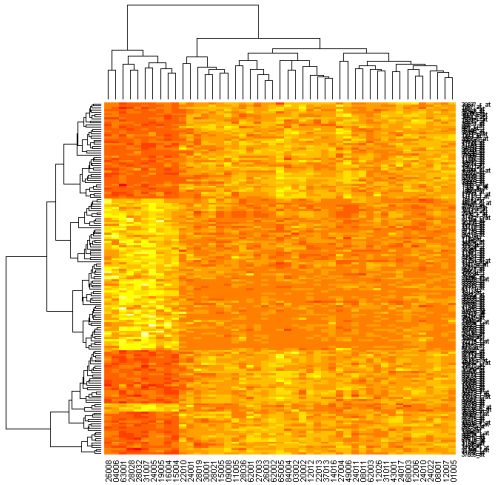
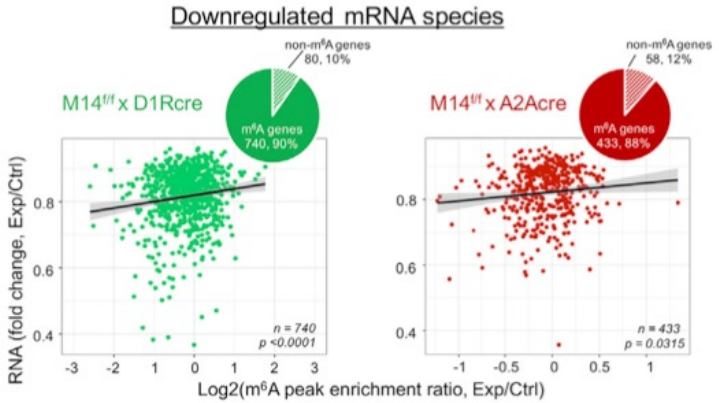
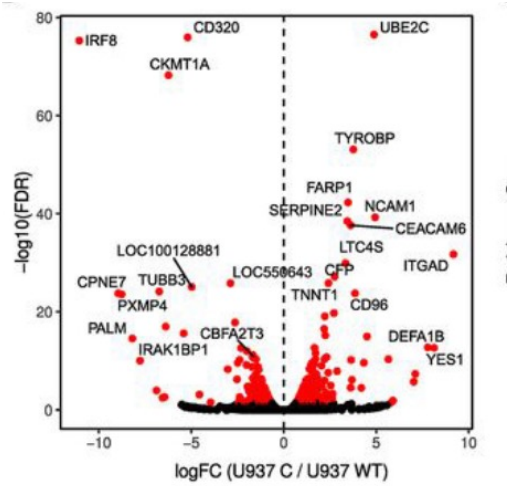
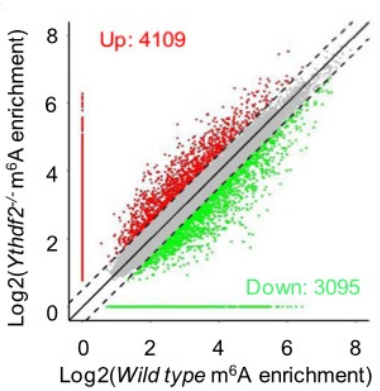
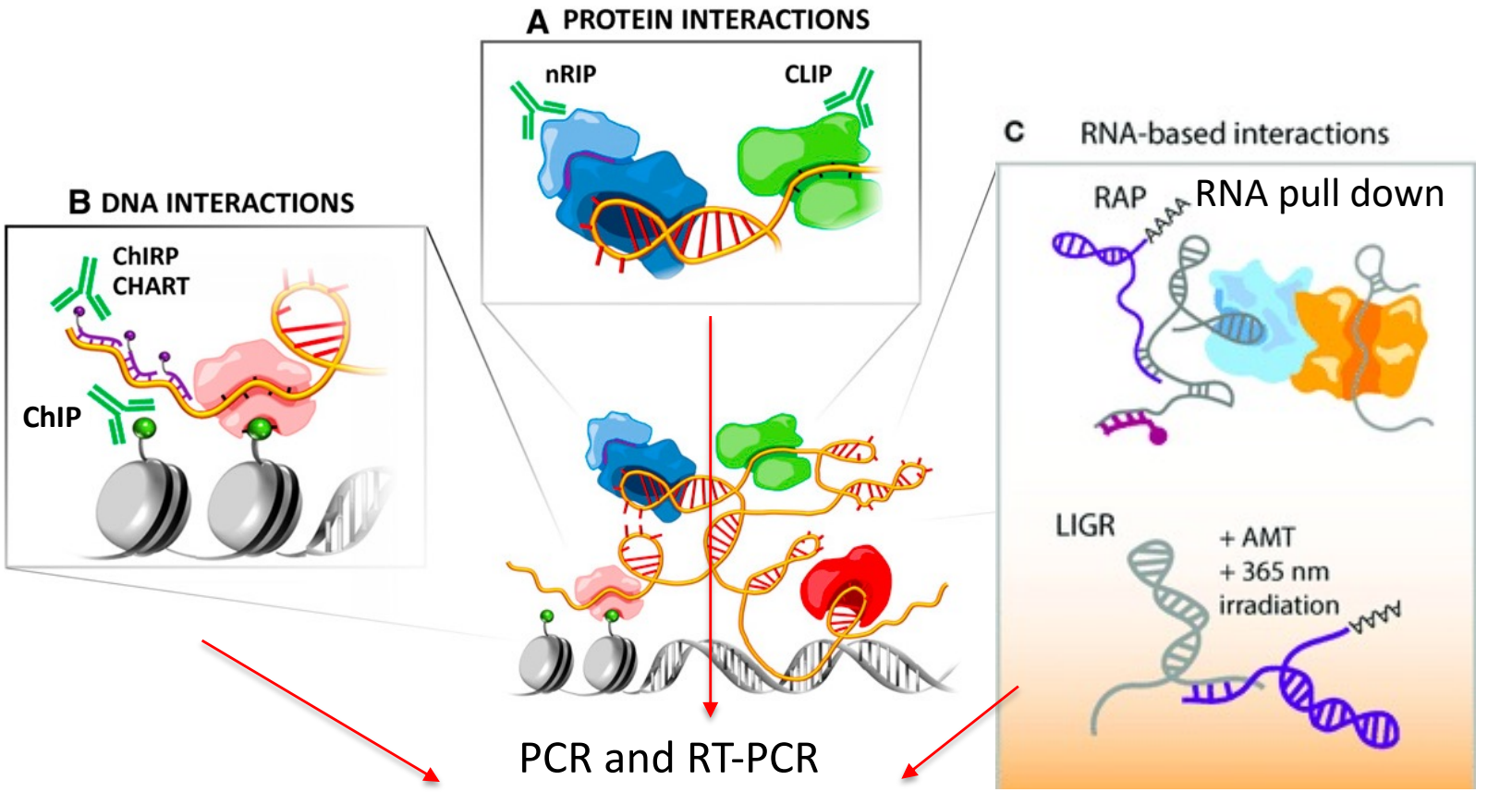
Next generation sequencing





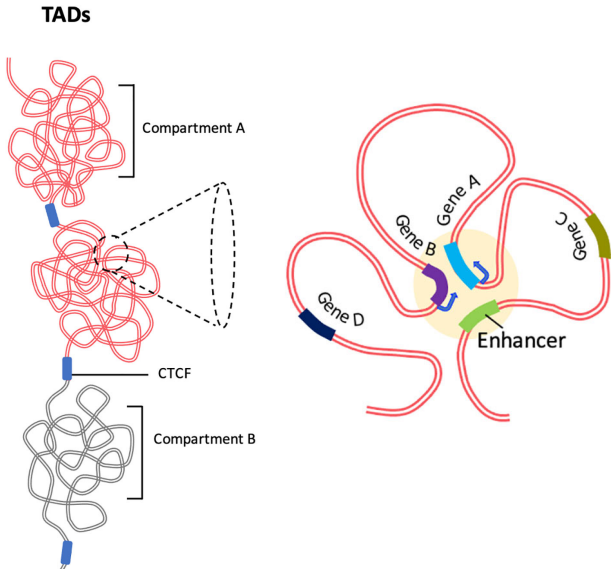




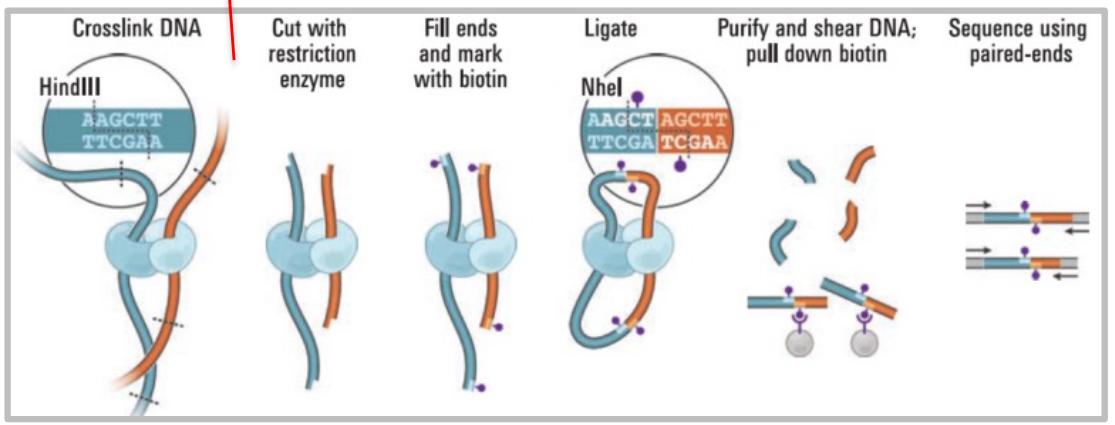


PCR and RT-PCR

DNA or RNA Sequencing



D. 3C or HiC



Catalyzing discoveries in gene regulation

1953 Resolution of DNA structure (Nobel price Watson/Crick)

Implications regarding the mechanisms of DNA replication and gene expression

1973 DNA cloning and DNA sequencing (Nobel price Gilbert/Sanger)

Definition of gene structure
- molecular definition of several pathologies

1987 PCR (Nobel price Mullis)

Huge improvement in DNA and gene expression analysis

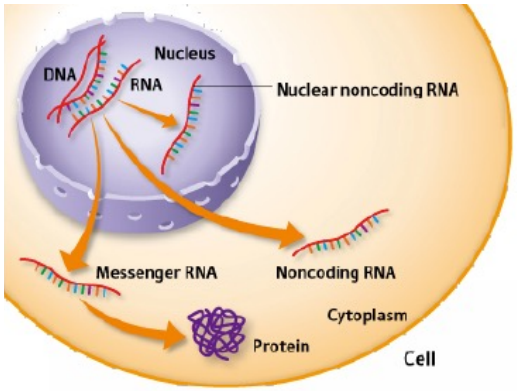
1990-2001 Genome sequencing

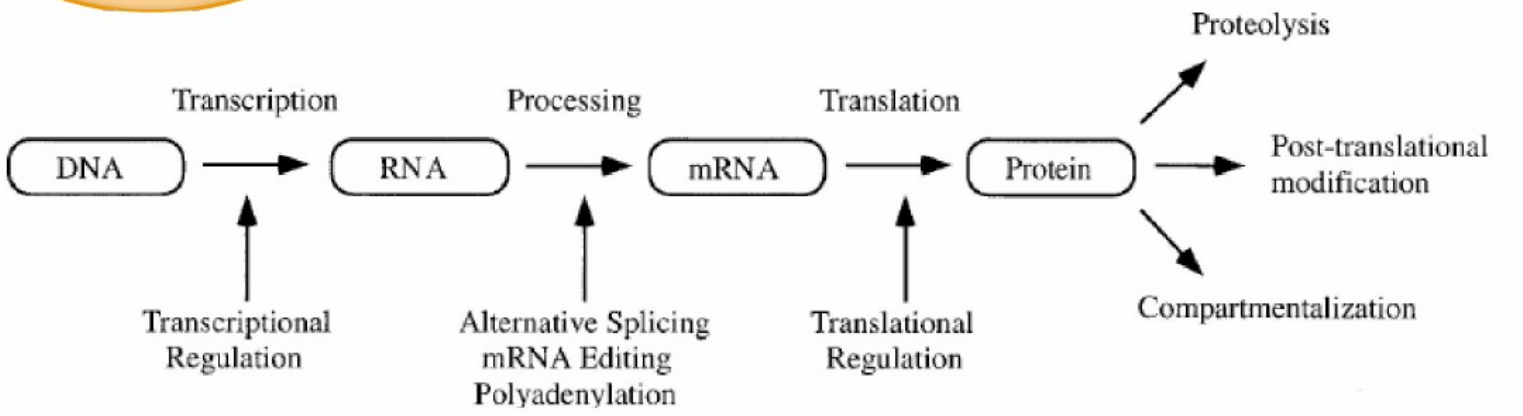
Identification of complex functions and analysis of multifactorial diseases.

2008-2009 RNA-seq with Next generation sequencing

Provides an accurate and comprehensive view of the transcriptome for any species

STEPS THAT ARE ANALYZED BY -OMICS





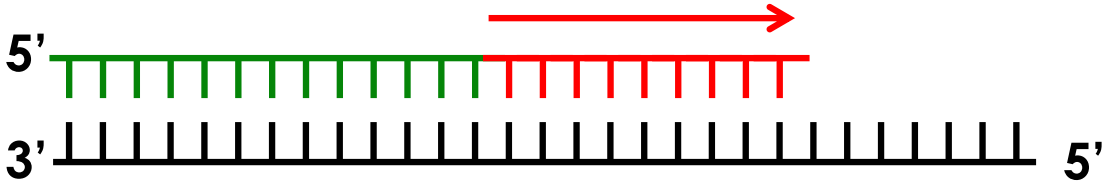
Genome Transcriptome Proteome

RNA-Seq

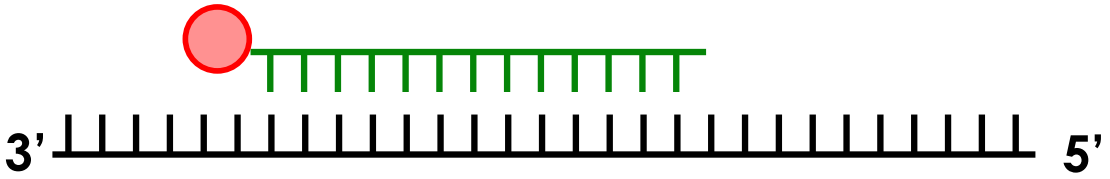
TRANSCRIPTOME ANALYSIS: HOW?

How to detect something that is unknown?

PCR / qPCR /
classic sequencing

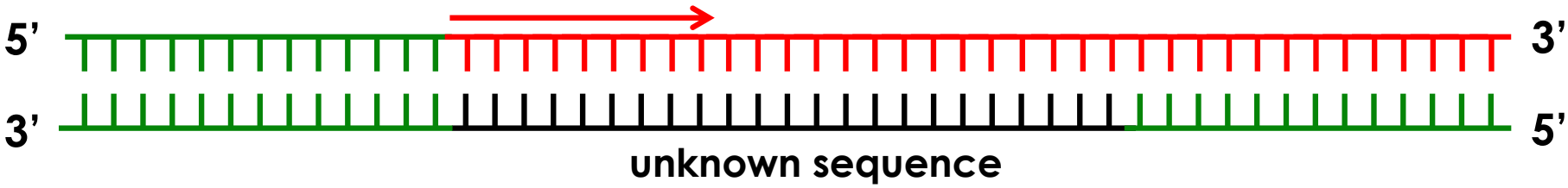


Northern blot /
Southern blot



We need to make detectable something that is not known

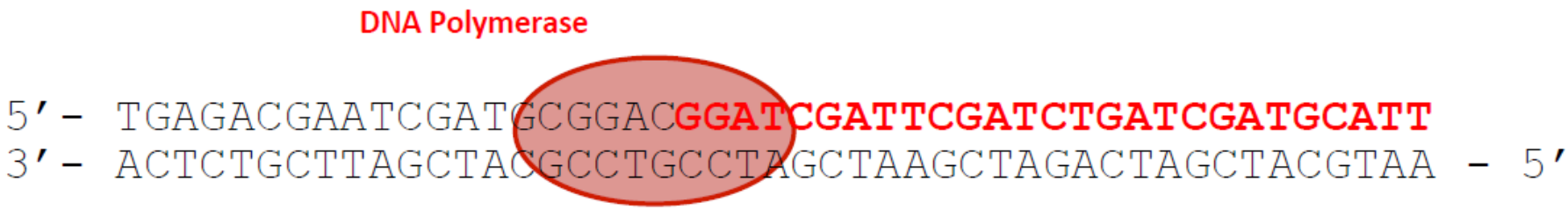
Next-Generation Sequencing (NGS)



SANGER METHOD FOR DNA SEQUENCING

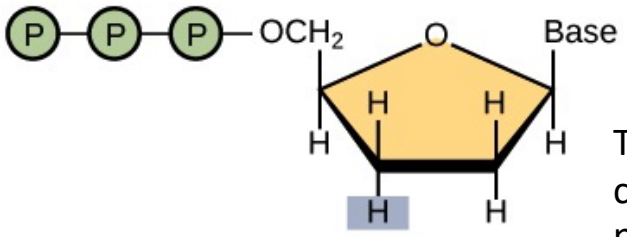
History of Sequencing: Sanger method for DNA sequencing

DNA Polymerase can add free nucleotides only to the 3' end of the newly forming strand. This results in elongation of the newly forming strand in a 5'-3' direction. No known DNA polymerase is able to begin a new chain (de novo). DNA polymerase can add a nucleotide only on to a pre-existing 3'-OH group, and, therefore, needs a primer at which it can add the first nucleotide.



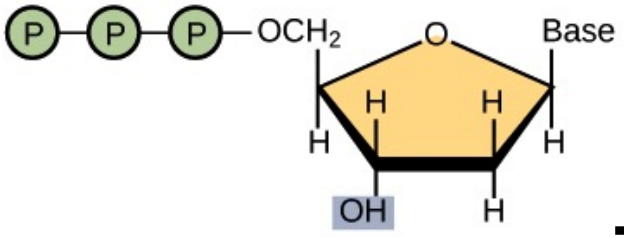
SANGER METHOD FOR DNA SEQUENCING

- "**Sanger Sequencing**" developed by Fred Sanger *et al.* in the mid 1970's
- Uses dideoxynucleotides for "chain termination", generating fragments of different lengths ending in ddATP, ddGTP, ddCTP or ddTTP



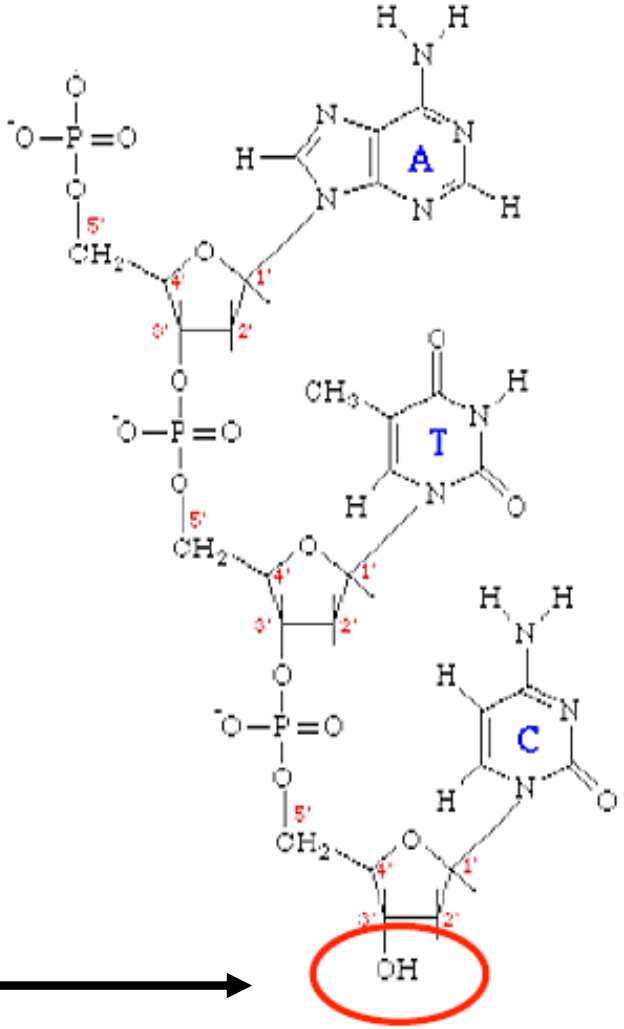
Dideoxynucleotide (ddNTP)

The dideoxynucleotide cannot form the phosphodiester bond with the next nucleotide

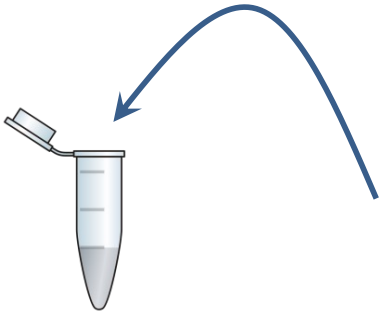


Deoxynucleotide (dNTP)

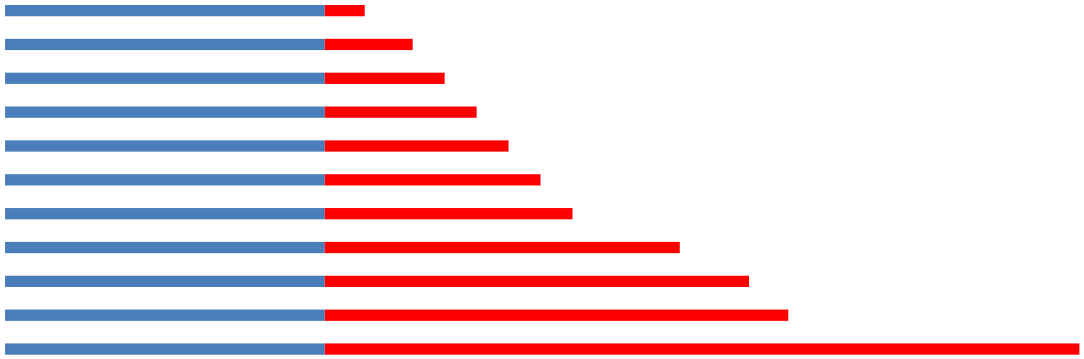
L'OH al 3' è richiesto per formare il legame fosfodiesterico con il nucleotide successivo



SANGER METHOD FOR DNA SEQUENCING

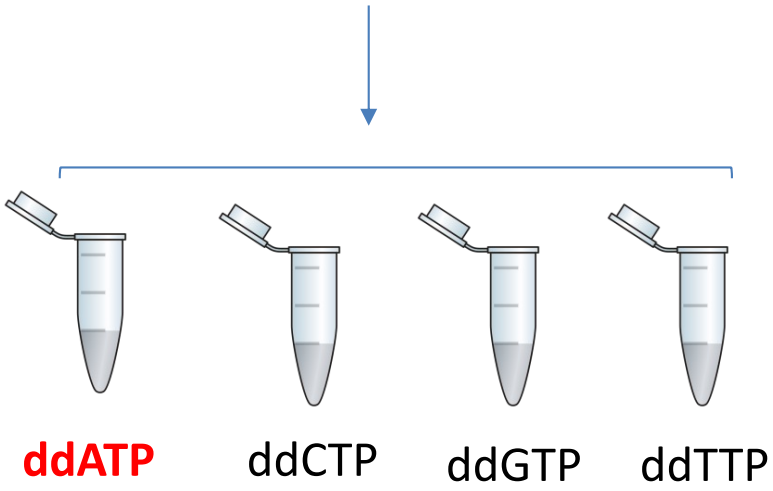


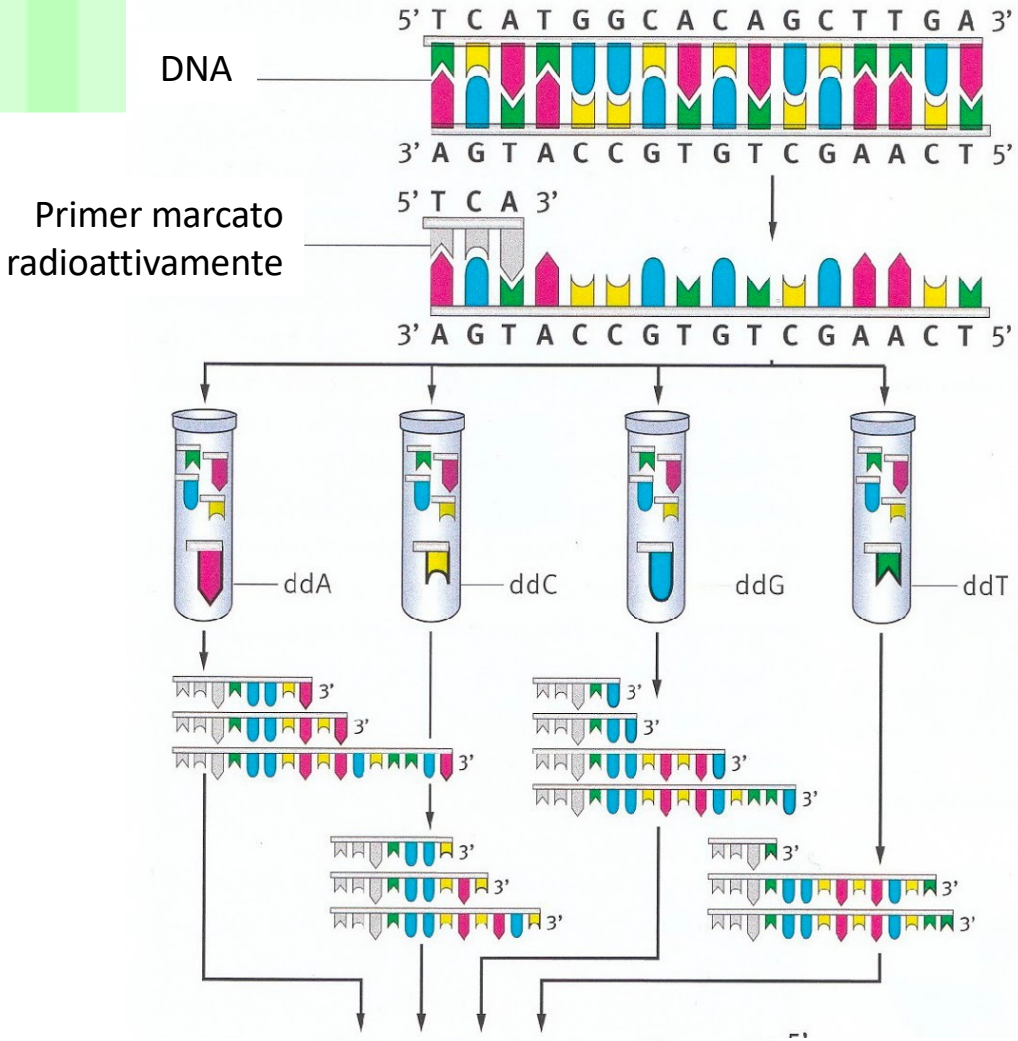
- Template DNA
- DNA Polymerase
- Primer
- dATP, dCTP, dGTP, dTTP
- **ddATP** (or ddCTP, ddGTP, ddTTP)



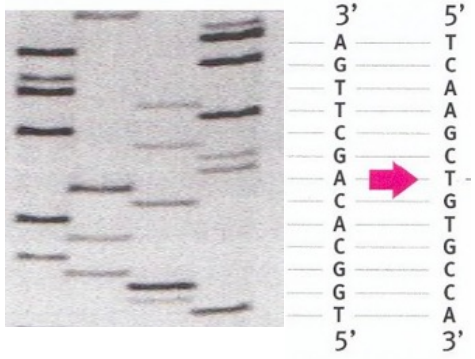
SANGER METHOD FOR DNA SEQUENCING

- Template DNA
- DNA Polymerase
- Primer
- dATP, dCTP, dGTP, dTTP





Elettroforesi su gel di acrilamminde

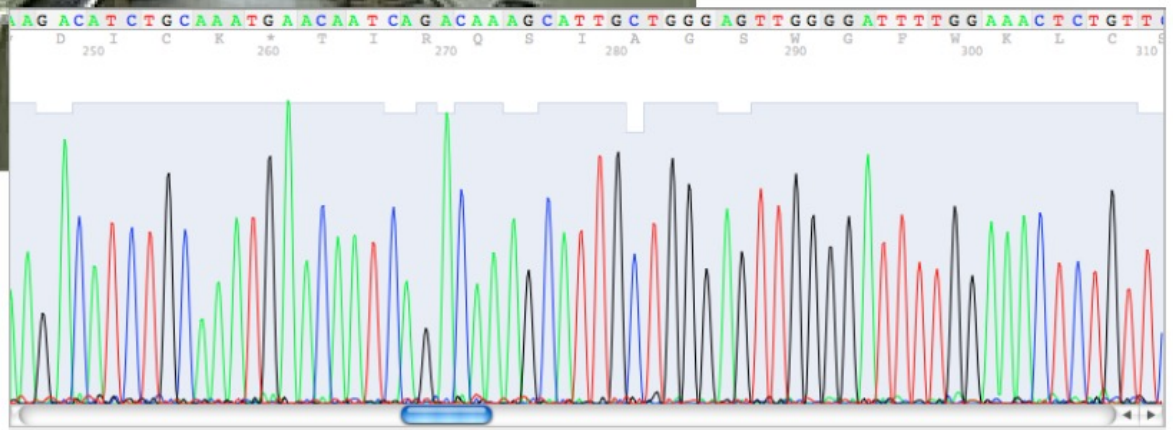


SANGER METHOD FOR DNA SEQUENCING

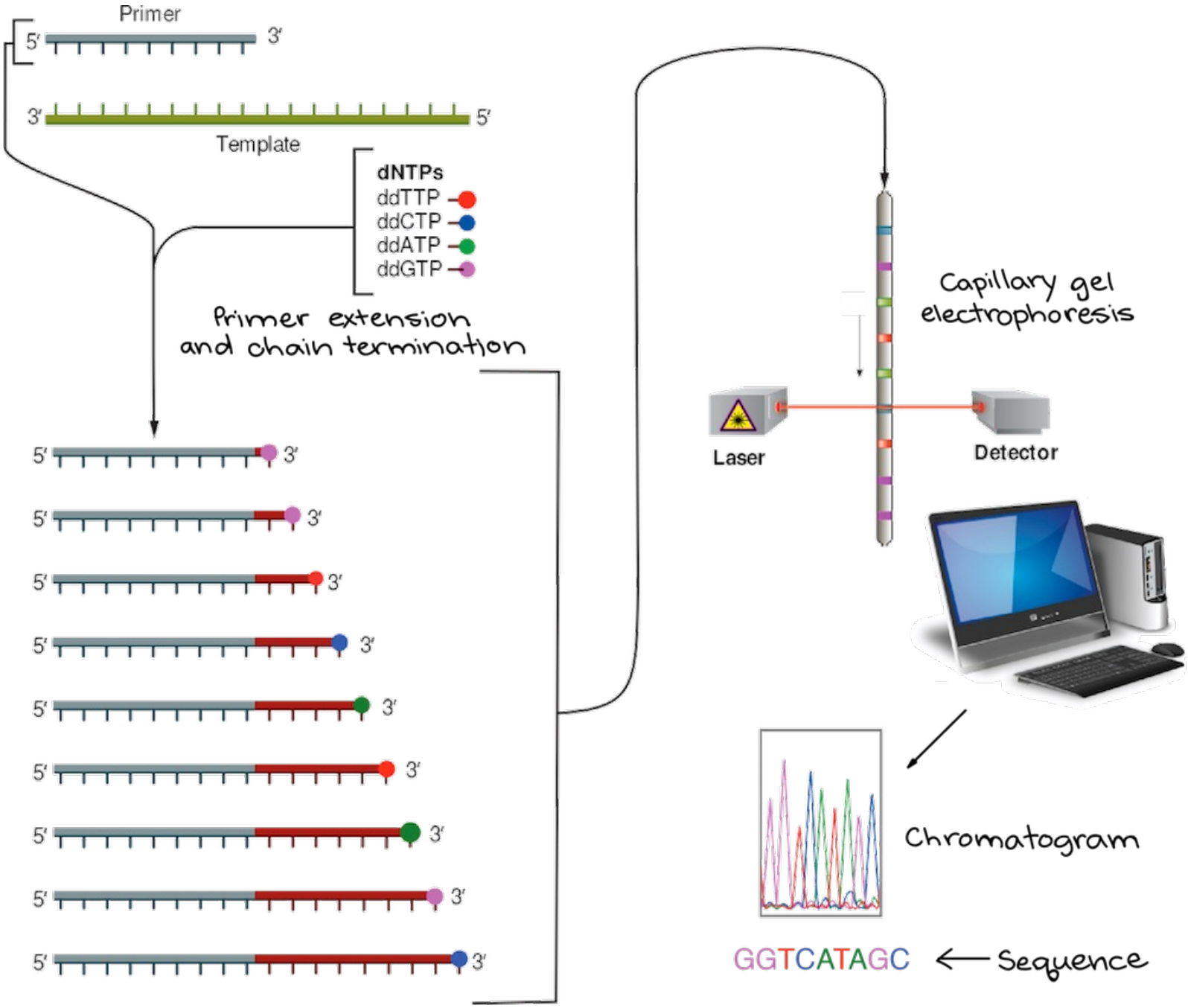
SANGER METHOD FOR DNA SEQUENCING

Automated Sequencing

- Sequencing technology was improved in the late 1980s by Leroy Hood who developed fluorescent color labels for the 4 terminator nucleotide bases.
- This allowed all 4 bases to be sequenced in a single reaction and sorted in a single gel lane



SANGER METHOD FOR DNA SEQUENCING



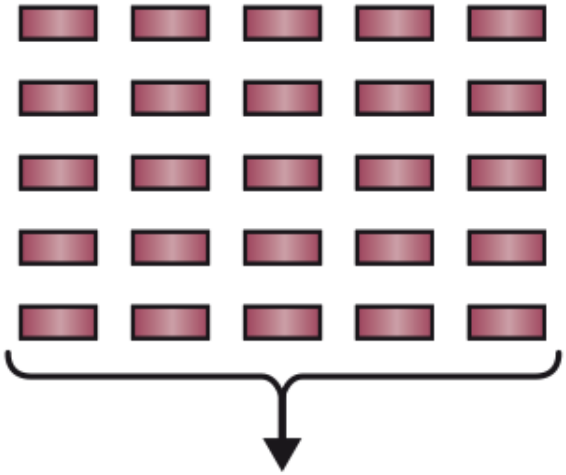
Chain-Termination Sequencing

Cloned insert or PCR product

TACGGATGGCATAGA

A single DNA sequence is generated

Second-Generation Sequencing



Library of DNA fragments

ACGTATCATGCGGATGG
TAGCATGACGTAGCGTT
GTAGCAGGTACGATGCC
GTAGACGATGCAGCATC
TAGGACCTAGCCGGACA

Many fragments are sequenced

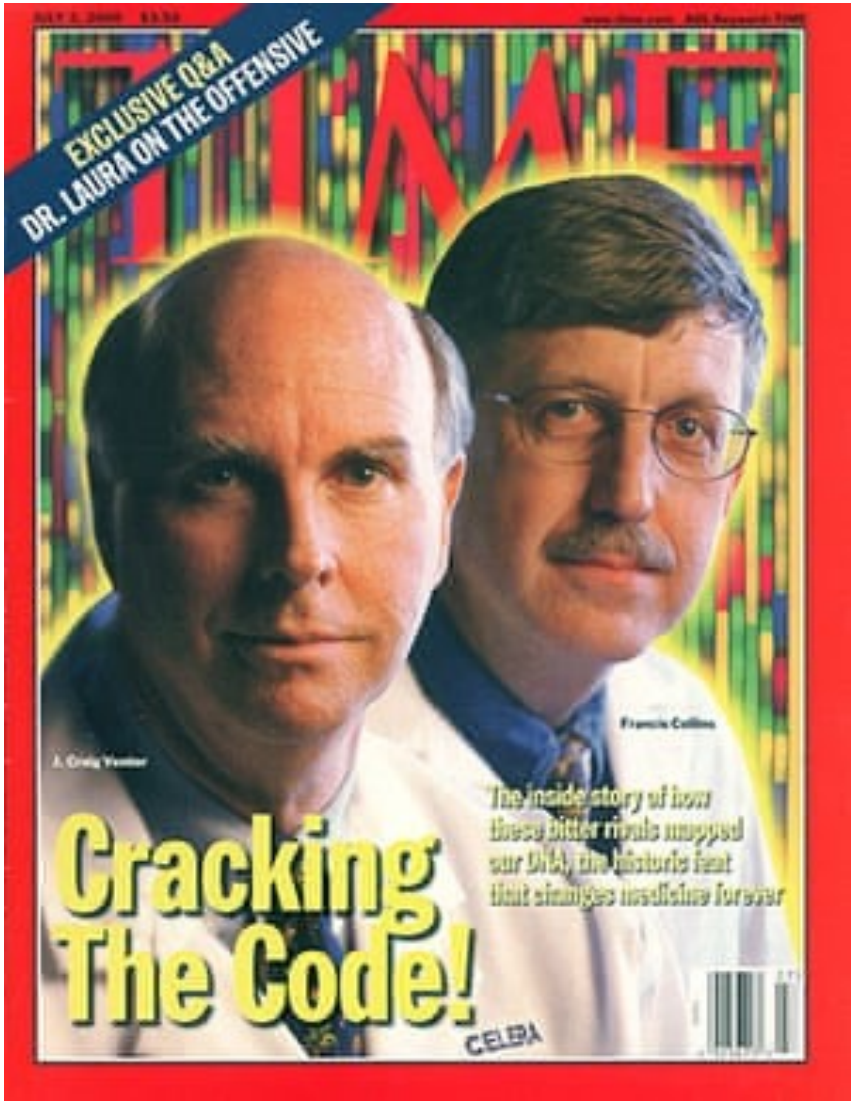
HUMAN GENOME PROJECT

Craig Venter

Celera Genomics

- Private company
- start in 1998
- 300 Milion $

No public access to data



Francis Collins

International Consortium

- 20 groups from USA, UK, China, Japan, Germany and France
- more than 1000 scientists
- start in 1990
- 2.7 billion $

Public access to data

Strategies

- Hierarchical shotgun approach
  - International Human Genome Sequencing Consortium (IHGSC)
- Whole-genome shotgun approach
  - Celera Genomics

HUMAN GENOME PROJECT: SHOTGUN SEQUENCING

•Sequencing technology allows for obtaining a sequence of about 800 bp at a time.

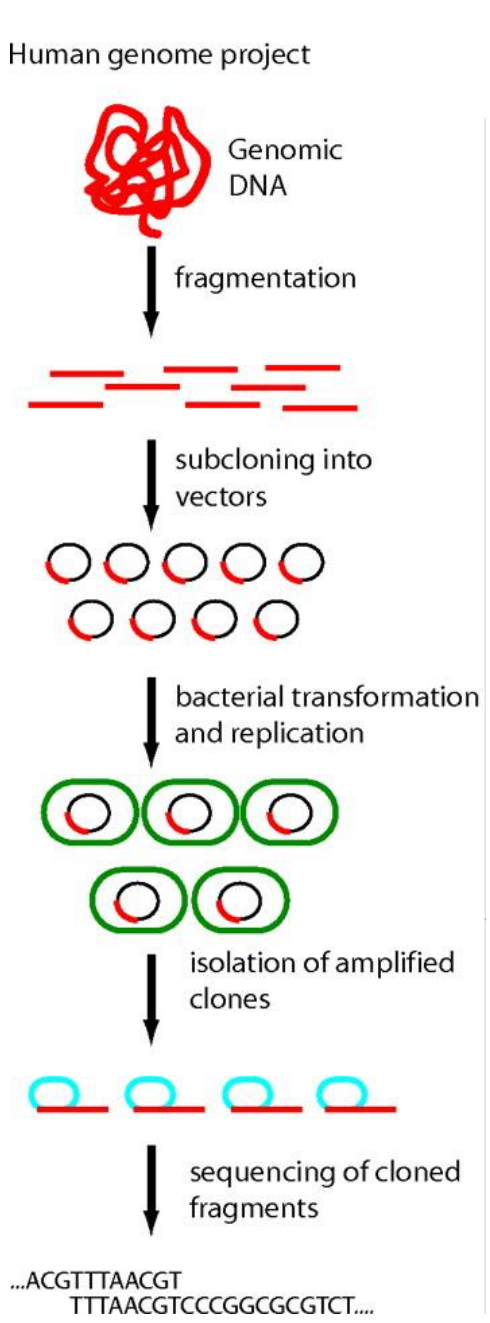
•Genomic DNA must be fragmented into small pieces for sequencing and then reassembled like a giant puzzle.

•Fragments of 150–350 kb are inserted into bacterial artificial chromosomes (BACs), which are then transformed into bacterial cells and replicated.

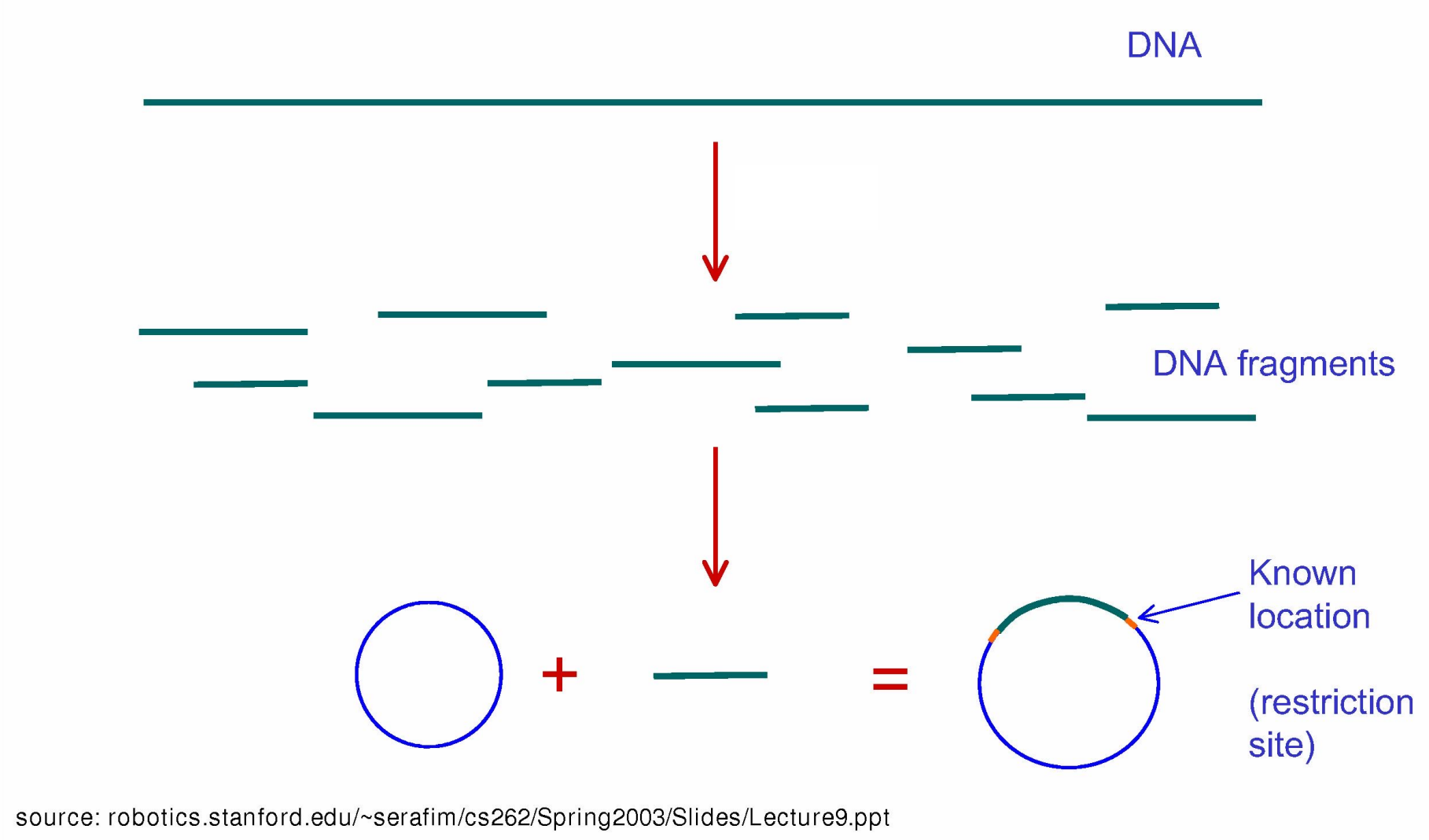
•The clones are fragmented into subclones of smaller sizes (4,000–6,000 bp) and reinserted into bacteria for amplification.

•DNA is extracted from the colonies.

•Sequenced using the Sanger method

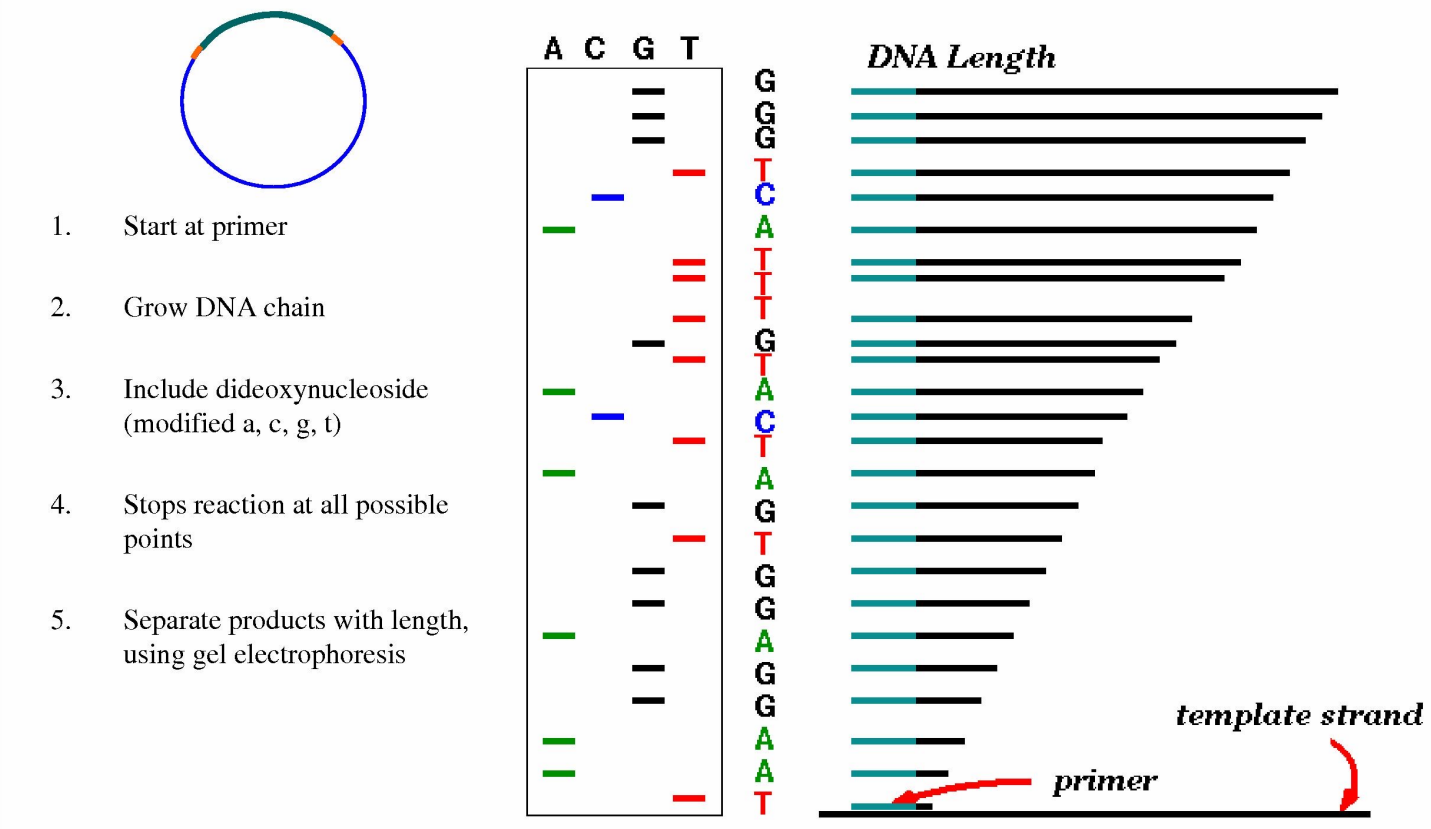


HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



source: robotics.stanford.edu/~serafim/cs262/Spring2003/Slides/Lecture9.ppt

HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



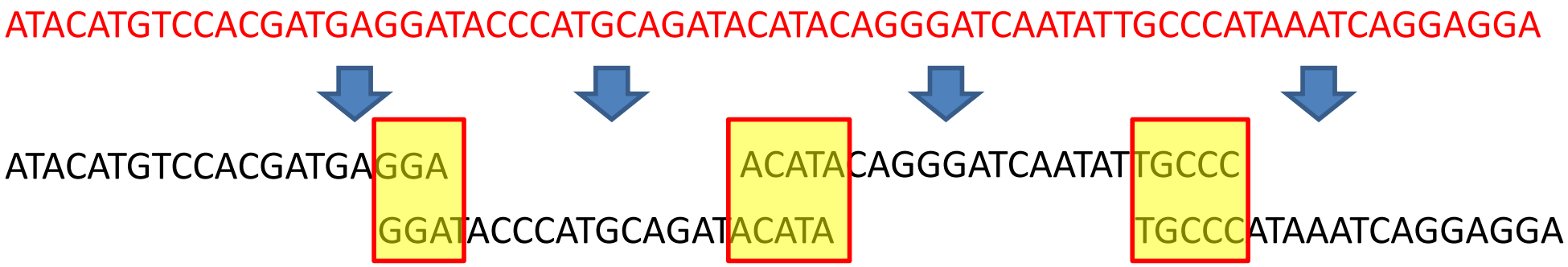
- Can produce DNA fragments 700-900bp long, but it's slow
- Lots of other problems including clone library generation and low-throughput
- The Human Genome Project used Sanger sequencing, completion took over 10 years

HUMAN GENOME PROJECT: SHOTGUN SEQUENCING

The principle is to obtain a series of overlapping DNA fragments that can be connected into a continuous map.

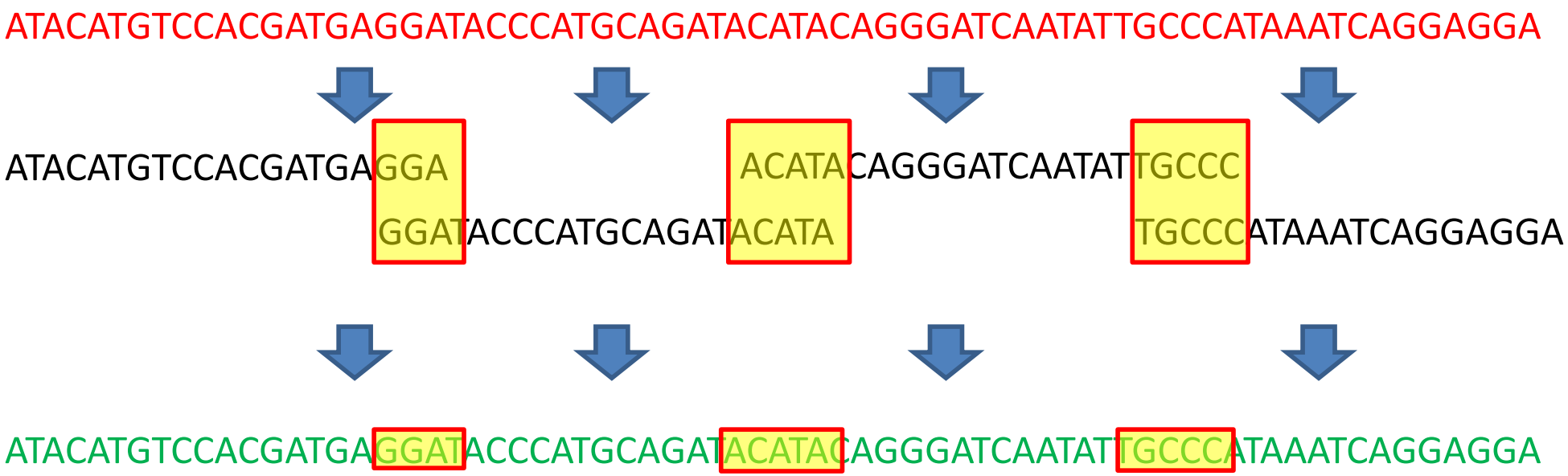
HUMAN GENOME PROJECT: SHOTGUN SEQUENCING

The principle is to obtain a series of overlapping DNA fragments that can be connected into a continuous map.

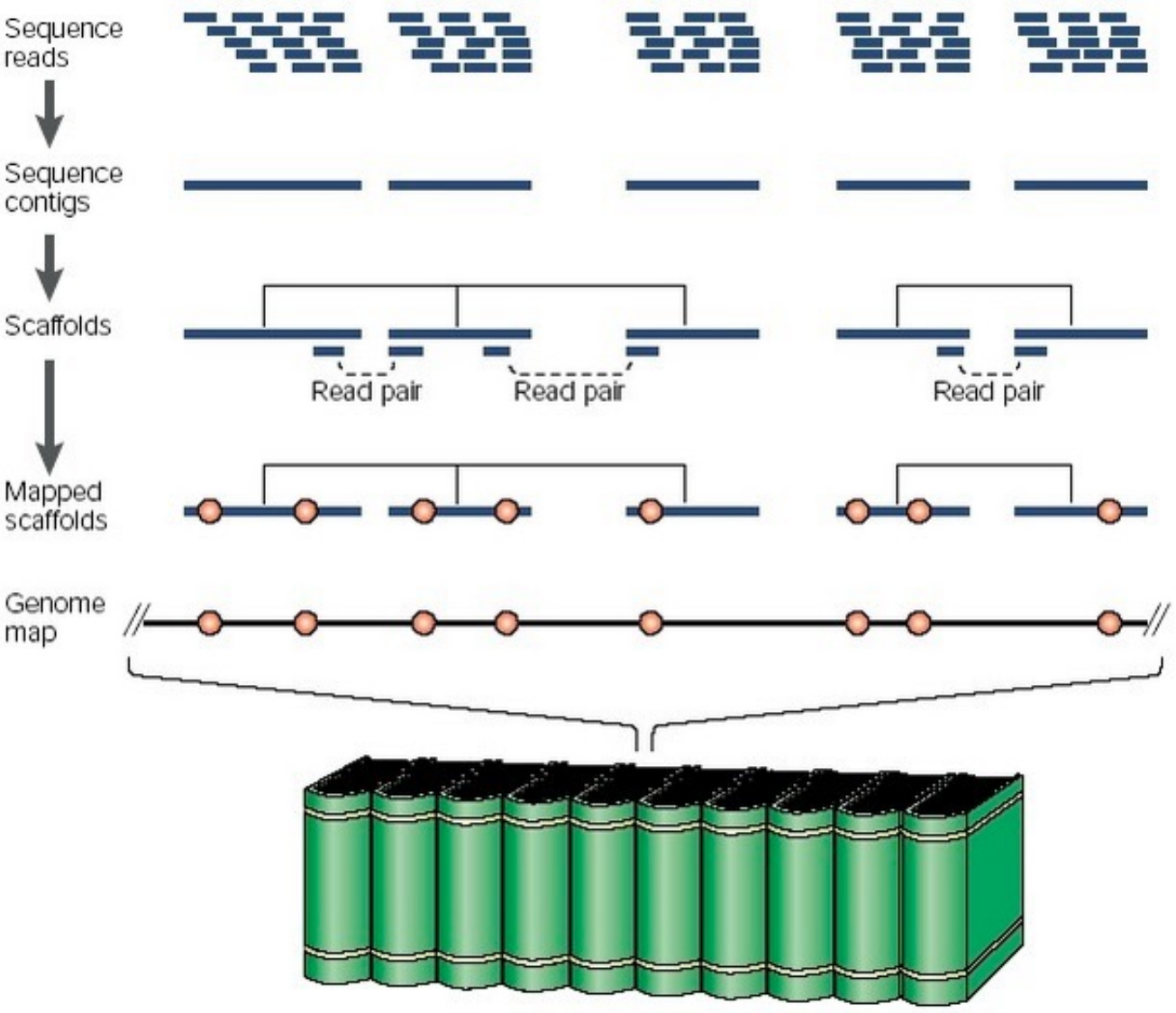


HUMAN GENOME PROJECT: SHOTGUN SEQUENCING

The principle is to obtain a series of overlapping DNA fragments that can be connected into a continuous map.

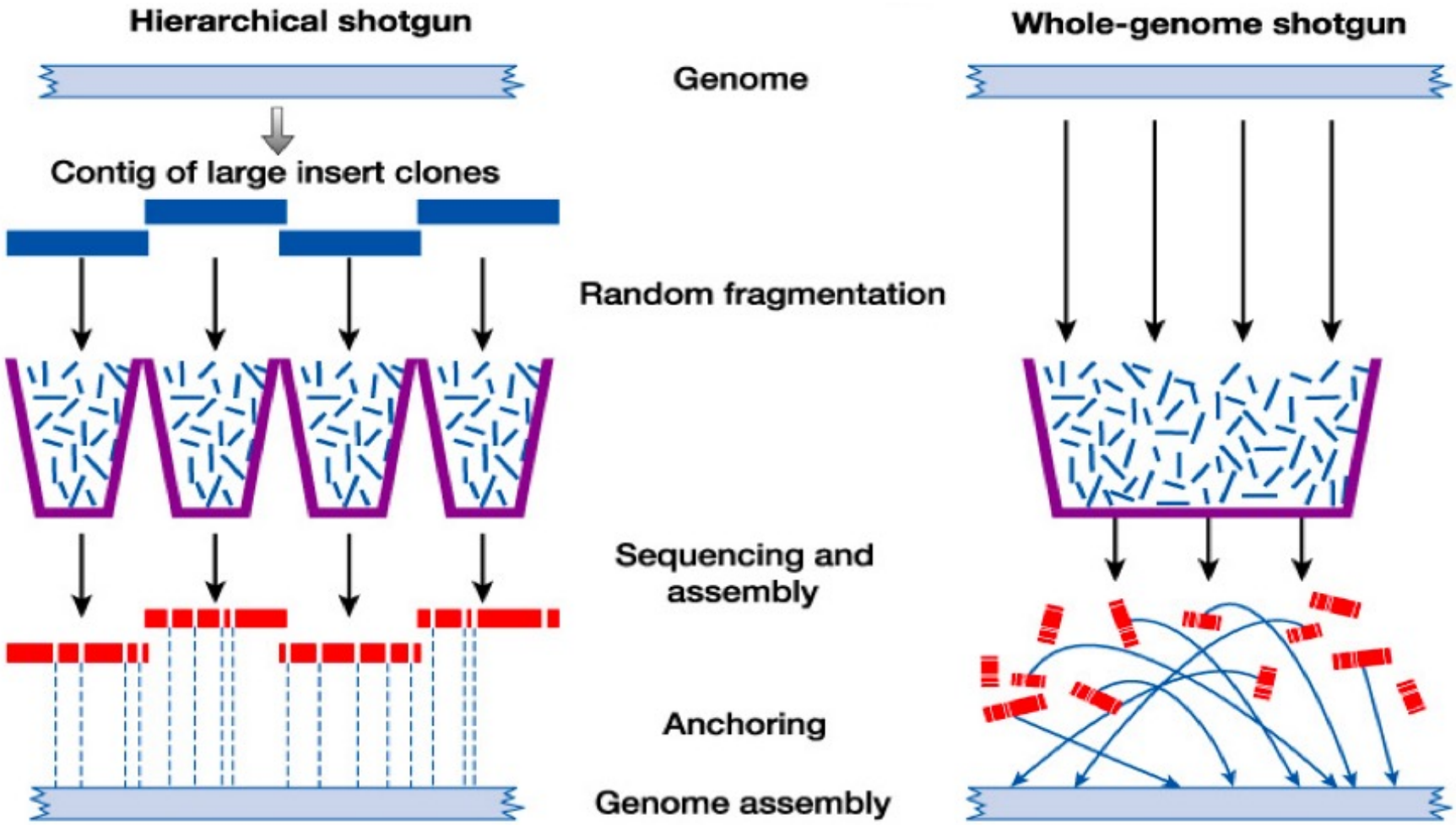


HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



- **"paired ends" sequencing**
- Sequence contigs from computational homology search
- "Scaffolds" use information from paired-end sequencing (not clone maps)
- More suitable for small genomes and/or those with few repetitive elements.

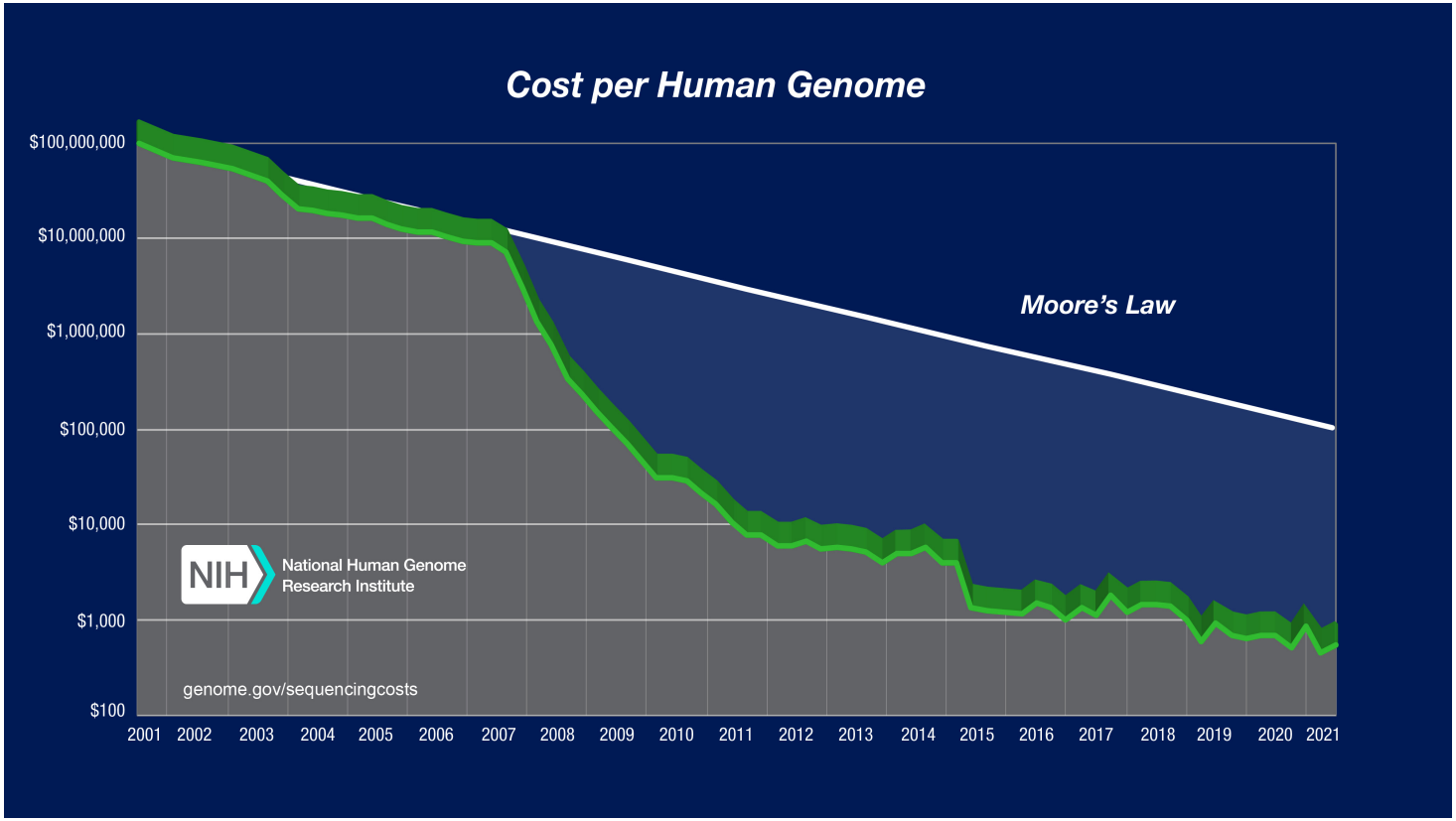
HUMAN GENOME PROJECT



The whole-genome shotgun approach simplifies and speeds up the preparation of a genomic library, making it more cost-effective. However, it requires more intensive computational processing. This has become feasible due to advancements in bioinformatic techniques and increased computational power.

SEQUENCING A HUMAN GENOME (3,2 BILLION BP)

300 million $



1000 $/genome

SEQUENCING A HUMAN GENOME (3,2 BILLION BP)

Costs and time for sequencing a human genome (3,2 billion bp)

2001	First human genome	13 years	300 million $
2005	Technology review	6 months	20-30 million $
2005	**454 Roche**	**1 month**	**900'000 $ (1X coverage)**
2009	Solexa (Illumina)	6 months	50'000 $ (30X coverage)
2010	Illumina		19'500 $ (30X coverage)
Today	**Personalized medicine**	Today 300$ (30x coverage)	

https://www.longdom.org/open-access/generations-of-sequencing-technologies-from-first-to-next-generation-0974-8369-1000395.pdf

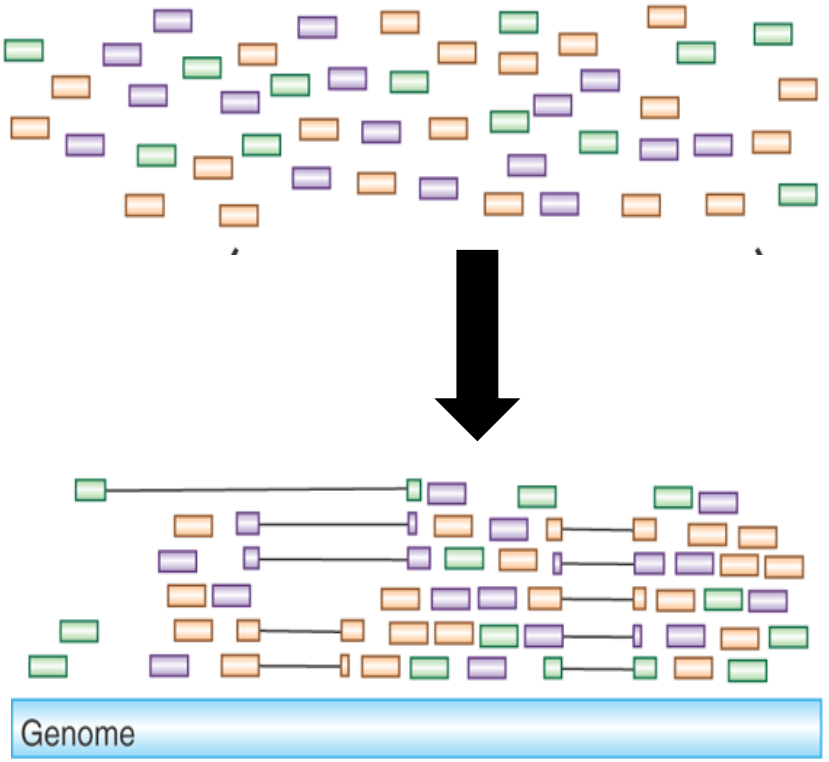
NEXT GENERATION SEQUENCING

What is it?

Set of new high throughput technologies:

- Allow millions of short DNA sequences from a biological sample to be "read" or sequenced in a rapid manner

- Computational power is then used to assemble or align the "reads" to a reference genome, allowing biologists to make comparisons and interpret various biological phenomena



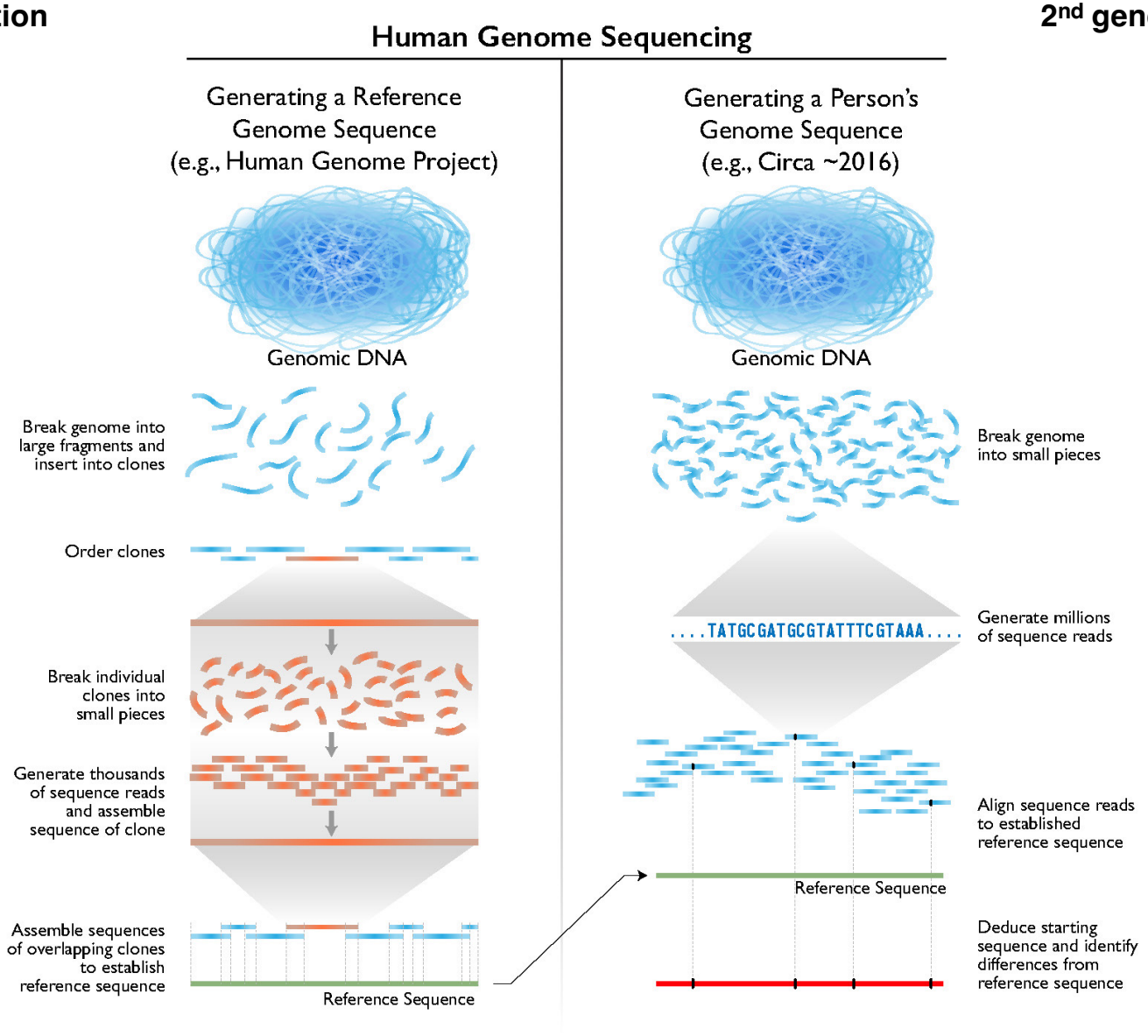
- Due to high depth of coverage (30-100x), accurate sequencing is obtained much faster and cheaper compared to traditional Sanger/Shotgun sequencing

Whitehead Institute, Center for Genome Research, Cambridge, MA

Nanopore sequencing

Sequencing methodologies

1st generation | **2nd generation (NGS)**



NEXT GENERATION SEQUENCING

What is it?

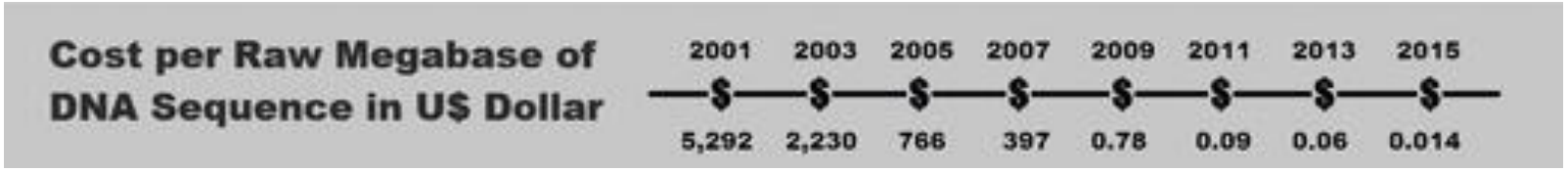
- **Set** of new high throughput technologies:

  • Allow millions of short DNA sequences from a biological sample to be "read" or sequenced in a rapid manner

  • Computational power is then used to assemble or align the "reads" to a reference genome, allowing biologists to make comparisons and interpret various biological phenomena

  • Due to high depth of coverage (30-100x), accurate sequencing is obtained much faster compared to traditional Sanger sequencing

SEQUENCING TECHNOLOGIES

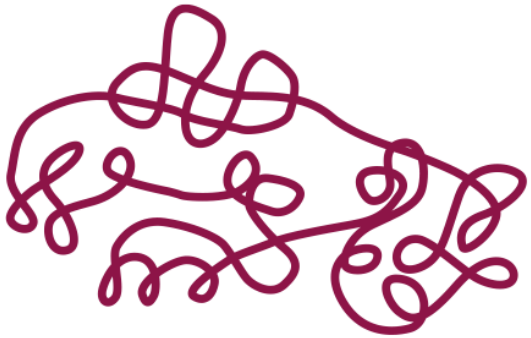


NEXT GENERATION SEQUENCING

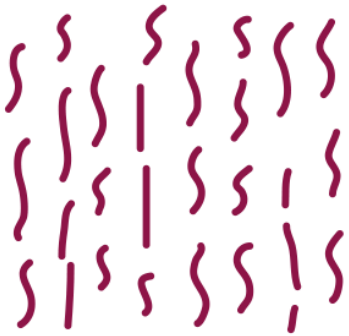
Just DNA sequencing or something more…

- **Mutation and SNP** identification or analysis (genome re-sequencing)
- Gene/Disease Linkage (genome re-sequencing)
- Pathogen identification (de novo sequence assembly or re-sequencing)
- DNA methylation study (medip-seq)
- Chromatin study (**ChIPseq**)
- Transcription factor study (ChIPseq)
- Genome structure (HiC)
- Transcriptome analysis (**RNAseq**)
- miRNAs, siRNA, piRNA, tRF, etc... (**small RNA seq**)
- Single cell transcriptome analysis

DNA sequencing



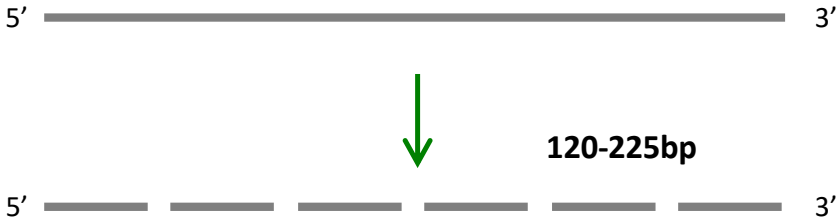
Sonication



Ligation of the the adaptors required for some sequencing platforms (i.e. Illumina)

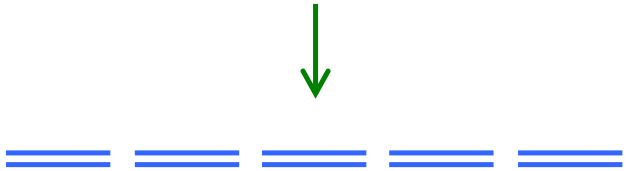
Sequenziamento dell'RNA

1.RNA fragmentation through the use of divalent cations and high temperatures

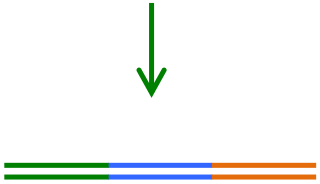


2. cDNA generation by retrotranscription

3. Generation of double strand DNA with random oligos



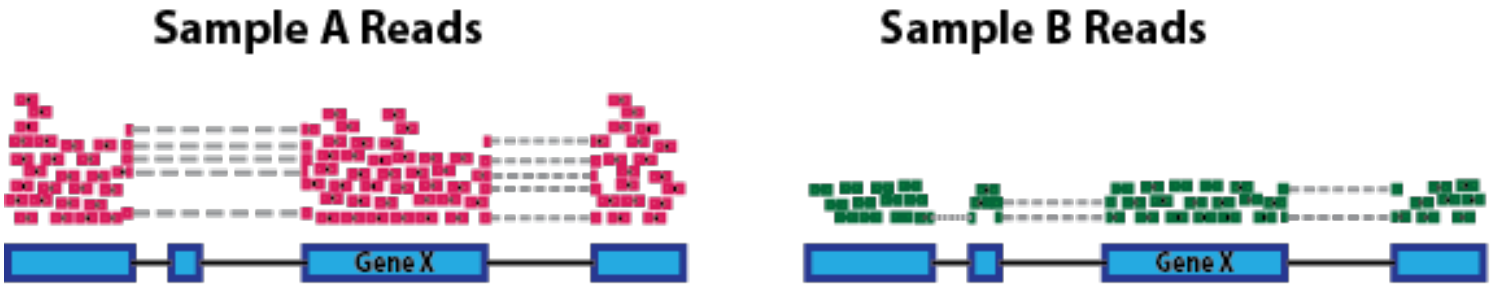
4 Ligation of the the adaptors required for some sequencing platforms (i.e. Illumina)



NEXT GENERATION SEQUENCING

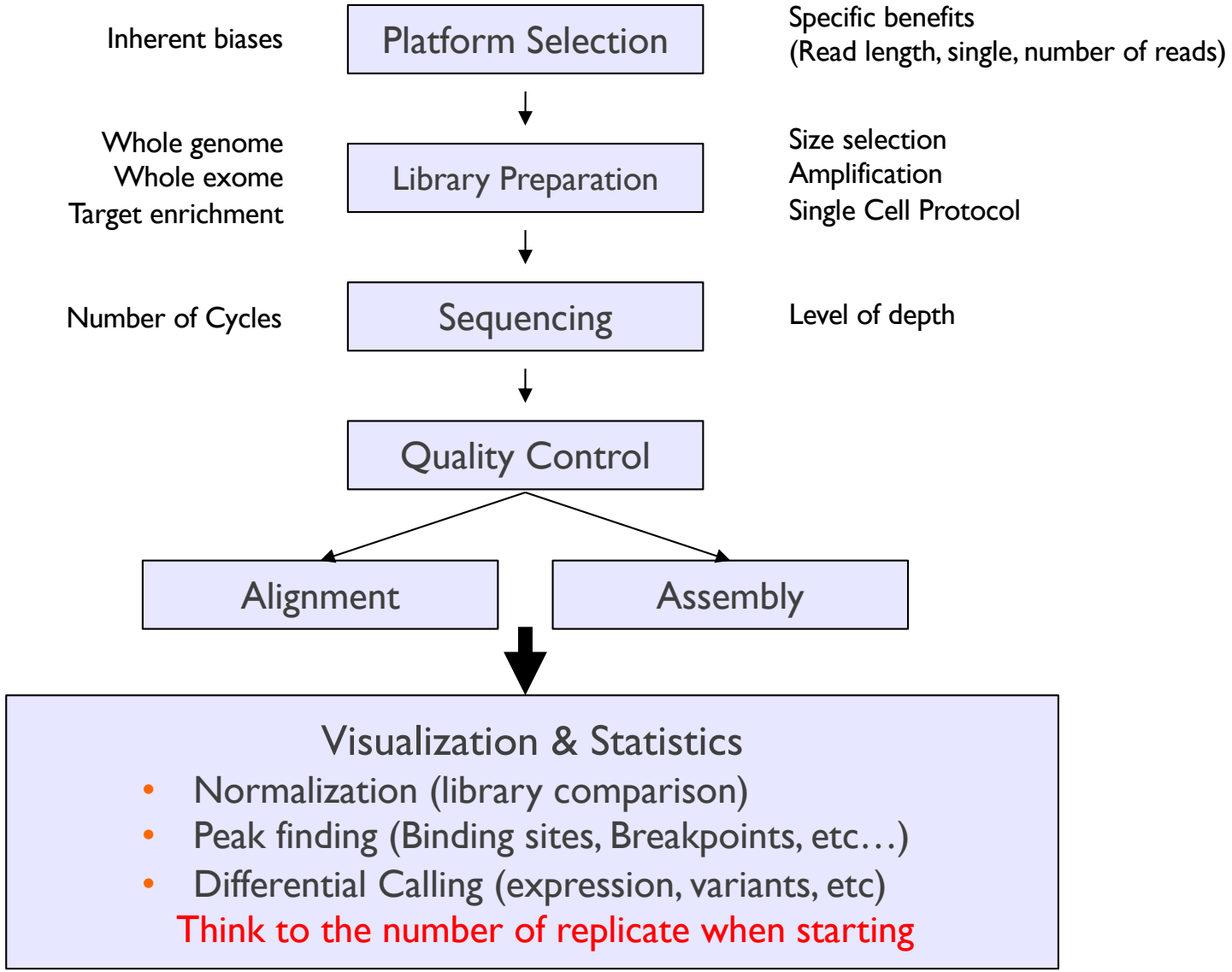
Deep sequencing → Qualitative information

Deep sequencing → Quantitative information



Example: RNA-Seq

NEXT GENERATION SEQUENCING



RNA-Seq

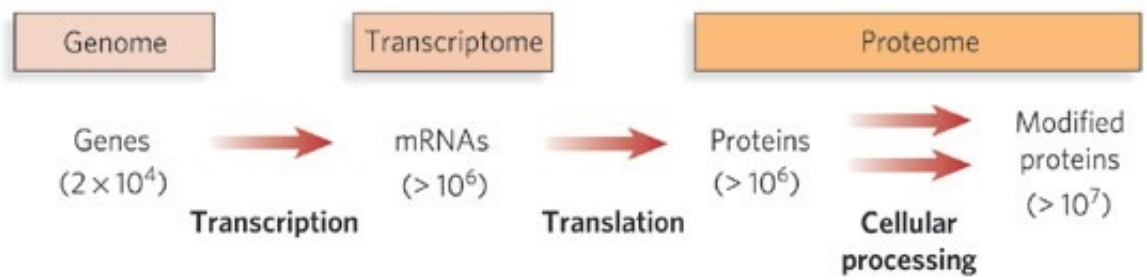
What is RNA-seq?

- RNA-seq is essentially **massively parallel sequencing of RNA** (or, in fact, the corresponding cDNA) and has heralded the second technical revolution in transcriptomics.
- It is **based on next-generation sequencing (NGS) platforms** that were initially developed for high-throughput sequencing of genomic DNA.
- Typically, **all the RNA molecules in a sample are reverse transcribed into cDNA,** and depending on the platform to be used, the **cDNA molecules may (amplification-based sequencing) or may not (single-molecule sequencing (SMS)) be amplified before deep sequencing.**
- After the sequencing reaction has taken place, **the obtained sequence stretches (reads) are mapped onto a reference genome** to deduce the structure and/or expression state of any given transcript in the sample.

From genomes to transcriptomes

- After genome sequencing, the second major branch of genomics is **analysis of the transcriptome**.

- The transcriptome is the complete set of transcripts and their relative levels of expression in particular cells or tissues under defined condition.



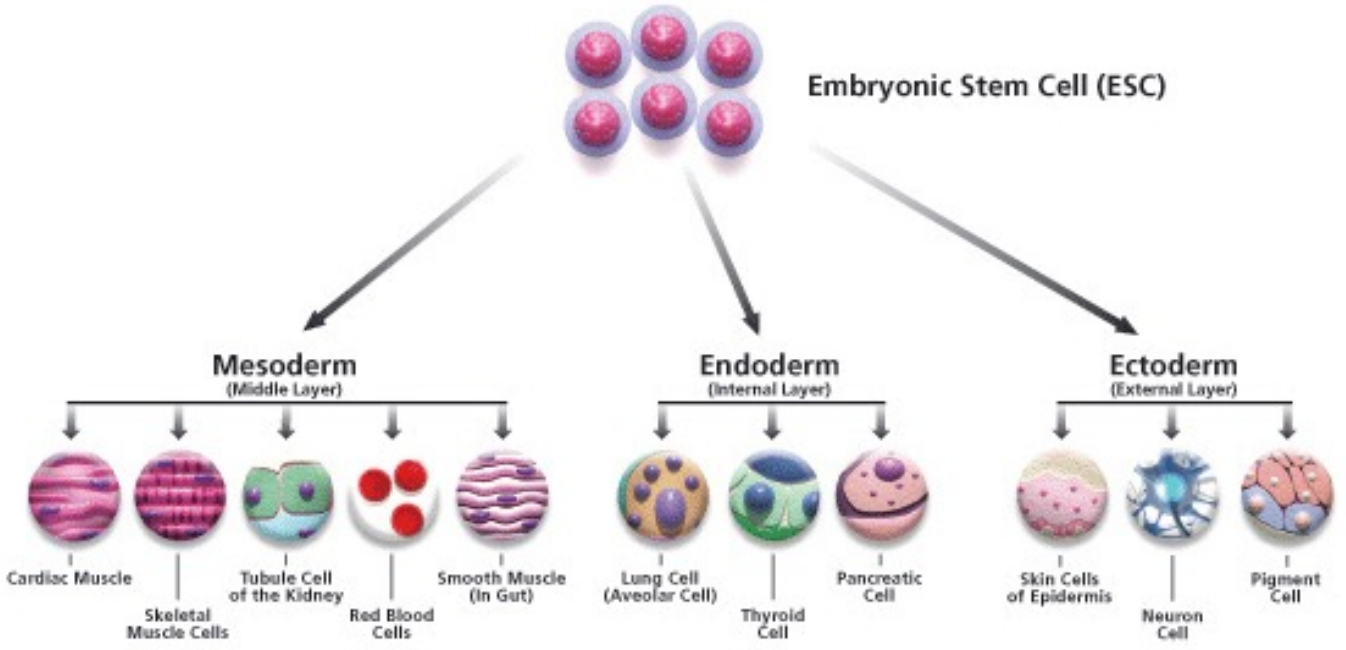
TRANSCRIPTOME ANALYSIS

Issues in the studies on Transcriptome

- The Transcriptome of a cell is a dynamic entity: unlike the Genome, it constantly changes.

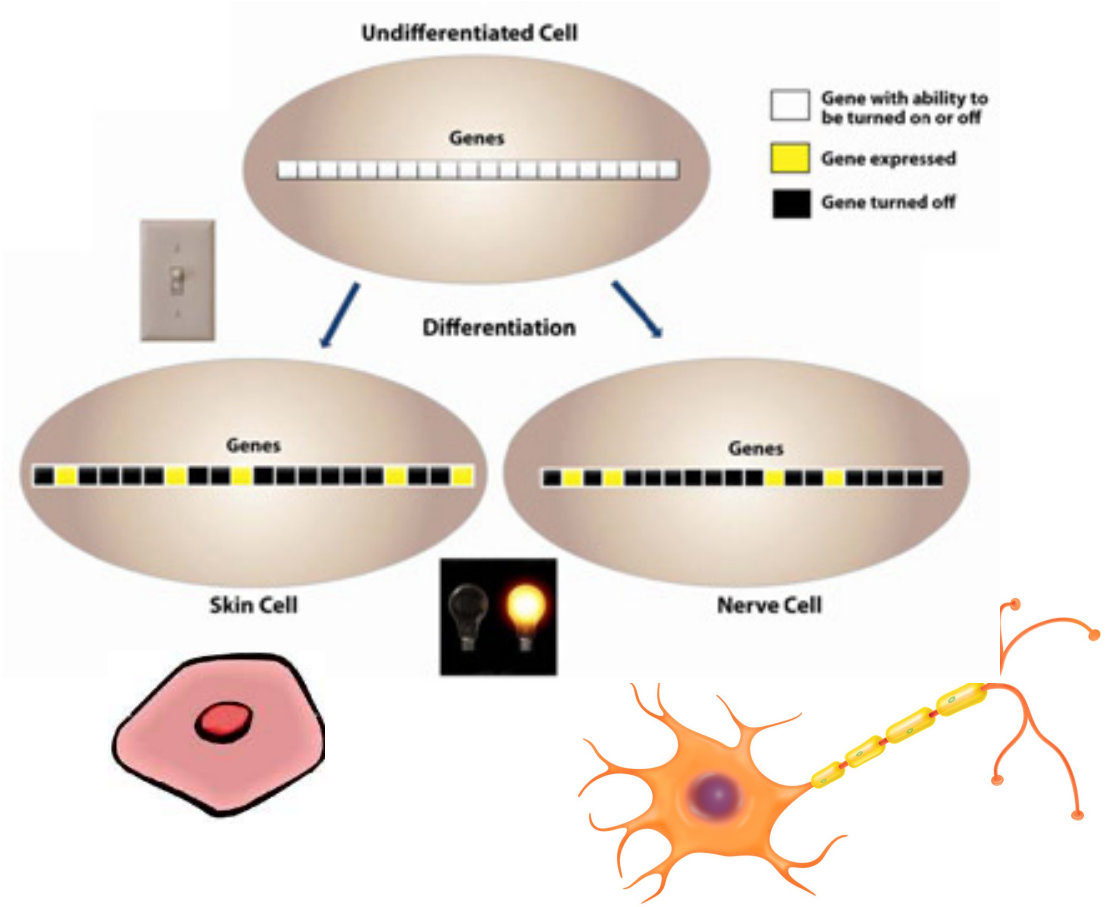
The transcriptome is cell type specific

In multicellular organisms, many genes are expressed in particular cell types or at certain developmental stages: same genome but different transcriptome.



Gene Expression Regulates Cell Differentiation

The particular combination of genes that are turned on (expressed) or turned off (repressed) dictates cellular morphology (shape) and function.



RNA- sequencing (RNA-Seq)

RNA-Seq provides the ability to look at:

- changes in gene expression
- alternatively spliced transcripts, alternative promoters and polyA sites
- post-transcriptional changes and RNA modifications
- gene fusions
- different populations of RNA (mRNAs, tRNAs, microRNAs, circRNAs etc.)

TRANSCRIPTOME ANALYSIS USING RNA-Seq

- **Traditional RNA-Seq**

  It allows to quantify the expression of:
  mRNAs and other polyadenilated RNAs (polyA+)
  all RNA species except for rRNAs (RiboMinus, Ribo-Zero).

- **Small RNA-Seq**

  Adapters are designed so that they can bind microRNA and other small RNAs which have a 3' hydroxyl group that is the result of the cleavage by Dicer or other RNA processing enzymes.

- **RIP-Seq and CLIP-Seq**

  All the RNAs which are bound by a protein are sequenced (using a standard protocol), so that they can be identified. CLIP-Seq also allows to find the localization of the binding site.

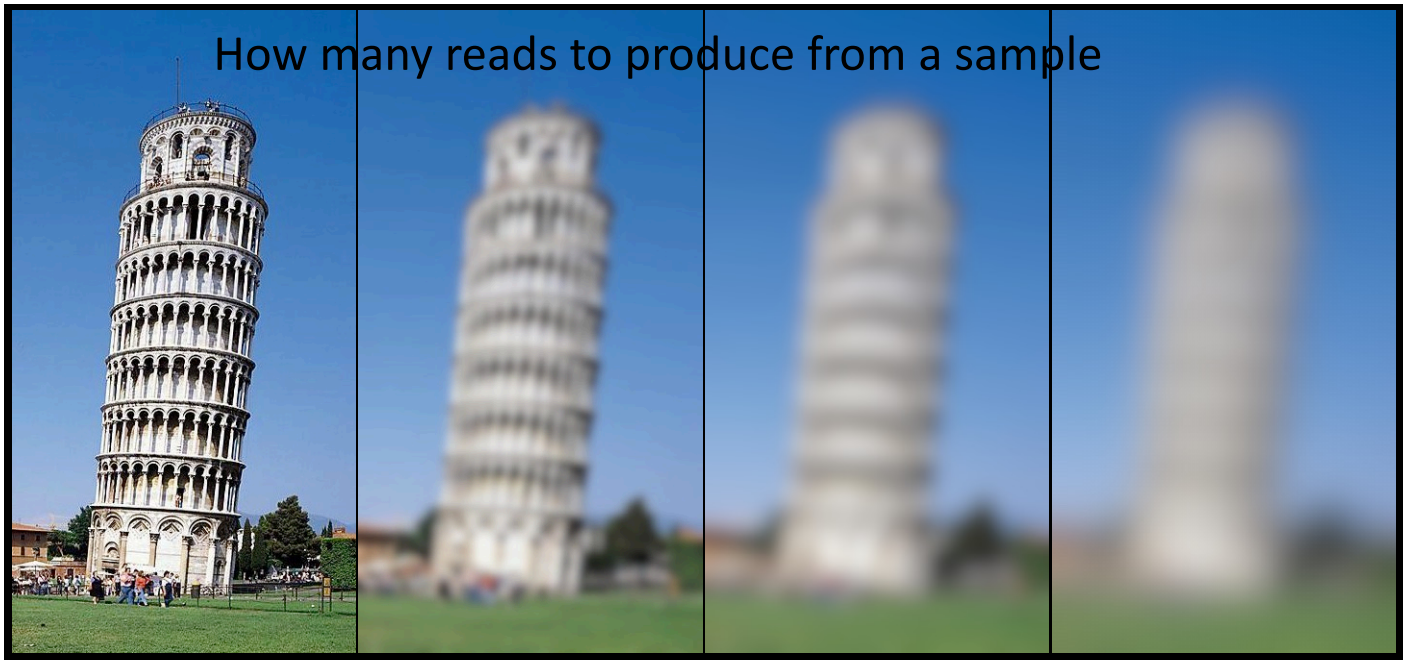
- **Nascent RNA-Seq**

  It allows the direct measurement of nascent RNA production.

NEXT GENERATION SEQUENCING

Sequencing Depth (Coverage)

Sequencing depth refers to the number of times a specific nucleotide base or region in a genome or transcriptome has been read during a sequencing experiment. Imagine a long piece of DNA is fragmented and sequenced. Each resulting short read is mapped back to a reference sequence. If a base in the reference sequence is covered by 30 unique reads, that base has a sequencing depth of 30X. Deeper sequencing is needed to accurately quantify low-abundance transcripts and reliably detect differential expression.

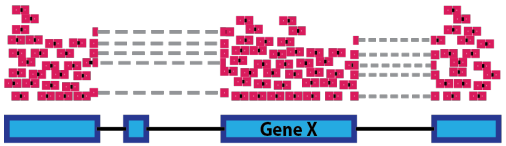


High resolution

↓

many information

many published human RNA-Seq experiments have been sequenced with a sequencing depth **between 20 M - 50 M reads per sample**



Low resolution

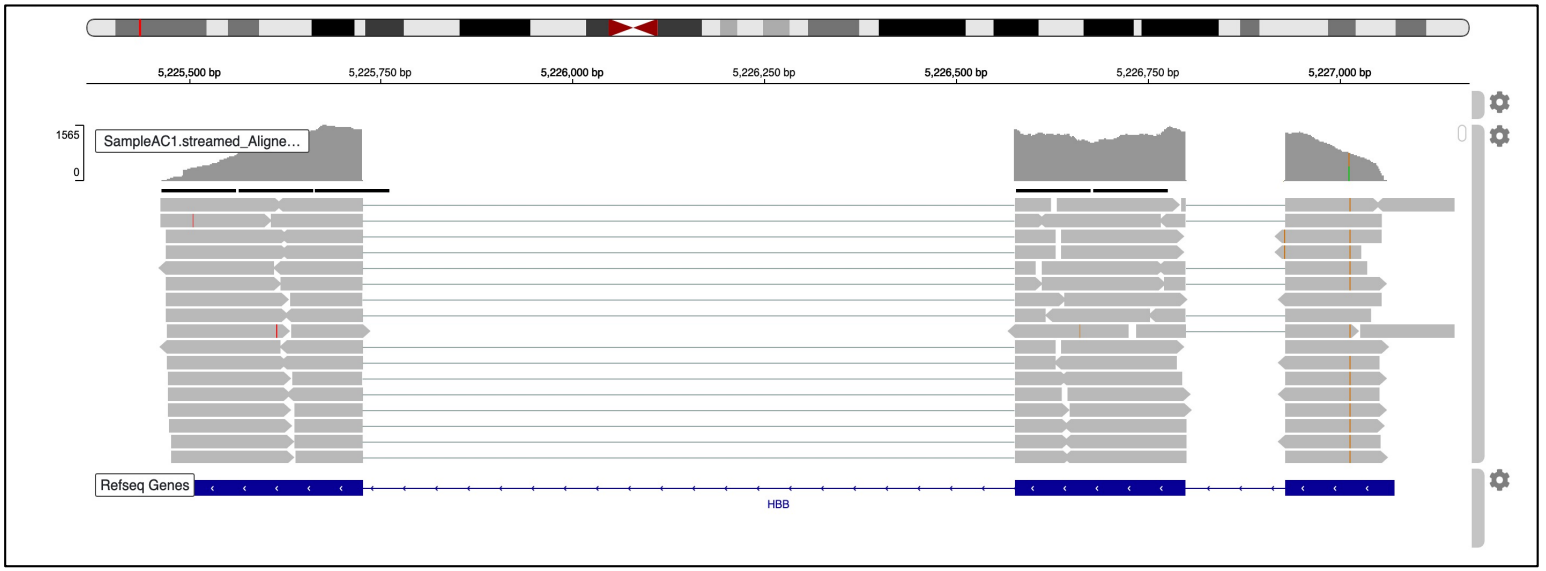
↓

few information

RNA-Seq

- Example of reads aligned to the reference genome

```
AACAAATGAGACGCTGTGCAATTGCTGA
AACAAATGAGACGCTGTGCAATTGCTGAGTAC
AACAAATGAGACGCTGTGCAATTGCTGAGTACCGTA
     ATGAGACGCTGTGCAATTGCTGAGTACCGT
               CTGTGCAATTGCTGAGTACCGTAGGTAGAAC
               CTGTGCAATTGCTGAGTACCGTAGGTAGAAC                    reference genome
GCATCCATCTTGGGGCGTCCCAATTGCTGAGTAACAAATGAGACGCTGTGCAATTGCTGAGTACCGTAGGTAGAACCCCGCAGGGTTAACGACTCATTGTTTACTCTGCGAC
```

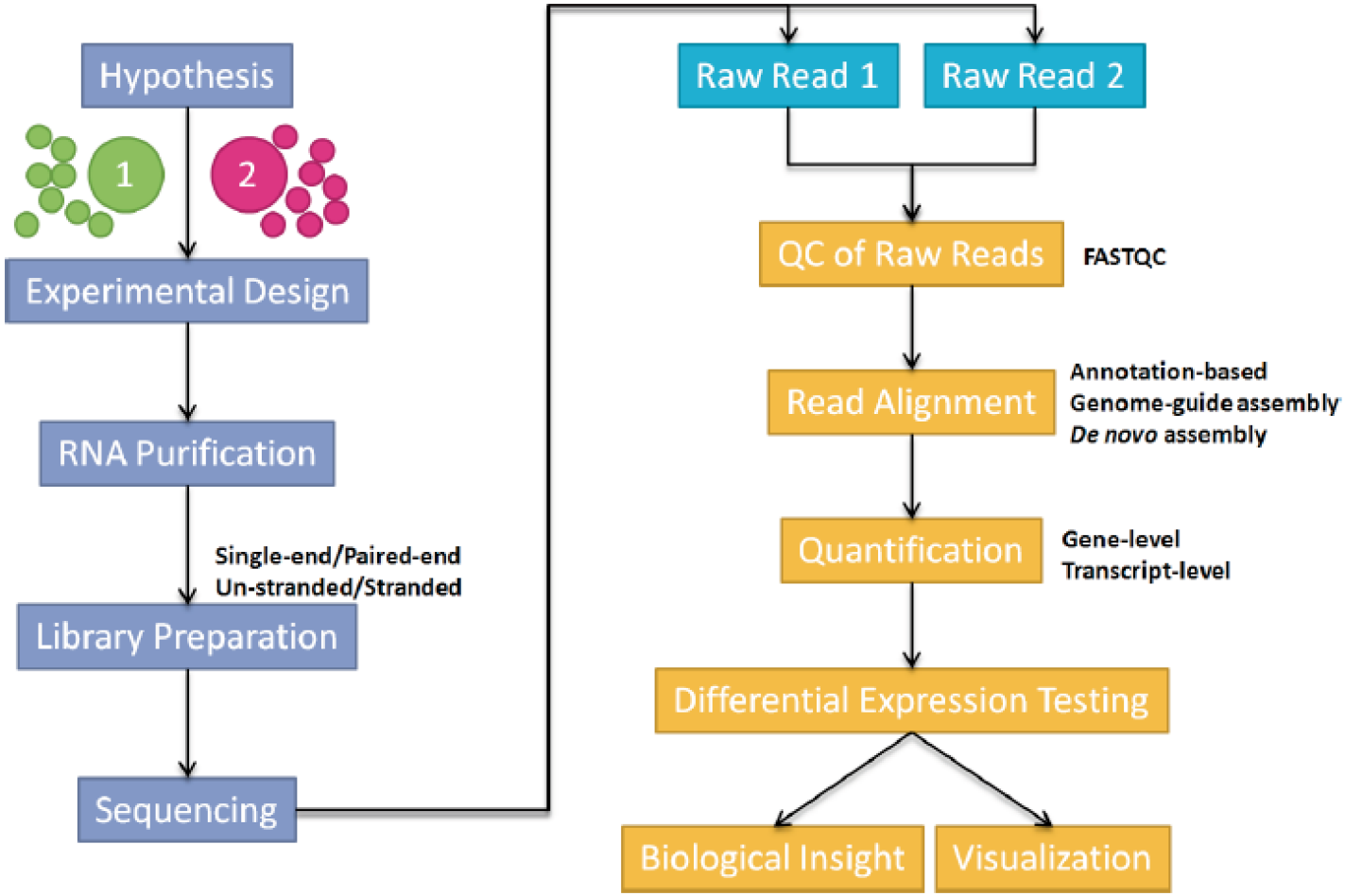


NEXT GENERATION SEQUENCING

Why is Coverage Important?

Application Type	Required Depth	Why it Matters
Germline Variant Calling (Standard Exome)	$\sim 30\text{X}$ to 50X	Necessary to confirm that a base call (especially a heterozygous one) is genuine and not a sequencing error.
Somatic Variant Calling (Tumor)	$\sim 100\text{X}$ to 500X	High depth is required to reliably detect **low-frequency alleles** (variants present in a small fraction of tumor cells).
Gene Expression (RNA-seq)	~ 20 million to 100 million reads	Deeper sequencing is needed to accurately quantify low-abundance transcripts and reliably detect differential expression.
High-Resolution De Novo Assembly	$\sim 80\text{X}$ to 100X	Extremely high coverage is necessary to span repetitive regions and ensure sufficient overlapping reads to accurately piece together a new genome without a reference.

WORK FLOW RNA-Seq



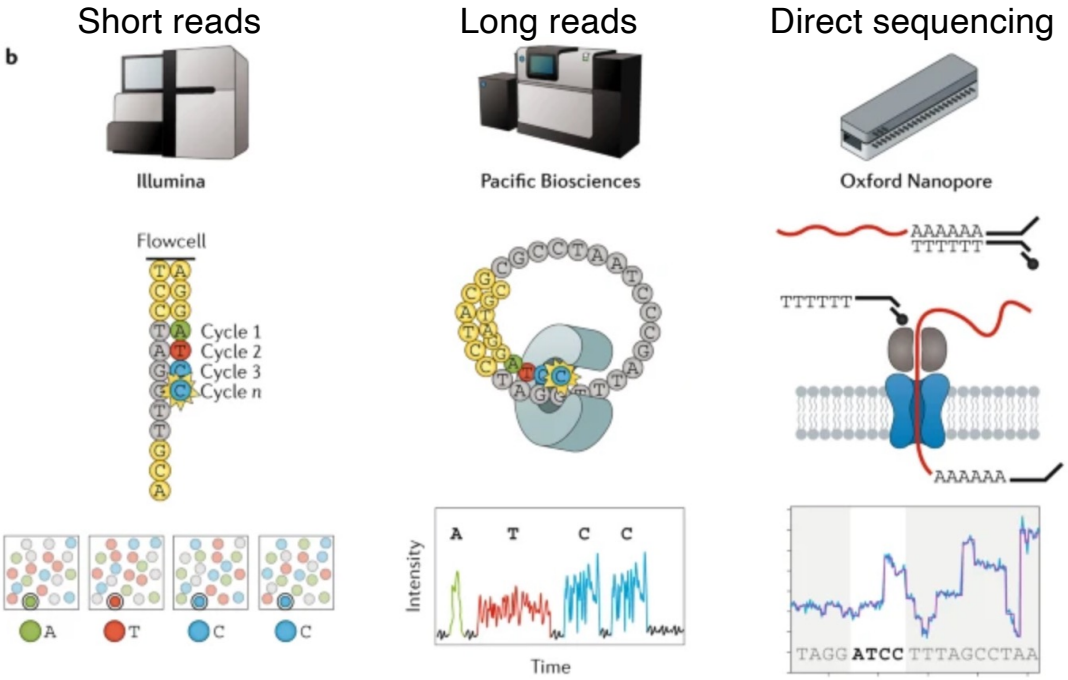
The library preparation methodology is specific to the type of sequencer used.

RNA-Seq: main sequencing technologies

The Illumina workflow: after library preparation, individual cDNA molecules are clustered on a flowcell for sequencing by synthesis using 3′ blocked fluorescently labelled nucleotides. In each round of sequencing, the growing DNA strand is imaged to detect which of the four fluorophores has been incorporated.

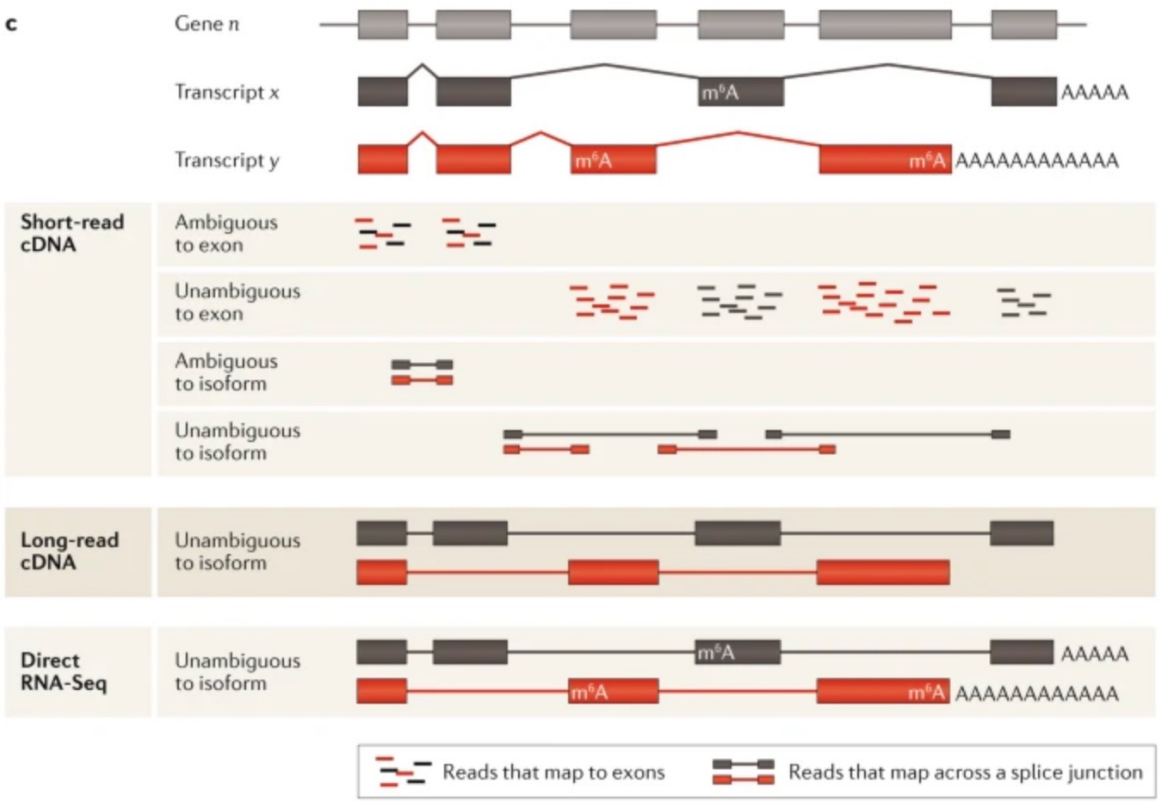
The Pacific Biosciences workflow: after library preparation, individual molecules are loaded into a sequencing chip (SMRT cell), where they bind to a polymerase immobilized at the bottom of a nanowell. As each of the fluorescently labelled nucleotides is incorporated into the growing strand, they fluoresce and are detected.

The Oxford Nanopore workflow: after library preparation, individual molecules are loaded into a flowcell, where motor proteins, which are attached during adaptor ligation, dock with nanopores. The motor protein controls the translocation of the DNA or RNA strand through the nanopore, causing a change in current.

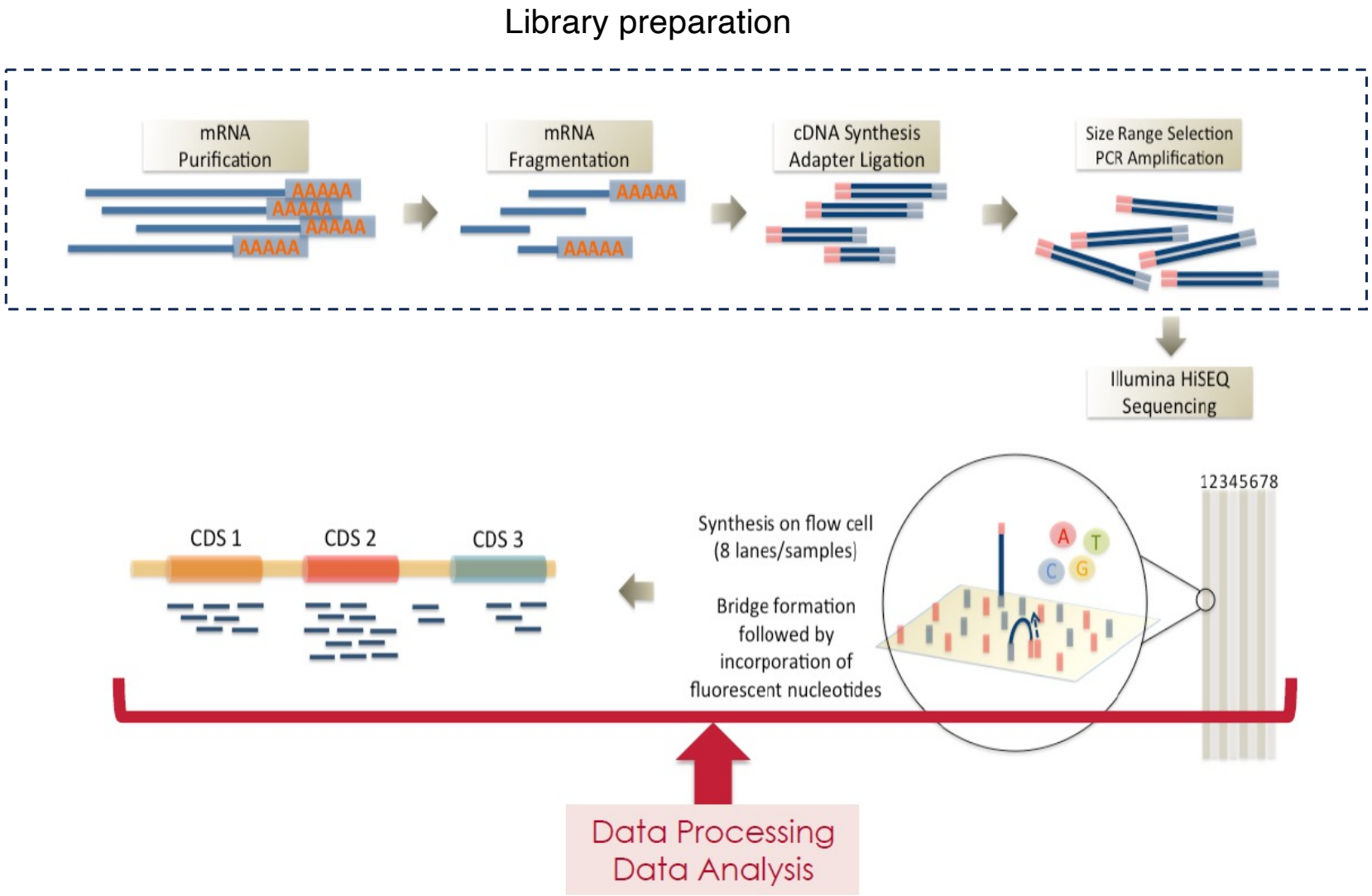


Short-read, long-read and direct RNA-seq analysis.

Over 90% of human genes are alternatively spliced to form two or more distinct and expressed isoforms (transcripts *x* and *y*). The complexity of information captured increases from short-read cDNA sequencing, where isoform detection can be compromised by reads that cannot be mapped unambiguously, to long-read methods that directly sequence isoforms. In short-read cDNA sequencing, a significant proportion of reads map ambiguously when an exon is shared between isoforms; reads that span exon–exon junctions can be used to improve the isoform analysis but can also be mapped ambiguously when a junction is shared between isoforms. These issues complicate analysis and the interpretation of results. Long-read cDNA methods can generate full-length isoform reads that remove, or substantially reduce, these artefacts and improve differential isoform expression analysis. However, these methods rely on cDNA conversion, which removes information about RNA base modifications and can only make crude estimates of polyadenylation (poly(A)) tail length. Direct RNA-seq enables full-length isoform analysis, base modification detection and poly(A) tail length estimation.



RNA-Seq (Illumina)



RNA-Seq: LIBRARY PREPARATION

coding RNAs — **mRNA** — **RNA-seq**

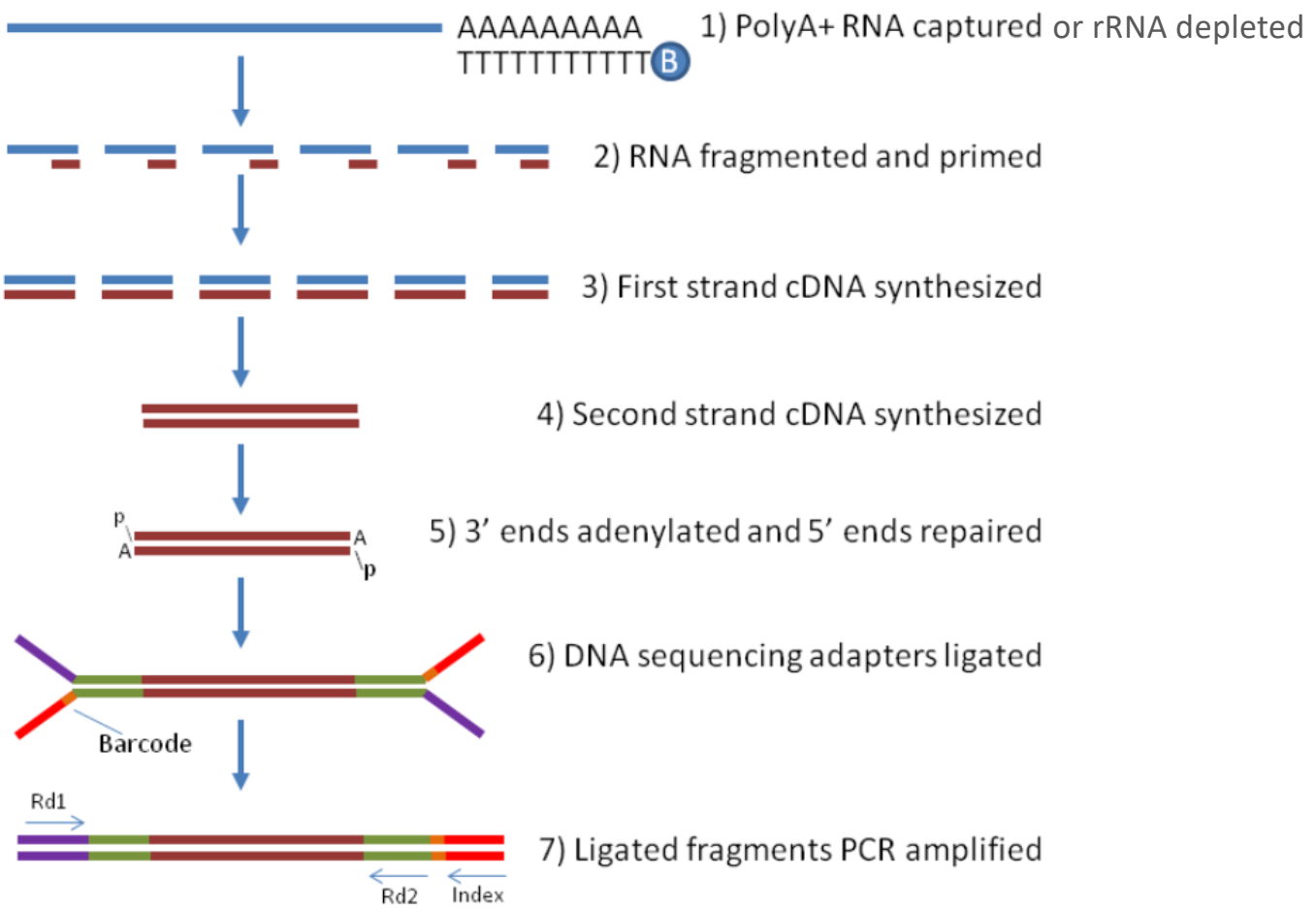
non-coding RNAs

large	**rRNA**		
	Xist		
	lincRNA		
	Pseudogenes		
	circular RNAs		
	……		
small	**tRNA**	translation	
	snRNAs	splicing	
	snoRNAs	modification	
	scRNAs	transl. control	
	gRNAs	editing	
	miRNAs	transl. control	**Small RNA-seq**
	siRNAs	RNA stability	
	rasiRNAs	chromatin	
	piRNAs	genome stability	
	……		

rRNA + tRNA → ~ 95%

RNA-Seq: LIBRARY PREPARATION

Example of library preparation: Illumina Truseq

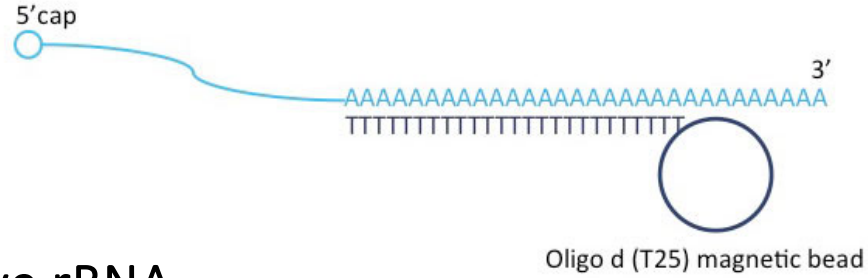


RNA-Seq: LIBRARY PREPARATION

Two ways to isolate long RNA molecules:

1a - Purify and Fragment mRNA

This process purifies the poly-A containing RNA molecules (mainly mRNA) using poly-T oligo-attached magnetic beads.



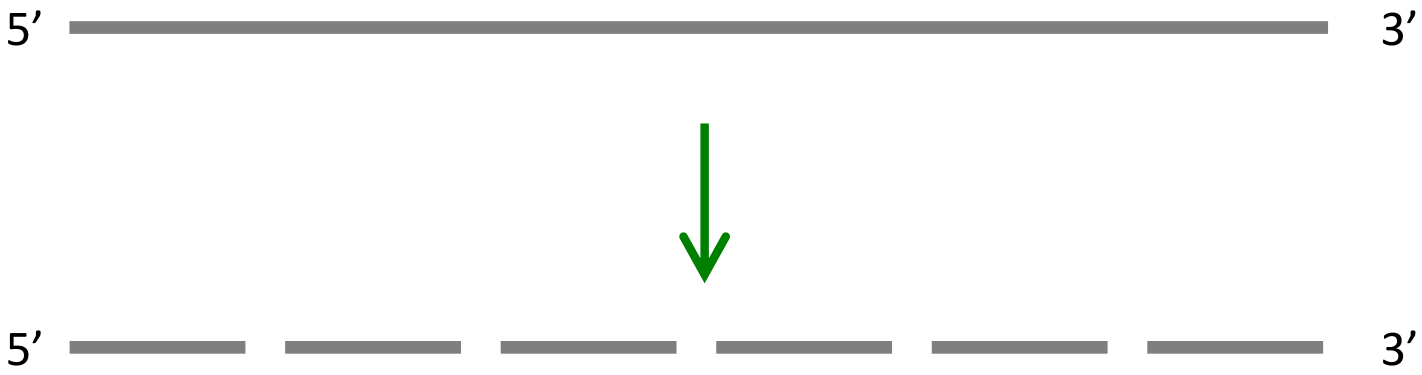
1b - Remove rRNA

After the ribosomal RNA is depleted, the remaining RNA (not only mRNA) is purified, fragmented and primed for cDNA synthesis. rRNA is removed using a hybridization-based technique.

RNA-Seq: LIBRARY PREPARATION

2 - RNA fragmentation

RNA molecules are fragmented into small pieces using divalent cations under elevated temperature

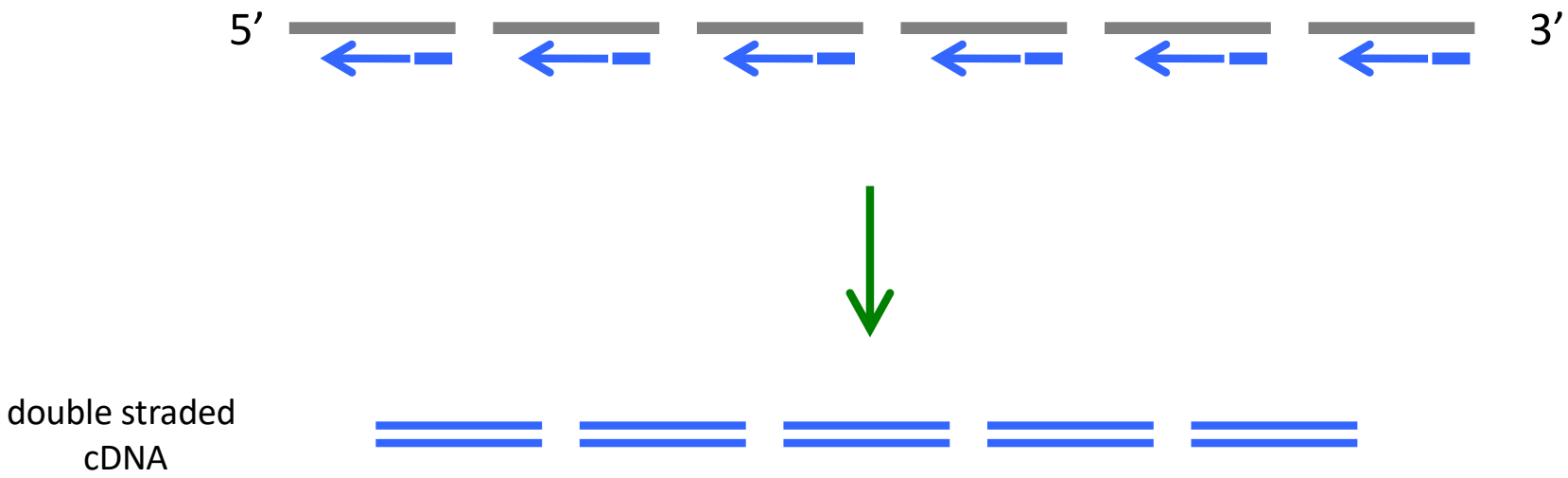


Range of fragments length: **120-225 bp**

RNA-Seq: LIBRARY PREPARATION

3 - Synthesize First Strand cDNA

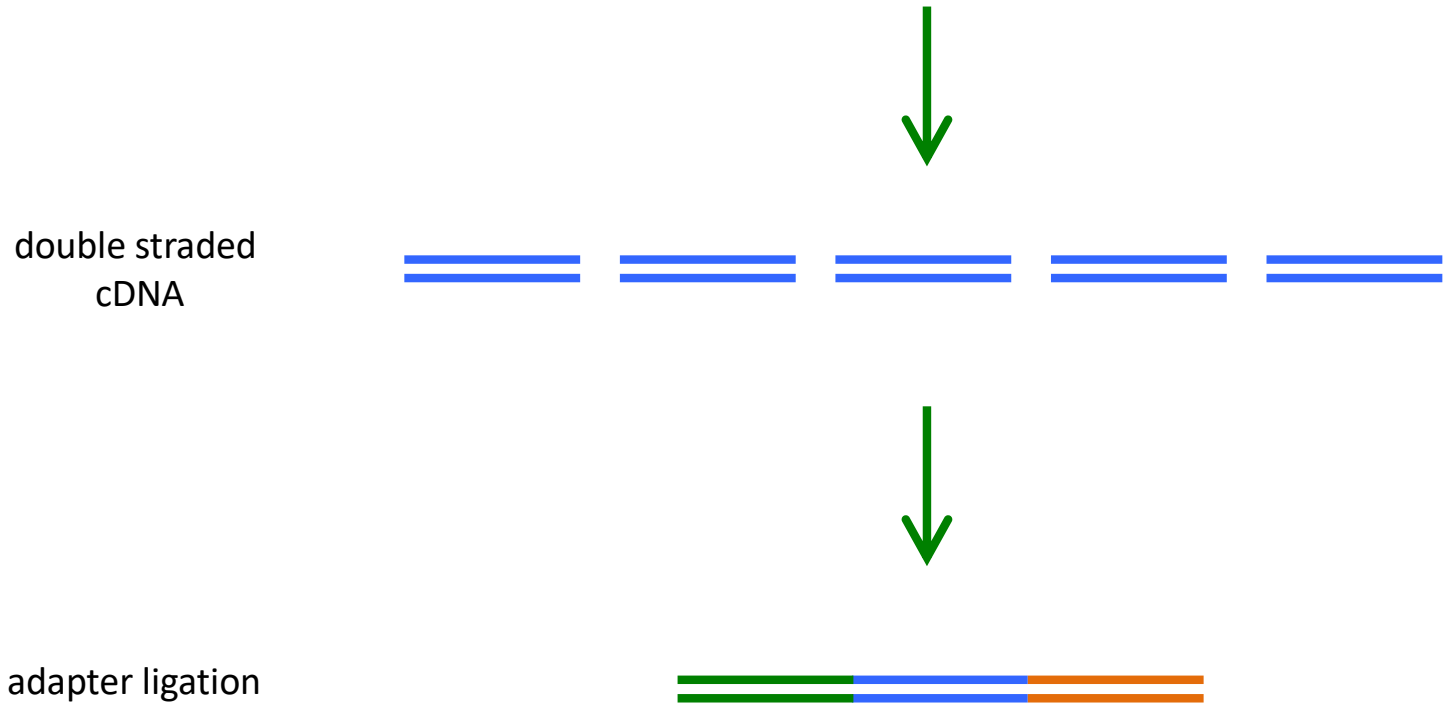
This process reverse transcribes the cleaved RNA fragments that were primed with random hexamers into first strand cDNA using reverse transcriptase and random primers.



RNA-Seq: LIBRARY PREPARATION

4 - SynAdapter ligation

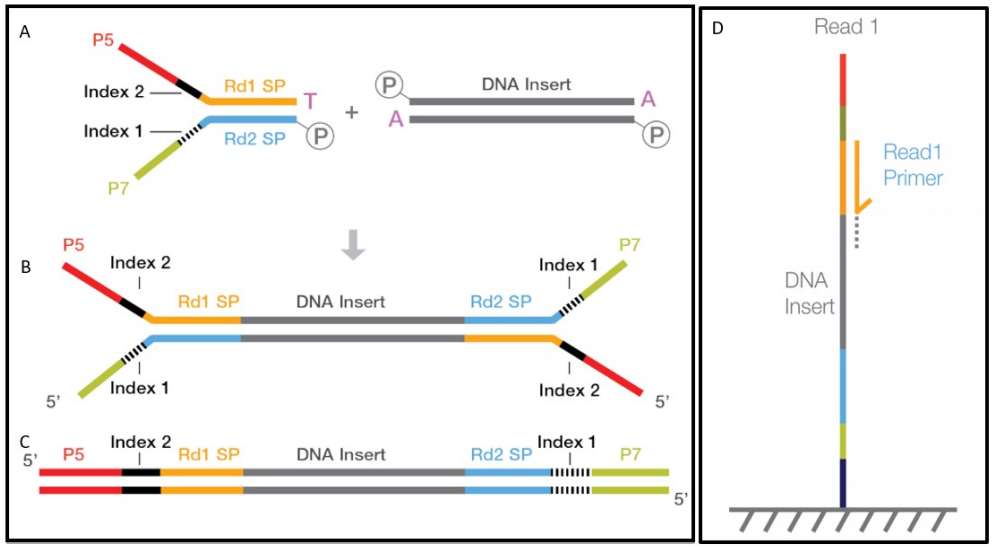
the attachment of **adapters** is the single most critical step that converts cDNA into a functional Illumina sequencing library. This process is essential because the Illumina system relies on these specific sequences to initiate sequencing, clustering, and multiplexing.



RNA-Seq: LIBRARY PREPARATION

Adapters are short synthetic DNA oligonucleotide (~50-70 bp) that serve three main functional roles in the Illumina workflow :

a) **Flow Cell Attachment:** They contain sequences (P5 and P7) that are complementary to the short DNAoligonucleotides fixed on the surface of the Illumina **flow cell**. This allows the cDNA fragments to bind to the flow cell for bridge amplification.
b) **Sequencing Primer Binding Sites:** They contain sites (R1 and R2) where the sequencing primers will anneal to initiate the read process in both the forward and reverse directions (Paired-End sequencing).
c) **Index/Barcode Sequences:** They include unique DNA sequences, called indexes or barcodes, which are sequenced separately. These indexes allow many different libraries (samples) to be mixed and sequenced in a single lane (multiplexing), and then computationally separated (demultiplexed) afterward.

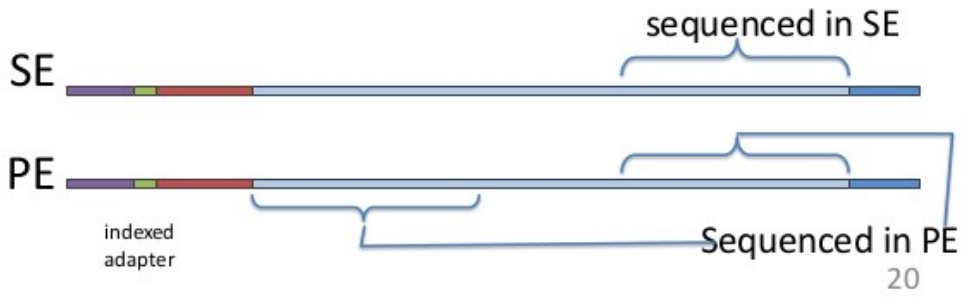


RNA-Seq: LIBRARY PREPARATION

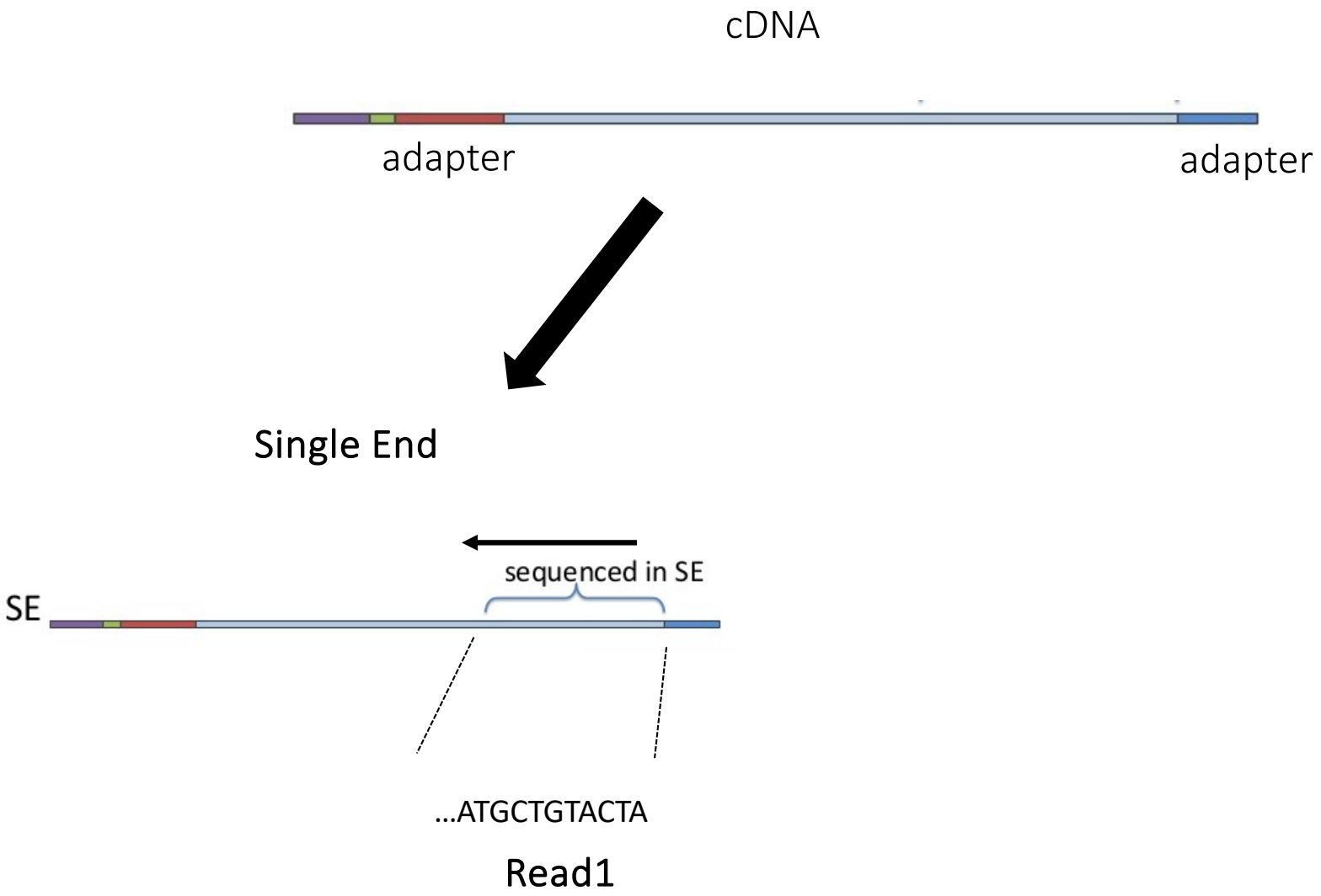
Single-end VS paired-end sequencing

- **Single-end sequencing (SE)**, involves sequecing of the fragment from only one end.

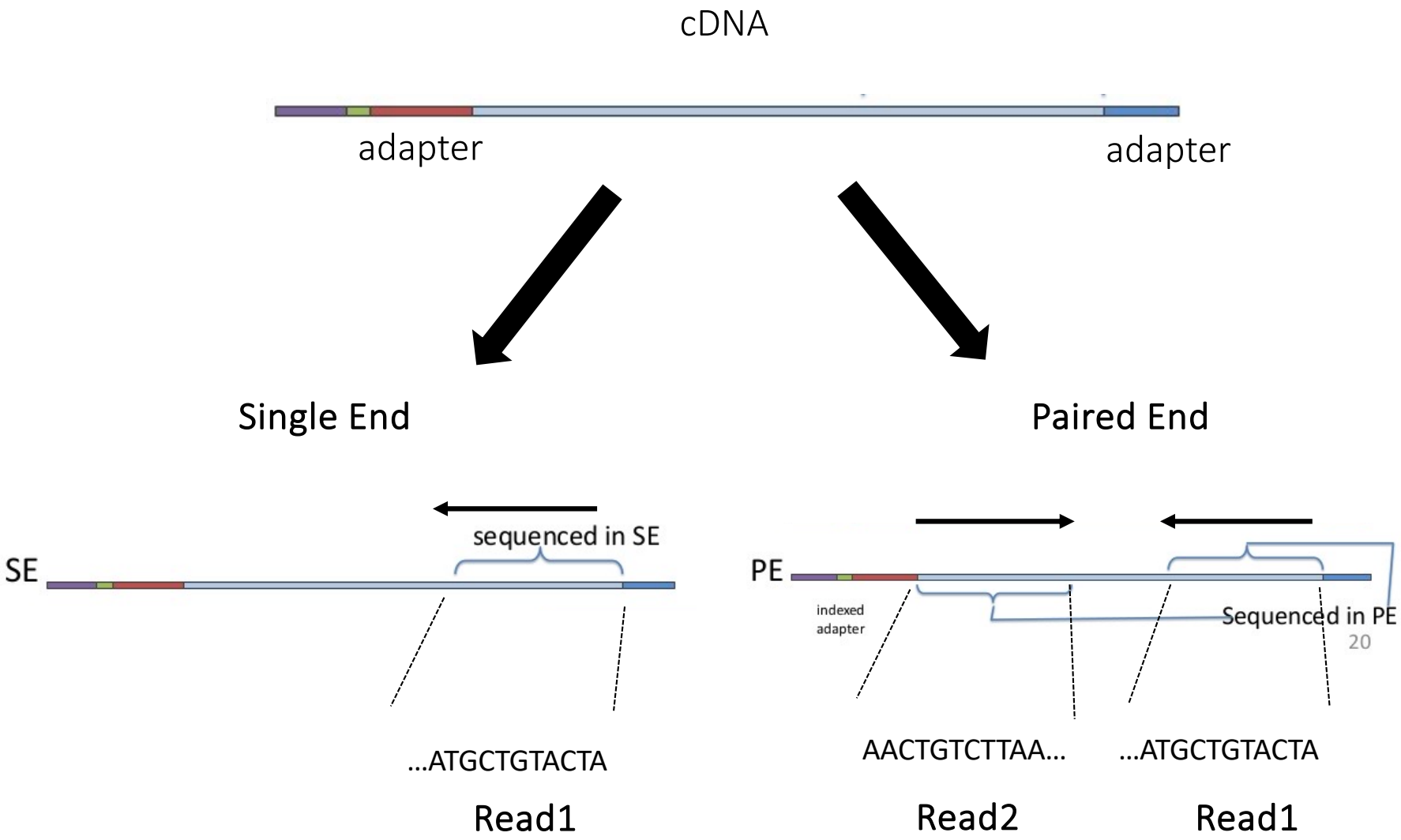
- **Paired-end sequecing (PE)**, involves sequencing both ends of a fragment, resulting in the production of read pairs. This allows to improve the alignment, to better identify and quantify splicing variants, and to detect rearrangements such as insertions, deletions, and inversions.



RNA-Seq: LIBRARY PREPARATION

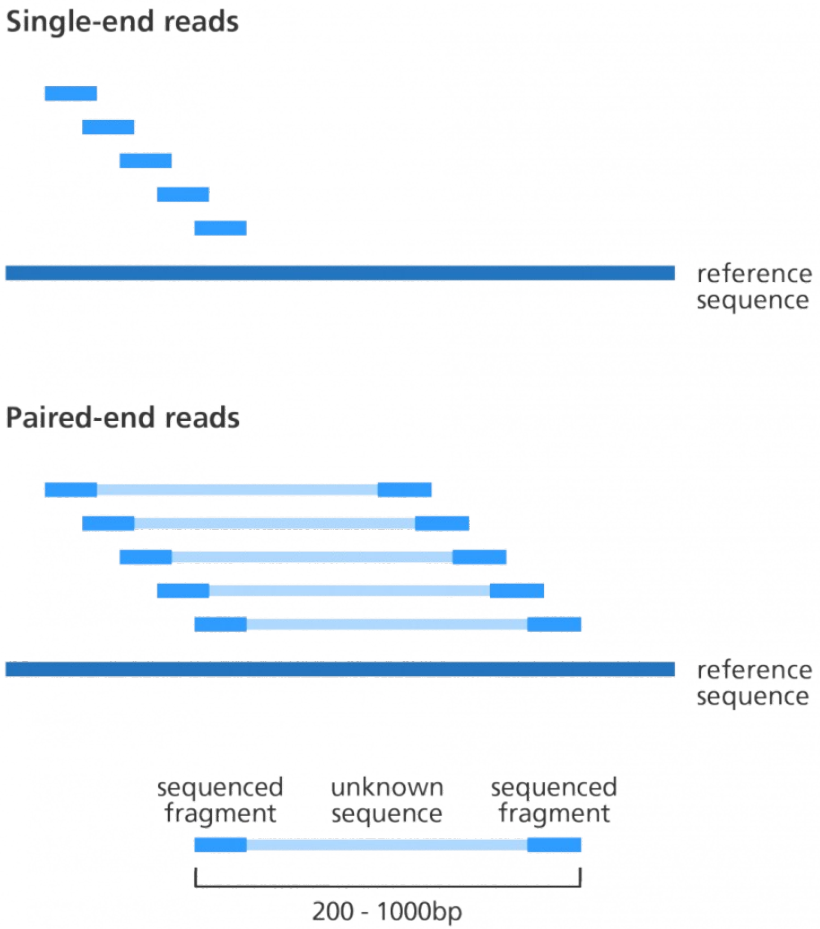


RNA-Seq: LIBRARY PREPARATION



RNA-Seq: LIBRARY PREPARATION

Single-end VS paired-end sequencing



RNA-Seq: SEQUENCING REACTION

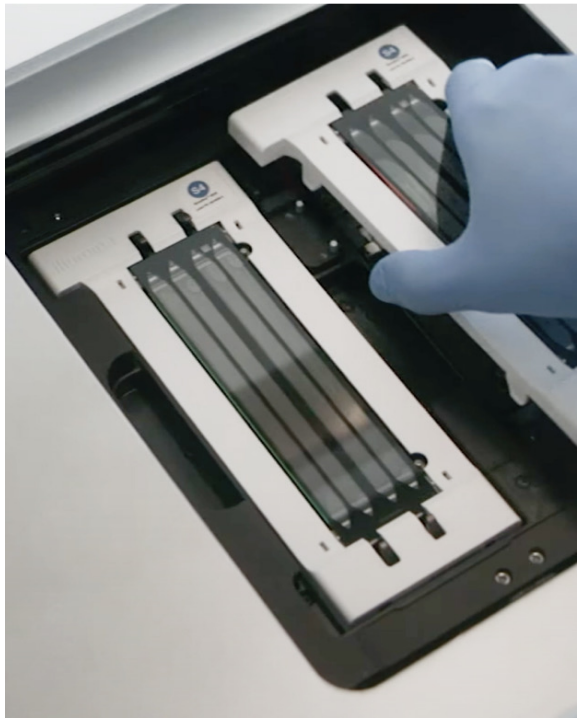



Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

NovaSeq 6000 Sequencing System

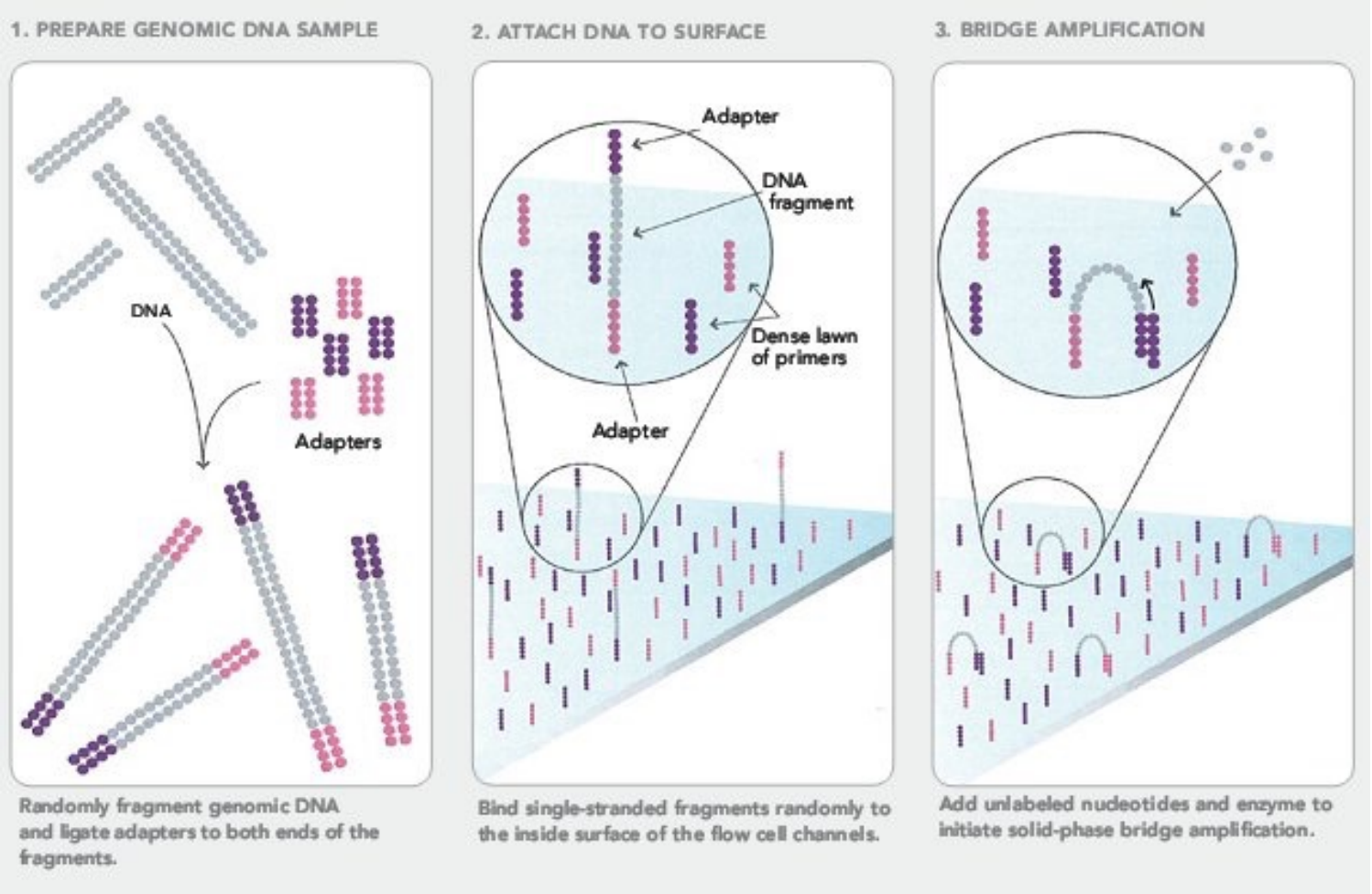
System specifications

- **Output range**
- 80–6000 Gb
- **Single reads per run**
- 650M–20B
- **Max read length**
- 2 × 250 bp
- **Run time**
- 13–44 hr



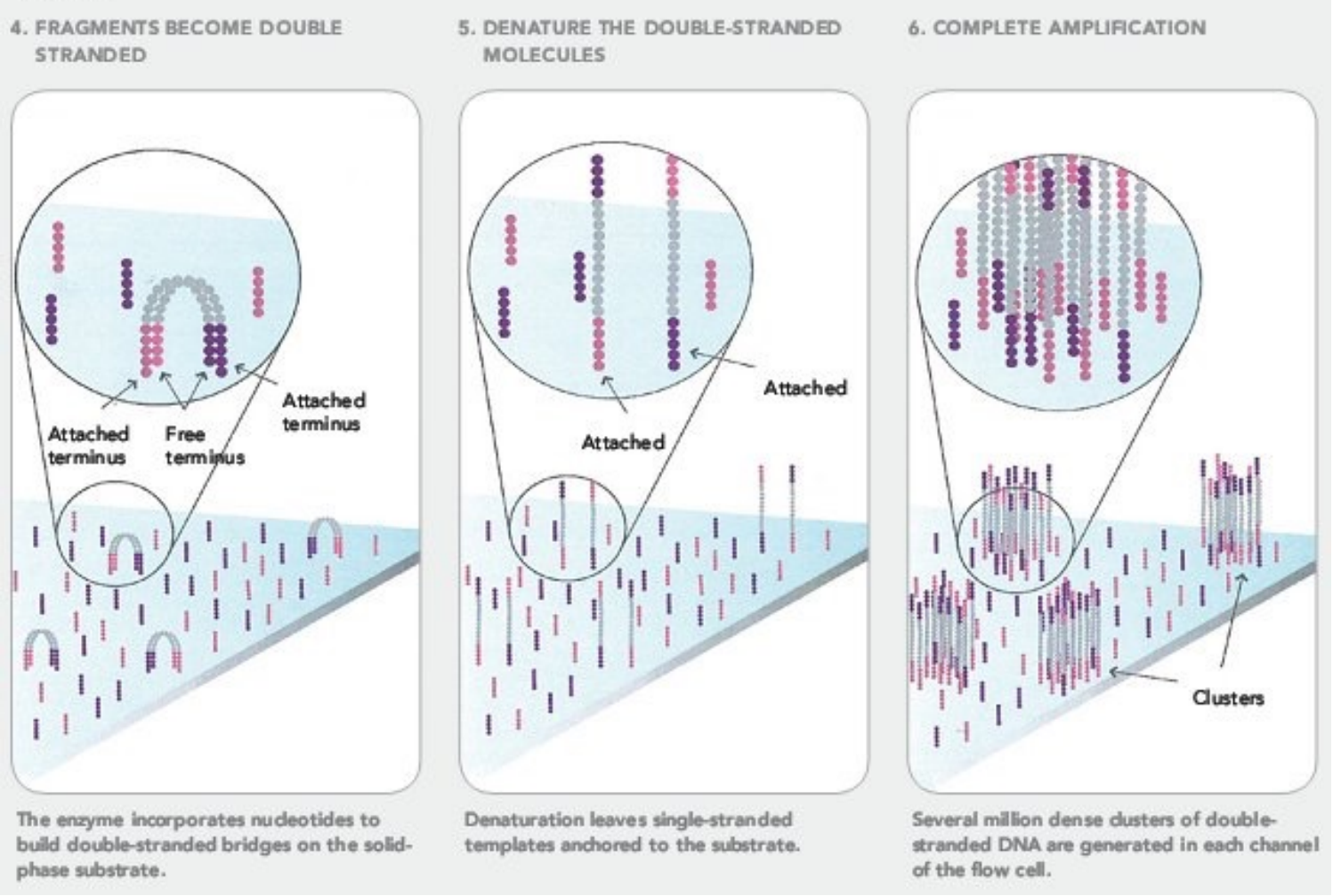
RNA-Seq: SEQUENCING REACTION

Illumina platform: Sequencing by Synthesis



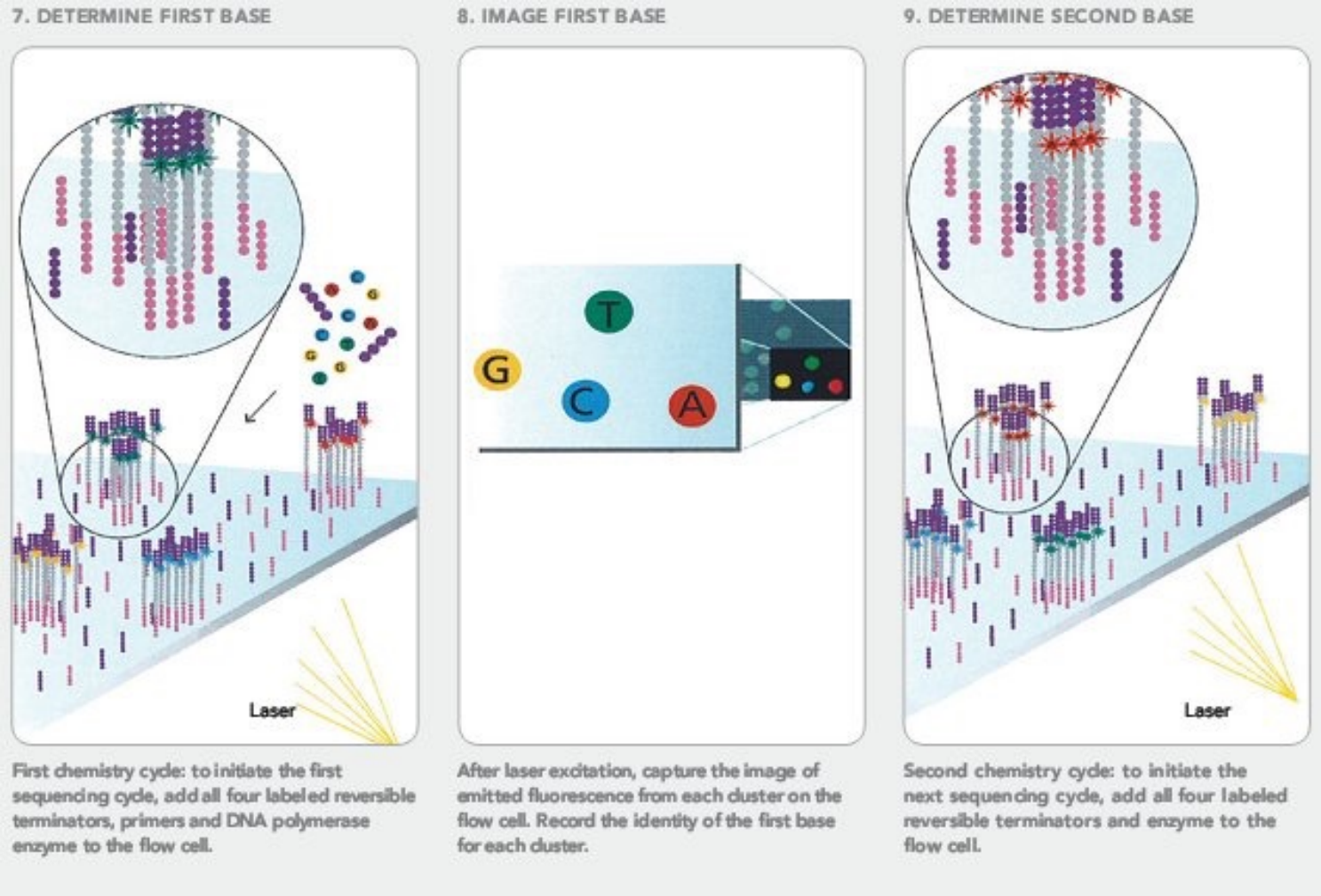
RNA-Seq: SEQUENCING REACTION

Illumina platform: Sequencing by Synthesis



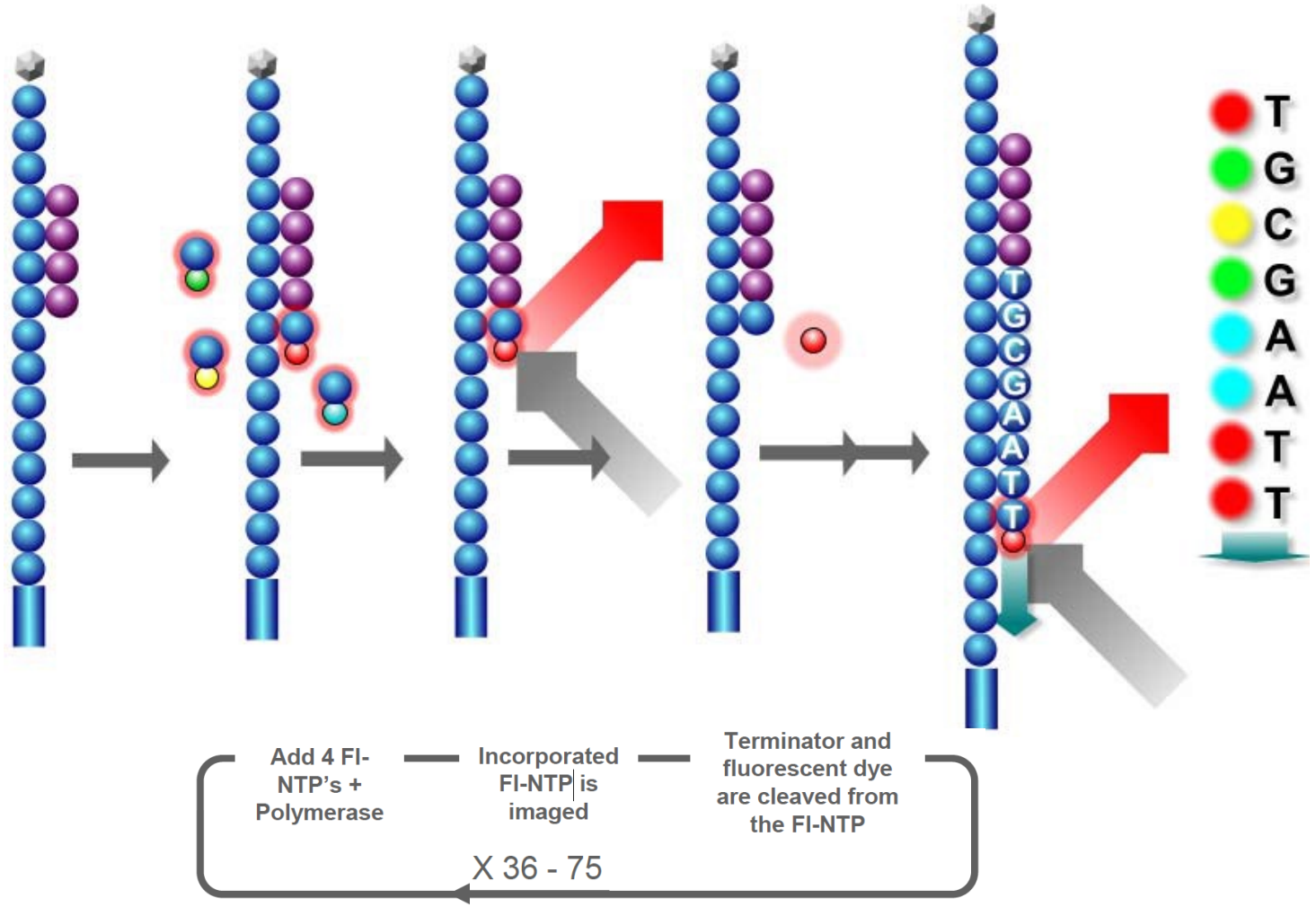
RNA-Seq: SEQUENCING REACTION

Illumina platform: Sequencing by Synthesis



TOTAL READS NUMBER = Number of clusters in flow cell

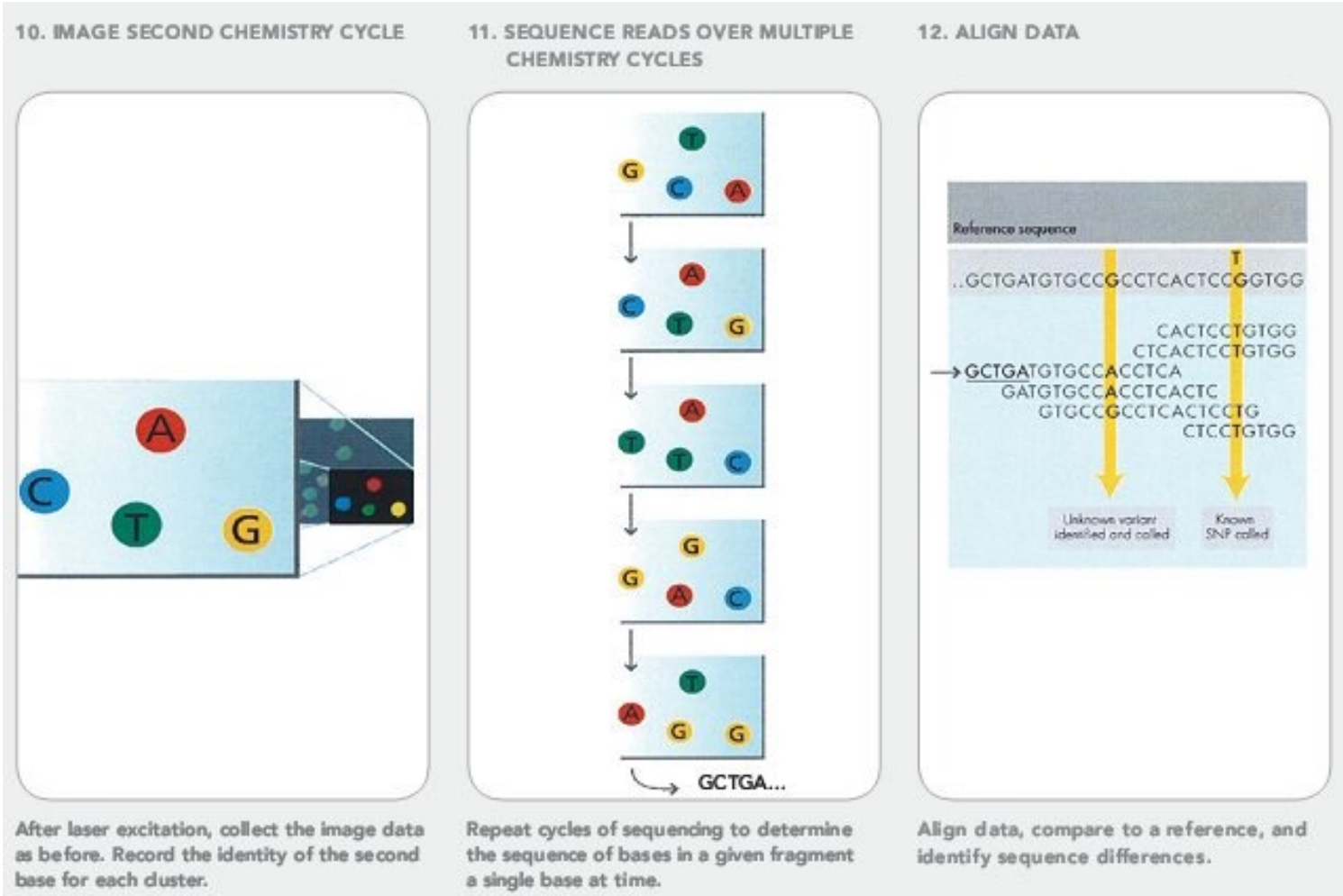
RNA-Seq: SEQUENCING REACTION



READ LENGTH = Number of reaction cycles

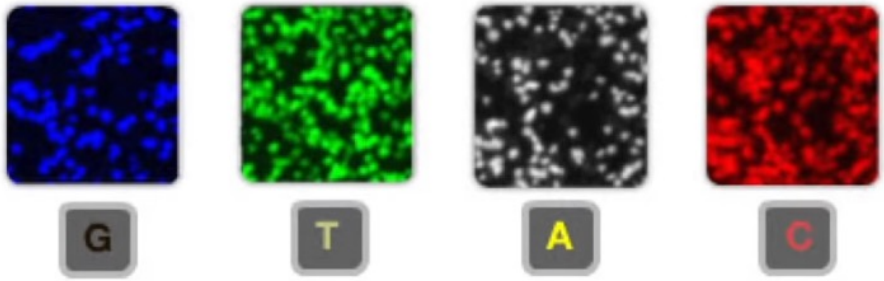
WET LAB: Reazione di Sequenziamento

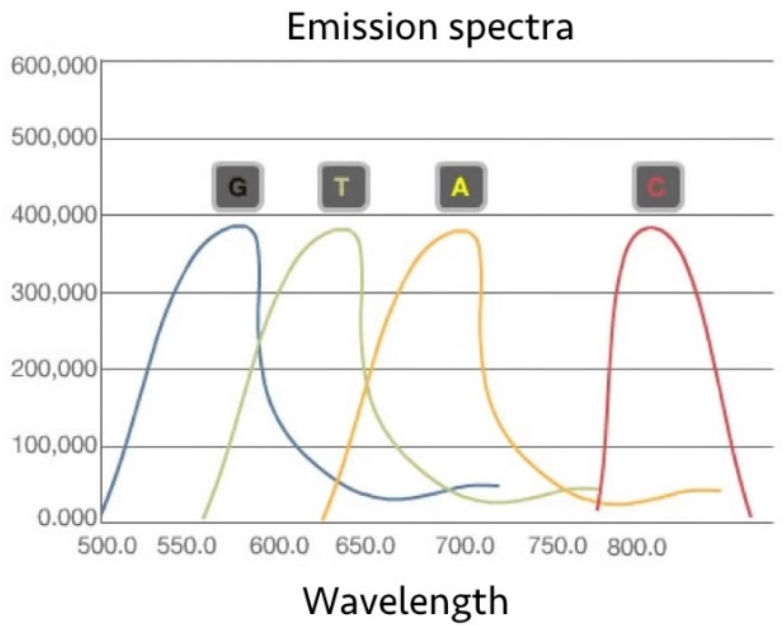
Illumina platform: Sequencing by Synthesis



RNA-Seq: SEQUENCING REACTION

4 colour chemistry





Illumina sequencing

Video:

https://www.youtube.com/watch?v=HMyCqWhwB8E

RNA-Seq: Illumina Sequencers

Key specifications	 NextSeq 1000 and 2000 Systems	 NovaSeq 6000 System	 NovaSeq X Series
Max output per flow cell	540 Gb[a]	3 Tb[b]	8 Tb[c]
Run time (range)[d]	~8–44 hr	~13–44 hr	~17–48 hr
Max reads per run (single reads)	1.8B[a]	10B (single flow cell)[b] 20B (dual flow cells)[b]	26B (single flow cell)[c] 52B (dual flow cells)[c,f]
Max read length	2 × 300 bp	2 × 250 bp	2 × 150 bp

Key applications and methods

	 NextSeq 1000 and 2000 Systems	 NovaSeq 6000 System	 NovaSeq X Series
Large whole-genome sequencing (human, plant, animal)		✓	✓
Small whole-genome sequencing (microbe, virus)	✓	✓	✓
Exome and large panel sequencing (enrichment-based)	✓	✓	✓
Targeted gene sequencing (amplicon-based, gene panel)	✓	✓	✓
Single-cell profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	✓	✓	✓
Transcriptome sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	✓	✓	✓
Chromatin analysis (ATAC-Seq, ChIP-Seq)	✓	✓	✓
Methylation sequencing	✓	✓	✓

RNA-Seq: Illumina Benchtop Sequencers

New

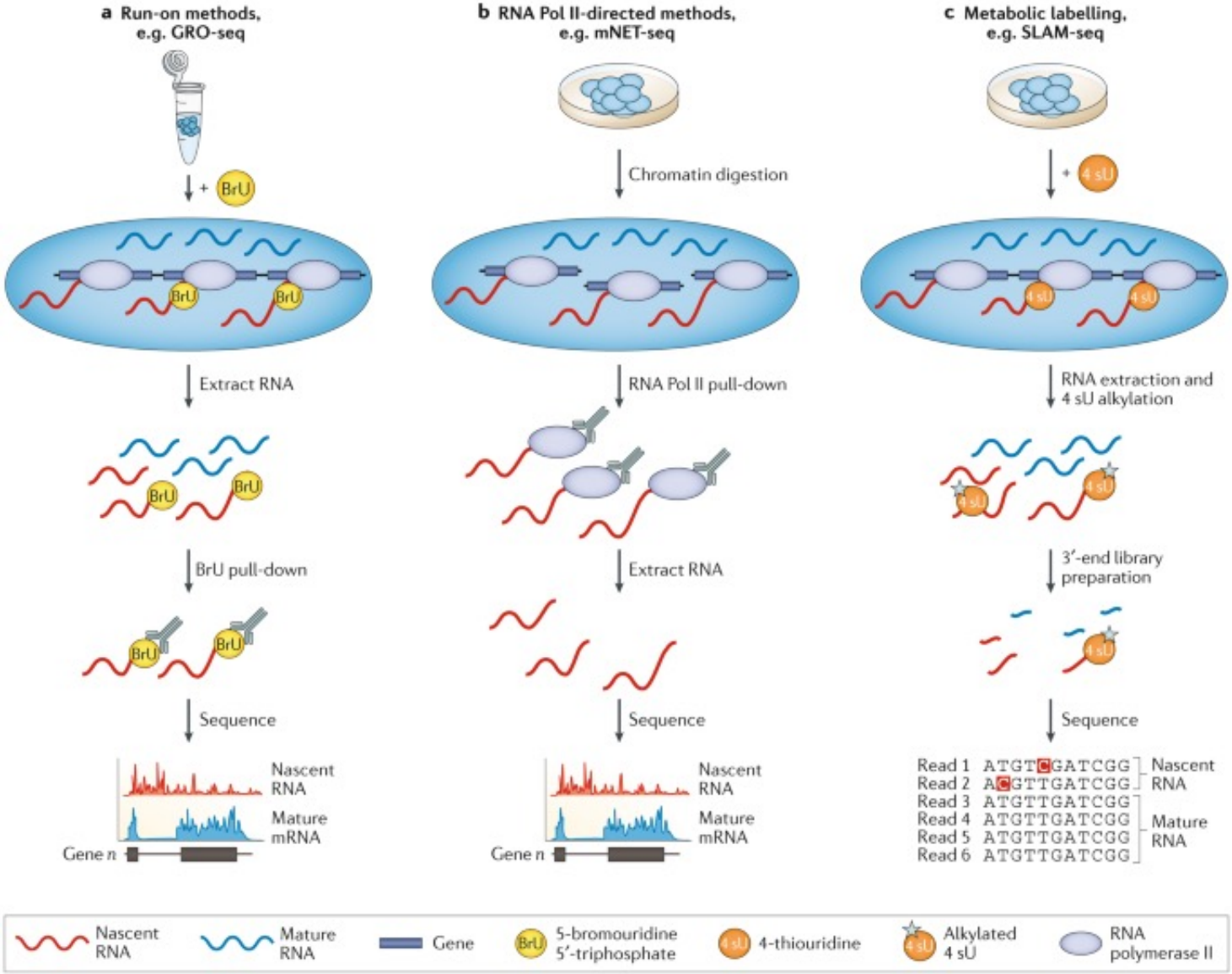
Key specifications	MiSeq i100 Series	NextSeq 550 System	NextSeq 1000 and 2000 Systems
Max output per flow cell	30 Gb [a]	120 Gb [b]	540 Gb
Run time (range)[c]	~4–24 hr	~11–29 hr	~8–44 hr
Max reads per run (single reads)	100M [a]	400M [b]	1.8B
Max read length	2 × 500 bp	2 × 150 bp	2 × 300 bp

Key applications and methods

	New  MiSeq i100 Series	 NextSeq 550 System	 NextSeq 1000 and 2000 Systems
Small whole-genome sequencing (microbe, virus)	✓	✓	✓
Exome and large panel sequencing (enrichment-based)		✓	✓
Targeted gene sequencing (amplicon-based, gene panel)	✓	✓	✓
Single-cell profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	✓	✓	✓
Transcriptome sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	✓	✓	✓
Targeted gene expression profiling	✓	✓	✓
miRNA and small RNA analysis	✓	✓	✓
DNA–protein interaction analysis (ChIP-Seq)	✓	✓	✓
Methylation sequencing	✓	✓	✓
16S metagenomic sequencing	✓	✓	✓

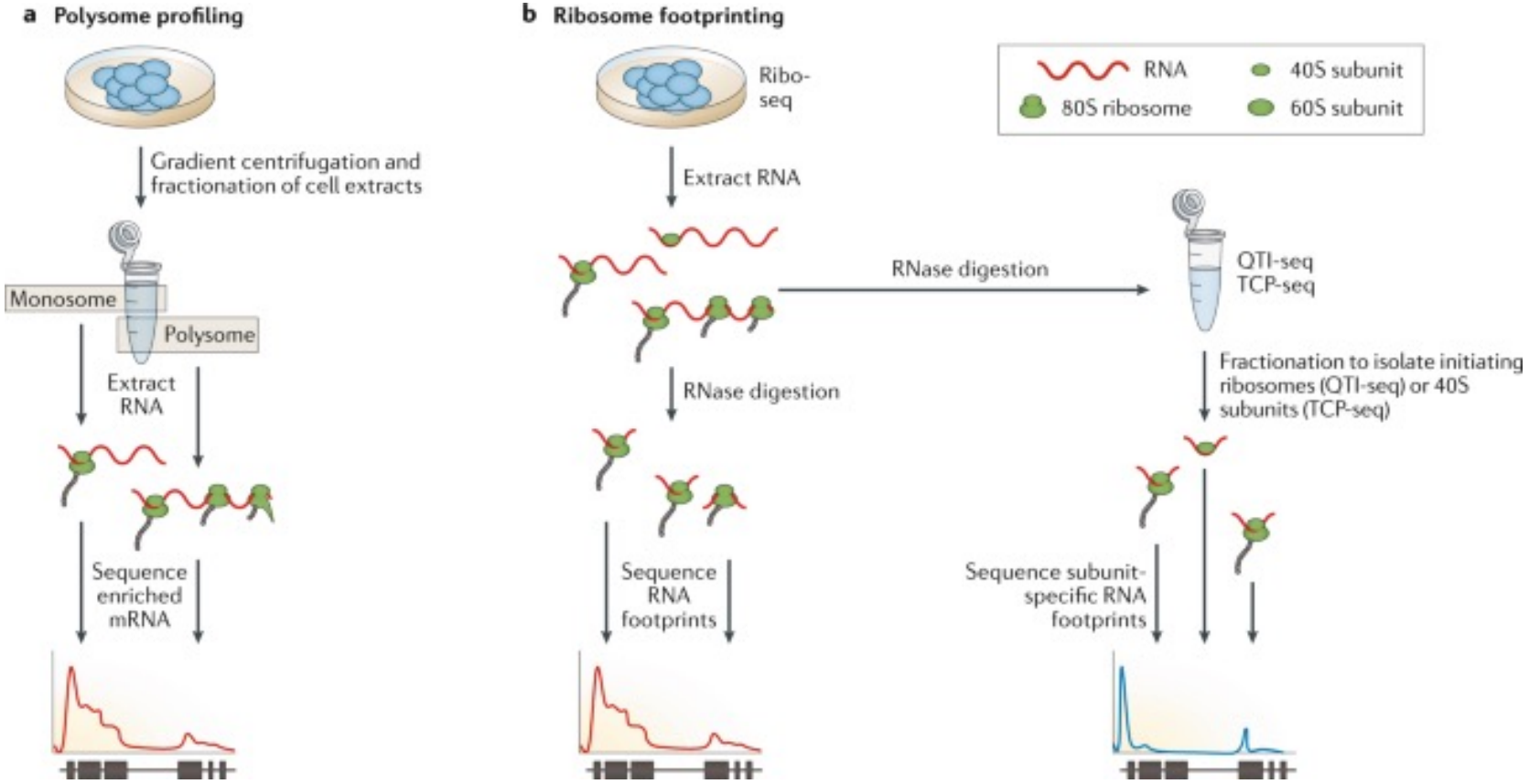
Nascent RNA sequencing

Nascent RNA analysis methods enrich newly transcribed RNAs from the other RNA in a cell and compare this to an unenriched (mature RNA) control, by one of three primary methods



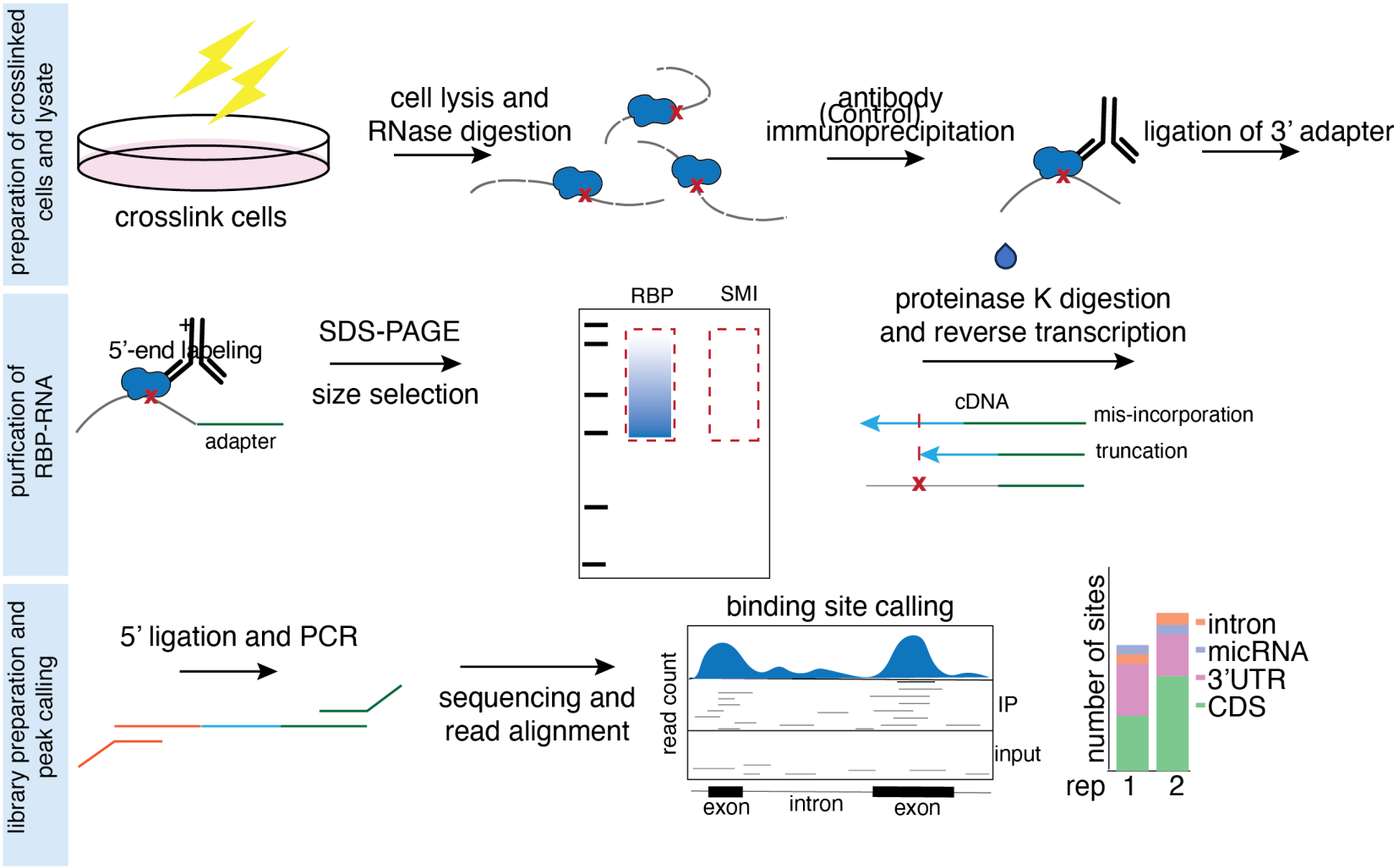
Translatome analysis methods

Translatome analysis methods generate RNA-seq data from ribosomally bound RNA, with an assumption that mRNA ribosome density correlates with the protein synthesis level.

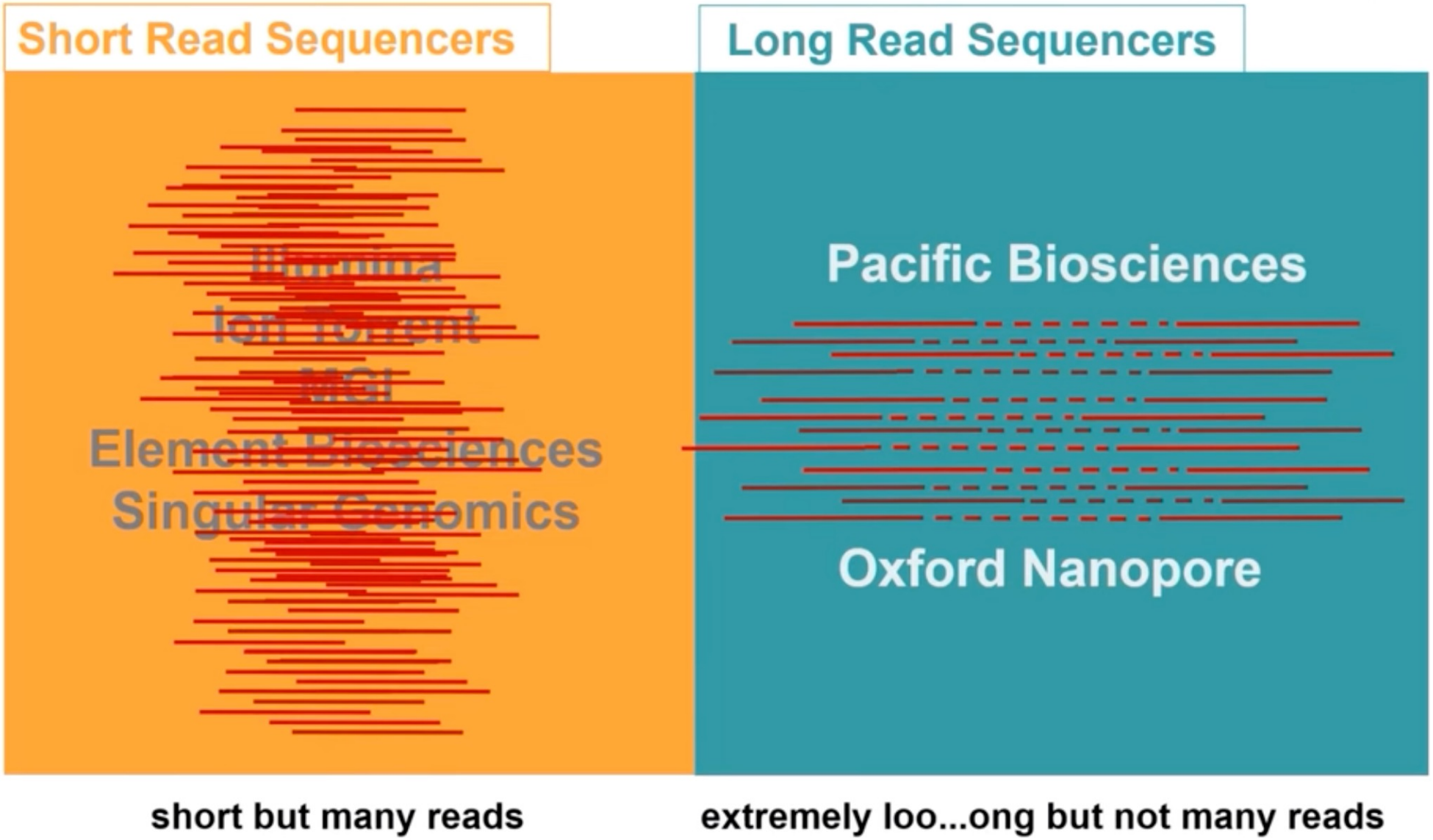


RNA–protein interaction analysis

CLIP-seq

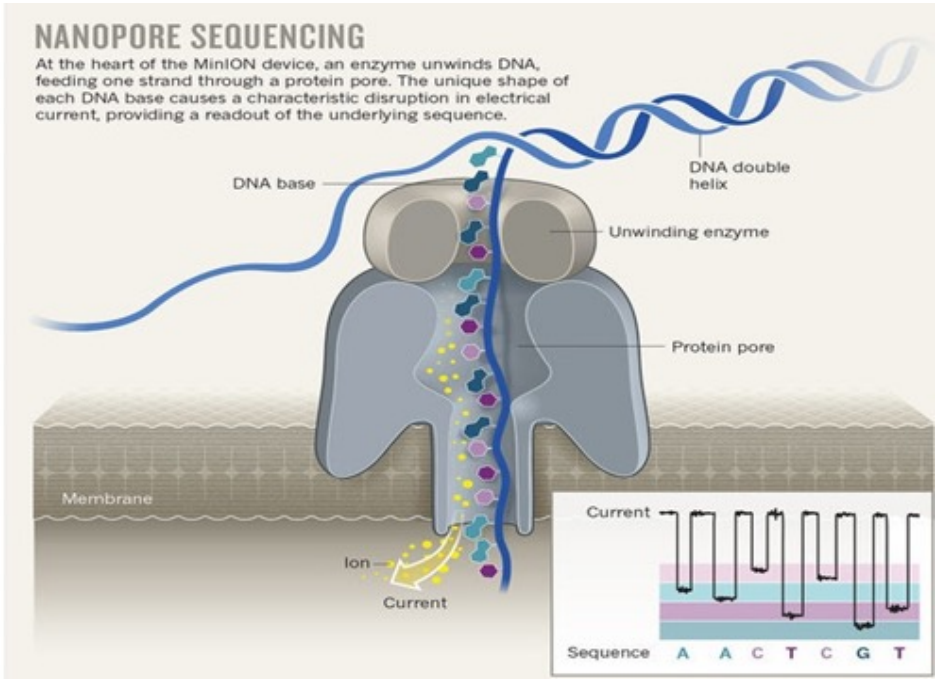


Short reads vs Long reads



NANOPORE

- **MinIon** Oxford Nanopore



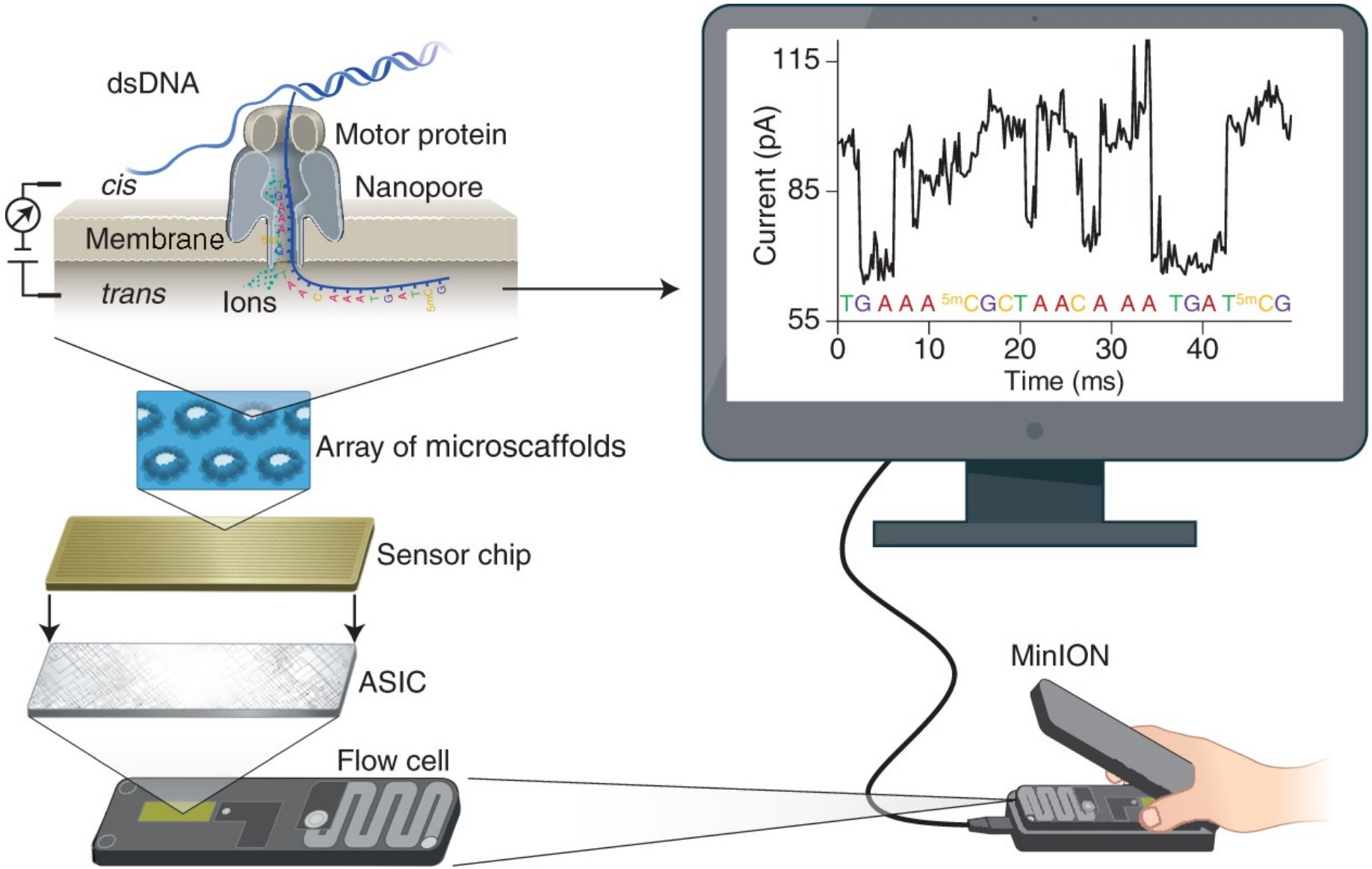
- High error rate (1%-2%)
  - Biased errors
- Really long reads (2 Mb)
- Can directly sequence RNA
- Maybe proteins in the future?

Easy sample preparation

Fast (450bases/sec) and cheap

Realtime data

Oxford nanopore: https://www.youtube.com/watch?v=E9-Rm5AoZGw



NANOPORE

Nanopore is extremely portable

Nature **521,** 15–16 (07 May 2015)

NANOPORE

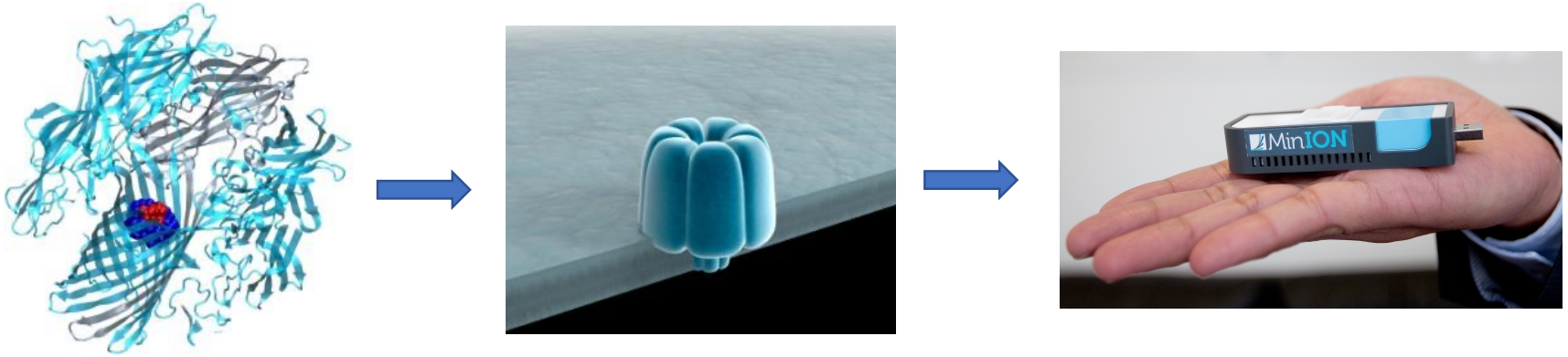
Nanopore is extremely portable

MinION: field deployable

The MinION device can sequence small genomes, such as those of bacteria and viruses, displaying the results as they are generated.

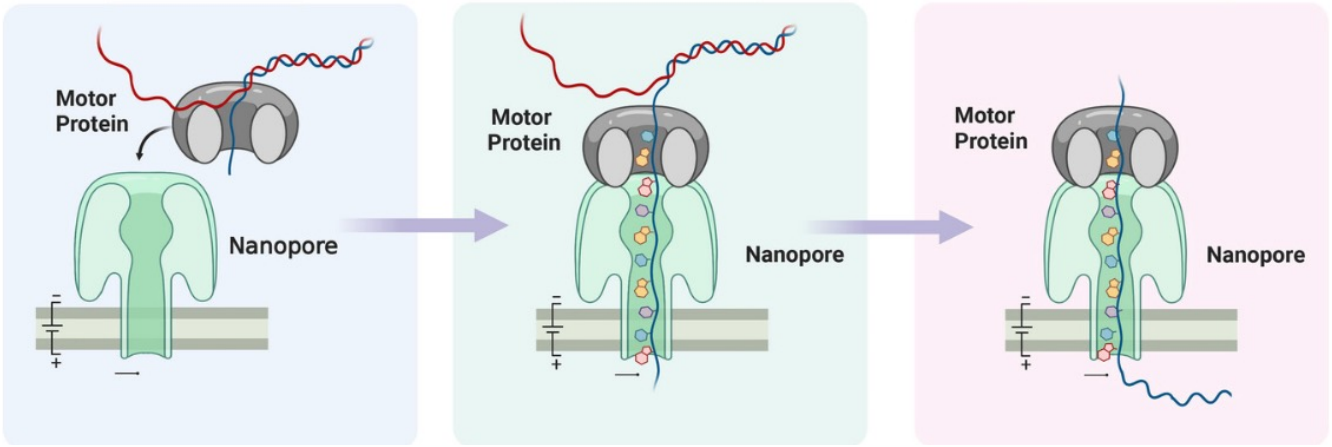
NASA Astronaut Kate Rubins sequenced DNA in space for the first time ever for the Biomolecule Sequencer investigation, using the MinION sequencing device.
Credits: NASA

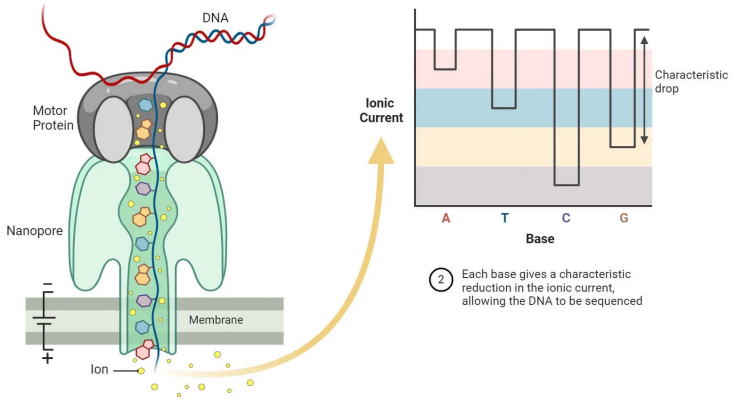
Oxford Nanopore



Oxford Nanopore

Motor proteins pull the nucleic acid through the nanopore, and the nucleic acid is detected and transmitted to the computer through the generated tiny current signal. Each base give a characteristic reduction of the ionic current, allowing the DNA to be sequenced.





PACIFIC BIOSCIENCE SEQUENCING

PACIFIC BIOSCIENCE SEQUENCING

From 500bp to 30000bp

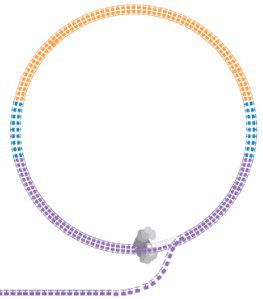
Start with high-quality double stranded DNA

- hairpin adaptors

Prepare SMRTbell libraries

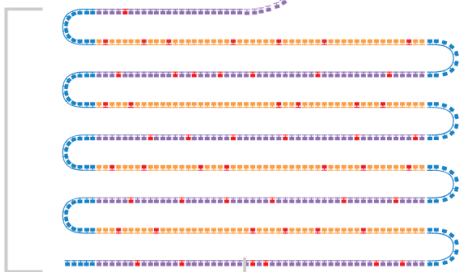
Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes



- many sequencing cycles

The polymerase reads are trimmed of adapters to yield subreads



Consensus and methylation status are called from subreads

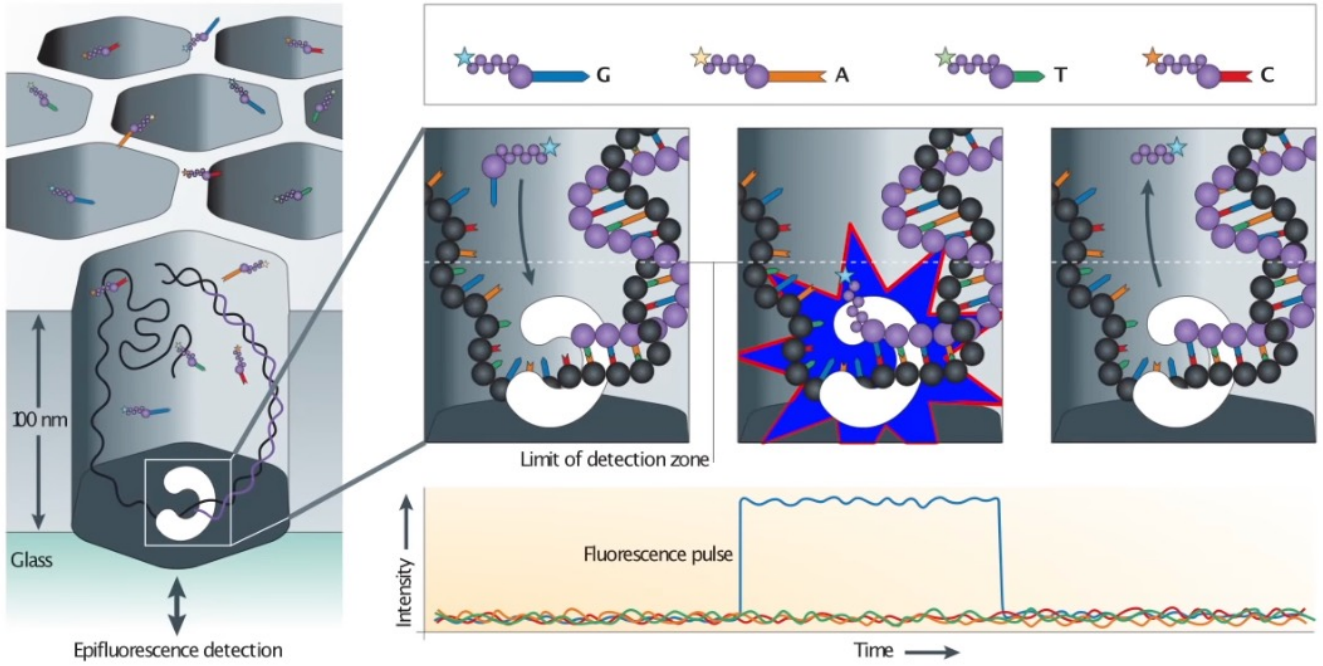
HiFi read
(99.9% accuracy)

- distinguish mutations and random errors

PACBIO sequencing machine

Sequencing calls on the PacBio RSII platform are based on the optical detection of the incorporation of phospholinked nucleotide. This is the principle behind the SMRT (Single Molecule Real Time) sequencing, and happens in tiny ZMWs microwells (Zero Mode Waveguides) directly on the sequencing flowcell (SMRTcell).

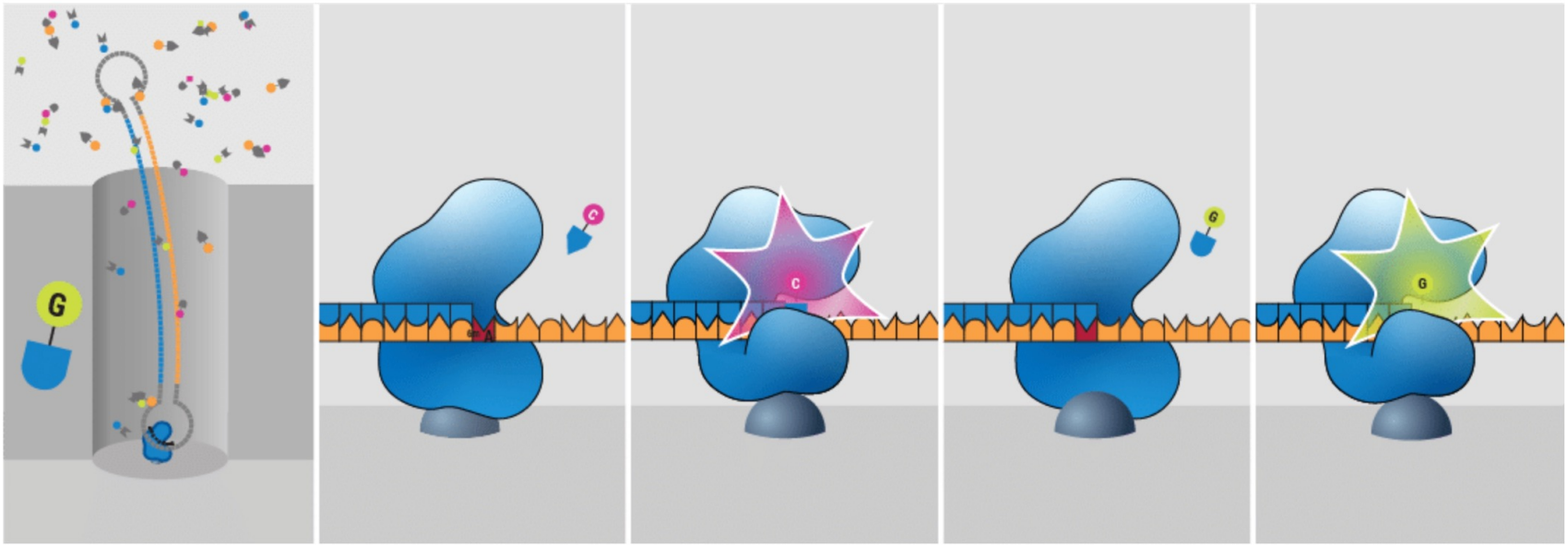
fluorescently labeled nucleotides, the unique fluorescent signal of each base (A, T, C, or G)



Zero-Mode Waveguides (ZMWs). Each ZMW contains a single DNA polymerase, which synthesizes the complementary strand of the DNA template.

Nature Reviews Genetics **11**, 31–46 (2010)

PACIFIC BIOSCIENCE SEQUENCING



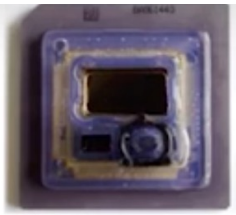
As the polymerase incorporates fluorescently labeled nucleotides, the unique fluorescent signal of each base (A, T, C, or G) is detected and recorded. The continuous observation of the polymerase allows for long reads, making it particularly useful for sequencing large and complex genomes and for detecting structural variants

PACIFIC BIOSCIENCE SEQUENCING

https://www.youtube.com/watch?v=v8p4ph2MAvI&t=99s

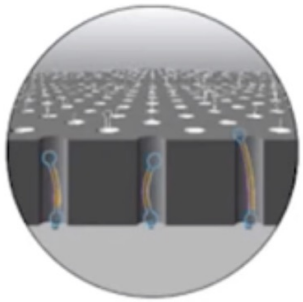
PACIFIC BIOSCIENCE SEQUENCING vs ILLUMINA

PacBio SMRT sequencing



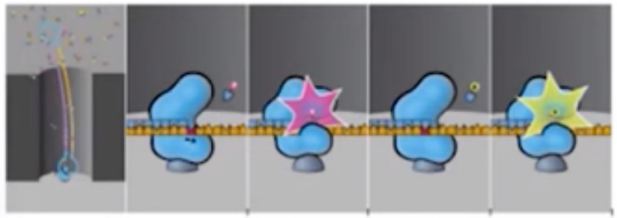
The SMRT™ Cell

INTE

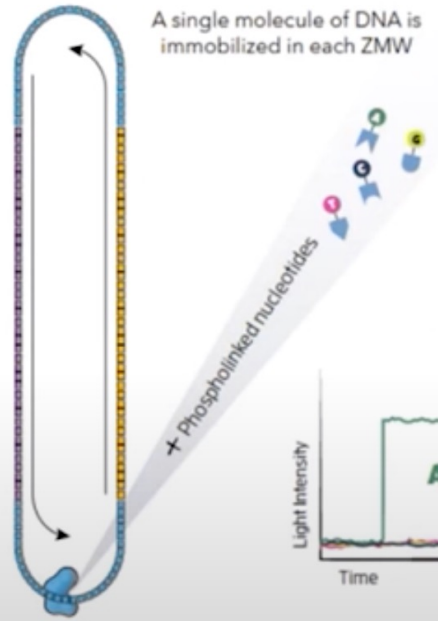


SMRT Cells contain millions of zero-mode waveguides (ZMWs)

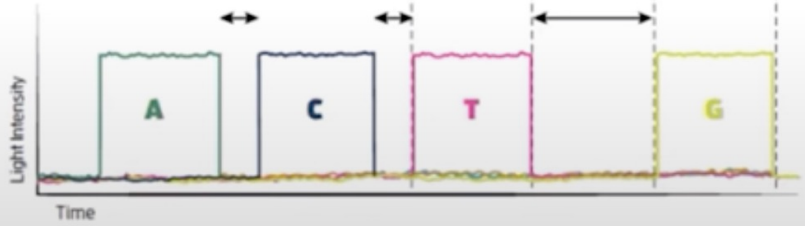
A single molecule of DNA is immobilized in each ZMW



As anchored polymerases incorporate labeled bases, light is emitted



SMRTbell® templates enable repeated sequencing of circular template with real-time detection of base incorporation



Directly detect DNA modifications during sequencing

Nucleotide incorporation kinetics are measured in real time

Speed: PacBio: 2 base incorporations / second (Illumina: 1 base incorporation / hour)

PACIFIC BIOSCIENCE SEQUENCING vs OXFORD NANOPORE

PacBio Sequel IIe

MinIONs

PACIFIC BIOSCIENCE SEQUENCING vs OXFORD NANOPORE

Long Read Sequencing

PacBio vs ONT in a nutshell

PacBio Sequencing:

- Long read lengths up to tens of kilobases for improved genome assembly and structural variant detection.
- High accuracy with HiFi sequencing technology.
- Capable of detecting DNA modifications for epigenetic analysis.
- Minimal GC bias and reduced impact of repetitive sequences.

Oxford Nanopore Sequencing:

- Portability and real-time analysis suitable for fieldwork and rapid surveillance.
- Ultra-long read lengths up to hundreds of kilobases spanning for comprehensive genome assemblies.
- Minimal sample preparation and rapid turnaround time for time-sensitive applications.
- Direct RNA sequencing without reverse transcription or amplification steps.
- Single-molecule sensitivity for detecting rare variants

Nanopore sequencing

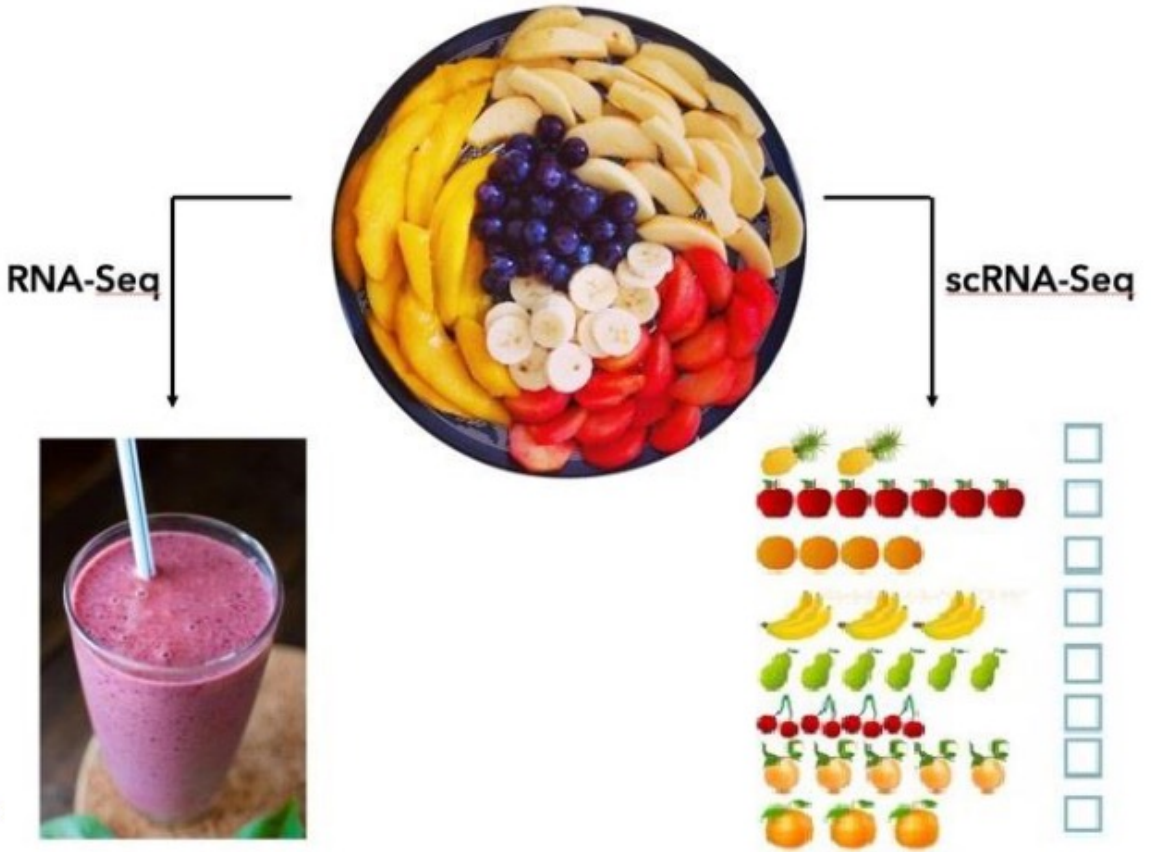
LONG READS

Why long reads?

- Downsides
  - Harder to prepare
  - Costs more
- Benefits
  - Easier to assemble genomes - bigger puzzle pieces
  - Identify structural variations
  - Phase variations - which variants are on which chromosomes?

Bulk vs Single cell RNA–Seq (scRNA-Seq)

Average expression level
- Comparative transcriptomics
- Disease biomarker
- Homogenous systems

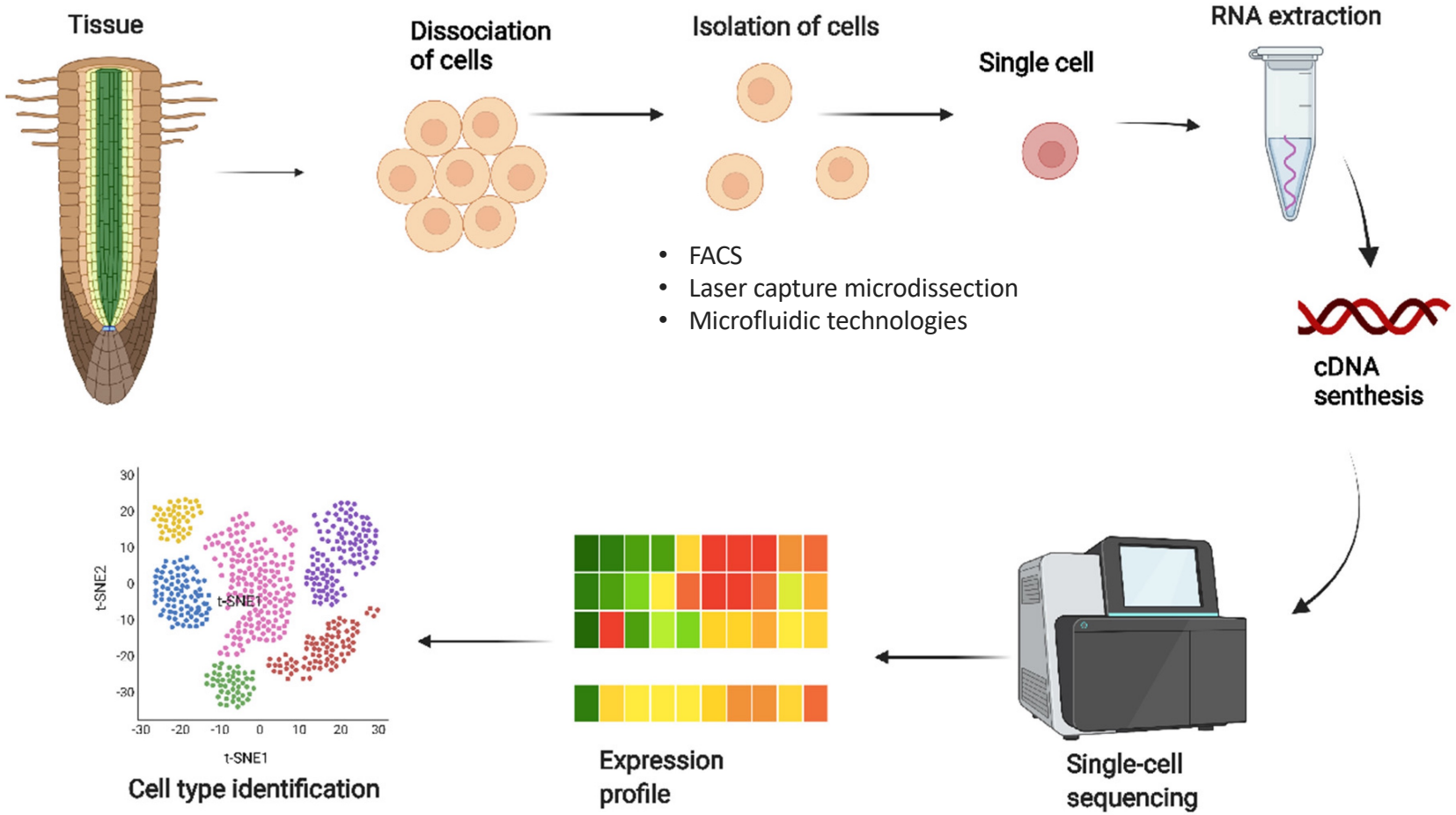


Separate populations
- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

Single cell RNA–Seq (scRNA-Seq)

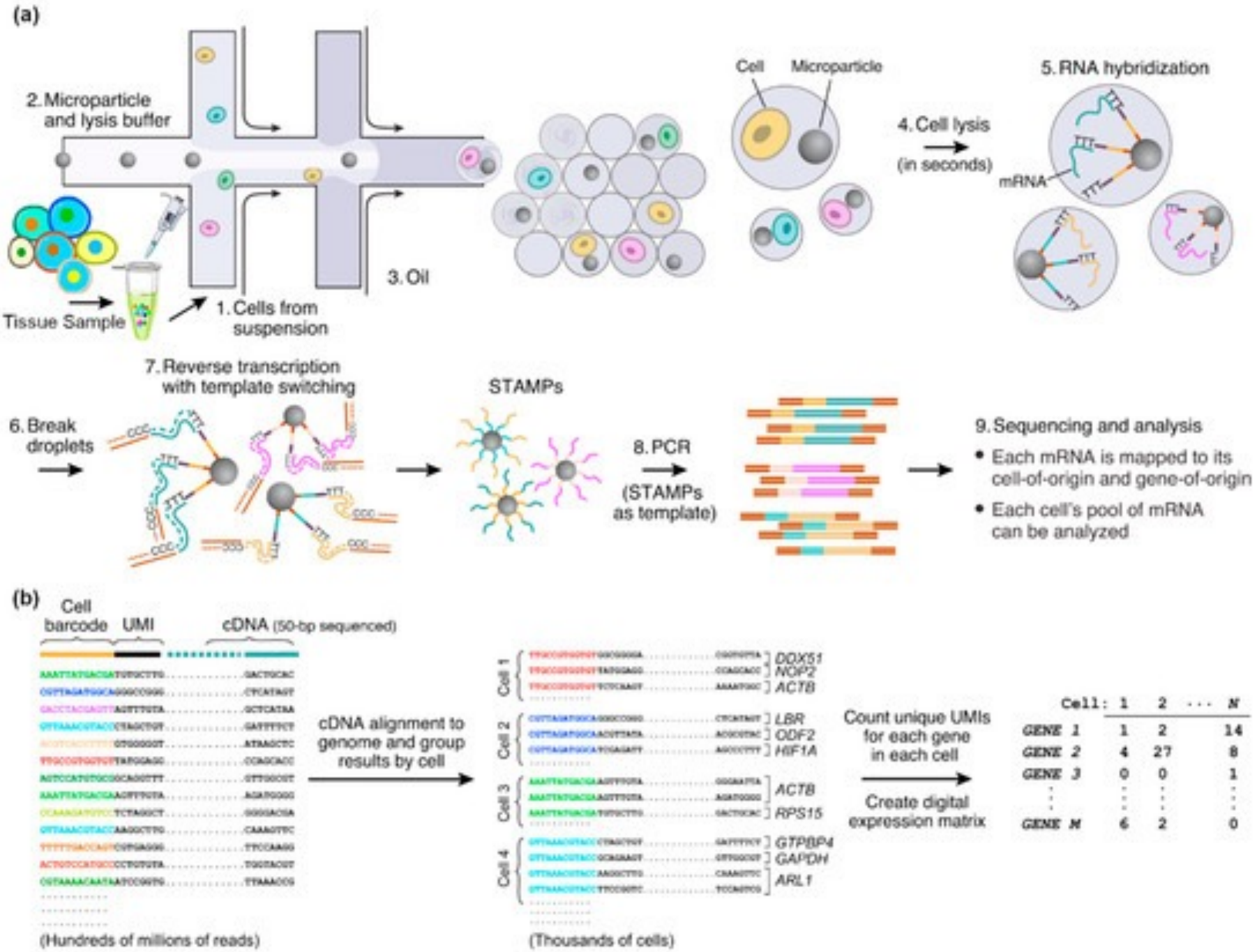
- A new technology, first publication by (Tang et al. 2009).
- Did not gain widespread popularity until ~2014/15 when new protocols and lower sequencing costs made it more accessible.
- Measures the distribution of expression levels for each gene across a population of cells.
- Currently there are several different protocols in use and there are also commercial platforms available.
- Several computational analysis methods from bulk RNA-seq can be used.

Single cell RNA–Seq (scRNA-Seq)

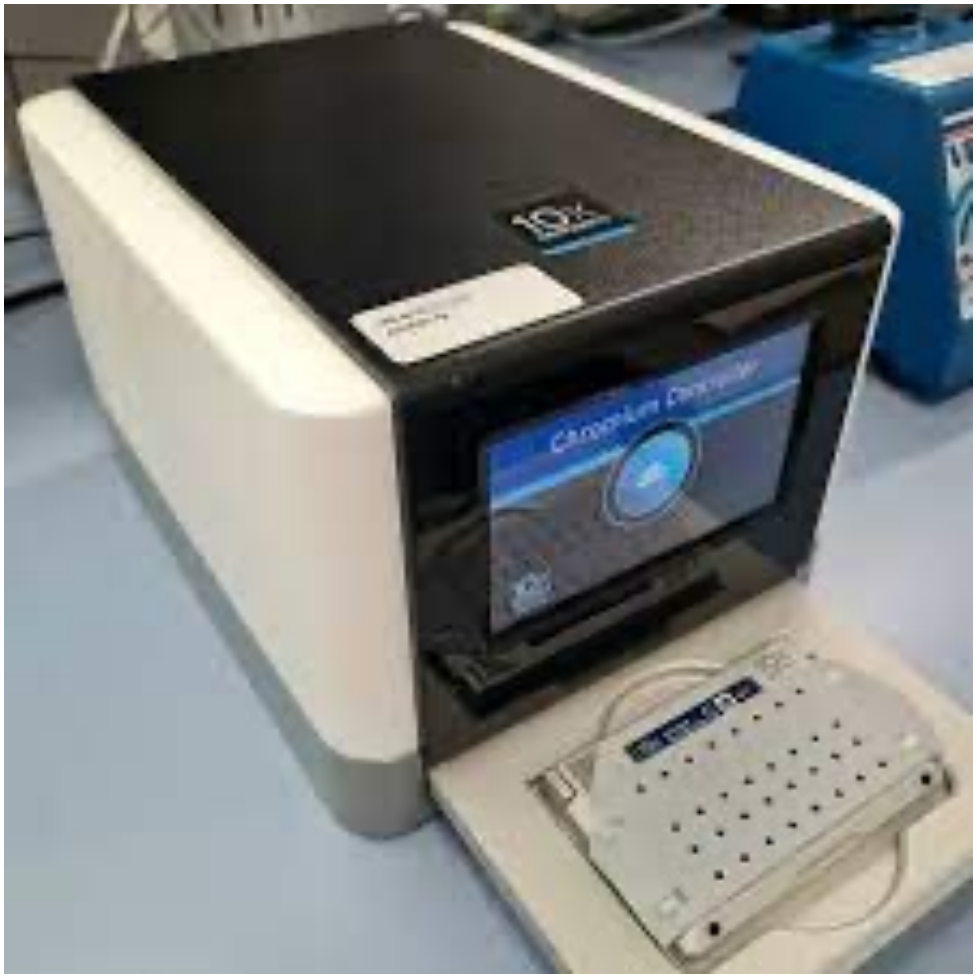


- FACS
- Laser capture microdissection
- Microfluidic technologies

Single cell RNA–Seq (scRNA-Seq)-*Drop-seq*

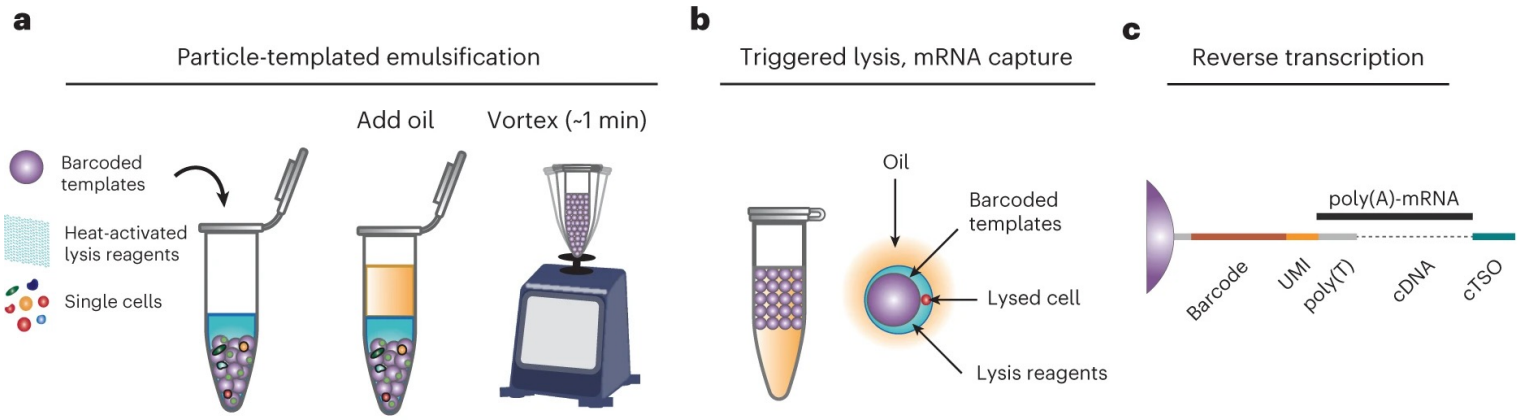


Single cell RNA–Seq (scRNA-Seq)-*Drop-seq*



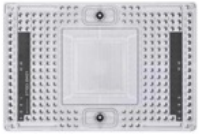
Single cell RNA–Seq (scRNA-Seq)
Particle-templated instant partition sequencing (PIP-seq)

PIP-seq and Drop-seq are both high-throughput methods for single-cell RNA sequencing that use microfluidic droplets to isolate and barcode individual cells. The core principle involves encapsulating a single cell and a barcoded bead in a droplet to tag the cell's entire mRNA content. The key difference is that PIP-seq uses an optimized microfluidic design for instant partitioning, allowing for faster cell lysis and mRNA capture immediately upon encapsulation.

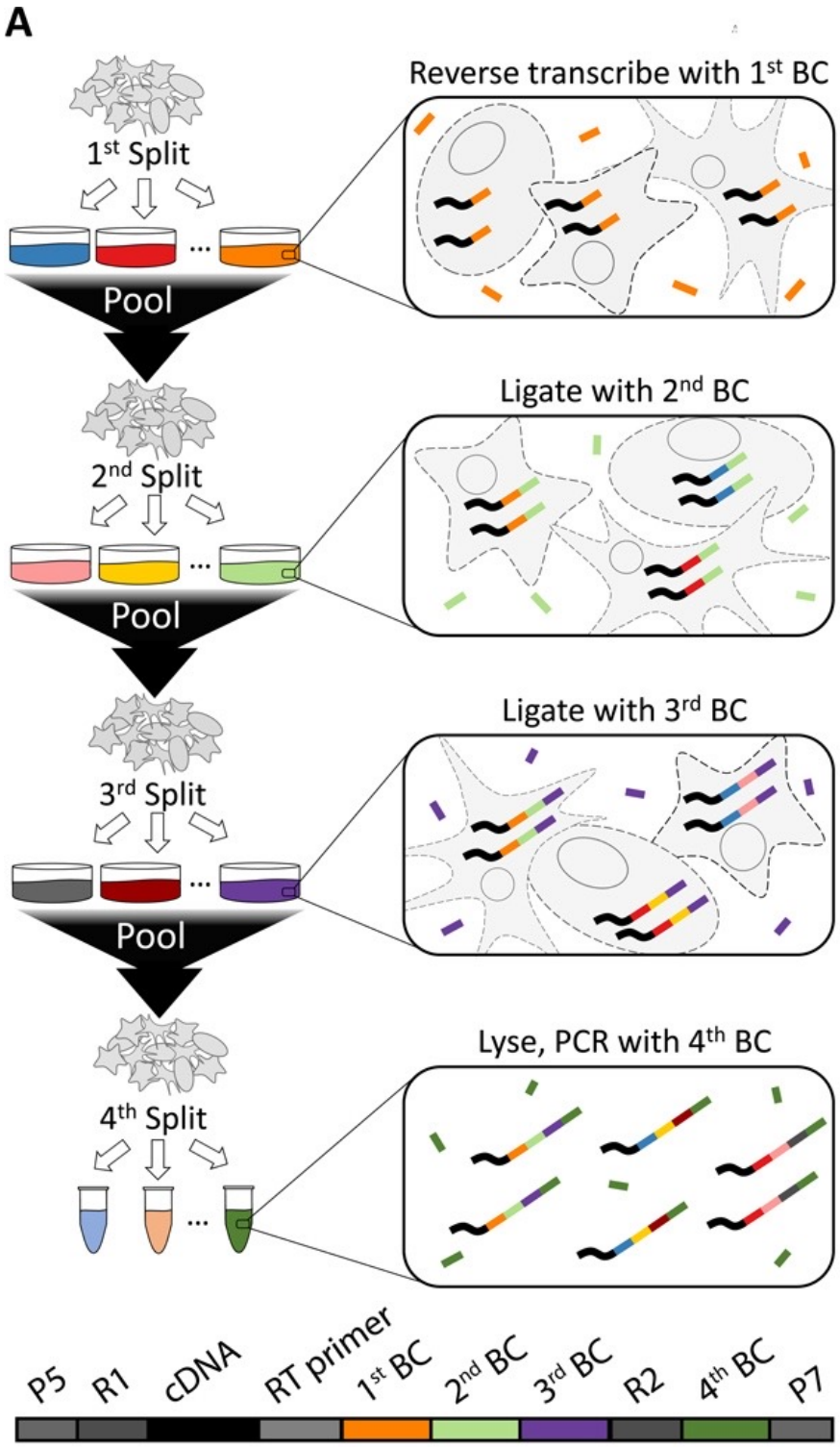


(no instrument required)

SPLiT-seq
Split-pool barcoding

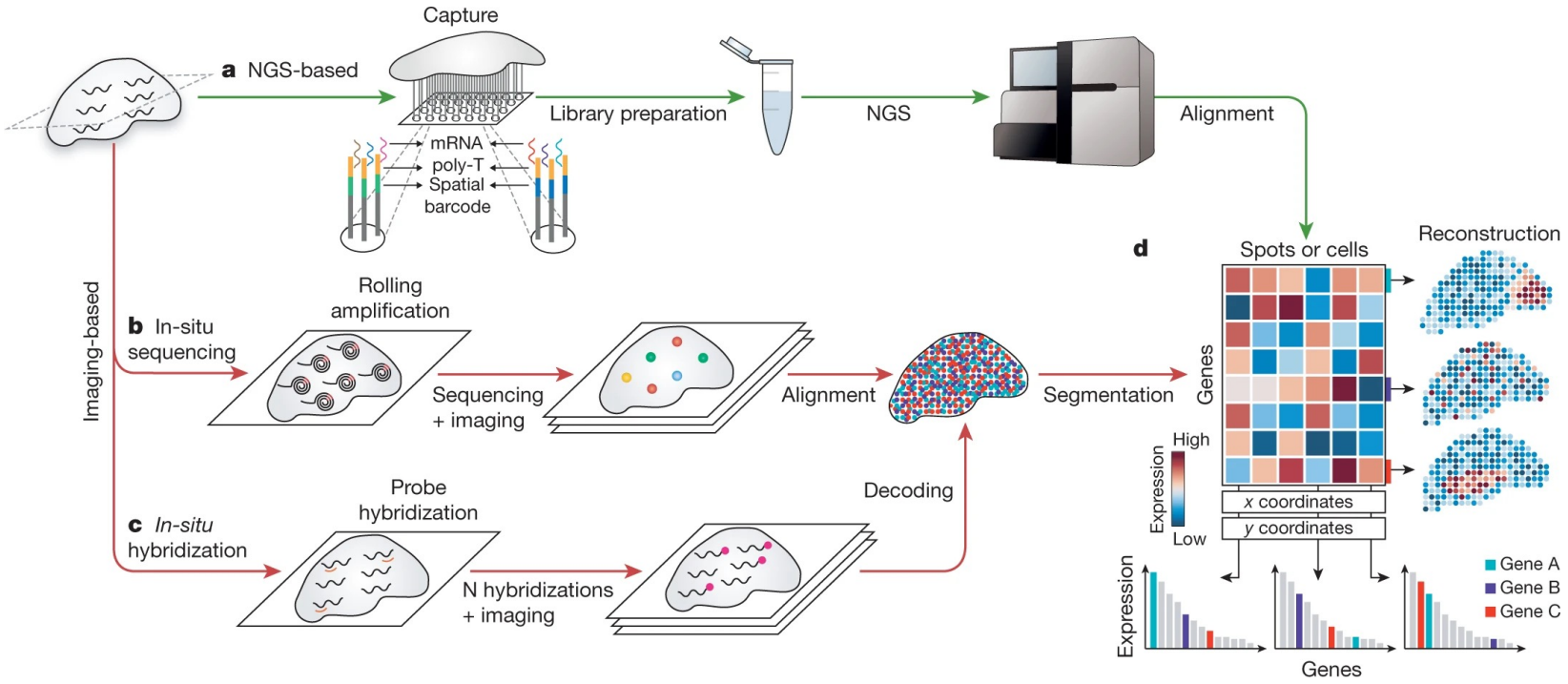


- Time flexibility – single experiment for samples collected on different dates (up to 6months storage)
- No instrument required for experiment. Computational pipeline available
- Up to 48 samples / 100k cells in total – kit has to be used at once
- Retail price of $9,800 per kit (+fixation kits)
- Doublet rate of 0.27% per 1000 cells (3.4% per library)
- No 3’/5’ bias – random hexamers method
- Median genes detection of about 12,000 genes
- Works with any species, any sizes of cells/nuclei & results in lower background noise



Spatial transcriptomics

Spatial transcriptomics technologies are primarily categorized as (1) next-generation sequencing (NGS)-based, encoding positional information onto transcripts before next-generation sequencing; and (2) imaging-based approaches, comprising in situ sequencing (ISS)-based methods—in which transcripts are amplified and sequenced in the tissue—and ISH-based methods—in which imaging probes are sequentially hybridized in the tissue



Rao et al., 2021 Nature