STAT 35510

Statistical Algorithms for Single-Cell Omics and Related Techniques

Winter, 2025

Jingshu Wang

Course logistics

- Focus of this course:
  - Introduction of the single-cell sequencing techniques and related biological concepts
  - Major scientific questions and types of analyses
  - Major analytical challenges and statistical / machine learning ideas and methods

- What is not taught in detail but can be important
  - Implementation details of each method
  - Data analysis practices
  - Benchmarking to compare performance of different methods
  - Specific applications / case studies

Course logistics

- Suggested prerequisite:
  - Some knowledge of common statistical methods: PCA, regression, clustering algorithms like k-means, spectral methods, permutation et. al.
  - Some knowledge of deep learning methods
- Homework:
  - no homework
- Final:
  - read papers on your chosen topic and perform a mini benchmarking study.

Lecture 1

Introduction: basic biology, single-cell omics, scRNA-seq technique

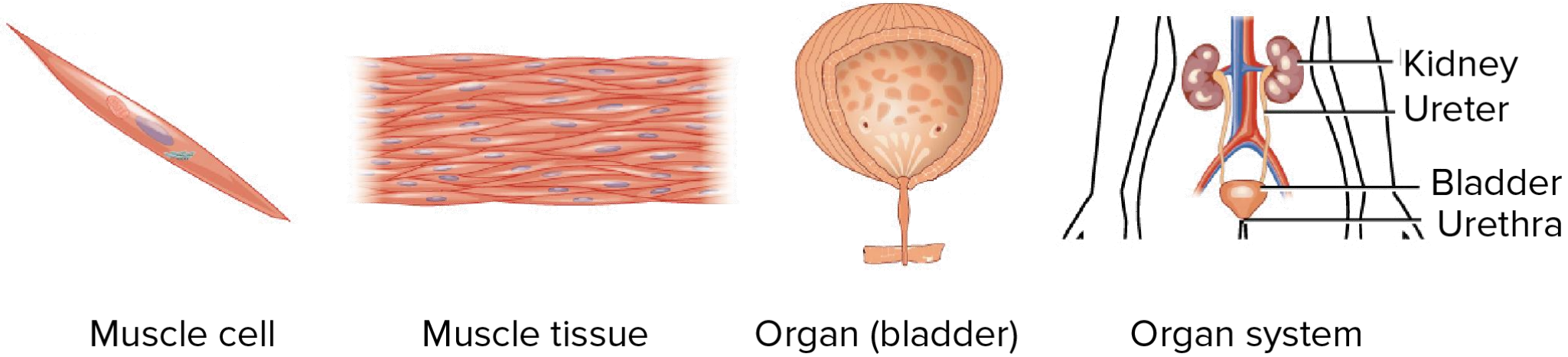
Outline

- Basic biological concepts:
  cell, gene, DNA, mRNA, transcription

- Brief introduction to single-cell omics

- Single-cell RNA sequencing technology
  - Cell capture

The cell is a building unit of an organism

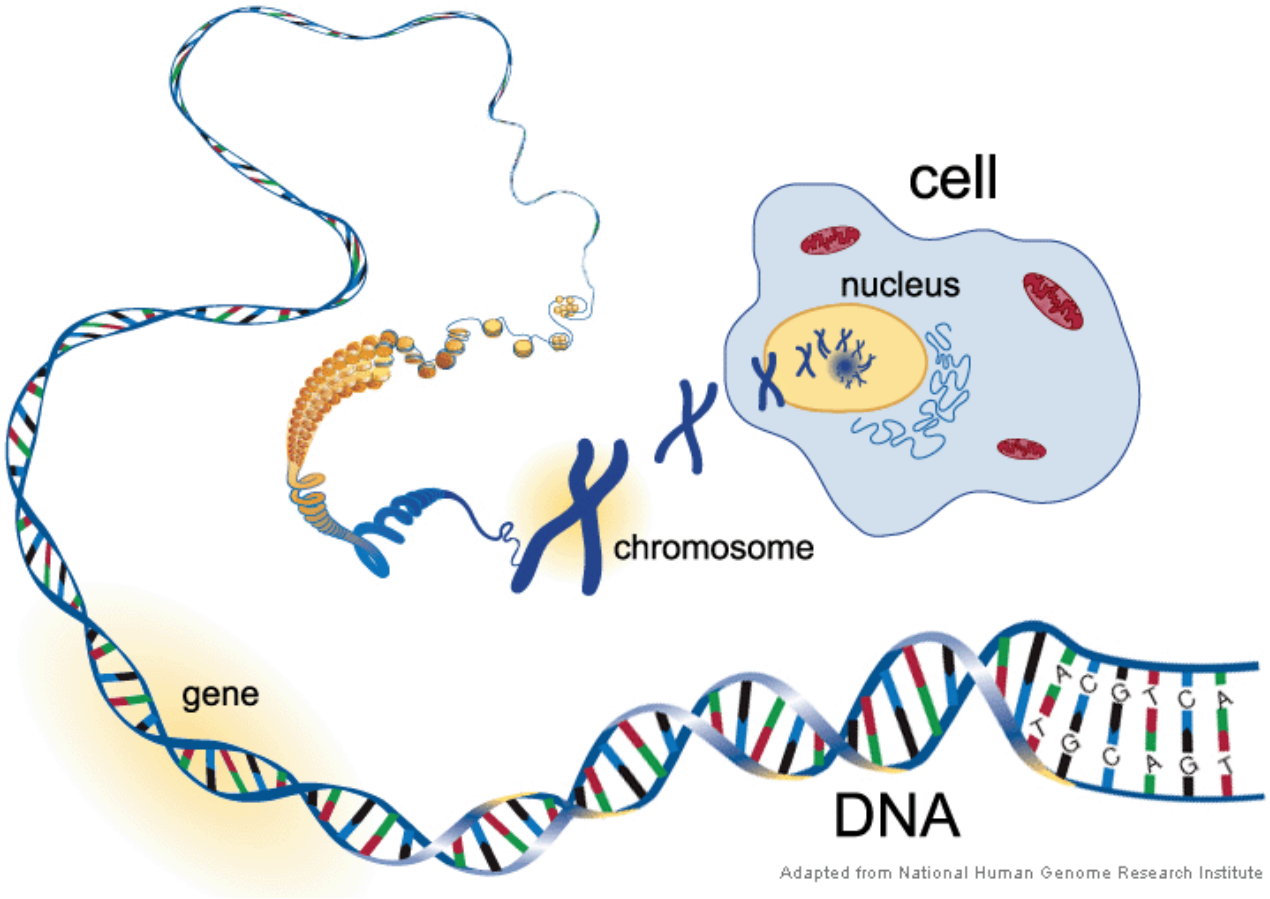


Muscle cell | Muscle tissue | Organ (bladder) | Organ system

- Each tissue can contain a heterogenous population of cells

Image credit: https://www.khanacademy.org/science/biology/principles-of-physiology/body-structure-and-homeostasis/a/tissues-organs-organ-systems, modified from *Levels of structural organization of the human body* by OpenStax College, Anatomy & Physiology, CC BY 4.0

The cell is a building unit of an organism



- Each (eukaryotic) cell has a nucleus
  - Prokaryotic cells (like bacteria) are simple cells without a nucleus
  - We focus on eukaryotic cells
- Nucleus of human somatic cell contains 46 chromosomes
  - 22 pairs of autosomes and one pair of sex chromosomes
  - Chromosome contains compacted DNA
- Gene: piece of DNA that encodes a protein

DNA double helix

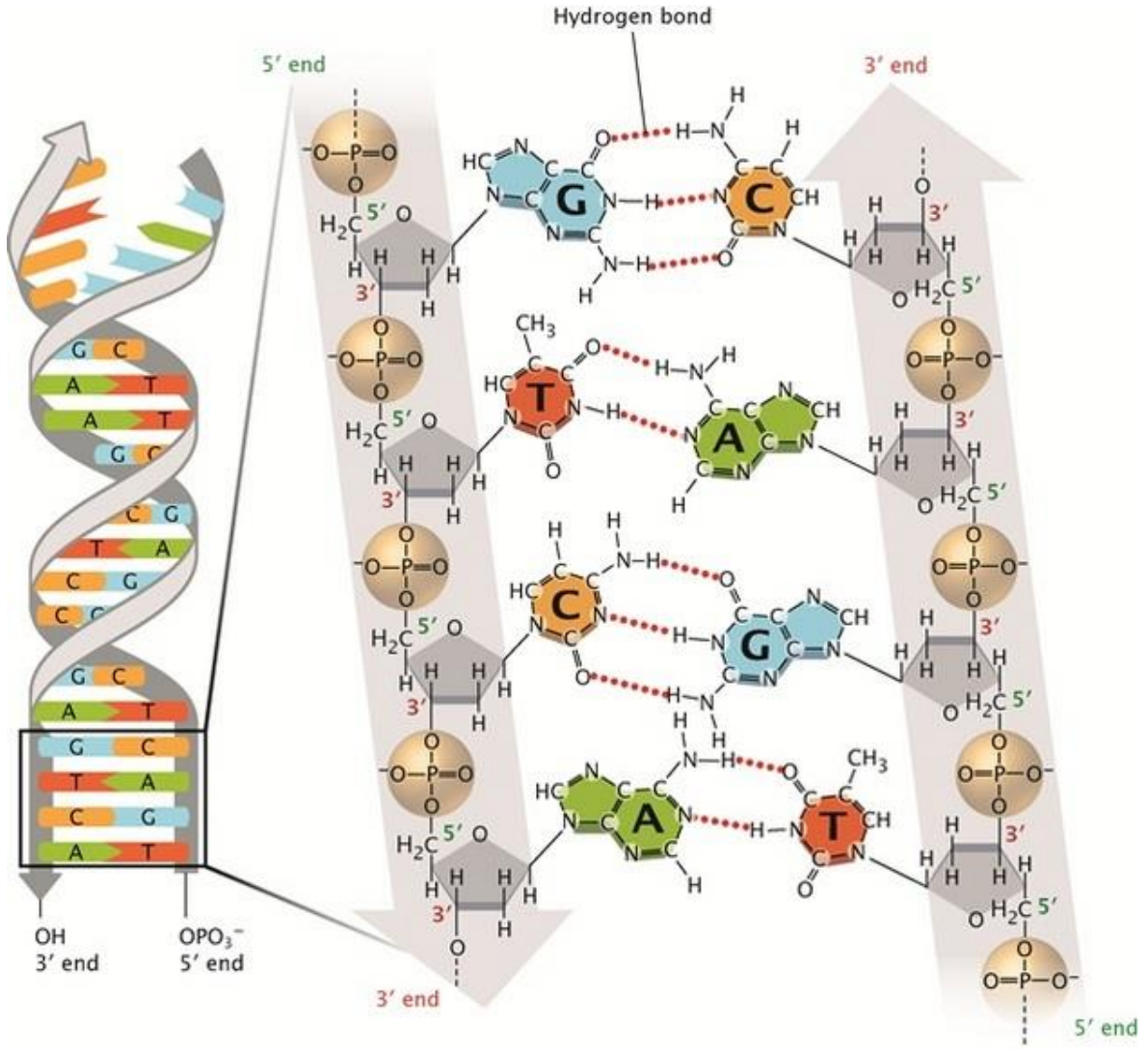
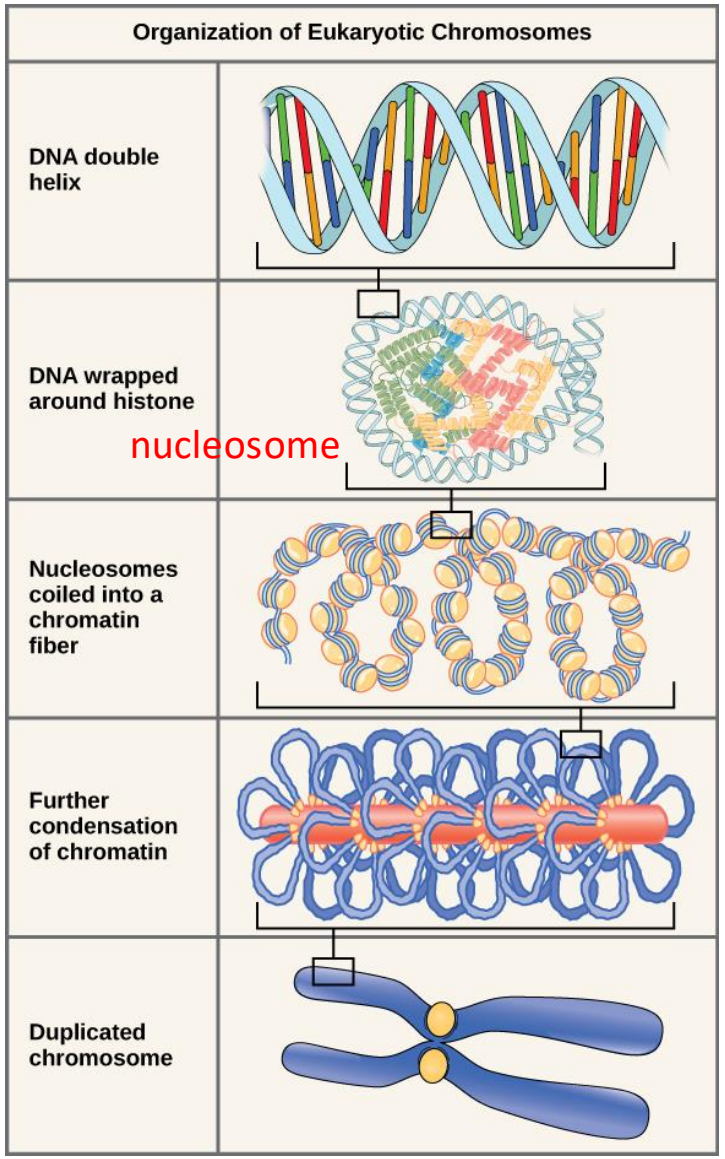


Figure source: https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

- DNA is double-stranded and each strand has a direction
- The DNA molecules never leave the nucleus
  - Use messenger RNA (mRNA) to communicate with the rest of the cell

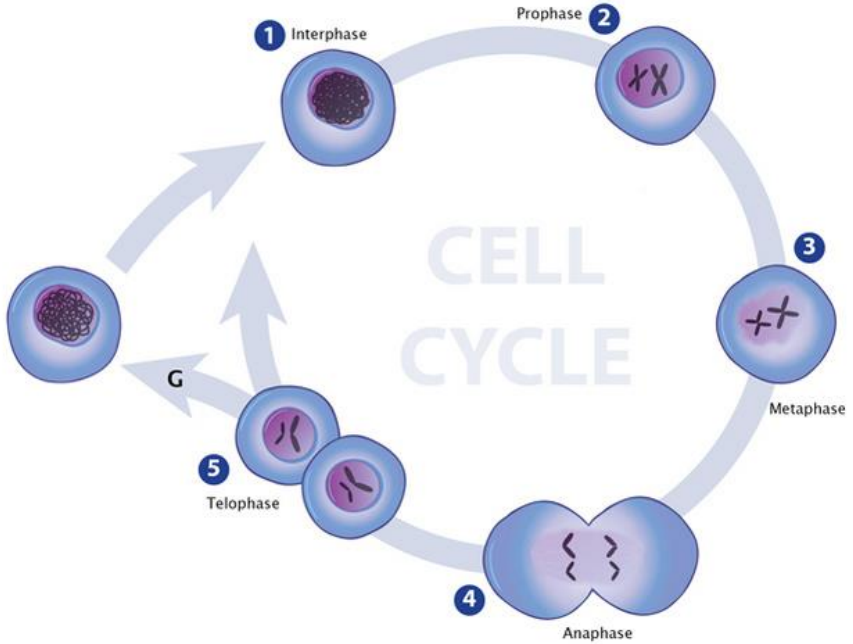
DNA organization inside nucleus



https://courses.lumenlearning.com/wm-biology1/chapter/reading-chromosome-structure/

- DNA is packaged inside the nucleus
  - Make DNA fit into the nucleus and stable
  - Controls the activity of DNA: inactive if tightly packed
- The basic unit is called nucleosome: DNA wrapped around 8 histone proteins

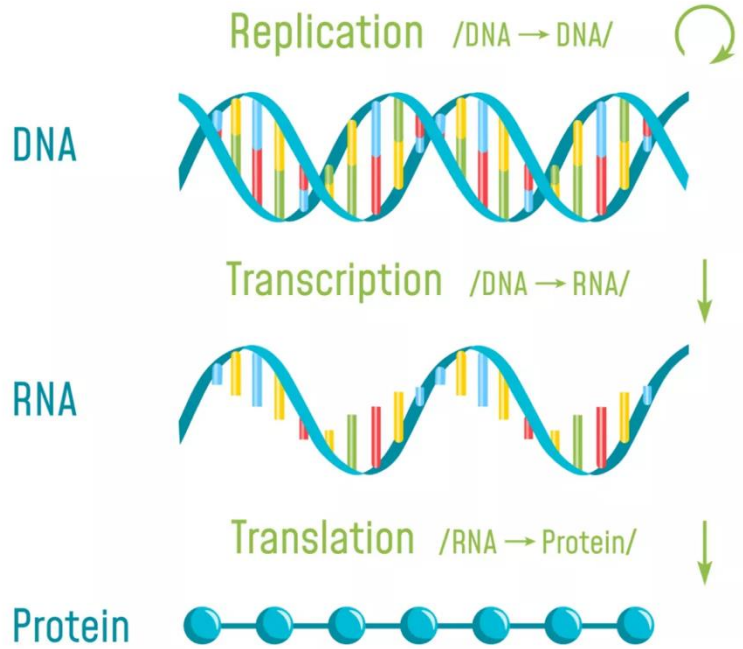
- Different level of condensation during the cell cycle

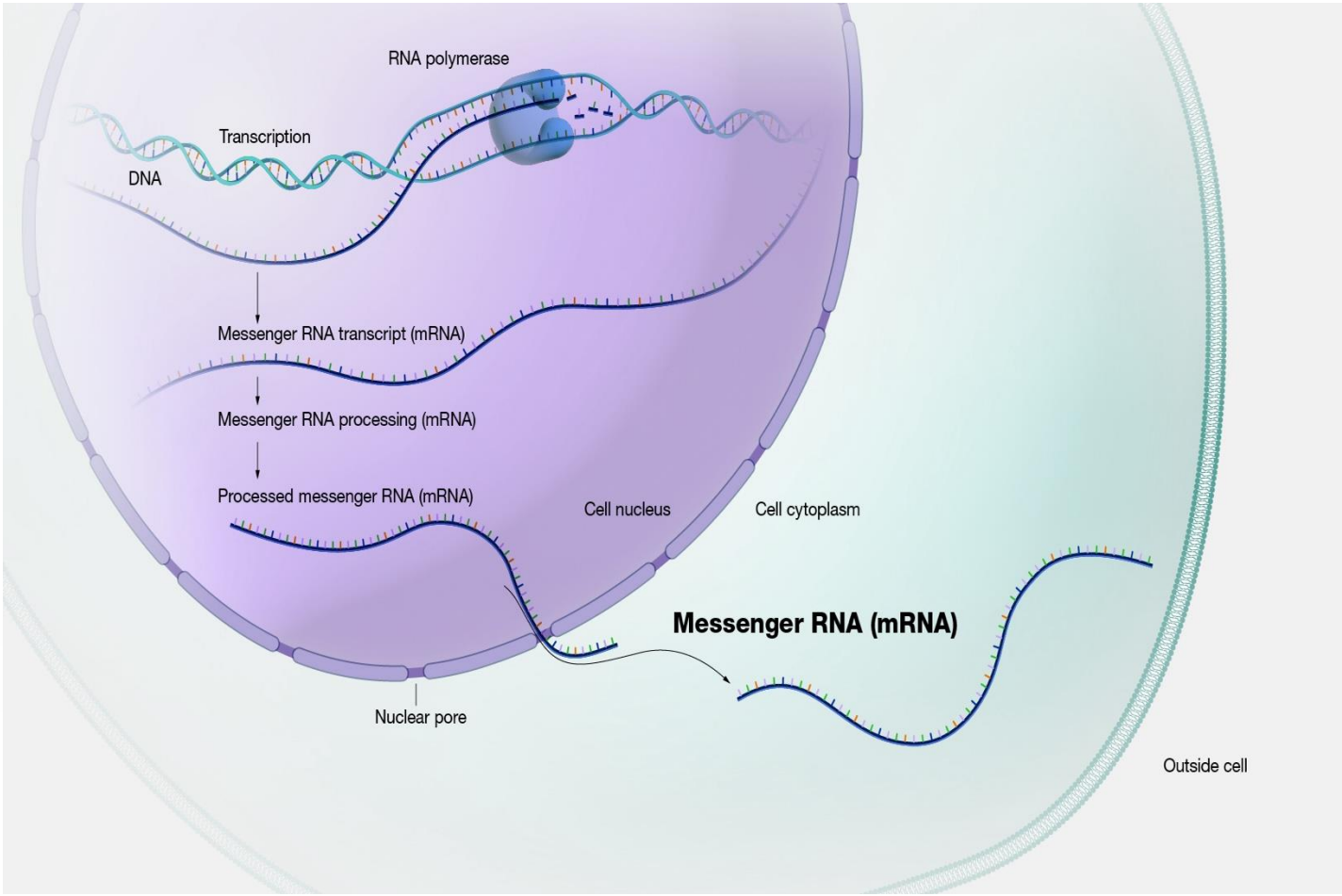


https://www.nature.com/scitable/topicpage/chromosomes-14121320/

- G0 and interphase: chromatin loosely distributed
- Mitosis phase: chromatin are highly condensed

mRNA and transcription



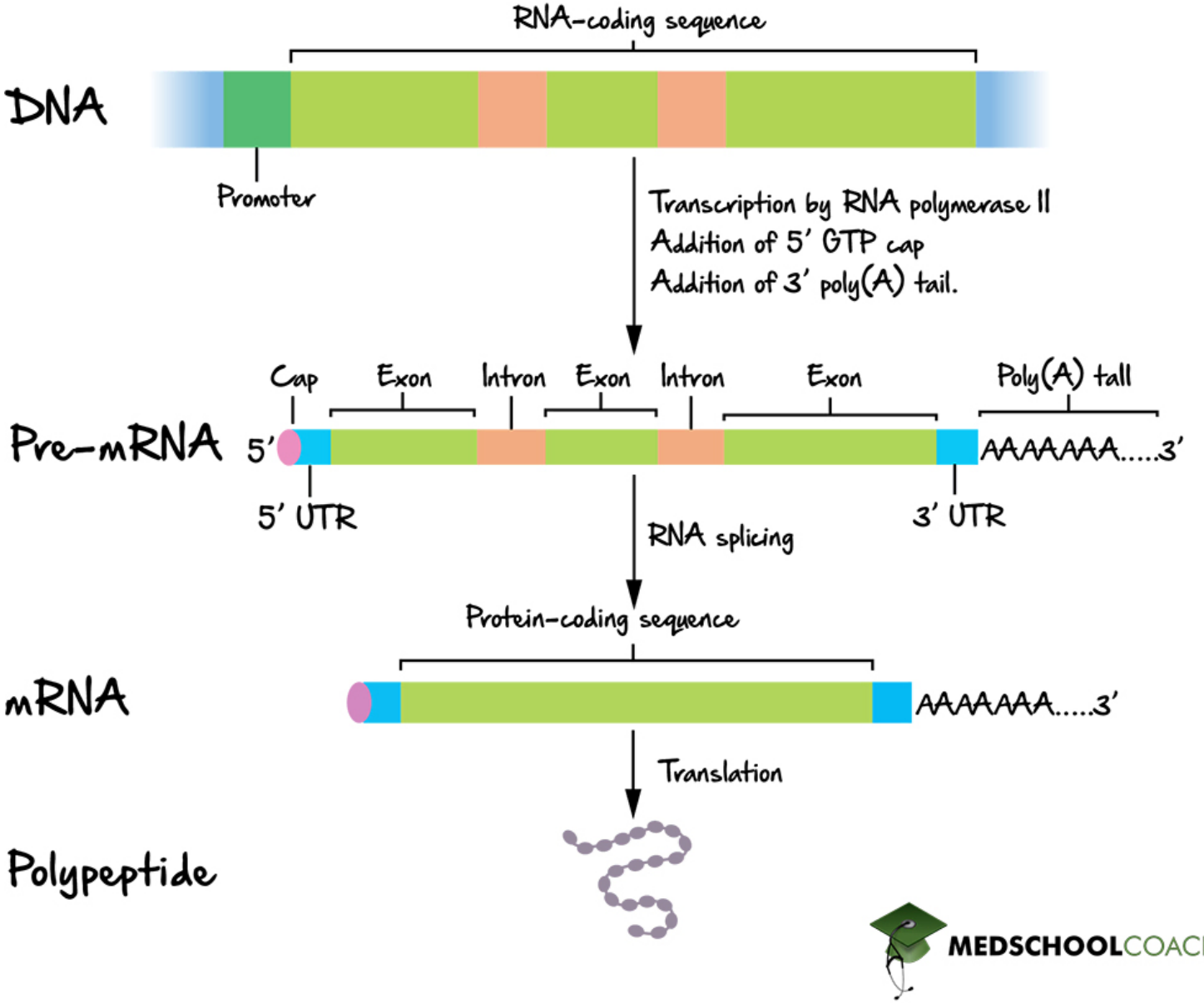


- mRNA: massager RNA that can translate into protein
- Gene expression level: amount of mRNA transcribed by the gene
- Why study mRNA level? Understand which gene is active, function of a cell
  - mRNAs are easier to measure than proteins
- Transcription occurs in the nucleus

Figure sources: https://www.thoughtco.com/dna-transcription-373398, https://www.genome.gov/genetics-glossary/messenger-rna

More details on transcription

- mRNA is single-strand and reads from 5’ to 3’
- DNA first transcribe into pre-mRNA which contains both exons and introns
  - A poly-A tail is added at the 3’ end
- Splicing: introns are removed, exons are stuck back together
  - Alternative splicing
- mRNAs are degraded in the cytoplasm



MEDSCHOOLCOACH

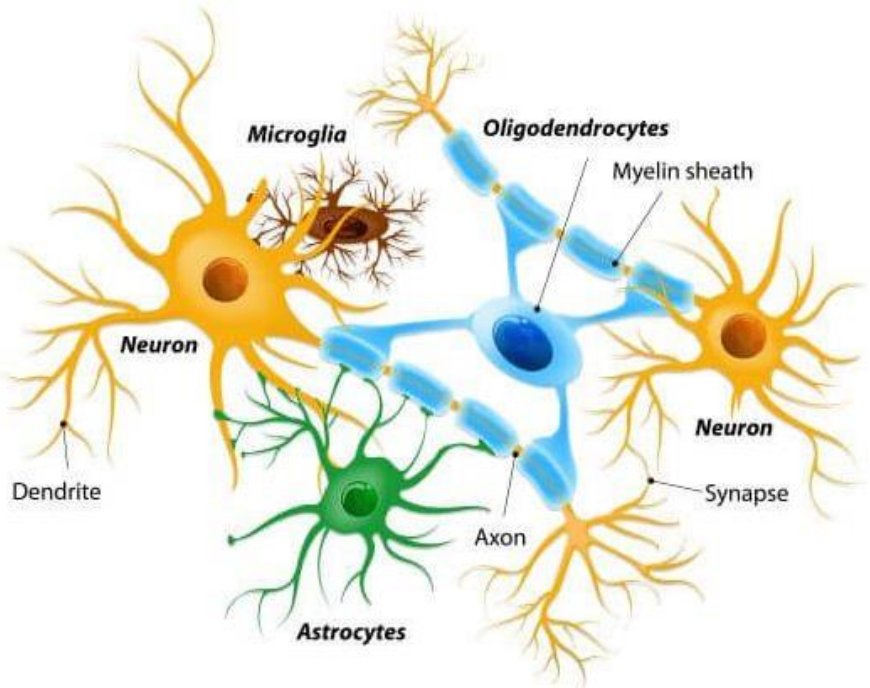
Single-cell omics

- Cellular heterogeneity lies at the heart of biological complexity

- Measure genetic information within for each single-cell for a (possibly biased) sub-sample of all cells in the tissue
  - Collect cells at a particular time point for a tissue and for one individual.
  - The whole experiment can collect samples from multiple time points, multiple tissues and multiple individuals

- Omics: collection of different disciplines such as genomics, transcriptomics, epigenomics, proteomics et. al.
  - Provide high-resolution characterization of individual cells

- Goal:
  - Characterize each cell and Understand cellular heterogeneity under different conditions
  - Understand gene functions



Layers of single-cell multiomics

- Genomics:
  single–cell DNA sequencing (scDNA-seq), understand somatic mutations, cancer intra-tumoral heterogeneity

- Transcriptomics:
  - Single-cell RNA sequencing (scRNA-seq), single-nucleus RNA sequencing (snRNA-seq, for cells that are difficult to isolate)
  - Most widely used and mature technologies
  - Measure the amount of mRNA copies for each gene (and more)

  - Spatial transcriptomics: spatial location of a “cell” + mRNA copies for each gene

  - Single-cell CRISPR screen: Random genetic perturbations in a cell + scRNA-seq
    - Randomized experiments to infer causal relationships

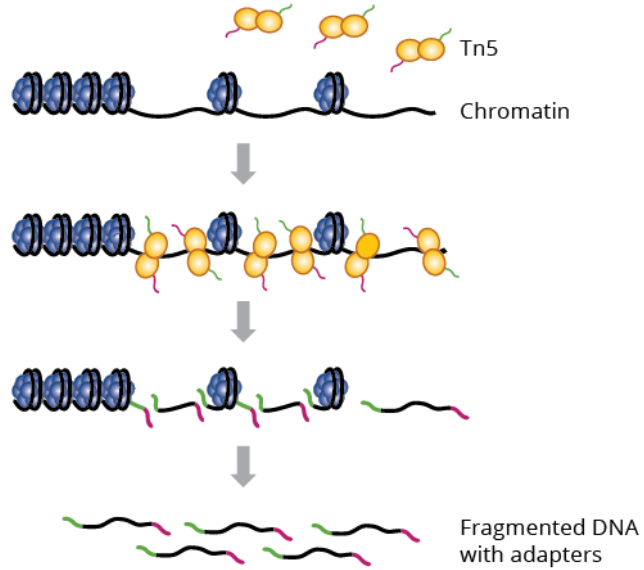
Layers of single-cell multiomics

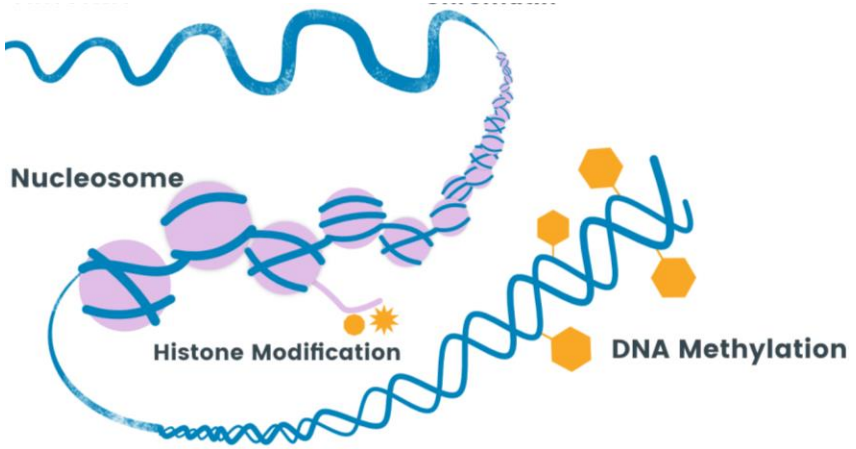
- Epigenomics:
  - Modification of DNA / histones that does not alternative the DNA sequence
  - Understand regulation of gene expression

  - Single-cell ATAC-seq:
    - measure the open regions of DNA (chromatin accessibility)
    - Understand how nucleosome positioning regulates gene expression (transcriptional activation)

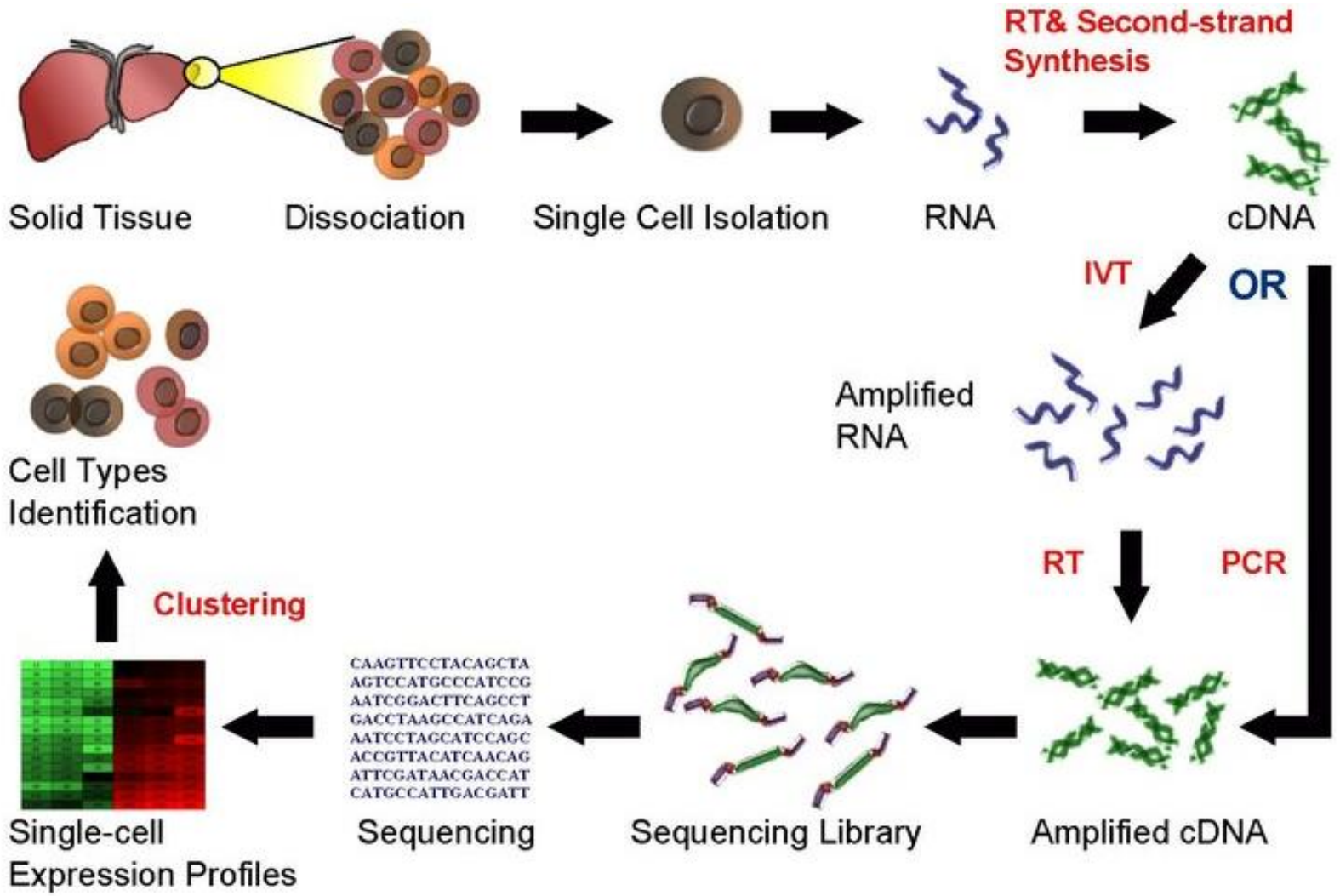
  - Single-cell DNA methylation
    - DNA methylation: a methyl group added to DNA
    - DNA methylation repress gene expression

- Multi-omics:
  - Simultaneously measure 2 or more omics in the same cell

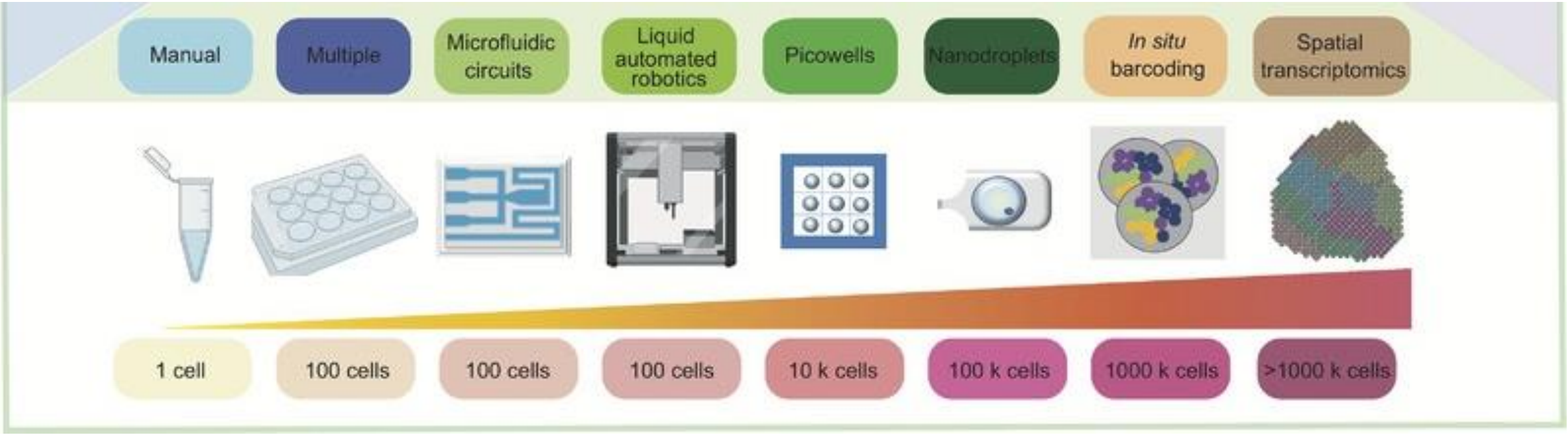




Single-cell RNA sequencing workflow



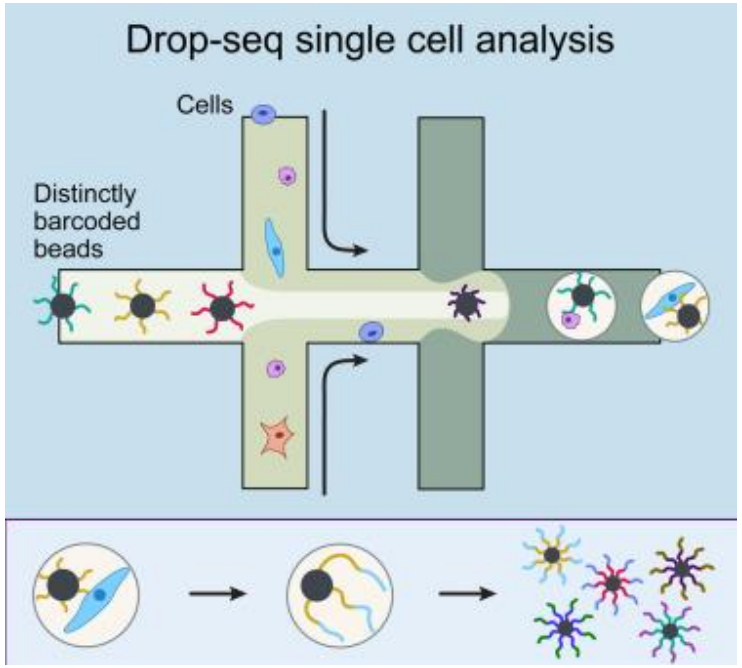
Single-cell capture and barcoding (multiplexing)

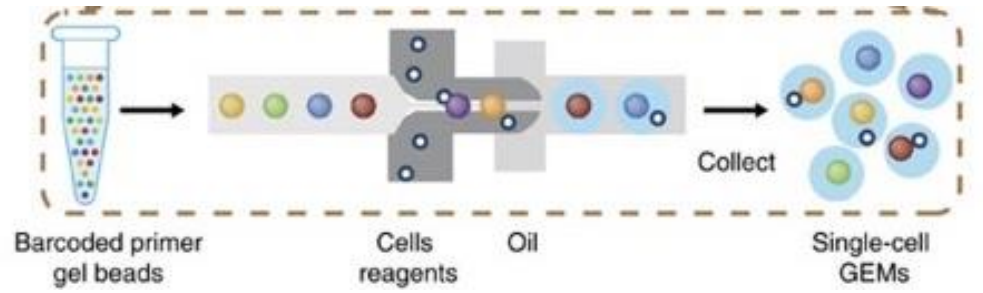


Jovic, Dragomirka, et al. "Single-cell RNA sequencing technologies and applications: A brief overview." *Clinical and Translational Medicine* 12.3 (2022): e694.

Droplet-based methods

- Drop-seq (Klein et al. Cell 2016), InDrop (Macosko et al. Cell 2016) and 10X genomics (Zheng et. al., Nature Comm. 2017)



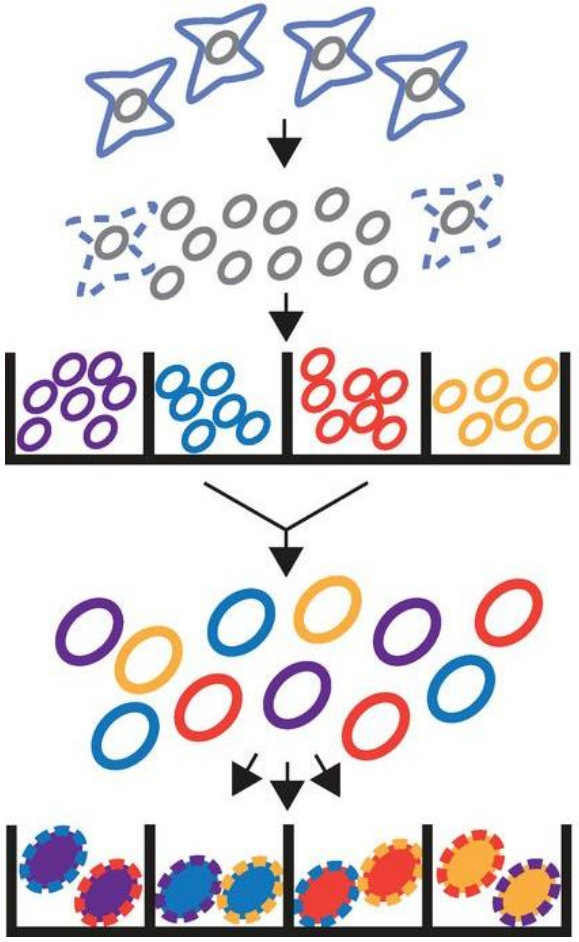


- Single-cell suspension and the bead suspensions are flowed at (equal) rate to form a droplet with one cell and one bead
  - Cells follow a Poisson process
  - Number of cells in a droplet follows a Poisson distribution → low rate to avoid doublets
  - Hydrogel Beads can achieve a "super-Poisson" process (inDrop and 10X) → increase cell capture efficiency
  - Can computationally remove doublets to increase efficiency

- Such bead contain tens of thousands primers (which RNAs will attach to) with the same cell barcode

Combinatorial Indexing method

- Use a combinatorial indexing method to uniquely label the transcriptomes of large numbers of single cells or nuclei (sci-RNA-seq, Cao et. al. Science 2017)



- Isolated cells/nuclei are first randomly distributed and tagged into different wells (96 or 384)
- Cells are then pooled and redistributed to a second set of wells to get a second barcode
- Rationale: the chance that two cells have exactly the same combinatorial barcode is relatively low
- Idea: Limit the number of cells per well to enter the second round to reduce collision rate

Combinatorial Indexing method

- Assume that there are exactly n cells per well to enter the second round and there are m wells
  - Expected number of combinatorial barcodes that have zero cells

$$m^2\left(1-\frac{1}{m}\right)^n$$

  - Expected number of combinatorial barcodes that have exactly one cell

$$m^2 \times n \times \frac{1}{m} \times \left(1-\frac{1}{m}\right)^{n-1} = nm\left(1-\frac{1}{m}\right)^{n-1}$$

  - Collision rate: proportion of combinatorial barcodes that have more than one cell

$$\frac{m^2 - m^2\left(1-\frac{1}{m}\right)^n - \left(1-\frac{1}{m}\right)^{n-1}}{m^2 - m^2\left(1-\frac{1}{m}\right)^n}$$

- If $n = 50, m = 384$, we can get $N = 19200$ cells with a collision rate around 6.2%

Combinatorial Indexing method

- What if we do not limit the number of cells in each well that enter the second round?
- Assume that there are N cells in total there are m wells, for both rounds we randomly assign cells into each well
  - Consider one cell, any other cell has the same combinatorial barcode with it has probability $\frac{1}{m^2}$
  - For this particular cell, number of cells that has the same barcode follows a Binomial distribution $\mathrm{Binomial}(N-1, \frac{1}{m^2})$
  - Expected proportion of cells that have non-unique barcode

$$1-(1-\frac{1}{m^2})^{N-1}$$

  - If $N = 19200, m = 384$, this expected proportion is 12.1%, collision rate should also be around 6%
- Statistically, it does not matter if one limit the total number of cells, or limit the number of cells per well in the first round