BIO214 Lecture 2

Bioinformatics-II

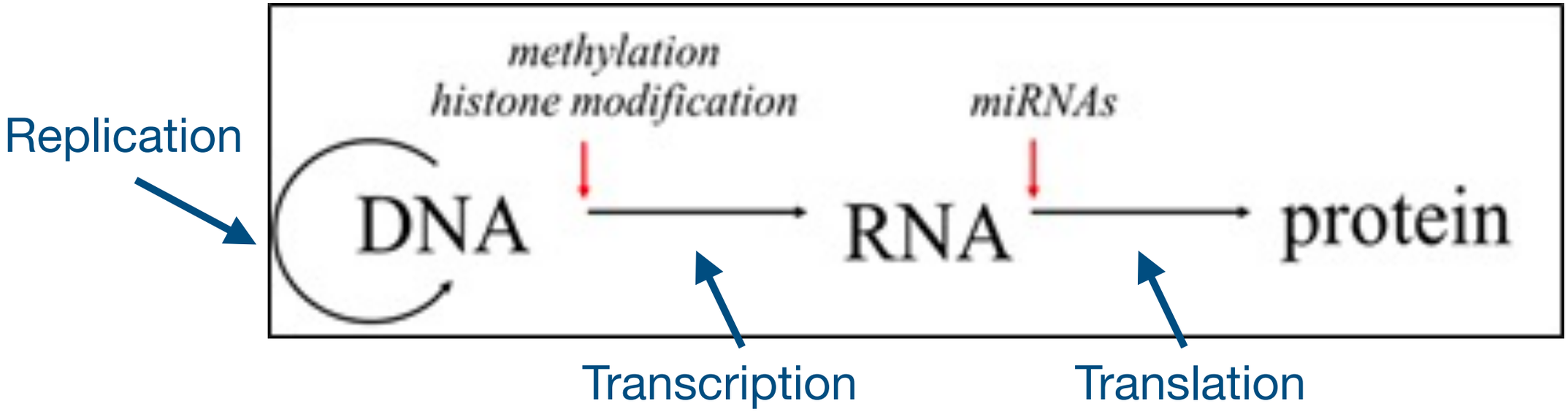
Introduction to Sequencing Technologies

Zhen Wei; 2023-Feb-14

Outline

- Sanger sequencing
- Next Generation Sequencing
- Measure RNA and epigenomes
- Single cell omics technique
- Real time sequencing

Flow of information in molecular biology



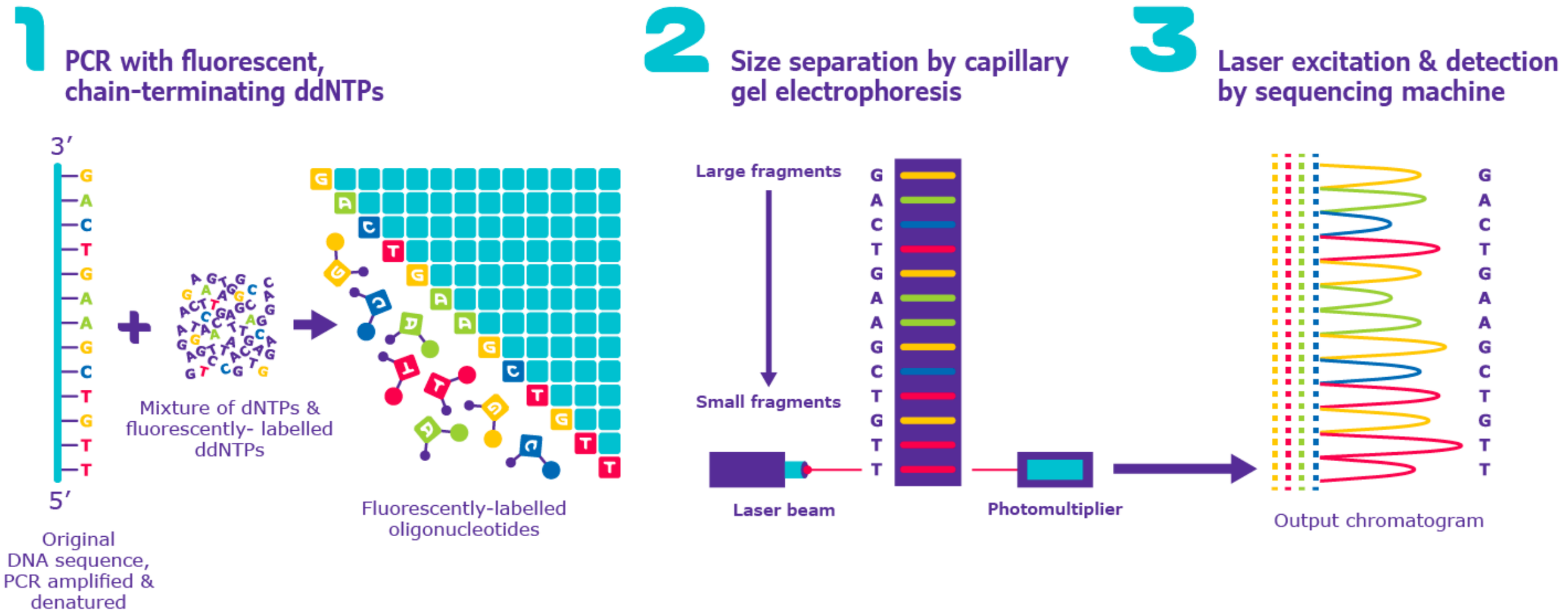
- Cell maintain and express genetic information through three main processes: **replication**, **transcription**, **translation**.
- Epigenetic markers, such as **histone modification** and **DNA methylation**, play a role in regulating transcription.
- Measuring the **genome**, **transcriptome** and **epigenome** is crucial for our understanding of life.

Sanger sequencing

How to experimentally determine the sequence of a DNA molecule?

- ddNTPs lack the 3'-OH group required for phosphodiester bond formation.

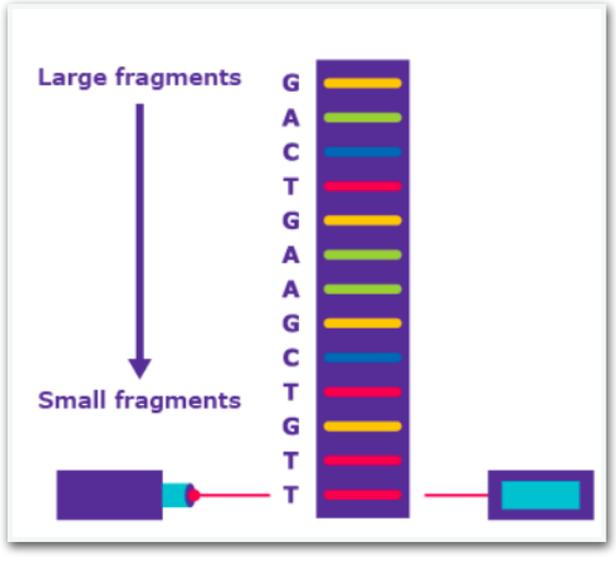
Sanger Sequencing



Steps for Sanger sequencing:

1. Mix a low amount of chain-terminating ddNTP with normal dNTP in PCR reaction, causing random termination of replication.
2. Use gel electrophoresis to separate chain-terminated oligonucleotides by size.
3. Each ddNTP has a unique fluorescent label, allowing the sequencer to read the sequencing results based on color.

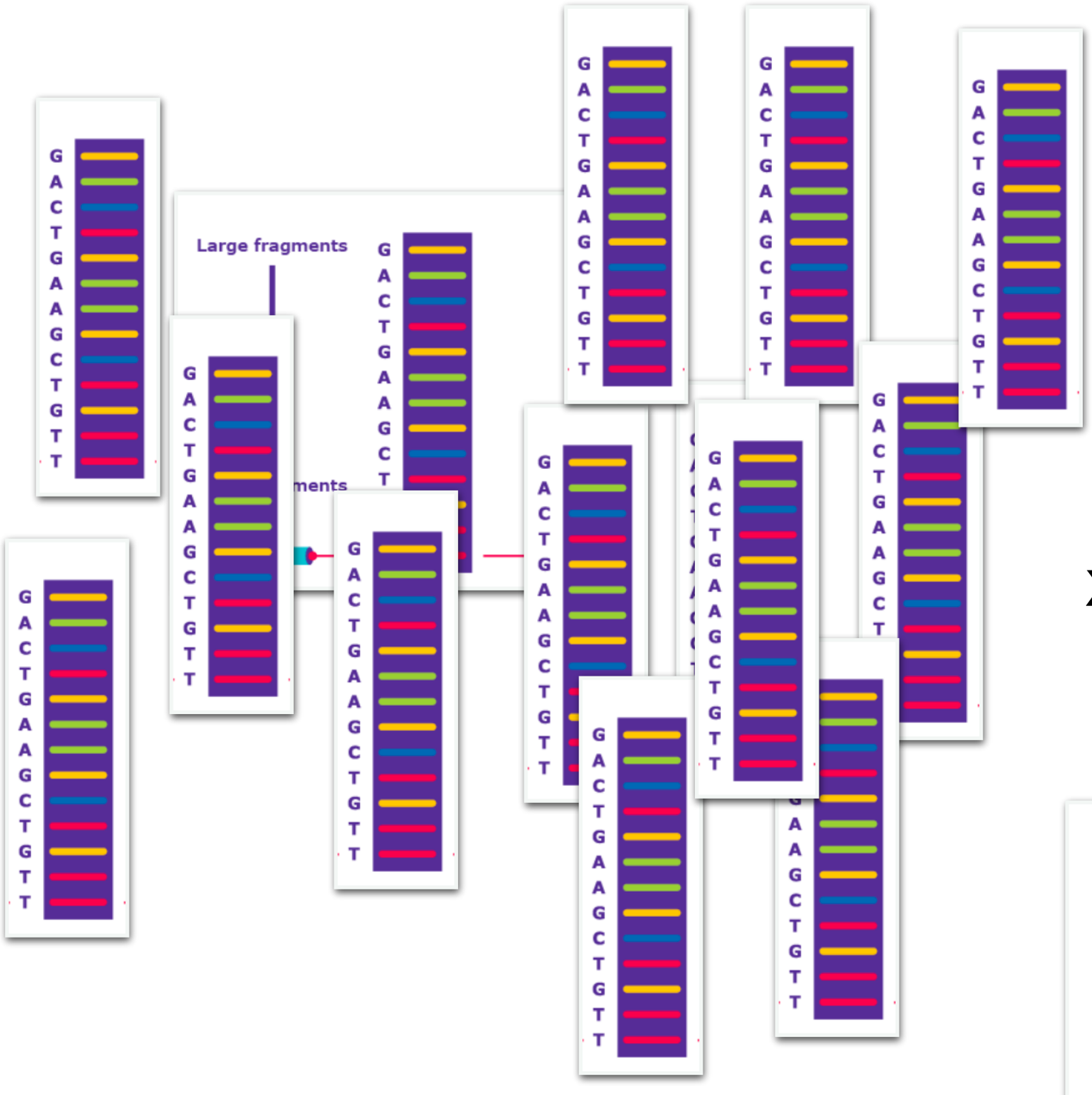
How to make the Sanger sequencing more efficient, i.e. able to measure multiple sequences at one time?



x 1

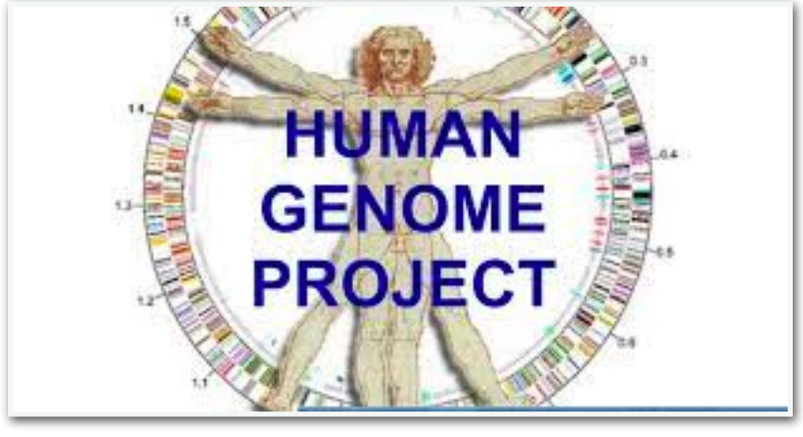
Speed is too slow using one Sanger sequencing machine

How about having multiple sequencers run together?



x N ...

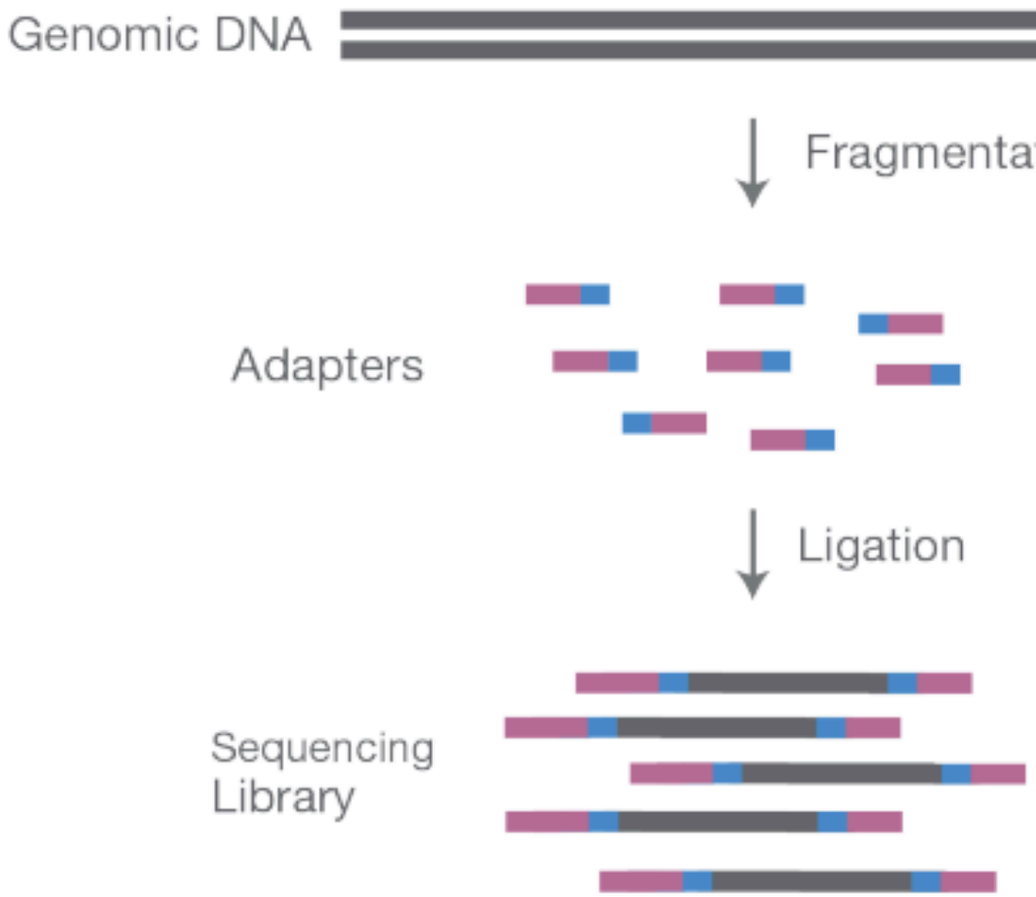
Back in 2000



Next generation sequencing

illumina sequencing (DNA-Seq)

Step 1. Library Preparation

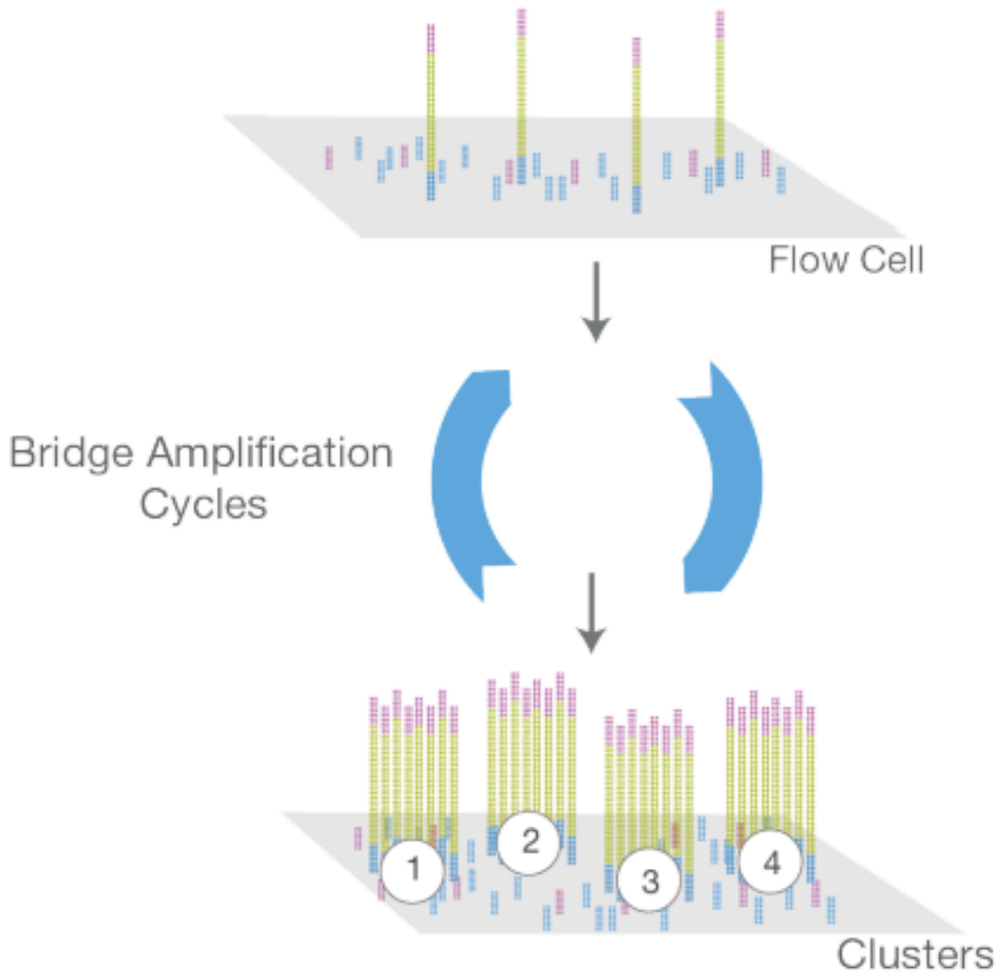


- In NGS library preparation, the DNA sample is firstly **fragmented**.

- Next, the specialized **adapters** are ligated to both fragment ends.

illumina sequencing

Step 2. Cluster Amplification

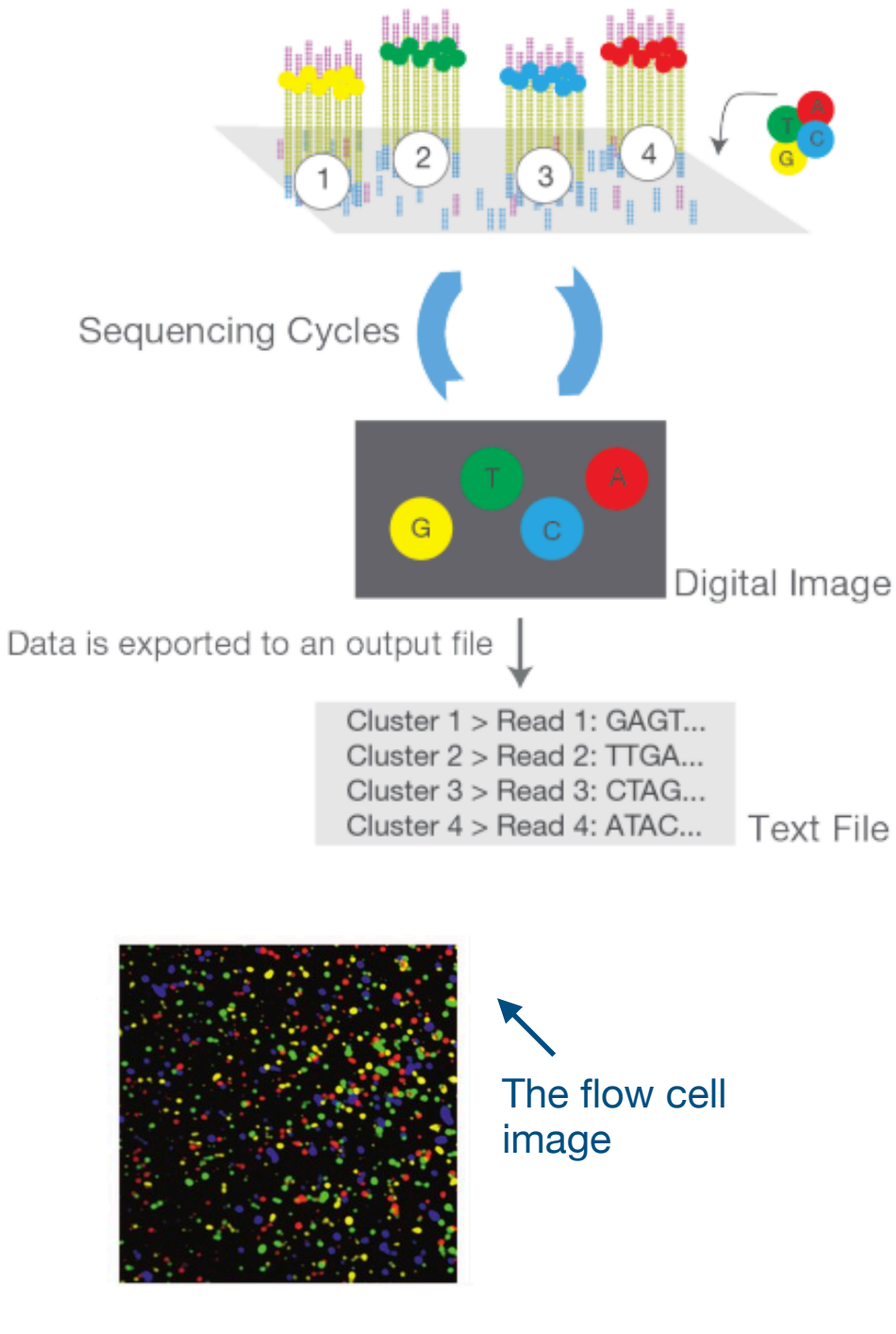


- The library is loaded into a **flow cell** and the fragments are hybridized to the flow cell surface.

- Each bound fragment is amplified into a colonel cluster through bridge amplification (a type of PCR act on flow cell).

illumina sequencing

Step 3. Sequencing

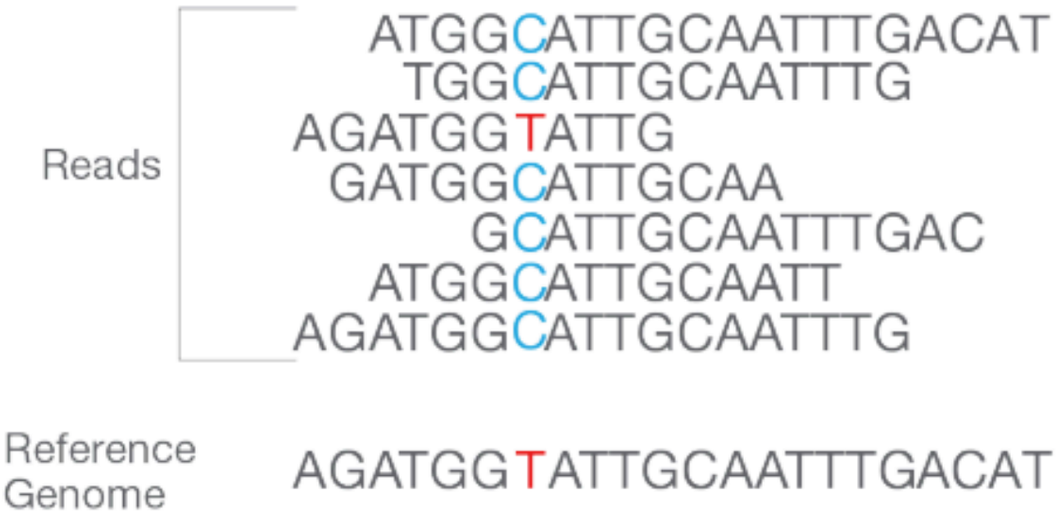


The flow cell image

- Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated.
- The flow cell is imaged and the emission from each cluster is recorded.
- The emission wavelength and intensity are used to identify the base.
- This cycle is repeated “n” times to create a read length of “n” bases.

illumina sequencing

Step 4. Alignment and Data Analysis



- Reads are aligned to a reference genome with bioinformatics software.

- After alignment, differences between the reference genome and the newly sequenced reads can be identified.

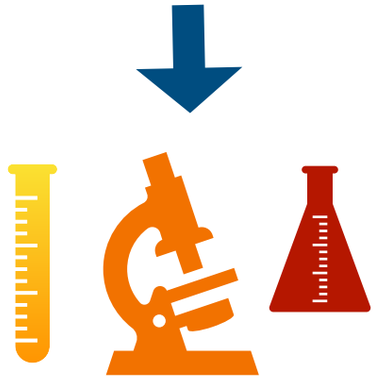
Measure RNA and epigenomes

How can we measure diverse ranges of genomic molecules with Illumina sequencing?

Total DNA/RNA from initial cellular sample

- Poly-A selection
- Antibody enrichment
- Tn5 treatment
- Bisulfite conversion

...

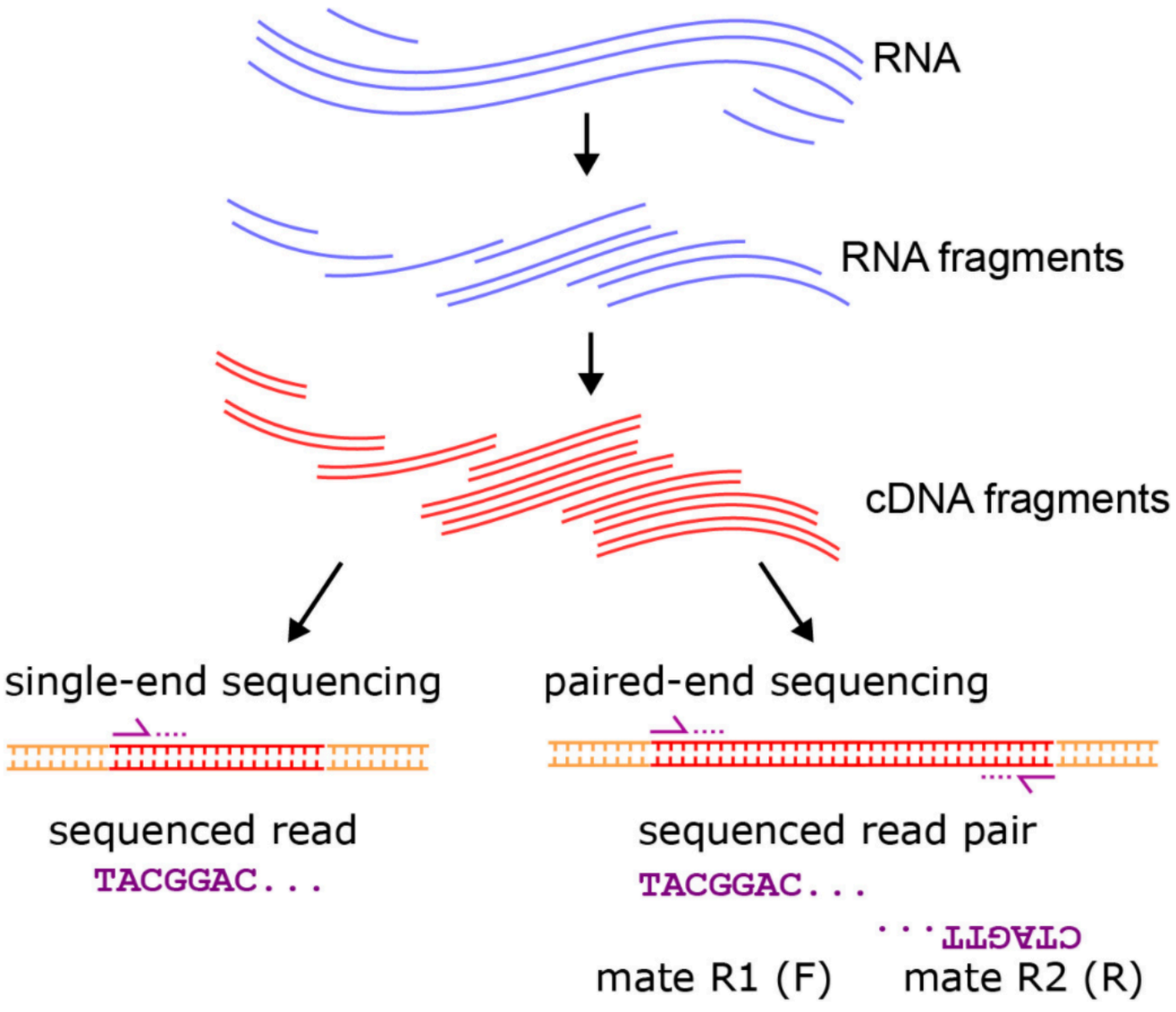


A set of molecular biology experiments that can “extract” a specific portion of genome/ transcriptome.

A DNA library is created to represent the signal of interest, such as mRNA, epigenetic markers, or chromatin conformation.

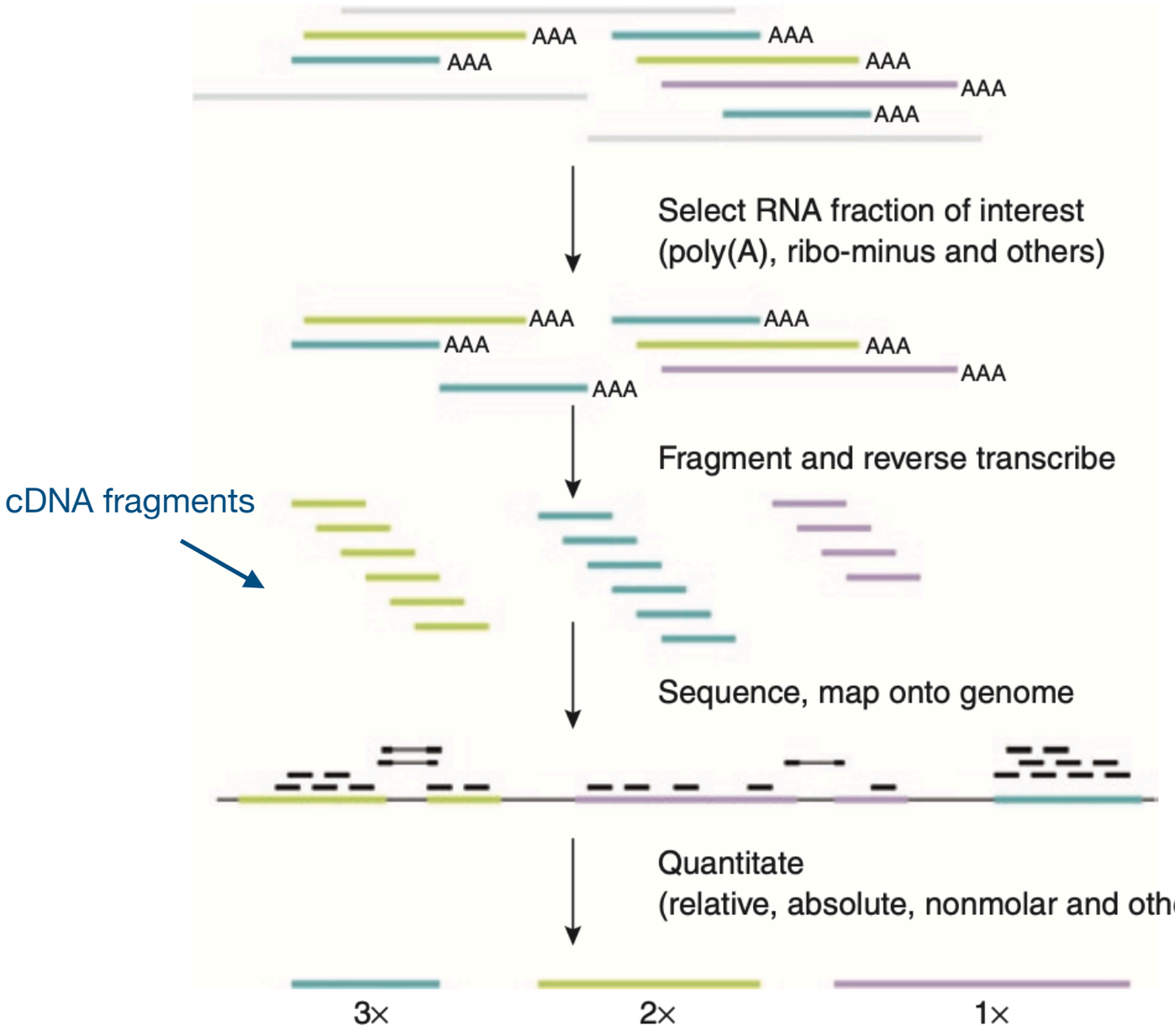
High throughput sequencing

Example: measuring the sequences of (expressed) mRNAs



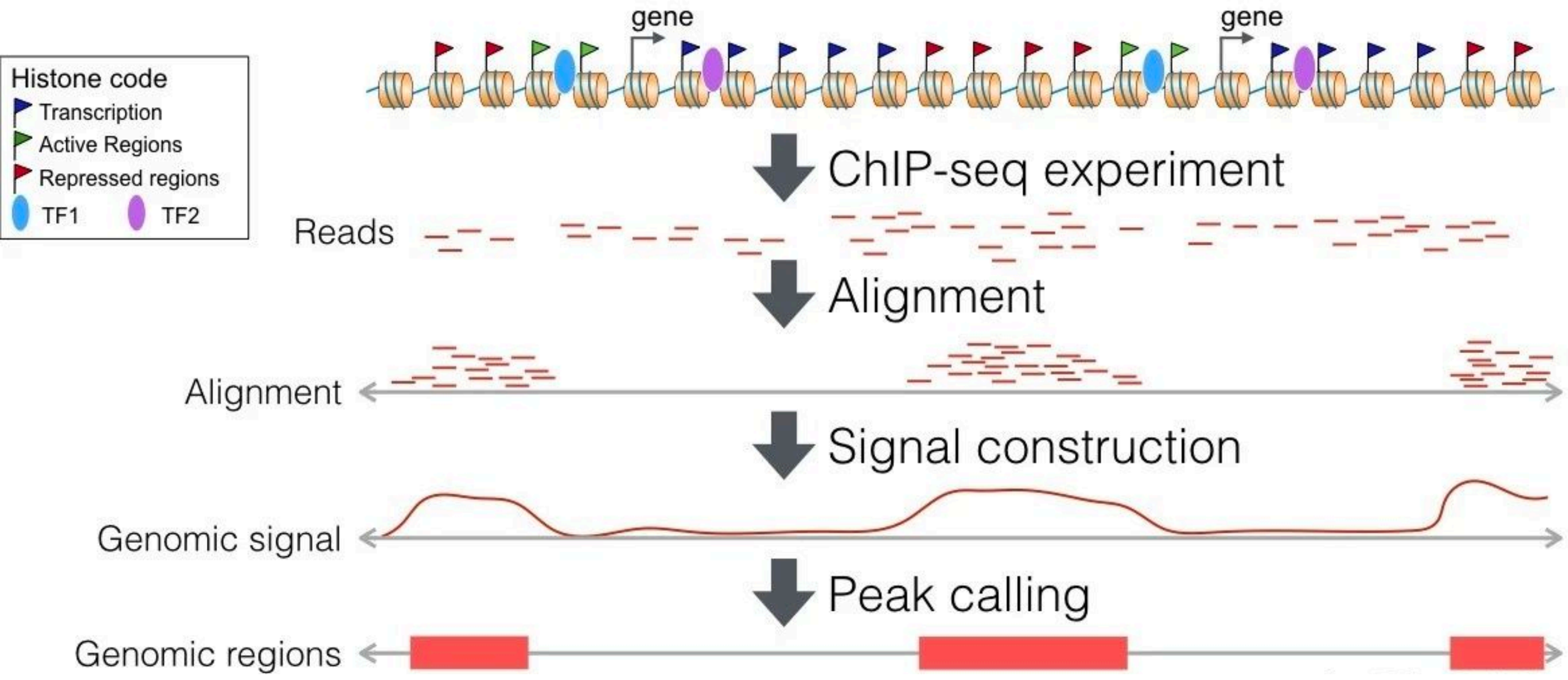
- The RNA transcripts are fragmented to ~200bp fragments.
- The fragments are reverse transcribed to cDNA, and the cDNA fragments are then **amplified by PCR**.
- 2 types of DNA libraries can be generated for illumina sequencing: single-end library and paired-end library**.**
- **Single-end sequencing** sequence one end of the cDNA fragment.
- **Paired-end sequencing** sequence both ends of the cDNA fragment. The mates within a pair are on different strands of the cDNA fragment.

RNA-seq



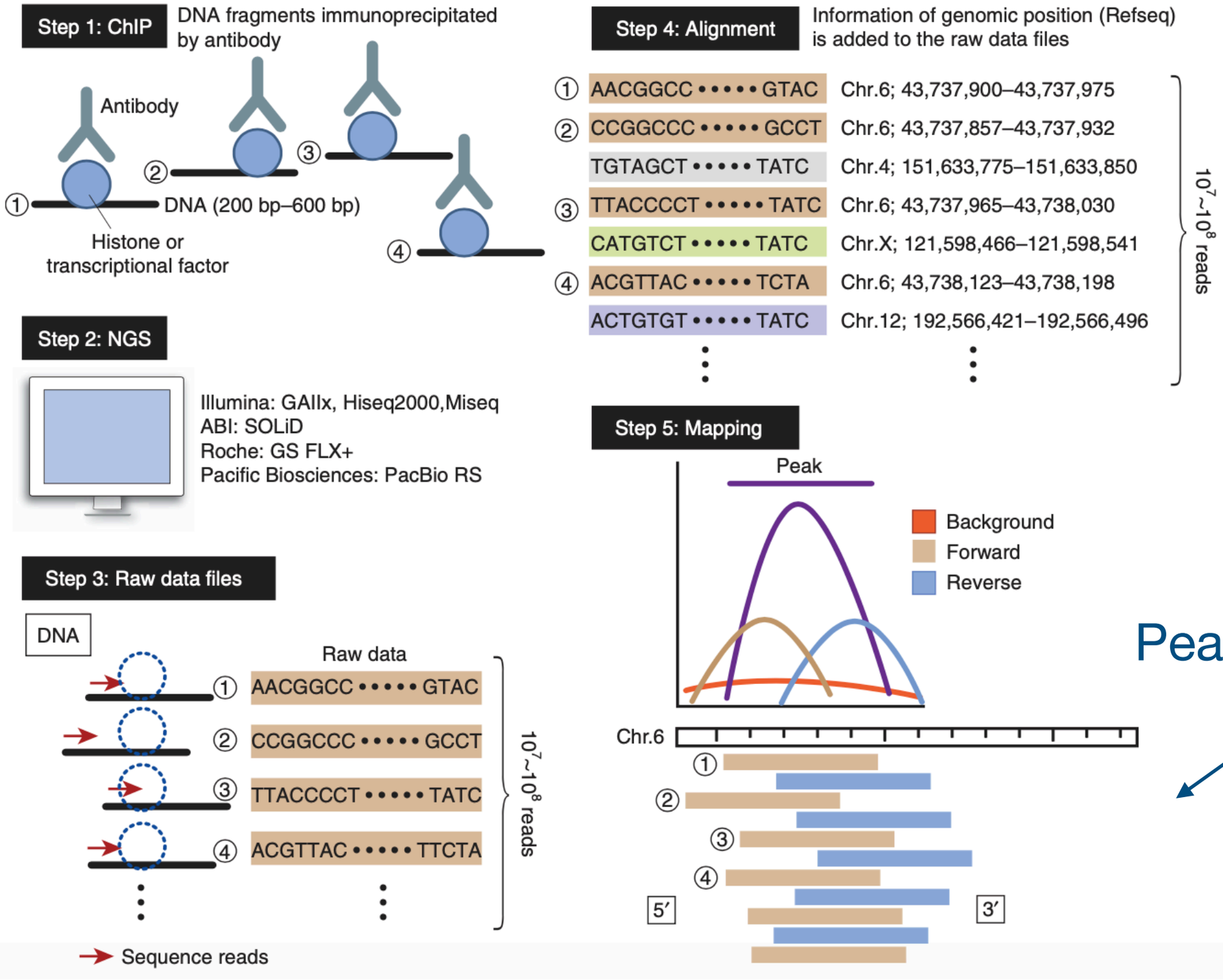
- **Poly-A selection** or **ribosomal RNA removal** are used to enrich the mRNA component of total RNA.
- Genome mapping software is used to align sequencing reads to the genome.
- Gene expression levels can be quantified by counting the aligned reads mapped to the gene annotations.

How to use illumina sequencing to profile epigenetic modifications?

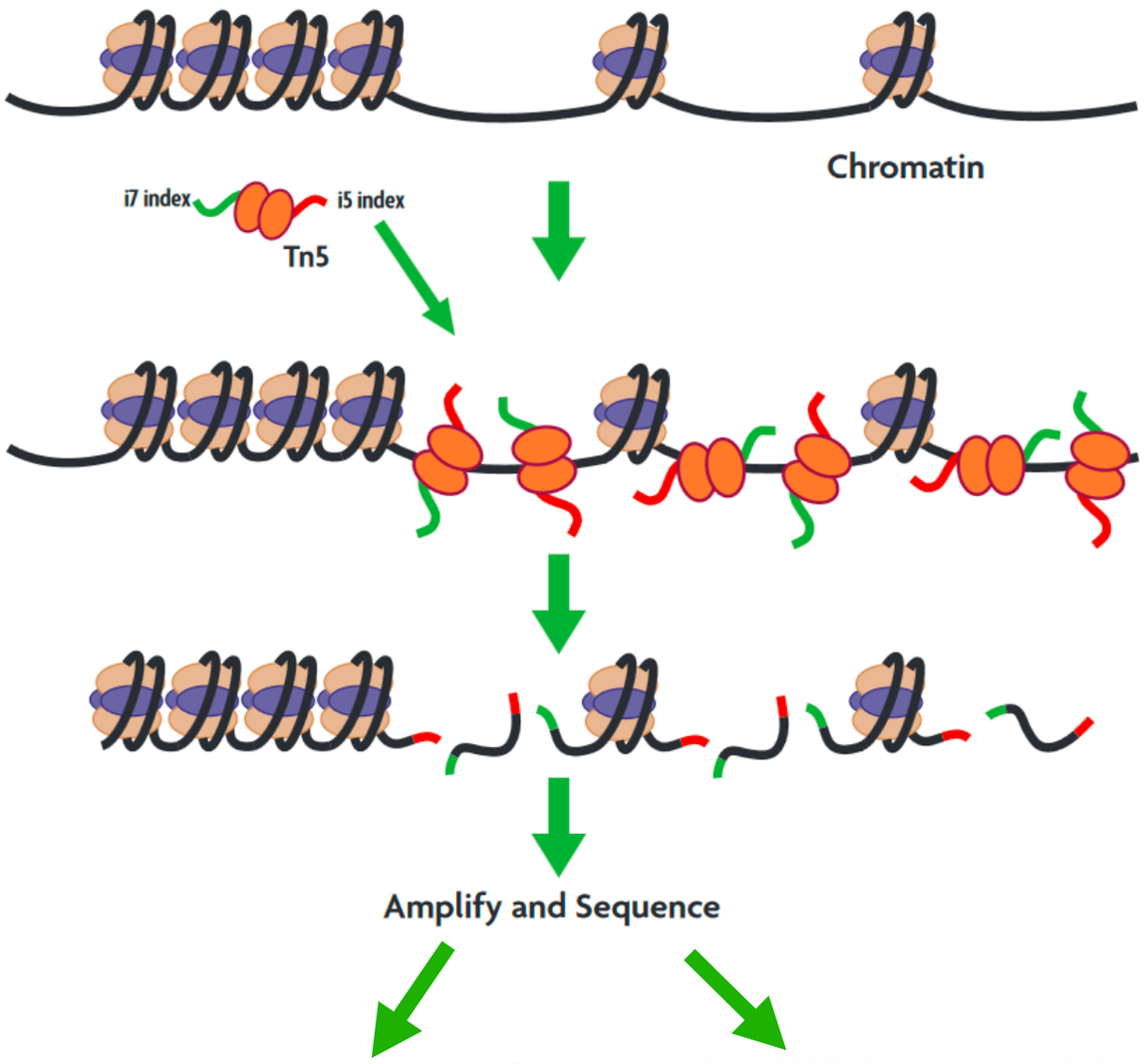


- The histone binding DNA fragments are enriched with immuno-precipitation.

ChIP-seq

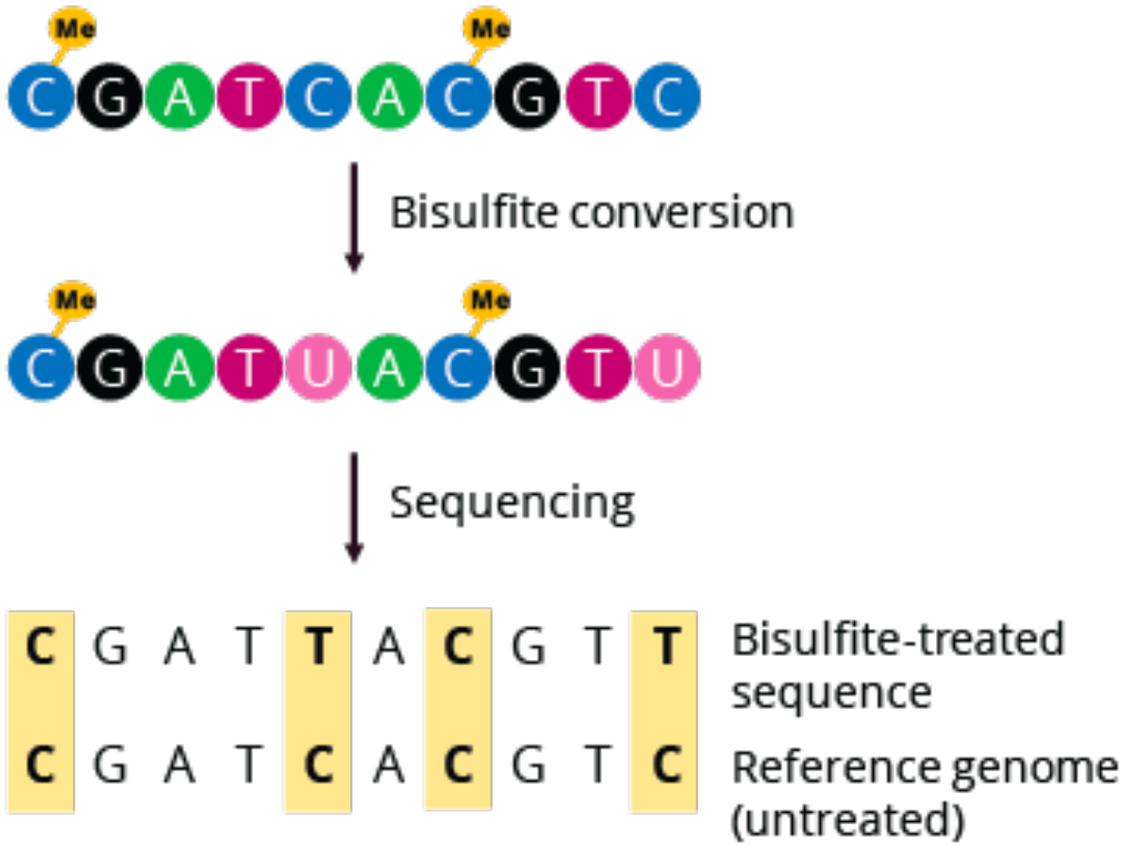

Step 1: ChIP
DNA fragments immunoprecipitated by antibody
Antibody
DNA (200 bp–600 bp)
Histone or transcriptional factor
Step 2: NGS
Illumina: GAIIx, Hiseq2000,Miseq
ABI: SOLiD
Roche: GS FLX+
Pacific Biosciences: PacBio RS
Step 3: Raw data files
DNA
Raw data
AACGGCC • • • • • GTAC
CCGGCCC • • • • • GCCT
TTACCCCT • • • • • TATC
ACGTTAC • • • • • TTCTA
10^7~10^8 reads
Sequence reads
Step 4: Alignment
Information of genomic position (Refseq) is added to the raw data files
AACGGCC • • • • • GTAC Chr.6; 43,737,900–43,737,975
CCGGCCC • • • • • GCCT Chr.6; 43,737,857–43,737,932
TGTAGCT • • • • • TATC Chr.4; 151,633,775–151,633,850
TTACCCCT • • • • • TATC Chr.6; 43,737,965–43,738,030
CATGTCT • • • • • TATC Chr.X; 121,598,466–121,598,541
ACGTTAC • • • • • TCTA Chr.6; 43,738,123–43,738,198
ACTGTGT • • • • • TATC Chr.12; 192,566,421–192,566,496
10^7~10^8 reads
Step 5: Mapping
Peak
Background
Forward
Reverse
Chr.6
5′
3′
Peak Calling

ATAC-Seq

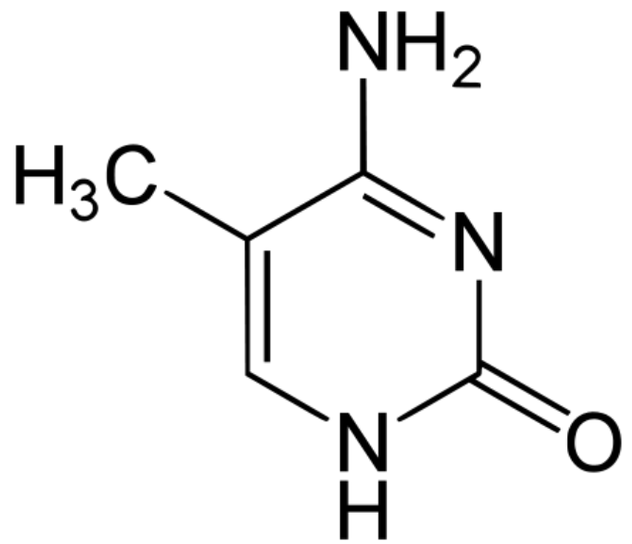


- **ATAC-seq** is a technique for epigenetic profiling that can detect open chromatin regions in a genome.
- The DNA sample is treated with **Tn5 transposase**, which introduces sequencing adapters into the accessible regions of the genome.
- The adapter-ligated fragments are then sequenced, and the sequenced library can be mapped to the accessible regions of the genome.

Bisulfite sequencing



5-Methylcytosine (m^5C)



- **Bisulfite sequencing** is used to profile DNA m^5C methylation.
- The DNA fragments are treated with bisulfite, which convert the C into T. However, methylation on C can protect the nucleotide from conversion.
- Methylation on reads is inferred by tracking nucleotide conversion events.

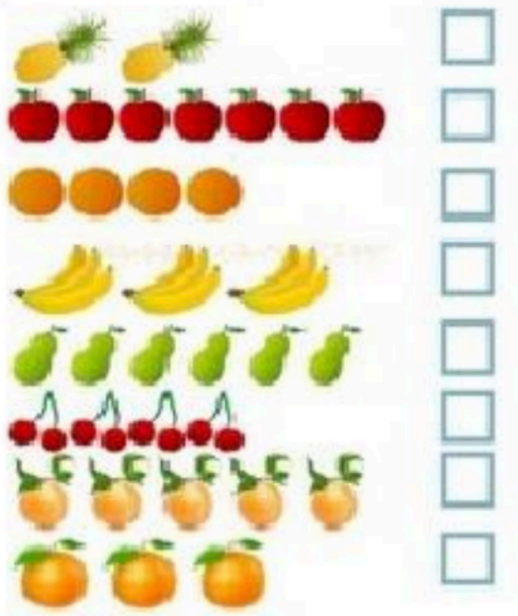
Single cell omics technique

How to obtain genomics data from individual cells rather than from a mixture of cells?

RNA-Seq

- Obtain average expression level.
- Homogenous in expression signals.

scRNA-Seq



- Separate cell populations.
- Detect heterogeneity.
- Identify rare cell populations.

scRNA-Seq

Bulk RNA sequencing

No change of expression of Gene X

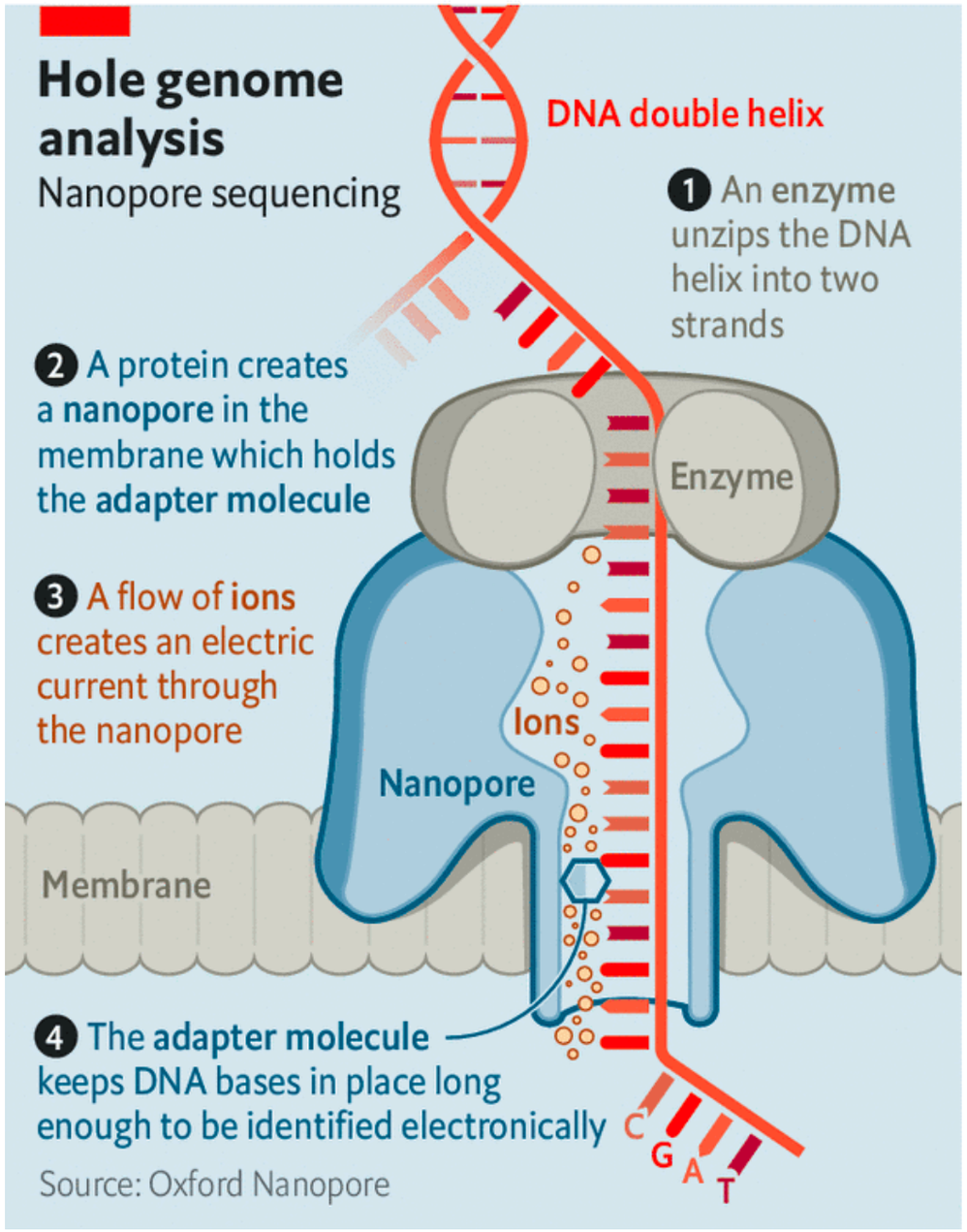
Single-cell RNA sequencing

Expression of Gene X is affected in cell type b only

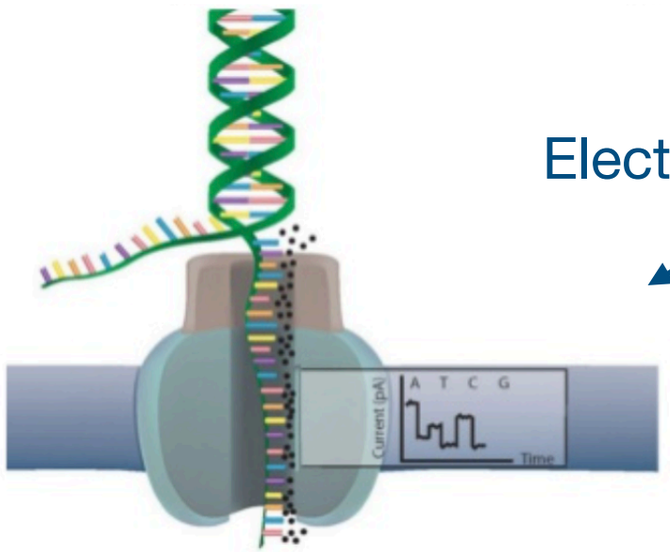
- Bulk RNA-Seq measures the average gene expression of a cell population, making it difficult to deconvolve individual cell expression profiles from a mixture.
- By using the cell specific library preparation techniques (e.x. cellularly unique barcodes), each sample in scRNA-Seq represents a single cell.
- There are also single cell assays to measure DNA sequences (scDNA-Seq), DNA methylation (scBS-Seq) and chromatin conformation (scATAC-Seq).

Real time sequencing

How to read sequences in “real time”?



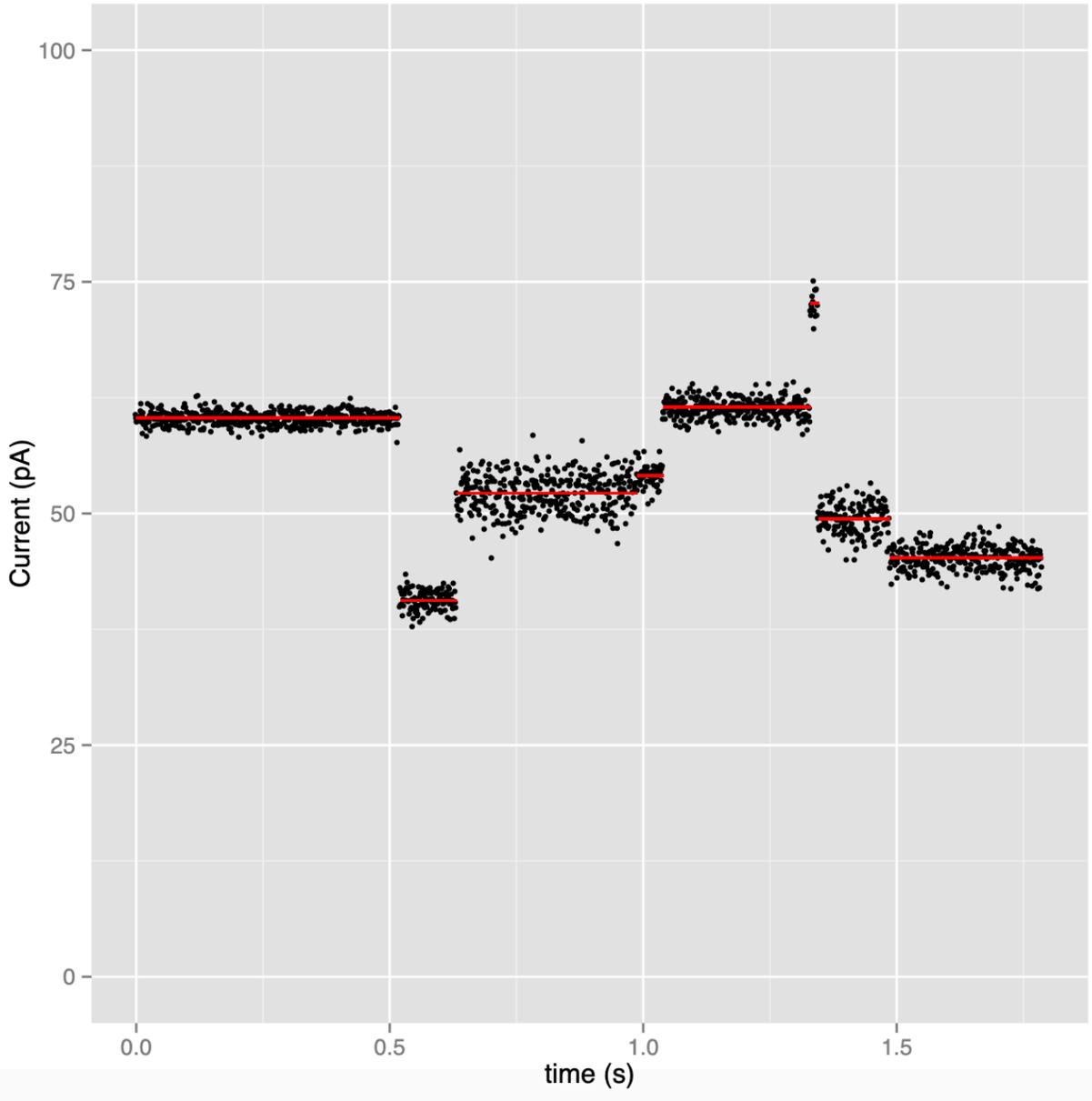
- Nanopore sequencing directly measures single DNA/RNA molecules without PCR amplification.
- The technique measures voltage changes as the molecule passes through a nanopore transmembrane protein.
- Maximum read length can be up to 100kbp.
- Base modifications can be detected, albeit with some noise, by analyzing signal alterations.



Electric signals read out

Oxford Nanopore Sequencing

Base calling with electric signal



5-mer	μ_k	σ_k
AAAAA	53.5	1.3
AAAAC	54.2	0.9
…	…	…
TTTTG	65.3	1.8
TTTTT	67.1	1.4

- The electric signal of a nucleotide is predicted by the 5 nucleotides upstream of the current position, which is called the 5-mer sequence.
- A classic statistical model is $\text{pA} \sim N(\mu_k, \sigma_k)$.
- pA is the current intensity within one nucleotide event, μ_k and σ_k are mean and standard deviation determined by the 5-mer sequence.