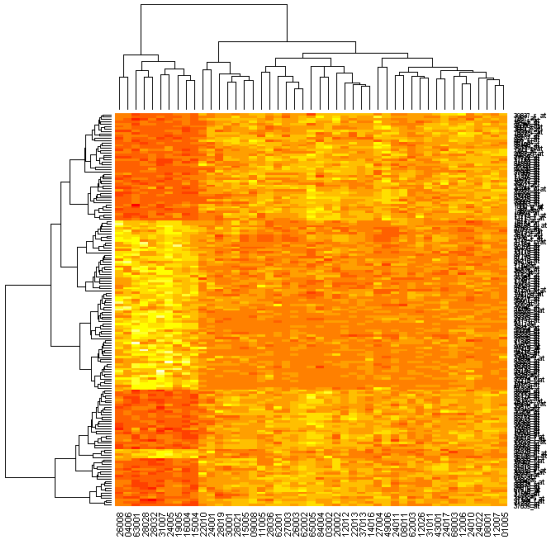
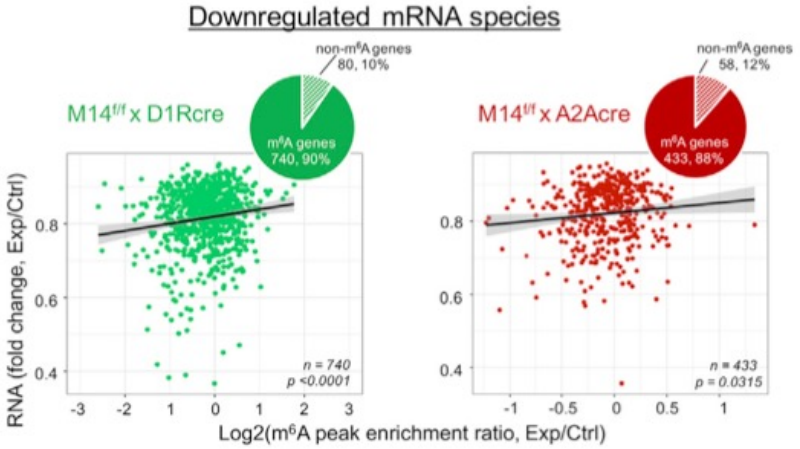
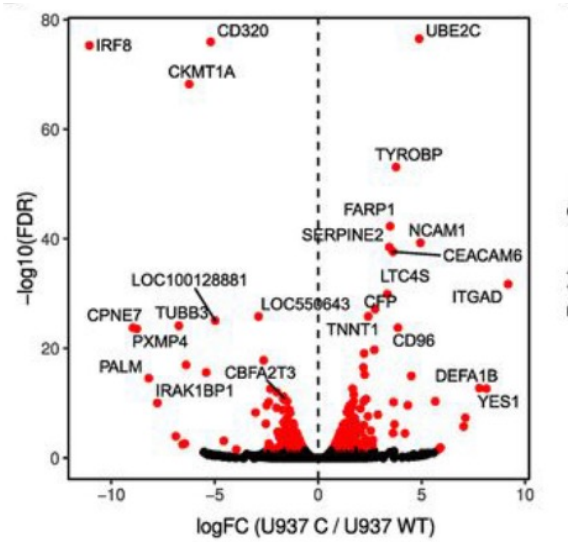
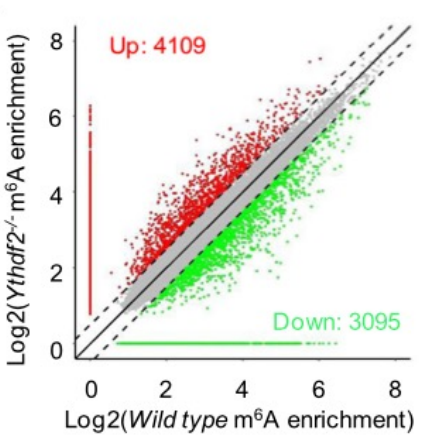
Next generation sequencing









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

The Genomic Era

- ***Genomics*** is the study of genomes, including large chromosomal segments containing many genes.

- The *initial phase of genomics* aims to map and sequence an initial set of entire genomes.

- *Functional genomics* aims to deduce information about the function of DNA sequences.

Questions about the genome

- Obtaining a genome sequence is a one step towards understanding biological processes

- Questions that follow from the genome are:
  - What is transcribed?
  - Where do proteins bind?
  - What is the epigenetic status?

- In other words, how does it work?

Human genome

General stats

Total No of Genes	60609	Total No of Transcripts	227462
Protein-coding genes	19965	Protein-coding transcripts	83986
Long non-coding RNA genes	17910	- full length protein-coding	57935
Small non-coding RNA genes	7576	- partial length protein-coding	26051
Pseudogenes	14749	Nonsense mediated decay transcripts	15811
- processed pseudogenes	10668	Long non-coding RNA loci transcripts	48351
- unprocessed pseudogenes	3556		
- unitary pseudogenes	228		
- polymorphic pseudogenes	42	Total No of distinct translations	62256
- pseudogenes	18	Genes that have more than one distinct translations	13734
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments	408		
- pseudogenes	237		

Long non coding RNA > 200 nt
Short non coding RNA < 200 nt

Mouse genome

General stats

Total No of Genes	55335
Protein-coding genes	21849
Long non-coding RNA genes	13201
Small non-coding RNA genes	6108
Pseudogenes	13681
- processed pseudogenes	10298
- unprocessed pseudogenes	2955
- unitary pseudogenes	74
- polymorphic pseudogenes	88
- pseudogenes	60
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments	494
- pseudogenes	206

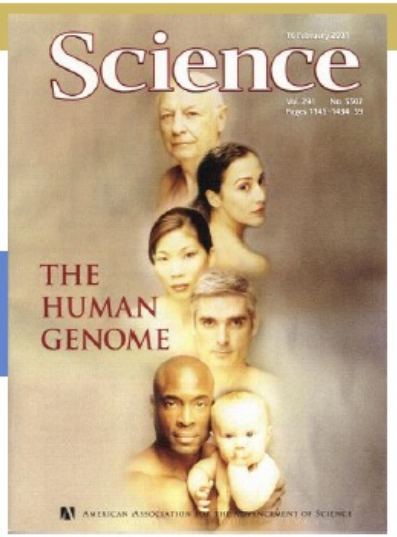
Total No of Transcripts	142351
Protein-coding transcripts	59188
- full length protein-coding	45391
- partial length protein-coding	13797
Nonsense mediated decay transcripts	7200
Long non-coding RNA loci transcripts	18800
Total No of distinct translations	45400
Genes that have more than one distinct translations	10990

Long non coding RNA > 200 nt
Short non coding RNA < 200 nt

HUMAN GENOME PROJECTS

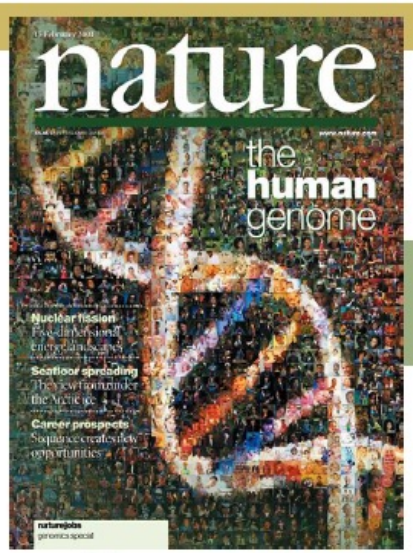
(performed by Sanger sequencing)

Celera Genomics



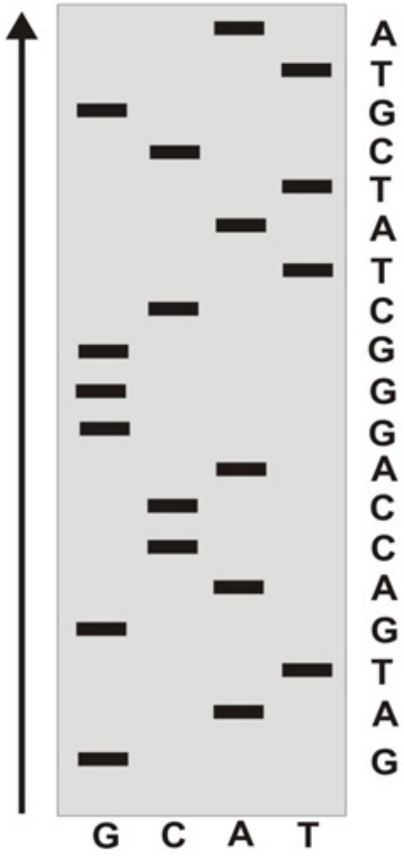
- Private company, founded in 1998
- Beginning of sequencing in 1998
- Patented sequence: **no public access to data**
- Estimated cost: 300 Million $

International Consortium



- 20 groups from USA, UK, Japan, France,
- Germany and China
- Over 1000 scientists involved
- Beginning of sequencing in 1991
- Open-access sequence: **immediate and free release of data**
- Estimated cost: 450 Million – 3 Billion $

SANGER METHOD FOR DNA SEQUENCING



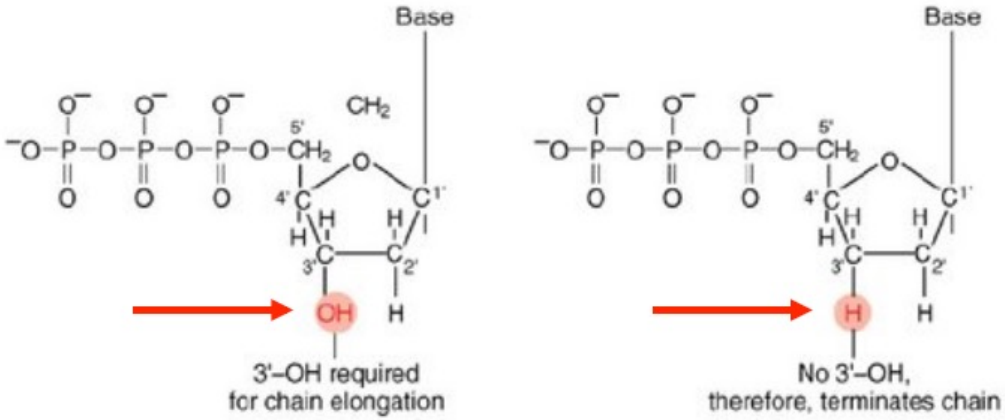
SANGER METHOD FOR DNA SEQUENCING

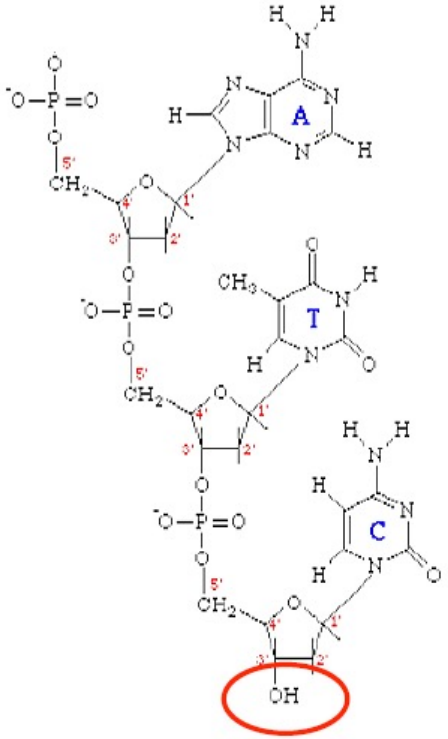
History of Sequencing: Sanger method for DNA sequencing

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

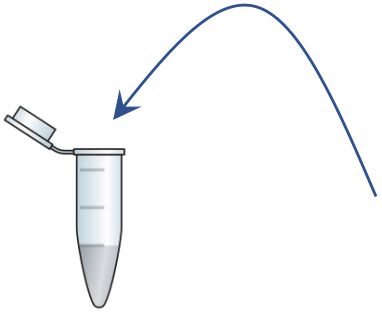
- "**Sanger Sequencing**" developed by Fred Sanger *et al.* in the mid 1970's
- Uses dideoxynucleotides for "chain termination", generating fragments of different lengths ending in ddATP, ddGTP, ddCTP or ddTTP



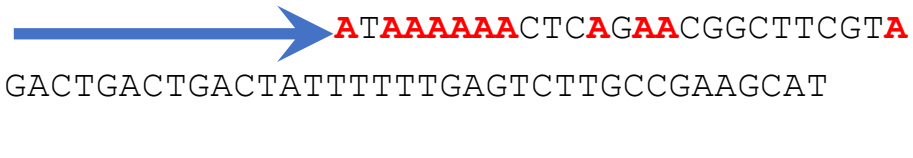


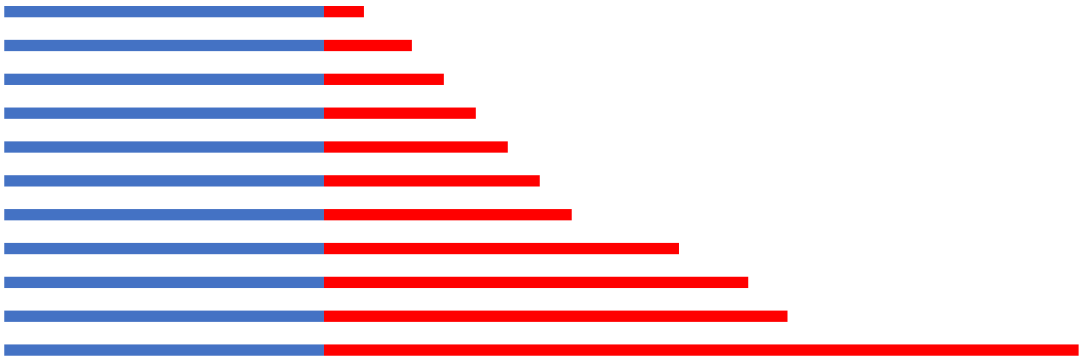
http://openwetware.org/wiki/BE.109:Bio-material_engineering/Sequence_analysis

SANGER METHOD FOR DNA SEQUENCING



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





Sanger sequencing uses ddNTPs (dideoxynucleotide triphosphates) which do not have a free 3' OH mixed in with dNTPs
Whenever the DNA polymerase incorporates a ddNTP it won't be able to add any other nucleotides
Then gel electrophoresis is used to separate the DNAvv

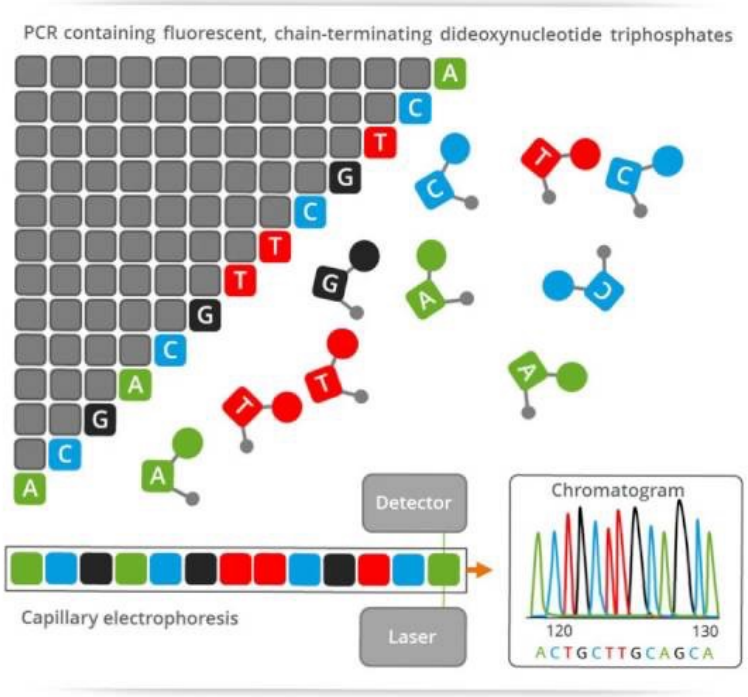
SANGER METHOD FOR DNA SEQUENCING

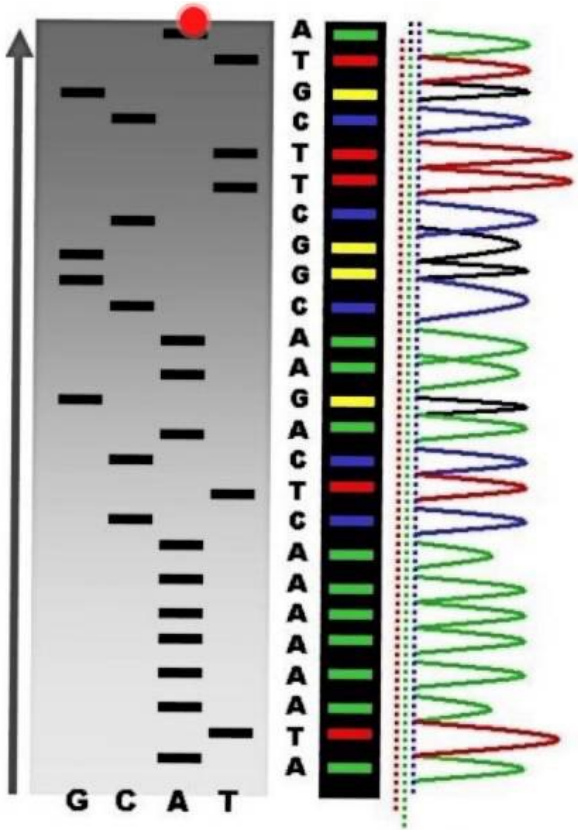
Automated Sequencing

- 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

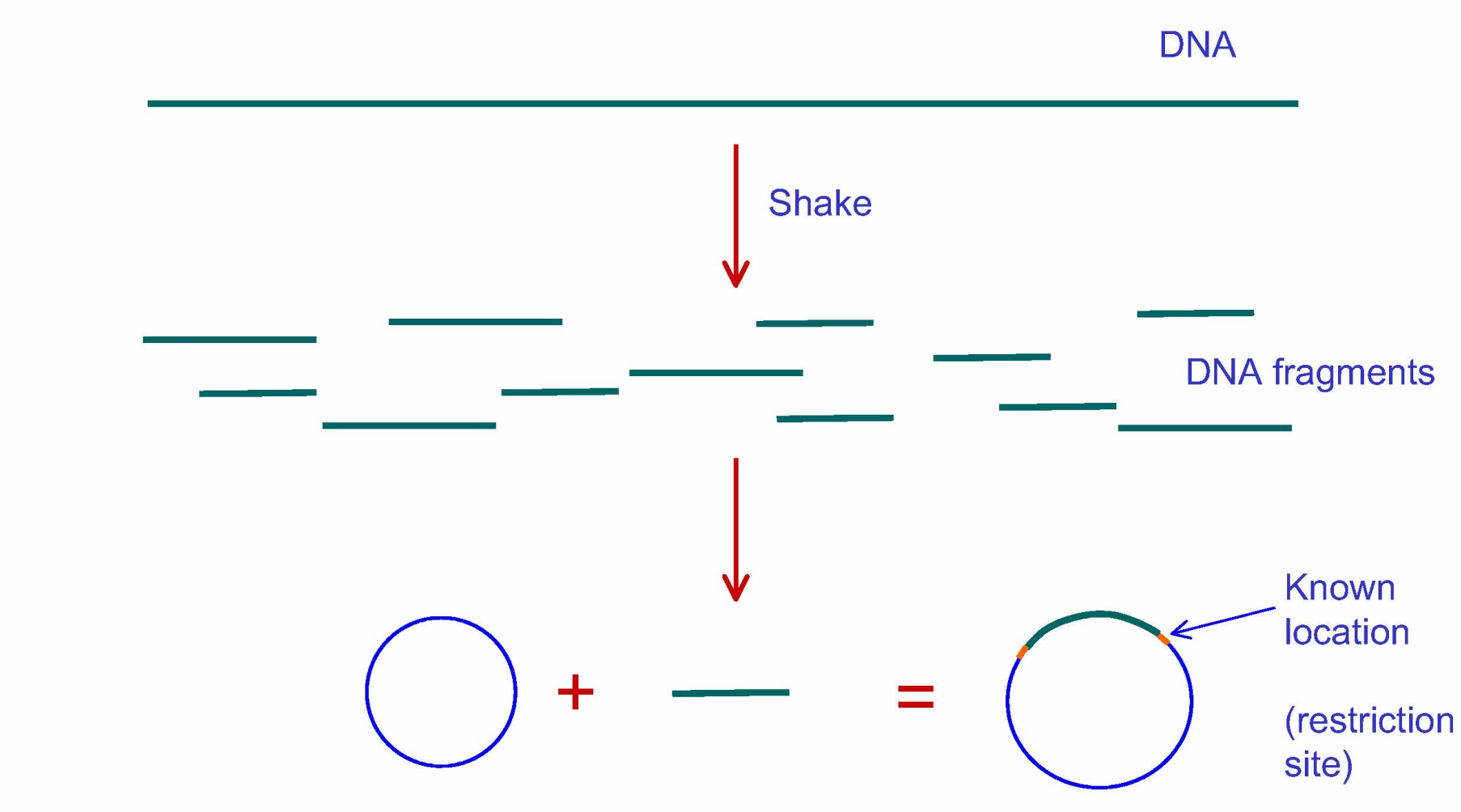
SANGER METHOD FOR DNA SEQUENCING

Automated Sanger sequencing

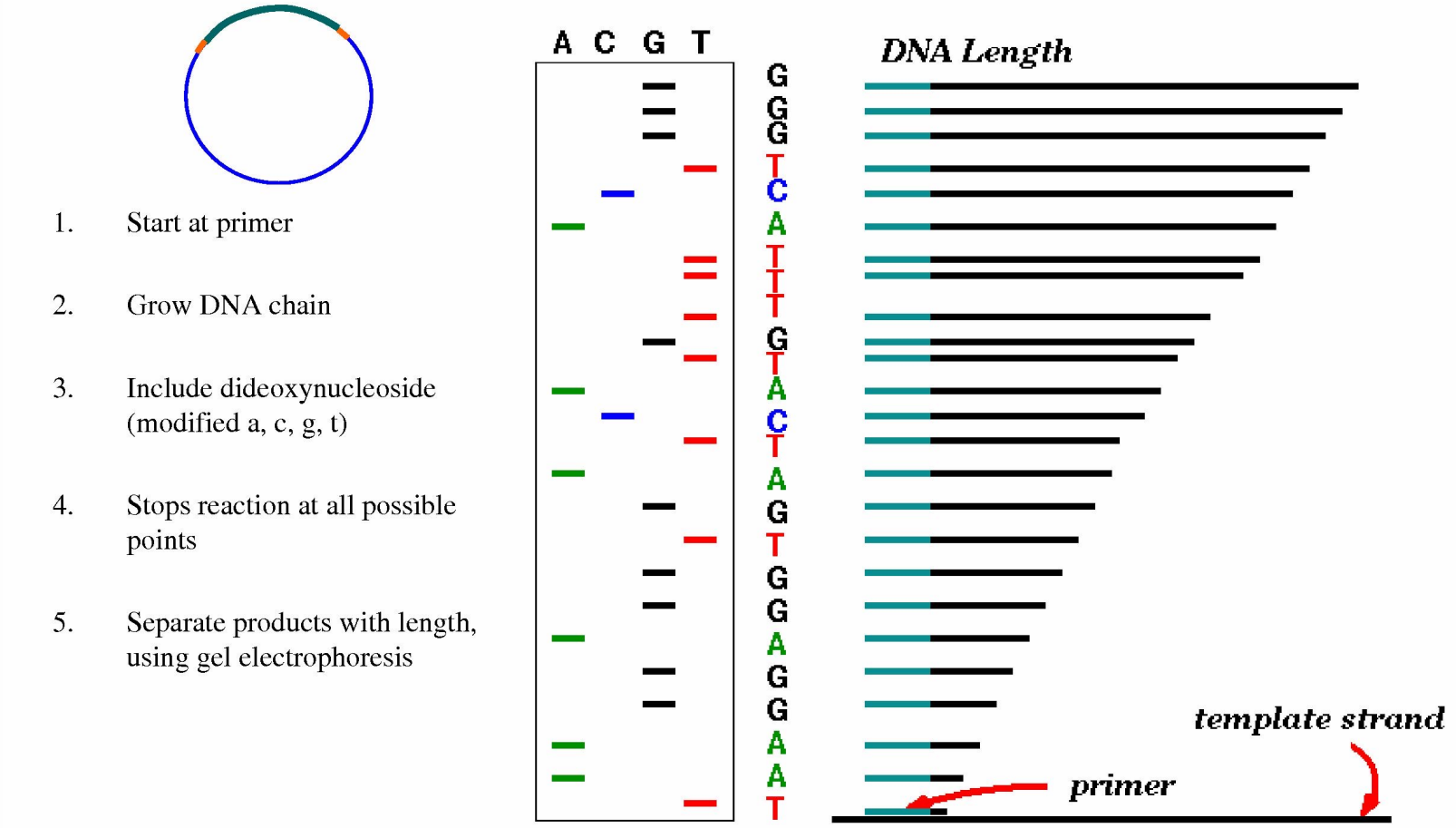




HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



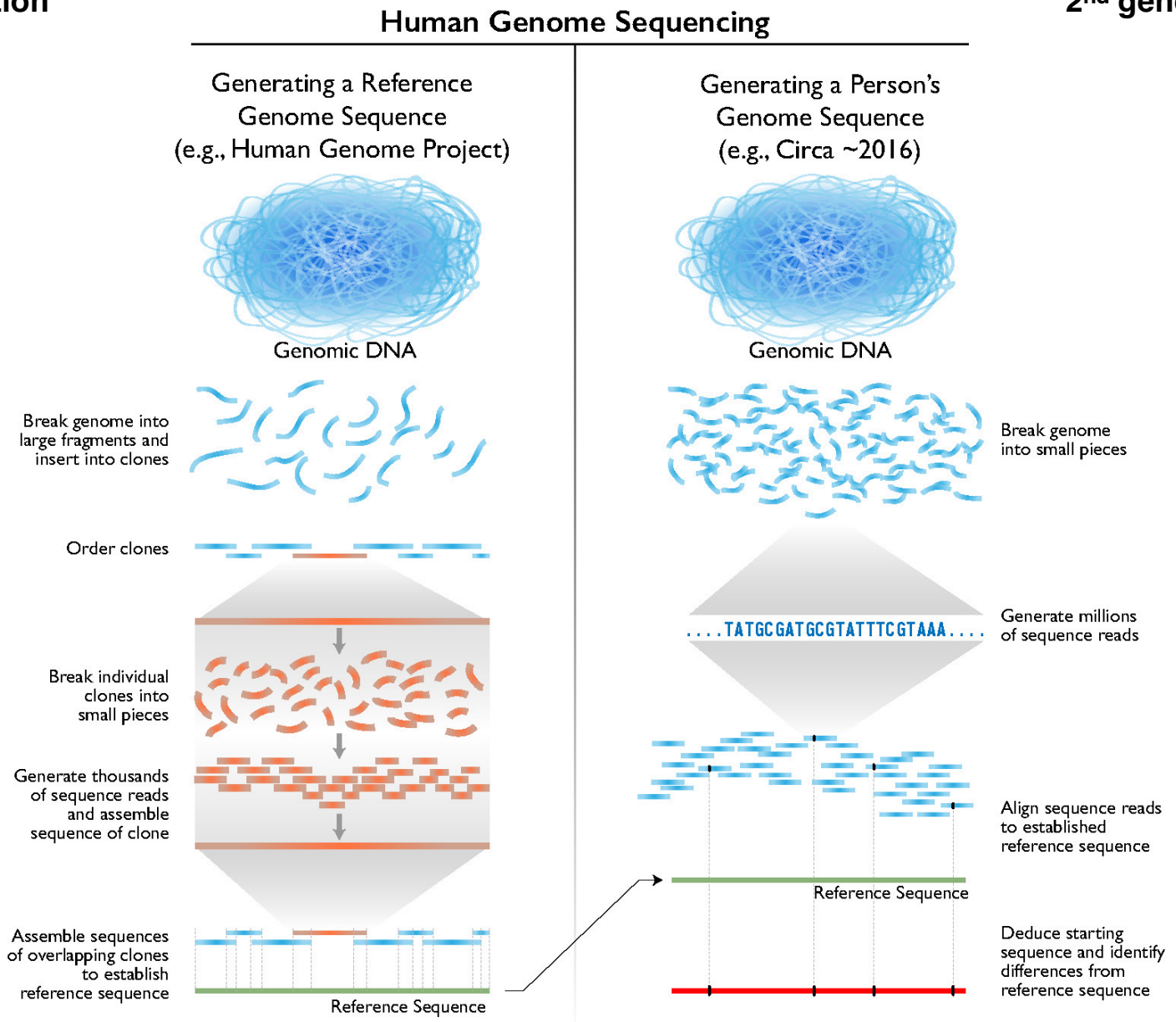
HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



- 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

Sequencing methodologies

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



NEXT GENERATION SEQUENCING

What is it?

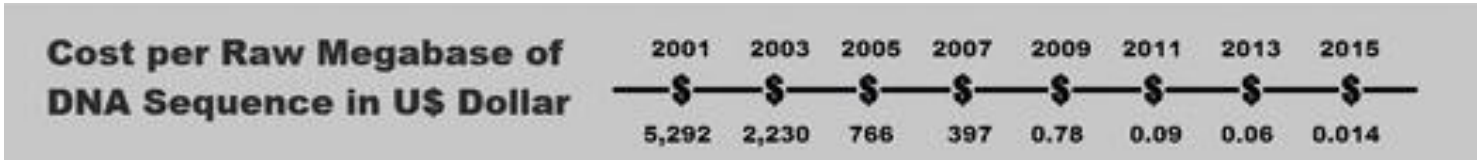
- **Set** of new high throughput technologies:

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

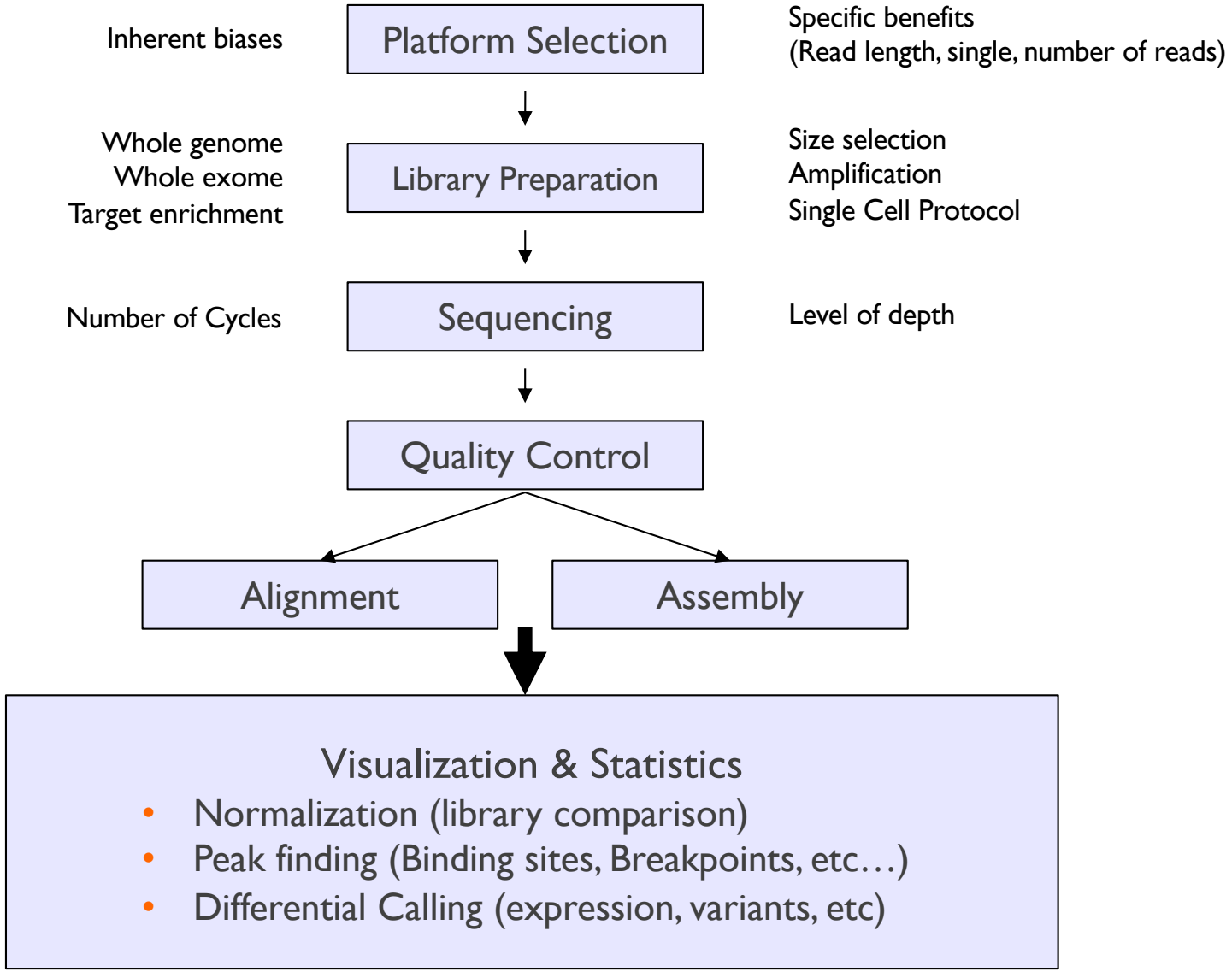
  • Computational power is then used to assemble or align the "reads" to a reference genome, allowing biologists to make comparisons and interpret various biological phenomena

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

SEQUENCING TECHNOLOGIES

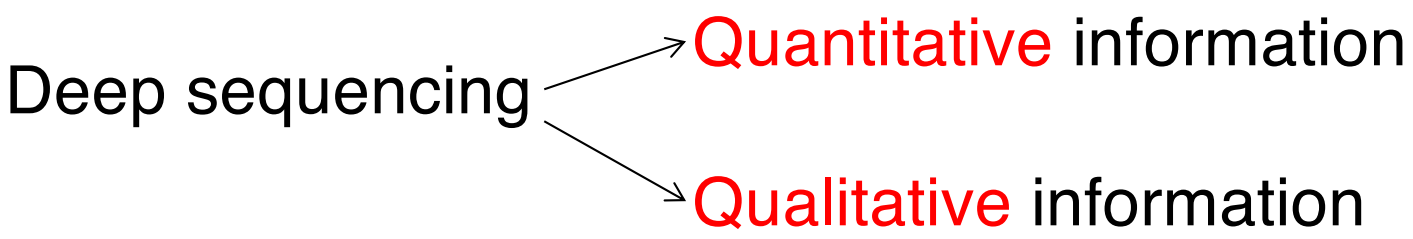


NEXT GENERATION SEQUENCING



NEXT GENERATION SEQUENCING

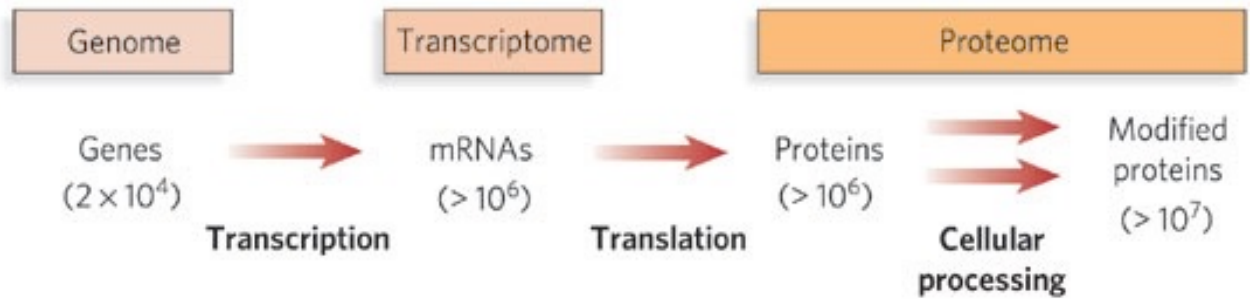
- Mutation and SNP identification or analysis (genome re-sequencing)
- Gene/Disease Linkage (genome re-sequencing)
- Pathogen identification (*de novo* sequence assembly or re-sequencing)
- **Transcriptome analysis (RNAseq)**
- DNA methylation study (medip-seq)
- Chromatin study (ChIPseq)
- Transcription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



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.



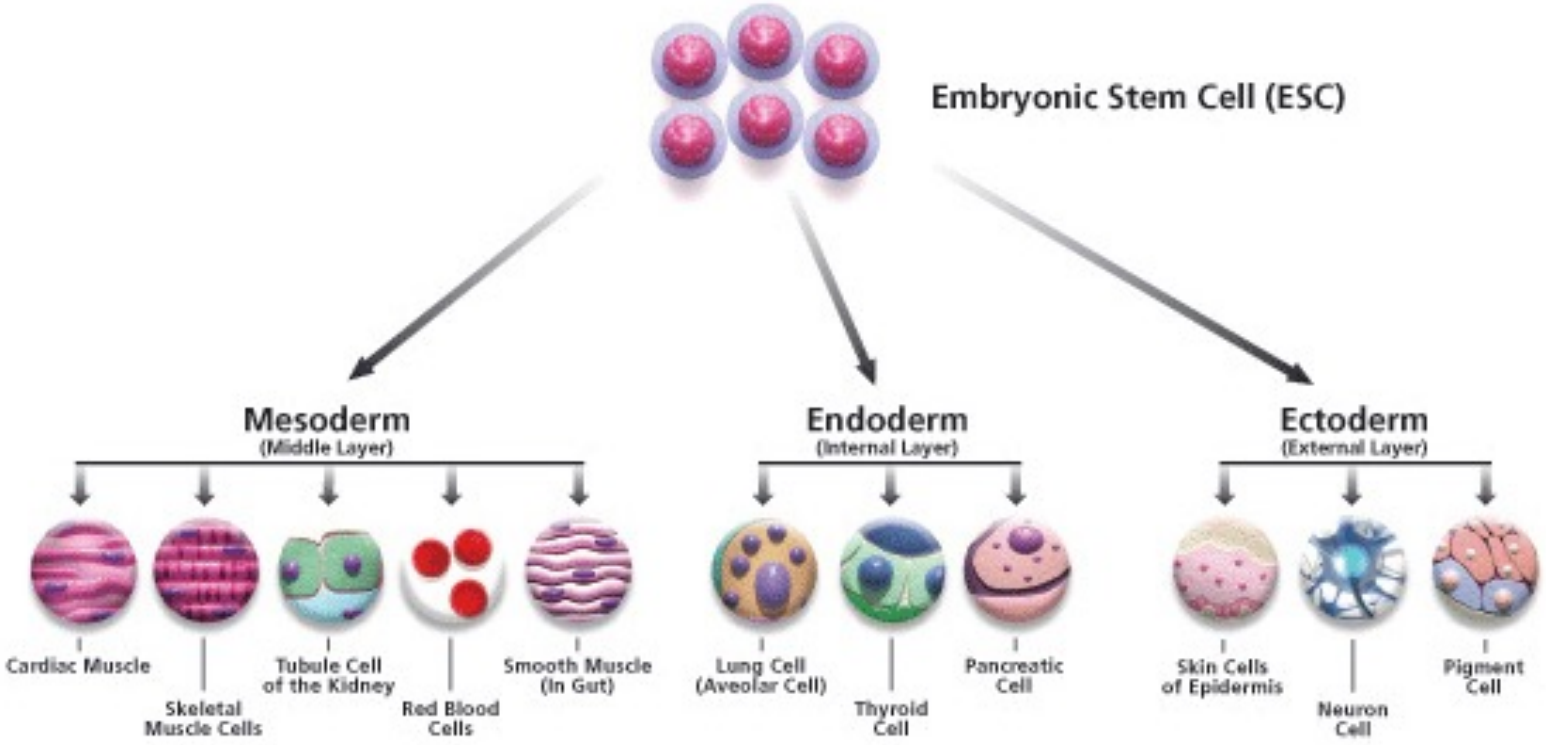
TRANSCRIPTOME ANALYSIS

Issues in the studies on Transcriptome

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

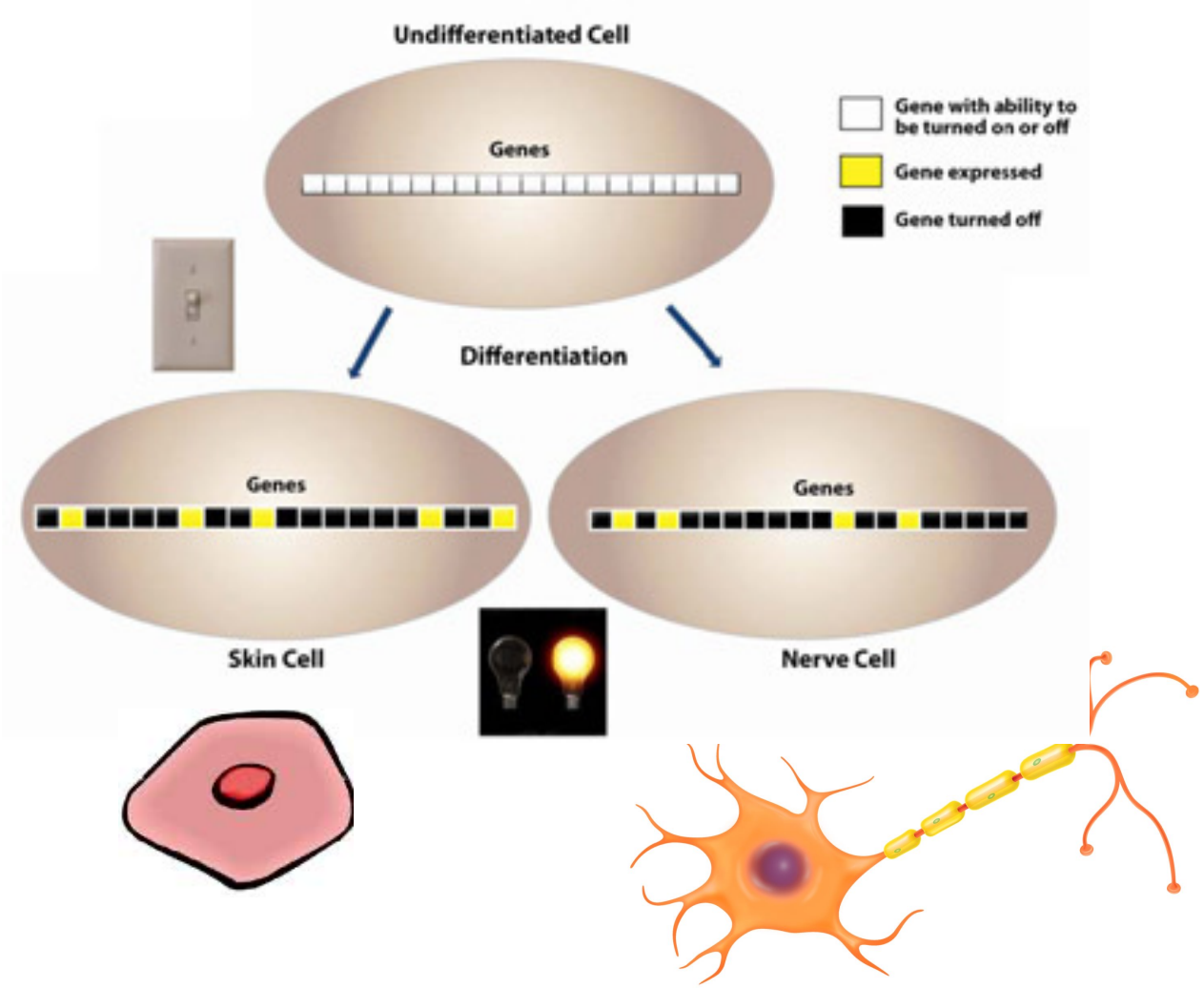
The transcriptome is cell type specific

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.

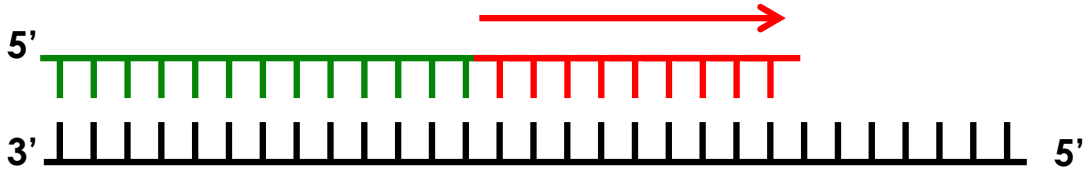


Transcriptome complexity

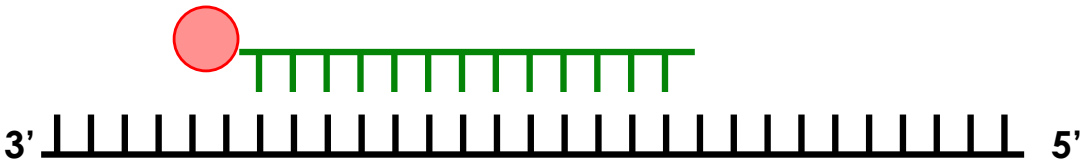
- Transcriptome is substantially more complex than the genome and transcripts may be:
  - Modified
  - Spliced
  - Edited
  - Degraded

TRANSCRIPTOME ANALYSIS

PCR / qPCR /
classic sequencing

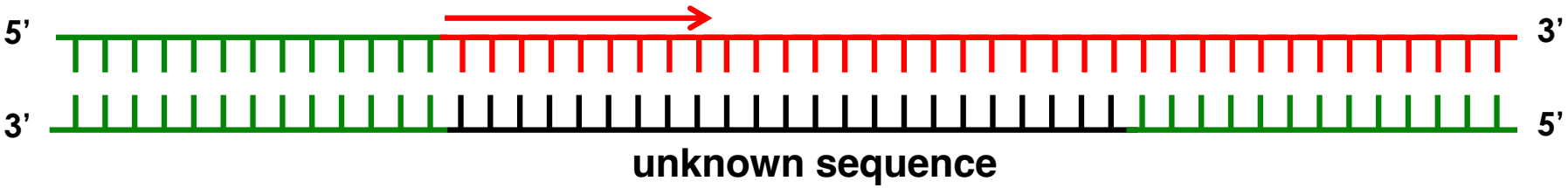


Northern blot



How to detect something that is unknown?

Next-Generation Sequencing (NGS)



RNA- sequencing (RNA-Seq)

What is RNA-seq?

- RNA-seq is essentially **massively parallel sequencing of RNA** (or 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.

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.

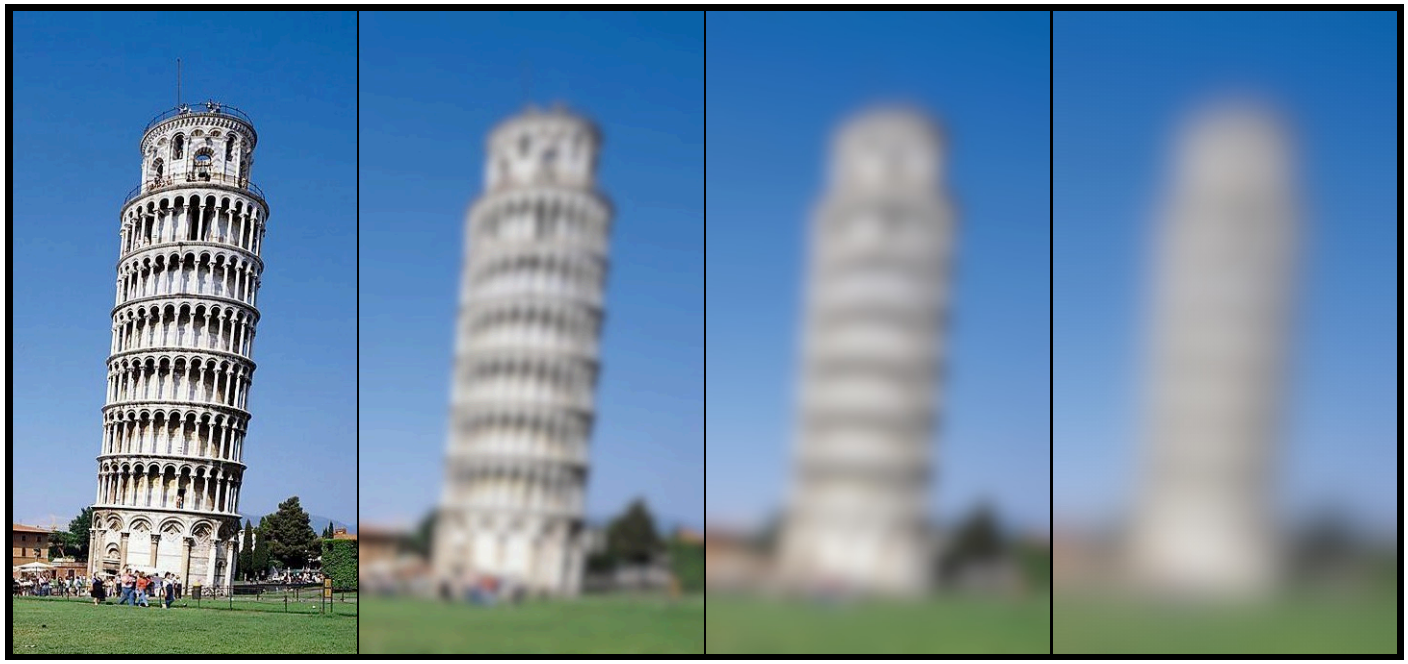
- **Nascent RNA-Seq**

  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

Low resolution

↓

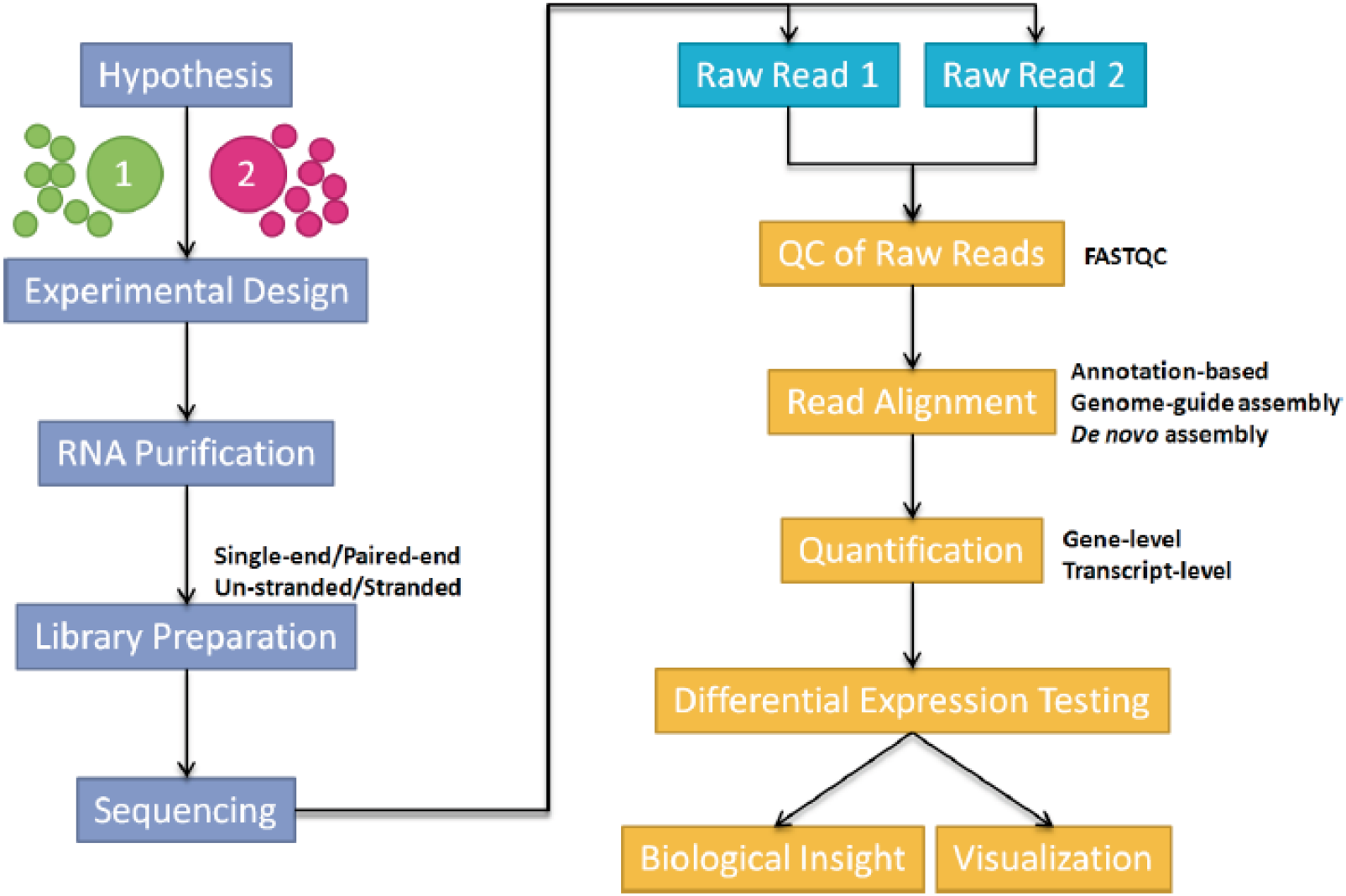
few information

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



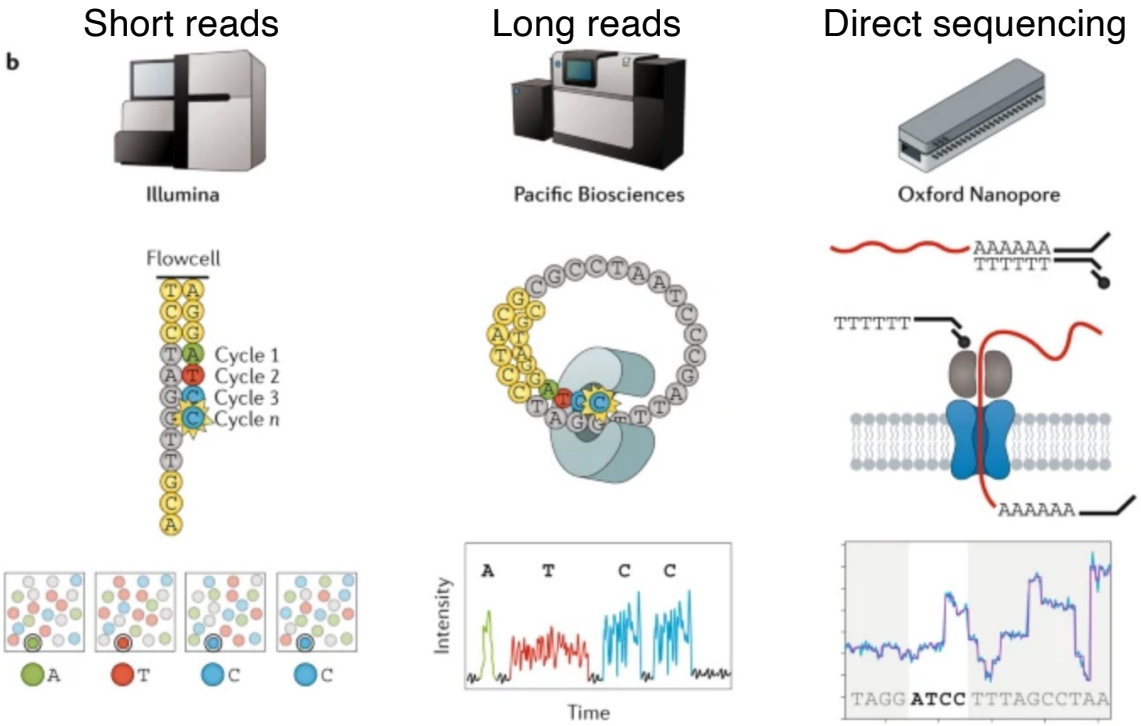
The library preparation methodology is specific to the type of sequencer used.

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.

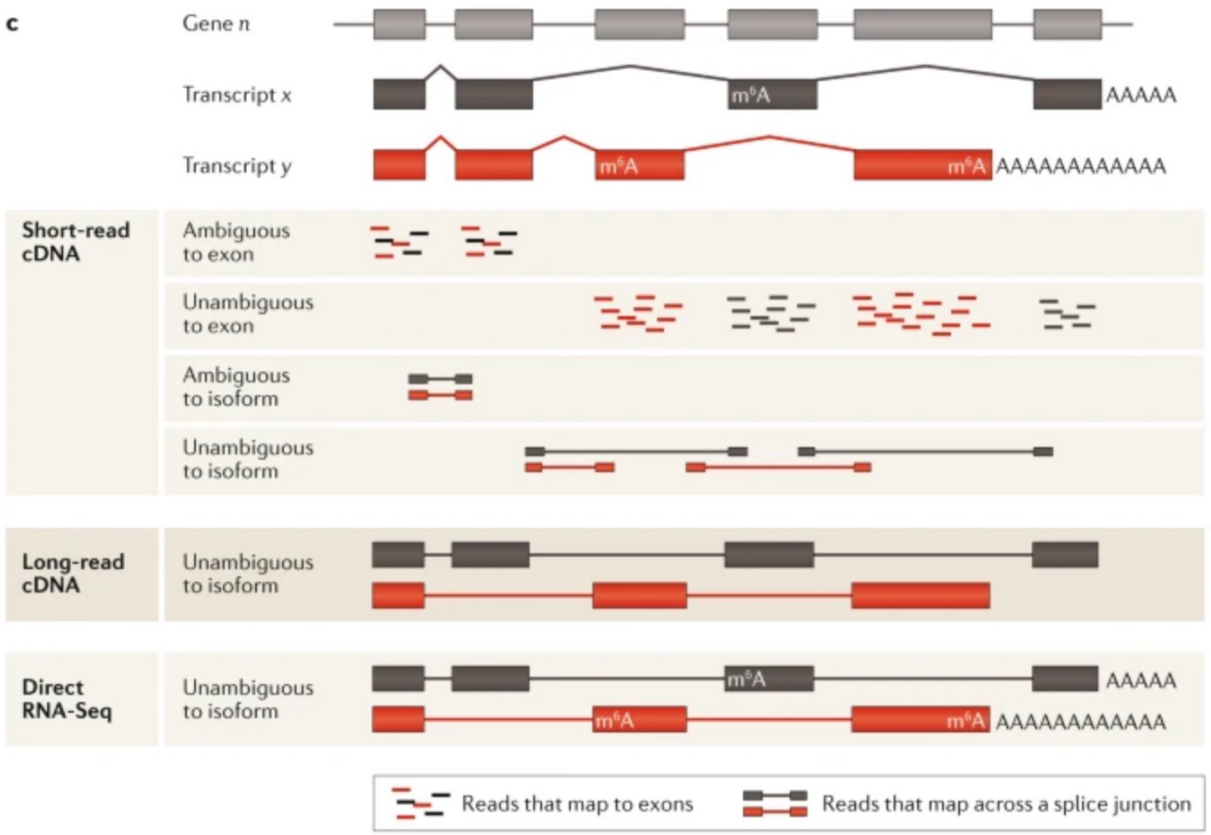
The Pacific Biosciences workflow: after library preparation, individual molecules are loaded into a sequencing chip (SMRT cell), where they bind to a polymerase immobilized at the bottom of a nanowell. As each of the fluorescently labelled nucleotides is incorporated into the growing strand, they fluoresce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

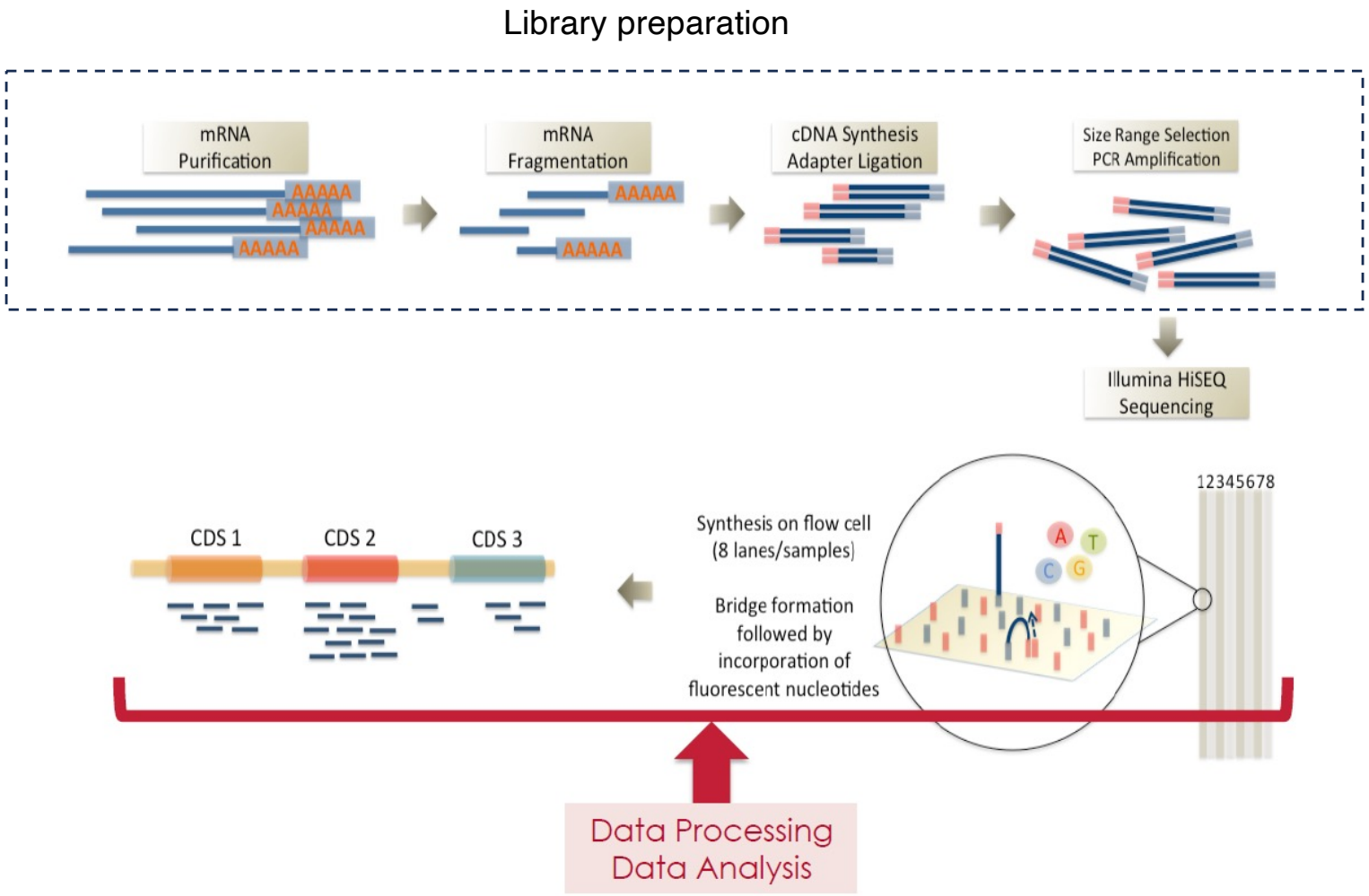


Short-read, long-read and direct RNA-seq analysis.

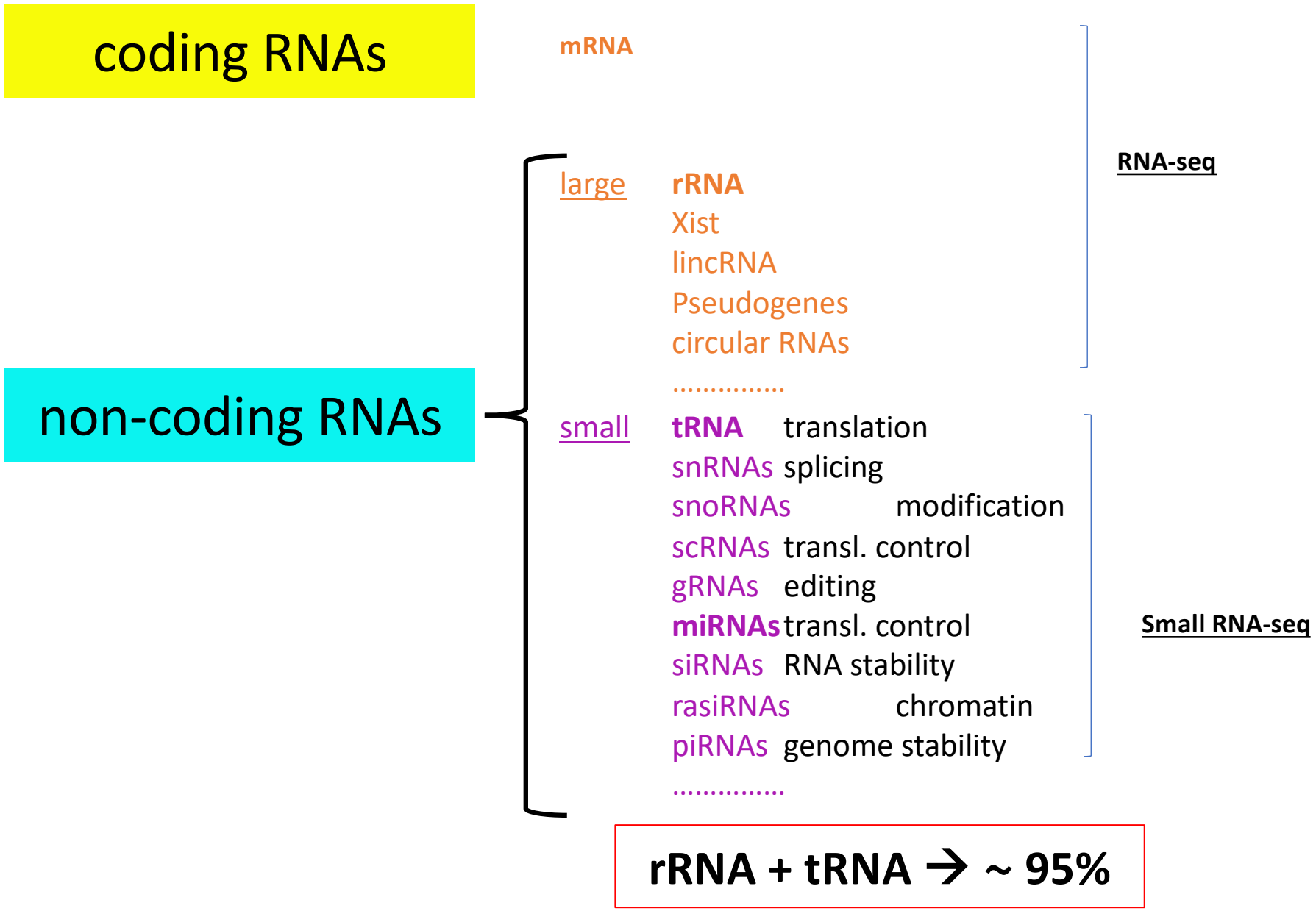
Over 90% of human genes are alternatively spliced to form two or more distinct and expressed isoforms (transcripts *x* and *y*). The complexity of information captured increases from short-read cDNA sequencing, where isoform detection can be compromised by reads that cannot be mapped unambiguously, to 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



RNA-Seq: LIBRARY PREPARATION

Two ways to isolate long RNA molecules:

- **Purify and Fragment mRNA**

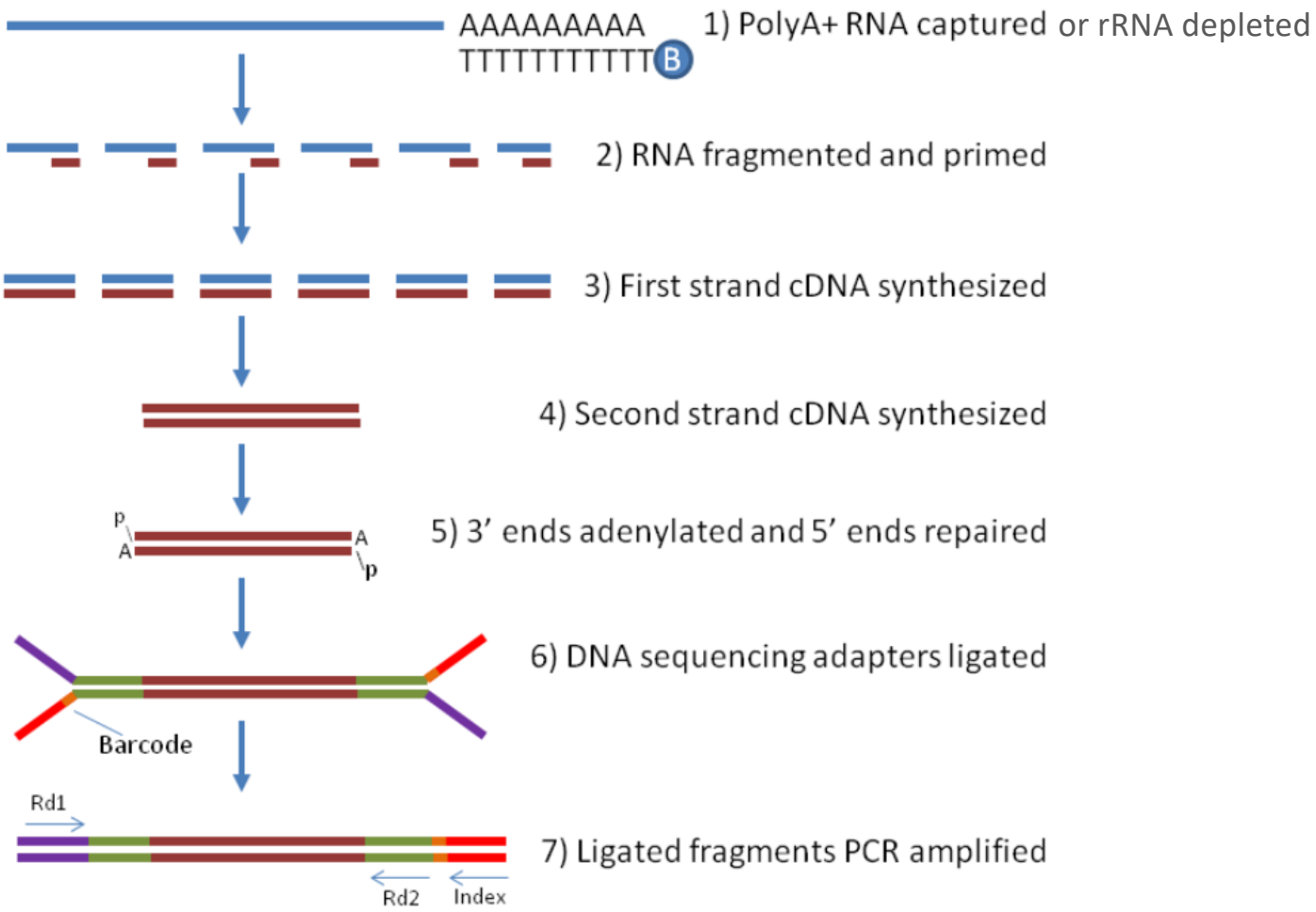
This process purifies the poly-A containing RNA molecules (mainly mRNA) using poly-dT oligo-attached magnetic beads.

- **Remove rRNA**

rRNA is removed using an hybridization-based technique. The method uses a cocktail of biotinylated oligonucleotide-attached magnetic beads that are highly specific to conserved rRNA sequences. These probes are designed to bind directly to the 18S, 28S, 5S, and 5.8S rRNAs.

RNA-Seq: LIBRARY PREPARATION

Example of library preparation: Illumina Truseq

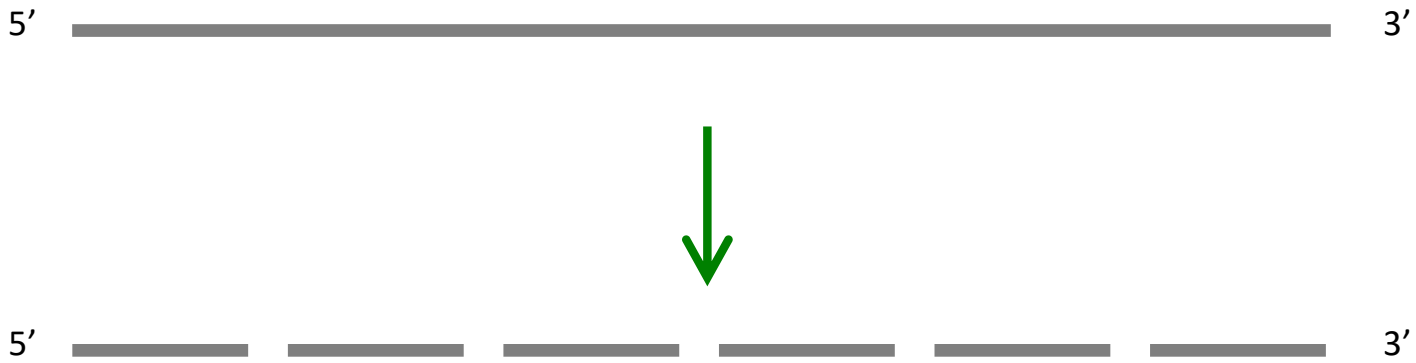


RNA-Seq: LIBRARY PREPARATION

1- RNA purification and selection (polyA+ or polyA-)

2 - RNA fragmentation

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

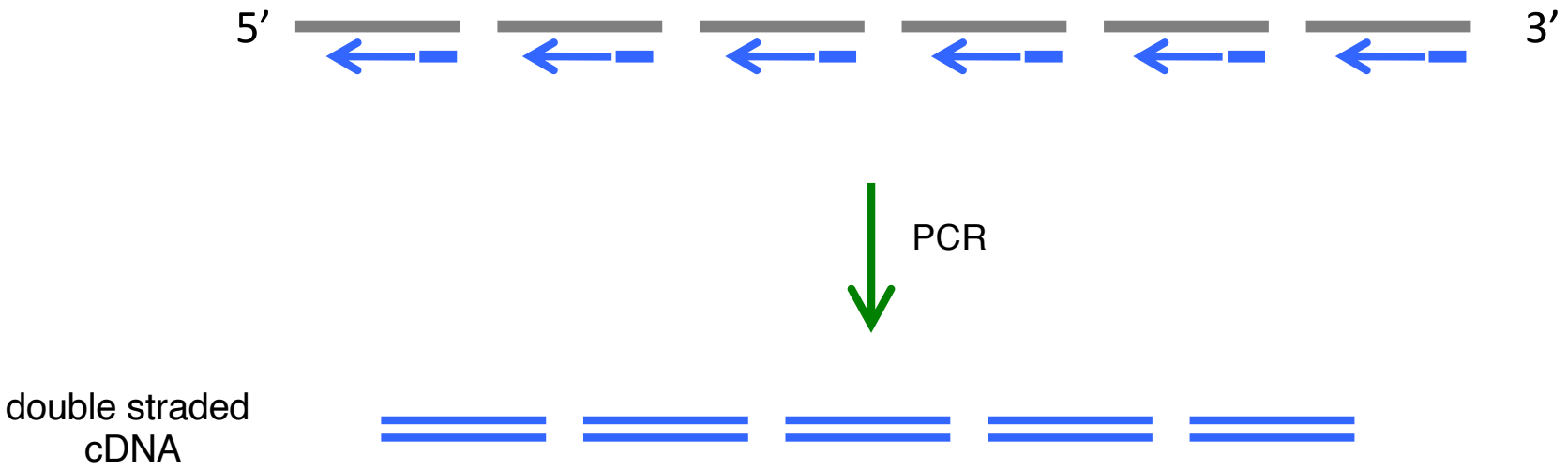


Range of fragments length: **100-300 nt**

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

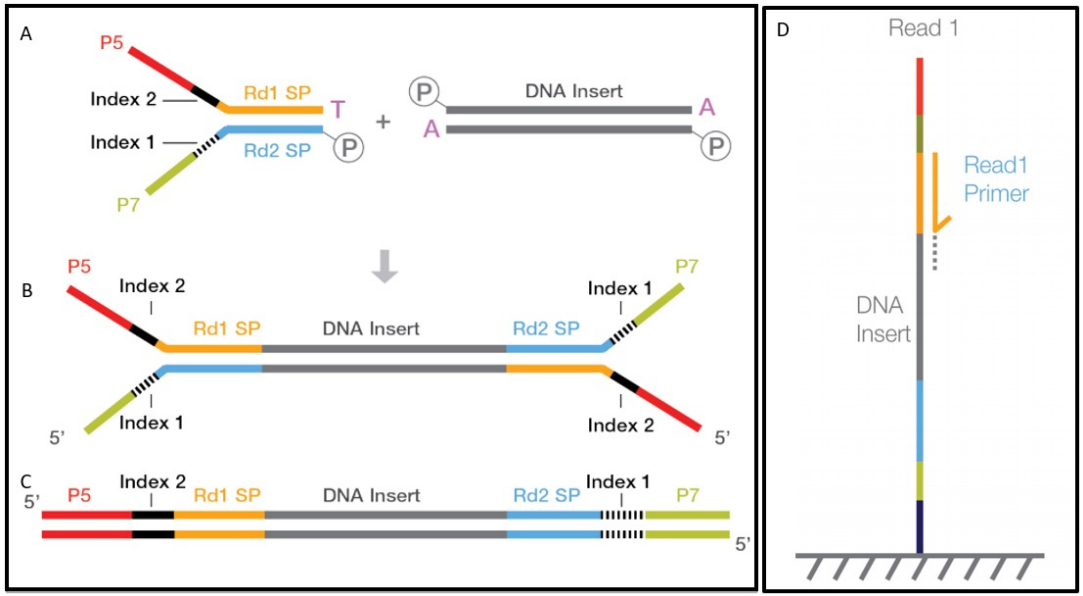
3 – Adapter ligation

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.

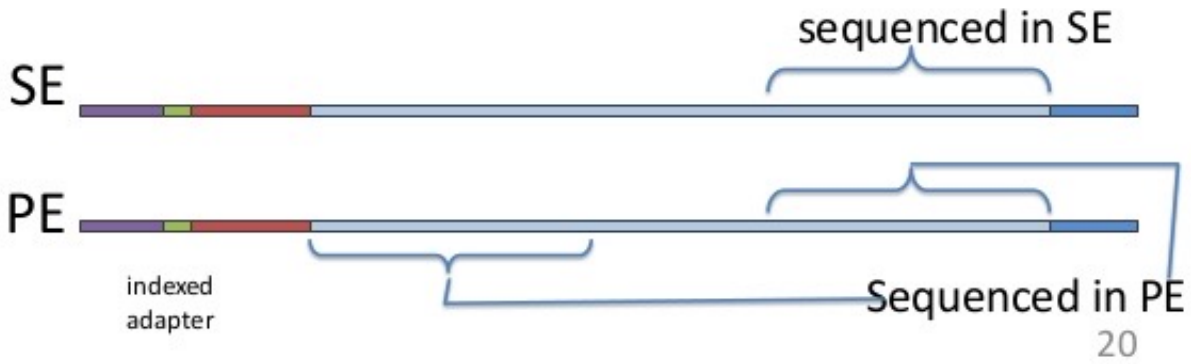


RNA-Seq: LIBRARY PREPARATION

Single-end VS paired-end sequencing

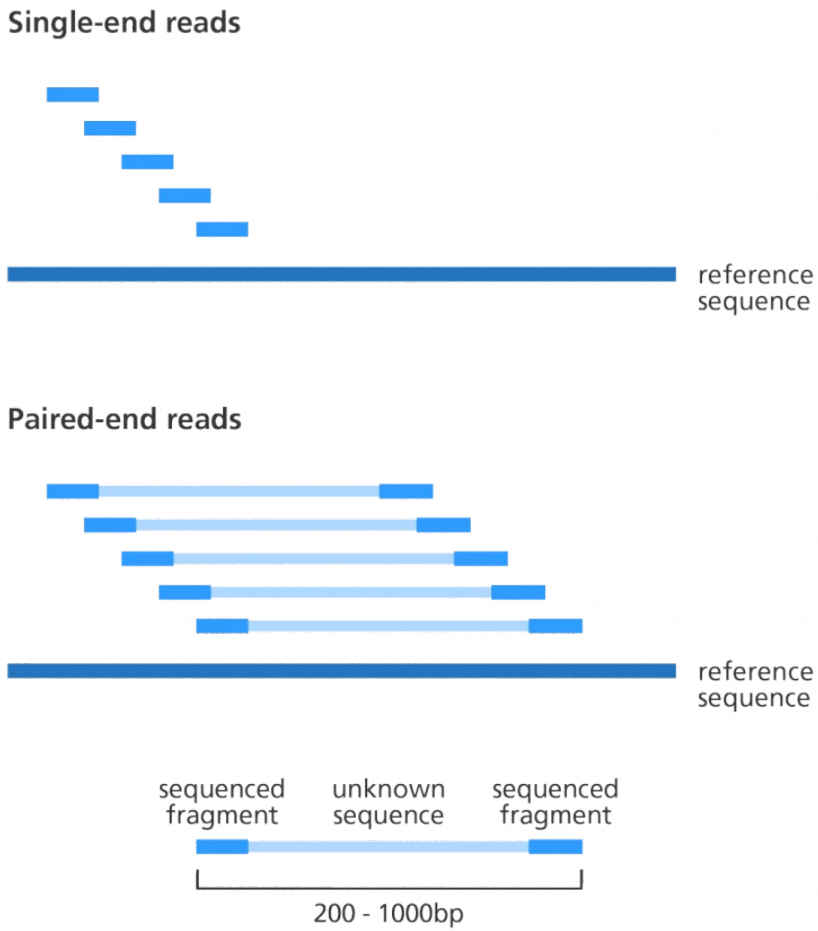
Single-end sequencing (SE), involves sequecing of the fragment from only one end.

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

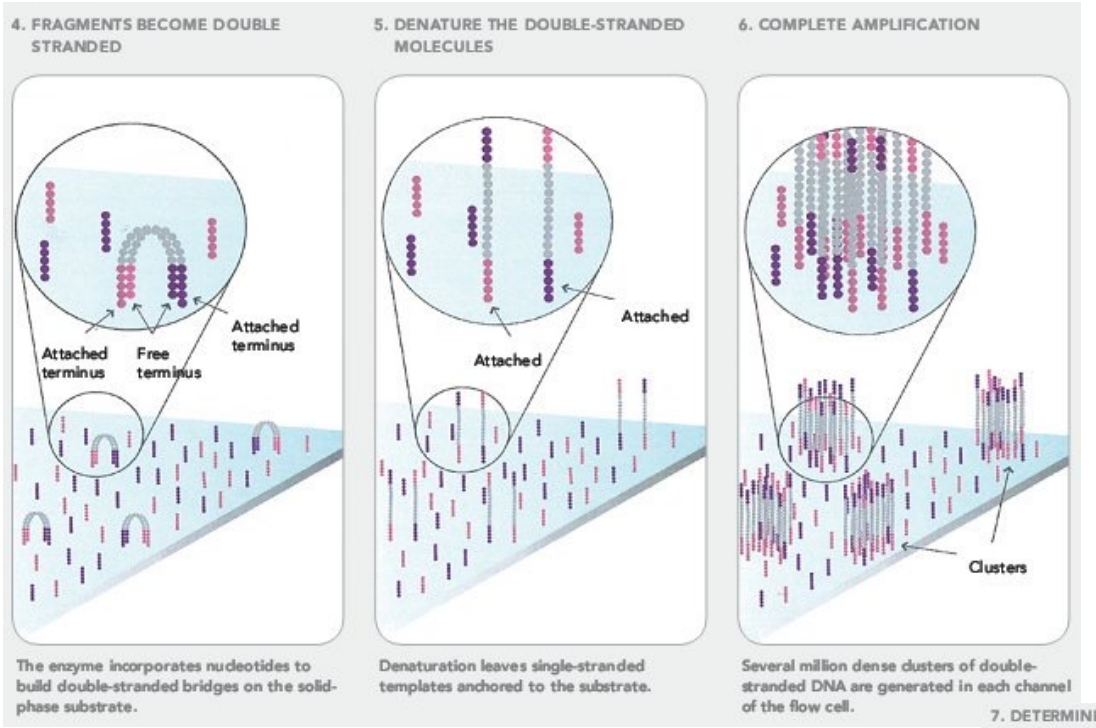
Single-end VS paired-end sequencing

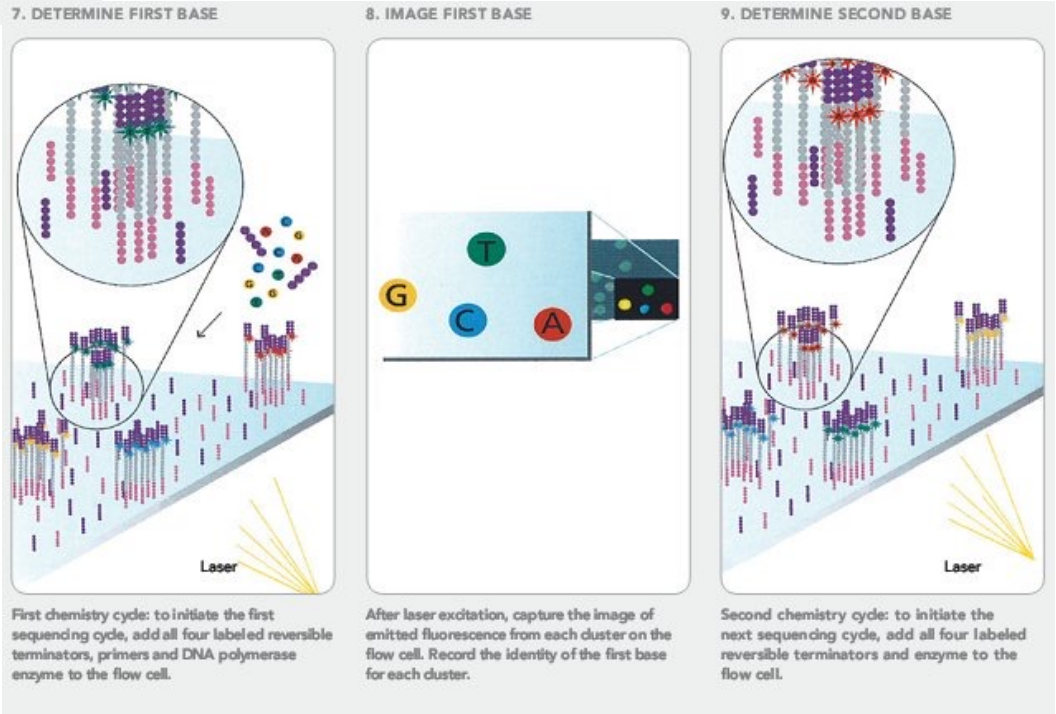


Illumina platform: Sequencing by Synthesis

https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=16s

Illumina platform: Sequencing by Synthesis





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

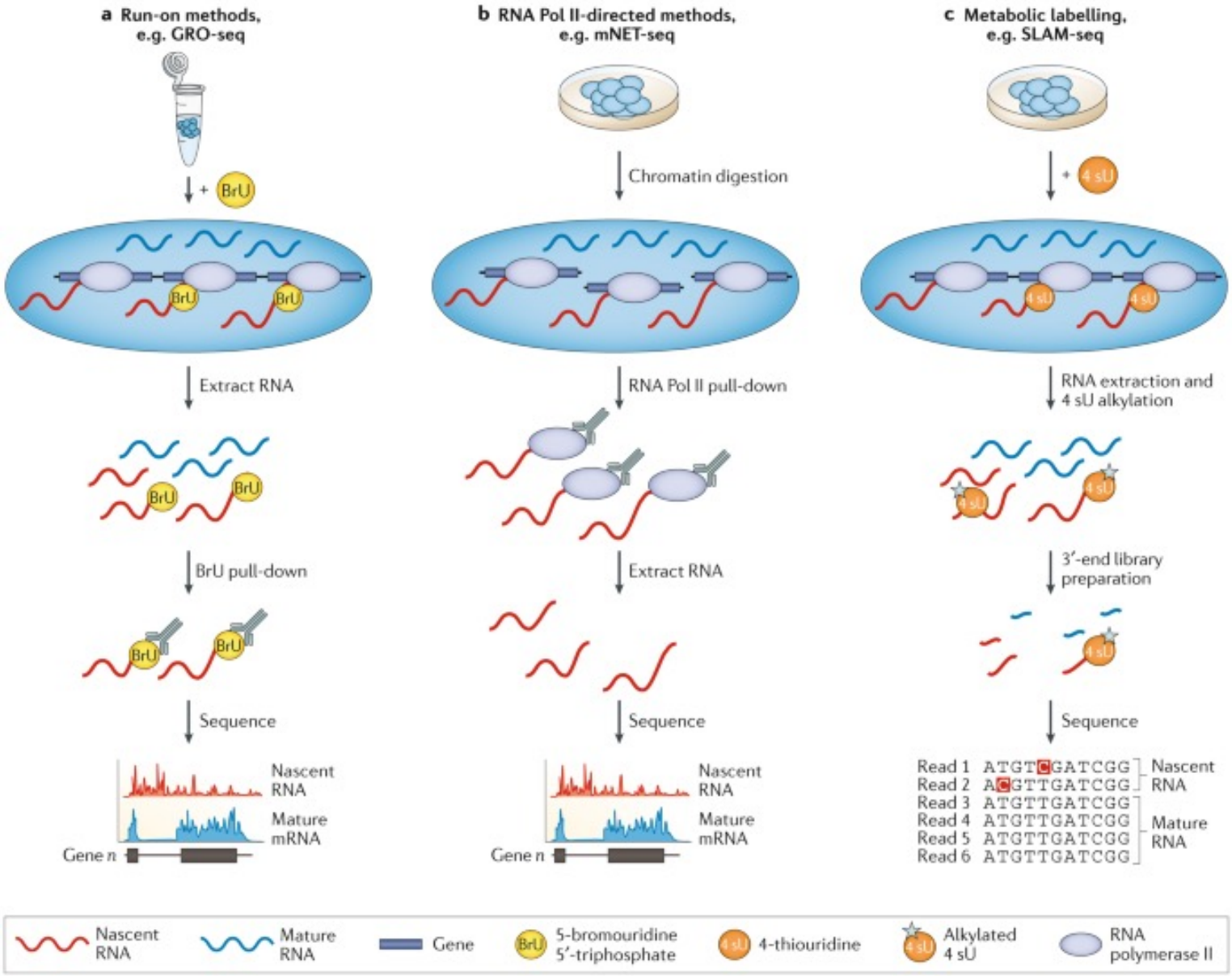
Key specifications	New MiSeq i100 Series	 NextSeq 550 System	 NextSeq 1000 and 2000 Systems
Max output per flow cell	30 Gb [a]	120 Gb [b]	540 Gb
Run time (range)[c]	~4–24 hr	~11–29 hr	~8–44 hr
Max reads per run (single reads)	100M [a]	400M [b]	1.8B
Max read length	2 × 500 bp	2 × 150 bp	2 × 300 bp

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	✓	✓	✓

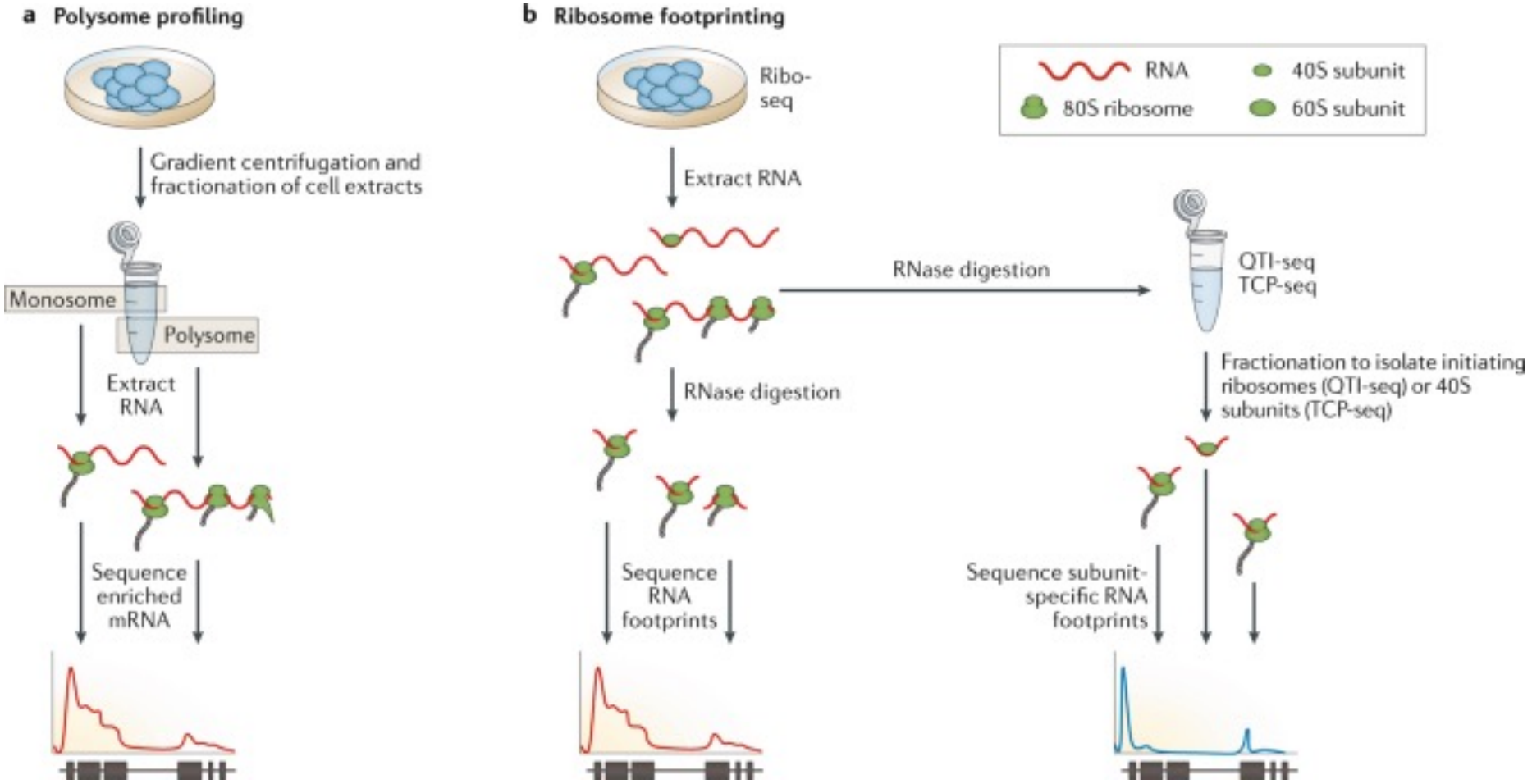
Nascent RNA sequencing

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



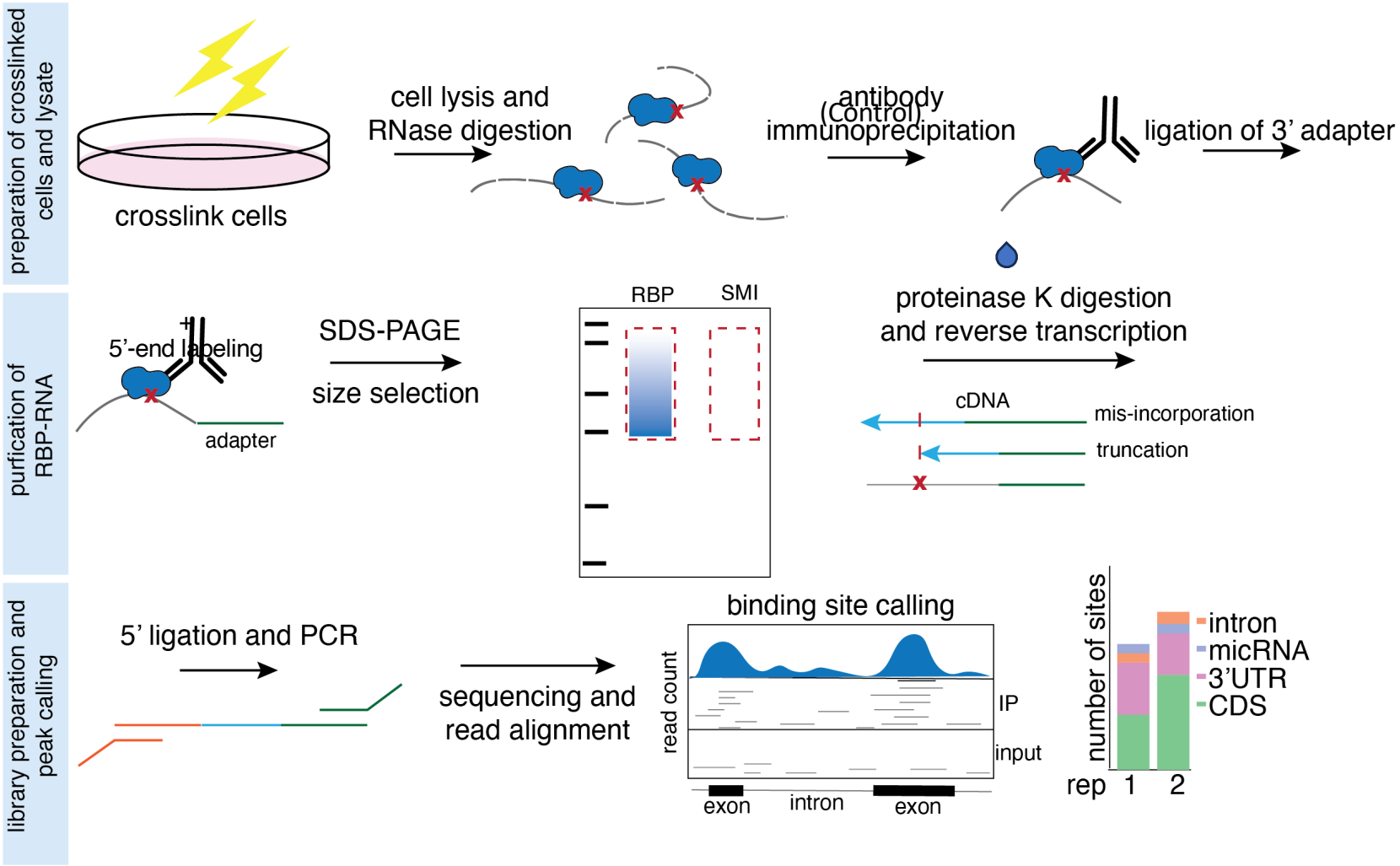
Translatome analysis methods

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



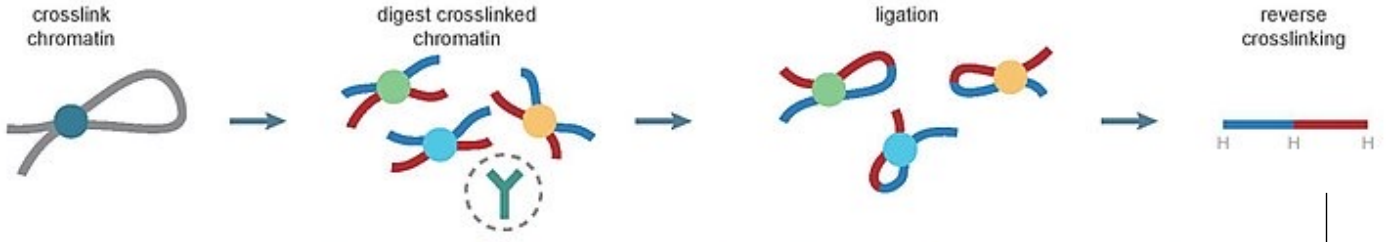
RNA–protein interaction analysis

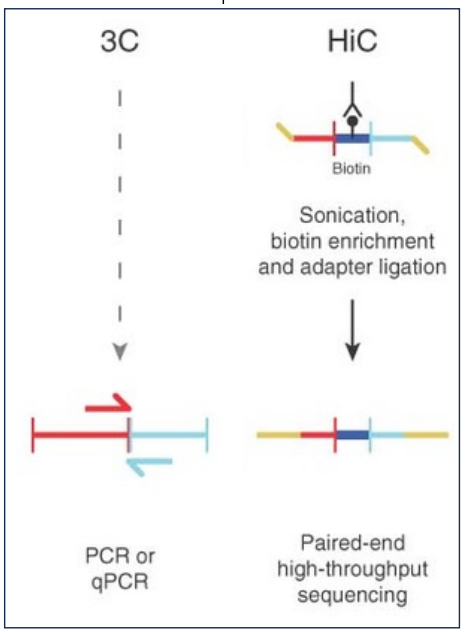
CLIP-seq



Chromosome Conformation Capture Methods

Common steps:



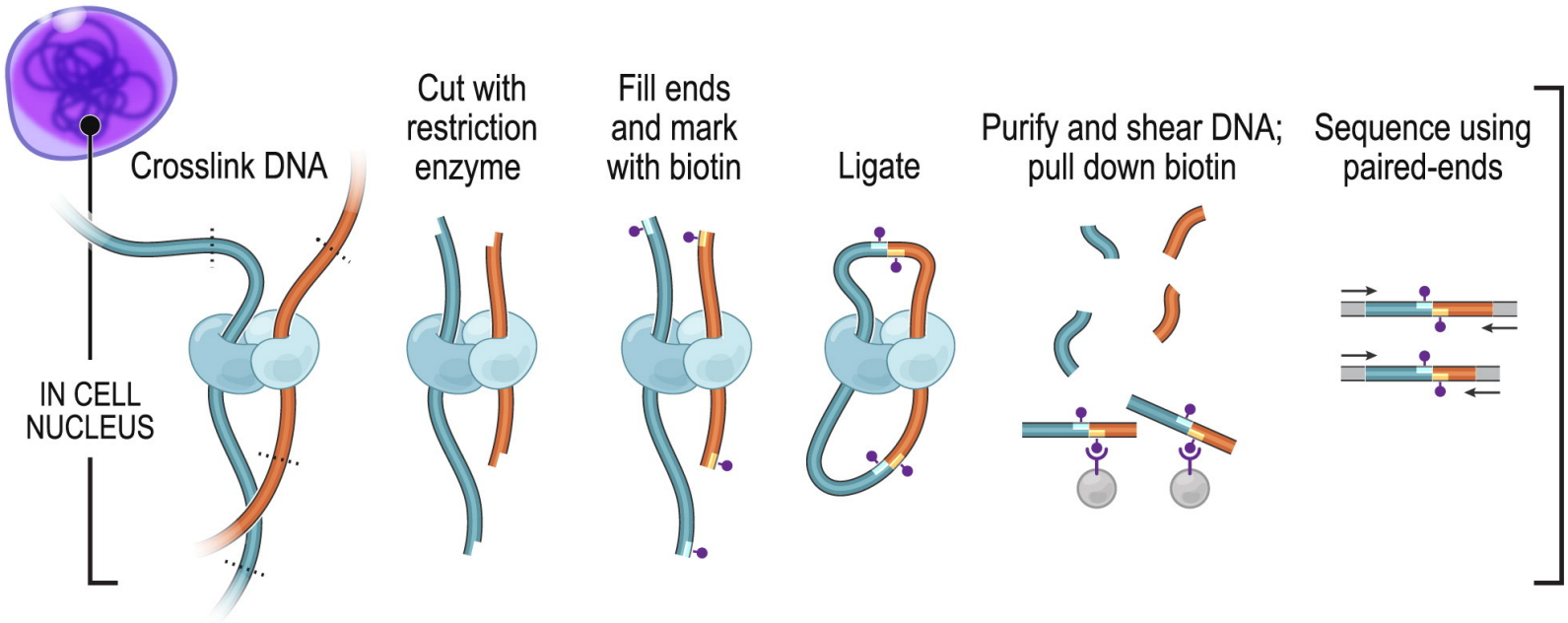


one-vs-one

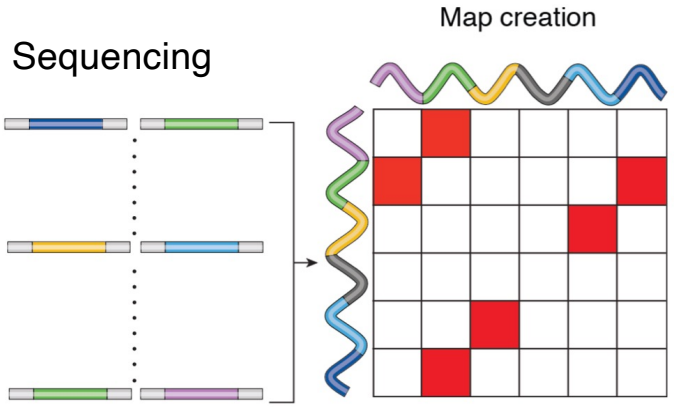
all-vs-all

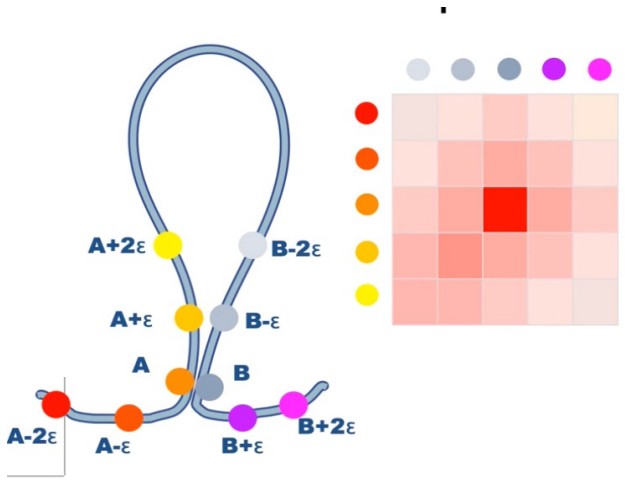
Hi-C Method

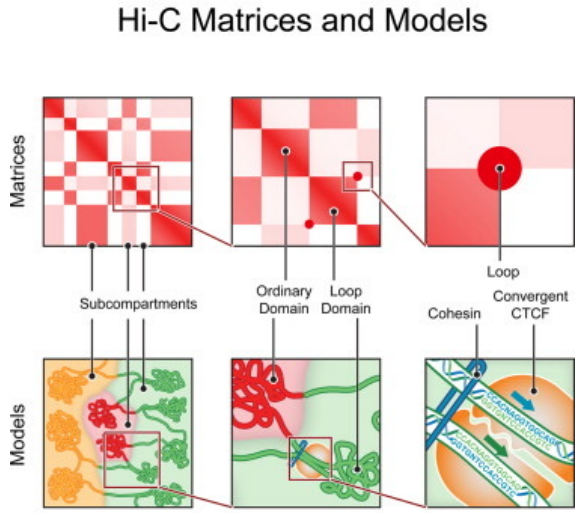
Hi-C protocol involves crosslinking cells with formaldehyde, permeabilizing them with nuclei intact, digesting DNA with a suitable 4-cutter restriction enzyme filling the 5′-overhangs while incorporating a biotinylated nucleotide, ligating the resulting blunt-end fragments, shearing the DNA, capturing the biotinylated ligation junctions with streptavidin beads, and analyzing the resulting fragments with sequencing.



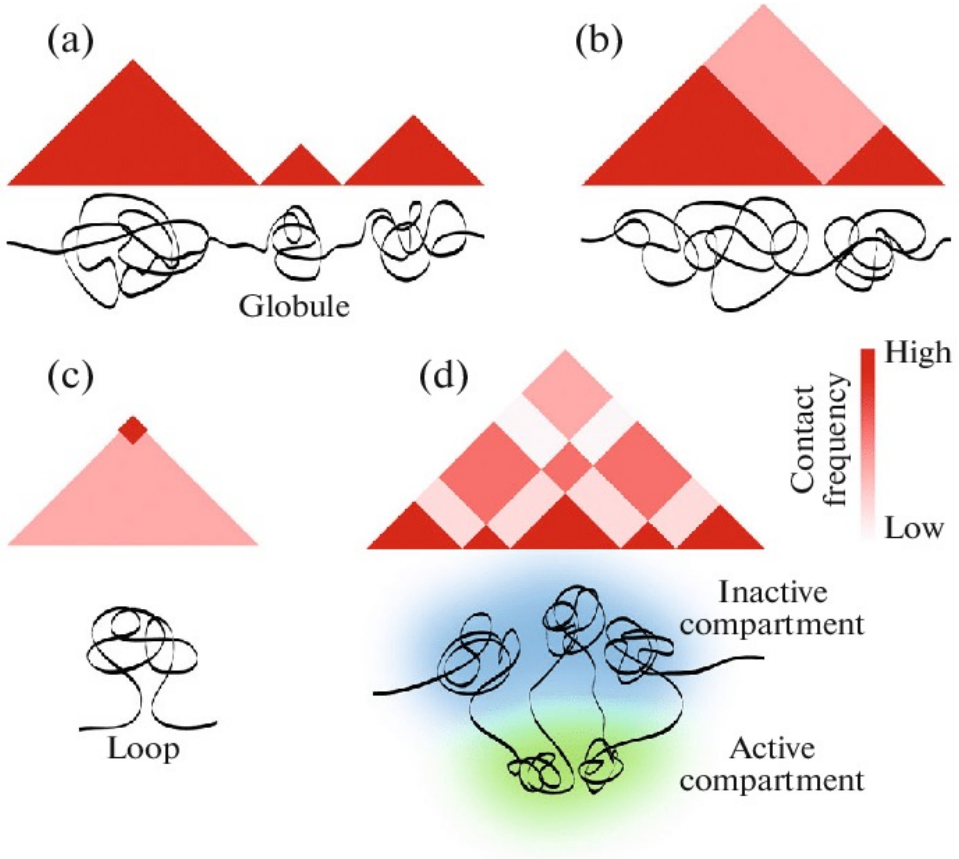
Visualization of genome interactions





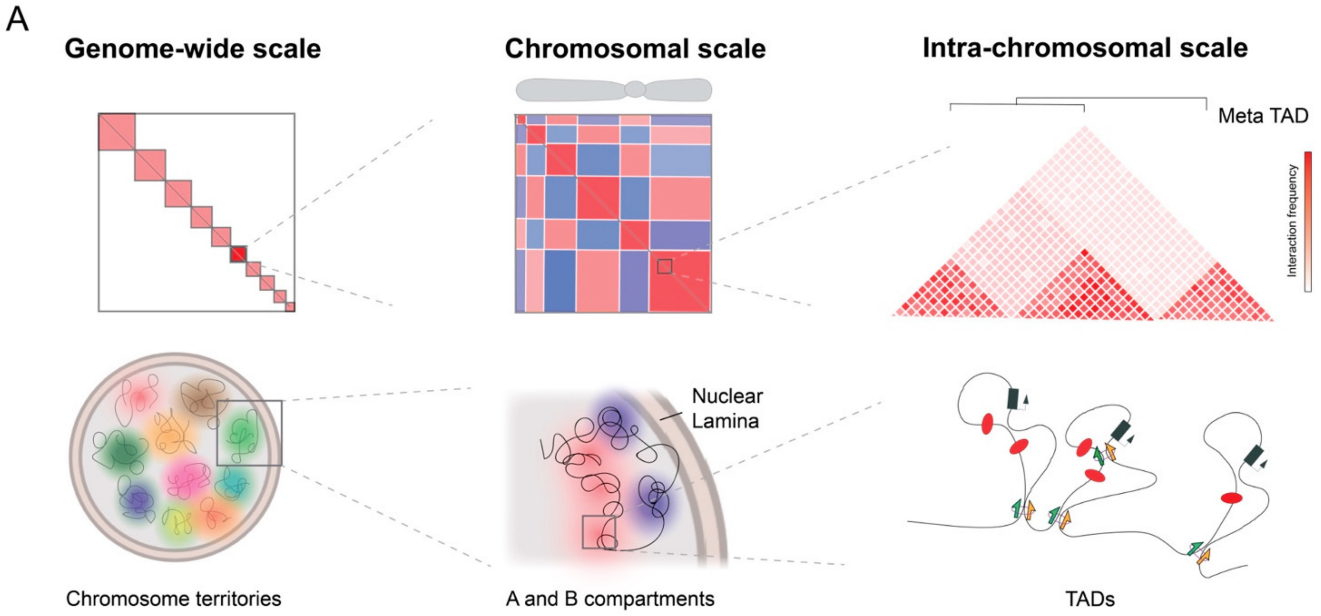


Visualization of genome interactions

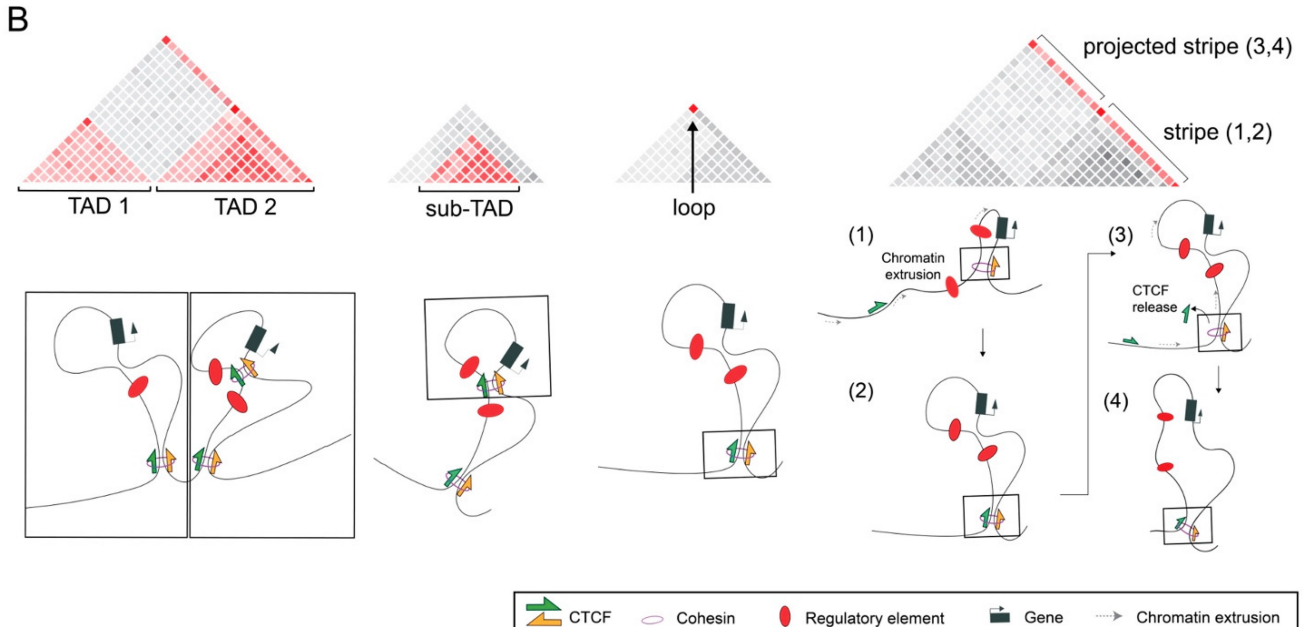


Hi-C map patterns (at the top) and their interpretation (at the bottom). (a) Isolated triangles on a heat map are commonly interpreted as chromatin globules. (b) A combination of small triangles into larger ones may reflect the hierarchic organization of globules (small globules fuse into larger ones). (c) An intense signal at the apex of a triangle suggests the interaction of TAD boundaries and the formation of a chromatin loop, internal regions of which may similarly interact with each other to form a globule. (d) A checkerboard pattern suggests the formation of chromatin compartments.

Visualization of genome interactions

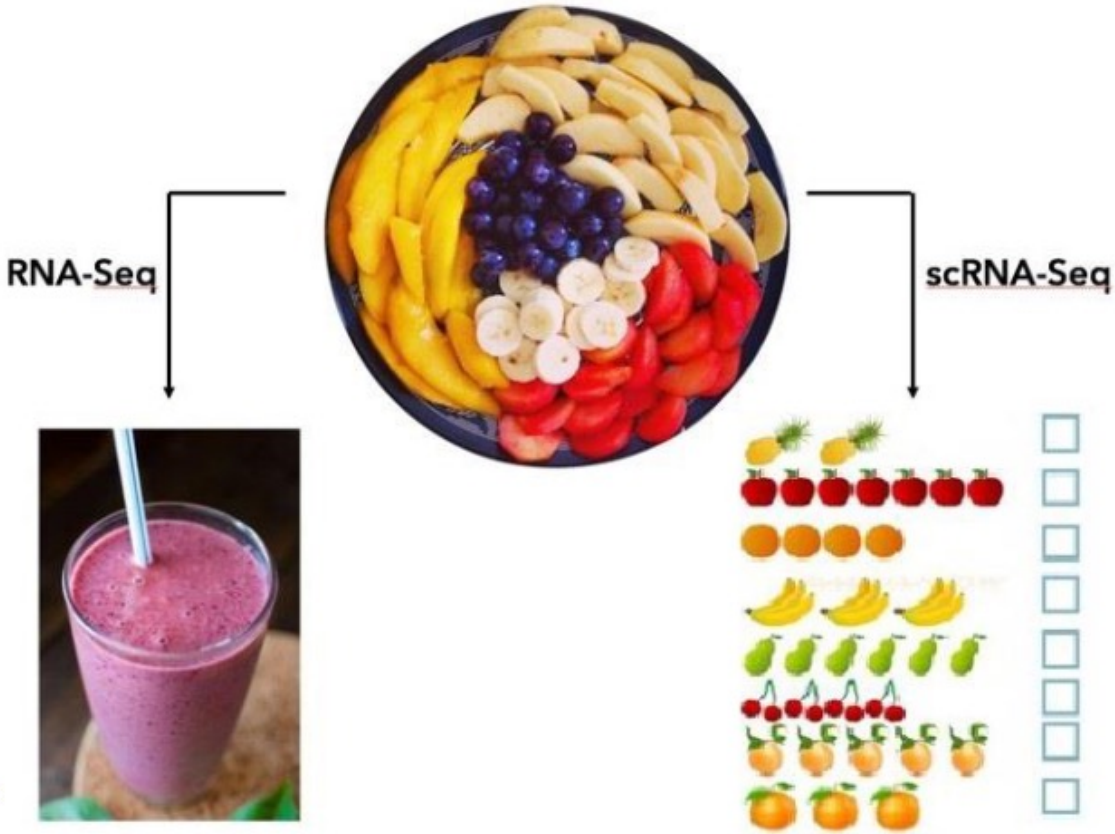


TADs: Topologically associating domains



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

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

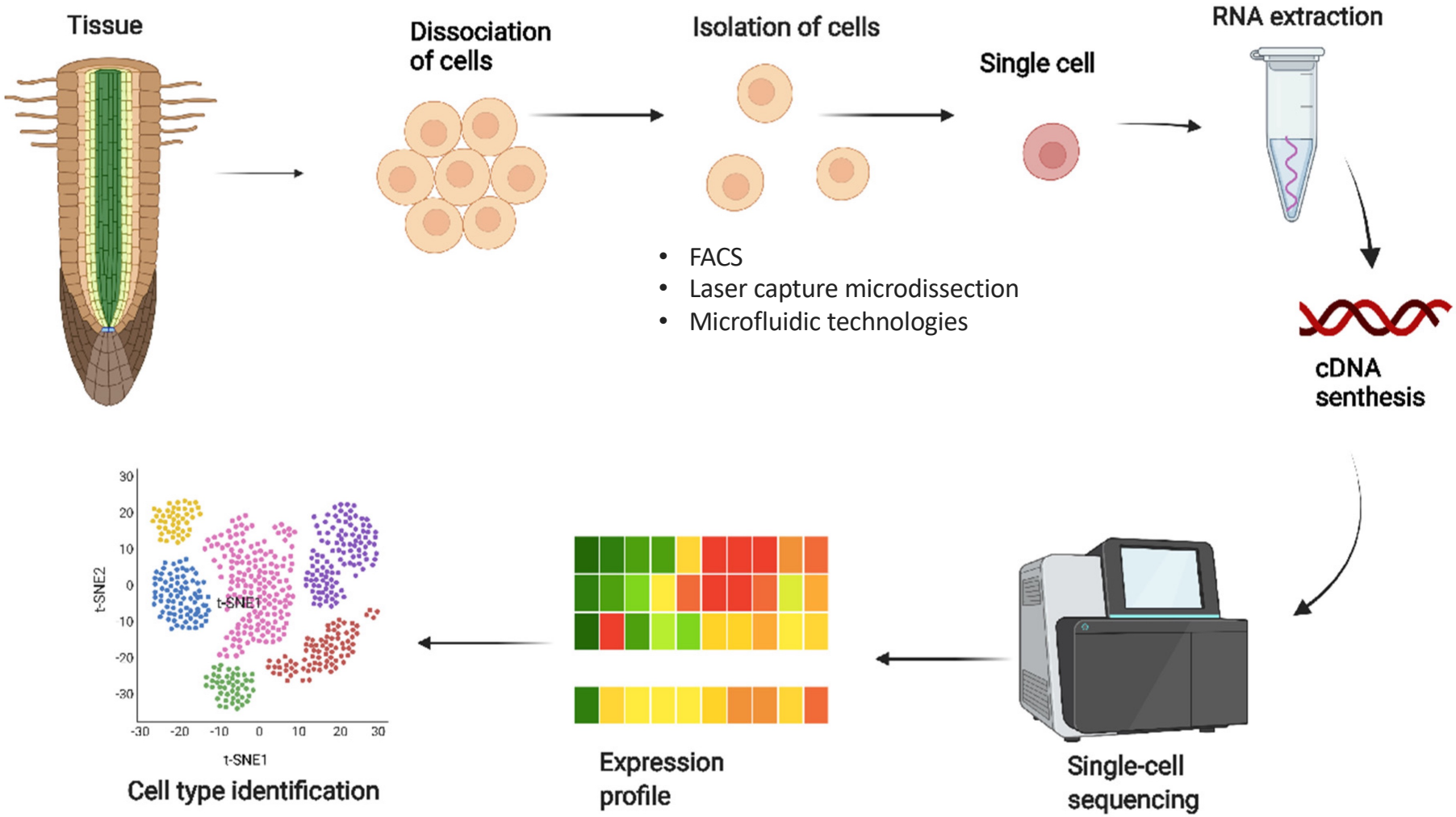


Separate populations
- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

Single cell RNA–Seq (scRNA-Seq)

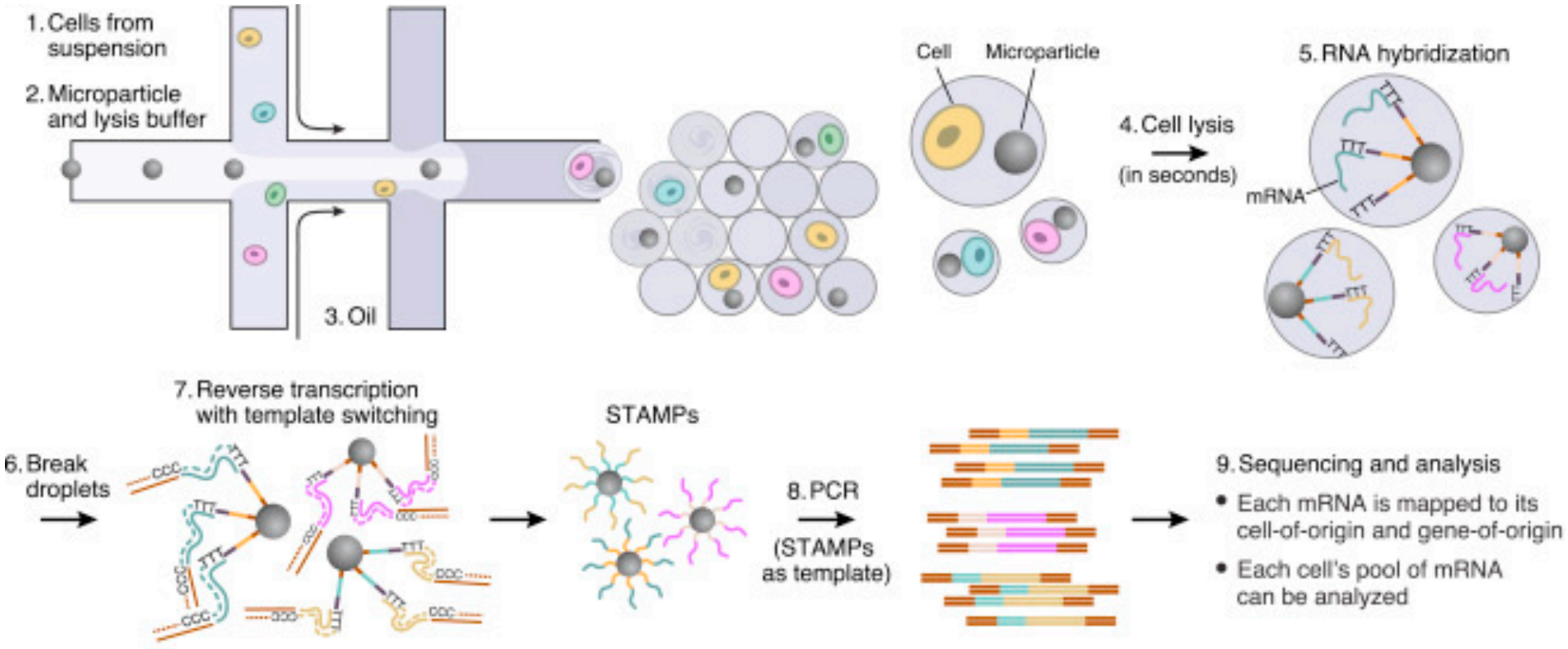
- A new technology, first publication by (Tang et al. 2009).
- Did not gain widespread popularity until ~2014/15 when new protocols and lower sequencing costs made it more accessible.
- Measures the distribution of expression levels for each gene across a population of cells.
- Currently there are several 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)



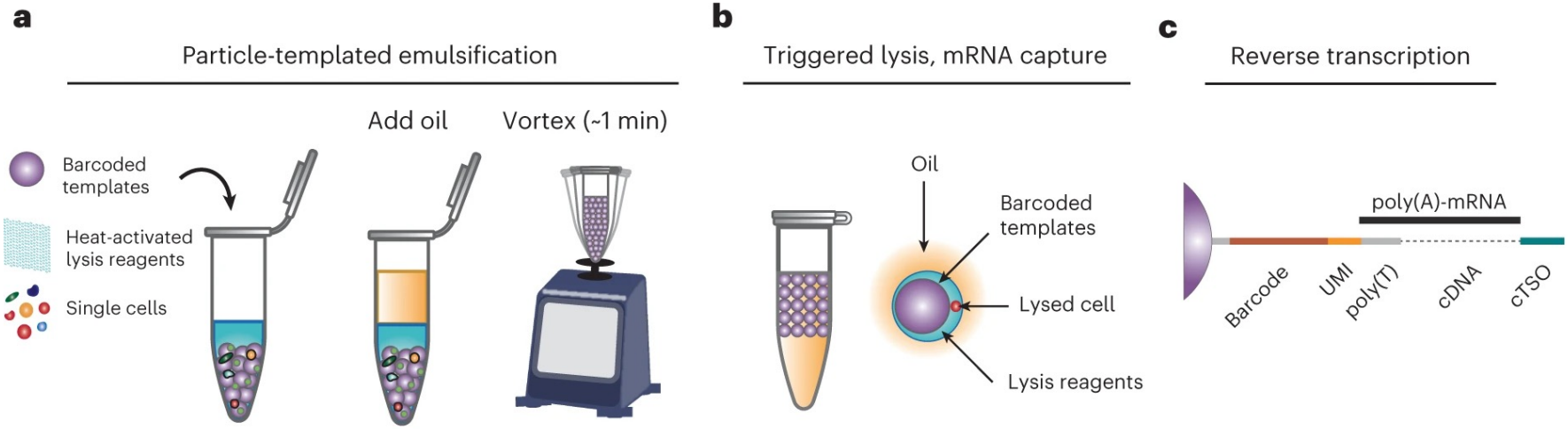
- FACS
- Laser capture microdissection
- Microfluidic technologies

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



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

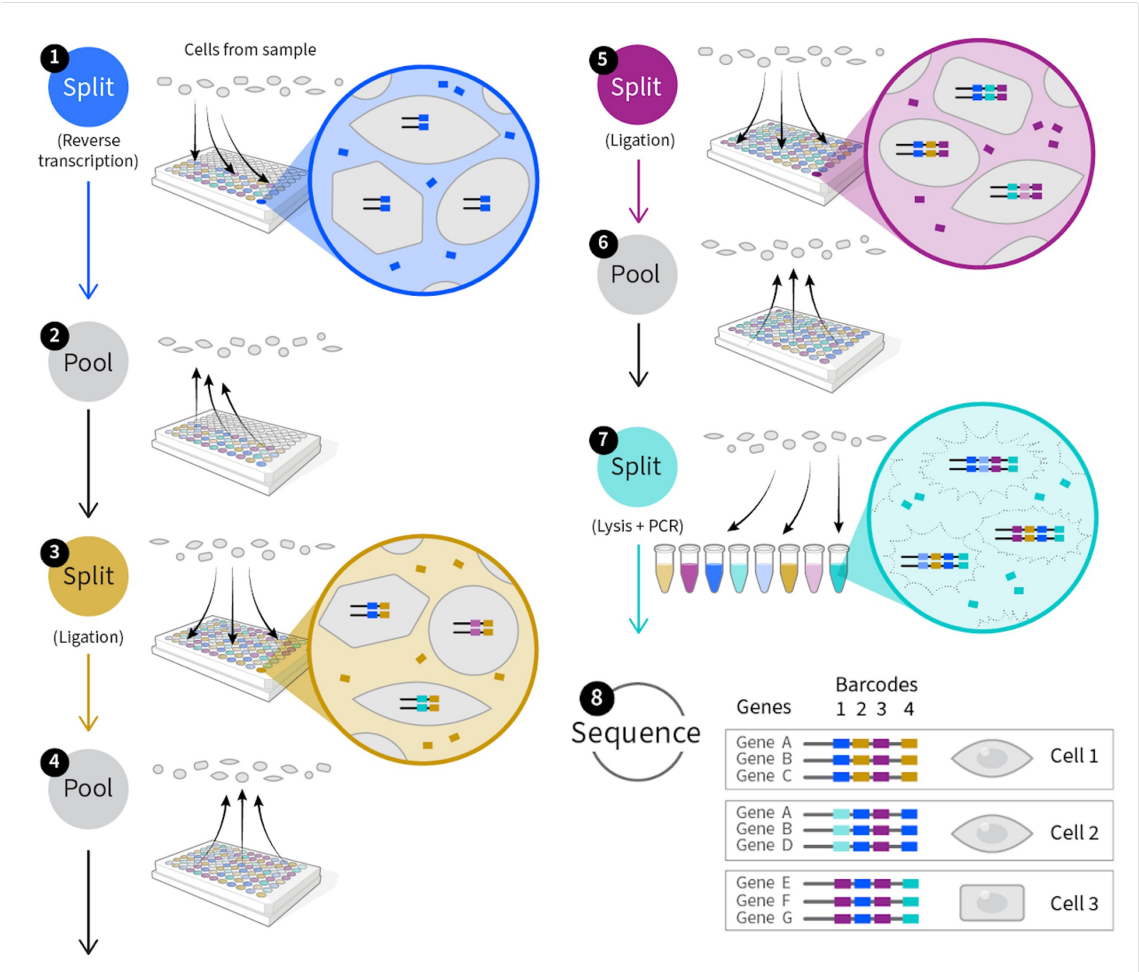
PIP-seq and Drop-seq are both high-throughput methods for single-cell RNA sequencing that use microfluidic droplets to isolate and barcode individual cells. The core principle involves encapsulating a single cell and a barcoded bead in a droplet to tag the cell's entire mRNA content. The key difference 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)

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

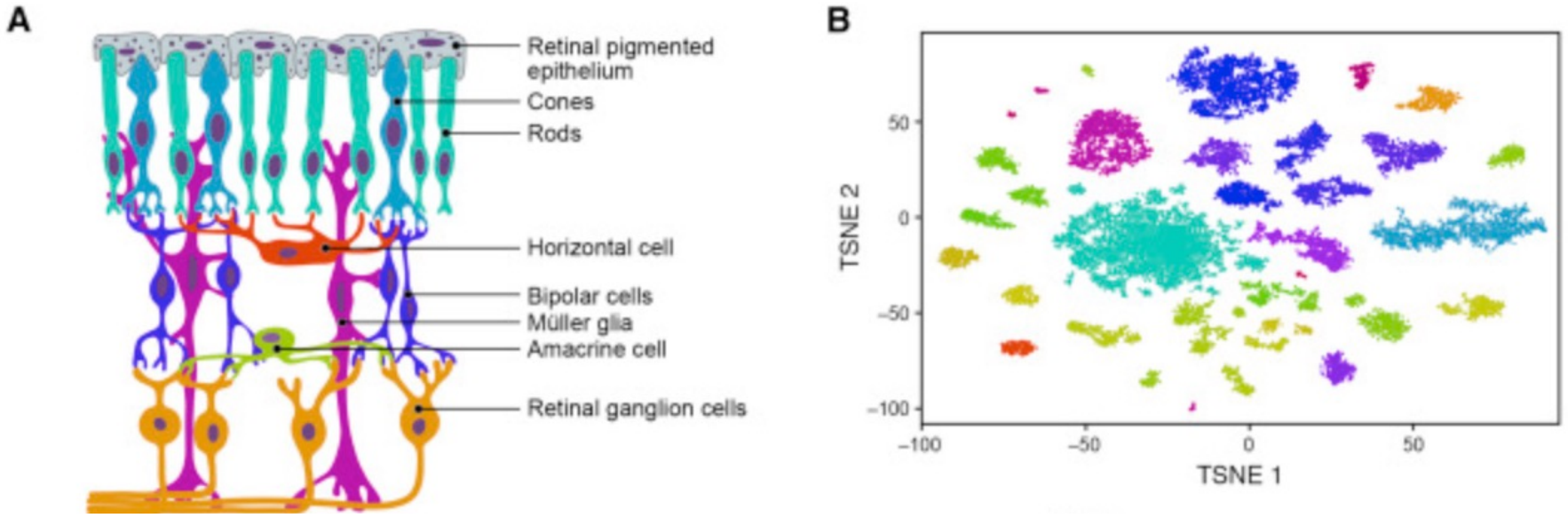
If you start with one million cells, and split them into 96 wells, each well has roughly 10,000 cells. Every cell in one well gets one barcode. That's 10,000 cells with one barcode. Then, the cells are pooled together and randomly distributed again into a second plate with 96 new wells. A second barcode is added to every well as each cell starts to take a unique path through these wells. This process is done over again through a third and fourth round. The probability that any two cells get the same combination of barcodes becomes smaller and smaller as they progress through the protocol.



(no instrument required)

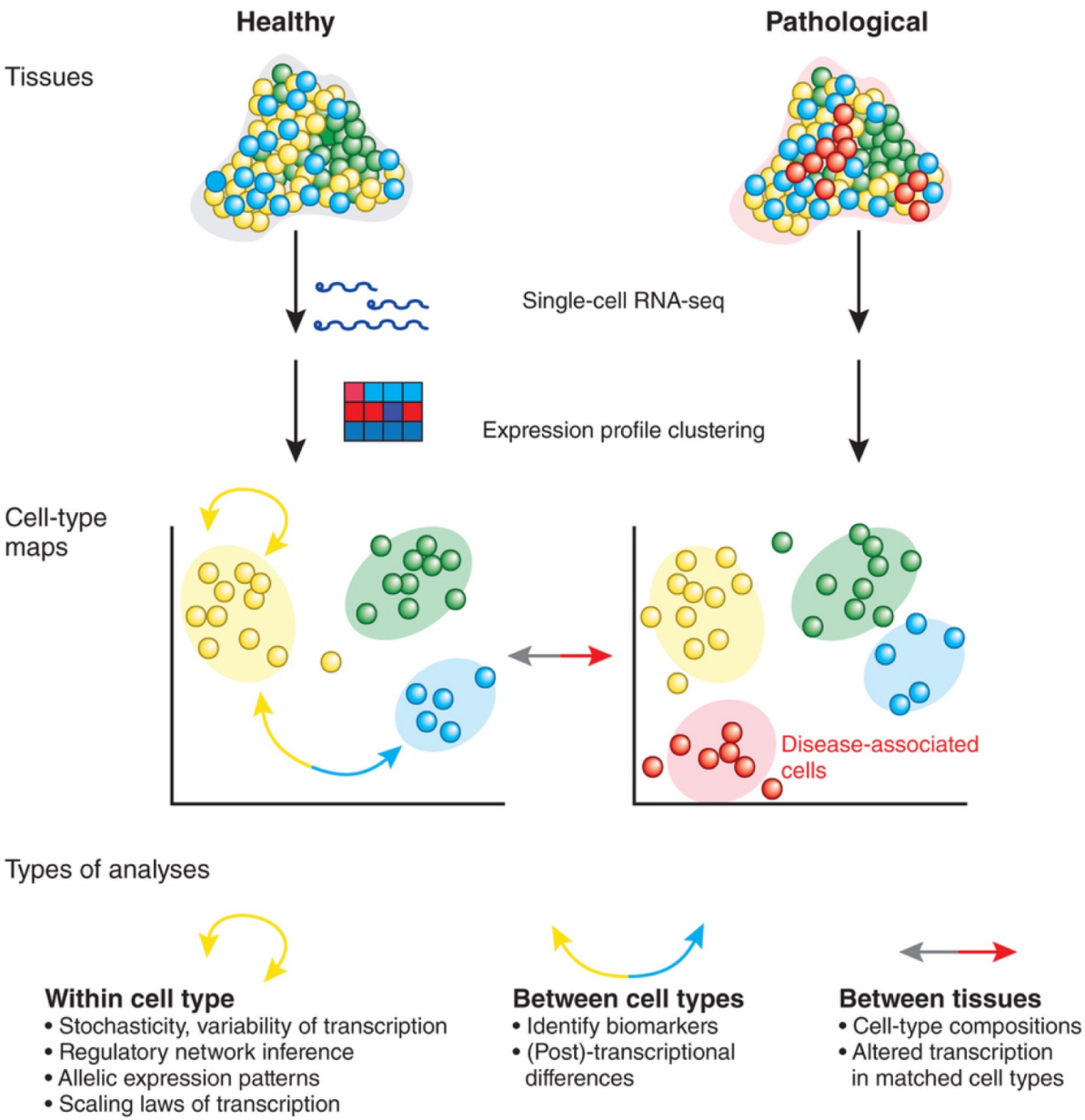
What type of biological questions can we address?

Reconstruction of Retinal Cell Types from 44,808 Single-Cell Transcription Profiles Prepared by Drop-Seq



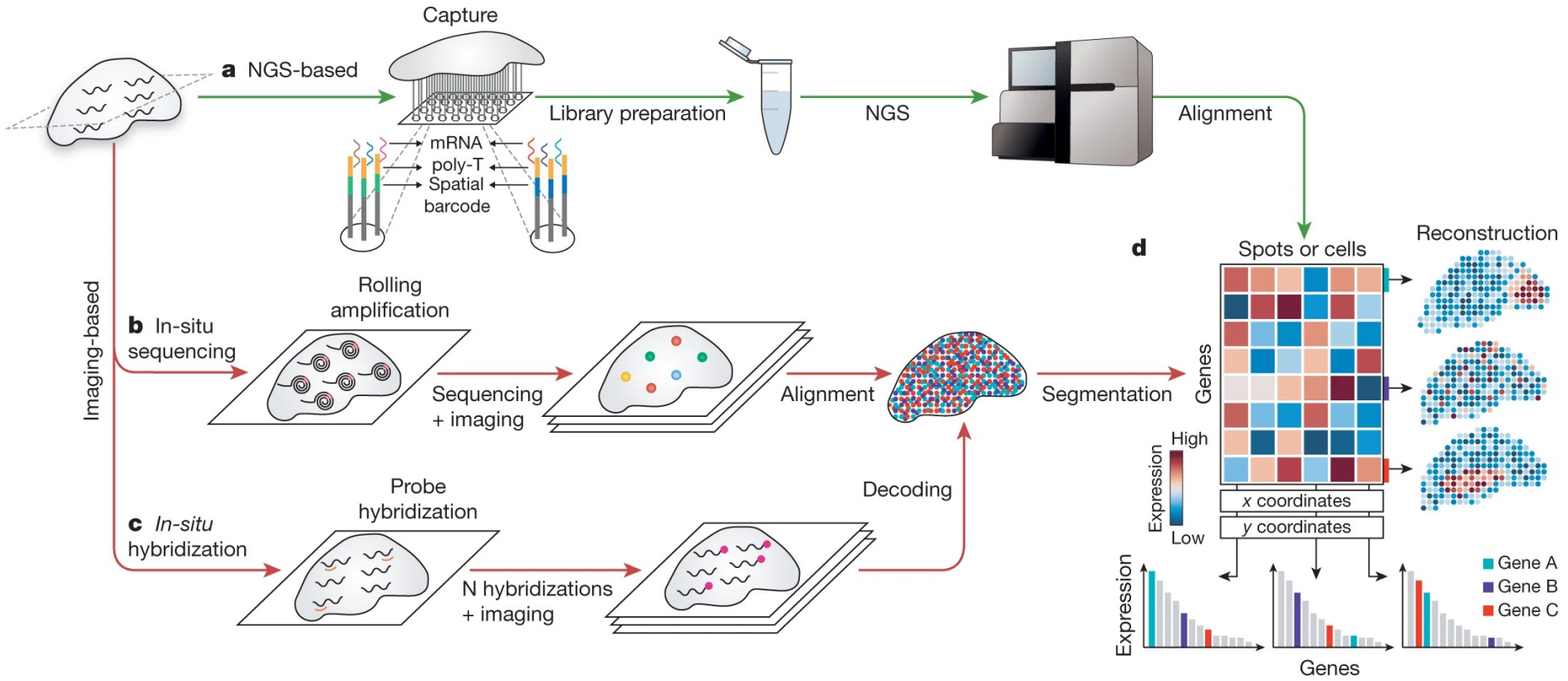
The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among 44,808 cells;

What type of biological questions can we address?



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

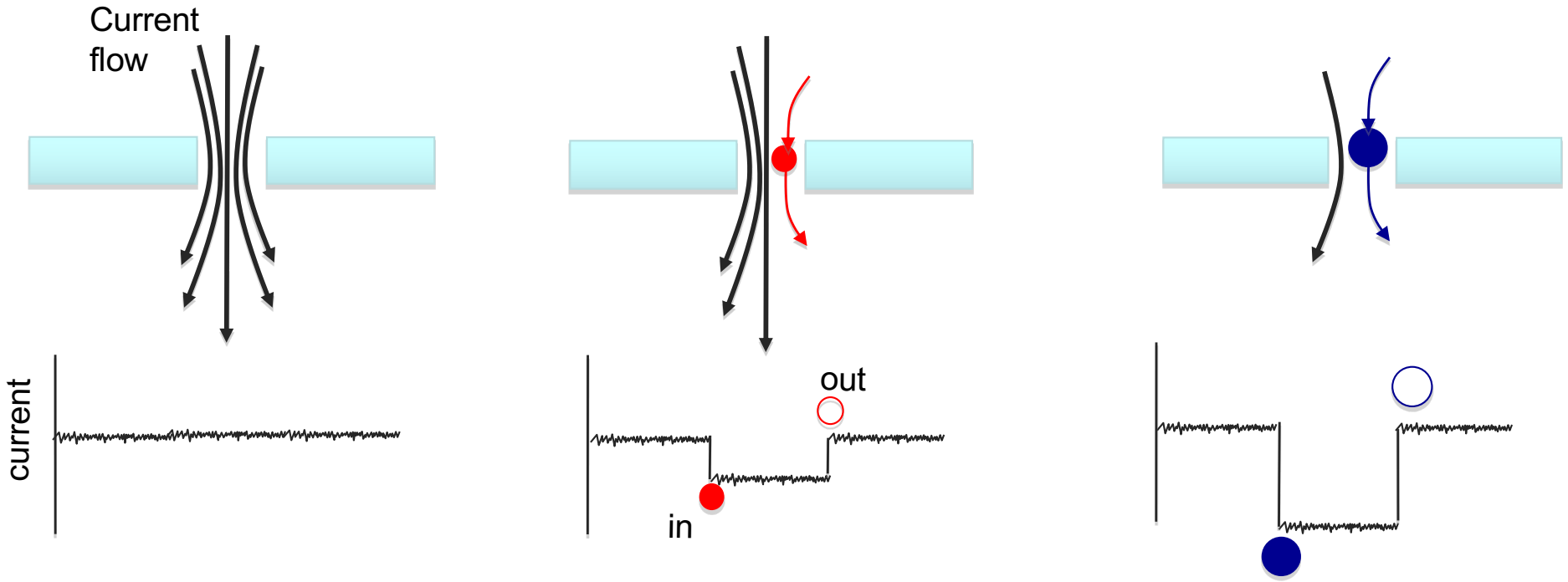


Rao et al., 2021 Nature

Oxford Nanopore

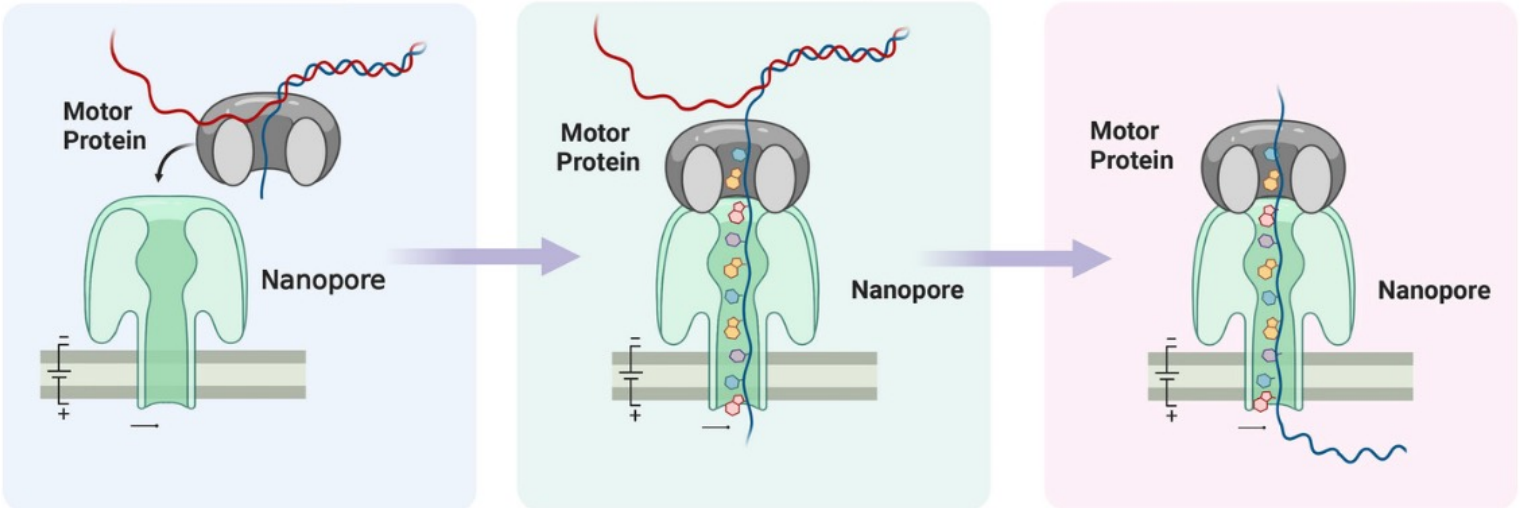
Determine the sequence of DNA/RNA fragments by passing them through a protein (or other) pore in a membrane

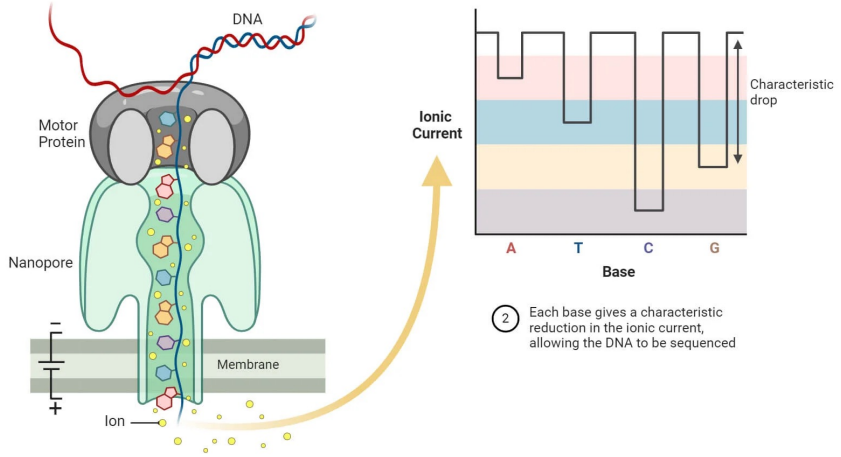
- Nanopore = ‘very small hole’
- Electrical current flows through the hole
- Introduce analyte of interest into the hole ➔ identify “analyte” by the disruption or block to the electrical current



Oxford Nanopore

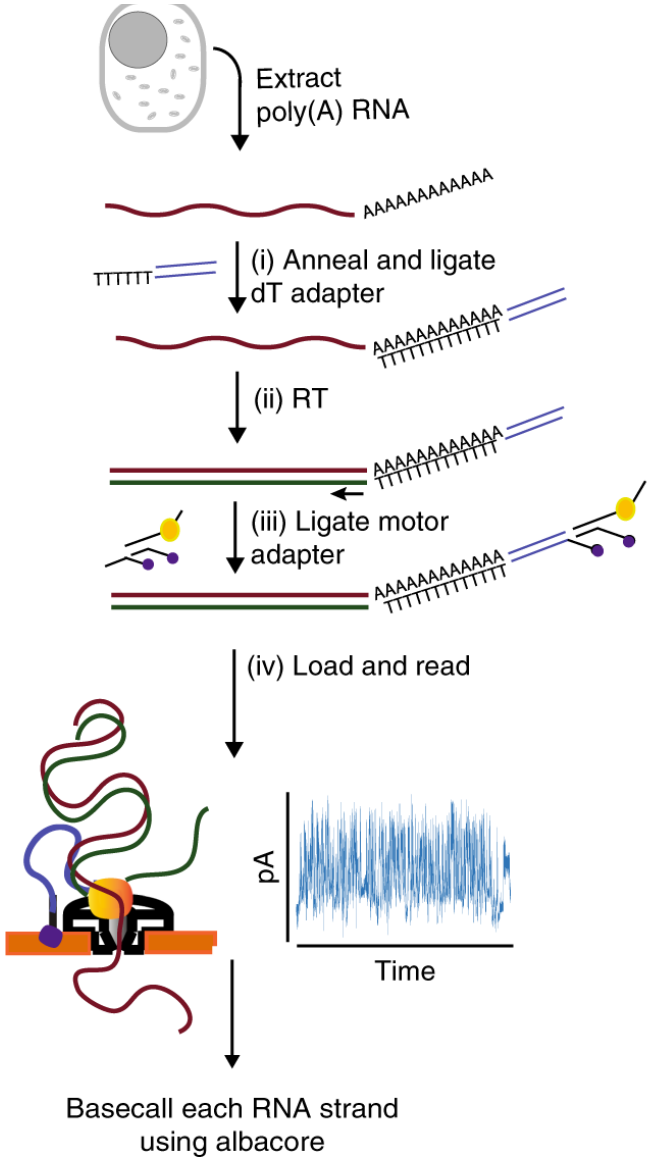
Motor proteins pull the nucleic acid through the nanopore, and the nucleic acid is detected and transmitted to the computer through the generated tiny current signal. Each base give a characteristic reduction of the ionic current, allowing the DNA to be sequenced.



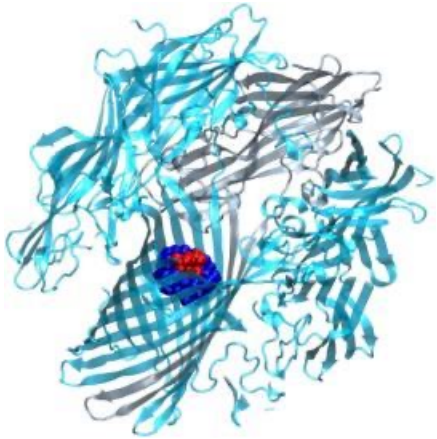


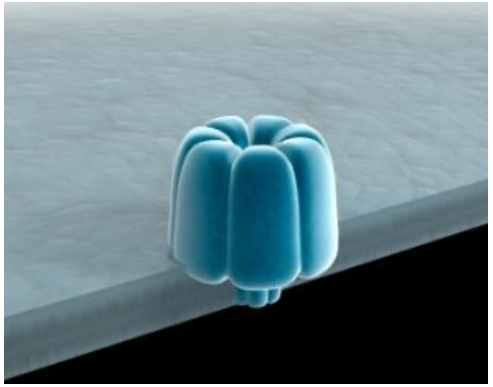
Nanopore *Direct* RNA Sequencing

Nanopore technology is capable of characterizing any polymer placed in the pore, not just DNA. For RNA sequencing, the direct approach is highly advantageous, providing long reads, isoform detection, and the capability for direct identification of base modifications (such as m6A), which is particularly beneficial when analyzing samples from variable shearing conditions.



Oxford Nanopore



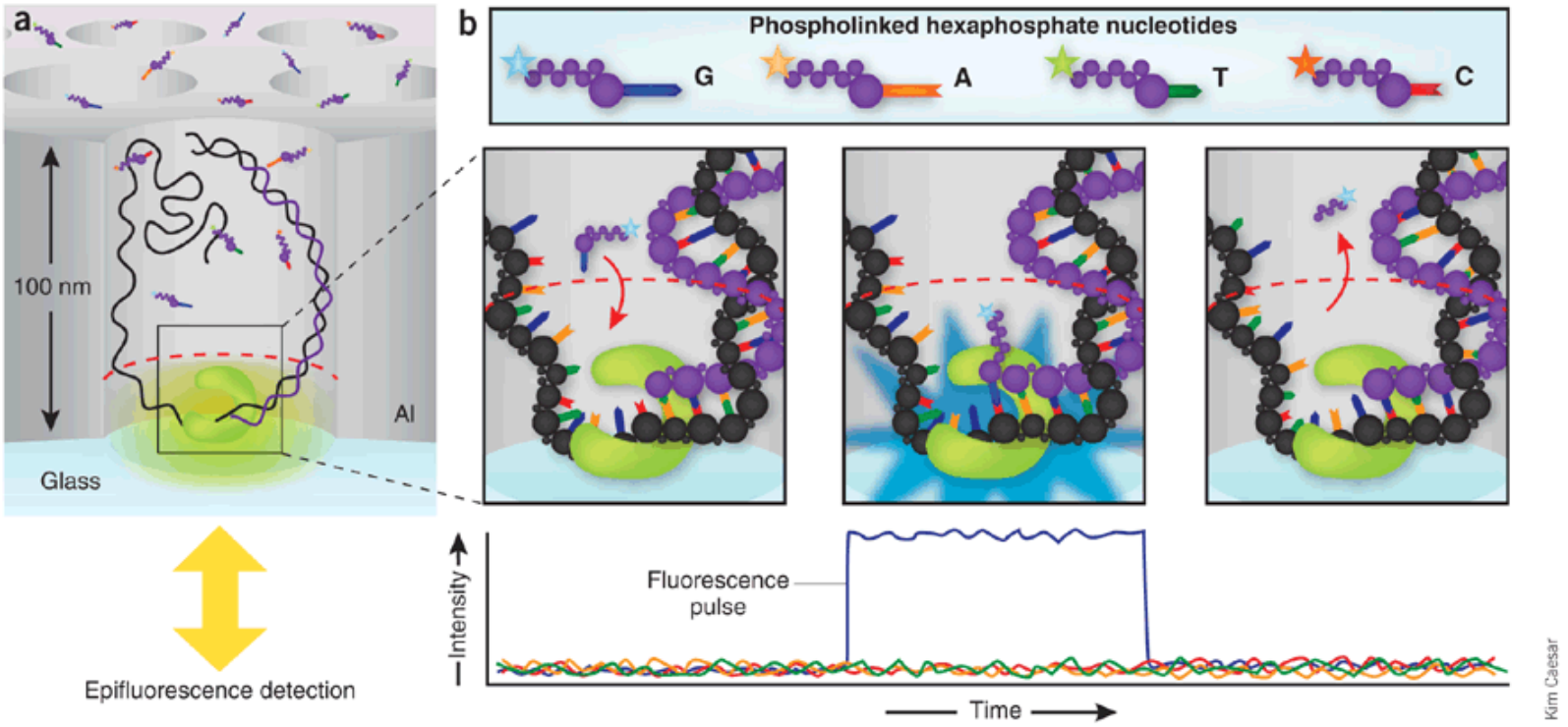


Oxford Nanopore

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

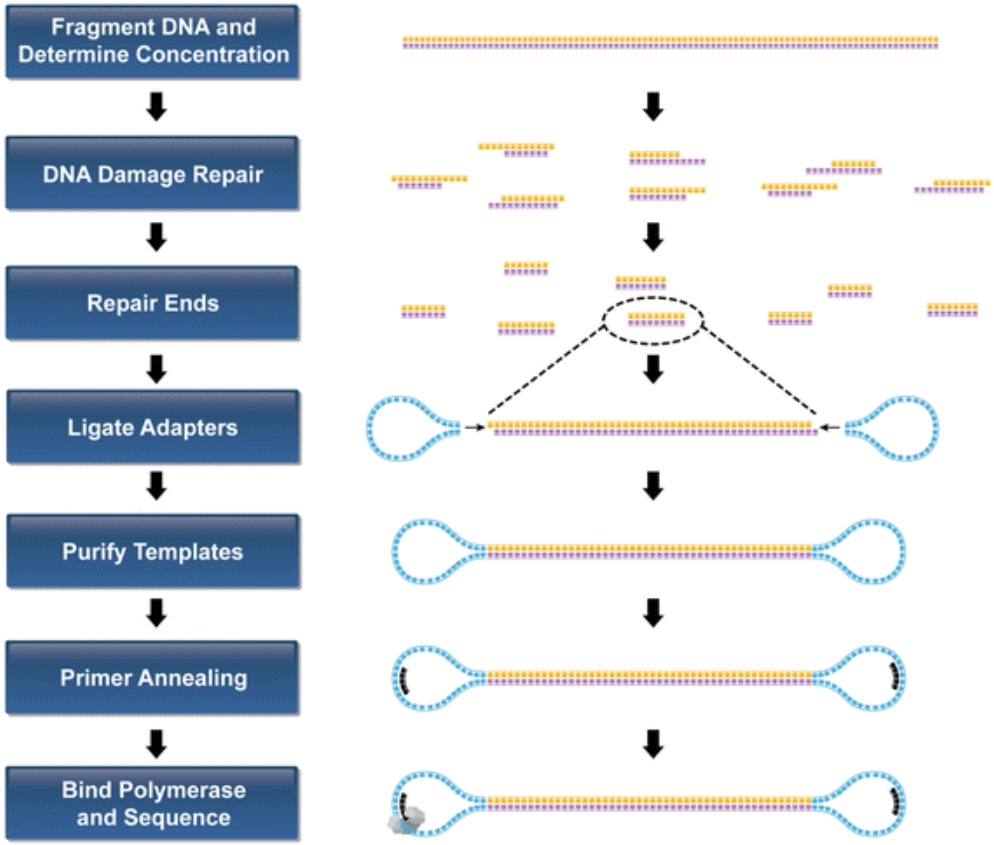
PACBIO sequencing machine

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).



PACBIO sequencing machine

https://www.youtube.com/watch?time_continue=99&v=v8p4ph2MAvI&feature=emb_logo

PACBIO sequencing machine

The polymerase continuously 'reads' the fragment to be sequenced, which has to be prepared with hairpin adapters (SMRTbell), where the polymerase can bind to start reading the template. SMRTbells are ligated on both sides of each double-stranded DNA fragment, forming a circular sequence: