modified DNA/RNA

High error rate

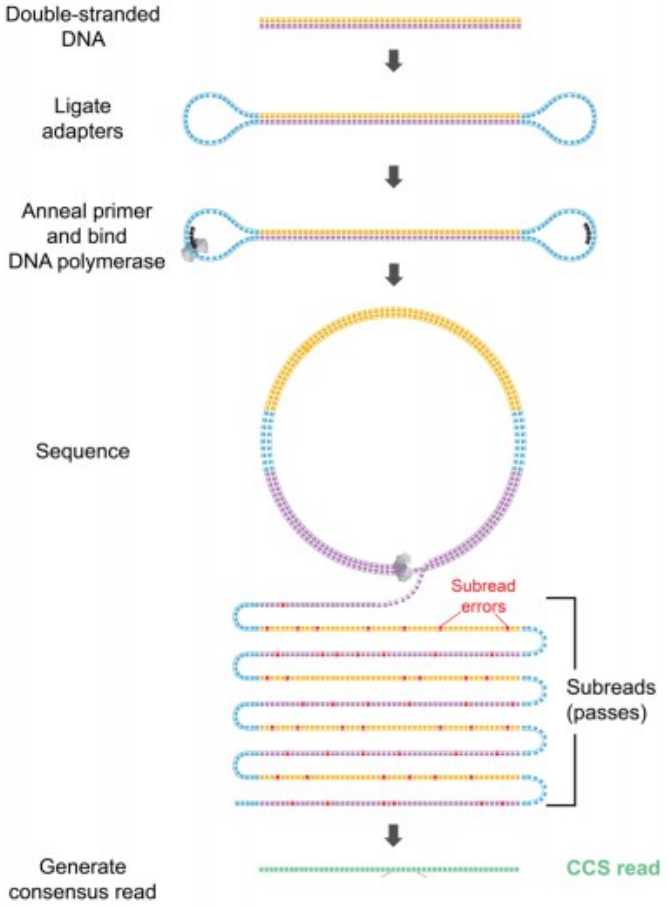


- 5-15% error compared to 0.01% of Illumina
- How do we solve this?

Circular consensus sequencing

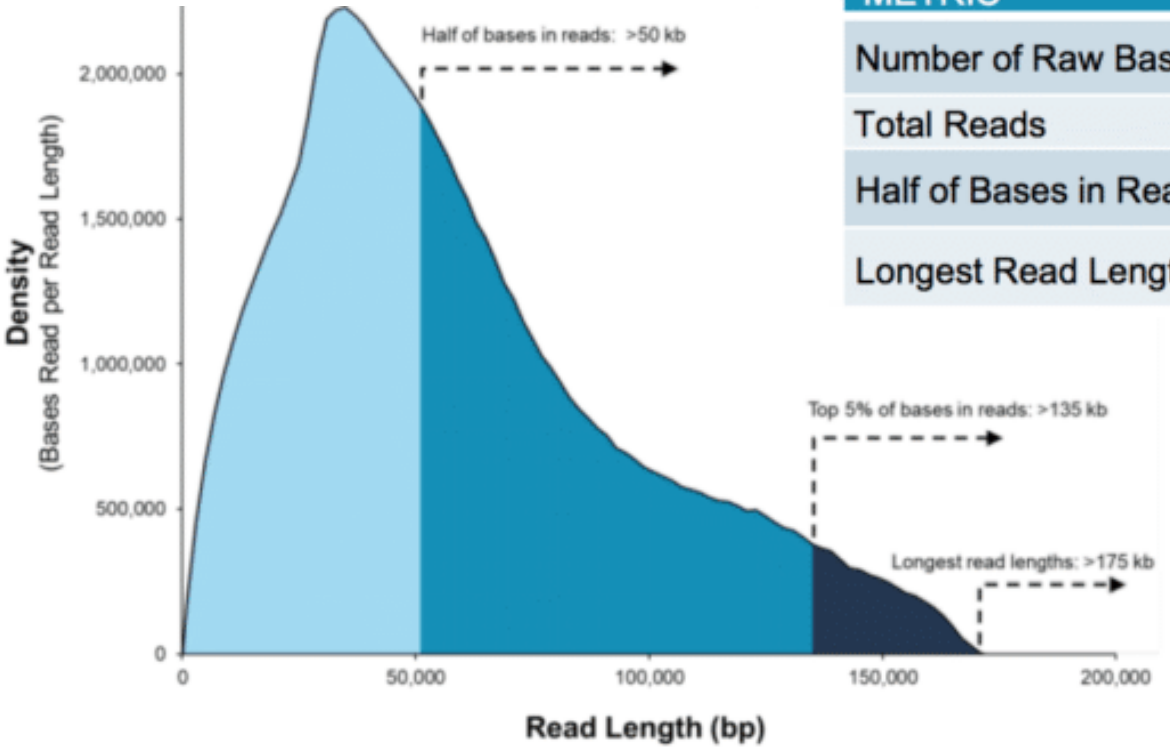


- Circular extension of each DNA molecule
- Read the extended molecules = multiple re-sequencing of the original sequence
- Take the consensus (majority vote)
- P(correct base in >*k* of *N* passes) ~ Binomial



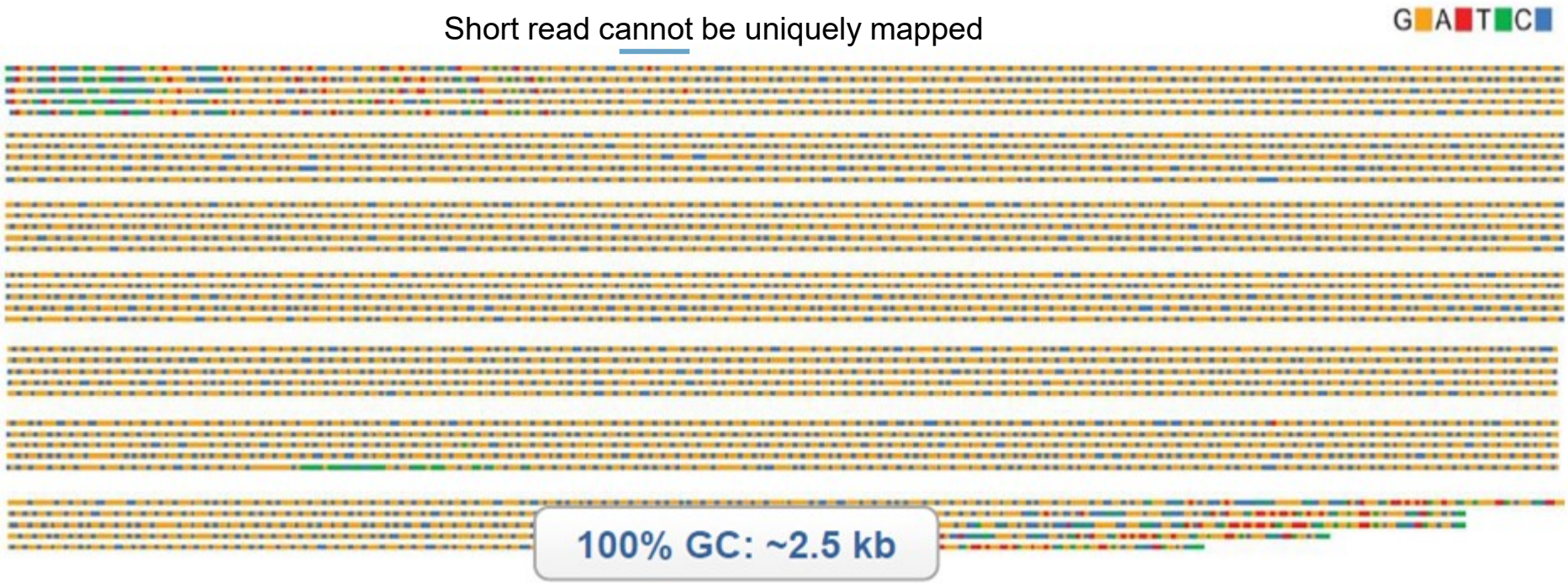
Images from Pacific Biosciences

Read length >> 10kb



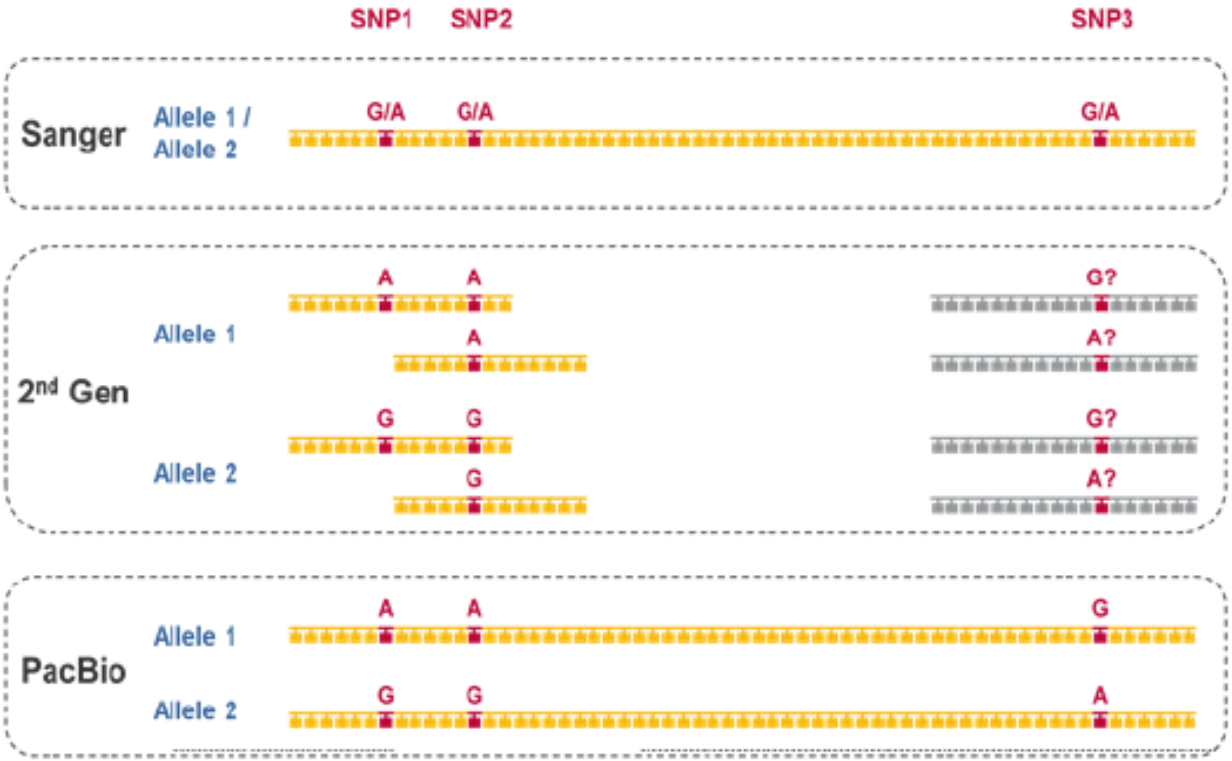
METRIC	
Number of Raw Bases	166 Gb
Total Reads	5,201,973
Half of Bases in Reads	>51,863
Longest Read Lengths	>175,000

Resolve repetitive region



- ~20% of human genome

Resolve haplotype

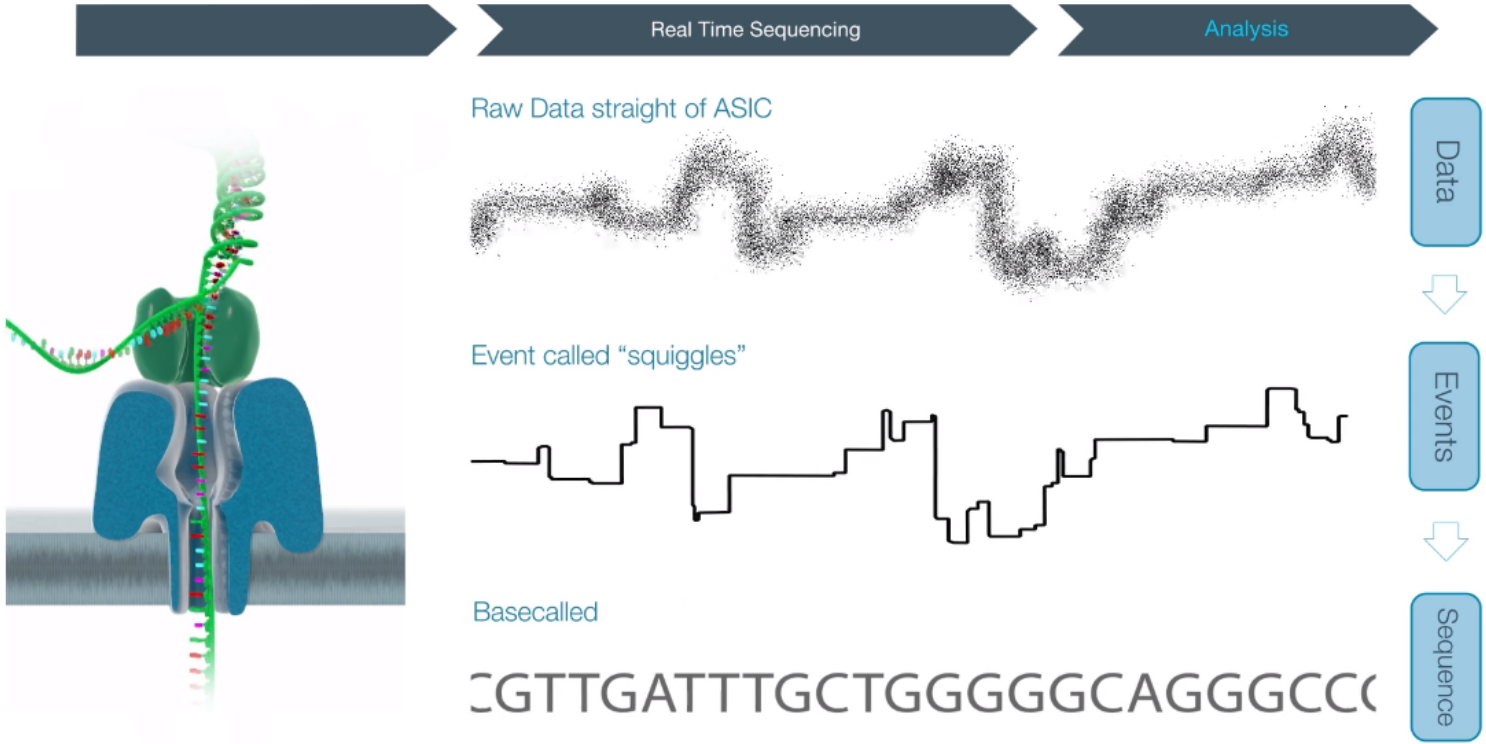


Sanger reads come from mixture of many molecules

2nd generation reads are too short to span the whole haplotype block

3rd generation reads are both single-molecule and long

Nanopore

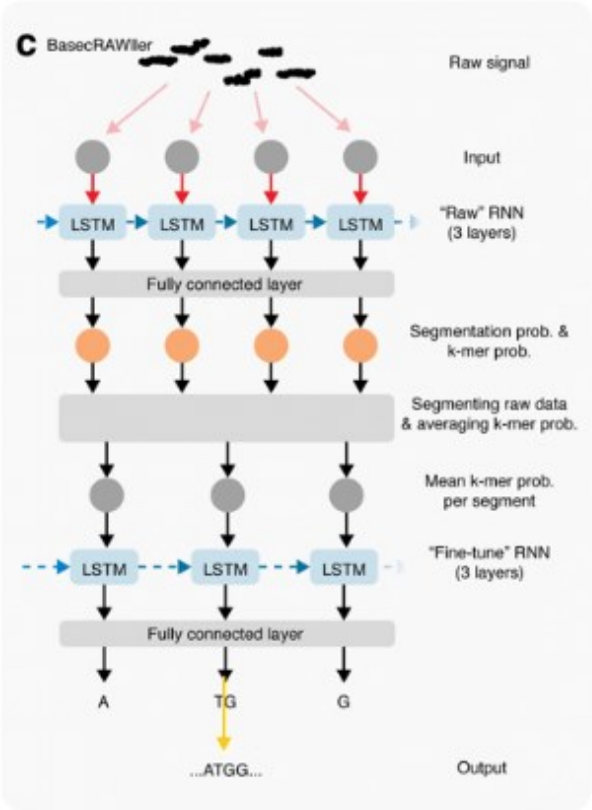


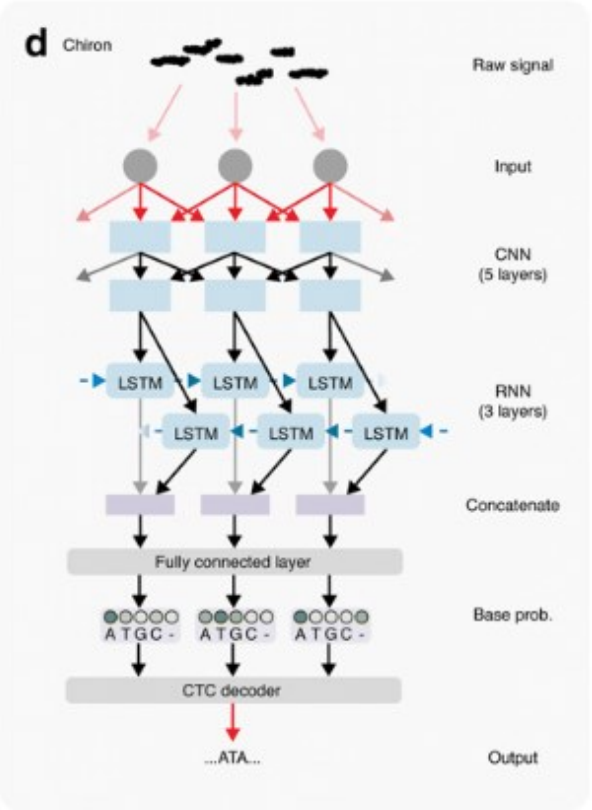
https://nanoporetech.com/applications/dna-nanopore-sequencing

Basecalling with deep neural networks

- Trained using data from synthetic DNA

- 14% baseline error
- Improved to 3-5% using bioinformatics and machine learning



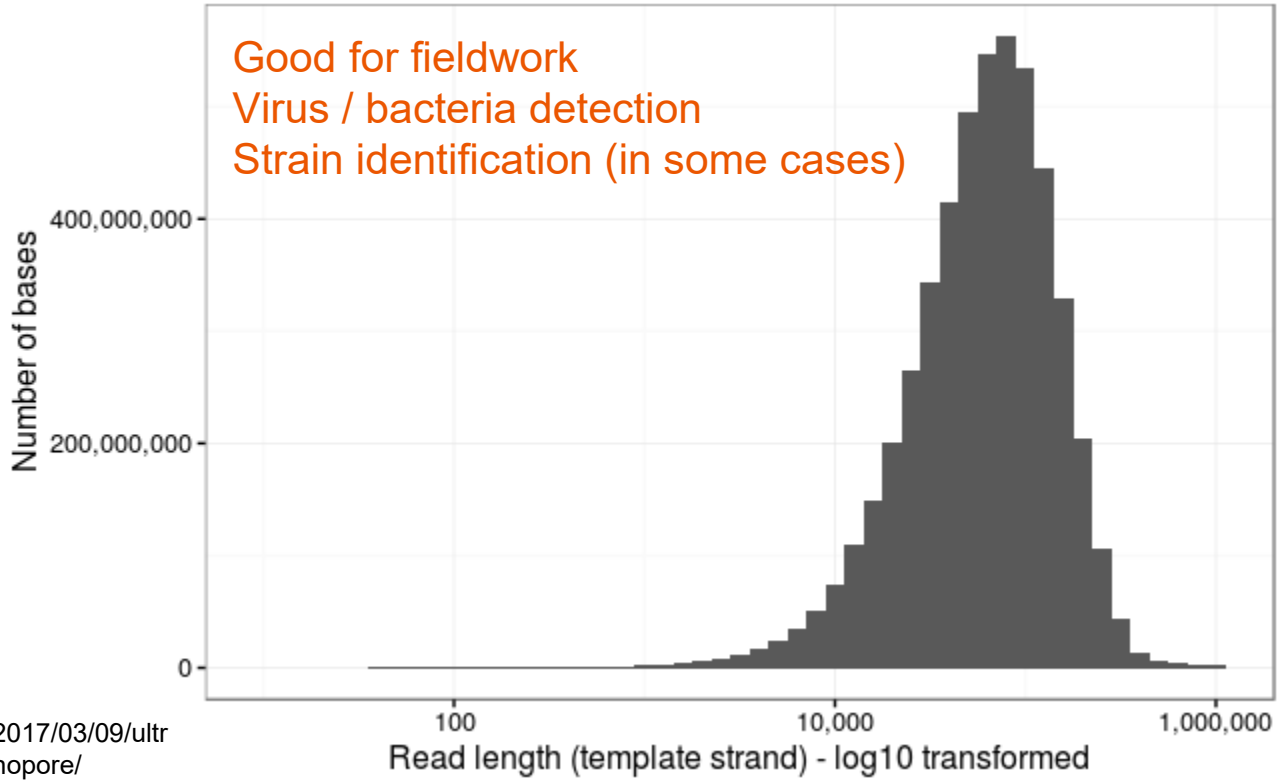


Rang, Kloosterman, and de Ridder.
Genome Biology 2018

Portability & fast turn-around time

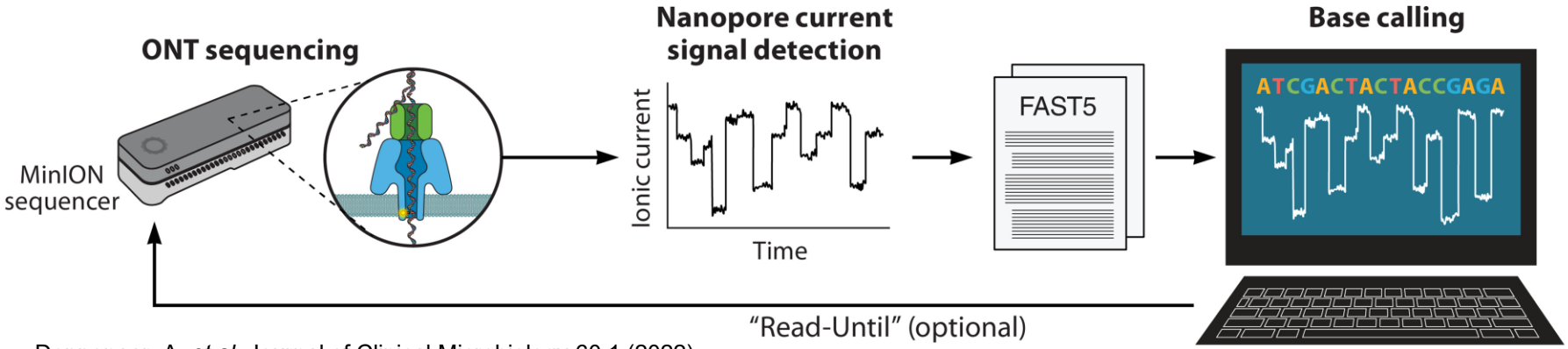
	Flongle	MinION	GridION (5 flow cells)	PromethION (48 flow cells)
				
Maximum run time	16 hours	72 hours	72 hours	64 hours
Theoretical 1D maximum yield	Up to 3.3 Gb	Up to 40 Gb	Up to 200 Gb	Up to 15 Tb
Current 1D maximum yield	Up to 2 Gb	Up to 30 Gb	Up to 150 Gb	Up to 8.6 Tb
Available channels	Up to 126	Up to 512	Up to 2,560	Up to 144,000

Read length up to Mb



https://lab.loman.net/2017/03/09/ultrareads-for-nanopore/

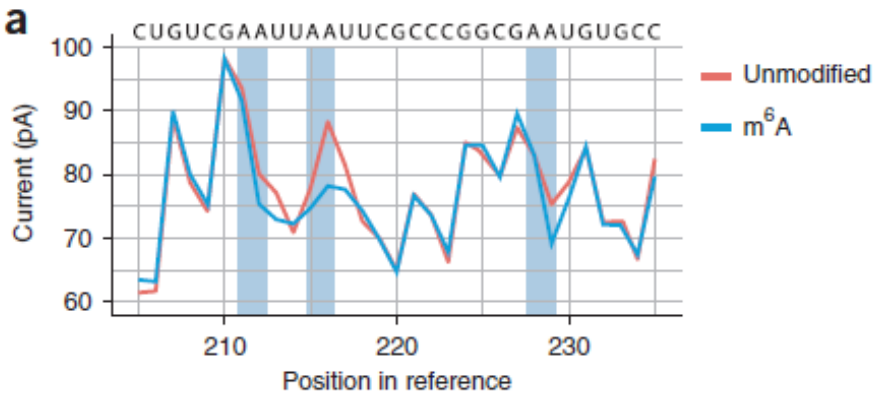
Real-time data

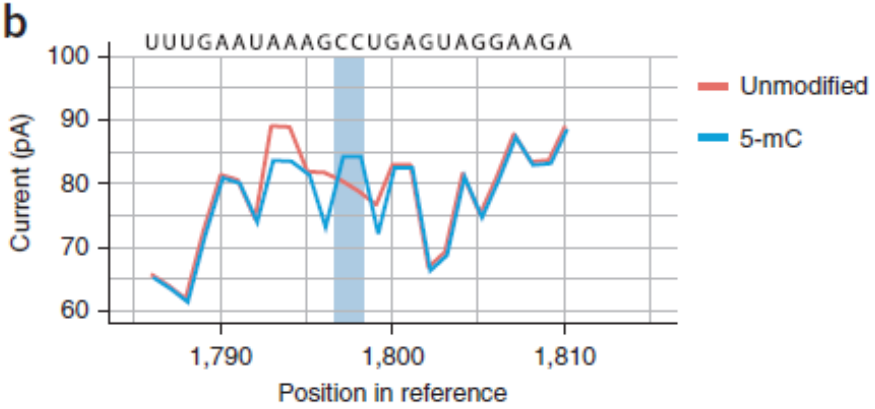


Deppenaar, A. *et al*. Journal of Clinical Microbiology 60:1 (2022)

- Real time ionic flow signals
- Ability to manipulate individual pore and terminate unwanted reads
- Rapid decision making (no need to wait for the full 16-72hr run)

Detection of modified nucleotides

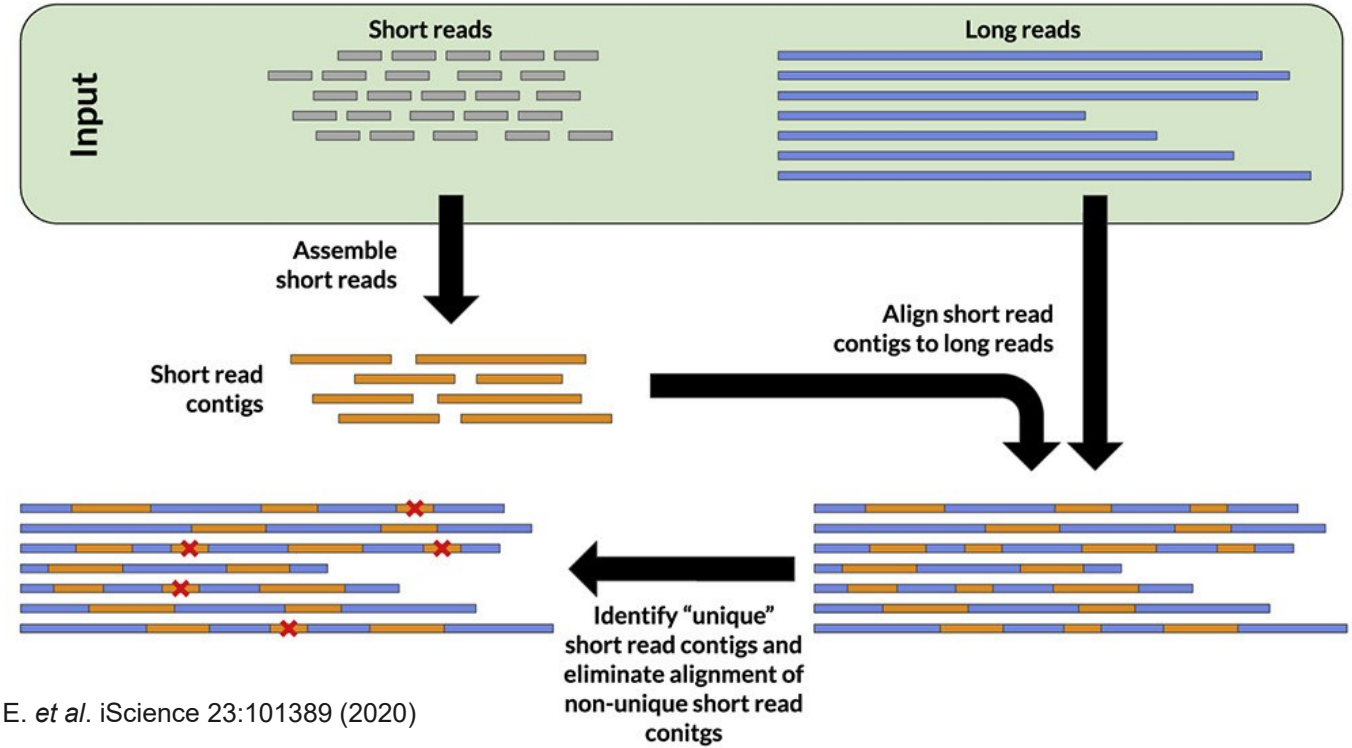




Geralde *et al*. Nature Methods 15, 201-206 (2017)

- Modified nucleotides = different 3D structure = different change in ionic flow
- Trained using synthetic nucleotides

Combining short and long read data



Haghshenas, E. *et al*. iScience 23:101389 (2020)

Applications of DNA/RNA sequencing

Sequencing scope

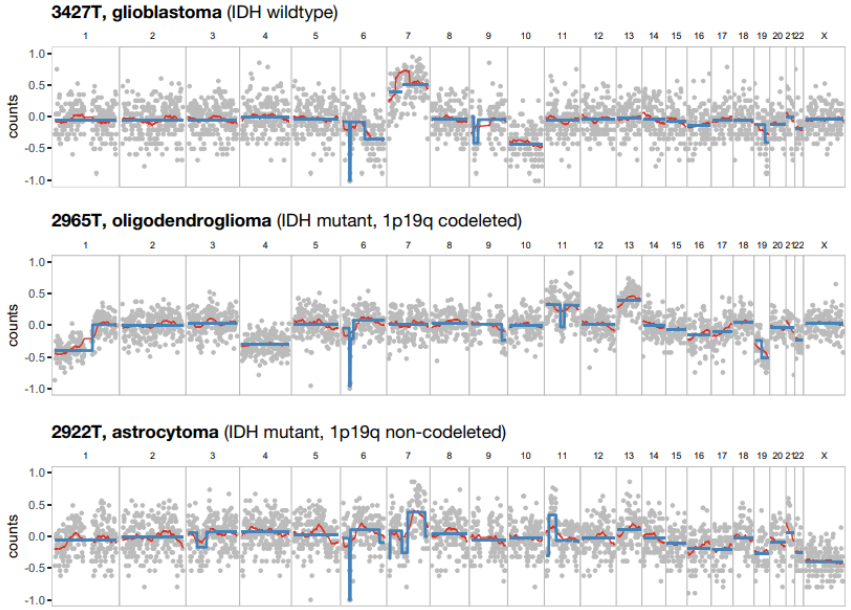
- Cost = Base Pair = Scope x Depth

Reduced scope

- Exome sequencing = exons only
- Amplicon sequencing = selected loci
  - 16S rRNA, RDRP gene
  - (Cancer) gene panels

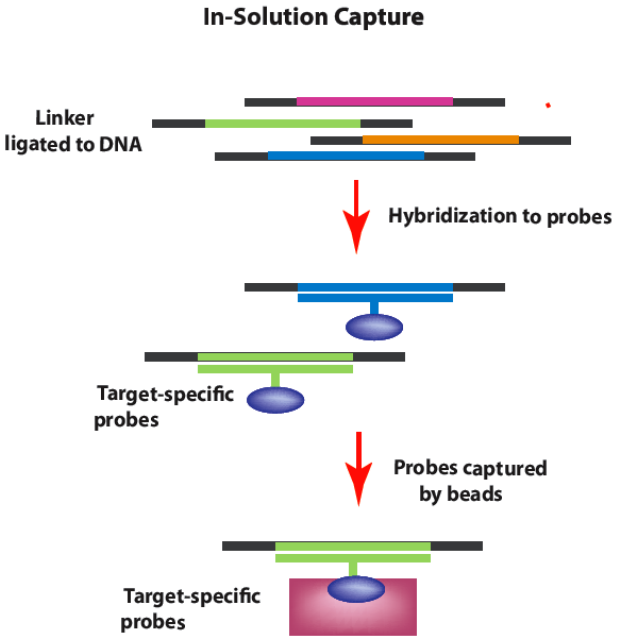
Reduced depth

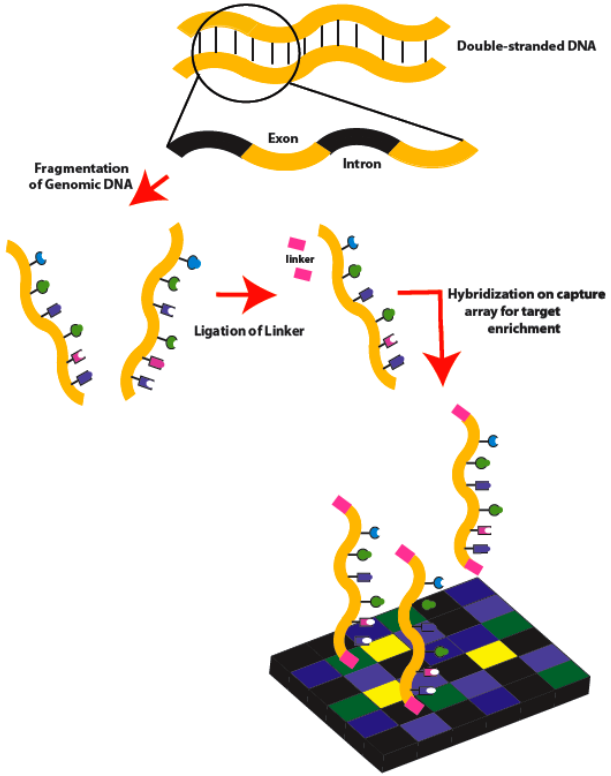
- Ultra-low pass
  - Detect chromosomal copy alternation
  - Estimate tumor fraction



Euskirchen, P. *et al*. Acta Neuropathol 134:691-703 (2017)

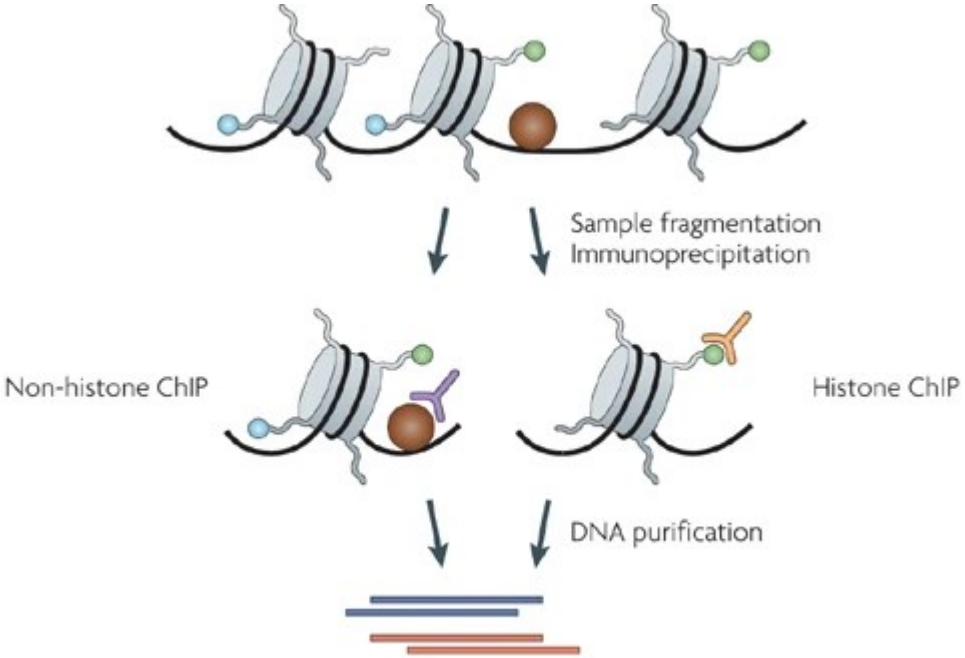
Enrichment for targeted sequencing

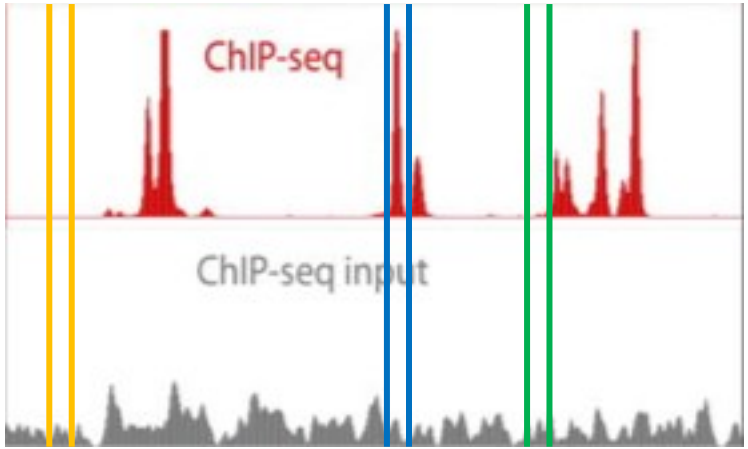




https://en.wikipedia.org/wiki/Exome_sequencing

Chromatin immunoprecipitation

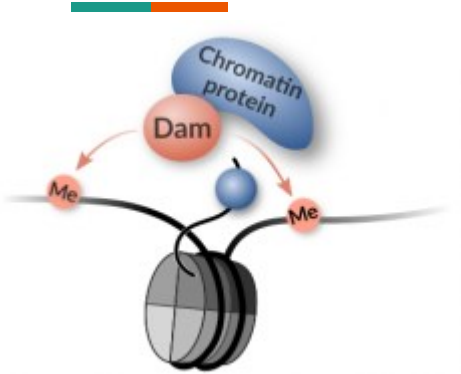




Park et al. Nat Rev Genet 10:669-680 (2009)

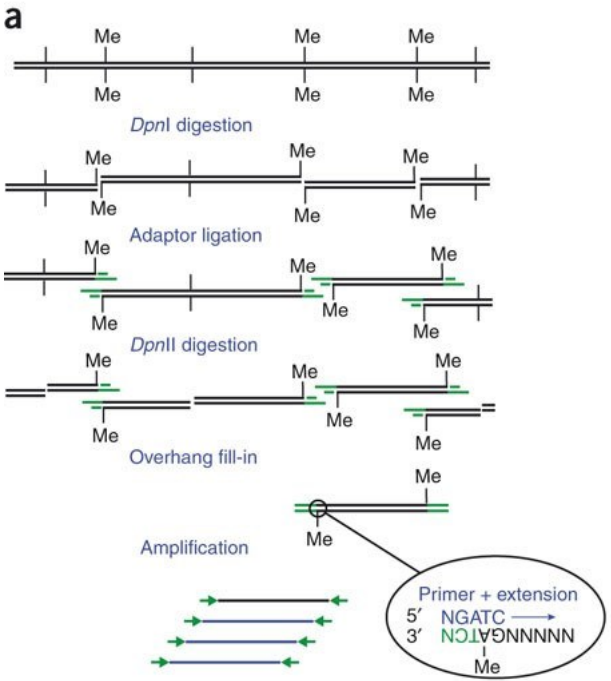
- DNA-bound protein / histone modification

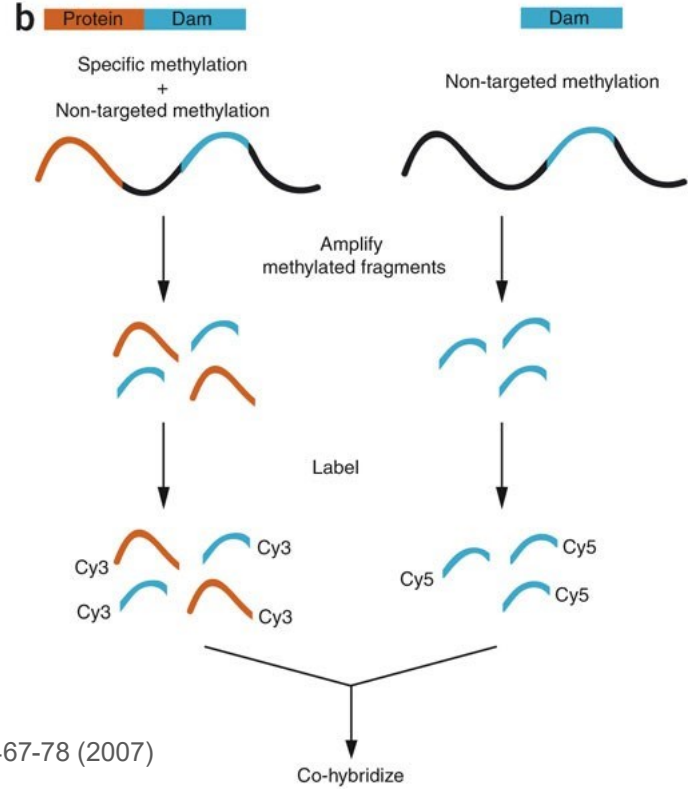
DNA adenine methylatransferase (DamID)



https://marshall-lab.org/damid/

- Dam attached to protein of interest
- Methylation of GATC
- DpnI/DpnII enzymes





Vogel, M.J. et al. Nat Protocols 2:1467-78 (2007)

Bisulfite sequencing

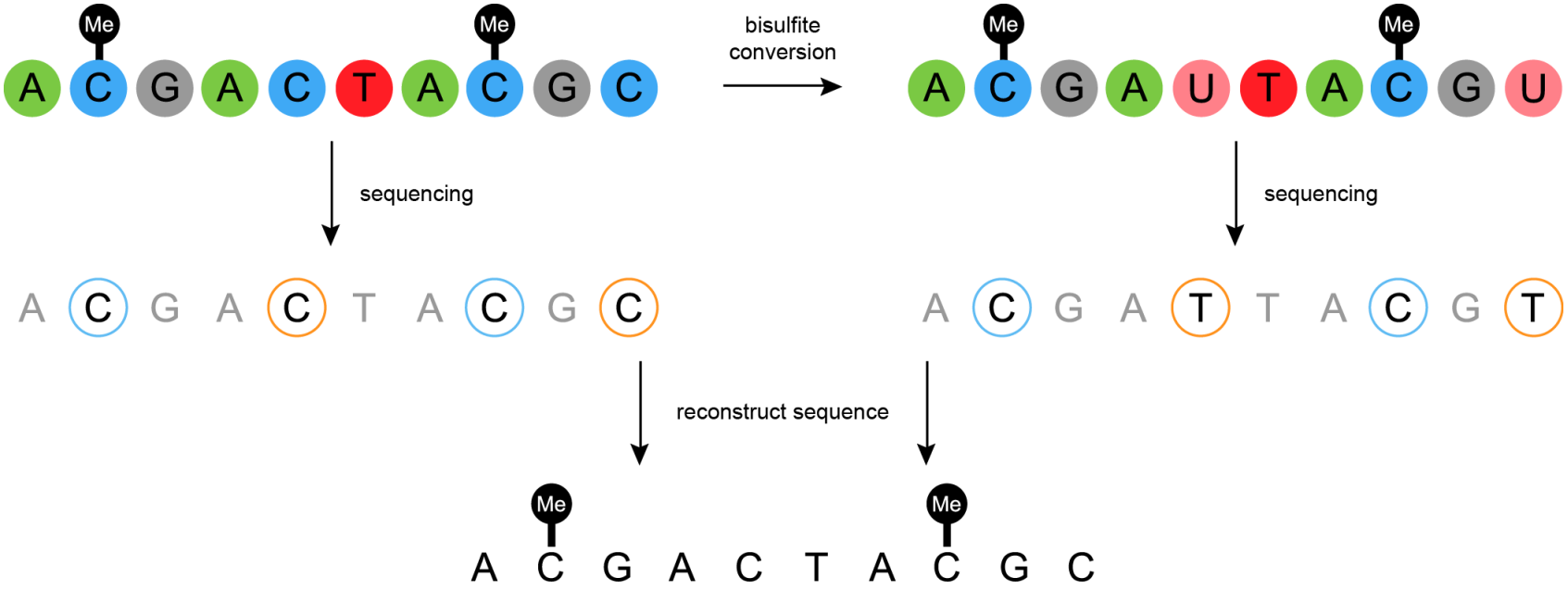
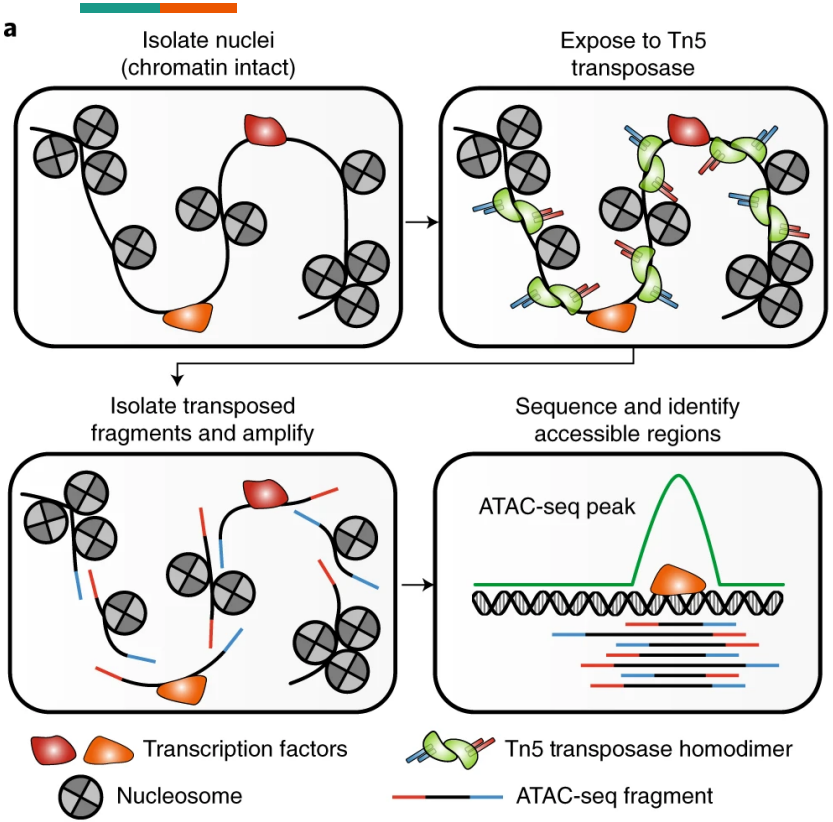
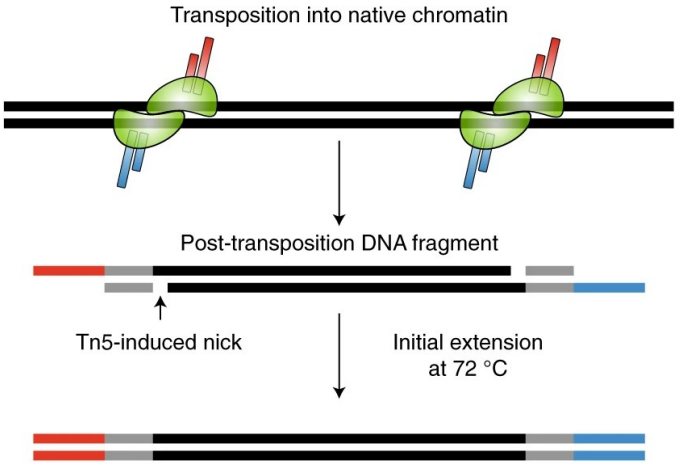


Image from http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis

Assay for transposase-accessible chromatin (ATAC)

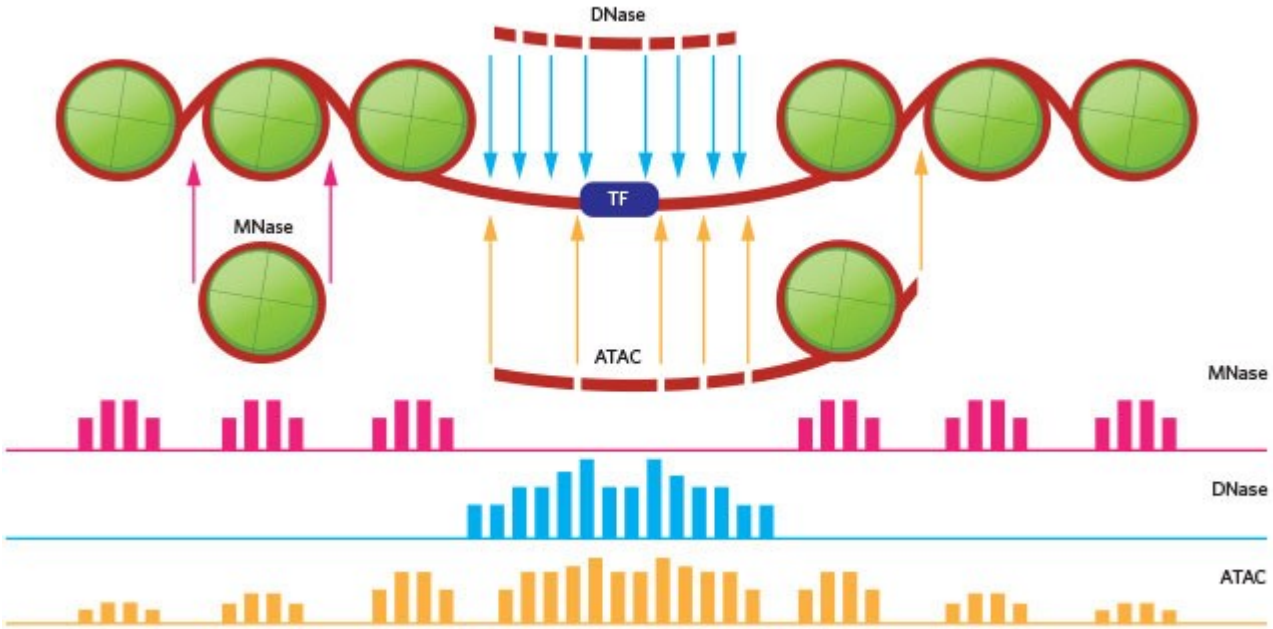




- Transposase with sequencing adapter insertion into open chromatin

Grandi, F.C. et al. Nat Protocols 17:1518-52 (2022)

Targetting bound or unbound chromatin



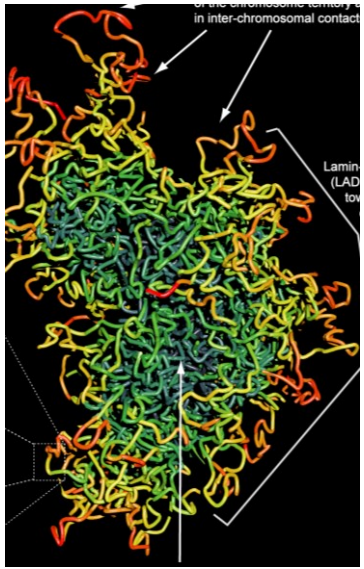
Bounded DNA

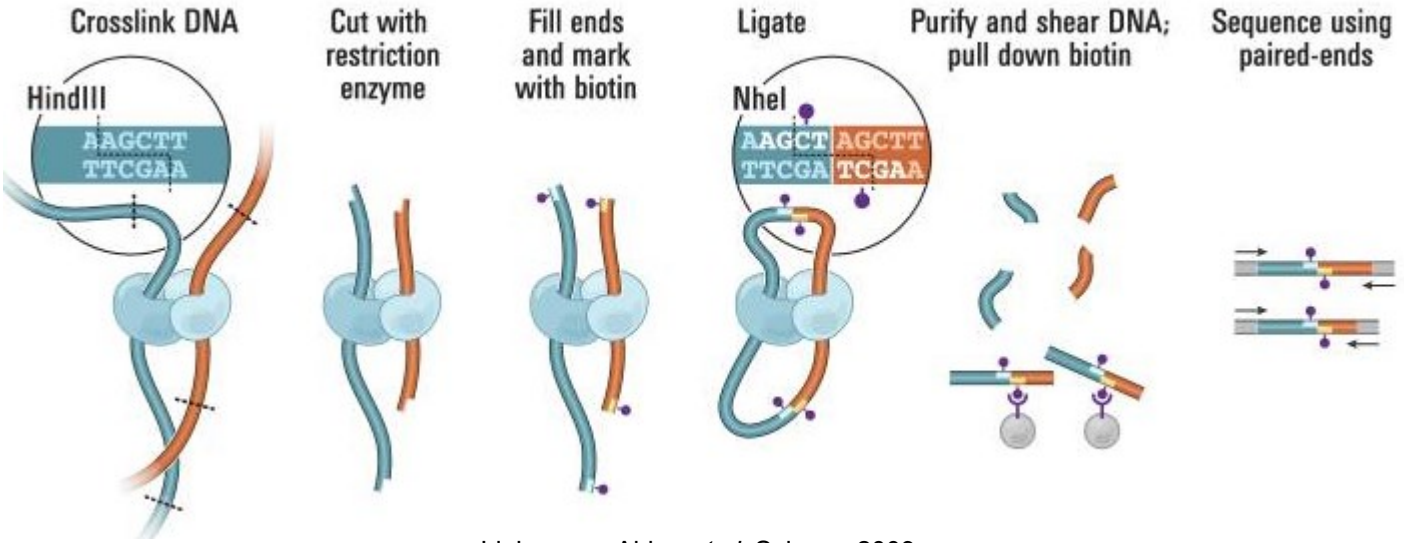
Unbounded DNA

Unbounded DNA + 1-2 nucleosomes

https://www.the-scientist.com/lab-tools/reveling-in-the-revealed-34261

Chromatin conformation capture



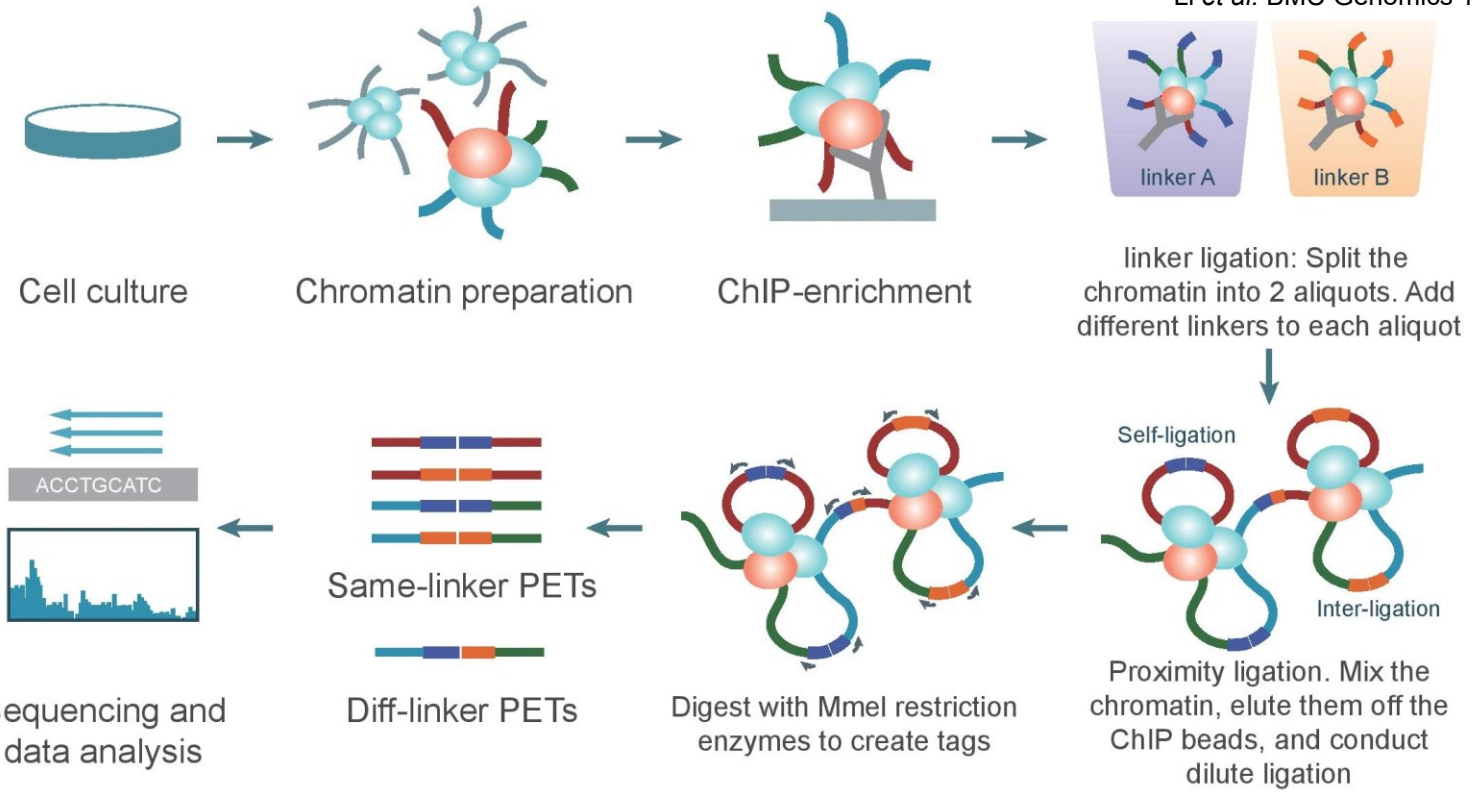


Lieberman-Aiden *et al*. Science 2009

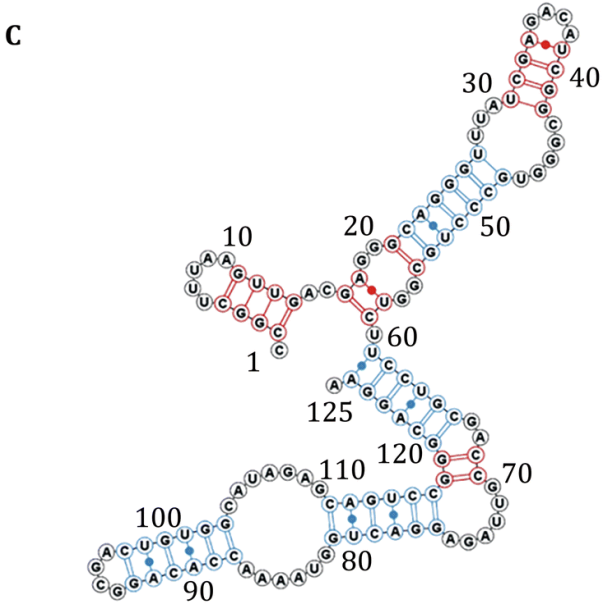
- Cross-link proximal DNA → join ends from different regions → sequencing

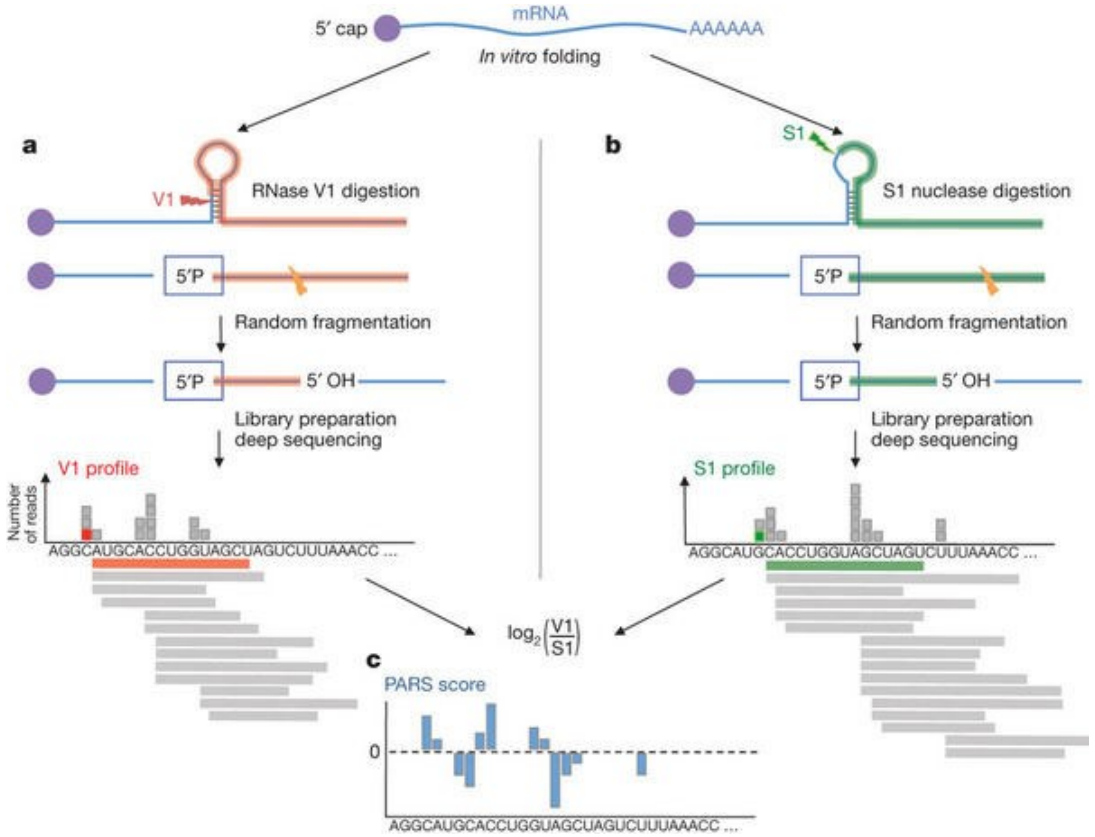
Chromatin interaction analysis with paired-end tag

Li *et al.* BMC Genomics 15:S11 (2014)



RNA secondary structure





Kertesz *et al*. Nature 467: 103-107 (2010)

Sequencing application key points

- Scope of sequencing = enrich target DNA
- DNA-binding protein = antibody pull-down
- Detection of DNA modification
- Targeting bound / unbound DNA
- Enzyme specificity

Any question?

- See you on August 31st
- **Reminder**: The next session will be via Zoom!