Single cell transcriptomics (2):

Complex design, trajectory inference, somatic mutations

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang / 黃淵華

School of Biomedical Sciences &

Department of Statistics and Actuarial Science

香 港 大 學

THE UNIVERSITY OF HONG KONG

Today’s learning objectives

1. Complex experiment designs and batch effects
2. Differentiation trajectory inference
3. Splicing and RNA velocity
4. Somatic mutations and its impact on gene expression

Resources:

- Analysis of single cell RNA-seq data (Sanger course) https://www.singlecellcourse.org/
- HKUMed Single-cell analysis tutorial workshop (HKU Med) https://statbiomed.github.io/HKU-single-cell-workshop/

Complex experiment designs

- Identify key abnormal cell types in Multiple sclerosis vs healthy donors
  - Design: 30 MS patient + 30 healthy donors, each with 2K cells

- Understand the cell type specific impact of a certain treatment
  - Design: 5 treated + 5 control, each with 5K cells

- Understand the genetic effects on cell differentiations
  - iPS cells differentiating to neurons, with 5 time points, 3K cells each time
  - For genetic, repeating for 100 cell lines

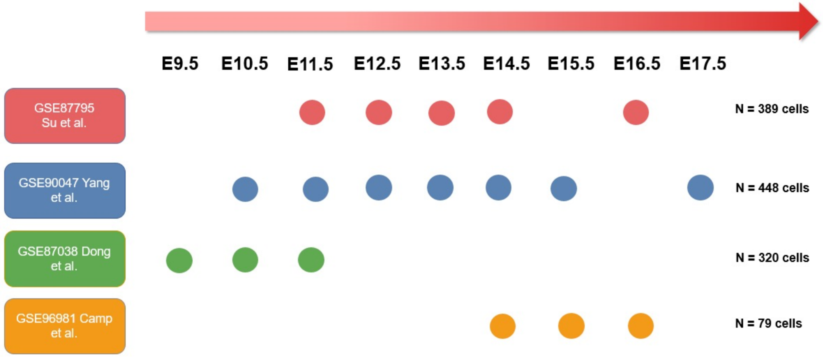
Multiple samples are needed: multiple batches are often needed;
Be careful what “sample” means: individuals or cells

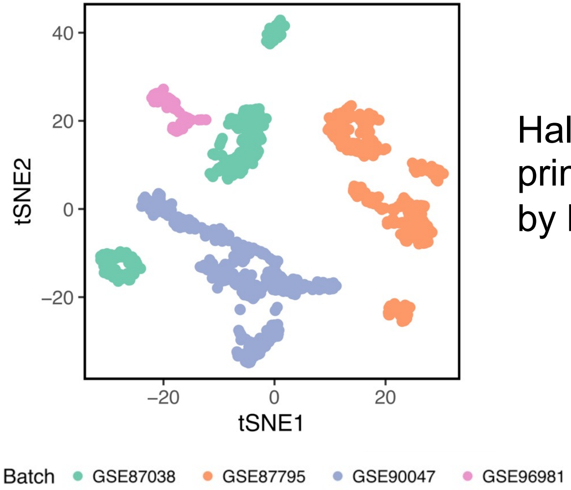
Batch effects

Batch 1

Batch 2

Batch effect: differences in gene expression caused by non-biological factors that may cause systematic difference in data generated in different batches of experiments

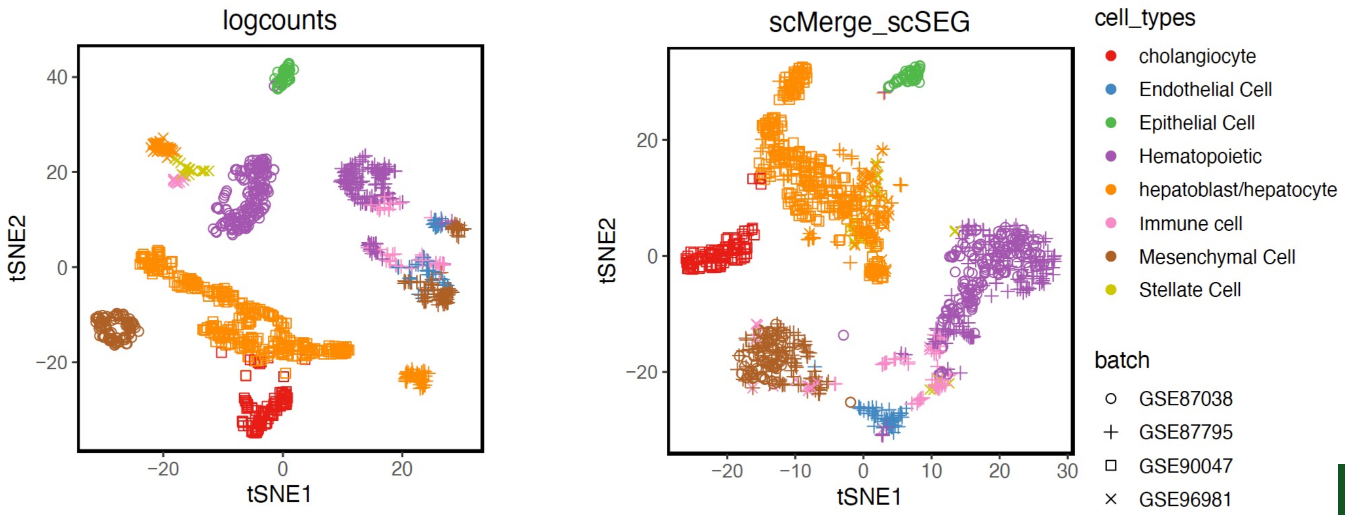




Hallmark: Cells are primarily grouped by batch

Batch effect correction in a post-step

- Preprocess each dataset separately (e.g., PCA), then align multiple datasets with statistical methods
- Many methods have been proposed to correct batch effects, e.g., scMerge
  - See a benchmark study: Tran et al, Gen Biol, 2020.
- Trade-off: under correction vs over correction.
  - Aim to regress out technical batch effects but retain biological differences



Lin et al (2019) PNAS

Less batch effect with join analysis

- Another approach is to join analyzing all multiple datasets, e.g., align them to a common (reference, often large) dataset
  - Conventional statistical methods: project all datasets into a common PCA space; Liu et al, Nat Biotech, 2021
  - Machine learning methods: transfer learning, e.g., Lotfollahi et al, nat biotech, 2021; methods not so mature, but shows good potential

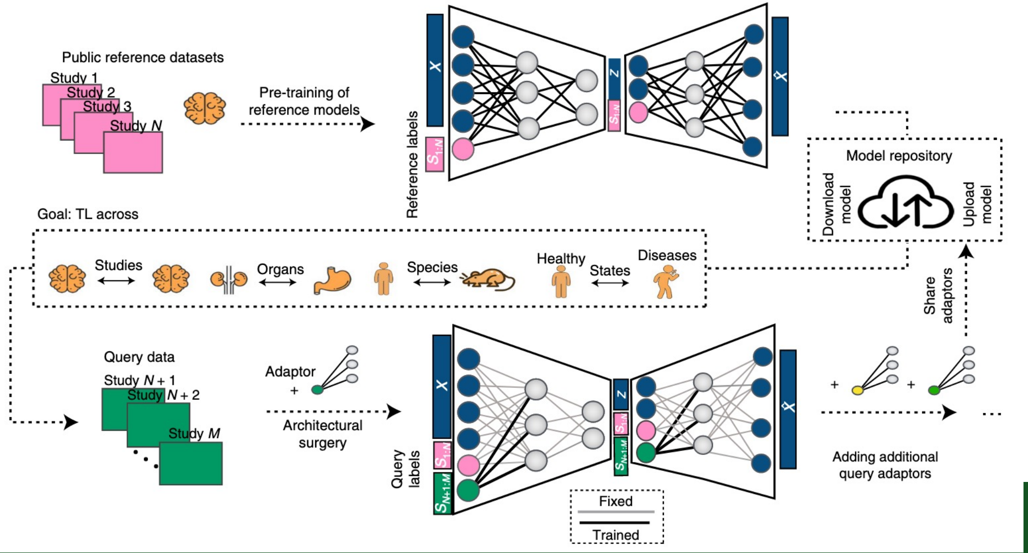
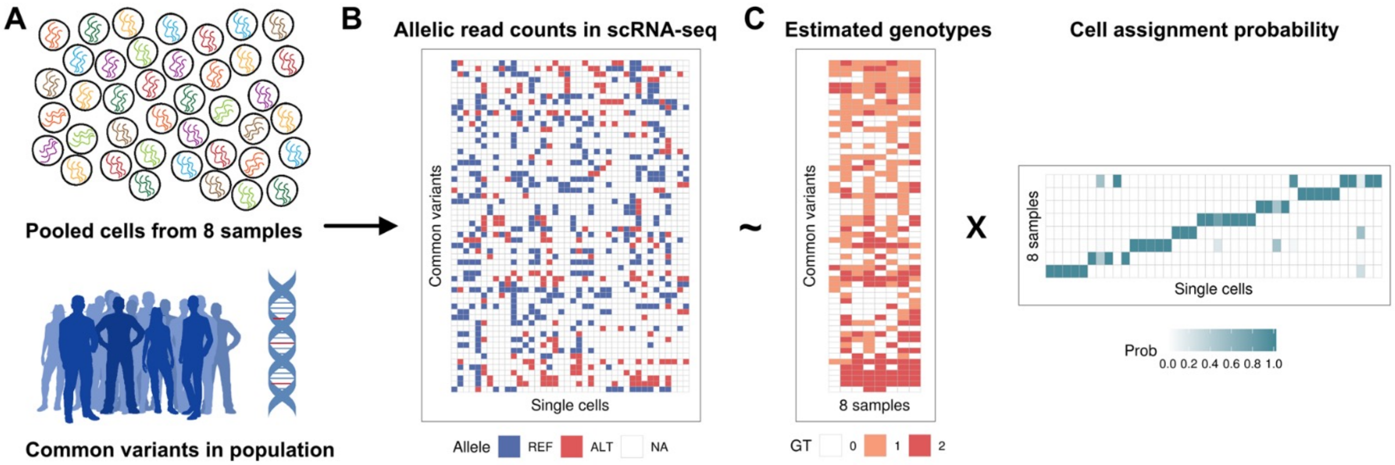


Illustration of a transfer learning method scArches

Multiplexed settings

- Given the cells are barcoded, multiple samples can be further multiplexed by using natural genetic makeups, or external molecular barcodes
- Batch effect in the sequencing step can be eliminated



Huang et al, 2019. Vireo: genetic makeups without reference;
Kang et al. 2018. Demuxlet: genetic makeup;
Shin et al, 2019. BSO: molecular barcoding.

Today's learning objectives

1. Complex experiment designs and batch effects

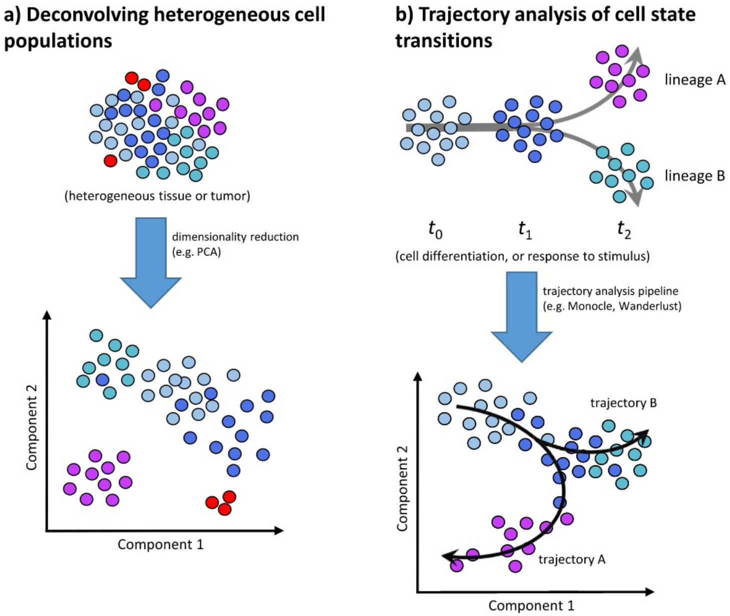
2. Differentiation trajectory inference

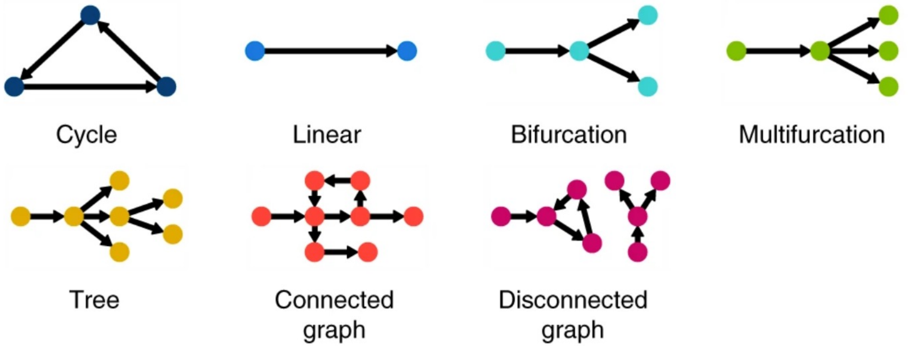
3. Splicing and RNA velocity

4. Somatic mutations and its impact on gene expression

Trajectory inference

- The pseudo-temporal ordering of the cells in development or differentiation


a) Deconvolving heterogeneous cell populations
(heterogeneous tissue or tumor)
dimensionality reduction (e.g. PCA)
Component 2
Component 1
b) Trajectory analysis of cell state transitions
lineage A
lineage B
t_0
t_1
t_2
(cell differentiation, or response to stimulus)
trajectory analysis pipeline (e.g. Monocle, Wanderlust)
trajectory B
trajectory A
Component 2
Component 1

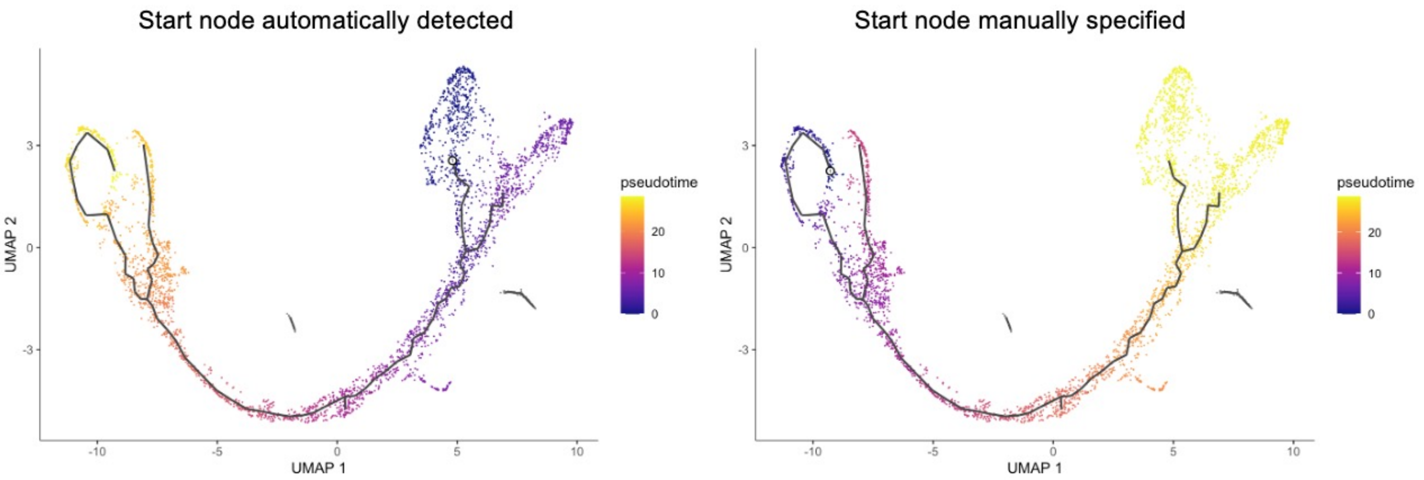


An inferred trajectory may or may not represent real developmental lineage (which required genetic linage tracing experiment)

Liu and Trapnell, F1000, 2019; Saelens et al. Nat biotech, 2019

Trajectory inference: directionality

- Commonly based on the Euclidean distance of transcriptome between two cells in raw genes or PCA space.
- Intrinsic challenge: very weak information on directionality along an inferred trajectory. No info for past and future states.



Start node automatically detected

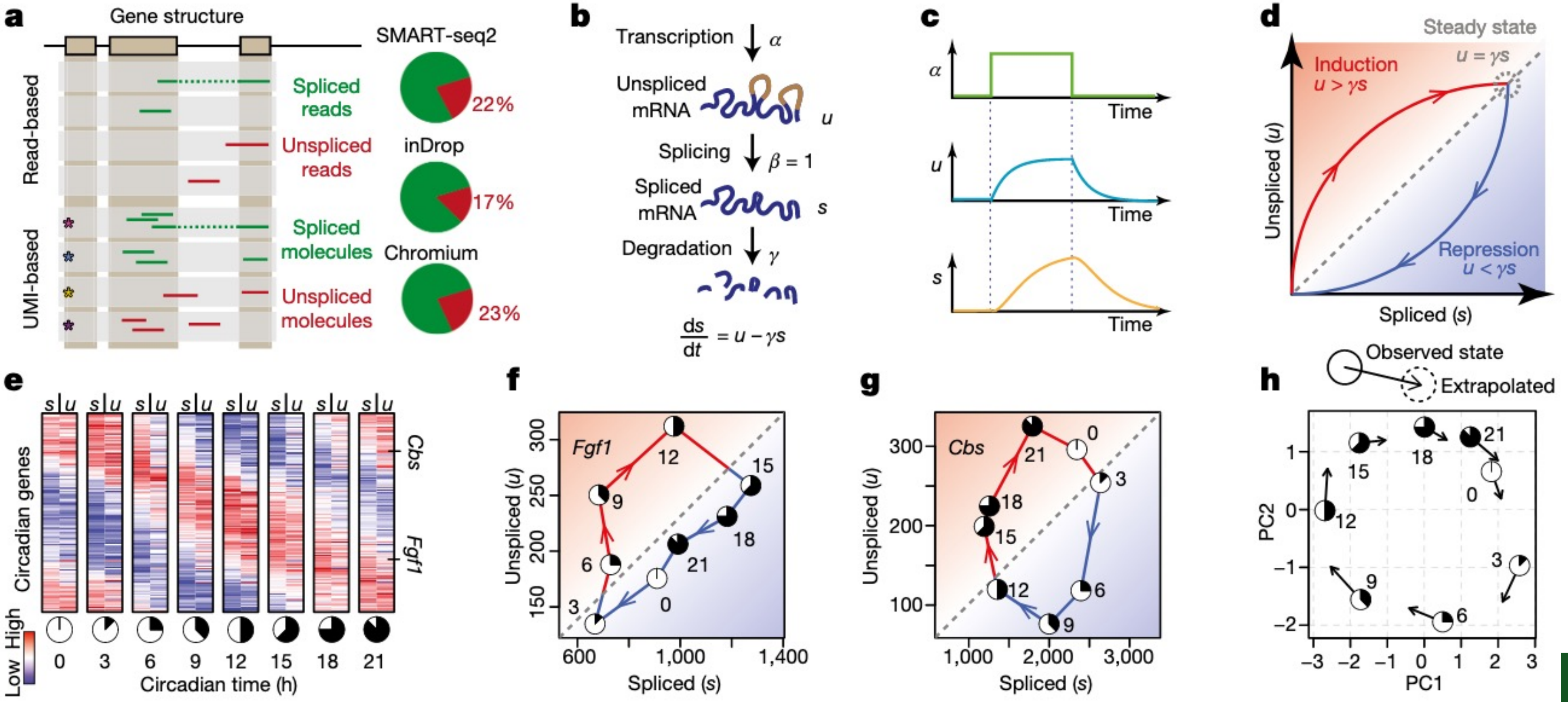
Start node manually specified

Monocl3: Pseudo-time requires specifying starting node

La Manno et al, Nature, 2018

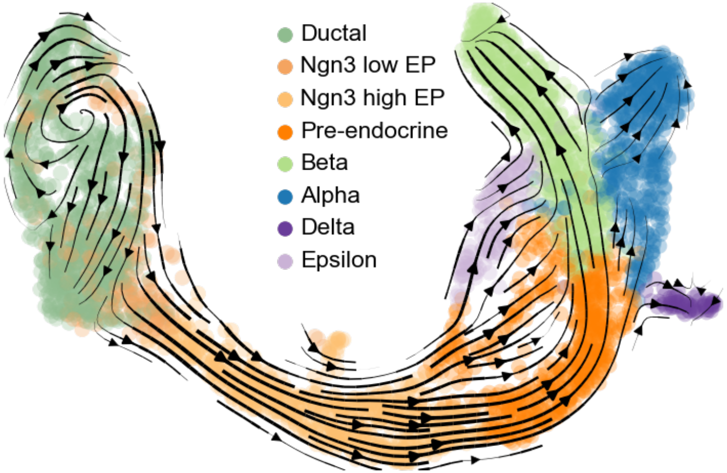
RNA velocity offers directionality

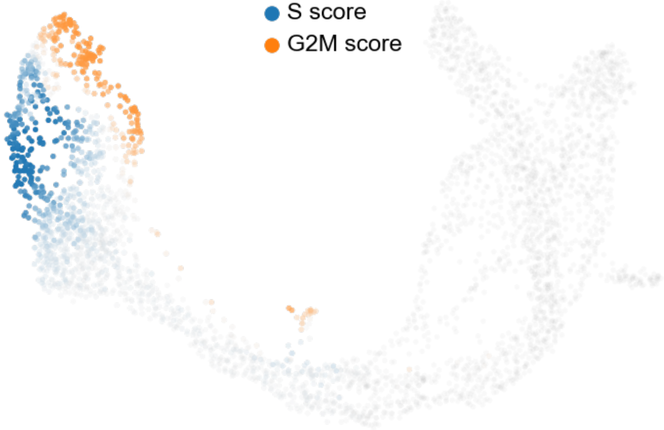
- Intrinsic dynamics of RNA processing (single direction)
- The level of unspliced RNAs indicate the future level of spliced RNAs



RNA velocity: projection of transitions

- Using unspliced RNAs to predict the near future of each gene
- Find the most similar transcriptome to the predicted one from unspliced RNAs, as the predicted cell state in near future



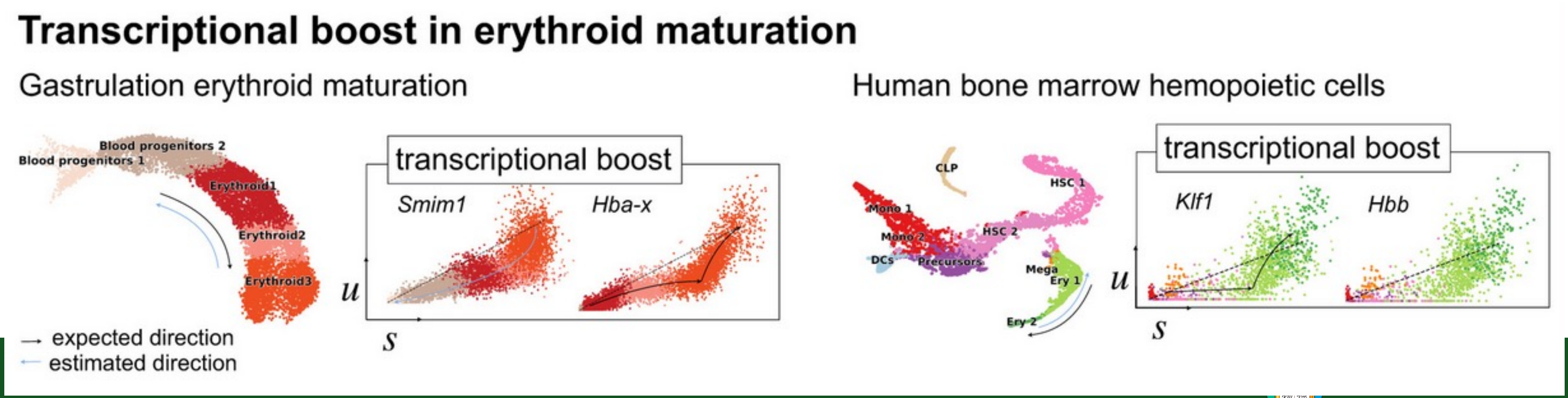


Bergen et al, Nat Biotech, 2020; scvelo, a Python package for RNA velocity analysis

Bergen, et al. MSB, 2021

RNA velocity: potential limitations

- Major limitations to be addressed, despite the appealing concept
  - Multiple kinetic rates (splicing rates, degradation rates) across populations
  - Transcriptional boost (far from steady state)
  - Complex kinetics (multiple branching)
- Potential ideas to improve the robustness and accuracy of RNA velocity:
  - Select informative genes: BRIE2 (Huang & Sanguinetti, 2021)
  - Denoise by projection on lower dimensions: veloAE (Qiao & Huang, 2021)



Today’s learning objectives

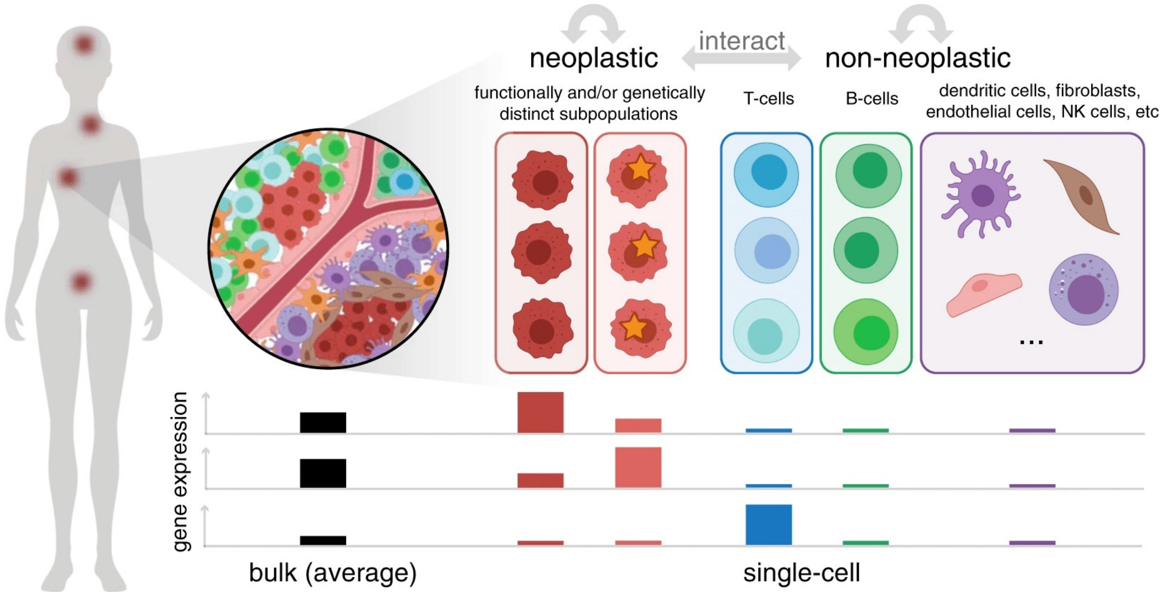
1. Complex experiment designs and batch effects

2. Differentiation trajectory inference

3. Splicing and RNA velocity

4. Somatic mutations and its impact on gene expression

Cellular composition of a tumour

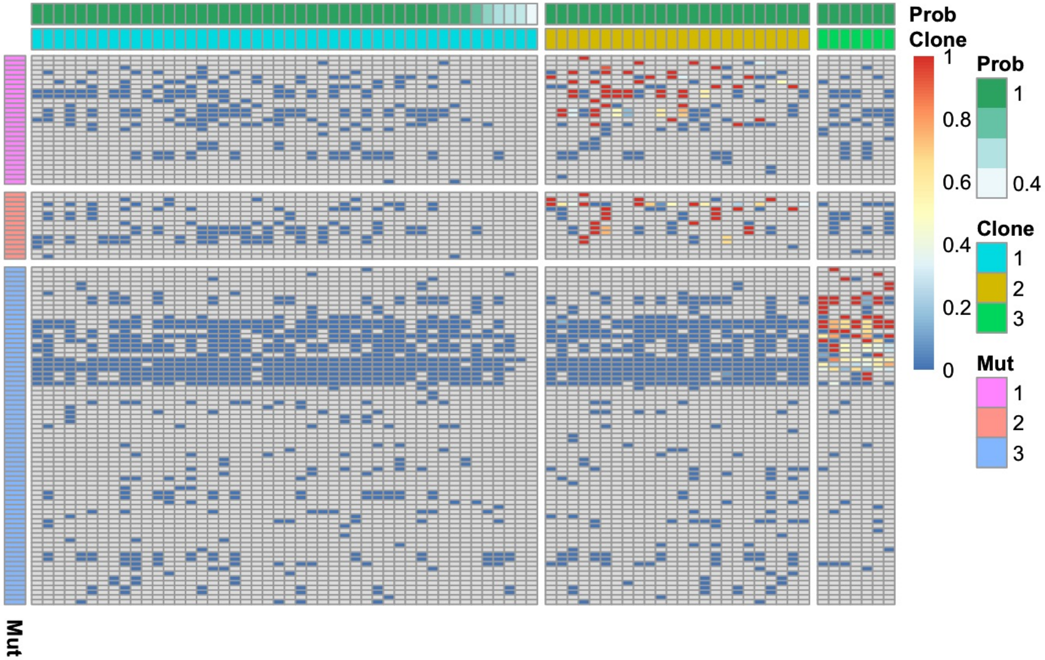

neoplastic
interact
non-neoplastic
functionally and/or genetically distinct subpopulations
T-cells
B-cells
dendritic cells, fibroblasts, endothelial cells, NK cells, etc
...
gene expression
bulk (average)
single-cell

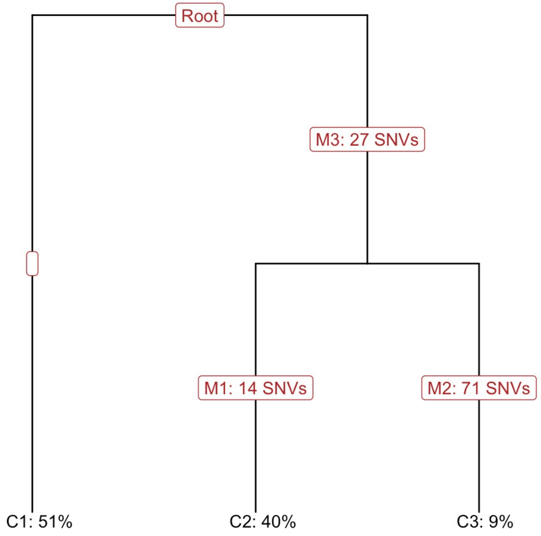
Neoplastic cells (cells with abnormal growth), usually affected by genetic alteration

Non-neoplastic cells (tumour microenvironment):
- Tumour infiltrating lymphocytes (TILs), including T cells, B cells, macrophages, etc
- Stromal cells, including cancer-associated fibroblast (CAFs)

Single-nucleotide variant (SNV)

- SNVs can be observed in the expressed RNAs
- Grouping cells by their carrying SNVs → clonal SNVs

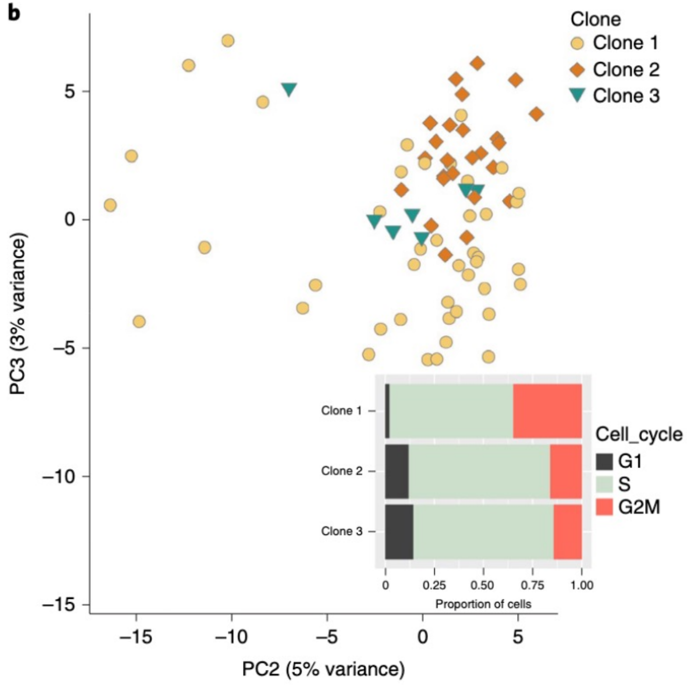


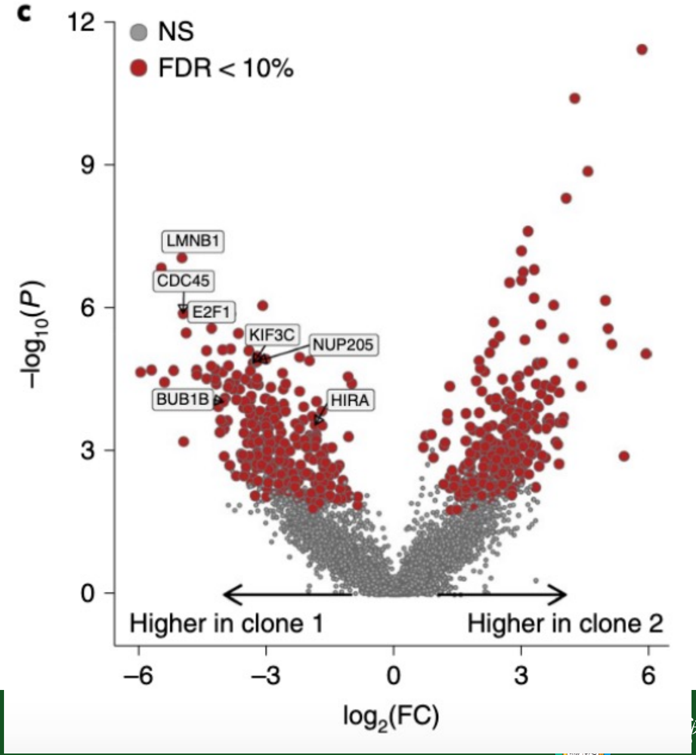


McCarthy, Rostom, Huang, et al, Nat Meth, 2020

Clonal SNVs and impact on gene expression

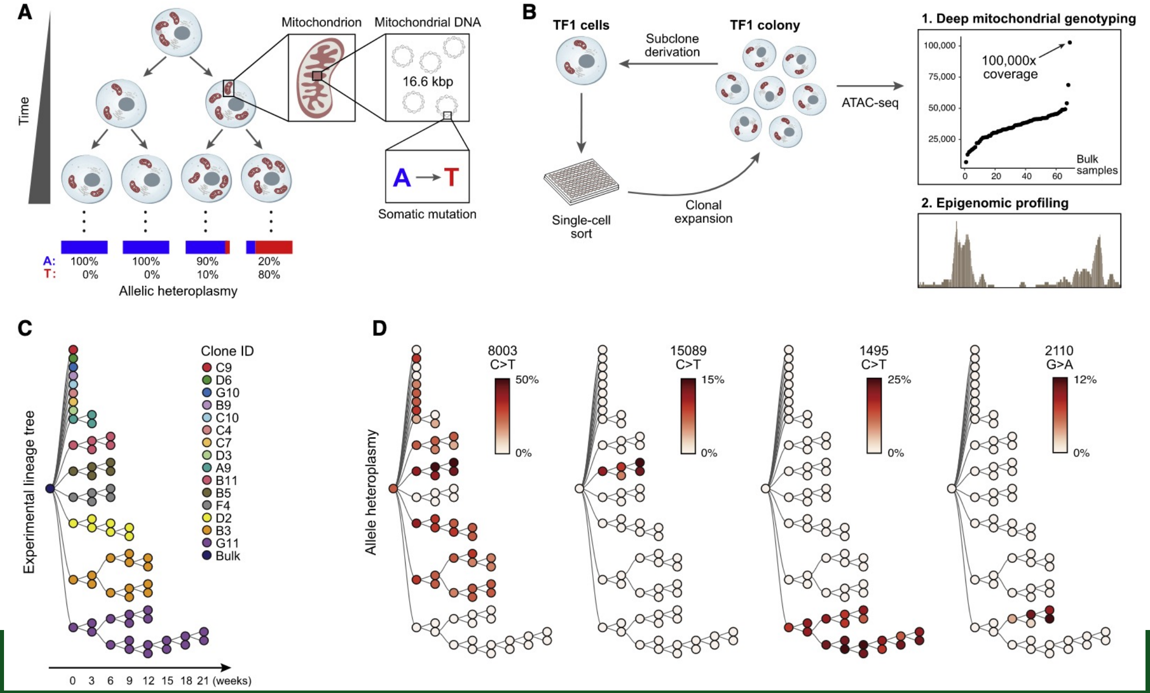
- Global impact on transcriptome; separation on PCA space
- Detecting differentially expressed genes between clones (recall how?)





Mitochondrial mutations

- Challenges in SNV: coverage is low (SMART-seq) or very low (droplet)
- mtDNA variants: much higher coverage (many copies)
- High mutation rate: potentials for lineage tracing

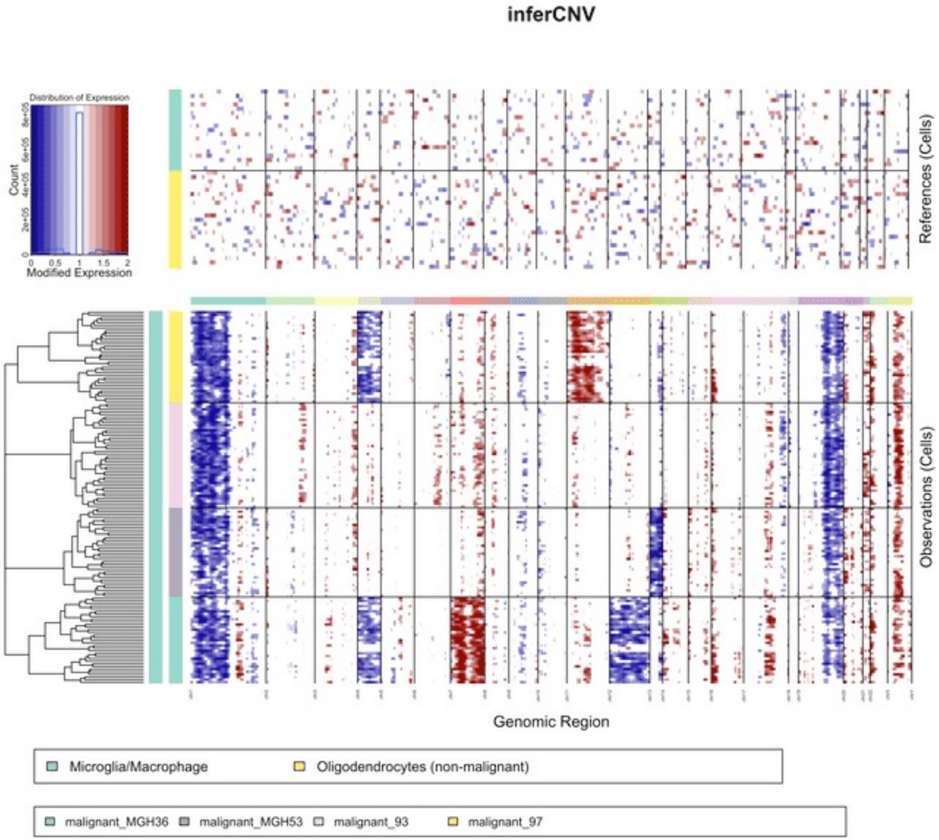


Accurate detection clonal mtDNA variants:

- mgatk: Lareau et al, 2020;
- MQuad: Kwok et al, 2021

Ludwig et al, Cell, 2019

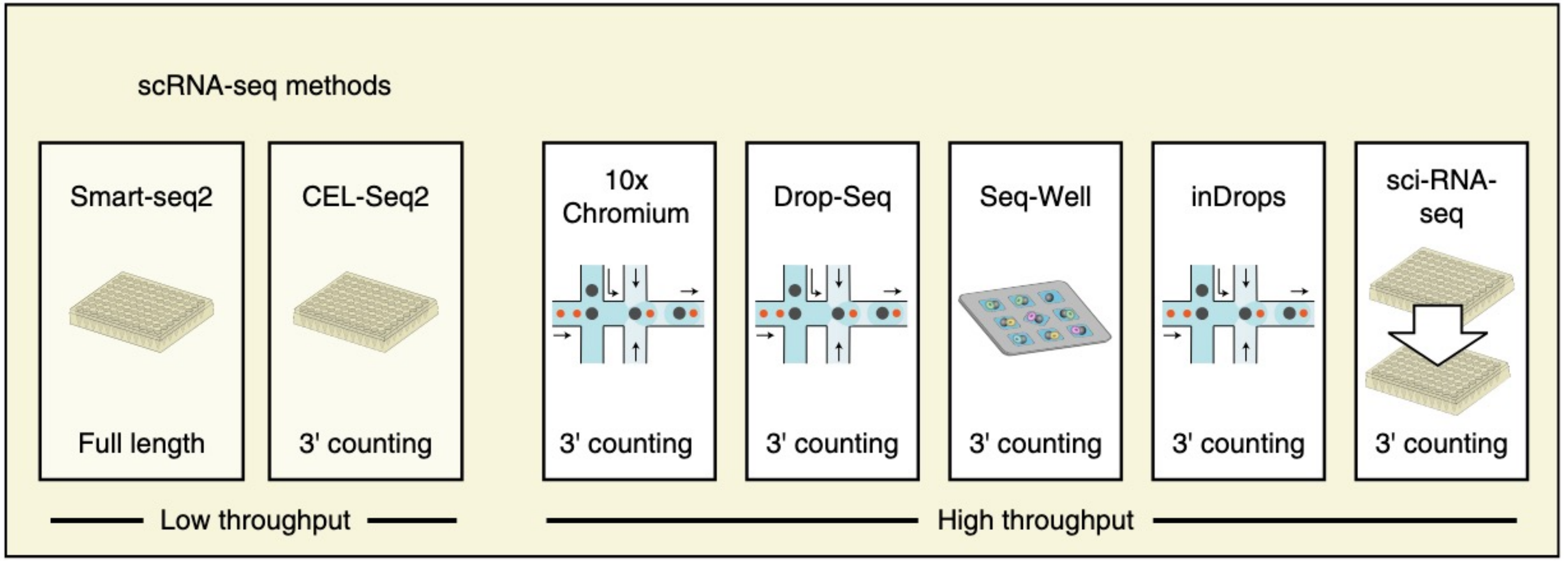
Copy number variation



- CNVs commonly exist in tumour cells
- CNV clonal structure
- Large range of CNV, whole chromosome or an arm
- Commonly used methods: inferCNV, CopyKat

- **Open challenges**:
  - Detecting loss of heterozygosity (allelic info)
  - Integrating with other assays, bulk WGS

Choice of protocols for somatic mutations



	Example protocol	coverage	cell numbers	missing regions	cost	CNV	SNV	mtSNV
scDNA-seq (well-based)		moderate	low	moderate	high	Very good	Good	Good
scDNA-seq (droplet)	10x CNV	low	moderate	high	high	Very good	Poor	Moderate
scDNA-seq (targeted)	Mission Bio	high (target)	high	non-targeted	moderate	Maybe	Good	Maybe
scRNA-seq (well-based)	SMART-seq2	moderate	low	non-expressed	moderate	Good	Moderate	Good
scRNA-seq (droplet)	10x Genomics	low	high	non 3' or 5'	moderate	Good	Very poor	Poor
scRNA-seq (targeted)	GoT	high (target)	high	non-targeted	moderate	Maybe	Good	Poor

Public resources for single-cell data

- Human Cell Atlas
  https://www.humancellatlas.org/

- Mouse Organogenesis Cell Atlas
  https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/landing

- Tabula Muris Senis
  https://tabula-muris-senis.ds.czbiohub.org/

- EBI Single Cell Expression Atlas:
  https://www.ebi.ac.uk/gxa/sc/

Any questions?

1. Complex experiment designs and batch effects
2. Differentiation trajectory inference
3. Splicing and RNA velocity
4. Somatic mutations and its impact on gene expression

Resources:

- Analysis of single cell RNA-seq data (Sanger course) https://www.singlecellcourse.org/
- HKUMed Single-cell analysis tutorial workshop (HKU Med) https://statbiomed.github.io/HKU-single-cell-workshop/

Contents and assessment level

- Quantitative skills for genomic data analysis 1
  - Only for assignment
- Quantitative skills for genomic data analysis 2
- Transcriptomics 1
- Transcriptomics 2
  - Multiple gene analysis: Only for assignment
- Expression QTL 1
- Expression QTL 2
  - Machine learning, gene regulatory network: only for introduction
- Single cell transcriptomics 1
  - Clustering and cell type annotation: only for introduction
- Single cell transcriptomics 2
  - Trajectory inference, RNA velocity: only for introduction

Introduction level means understand the concept

Focus:
Hypothesis testing and its application in DEG, eQTL, and highly variable genes

Drop-in session (Zoom or my office)

You can join either or both slots; no appointment is needed.

Slot 1: 30th Sep (Thursday), 3-4pm

Slot 2: 13th Oct (Friday), 2-3pm

Join from Zoom:

Topic: BBMS3009 open office hour - Dr. Huang

Join Zoom Meeting

https://hku.zoom.us/j/91480501144

Meeting ID: 914 8050 1144

Or come to my office:

My office: 1-05E, 1/F, JCBIR, 5 Sassoon Road