3000788 Intro to Comp Molec Biol

Lecture 3: DNA sequencing techniques and applications

Fall 2025

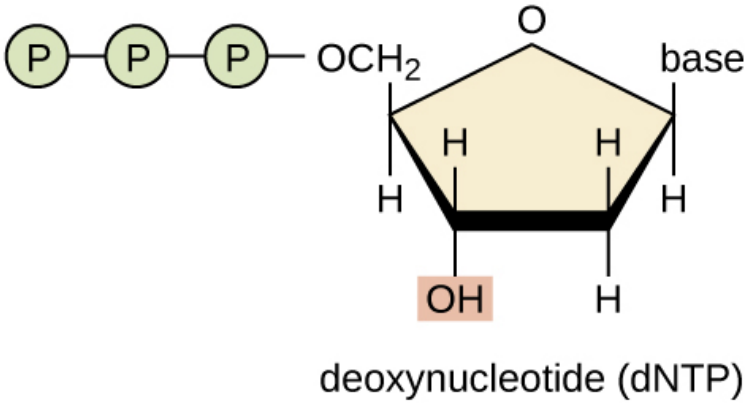
Sira Sriswasdi, PhD

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

Today’s agenda

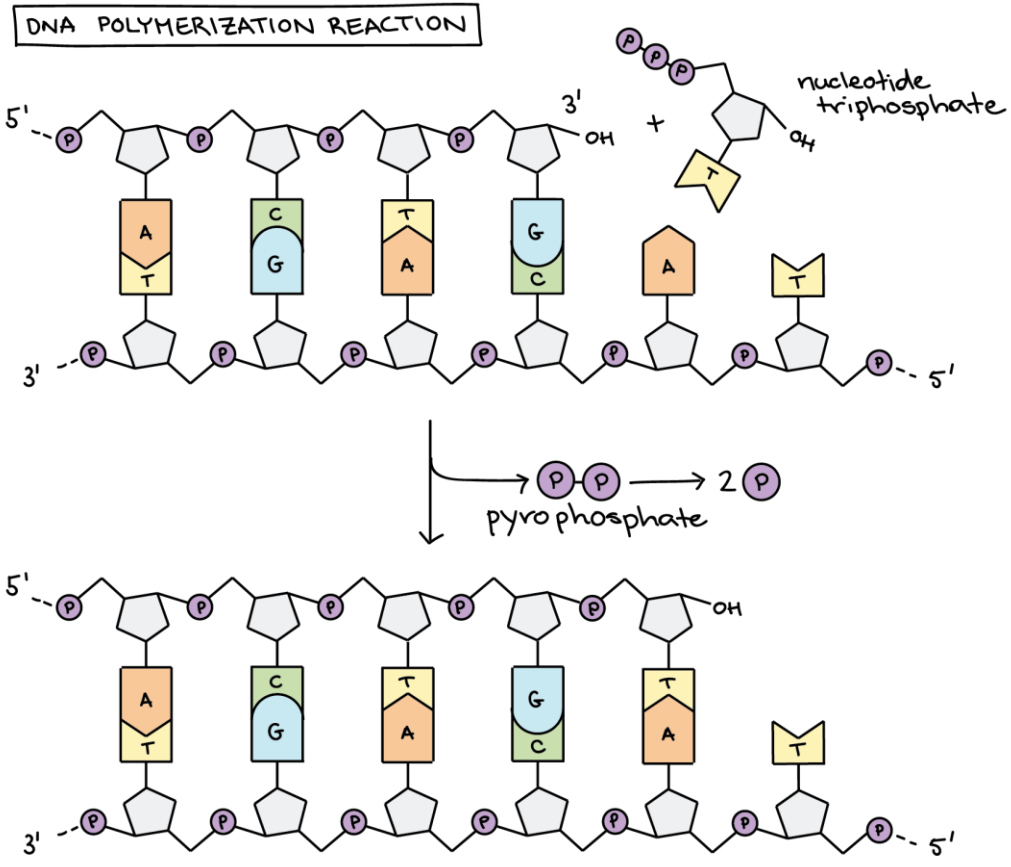
- DNA sequencing platforms
- Applications of DNA sequencing as molecular assays

DNA polymerization



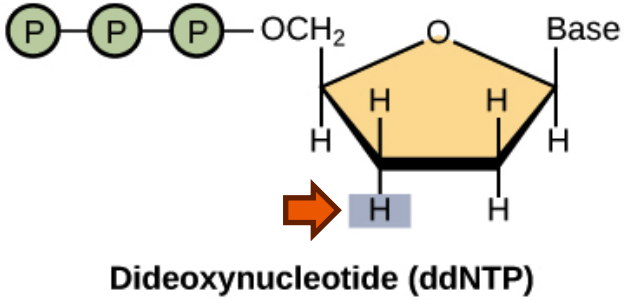
deoxynucleotide (dNTP)

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

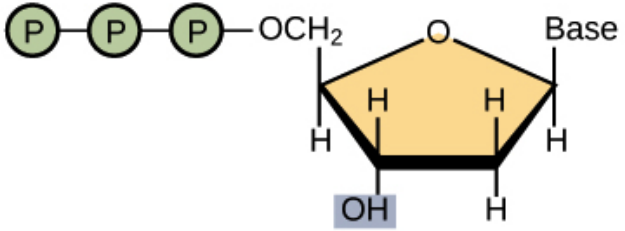


https://www.khanacademy.org

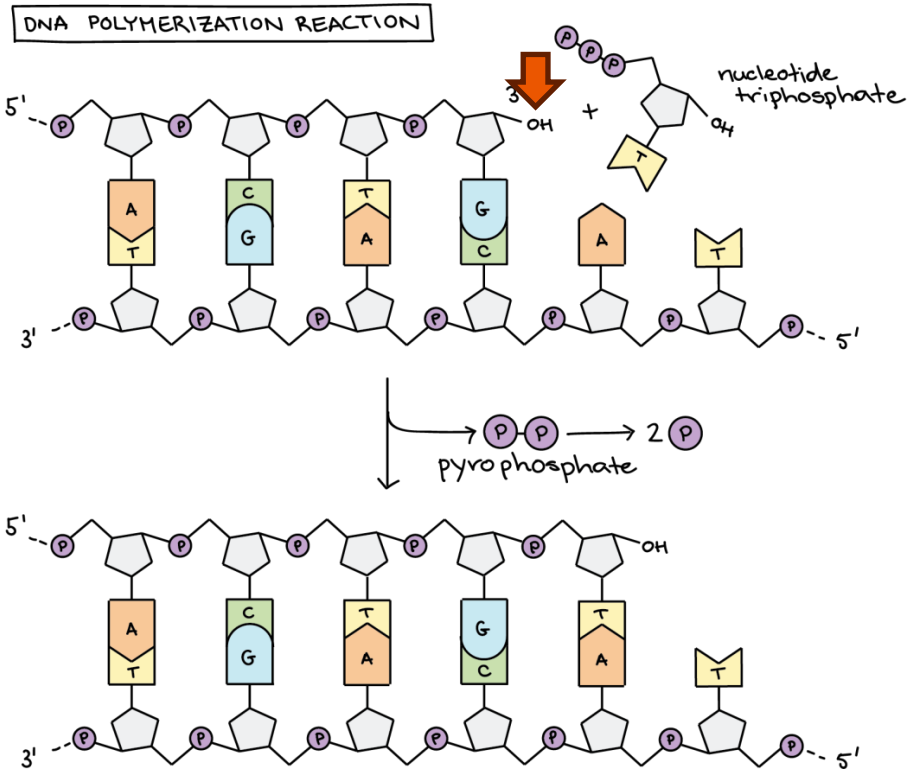
ddNTP terminate polymerization



Dideoxynucleotide (ddNTP)

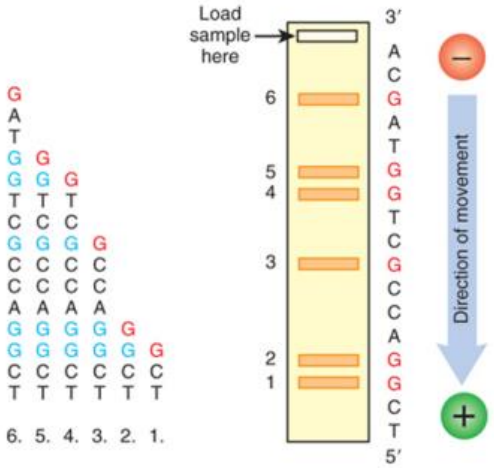


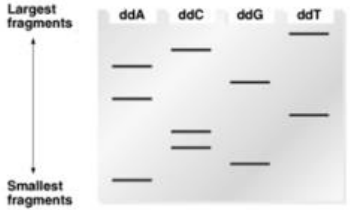
Deoxynucleotide (dNTP)



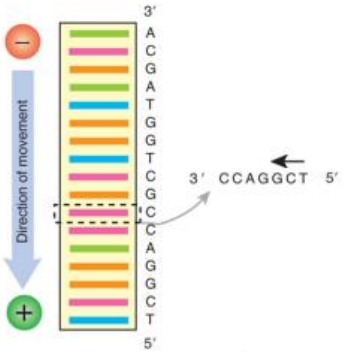
https://www.khanacademy.org

Sanger sequencing

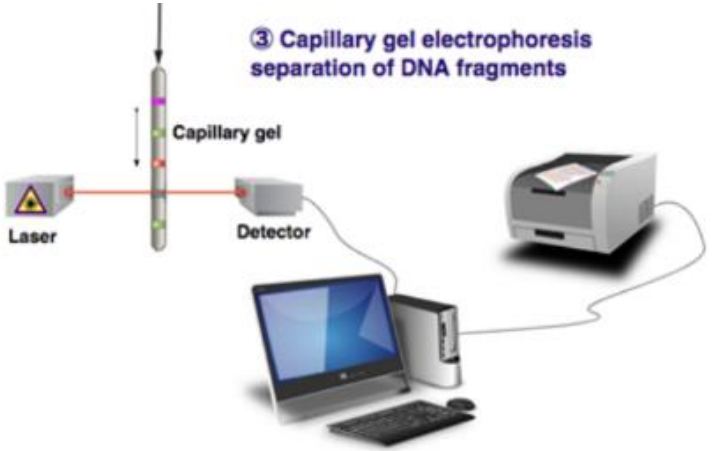


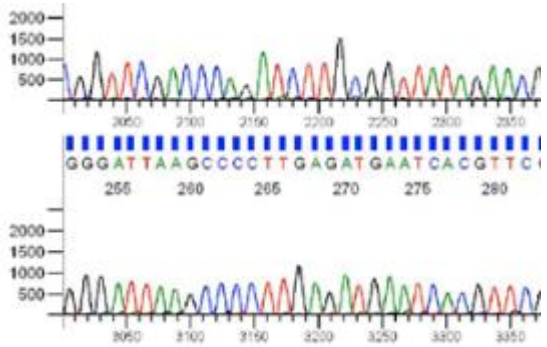


What is the sequence?



Fluorescence-labeled ddNTP

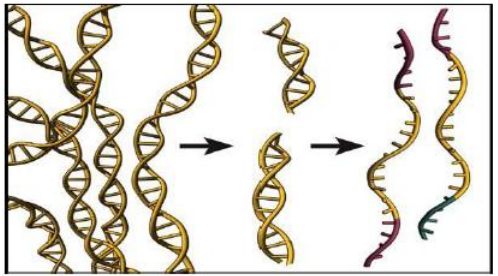




Images from https://en.wikipedia.org/wiki/Sanger_sequencing

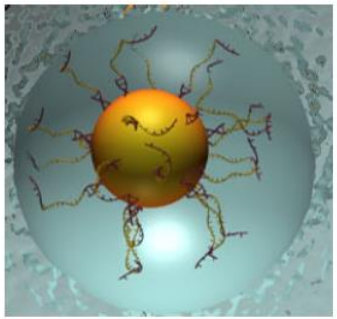
Next-Generation Sequencing (NGS)

High throughput with parallel reactions

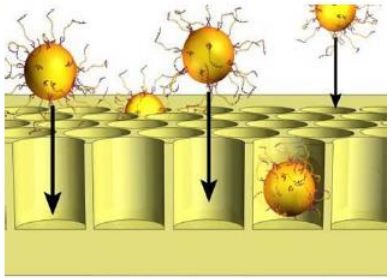


1) Adapter-ligated ssDNA library

Roche & Ion Torrent wells

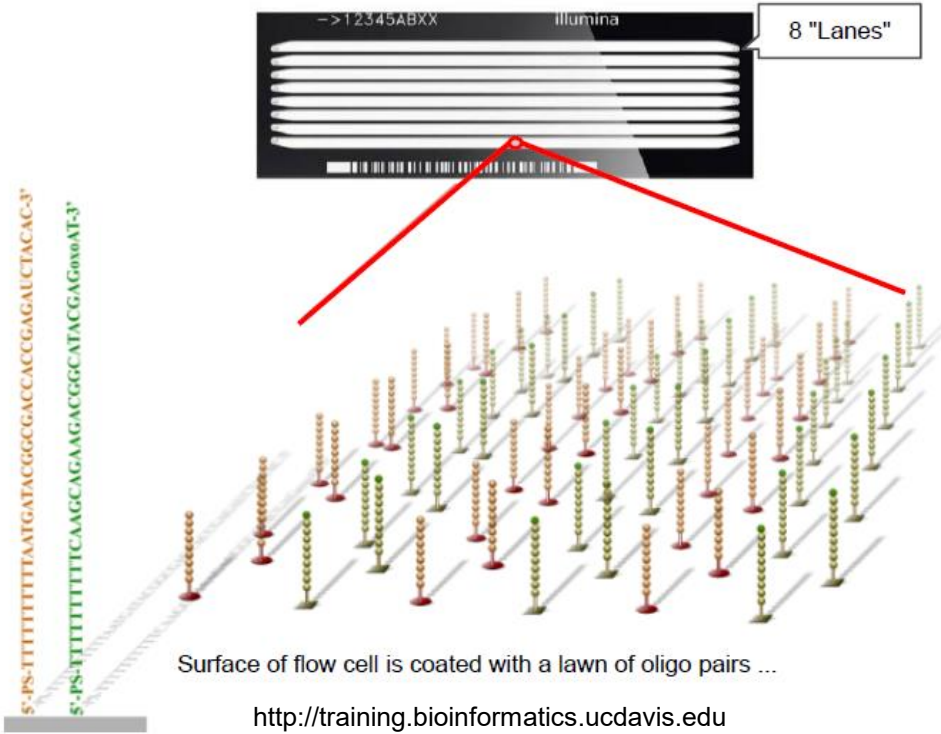


2) Clonal amplification on 28 micron beads ... emulsion PCR

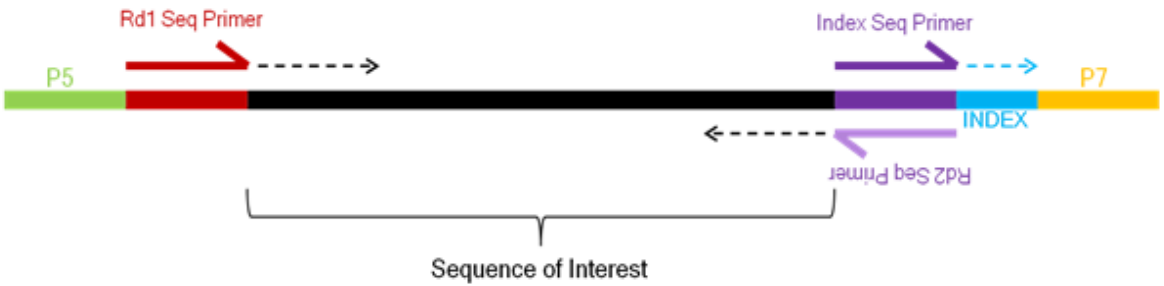


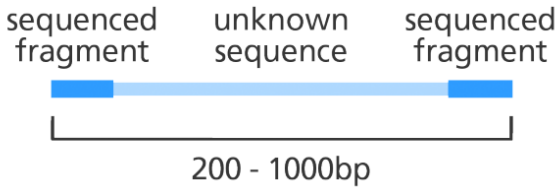
3) Beads deposited on PicoTiterPlate wells

Illumina's flow cell

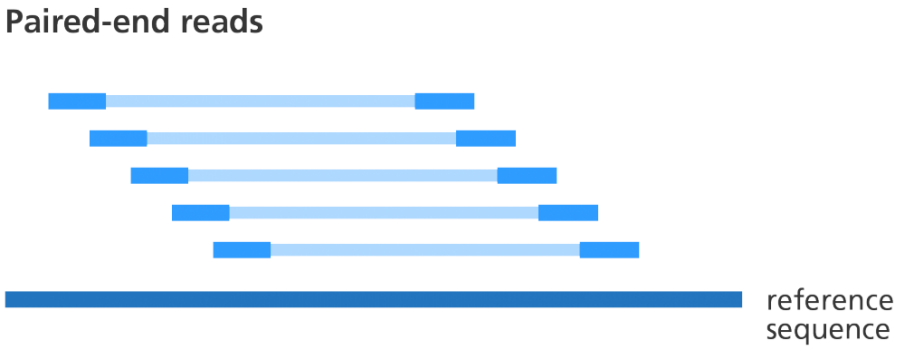


Paired-end sequencing



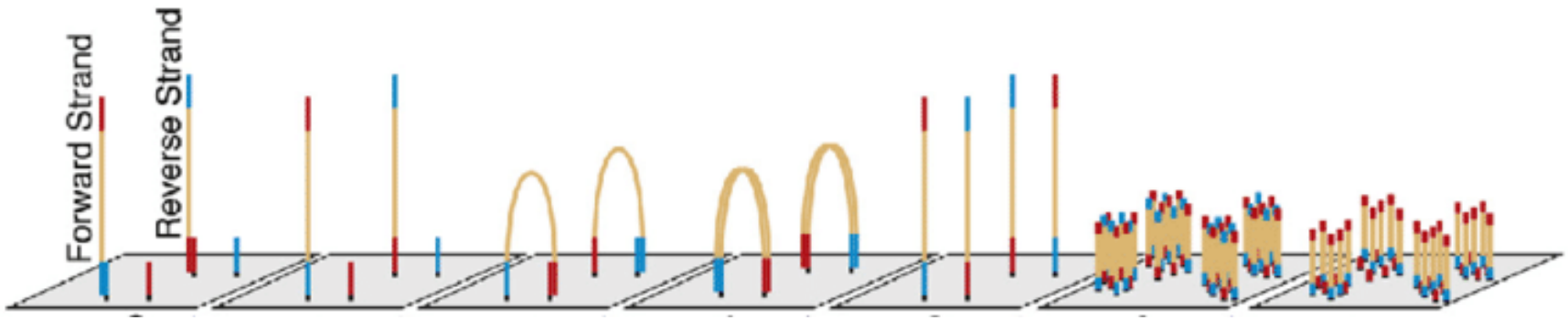


- Add adapters to both ends of a DNA fragment
  - Improve mappability
  - Identify splice junction
  - Identify genomic translocation
  - Useful in chromatin conformation capture



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

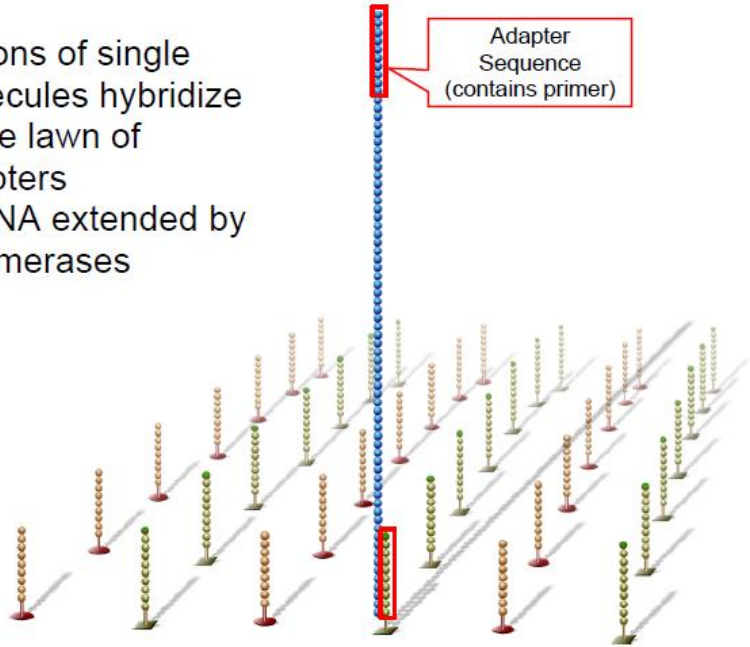
DNA amplication for sequencing



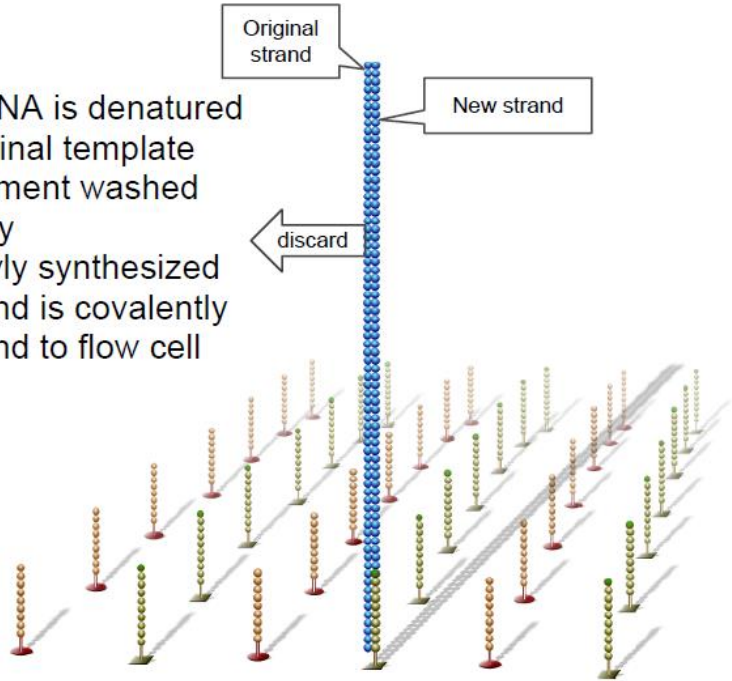
- Amplified DNA molecules into clusters of identical sequences
  - Improve signal-to-noise / sensitivity
  - Can introduce amplification bias

Amplification step 1

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

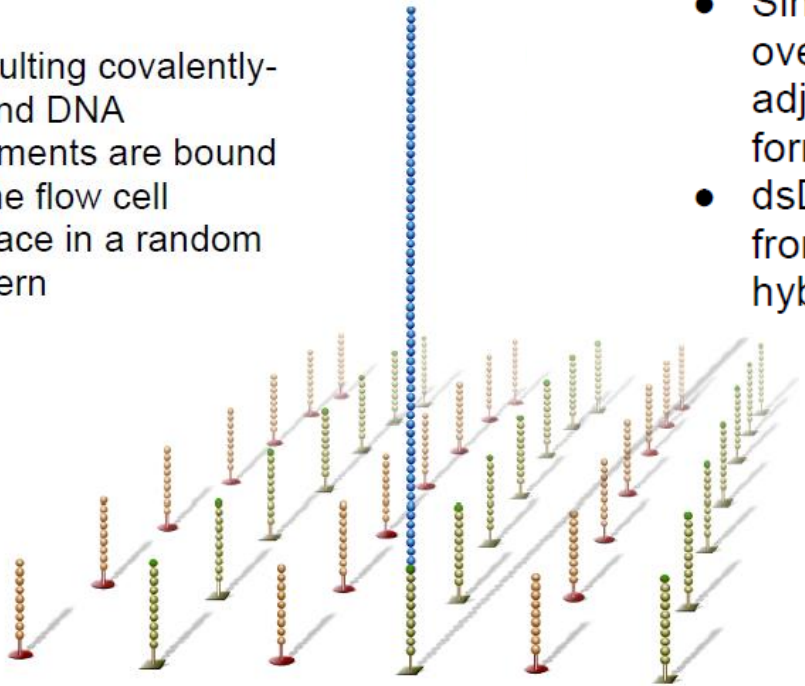


- dsDNA is denatured
- Original template fragment washed away
- Newly synthesized strand is covalently bound to flow cell

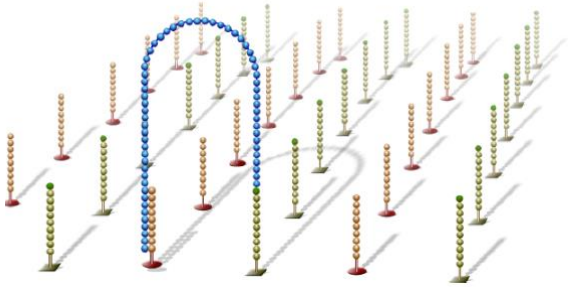


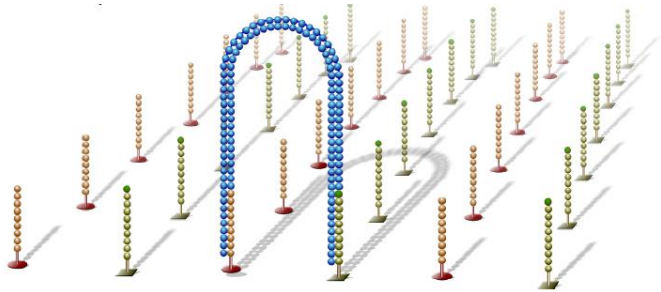
Amplification step 2

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

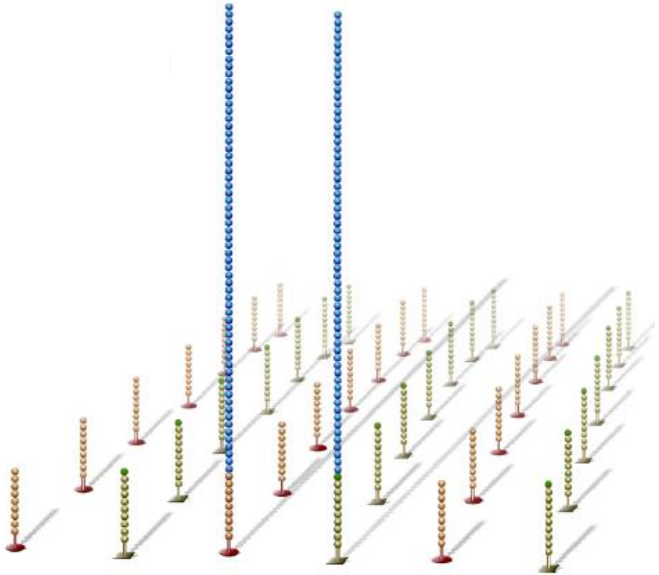


- Single-strand flops over to hybridize to adjacent adapter, forming a bridge
- dsDNA synthesized from primer in hybridized adapter

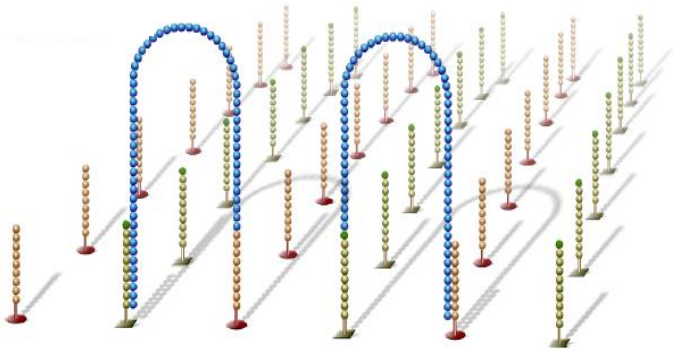




Amplification step 3



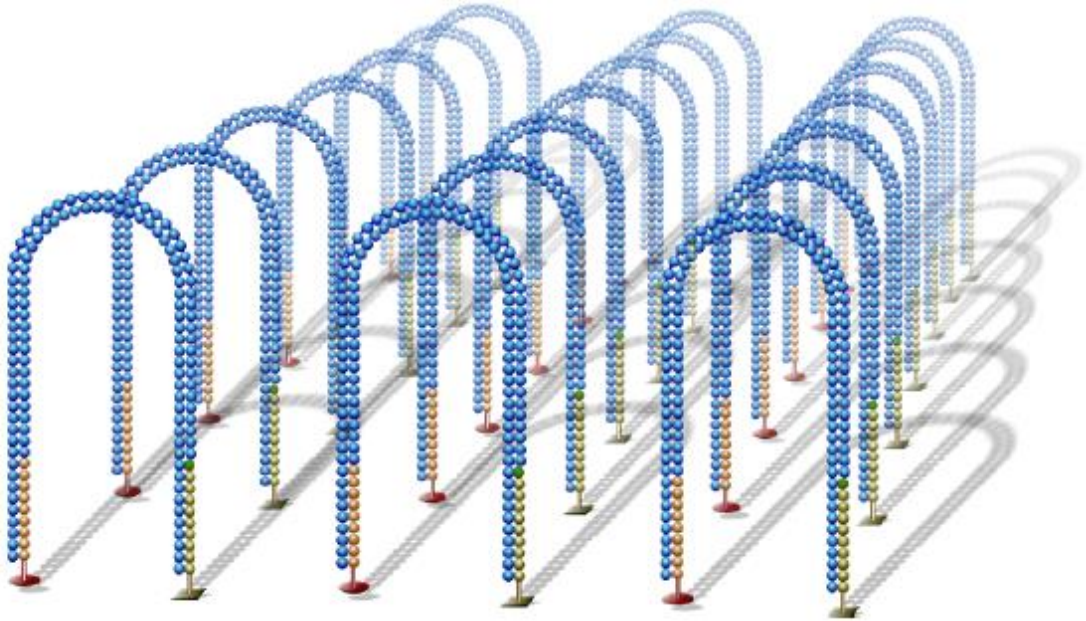
- dsDNA bridge is denatured



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

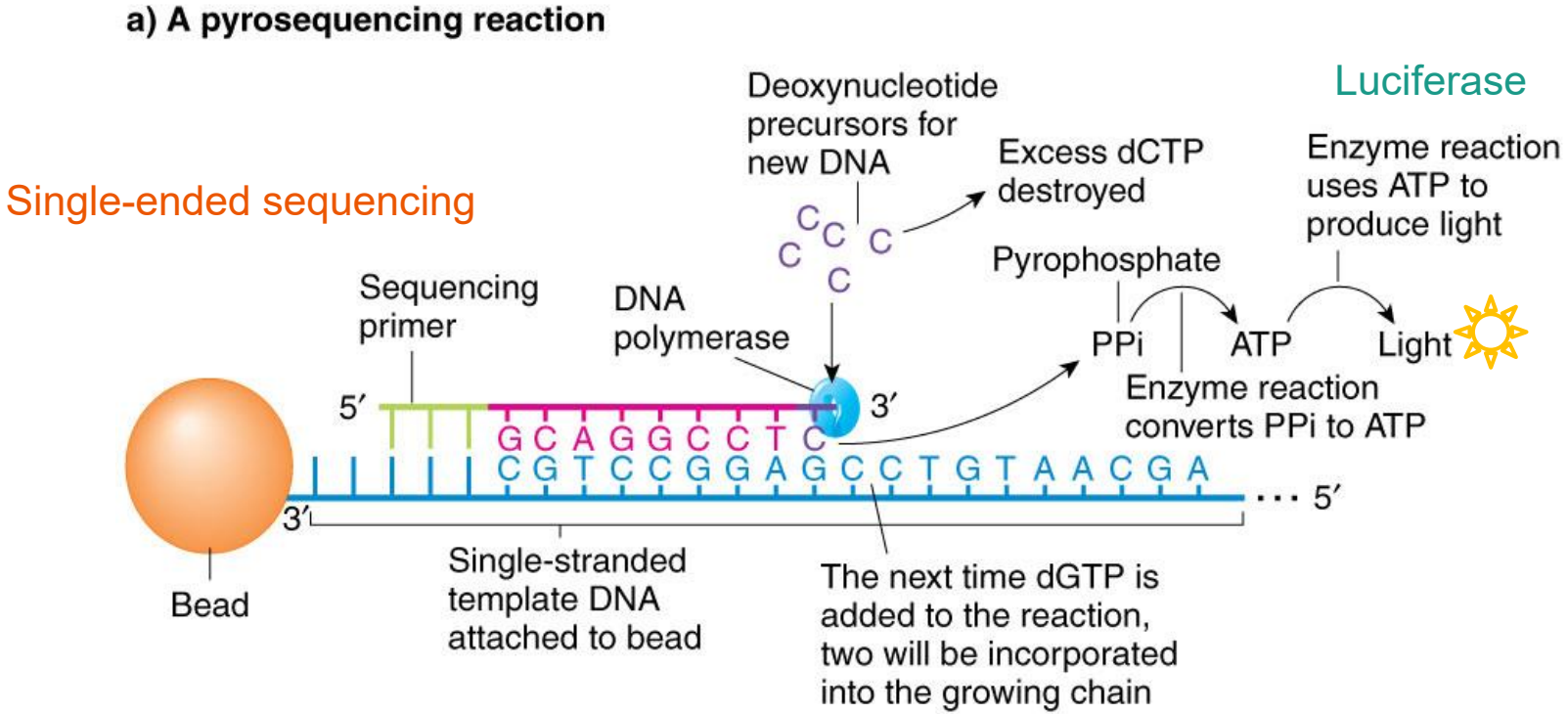
Amplification step 4

- Bridge amplification cycles repeated many times

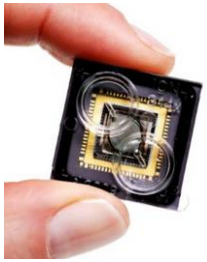


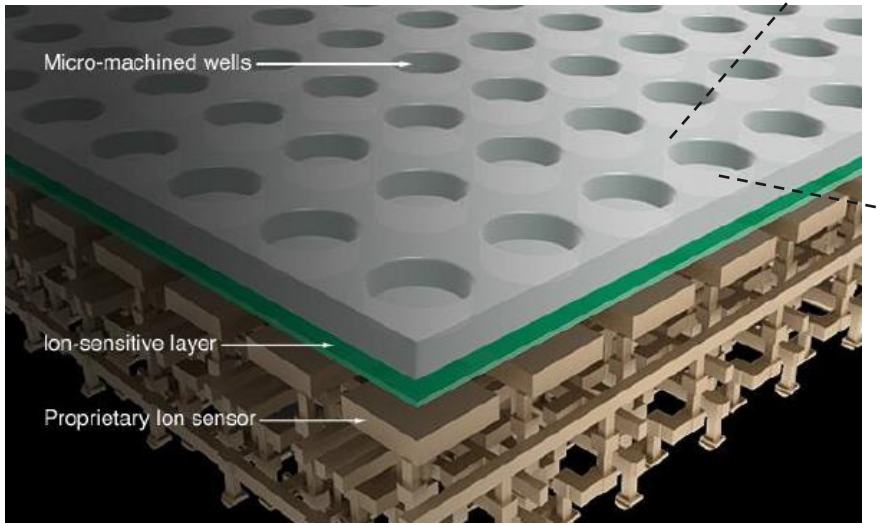
454 Pyrosequencing

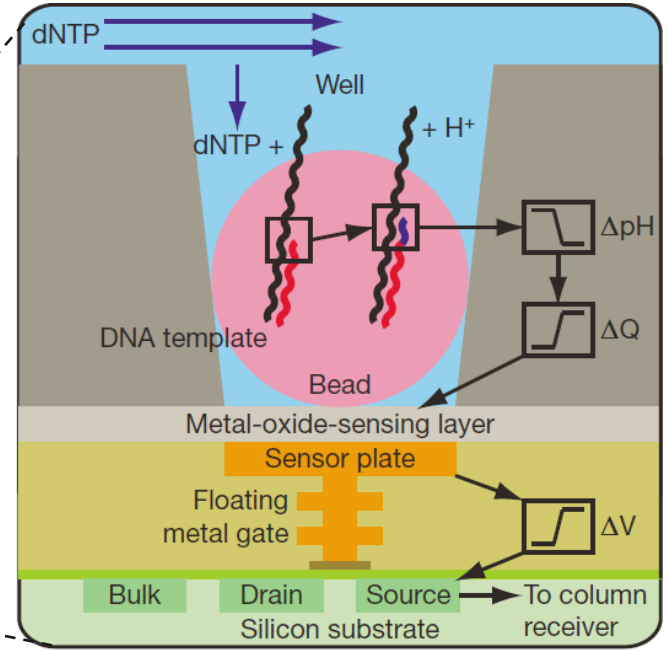
Single-ended sequencing



Ion Torrent

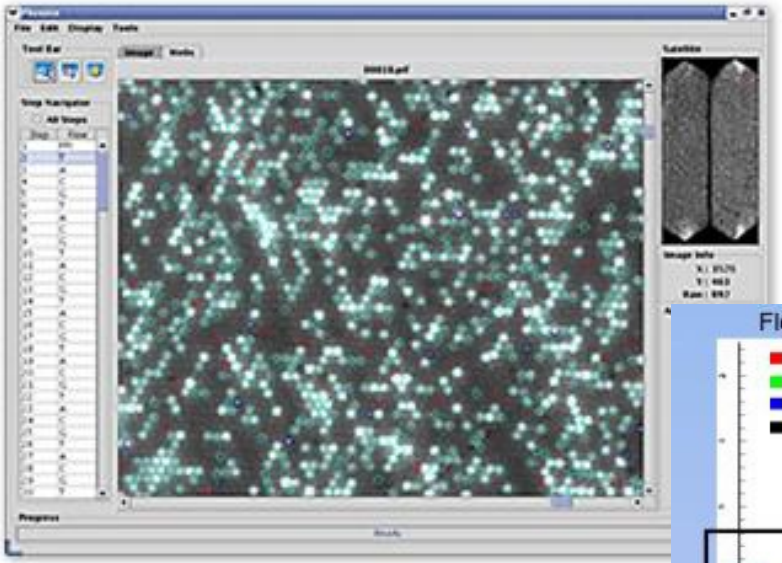






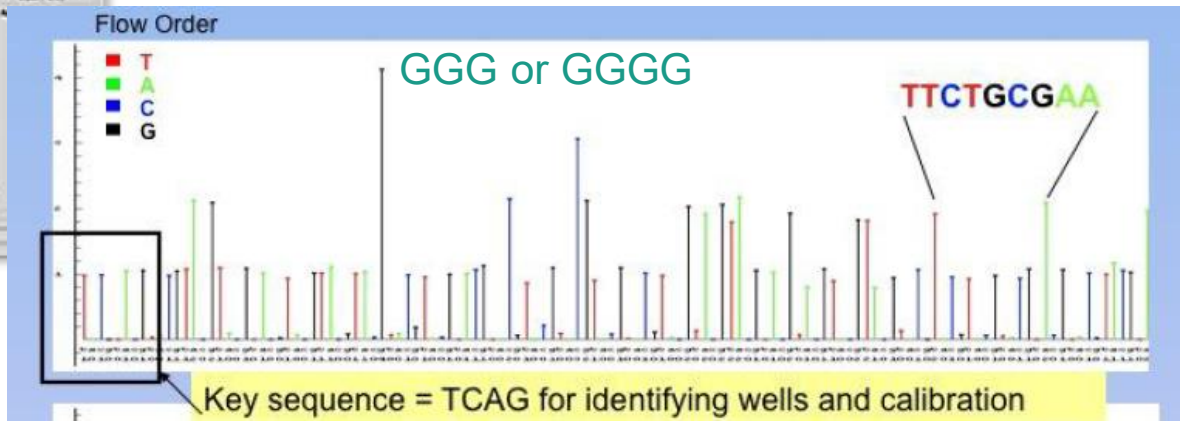
- Measure changes in pH
- Also has **homopolymer** limitation

Limitation of 454 pyrosequencing and Ion Torrent



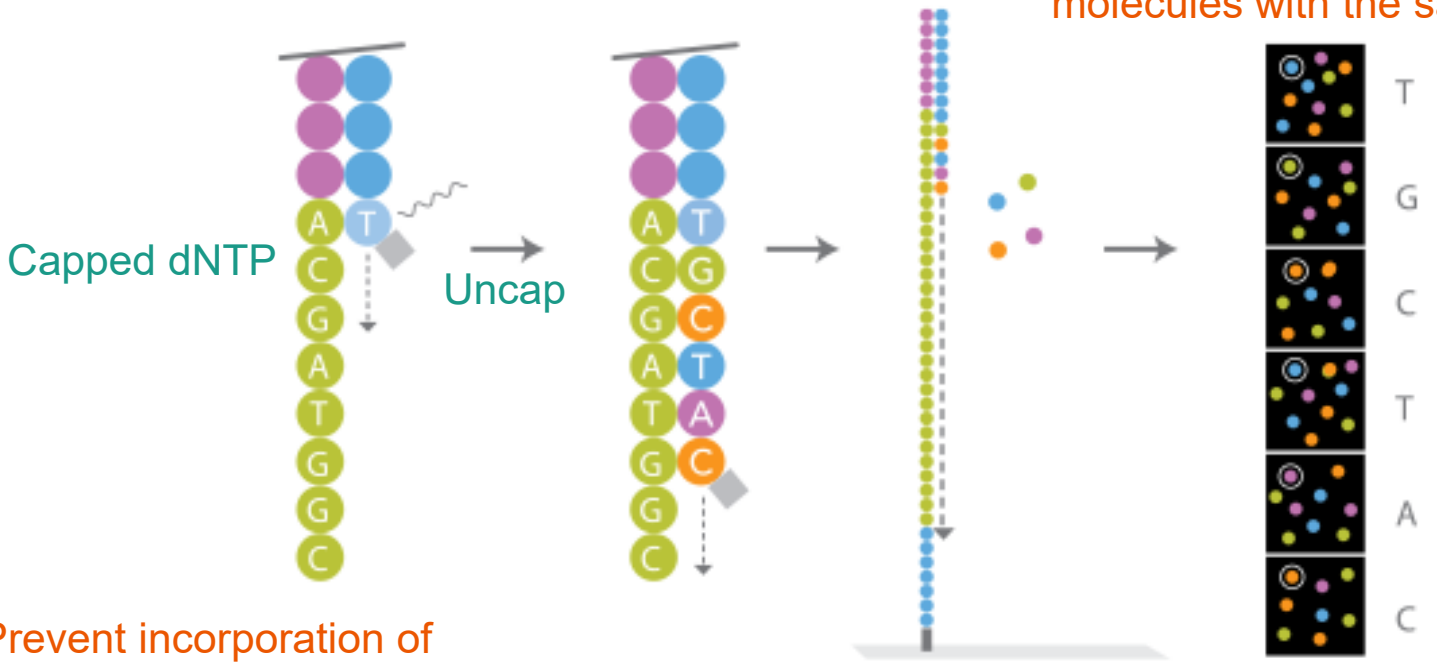
DATA ANALYSIS: OUTPUT PACKAGE

- Light intensity and pH signals are non-linear and may saturate
- Difficult to distinguish **homopolymer**
  - **AAAA** or **AAAAA**



Illumina / Solexa

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



Prevent incorporation of multiple dNTP

http://www.bioinformatics.nl/

Pros and cons of NGS techniques

Platform	Read Length	Run Time	Gb/ Run	Advantage	Disadvantage
454 Pyrosequencing	400+	1 day	0.7	Long read length	Homopolymer error
Illumina	50-300	5 days	600	Low cost per base	Short read length Long run time
Ion Torrent	200-400	2 hrs	100	Fast run times	Homopolymer error

Tradeoffs

Sanger

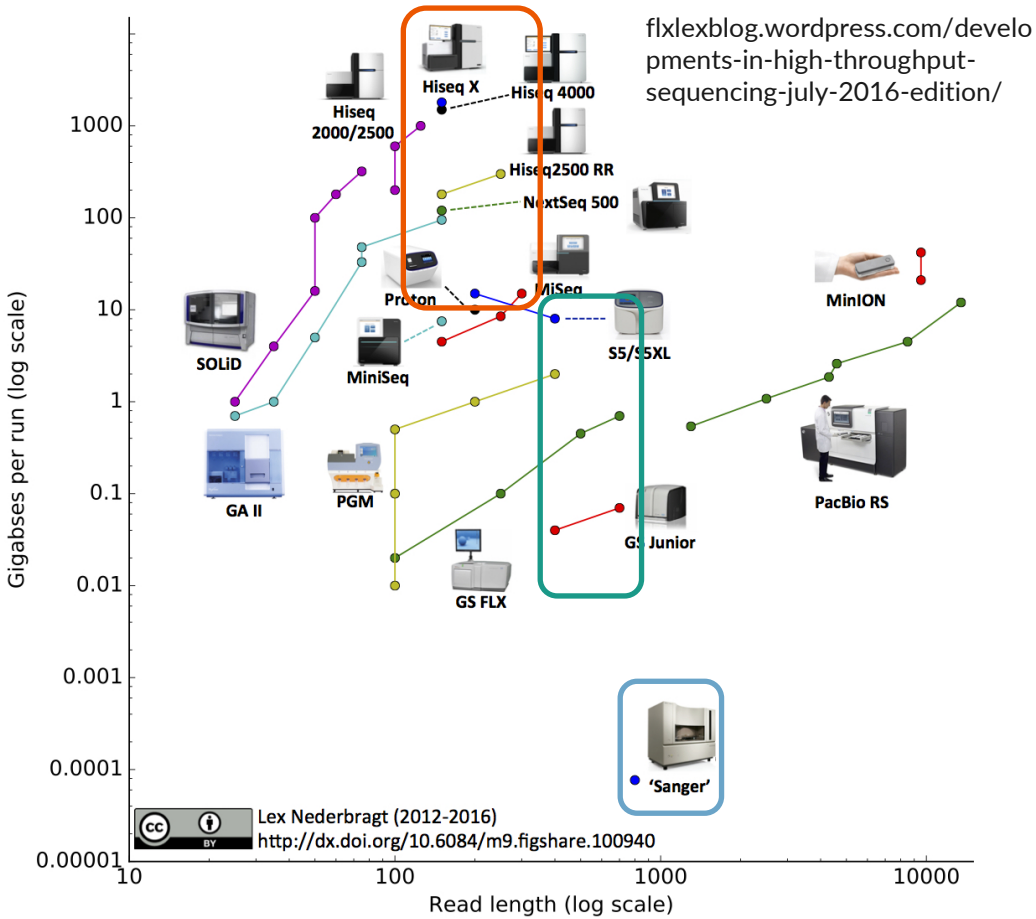
- 1000 bp, low throughput
- Use to validate small DNA

(454) and Ion Torrent

- 400+ bp, medium throughput
- Use when fast turn-around time is needed

Illumina

- <300 bp, high throughput
- Primary technology today



3rd Generation Sequencing (Long-Read)

Single-Molecule Real-Time (SMRT) sequencing

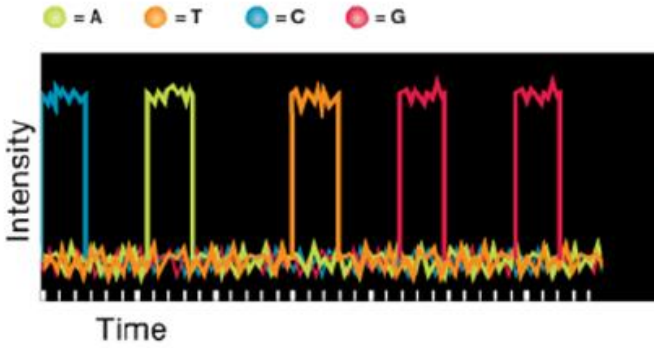
Zero-mode waveguide (ZMW)

Phospholinked nucleotide

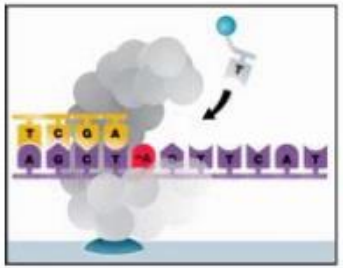
Images from Pacific Biosciences

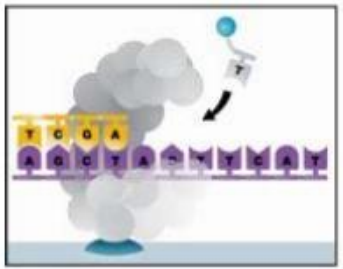
- Faster, more durable DNA polymerase
- Very small wells each containing a single DNA molecule
  - Zero-mode waveguide = nanophotonic confinement structure
  - Allow monitoring of fluorescence signal from individual reaction
- No amplification = direct quantification of DNA/RNA abundance

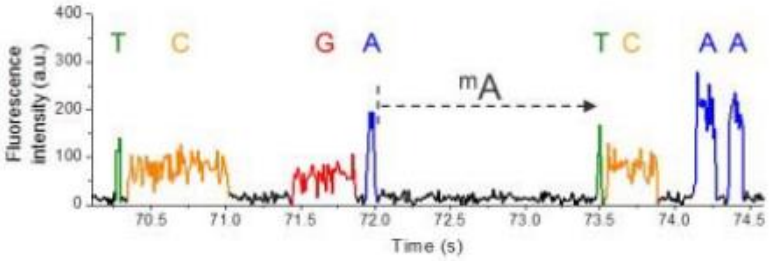
Video data from SMRT-seq

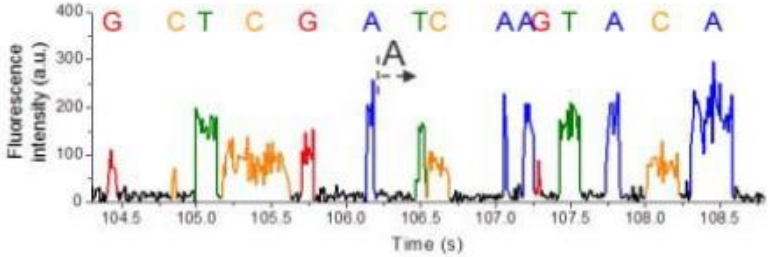


Images from Pacific Biosciences



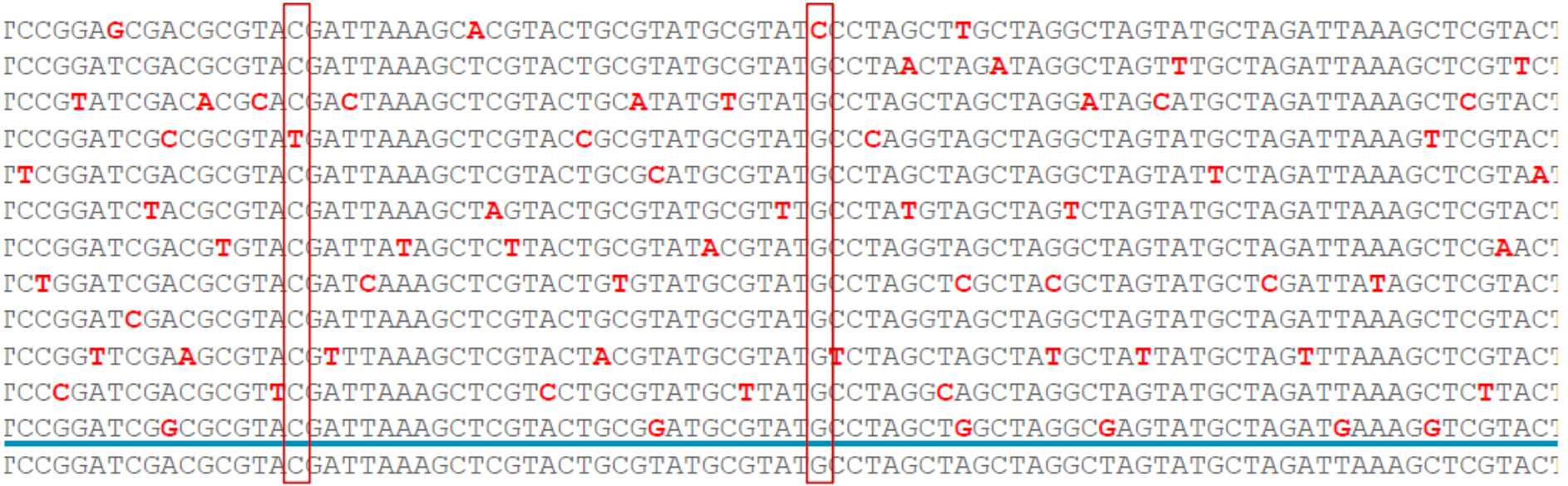






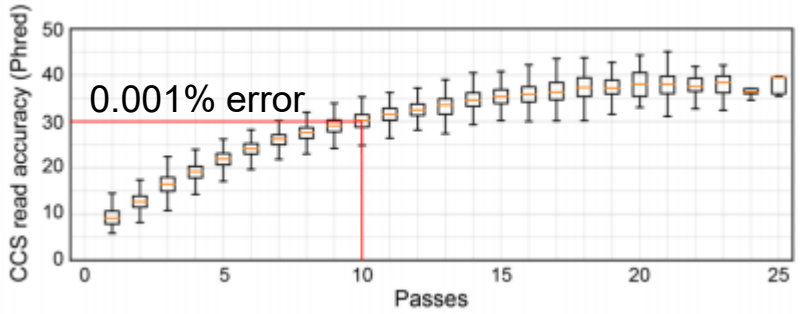
- Compared to image data from Illumina platform
- Video gives time information → identification of modified DNA/RNA

High (random) error rate

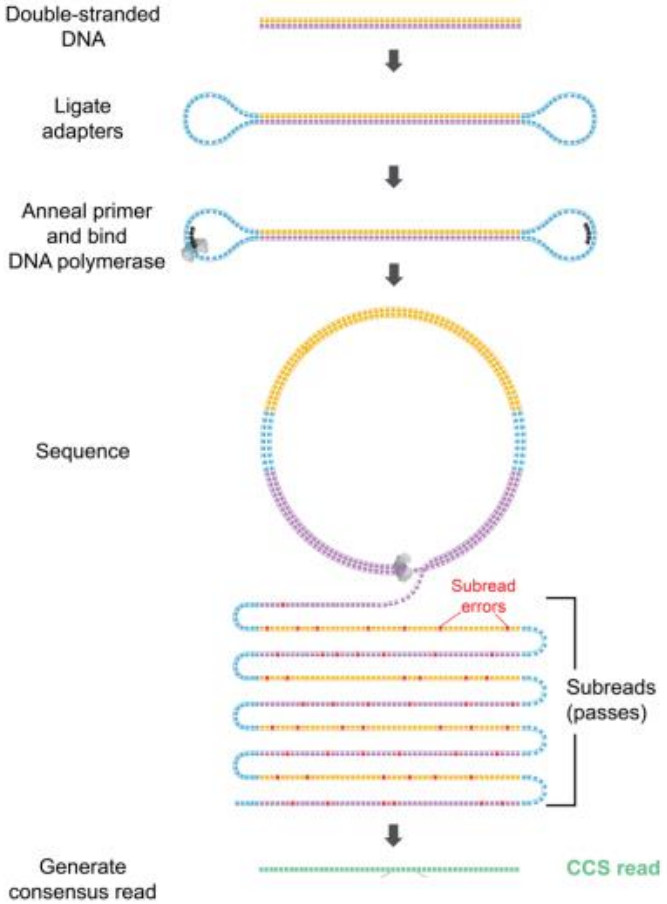


- 5-15% error compared to 0.01% of Illumina
- How to solve this problem?

Circular consensus sequencing

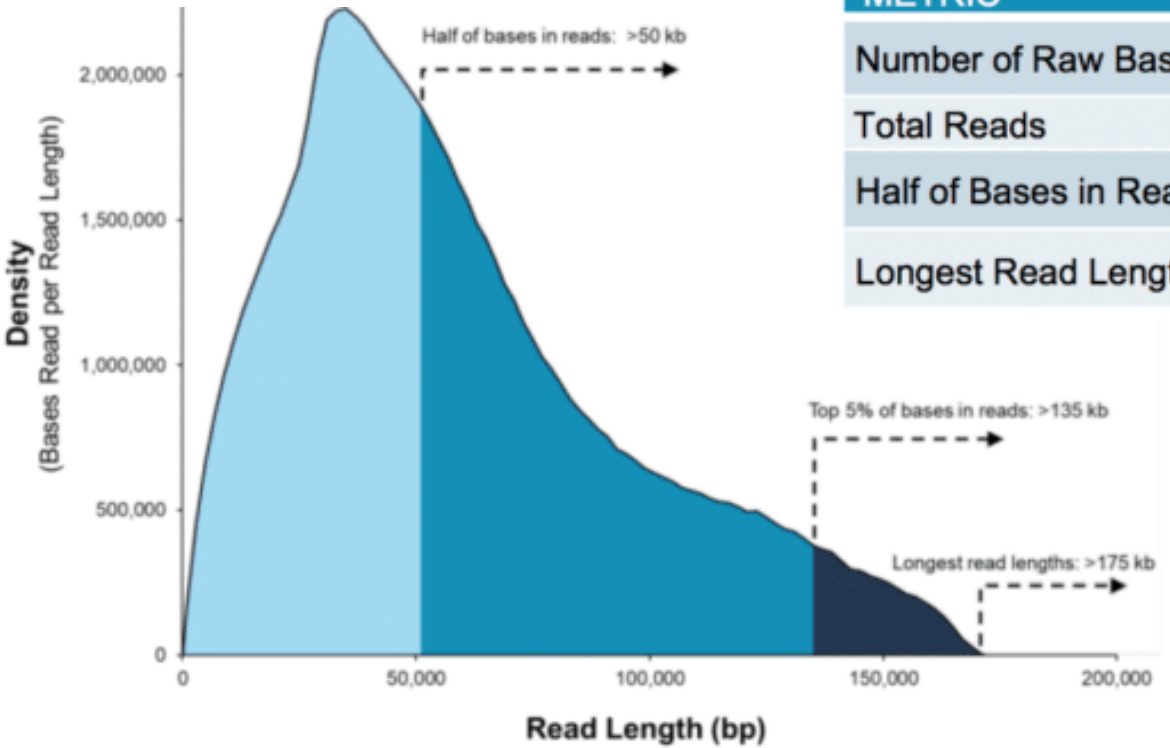


- Circular extension of each DNA molecule
- Read the extended molecules = multiple re-sequencing of the original sequence
- Take the consensus (majority vote)
- P(correct base in >*k* of *N* passes) ~ Binomial



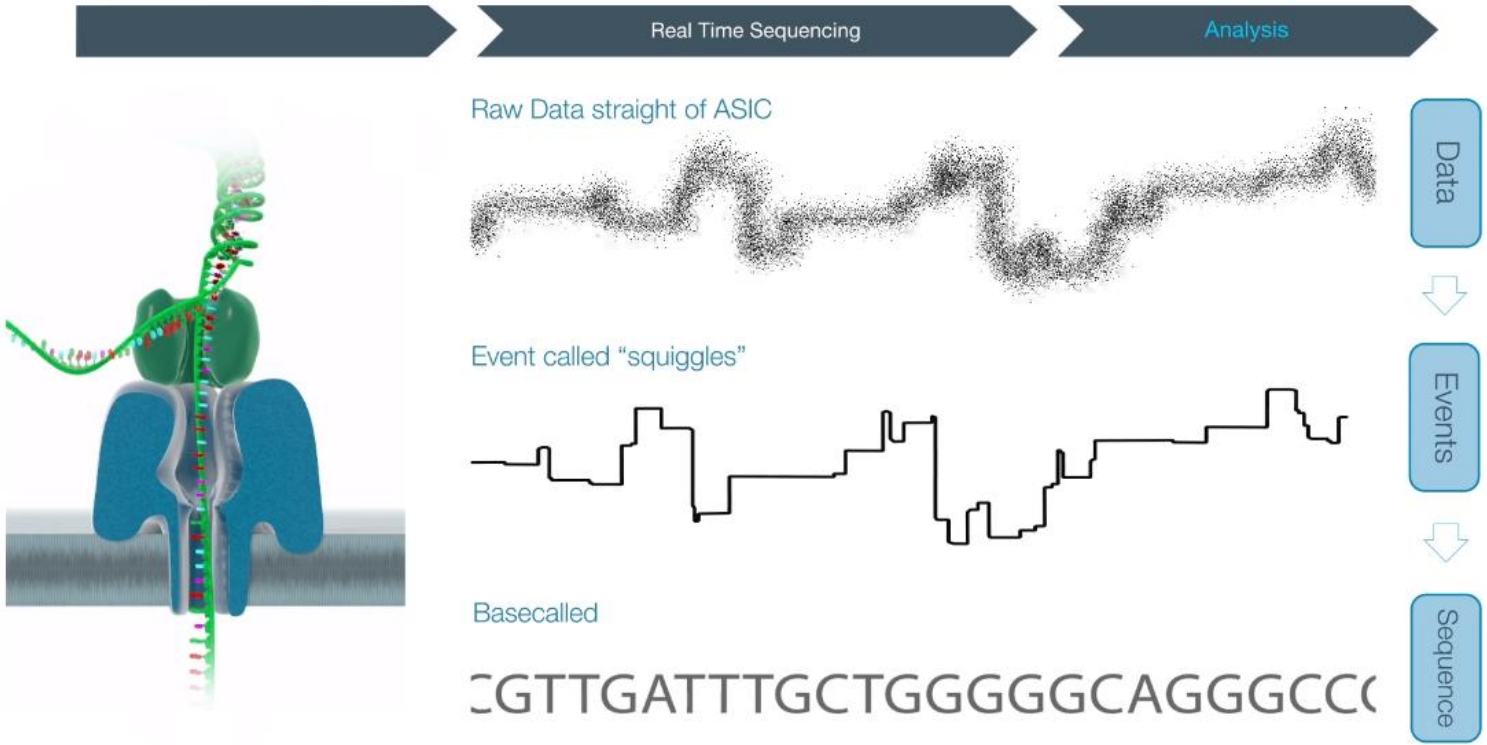
Images from Pacific Biosciences

Long 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

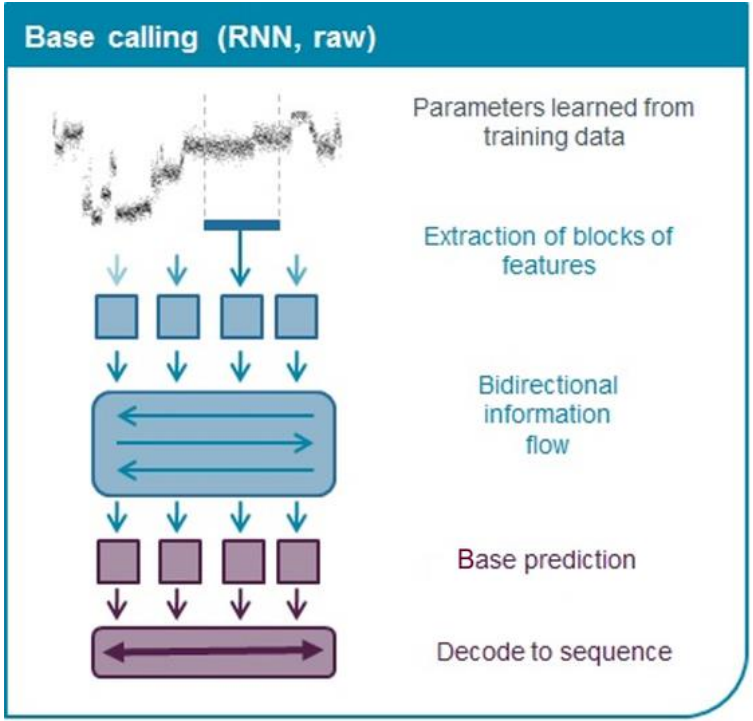
Nanopore



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

Basecalling with deep neural networks

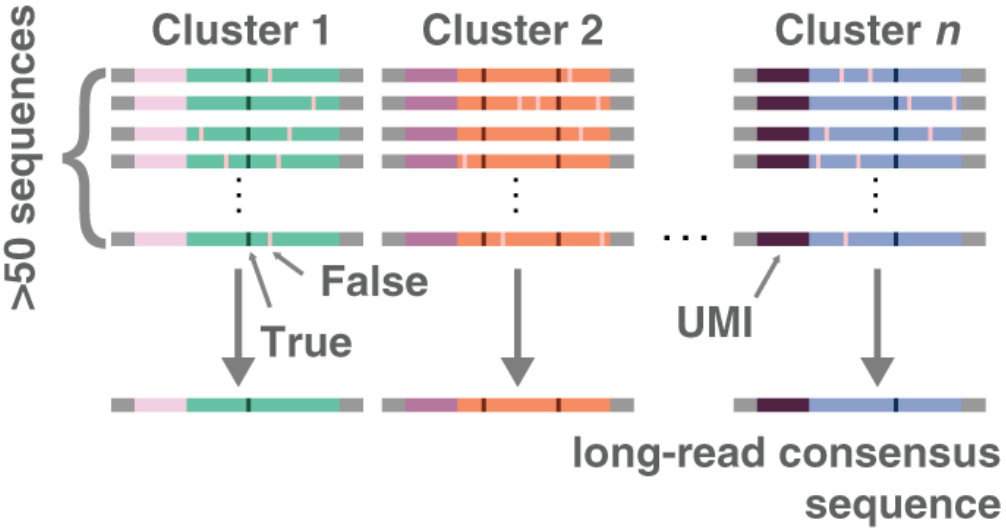
- Trained using data from synthetic DNA (known sequences)
- 14% baseline error
- Improved to 1-5% using bioinformatics and machine learning
- Require computer with GPU



https://nanoporetech.com/document/data-analysis

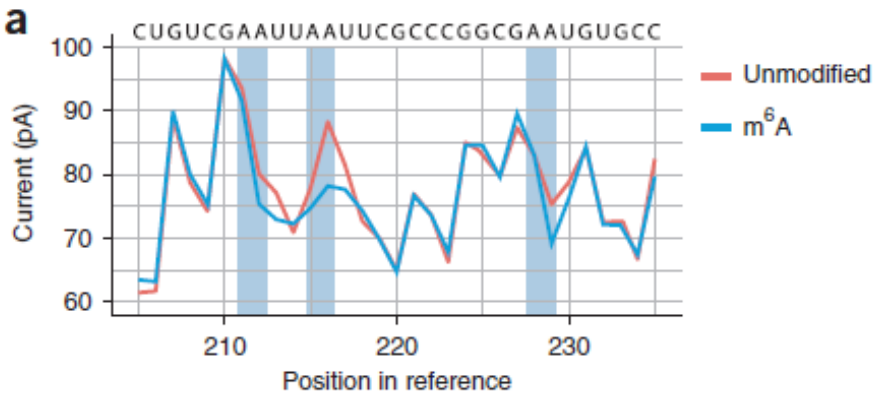
Consensus sequencing with UMI

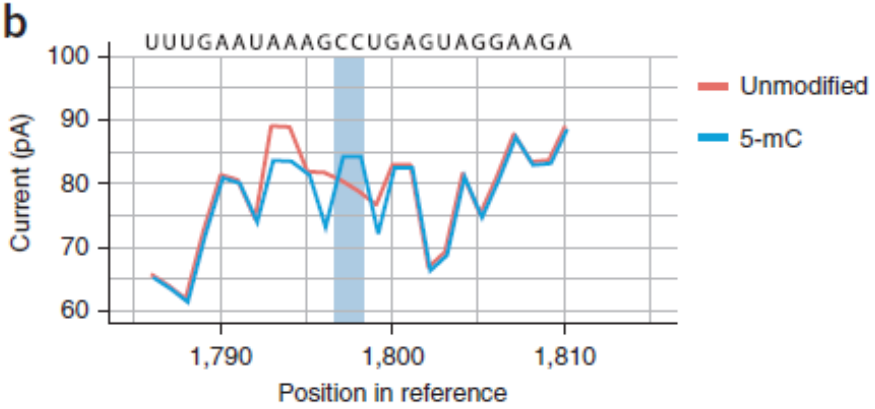
- Attach unique molecular identifier (**UMI**) tag to DNA molecules
- Amplify and sequence
- Cluster reads with the same UMI and call consensus sequences



Zurek, P.J. et al. Nature Communications 11:6023 (2020)

Detection of modified nucleotides

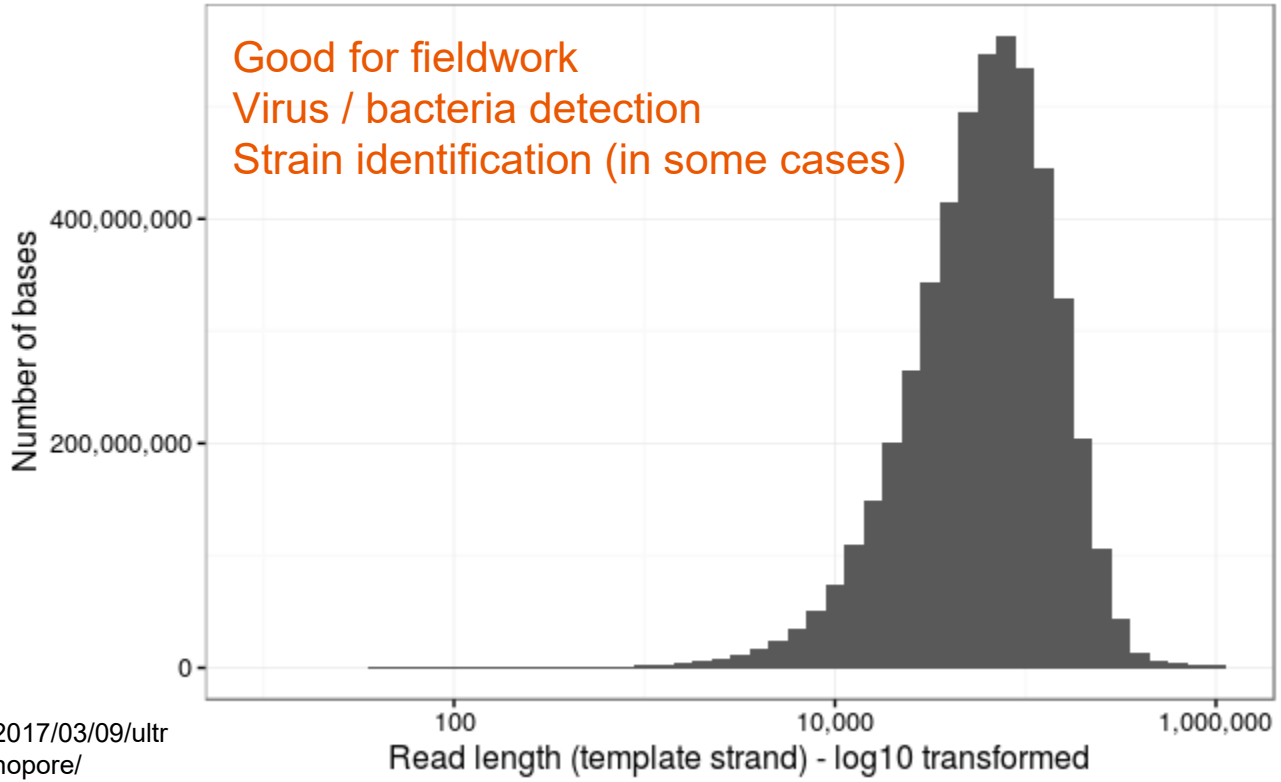




Geralde *et al*. Nature Methods 15, 201-206 (2017)

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

Very long read length up to Mb

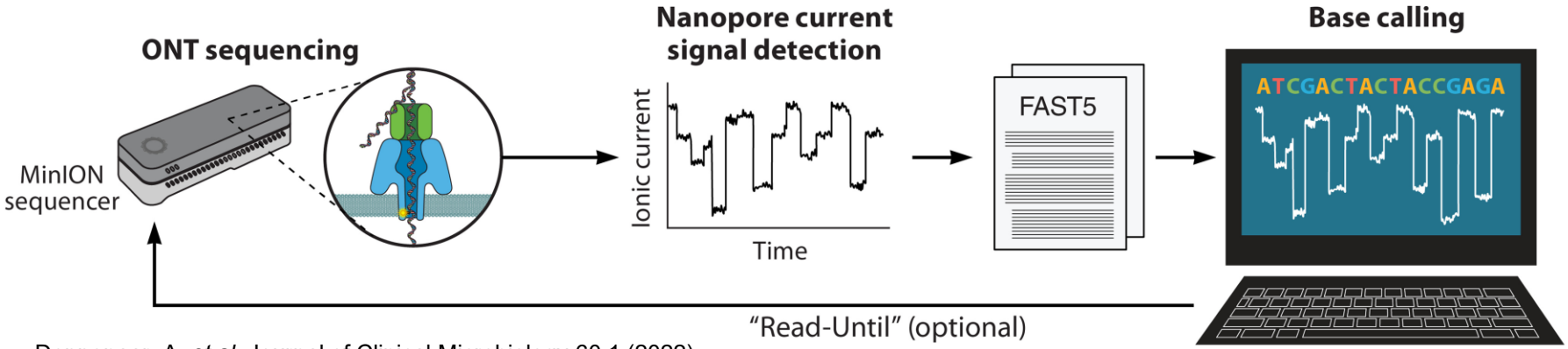


https://lab.loman.net/2017/03/09/ultrareads-for-nanopore/

Portability & fastest 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

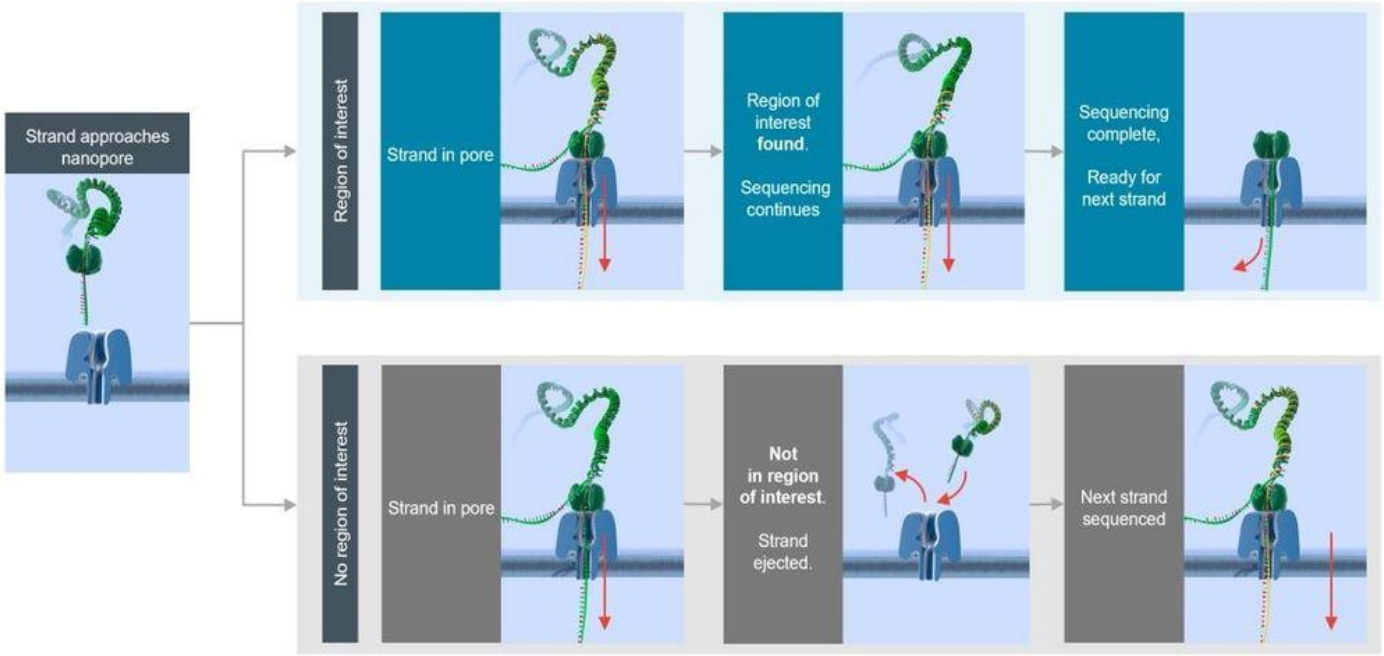
Real-time data with Nanopore



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)

Adaptive sampling

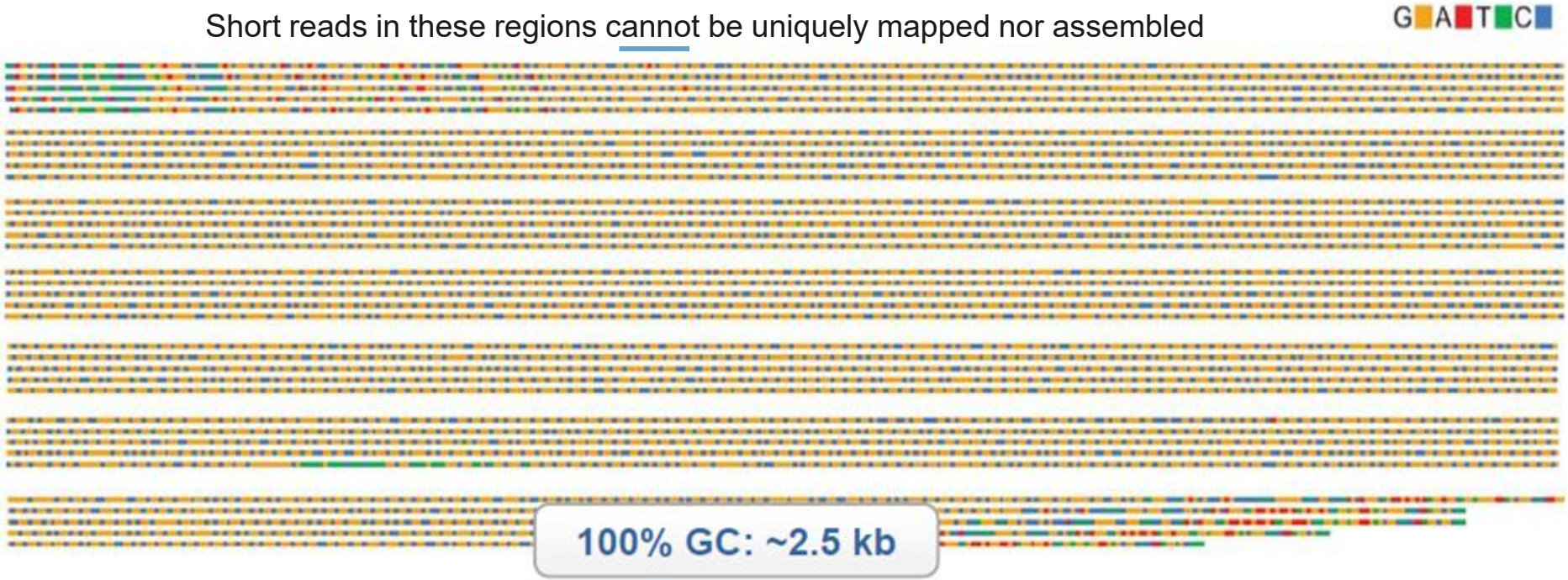


https://nanoporetech.com/document/adaptive-sampling

- Nanopore signal can be decoded in real-time to control the pores

Long read resolves repetitive regions

Short reads in these regions cannot be uniquely mapped nor assembled



- As much as 20% of human genome is like this!

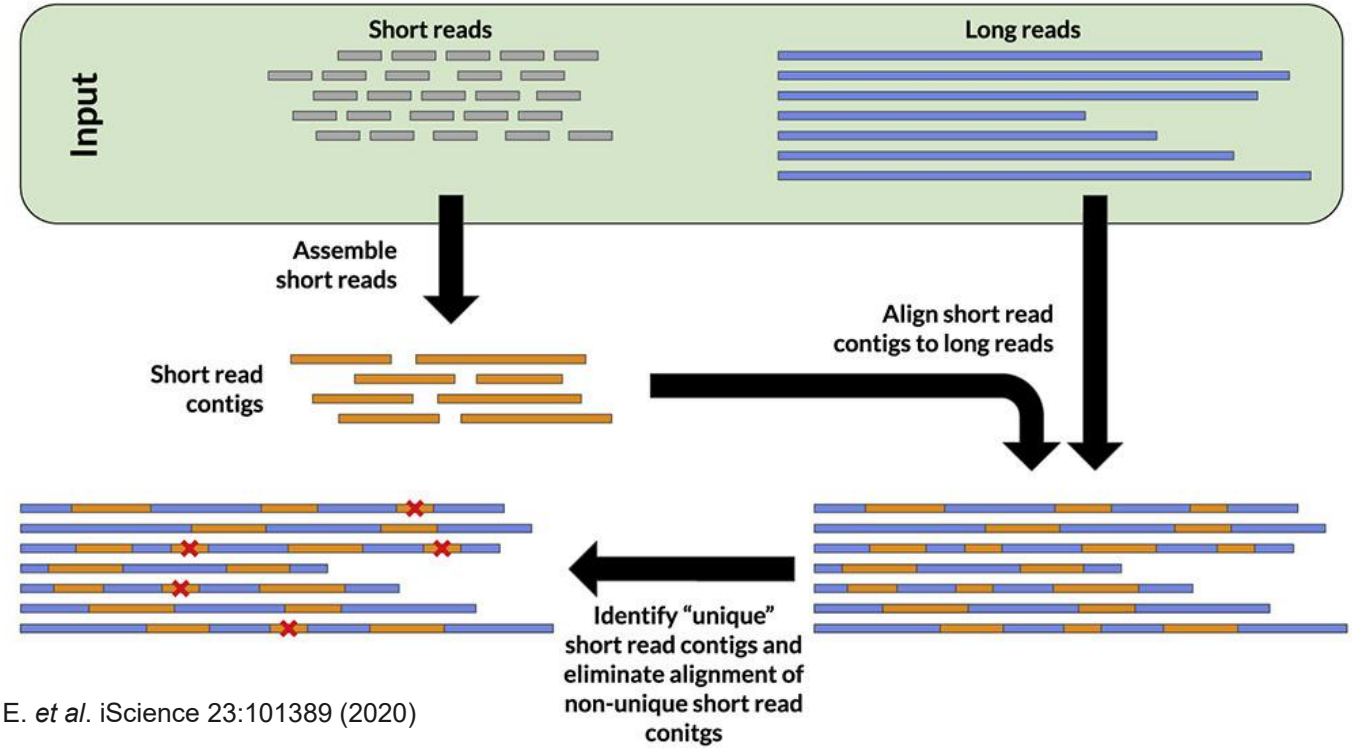
Long read resolves haplotype

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

Combining short and long read data



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

Pros and cons of NGS/long-read techniques

Platform	Read Length	Run Time	Advantage	Disadvantage
Illumina	50-300	5 days	Low cost per base High throughput	Short read length
PacBio SMRT-seq	10-25kb	10 days	Long read with high accuracy	Expensive Low throughput
Nanopore	30-100kb	<10 hours	Quick turn-around time Portable Ultra-long read Low instrument cost	Low accuracy

Choosing the right platform for your task

- Identify low frequency variants (mutations) — **Illumina**
- Gene expression
- Targeted sequencing

- Assemble new genomes — **Mixed**

- Clinical service time-scale — **Nanopore**
- Field study

- Structural variant, gene copy number — **PacBio**
- Identify new transcript (RNA) splice isoforms

Applications of DNA sequencing

Sequencing cost breakdown

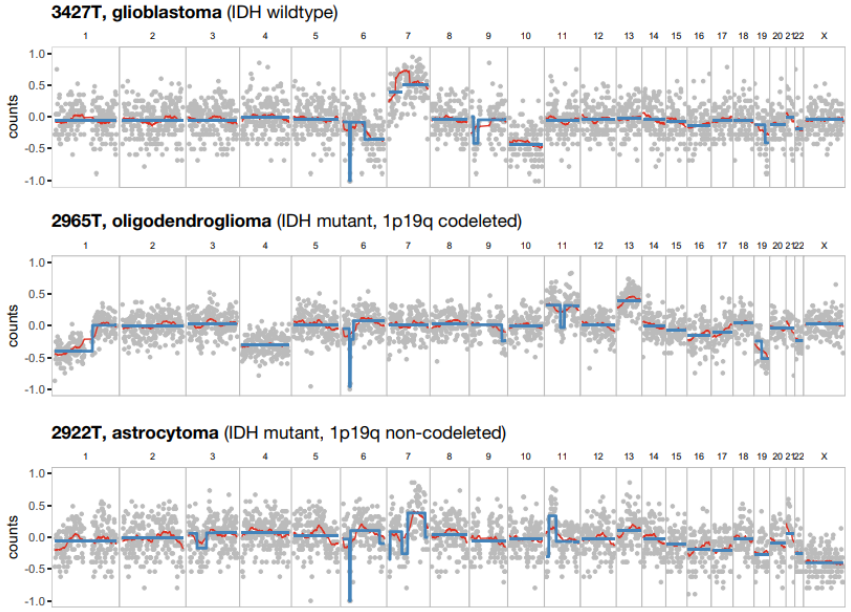
- Cost = Base Pair = **Scope** x **Depth**

Reduced scope

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

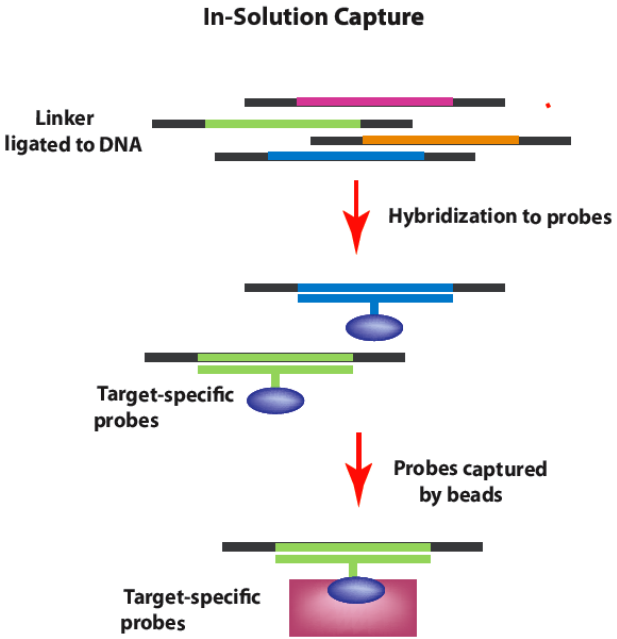
Reduced depth

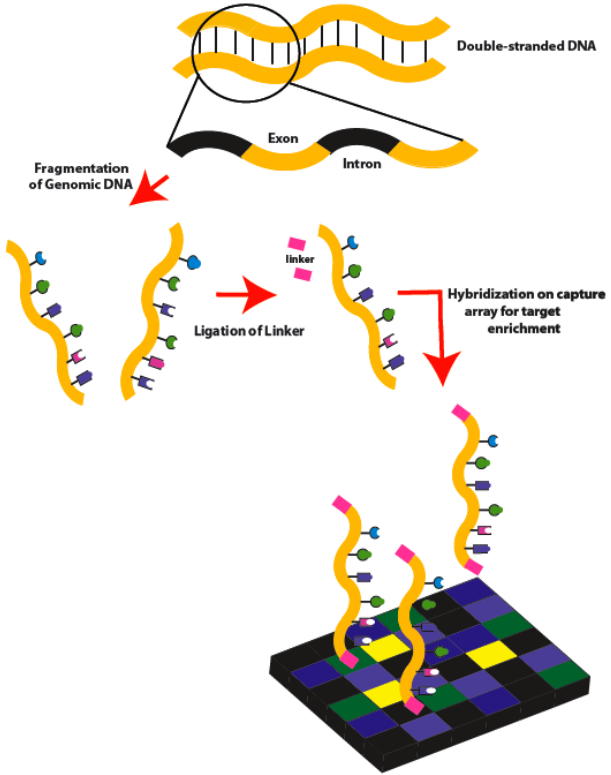
- Ultra-low pass
  - Detect chromosomal copy alternation
  - Estimate tumor fraction



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

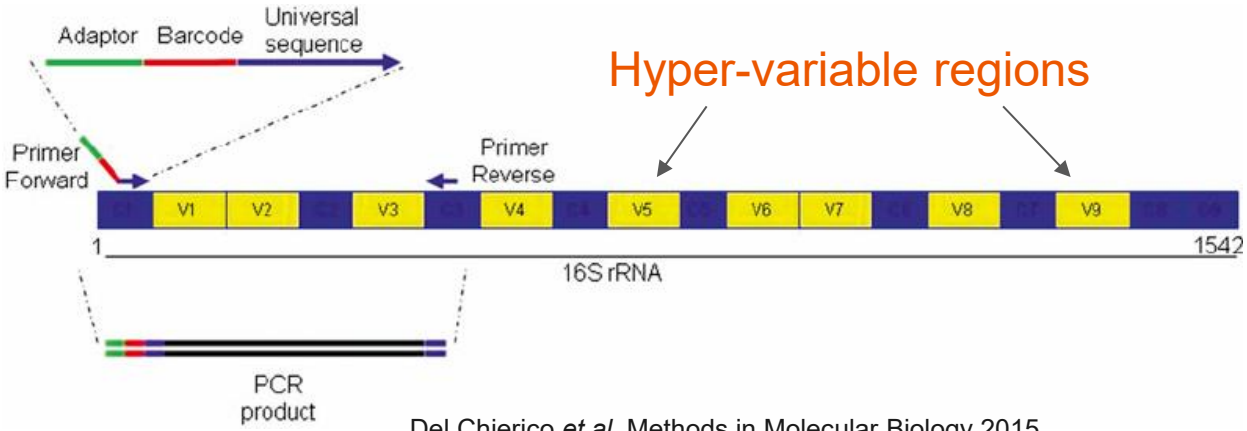
Exon enrichment for exome sequencing





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

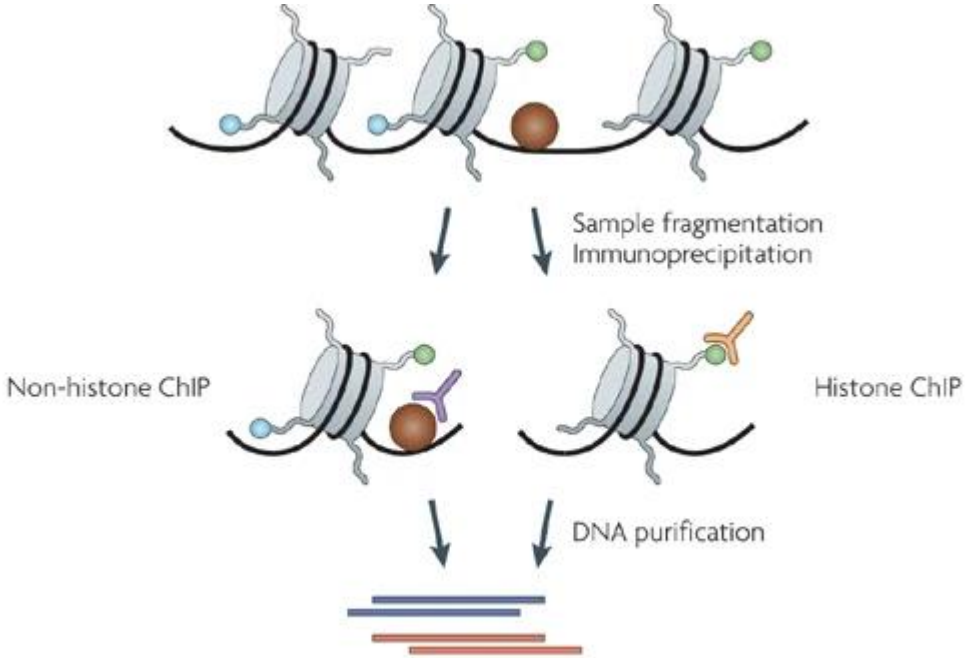
rRNA amplicon sequencing

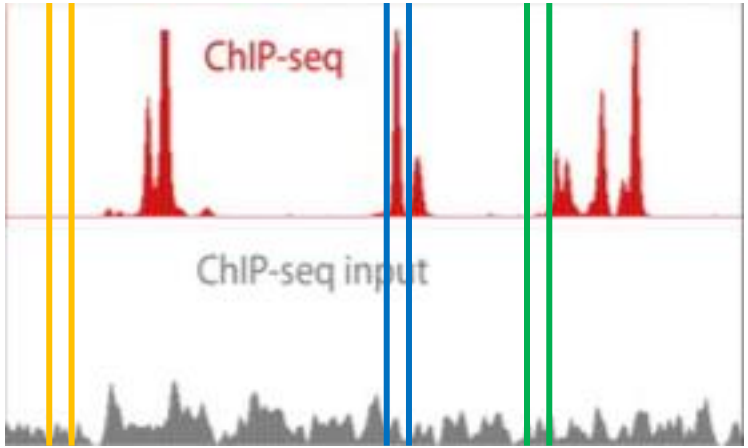


Del Chierico *et al*. Methods in Molecular Biology 2015

- **Bacteria**: 16S rRNA
- **Fungi**: internal transcribed spacer (ITS) located between rRNA genes
- Provide taxonomy, composition details

Chromatin immunoprecipitation



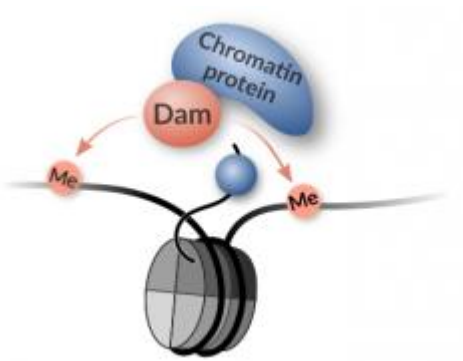


Park et al. Nat Rev Genet 10:669-680 (2009)

- DNA-bound protein / histone modification

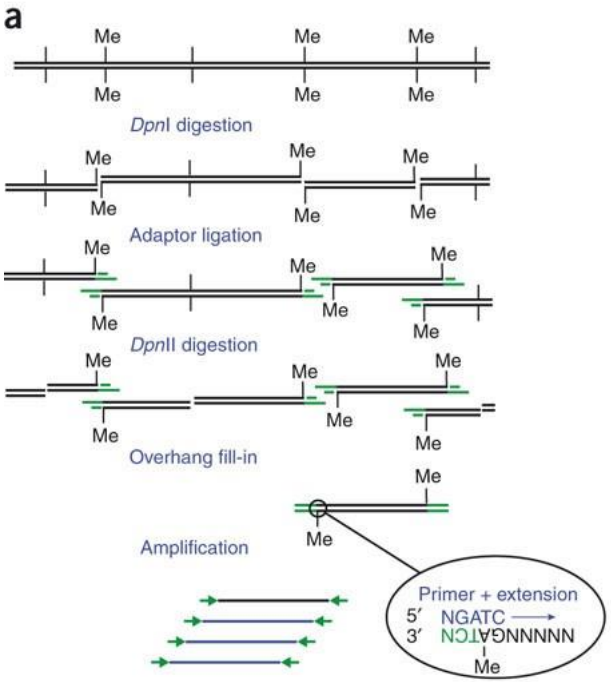
DNA sequencing coupled with molecular techniques (labeling, modification)

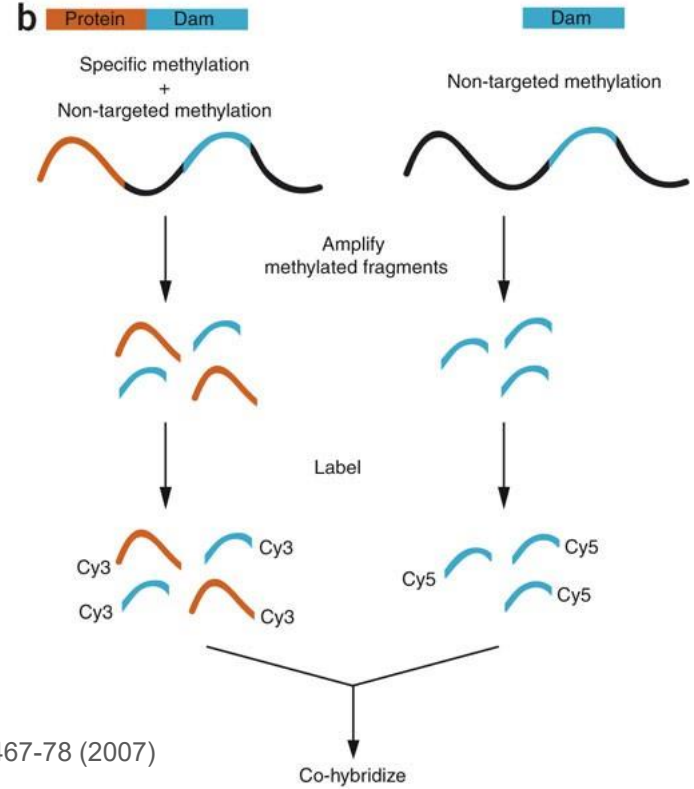
DNA adenine methylatransferase (DamID)



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

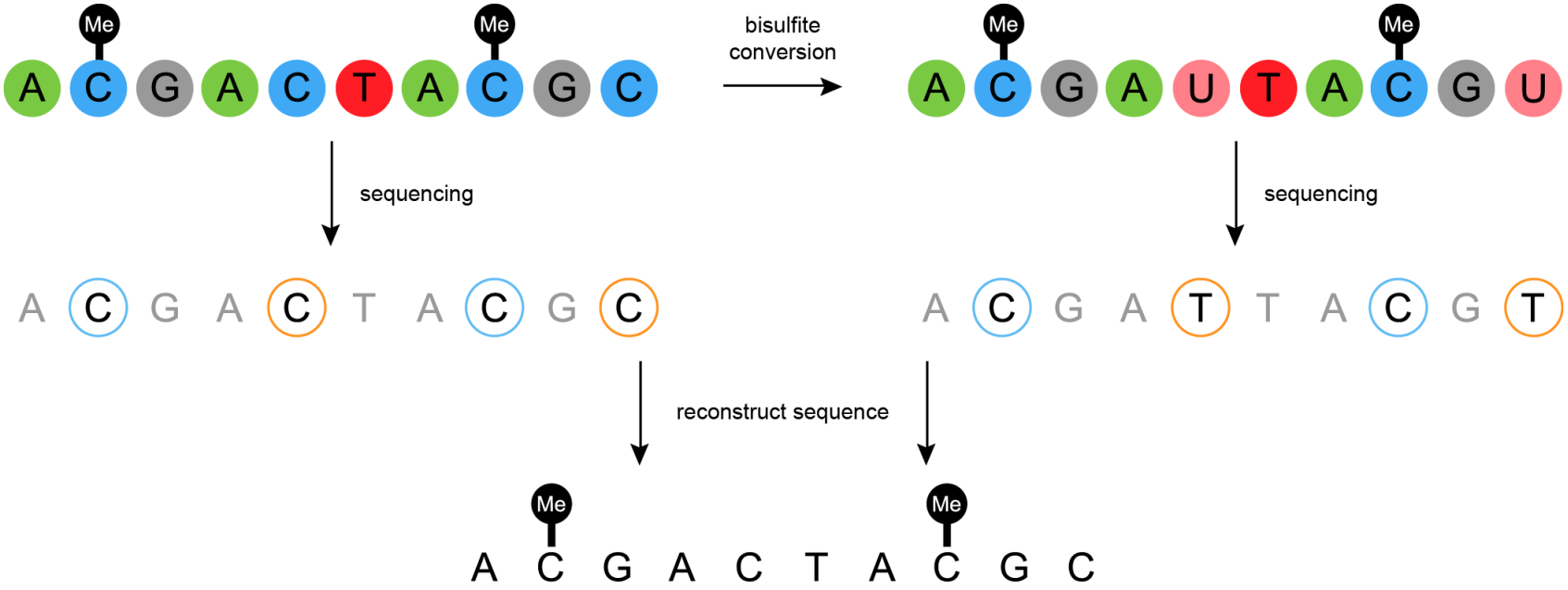
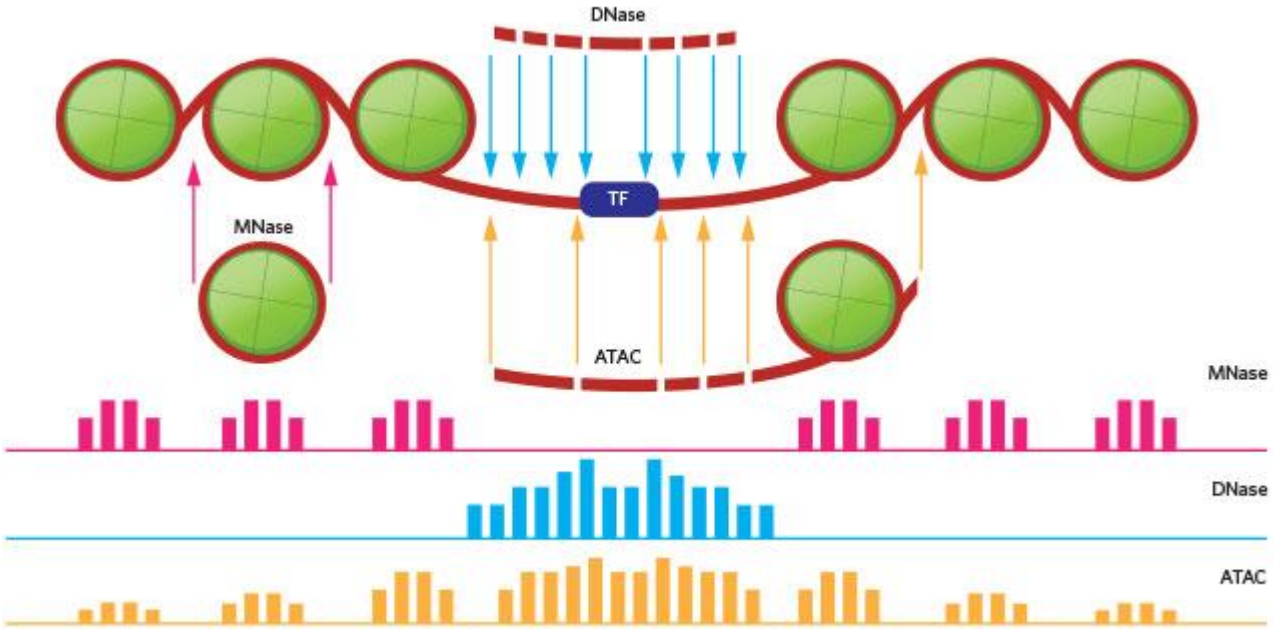


Image from http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis

Targetting bound or unbound chromatin



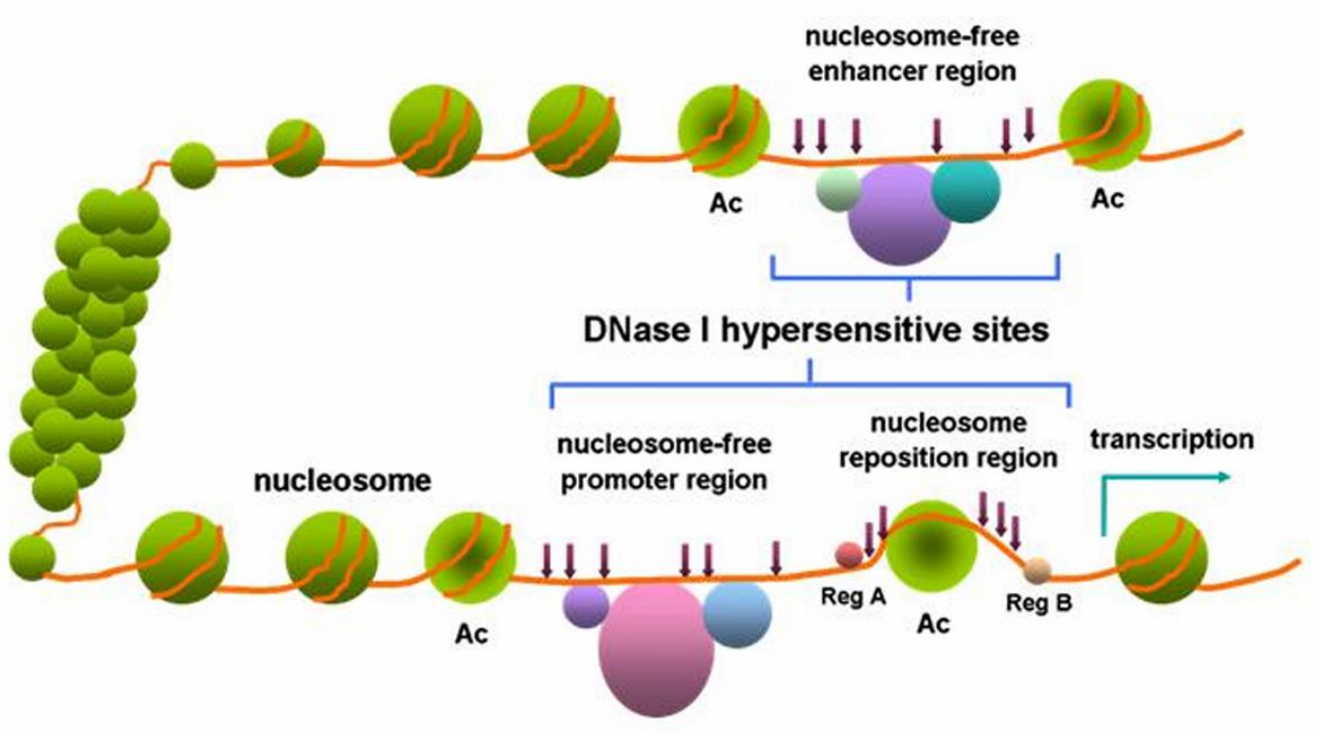
Bounded DNA

Unbounded DNA

Unbounded DNA + 1-2 nucleosomes

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

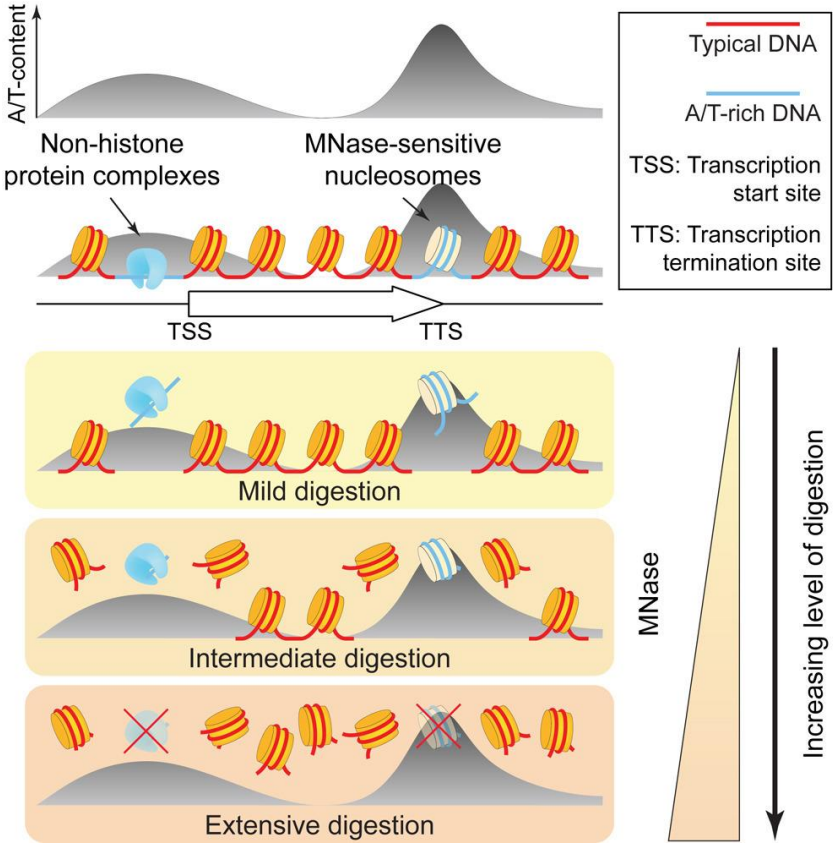
DNAse I cuts free chromatin → size-selection



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

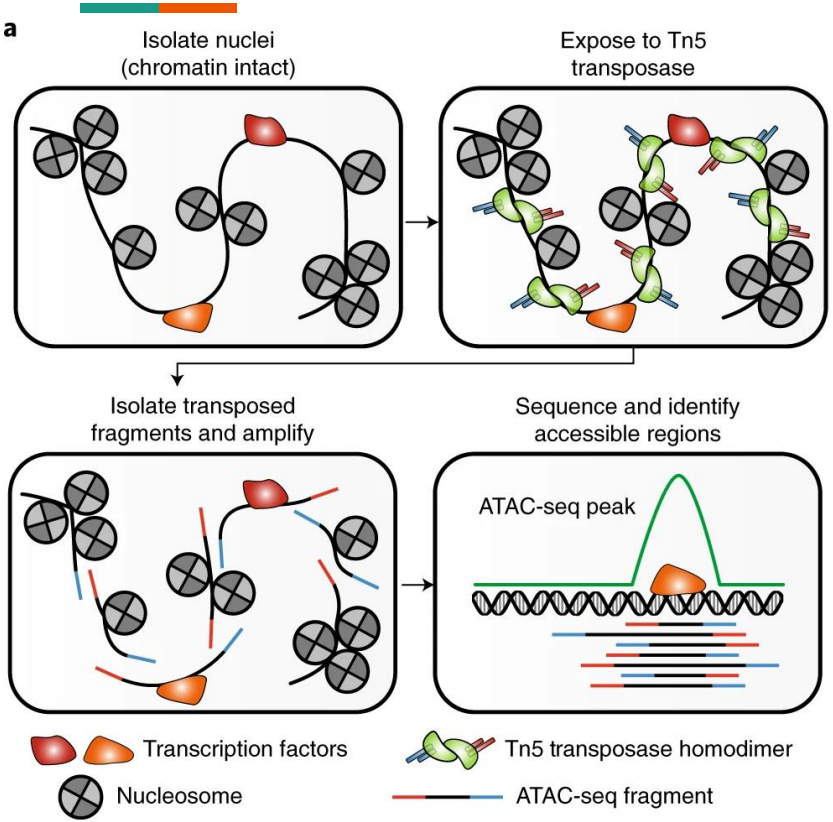
MNase

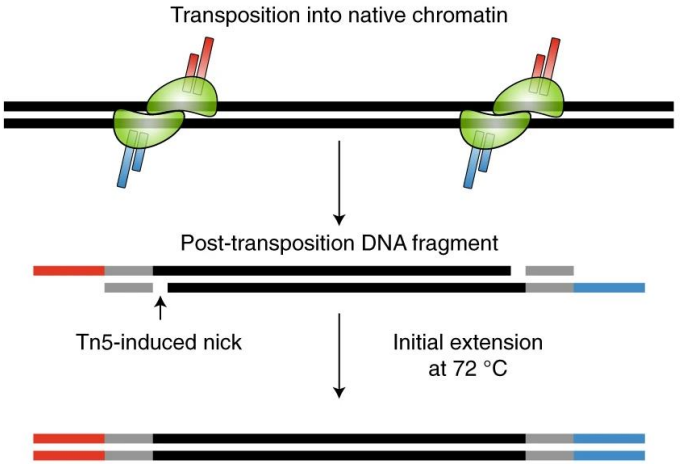
- Micrococcal nuclease = endo-exonuclease from *S. aureus*
- Target free chromatin more efficiently than DNase (smaller molecule)
- Less able to digest nucleosome-bound DNA than DNase
- **Good for mapping nucleosome locations**



Chereji, R.V. et al. Molecular Cell 65:565-577 (2017)

Assay for transposase-accessible chromatin (ATAC)

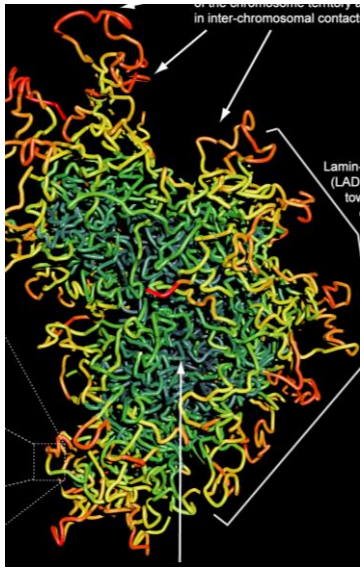


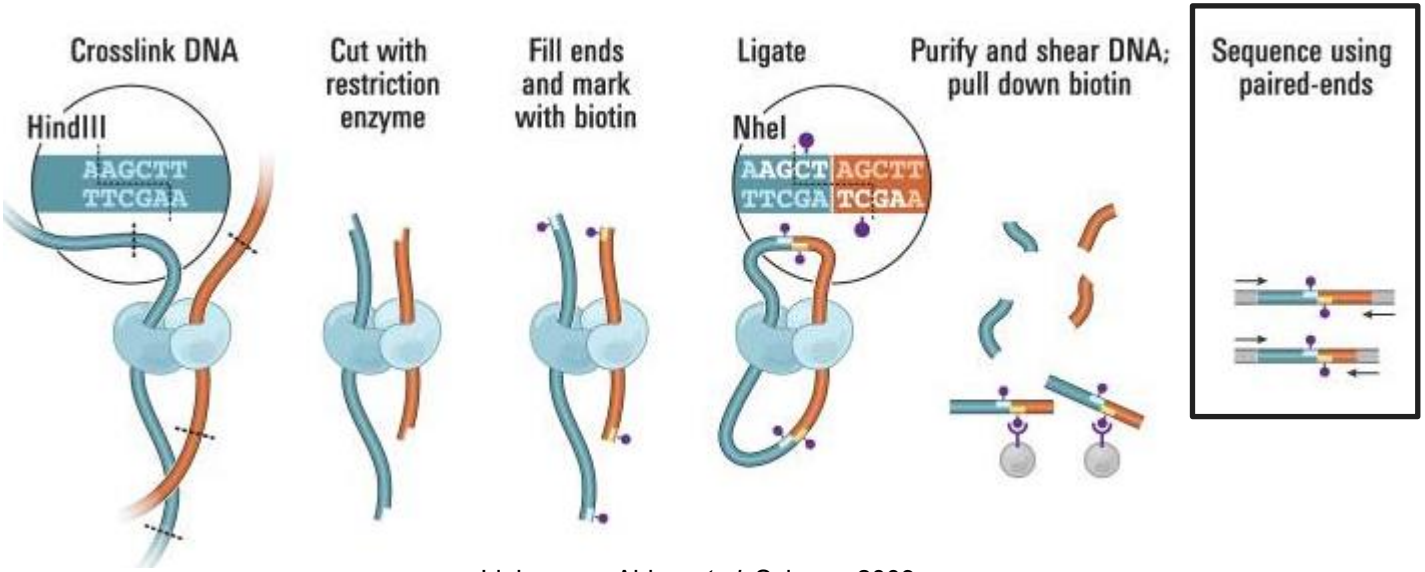


- Transposase Tn5 inserts sequencing adapter on open chromatin

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

Chromatin conformation capture

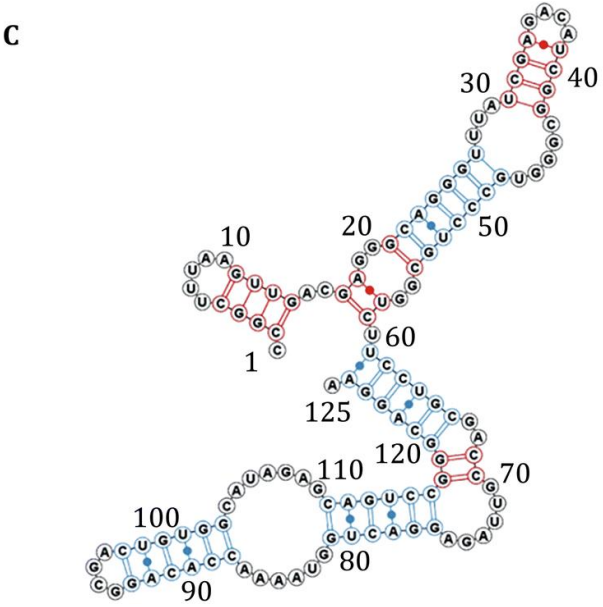


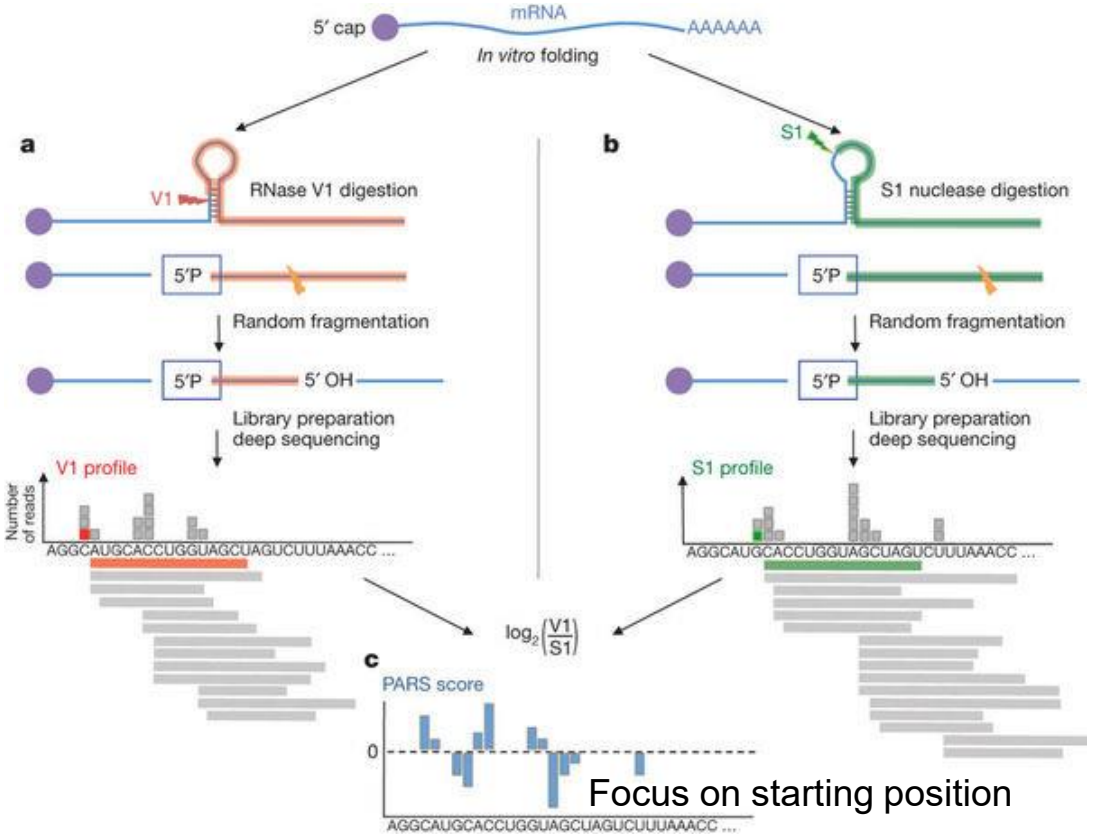


Lieberman-Aiden *et al*. Science 2009

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

Probing RNA secondary structure





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

Focus on starting position

Any question?

- See you next time