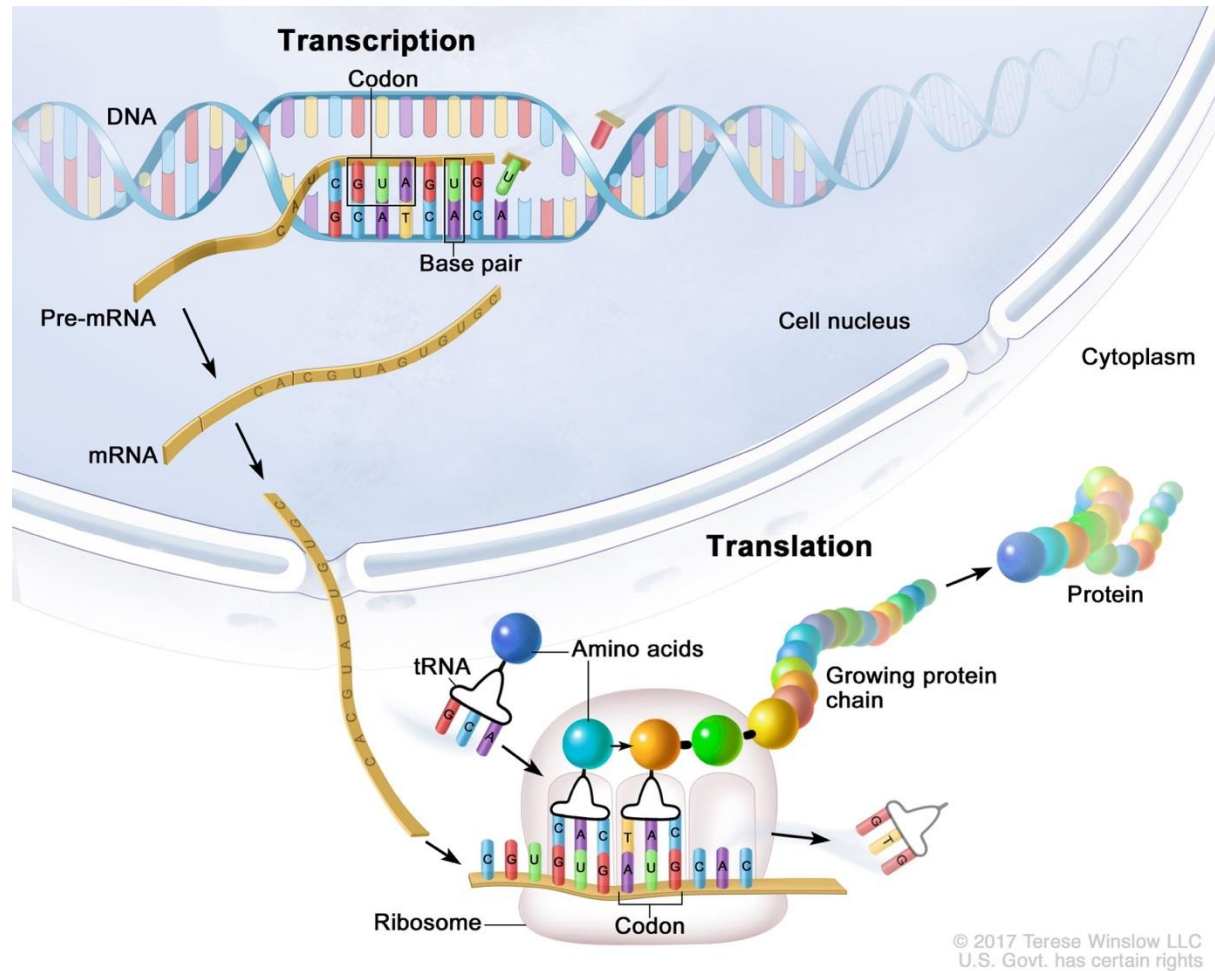


Transcriptomics 101

Zulekha A. Qadeer, PhD
Postdoctoral Fellow
William A. Weiss Lab, UCSF

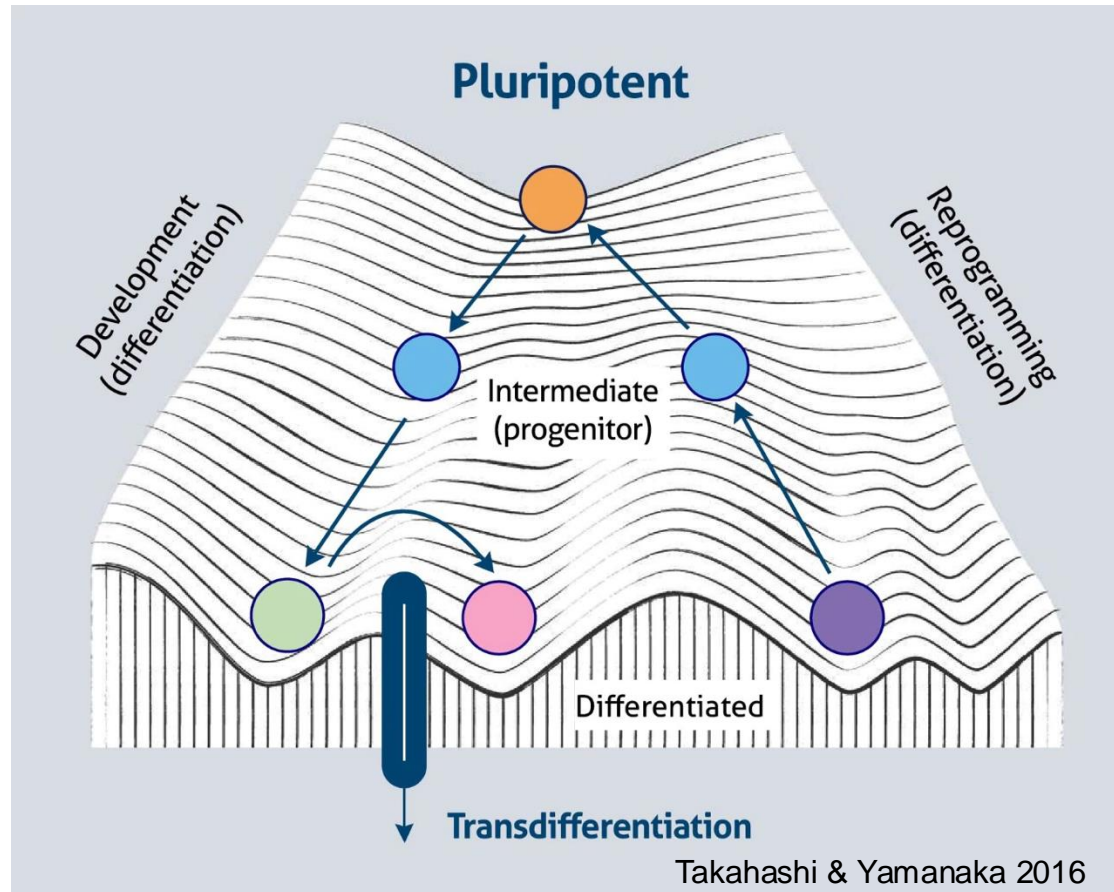
PROPEL Literature Review
December 5, 2024

Overview of Gene Regulation



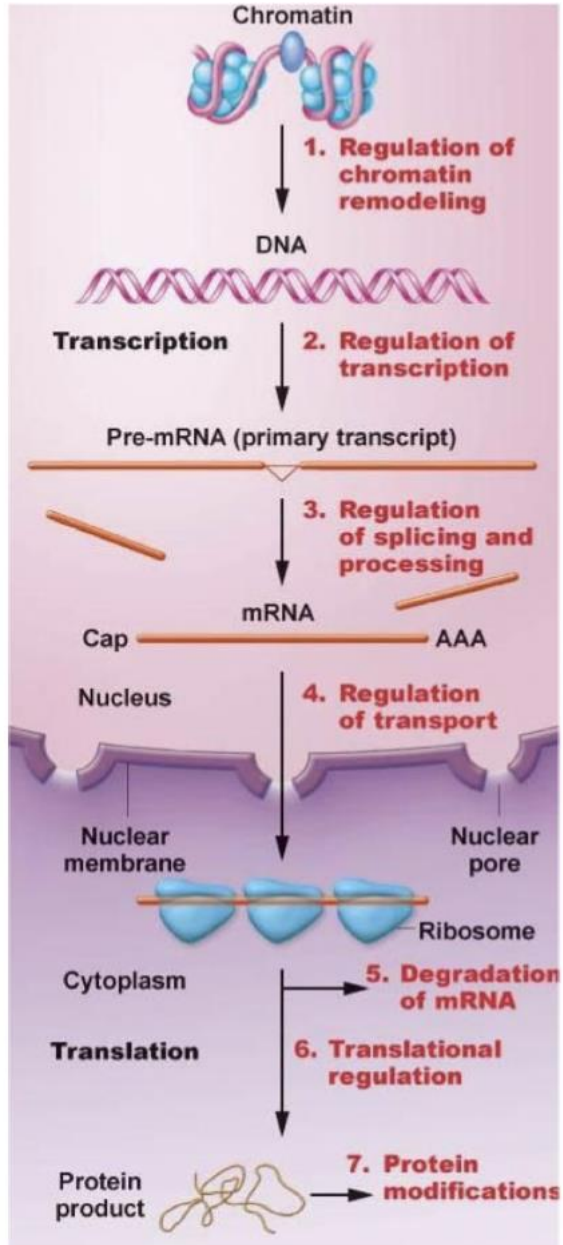
- **Transcription** is converting the coding information or DNA → RNA
 - mRNA then leaves the nucleus via nuclear pores to the ribosome.
- **Transcriptomics** is studying the complete set of RNA or transcripts in a cell and their quantity for a specific developmental stage or physiological condition.

Conrad Waddington's Model for the Epigenetic Landscape



- In 1957, Conrad Waddington described that mammalian development is unidirectional-- embryonic stem cells develop into a more mature differentiated state.
- Expression of genes shape the early landscape of one lineage to maintain a certain cell state or differentiate to another lineage.

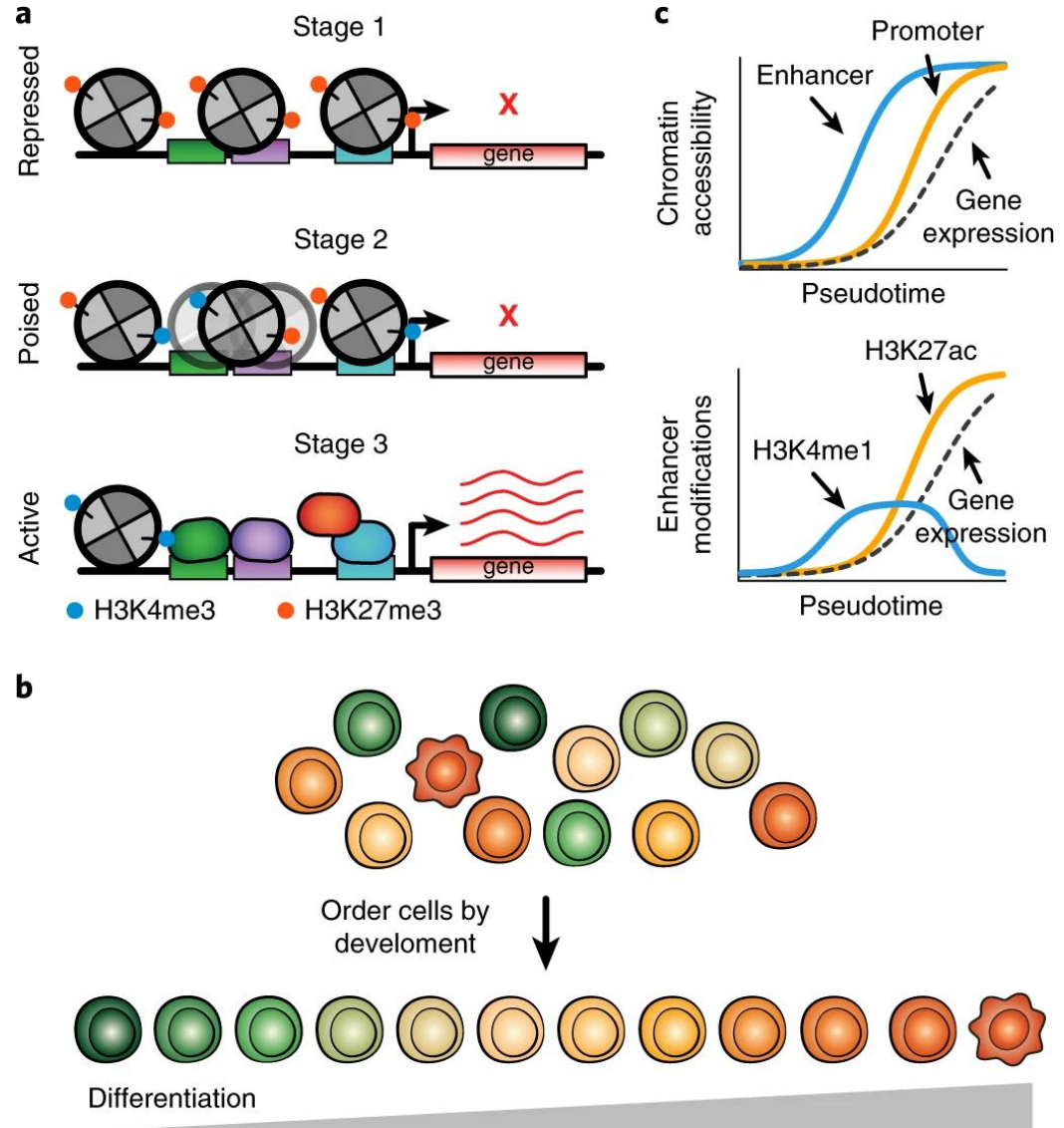
Chromatin remodeling and Transcription



- Chromatin is the term coined for the complex structure of DNA wrapped around core histone proteins.
- Open chromatin allows RNA polymerase and other transcription factors to bind to genes to initiate transcription.
- The **transcriptome** is ultimately the byproduct of a highly coordinated program of gene expression.
- This program is regulated by the epigenome, DNA and histone modifications that impact transcription factors interact with genomic DNA, leading to activation or repression of gene expression.

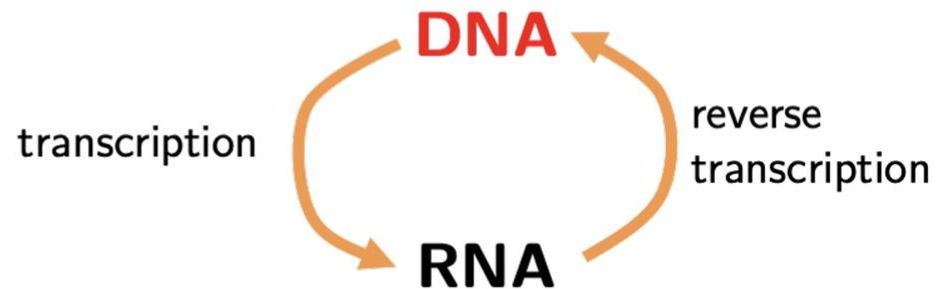
Dynamic Regulatory Changes in Development

- Epigenomic regulation contributes to the complexity of biology by driving developmental event and differentiation.
- This enable cells with the same DNA to take on unique identities and form diverse tissue and organ types.
- This can also drive the cellular heterogeneity that promotes complex disease biology.



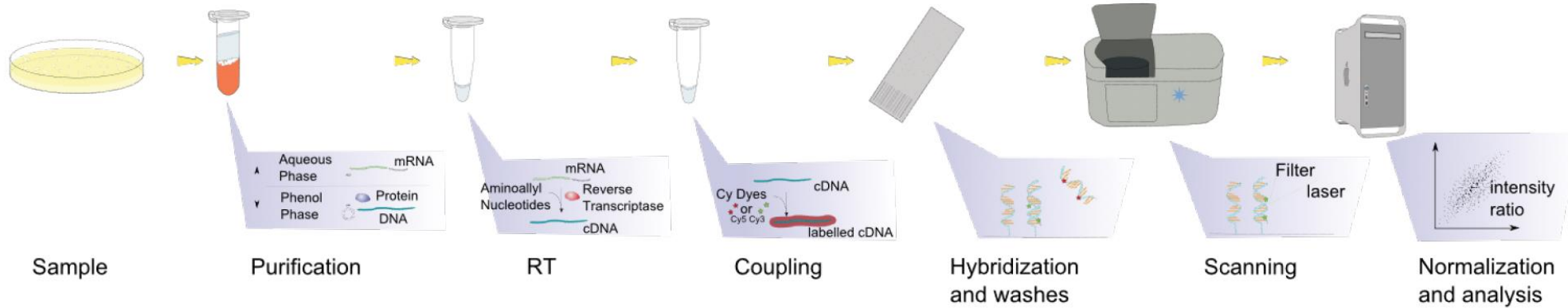
Study of the Transcriptome

- RNA and DNA molecules can be sequenced in a high throughput manner.
- Elucidating what genes are expressed and their relative time course and/or localization allows to determine their biological role in a specific cell or pathway or disease.
- Several methods have been developed to study the transcriptome, relying on generating complementary DNA from mRNA templates using reverse transcriptase.



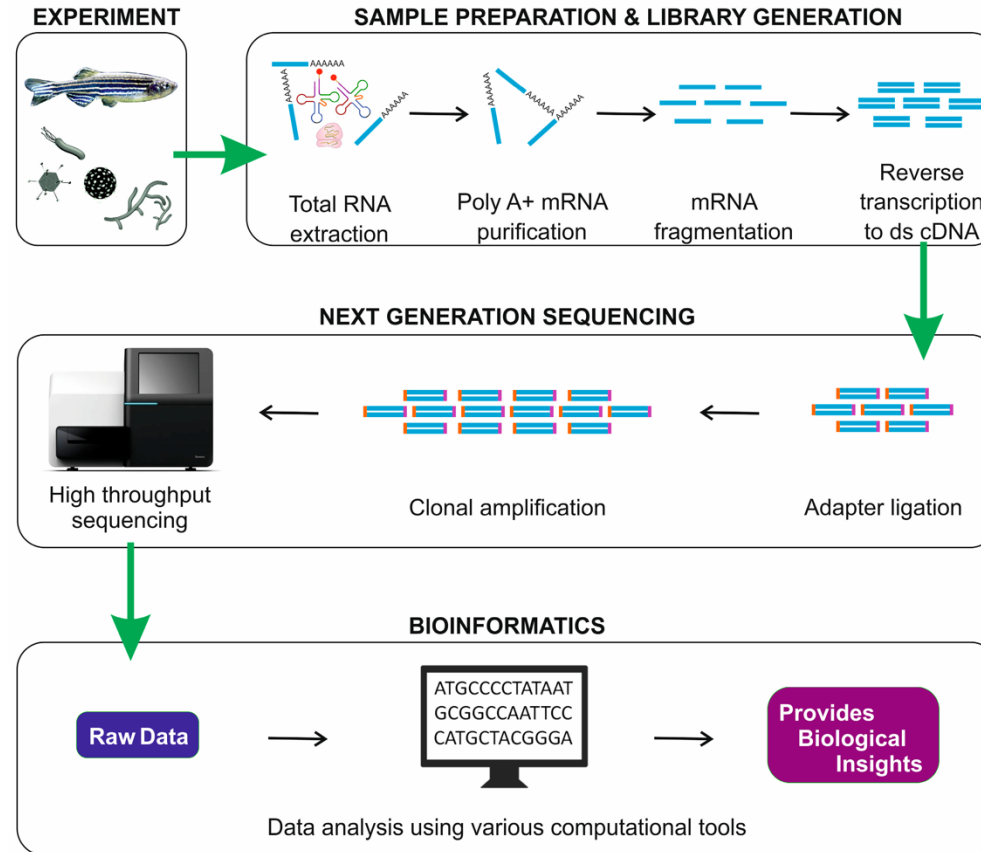
- The cDNA can be used as a probe to hybridize with a specific gene sequence (**microarray and in situ hybridization**) or amplified by PCR and sequenced (**RNA seq**).

Microarray



- A collection of microscopic DNA spots or probes attached to a solid surface (membrane or glass). The probes are designed to bind to the mRNA of interest.
- Probe-target hybridization is usually detected and quantified by detection of fluorophore- or chemiluminescence-labeled targets to determine relative abundance of target sequences.
- The microarrays are scanned by a machine that uses a laser to excite the dye and measures the emission levels with a detector. Raw data is normalized and compared between 2 samples.
- Main limitations are a reference genome and transcriptome need to be available and you're limited to preselected probes.
- Sensitivity as gene expression measurement is also limited by background at the low end and signal saturation at the high end.

Overview of Bulk RNA-seq

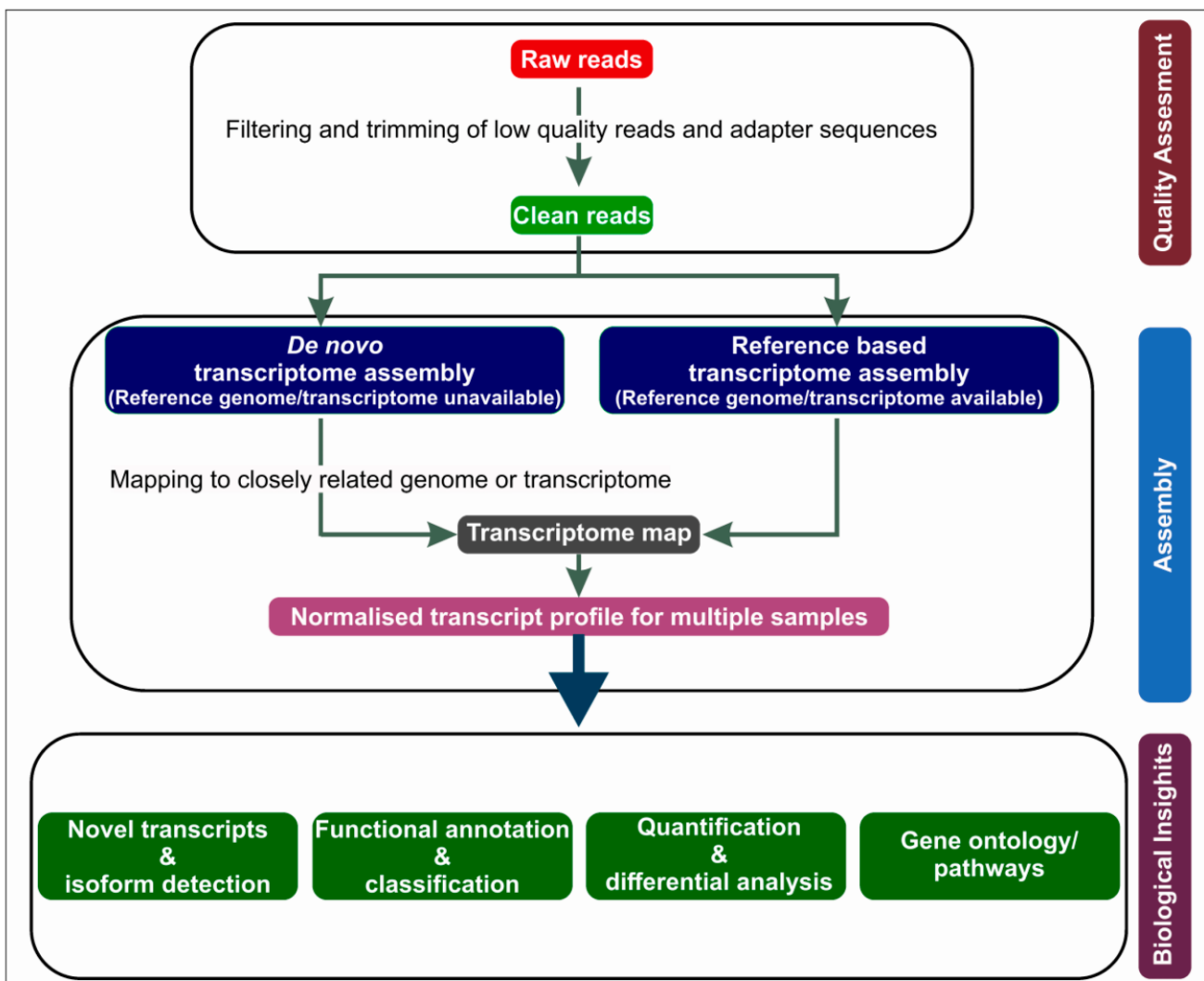


- Next generation sequencing (NGS) refers to highly parallelized sequencing of millions of cDNA fragments at the same time in a **quantitative** manner.
- Unlike hybridization techniques such as microarray, RNA-seq analysis is not limited to known genomic sequences. Can detect novel transcripts, SNPs or other alterations.

RNA-seq Experimental Design

- The experimental design and analysis is an important criteria for RNA-Seq-based transcriptome research.
- Technical and biological replicates of your samples are needed to understand the level of noise.
- After the experimental design, the next step is the RNA preparation and library construction, which starts with total RNA isolation from the tissue or cell of interest.
- It is important to check the quality of RNA before proceeding to the further downstream step, and this can be done by using the Agilent Bioanalyzer, which provides representative values as RNA integrity numbers (RIN).
- Total RNA is comprised of rRNA, pre-mRNA, mRNA, and various ncRNA. The total RNA samples could be either enriched or depleted to obtain particular species of RNA.
 - Poly-A selection and enrichment are done for sequencing mRNA
 - Ribo-depletion is performed for the sequencing of mRNA, pre-mRNA, and ncRNA.

RNA-seq Analysis Pipeline

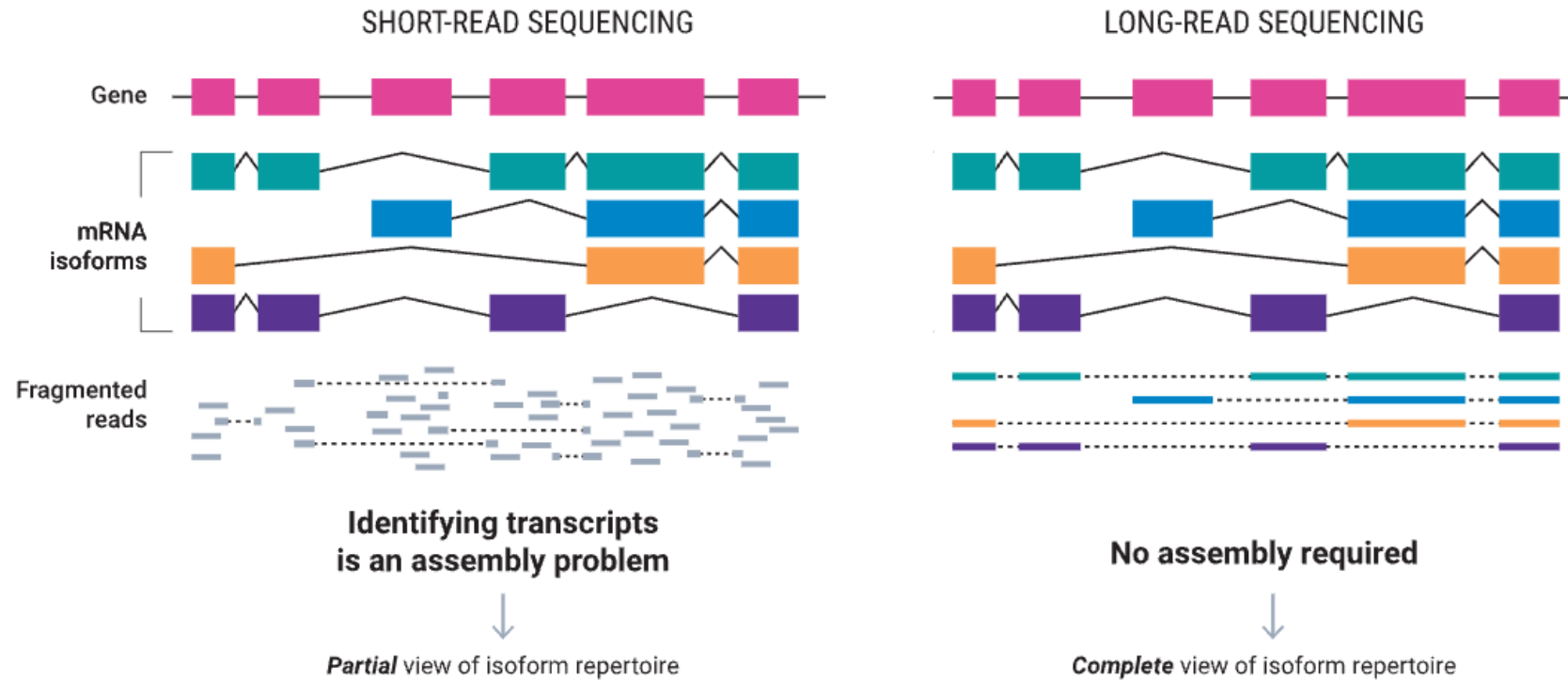


- The bioinformatics analysis starts with the refinement of the sequence reads using FastQC, which involves the removal of the adapter sequences, the trimming and discarding of reads based on quality, and the filtration of the sequence reads based on K-mer coverage.
- Filtered sequence reads are de novo assembled or aligned to a reference genome. Various choices of alignment tools and assemblies are available, including STAR, GSNAP, GEM, MAPslice etc.
- Cufflinks, featureCounts, or HTSeq are employed for transcriptome assembly and quantification.
- Differential expression analyses could be performed by employing tools such as DESeq, DEGseq, Bayseq, EdgeR etc.
- The desired choice of bioinformatics tools is always based on a researcher's preferences and objectives.

RNA-seq Impact on Field

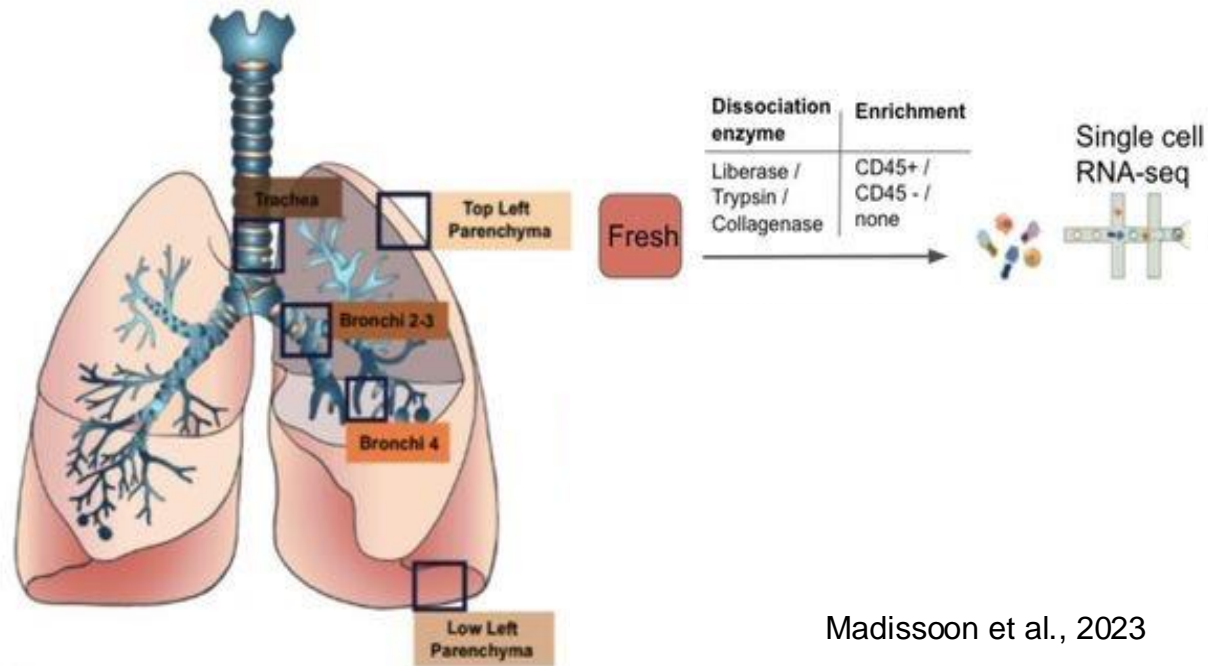
- RNA-seq methods have revolutionized modern biology and clinical applications, driven by continuous efforts of the bioinformatics community to develop accurate and scalable computational tools.
- Ability to analyze a large and wide range of biological datasets, enabling new explorations of novel and existing biological problems.
- Robust accuracy, precision, and reliability of gene expression analyses in cells and tissue, which lead to the improved downstream application in studying disease and development.
- However, transcription is, fundamentally, a stochastic process and mammalian cells are known to have non-continuous, bursting transcription, which inherently leads to variable cellular states.
- Thus, bulk RNA-seq may not give a comprehensive picture of the cellular states within a tissue or organ.

New technologies in long-read sequencing



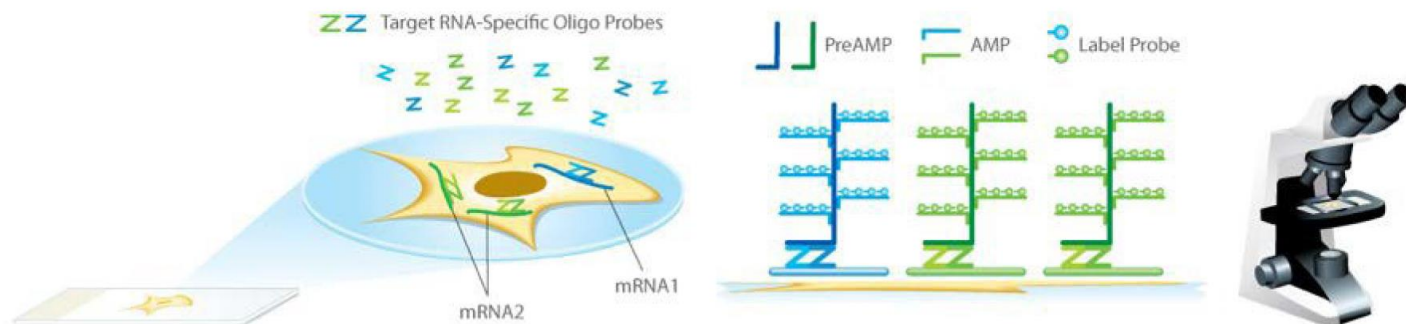
- Current limitations have biased the annotation toward multi-exonic protein coding genes and require good genome annotation.
- Pacific Biosciences and Oxford Nanopore Technologies have developed better technologies to improve transcript assembly of many organisms
 - Up to 20 kb for PacBio and up to 4 MB for Nanopore
- Long-read sequencing has better potential to capture full-length and novel transcripts, alternative splicing events, gene fusions in tumor samples, and differentially expressed or allele-specific isoforms.

Why single-cell RNAseq?



- Bulk RNA-seq returns the average expression of an entire cell population. The average behavior measured in populations of cells does not always reflect the behavior in individual cells.
- Tissues/organs are usually made up of very different types of cells that are often difficult to separate prior to the experiment.
- Can identify heterogeneity at the cellular level in disease states.
- Single-cell analyses are needed to fully understand the cellular specificity and complexity of tissue microenvironments

Traditional scRNA-seq methods



1: Tissue section

Start with properly prepared tissue sections and pretreat to allow access to target RNA.

2: Hybridize to target RNA

Hybridize multiple sets of gene-specific probe pairs to target mRNAs.

3: Amplify signal

Use up to four signal amplification systems to detect multiple target RNAs. Probes are hybridized to a cascade of signal amplification molecules, culminating in binding of dye-labeled probes visible in different fluorescent channels.

4: Image

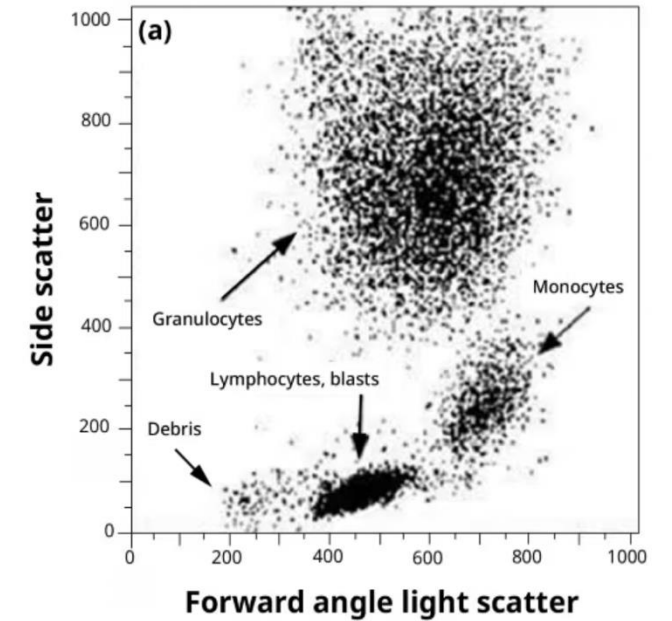
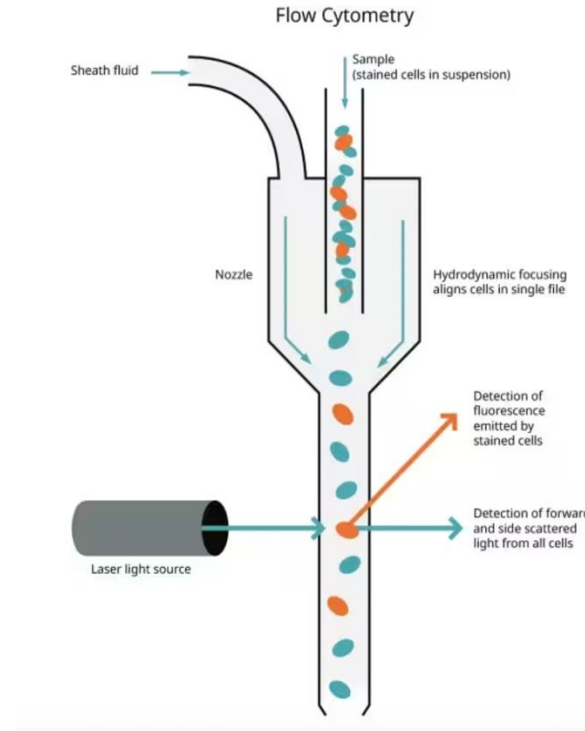
Visualize target RNA using a standard fluorescent microscope.

• In situ hybridization (ISH)

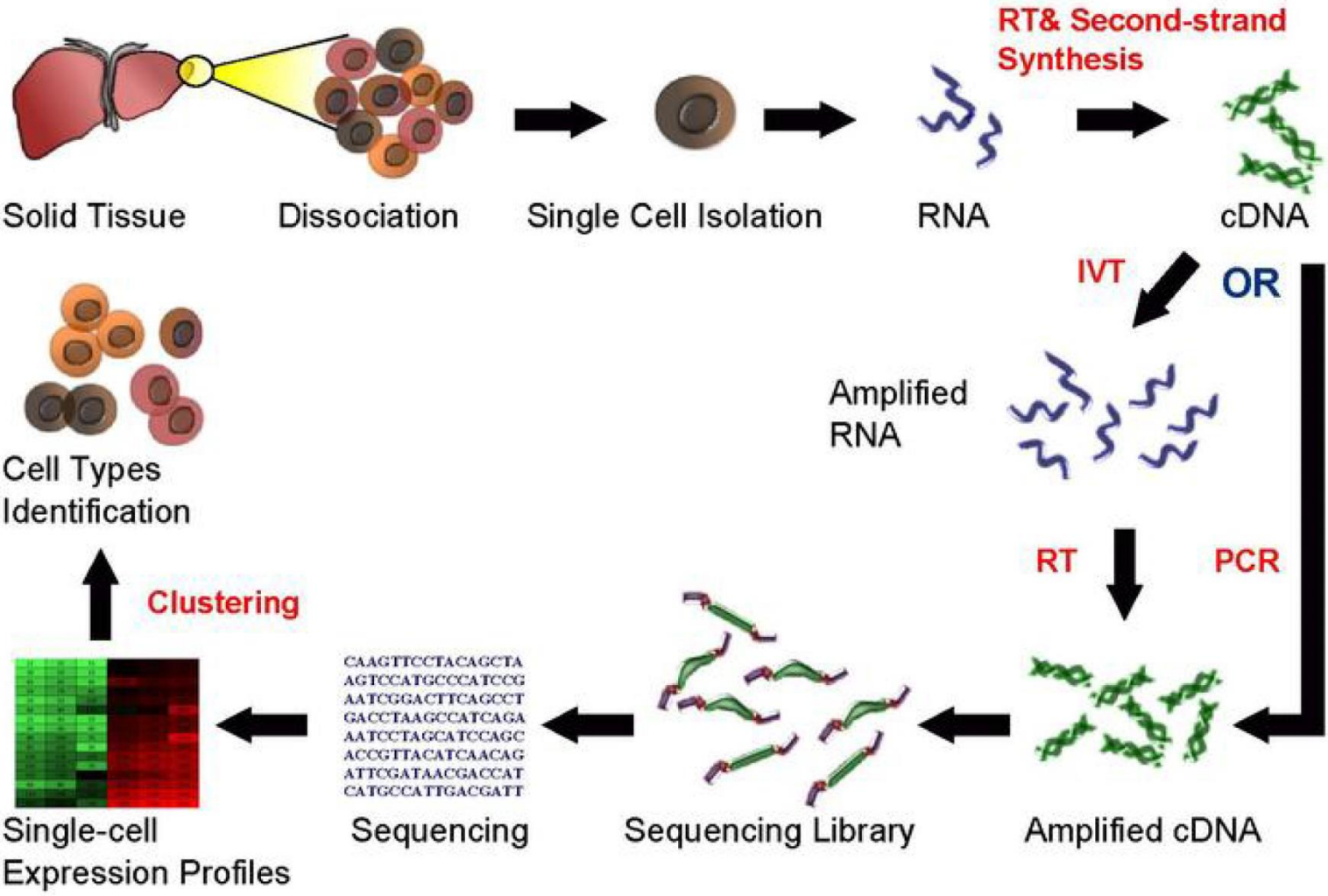
- Microscopy based assay that uses a labeled complementary DNA, RNA or modified nucleic acids probe to localize a specific DNA or RNA sequence in a section of tissue.
- It's complementary to RNA-seq as it allows to spatially detect the expression of genes.
- Allows the detection of mRNA for target with no or poor antibodies.

Traditional scRNA-seq methods

- **Flow cytometry**
 - Laser-based method for analyzing the expression of cell surface and intracellular molecules at single cell level.
 - Measure fluorescence intensity produced by fluorescently labeled antibodies specific to proteins on or in cells or ligands that bind to specific cell-associated molecules.



ScRNAseq Methods

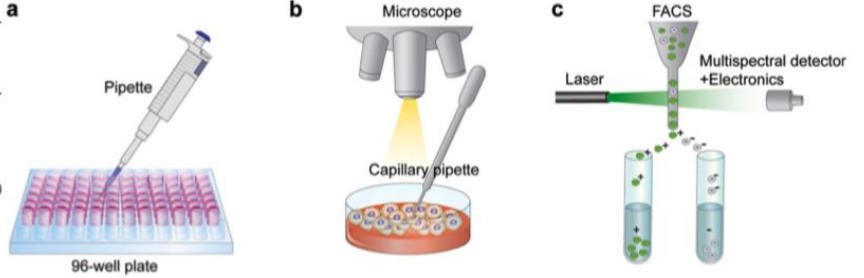


- Innovative methods to isolate single cells focus on using either flow cytometry or microfluidics.
- Choice of method depends on # of cells you want to analyze, type of tissue, cost effectiveness, and how much sequencing depth you want.
- Given a fixed population of cells and a total number of reads available → reads can either be used to sequence **fewer cells more deeply** or to **sequence more cells at a shallower depth**.

How to choose between different scRNA-seq platforms?

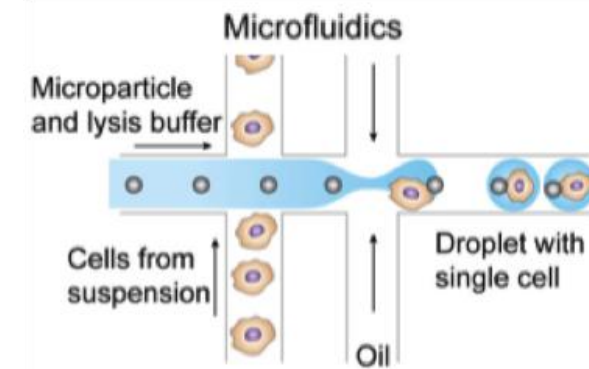
Hwang et al. (2018)

SMART-seq



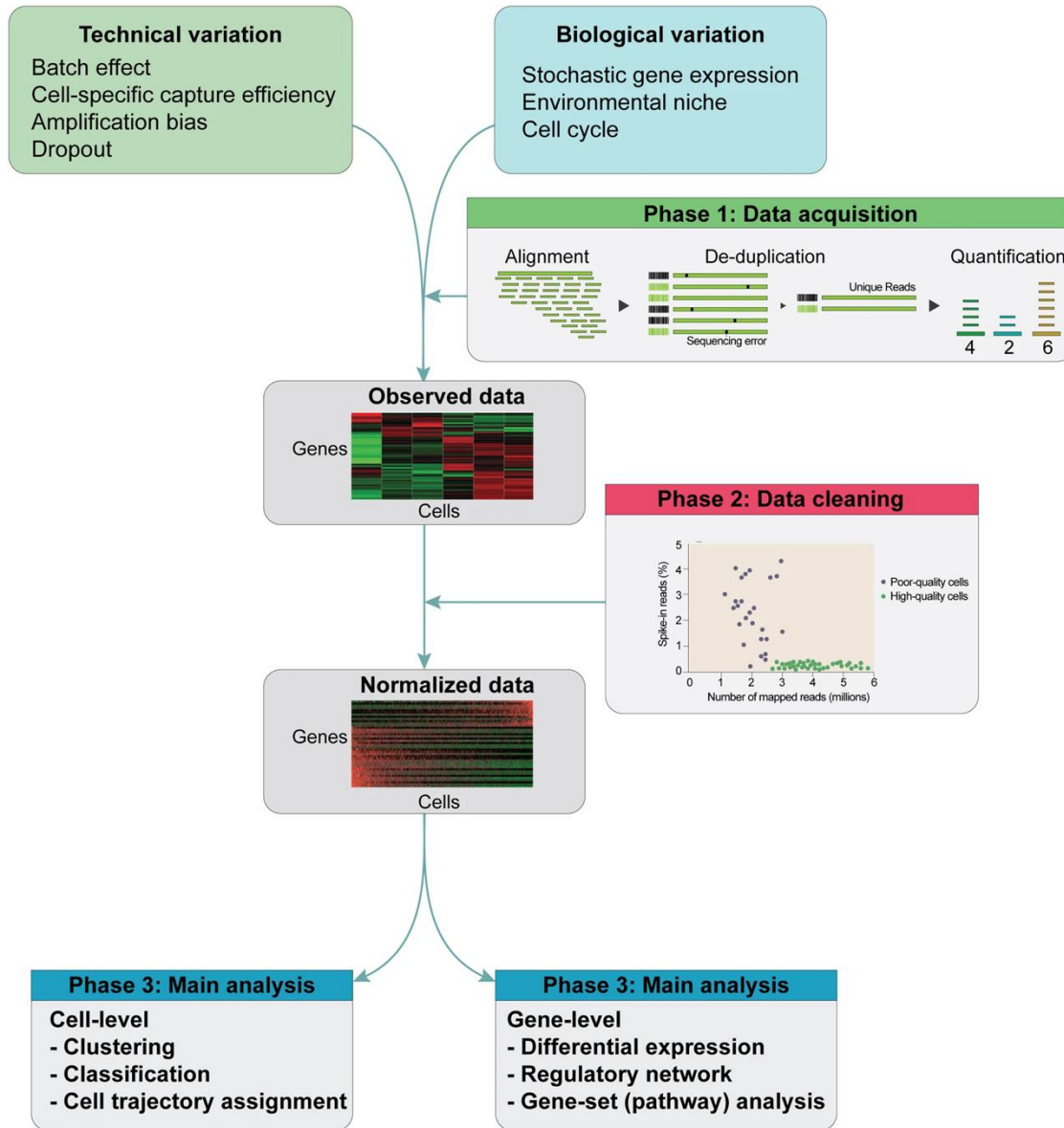
- Manual cell isolation and sorting into 96 well plates.
- Labor intensive and more costly.
- Low throughput with less cells.
- Allows for full-length transcripts and can investigate greater gene diversity.

10X Chromium



- Automated and kit based.
- Using microfluidics, individual cells are captured together with a large set of barcoded poly (dT) primers.
- High throughput and can do 10,000s of cells.
- Less reads per cell, limiting types of transcripts analyzed.

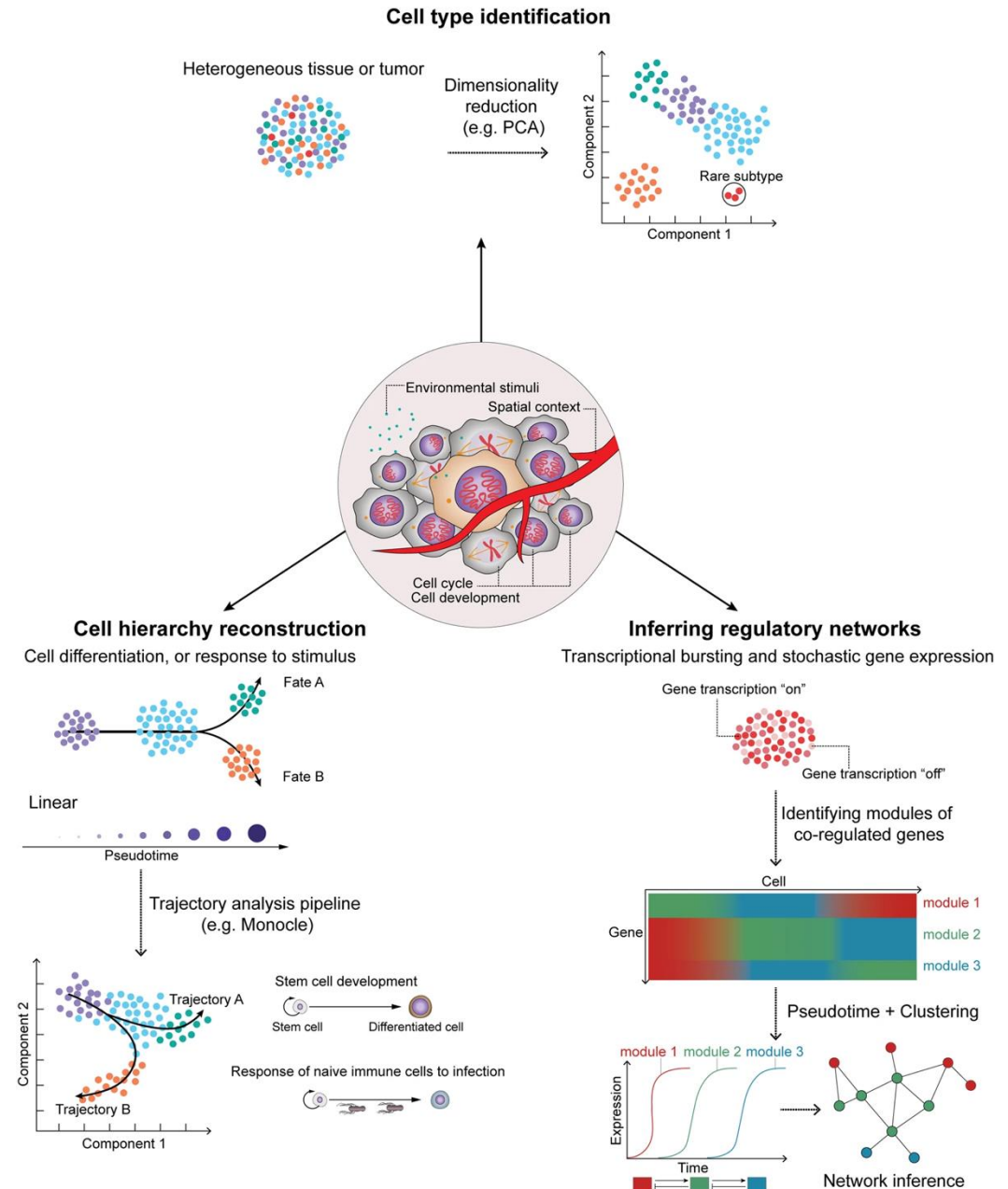
Analysis of scRNA-seq



- The bioinformatics analysis is similar to bulk RNA-seq analysis.
- A distinctive feature of scRNA-seq data is the presence of zero-inflated counts due to dropout or transient gene expression. Normalization of data is necessary to remove cell-specific bias, which can affect downstream applications (e.g., determination of differential gene expression).
- The read count for a gene in each cell is expected to be proportional to the gene-specific expression level and cell-specific scaling factors.

Processing and Analysis of scRNA-seq

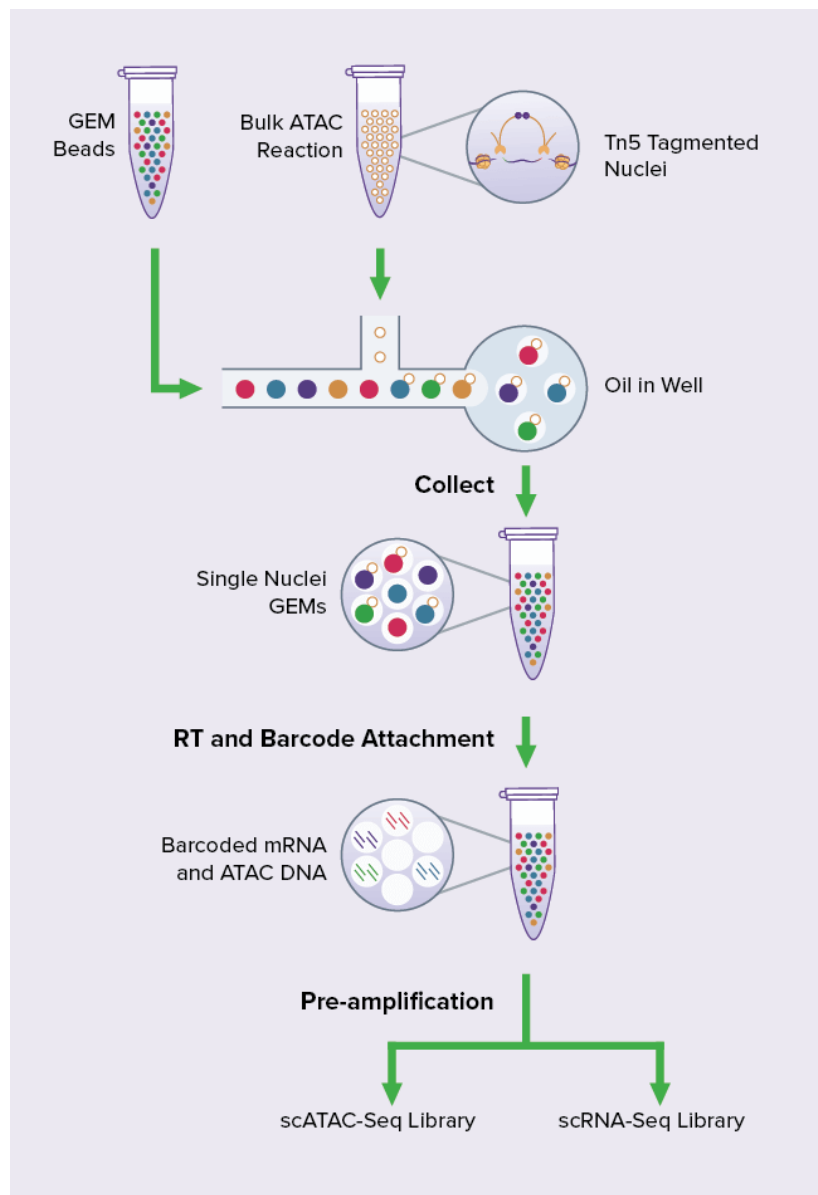
- Principal component analysis (PCA) is a widely used unsupervised linear dimensionality reduction method to cluster individual cells to common "states".
- Cell Ranger (10X Genomics) or Seurat in the R package are softwares that perform cell clustering of your dataset and identifying gene regulatory networks.
- Software like Monocle can reconstitute cell hierarchy and help better understand cellular dynamics in tissue.



scRNA-seq Impact on Field

- scRNA-seq is revolutionizing our fundamental understanding of biology, going beyond descriptive studies of cell states.
- Tumor heterogeneity is a common phenomenon that can occur both within and between tumors, and scRNA-seq can be applied to better understand drug resistance, metastasis, and overall cancer evolution.
- scRNA-seq helps generate models of developing organs, such as the brain and lung. This technique could also be applied to reconstruct clonal and phylogenetic relationships between cells by modeling transcriptional kinetics.
- Lineage tracing is a long-standing fundamental question in biology aimed at understanding how a single-celled embryo gives rise to various cells types that are organized into complex tissue and organs. scRNA-seq identifies new markers that can better identify novel subpopulation of cells.
- Future applications of scRNA-seq in biology and biomedical research will also provide novel insights into physiological structure–function relationships in various tissue and organs.

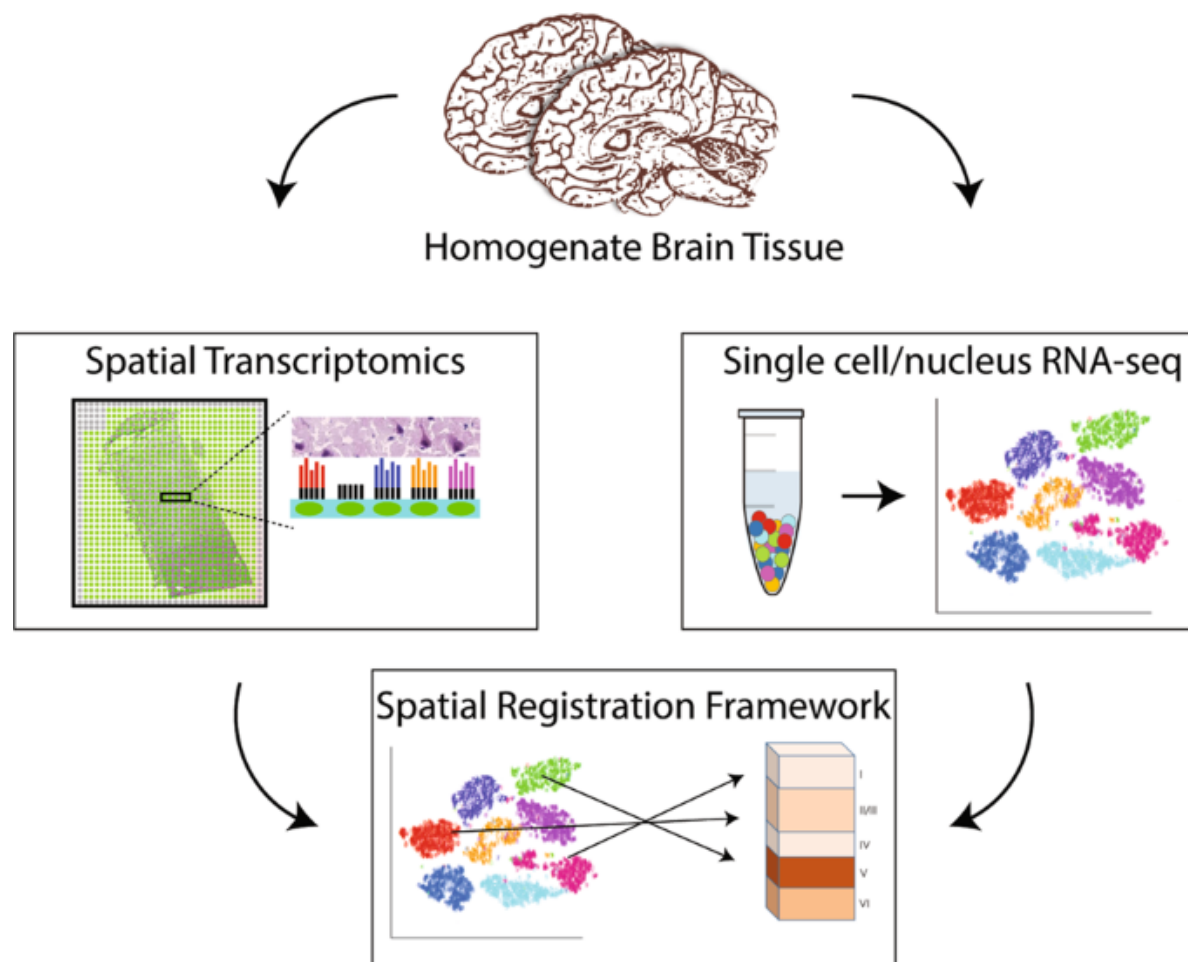
Future of Transcriptomics : Multi-omics



- Multi-omics provides an integrated approach to power discovery across multiple levels of biology by combining data from genomics, transcriptomics, epigenetics, and proteomics.
- Many multi-omics methods function with transcriptome profiling as their 'anchor' to facilitate single-cell multi-omics interrogation.
- Single-cell assay for open chromatin with sequencing (scATAC-seq) profiles the transcriptome together with the epigenome. These are known as SNARE-seq and 10X Multiome.
- Limited coverage of single cells, for which each omics layer can only be partially profiled, thus resulting in loss of important data.

Future of Transcriptomics: Spatial Transcriptomics

- Combines different single-cell techniques to generate transcriptome-wide profiling in a defined spatial area.
- The position of any given cell, relative to the structures surrounding it can provide helpful information for defining cellular phenotypes and cell and tissue function.
- Spatial resolution and mRNA recovery rates are tissue dependent and can be lower than ideal.
- Dependent on the quality of tissue as RNA degrades much faster than DNA over time.



Maynard and Martinowich et al., 2019


nature communications

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature communications](#) > [articles](#) > article

Article | [Open access](#) | [Published: 21 April 2020](#)

Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis

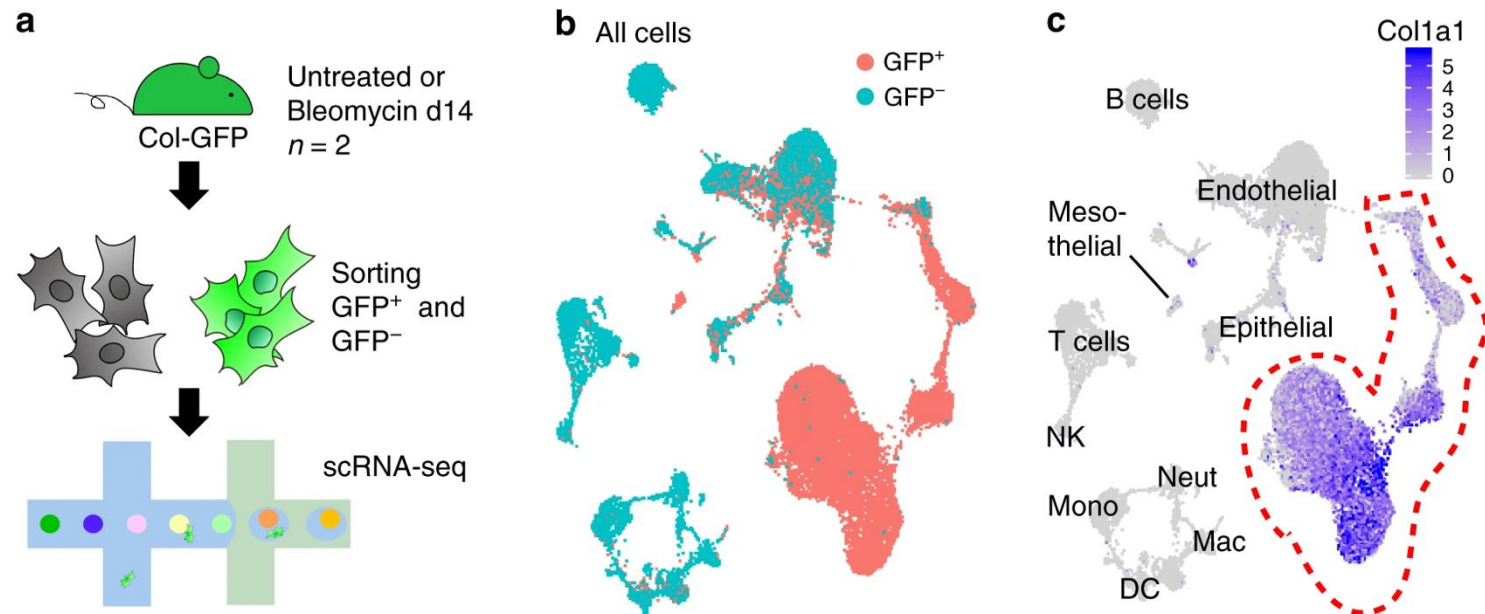
[Tatsuya Tsukui](#), [Kai-Hui Sun](#), [Joseph B. Wetter](#), [John R. Wilson-Kanamori](#), [Lisa A. Hazelwood](#), [Neil C. Henderson](#), [Taylor S. Adams](#), [Jonas C. Schupp](#), [Sergio D. Poli](#), [Ivan O. Rosas](#), [Naftali Kaminski](#), [Michael A. Matthay](#), [Paul J. Wolters](#) & [Dean Sheppard](#) 

[Nature Communications](#) **11**, Article number: 1920 (2020) | [Cite this article](#)

45k Accesses | **249** Citations | **46** Altmetric | [Metrics](#)

Introduction the paper

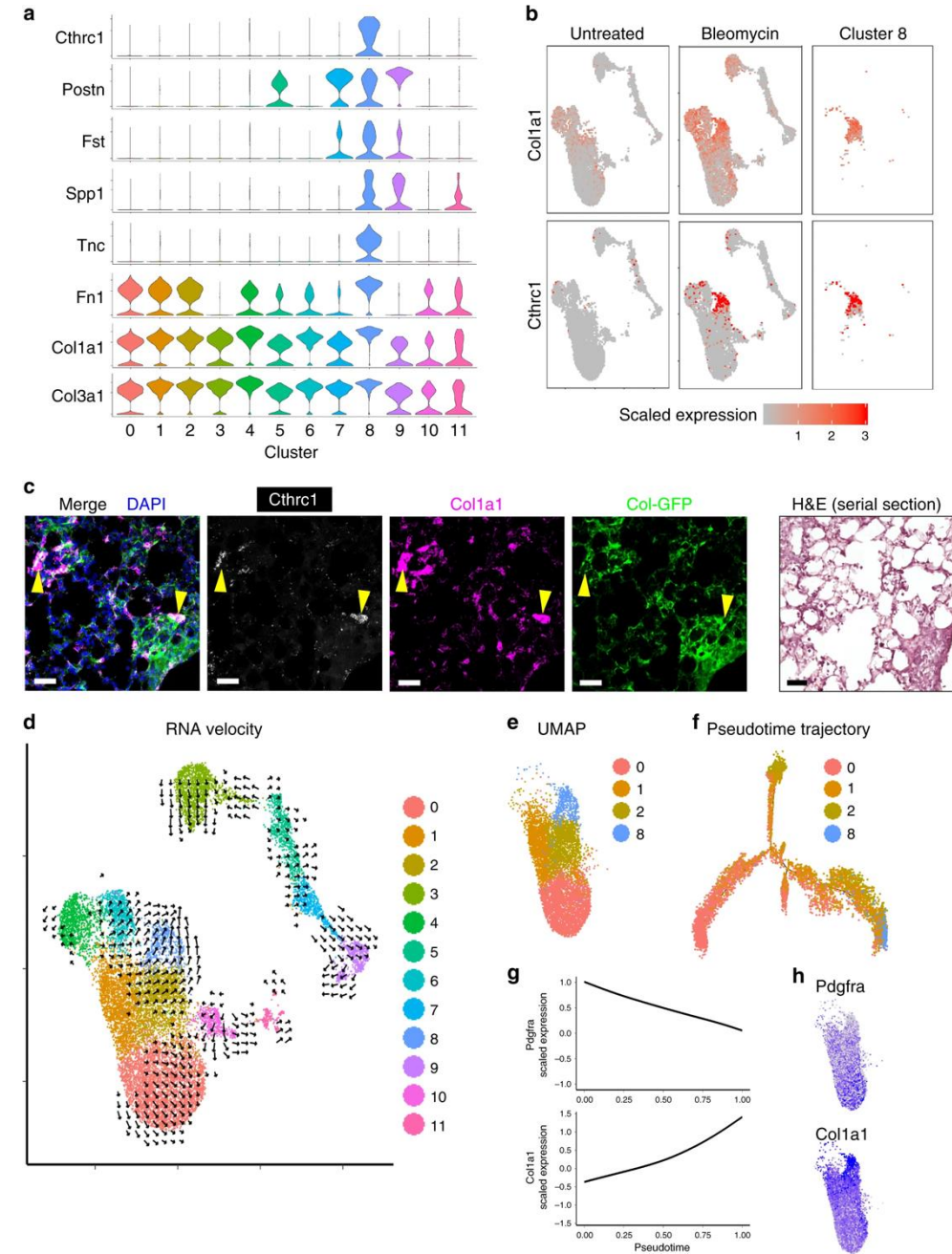
- The authors performed scRNA-seq analysis using FACS sorting and the 10X Chromium platform.



- They identified all collagen-producing cells in normal and fibrotic lungs and characterized multiple collagen-producing subpopulations with distinct anatomical localizations in different compartments of murine lungs.
- They also identified a unique population of **Cthrc1+** (collagen triple helix repeat containing 1) fibroblasts, which are mostly found in fibrotic lungs in both mice and humans and express the highest levels of type 1 collagen and other ECM genes.

Introduction the paper

- Their scRNA-seq results provide the first systematic atlas of the molecular characteristics and anatomic locations of collagen-producing cells in the adult human lung.
- The authors used two different computational tools—RNA velocity and pseudotime trajectory analysis via R package software. Both identified unique cell of interest that were most likely to have differentiated from a population of alveolar fibroblasts.
- They found that *Cthrc1* is a marker for pathologic fibroblasts in human pulmonary fibrosis. *Cthrc1* is expressed in injured tissue and promotes cell migration.
- Their conclusions have important implications for lung fibrosis and cancer.



Thank you!