Single Cell RNA-seq Preprocessing and QC

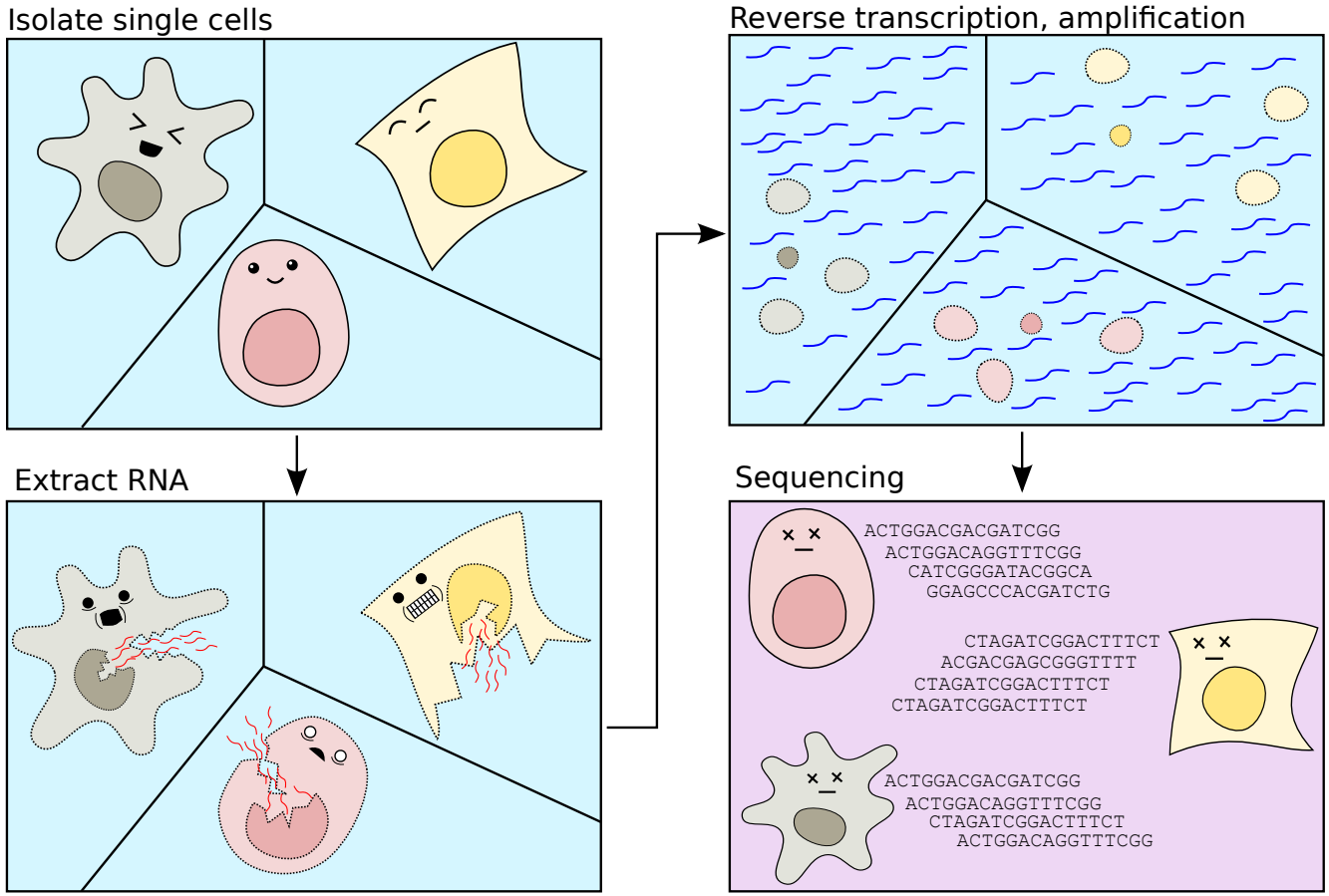
Hubert Rehrauer

(with slides from Ge Tan)

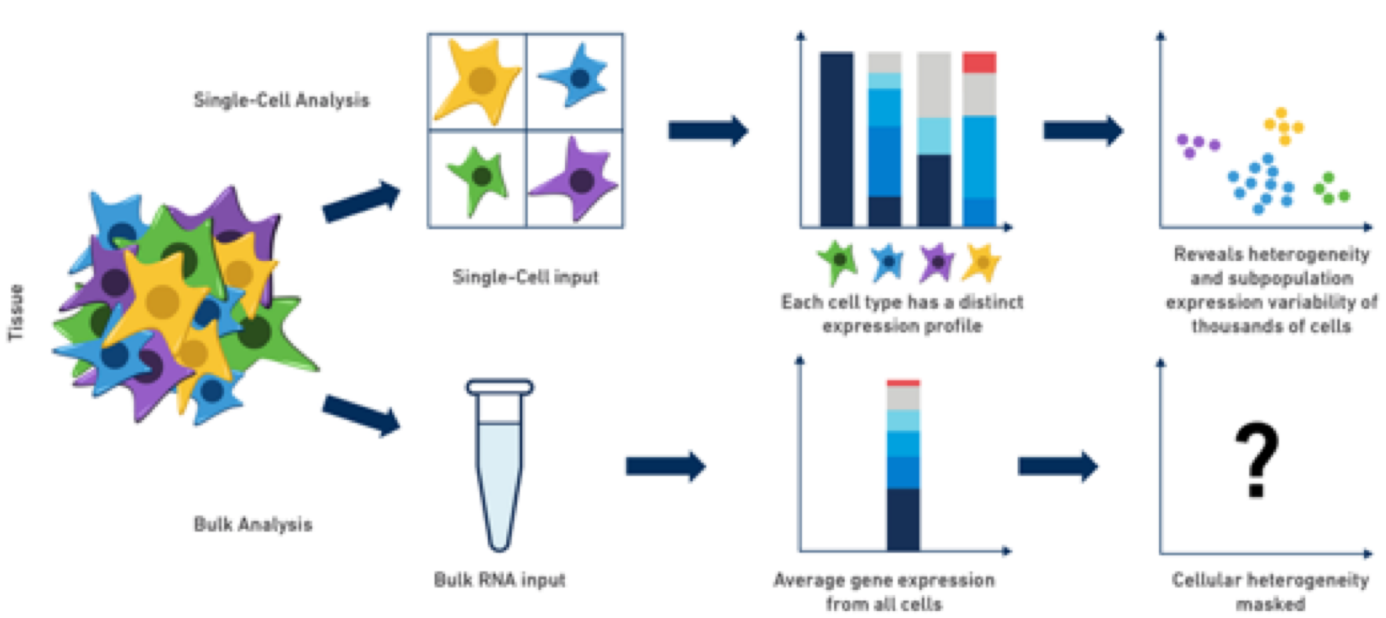
Outline

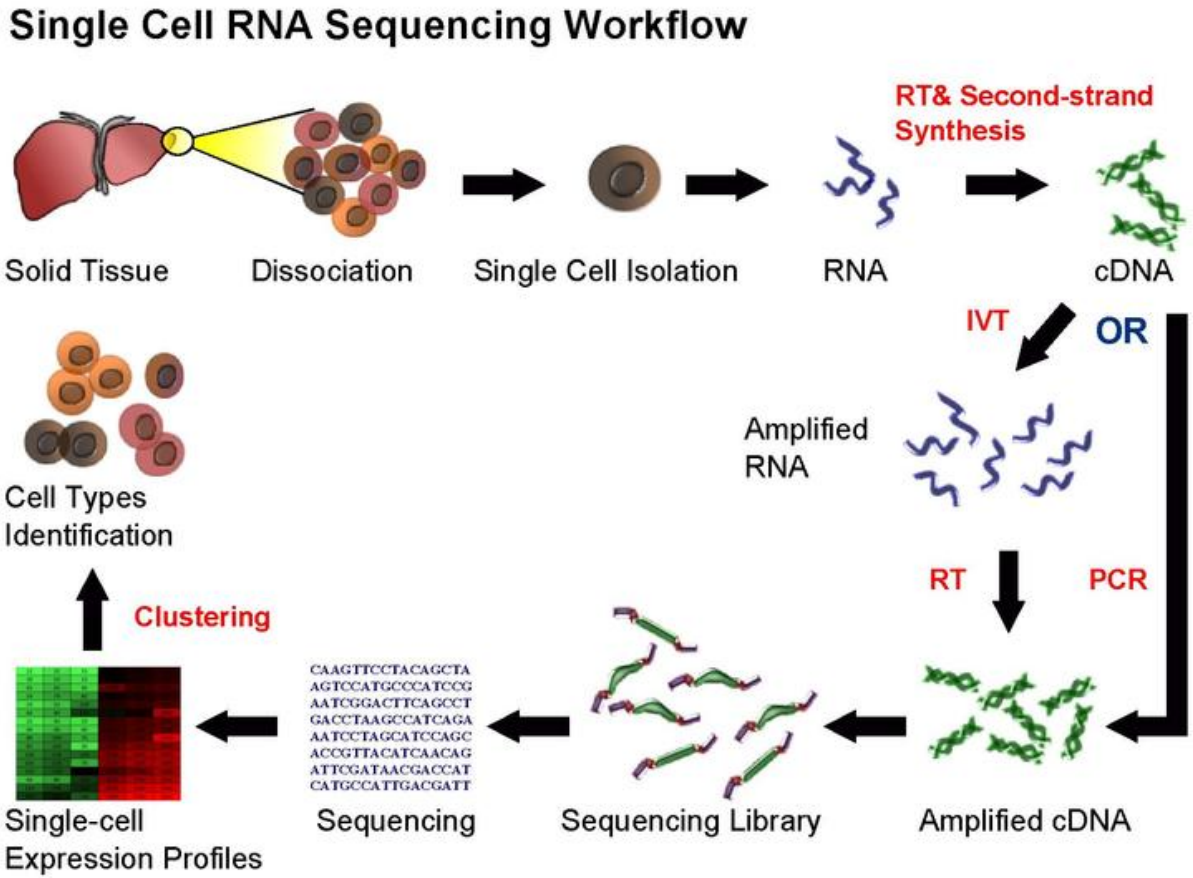
- Biological need, technical solution, noise characteristics

- QC/Issues: mitochondrial genes, ribosomal genes, few genes, emptyDrops, doublets, dopouts
- normalization
- dimensionality reduction
  - Matrix factorization
  - graph-based (t-SNE, UMAP)
  - Autoencoder
- batch correction, batch integration
- clustering



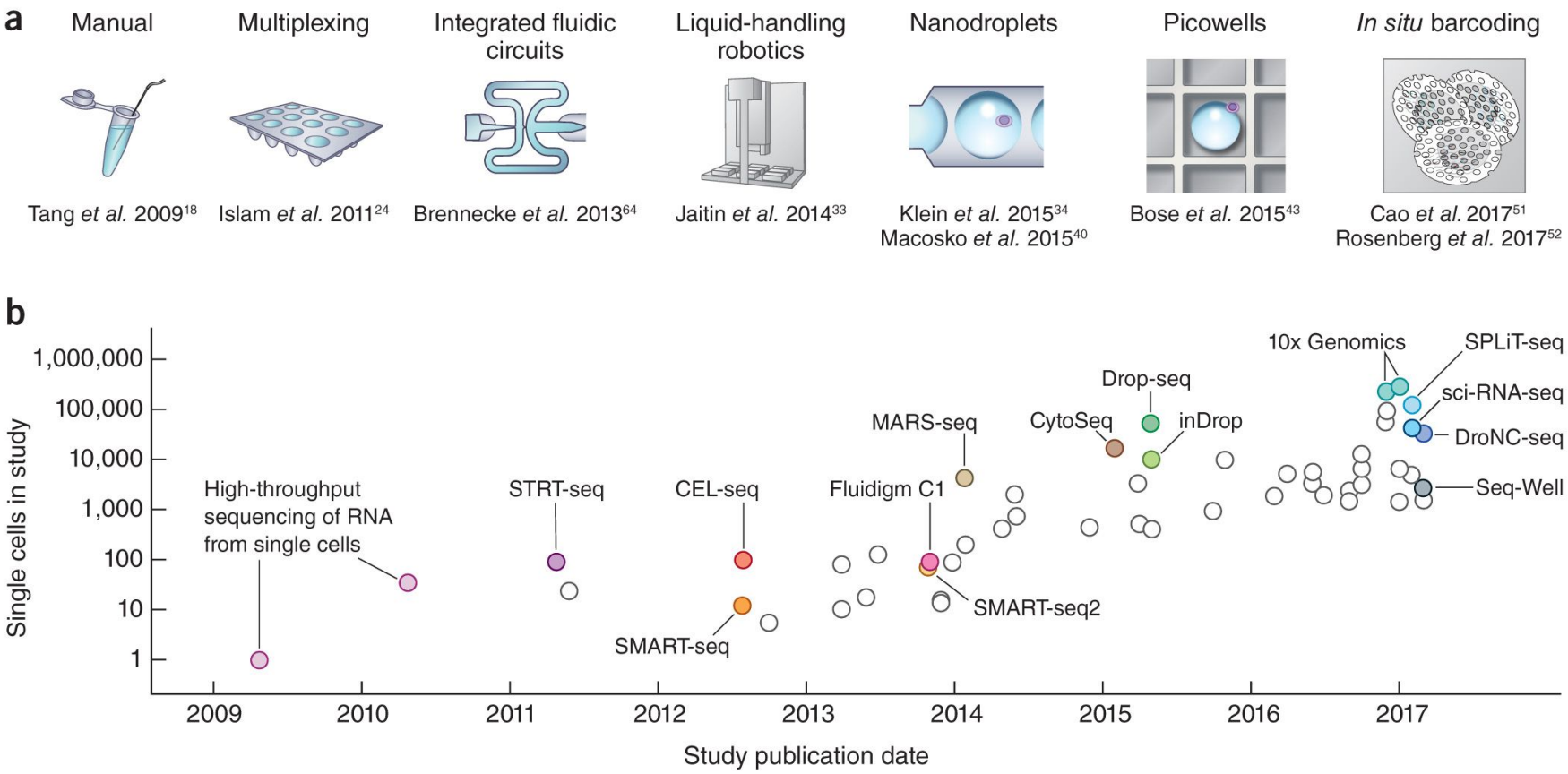
Adapted from Aaron Lun


Single-Cell Analysis
Tissue
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked



https://en.wikipedia.org/wiki/Single_cell_sequencing

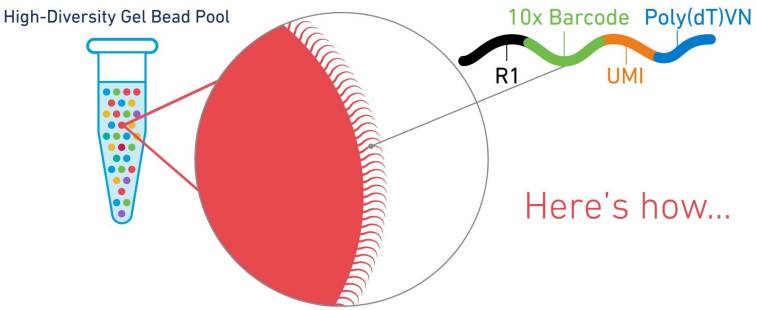
Technologies

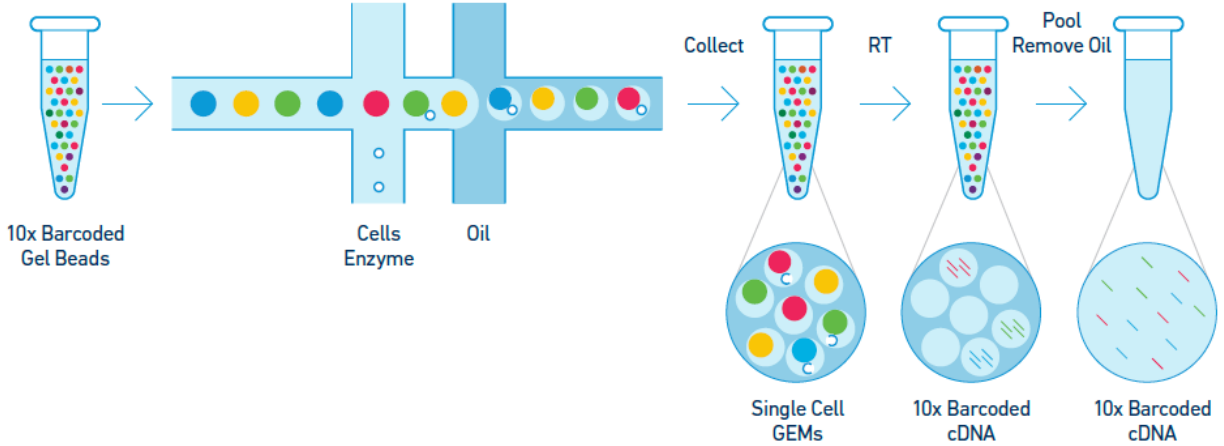


Svensson, 2018, Nature Protocols

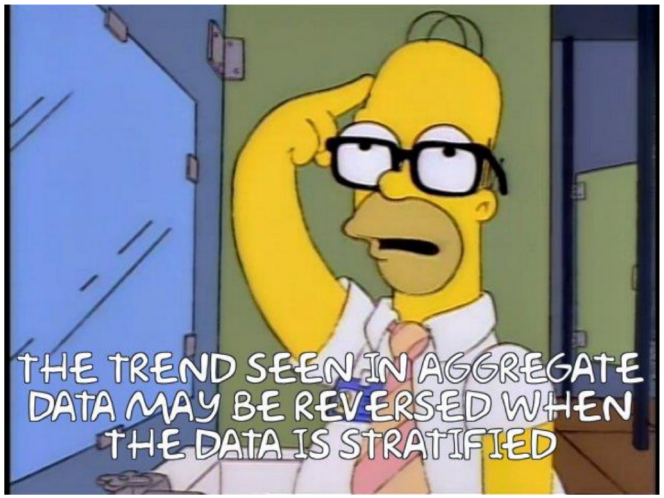
Nanodroplets systems – e.g. 10X Genomics

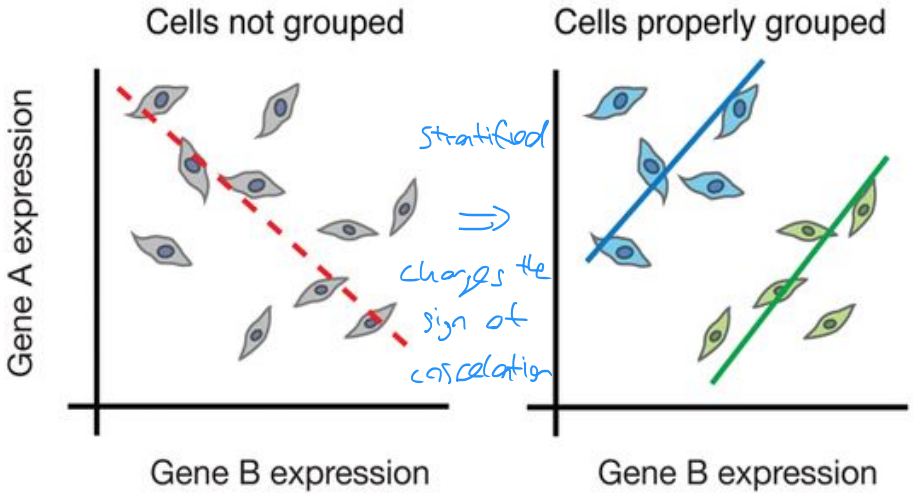
- High throughput (up to 80,000 cells per sample).
- Current standard in the field.
- Restricted to certain cell-types / sizes.





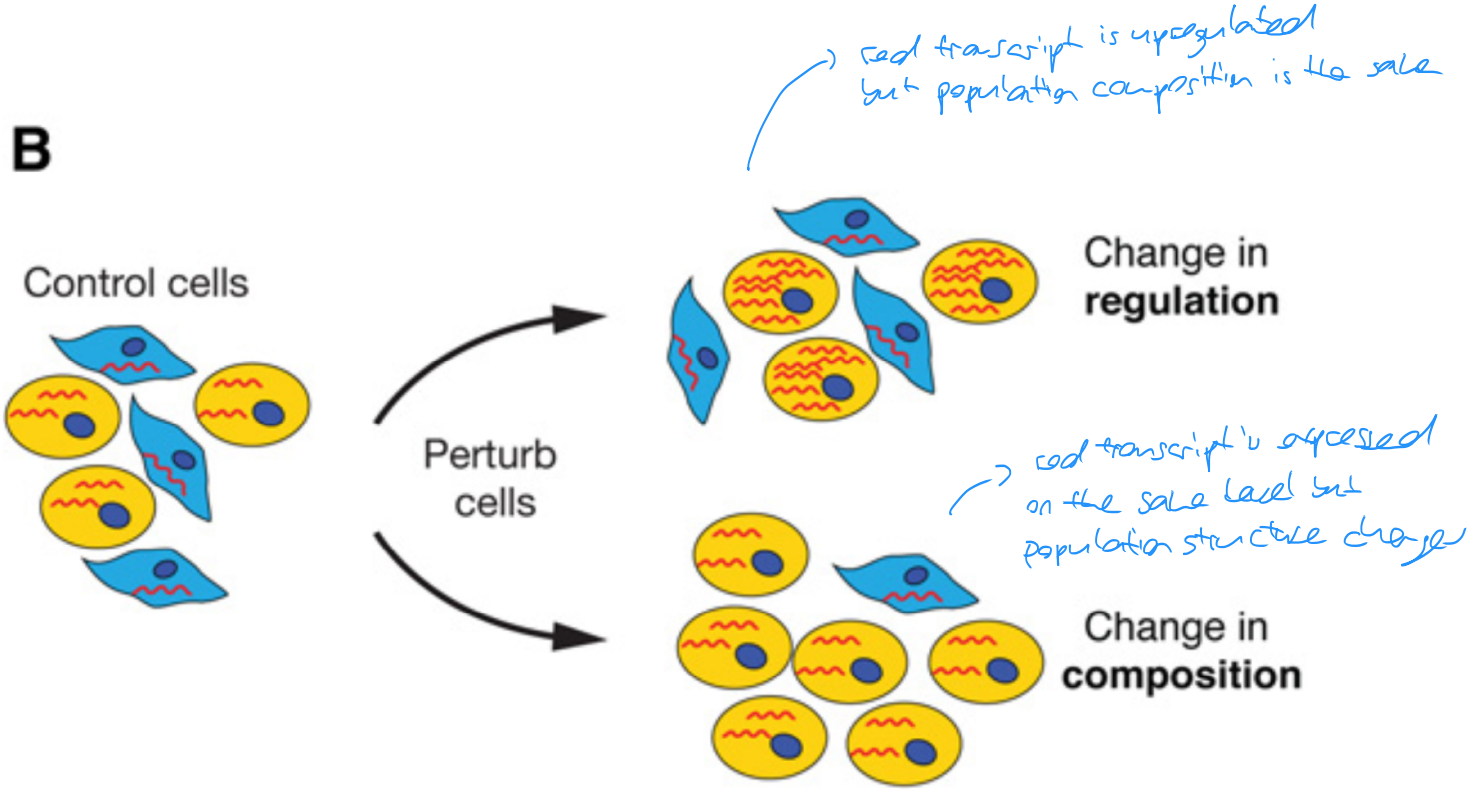
Single cell measurements preserve crucial information that is lost by bulk measurements





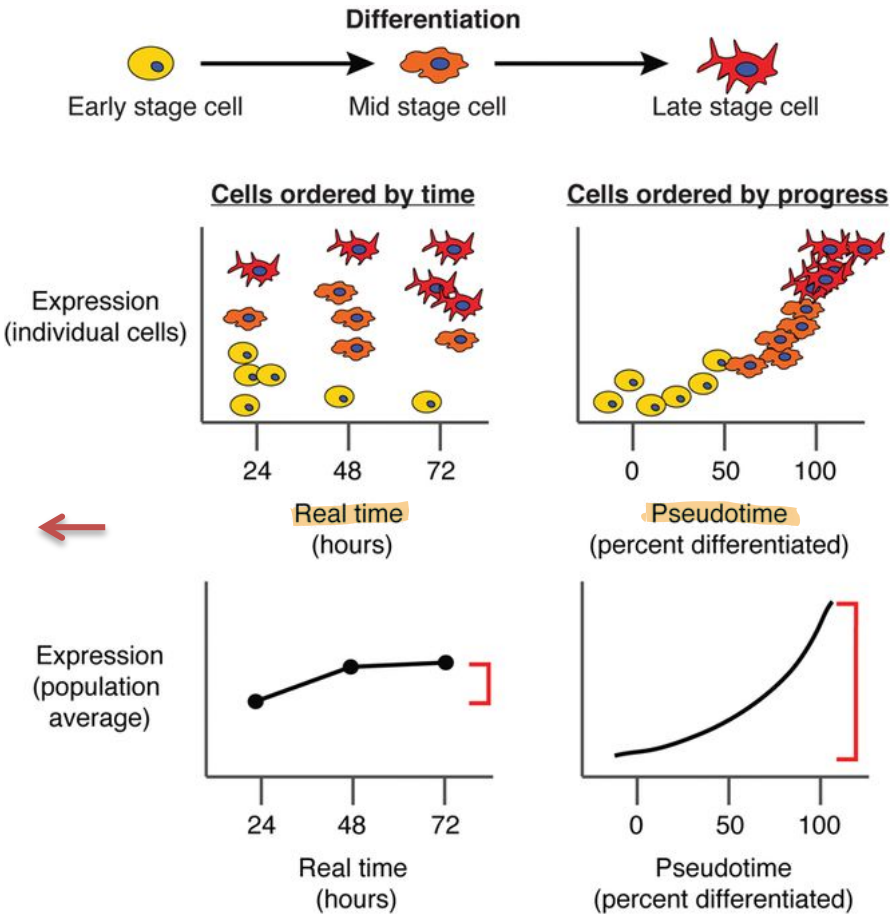
Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals

Trapnell Genome Res. 2015


B
Control cells
Perturb cells
Change in regulation
Change in composition
cool transcript is upregulated but population composition is the same
cool transcript is expressed on the same level but population structure changes

Time series experiments are affected by averaging when cells proceed through a biological process in an unsynchronized manner

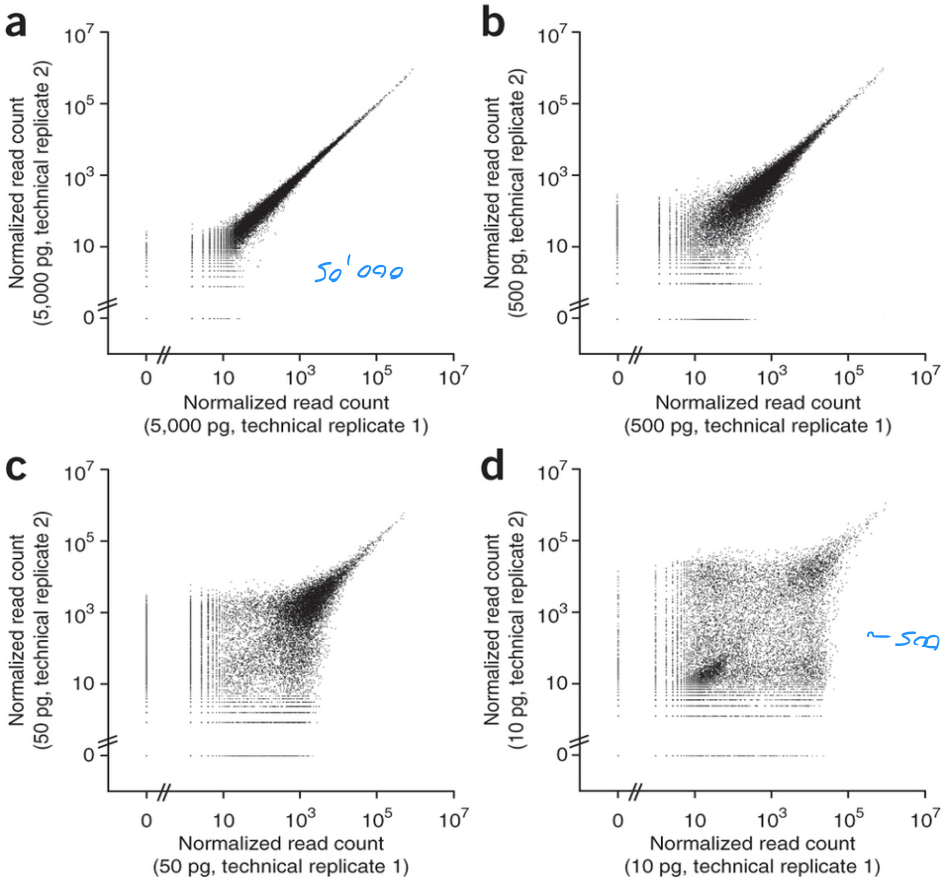
A single time point may contain cells from different stages in the process, blurring the dynamics of relevant genes



Reordering the cells in "pseudotime" according to biological progress eliminates averaging and recovers the true signal in expression

Trapnell Genome Res. 2015

Challenge in scRNAseq: high technical noise



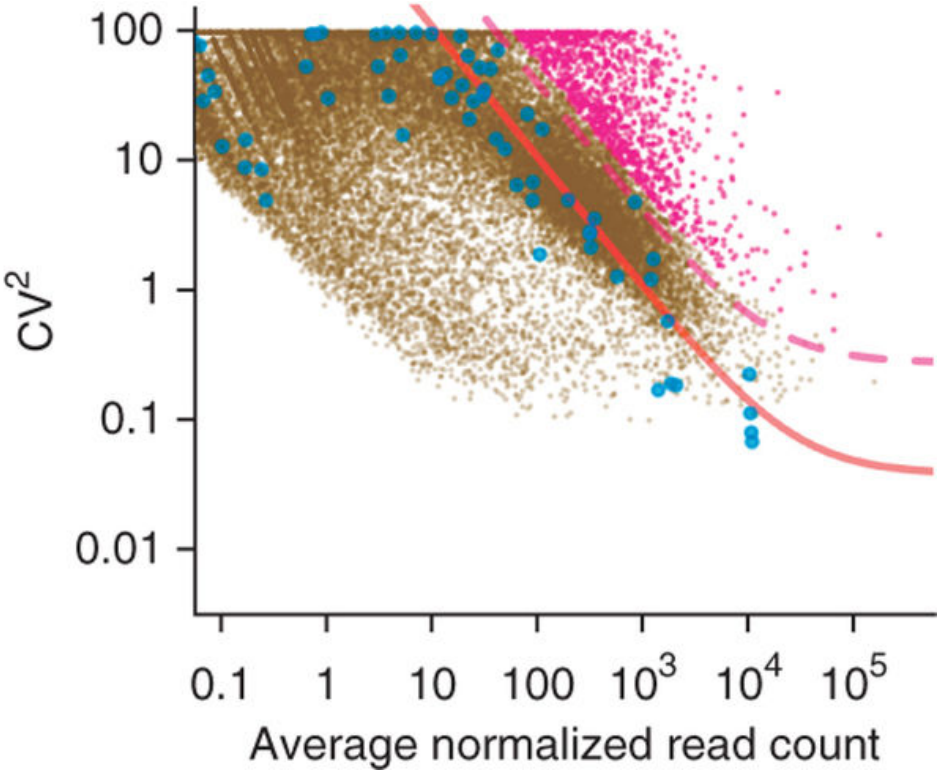
- low input means high amplification and high technical noise

⇒ More noise in low abundance species

Brennecke et al. Nature Methods 2013

Experiments with different input material - dilution series of total A. thaliana RNA

Variability of scRNA-seq counts

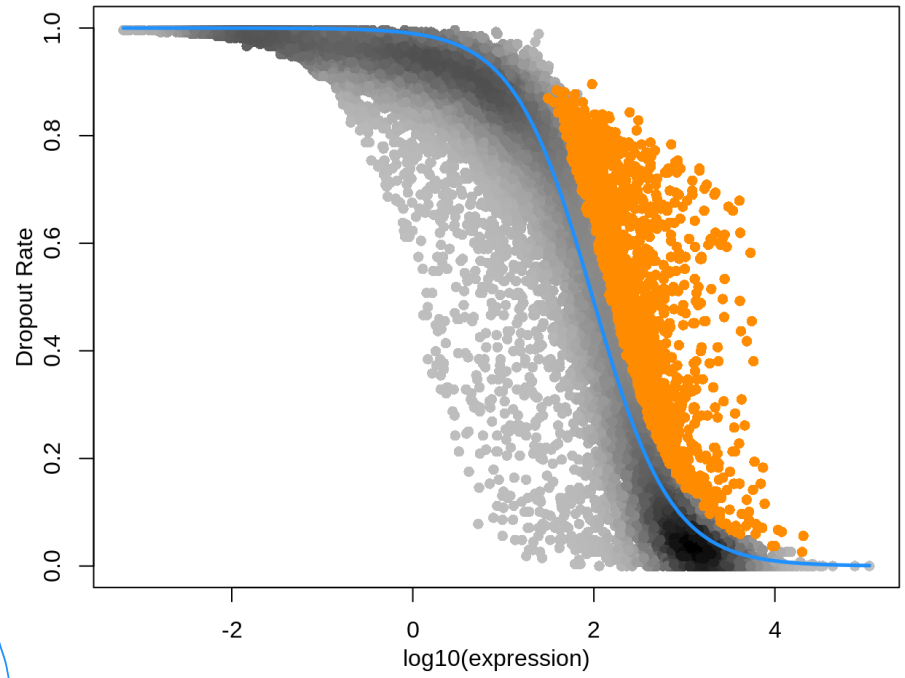


- Blue: spike-ins
- Pink: genes with potentially differential expression between cells

Brennecke et al. Nature Methods 2013

CV = st.dev/mean

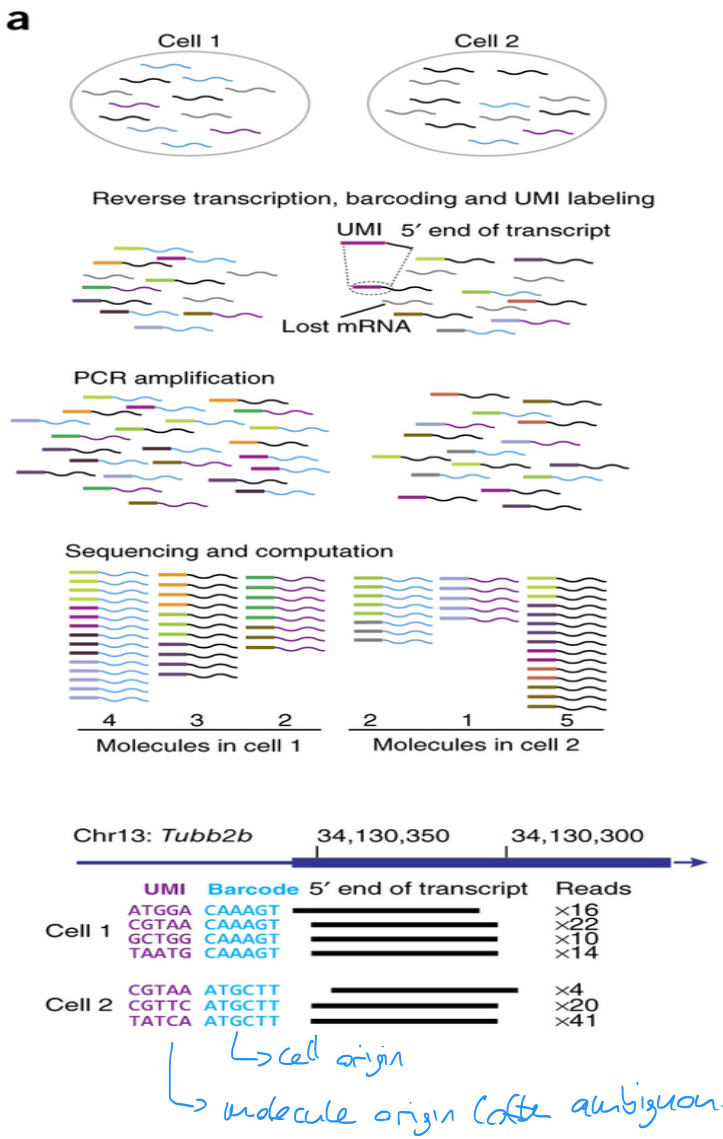
Dropouts: Genes not expressed



fraction of the cell in which the individual gene is not detected.

- Dropout Rate: Fraction of cells where gene is not detected
- genes with low average expression have high dropout rate
- 1000 – 10‘000 genes detected per cell (depends on single cell protocol and sequencing depth)
- > 15‘000 genes detect in bulk tissue
- Reasons for dropout
  - stochastic expression bursts
  - not enough reads in sequencing – lost by the sampling process
  - library prep failed to capture the transcript

Unique Molecular Identifiers (UMIs)

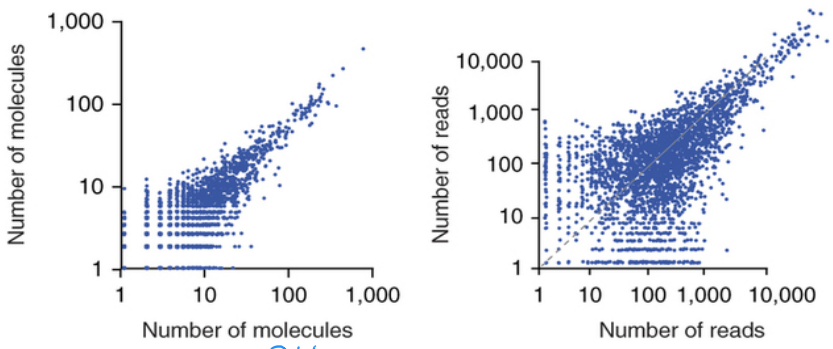


PCR duplicates can be identified

Two reads are considered duplicates if they have the same UMI, the same cell barcode and map to the same gene.

UMIs are only available for 3'-tagging protocols, e.g. the 10X.

Islam et al. Nature Methods 2014

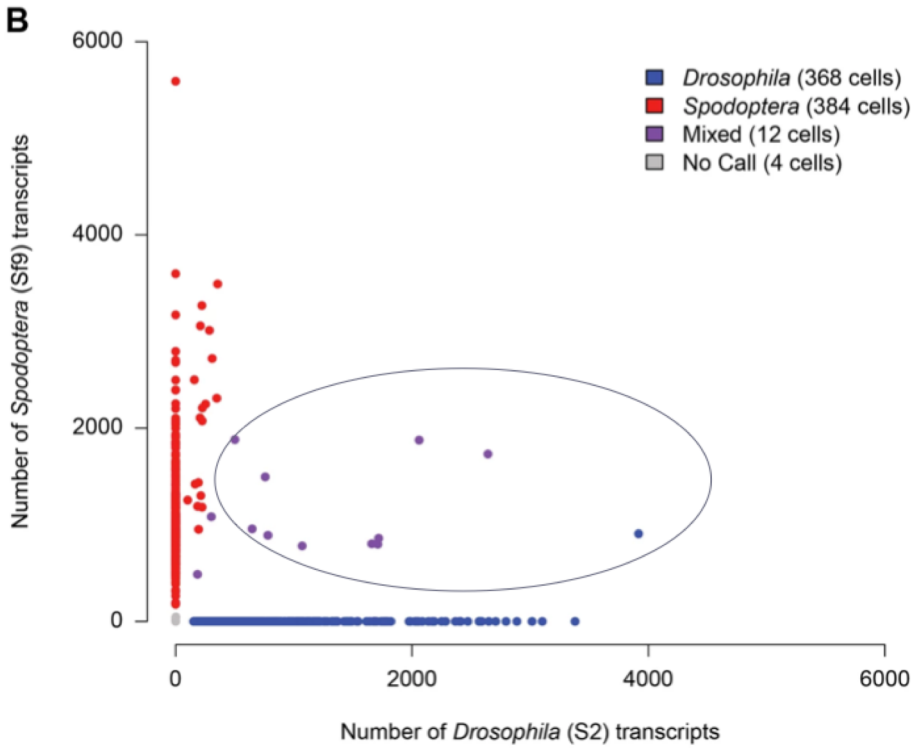


3'- AAAA ~2kb 5' ← not targeted
UMI — TTTT
20nt
=> only 20nt are sequenced / read
thus the UMI protocol can only be applied to 3-end sequencing of RNA transcripts

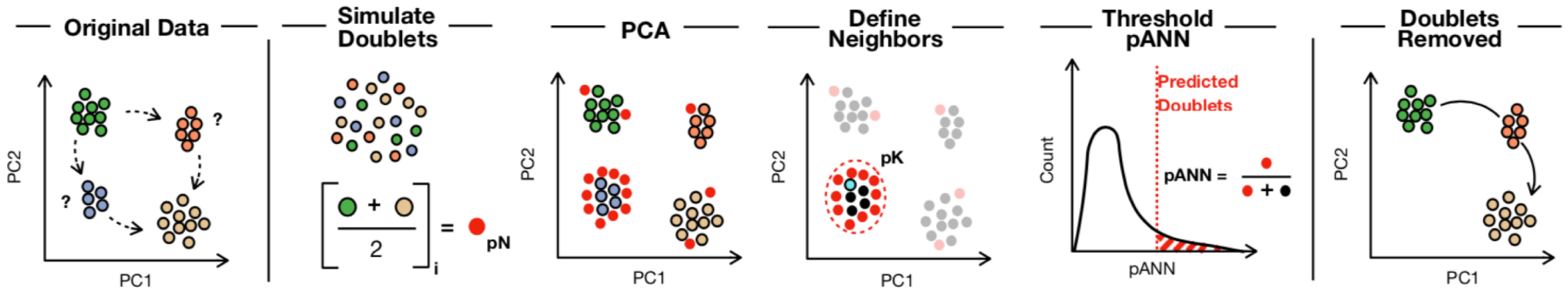
Quality Control: Doublets

- Barcode collisions: not enough barcodes
- Technical doublets: two cells in the same droplet (10X specs: +1% per 1000 cells)
- Biological doublets: two cells sticking tightly together and form a unit; need to do nuclear single-cell RNA-seq

- Test datasets for doublets consist of cells from two species



Doublet Detection



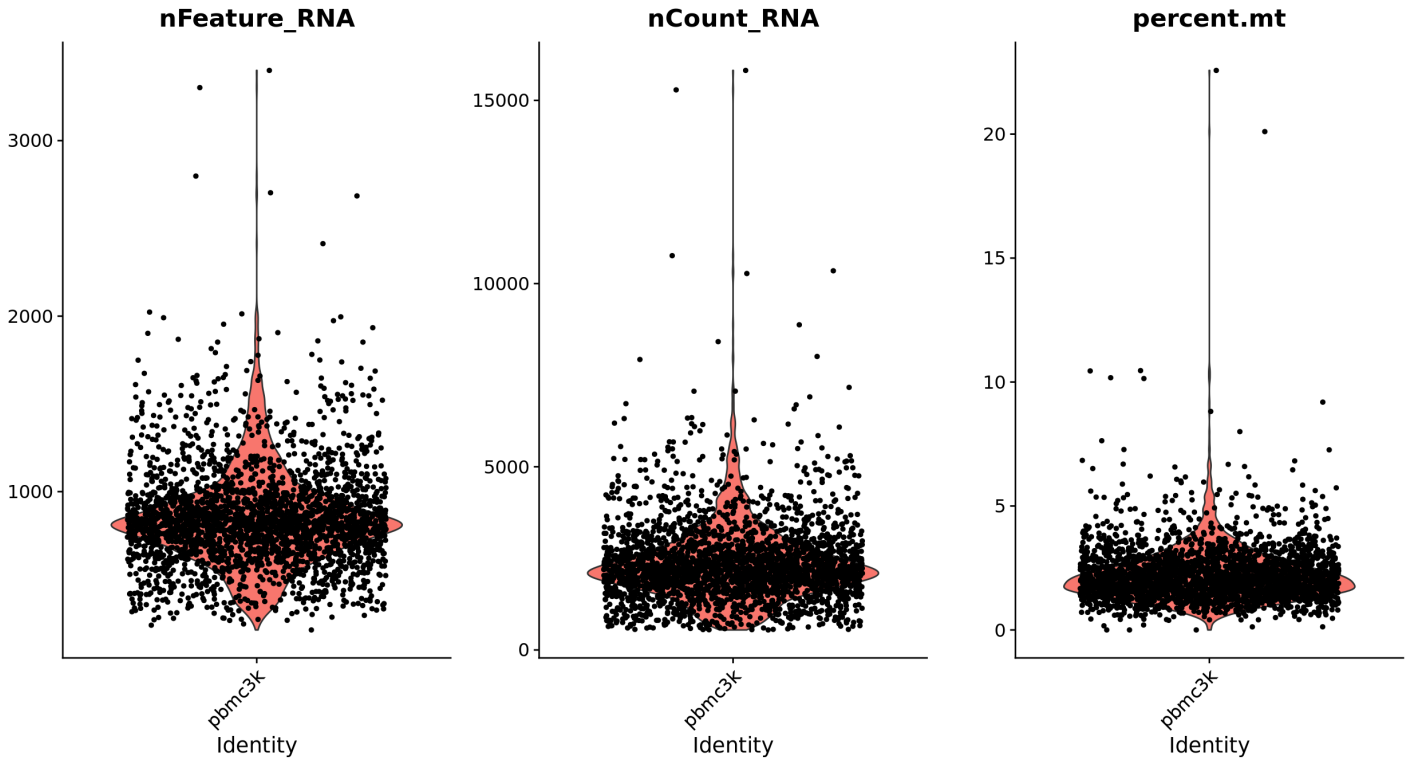
- see also:
  - https://github.com/plger/scDblFinder

https://github.com/chris-mcginnis-ucsf/DoubletFinder

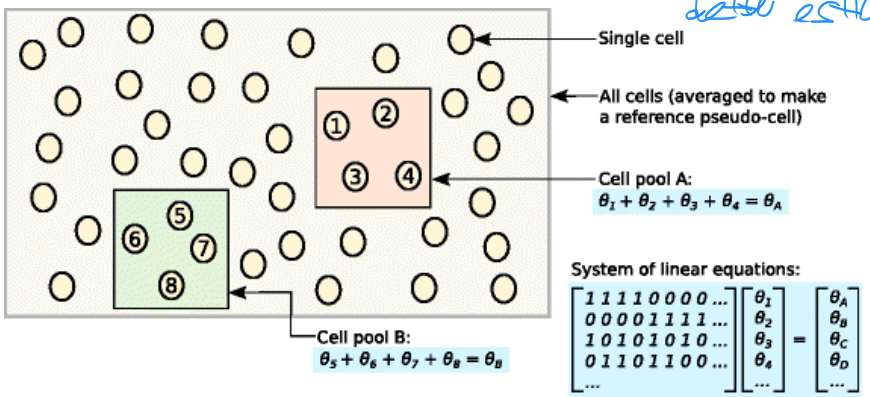
Other Quality Control Metrics

- high content of mitochondrial RNA
  - apoptotic cell
- high content of ribosomal RNA
  - failed library prep
- few reads sequenced or few genes detected
  - **empty drop**
  - failed library prep
  - alternative explanation: cell with low transcriptional activity

QC Filtering Plots



scran: Normalization using pools of cells



- Define a pool of cells
- Sum expression values across all cells in the pool
- Normalize the cell pool against an average reference, using the summed expression values
- Repeat this for many different pools of cells to construct a linear system
- Deconvolute the pool-based size factors to their cell-based counterparts

Lun, Genome Biology, 2016, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7

Normalization: Performance comparison papers

nature methods

Analysis | Published: 26 February 2018

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson & Mark D Robinson

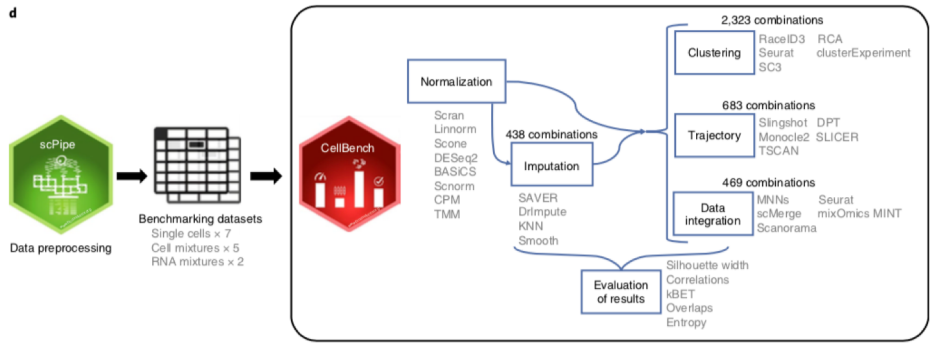
Nature Methods **15**, 255–261(2018) | Cite this article

10k Accesses | **61** Citations | **158** Altmetric | Metrics

nature methods ANALYSIS
https://doi.org/10.1038/s41592-019-0425-8

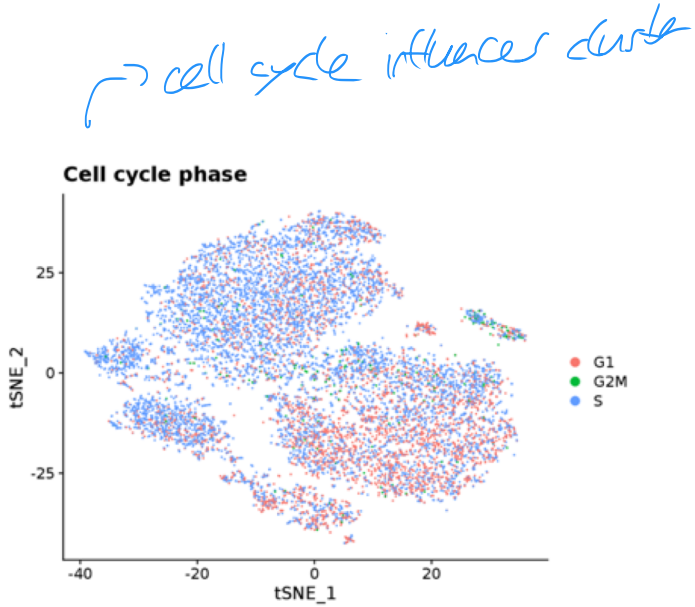
Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments

Luyi Tian[1,2*], Xueyi Dong[1,3], Saskia Freytag[1,4], Kim-Anh Lê Cao[5], Shian Su[1], Abolfazl JalalAbadi[5], Daniela Amann-Zalcenstein[1,2], Tom S. Weber[1,2], Azadeh Seidi[6], Jafar S. Jabbari[6], Shalin H. Naik[1,2] and Matthew E. Ritchie[1,2*]



Normalization: Cell Cycle causes unwanted variation

- Requires normalization considering cell cycle as latent variable



Dimensionality Reduction

- Removes redundancy in data
- Makes subsequent analyses more efficient and robust
  - clustering
  - classification
  - cell characterization
- Many features and high dimensionality make classification more erroneous

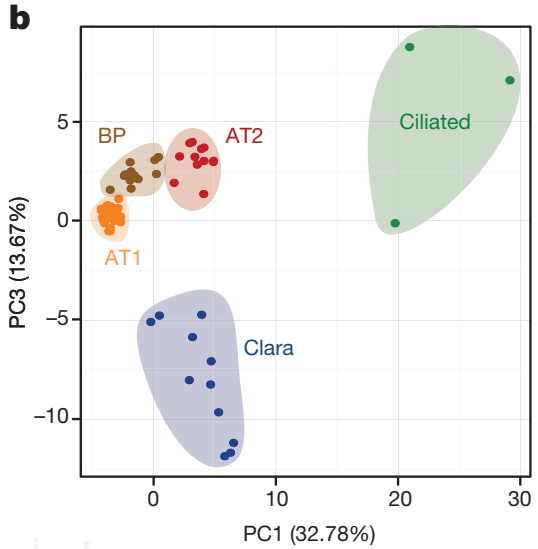
Dimensionality Reduction Methods

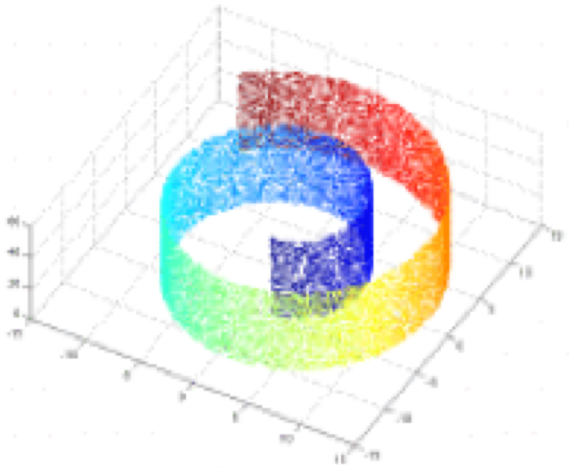
- Matrix Factorization
  - PCA
  - MDS – multi-dimensional scaling
  - ….
- Graph-based
  - t-SNE
  - UMAP

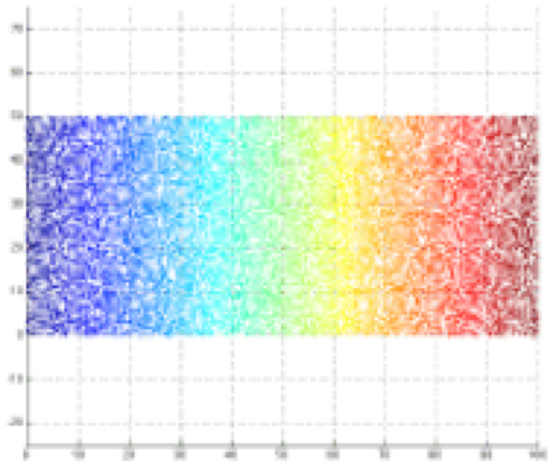
- Autoenconder (based on matrix factorization)

PCA

- PCA is simple and efficient
- But can not cope with complex structures because it is a linear projection.



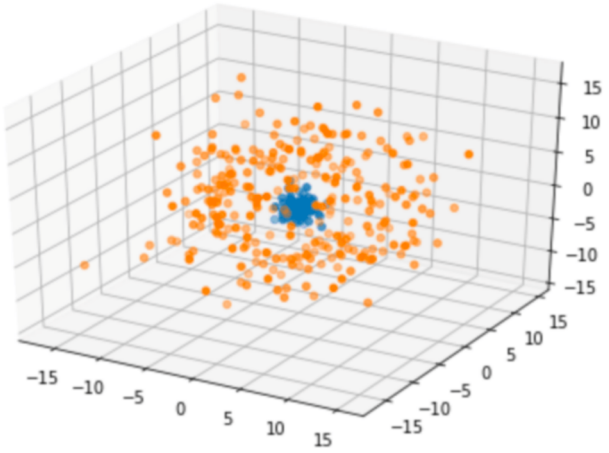




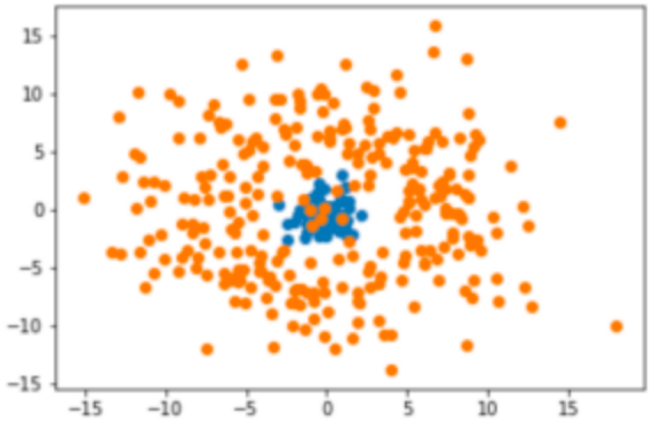
PCA vs t-SNE

original space:

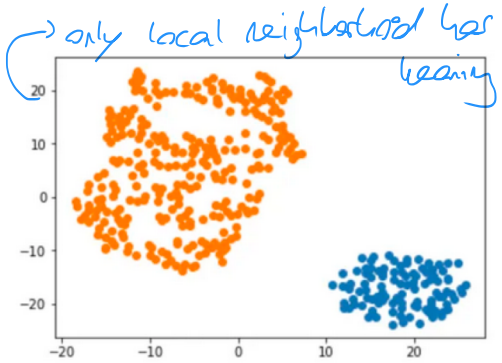
- blue center
- orange hull

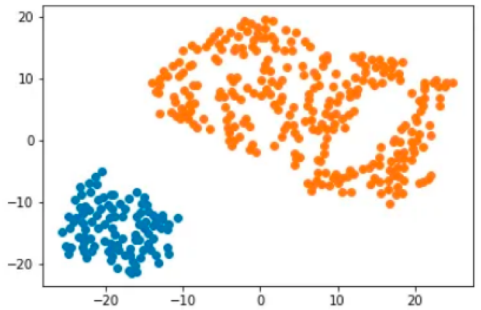


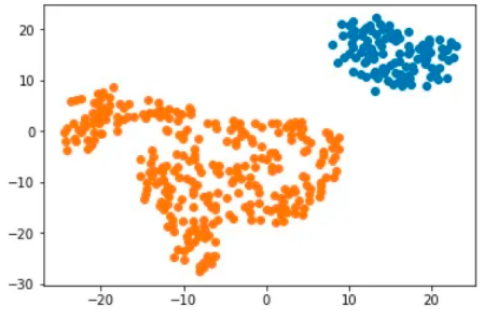
PCA



t-SNE







- t-SNE is not deterministic
- only local neighborhood has a meaning

https://mlexplained.com/2018/09/14/paper-dissected-visualizing-data-using-t-sne-explained/

t-SNE

- Step 1: In the high-dimensional space, create a probability distribution that dictates the relationships between various neighboring points
- Step 2: Recreate a low dimensional space that follows that probability distribution as best as possible.

- the “t” in t-SNE comes from the t-distribution, which is the distribution used in Step 2. The “S” and “N” (“stochastic” and “neighbor”) come from the fact that it uses a probability distribution across neighboring points.

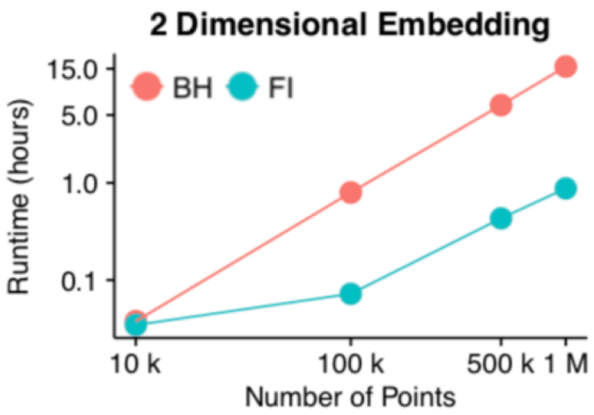
https://mlexplained.com/2018/09/14/paper-dissected-visualizing-data-using-t-sne-explained/

t-SNE

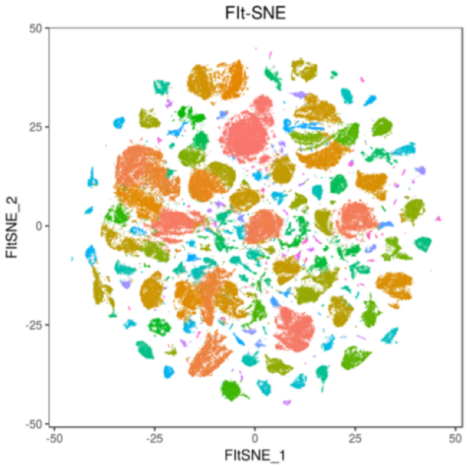
- Finds a low-dimensional representation of high-dimensional data
  - preserve distances to neighbouring cells
  - non-linear, different transformations on different regions
- Powerful, but need to fiddle with random seed and perplexity
  - 5-50 usually; default value for 10X: 30
  - Often on PC space, but not mandatory
- Nice blog about t-SNE: https://distill.pub/2016/misread-tsne/
- Implementation: Rtsne or much faster FIt-SNE

- Fast Fourier Transform-accelerated Interpolation-based t-SNE - $O(n)$

Linderman et al (2017) *BioRxiv*



250k cells
1 hour

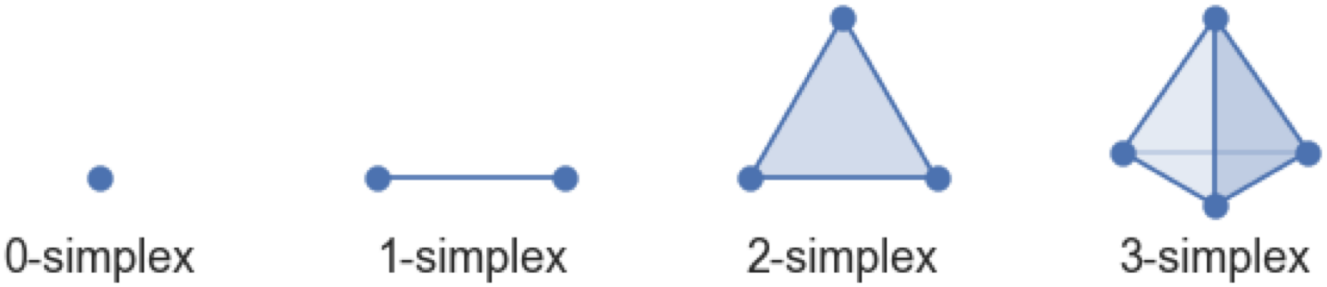


Linderman et al 2017 / Seurat Pipeline

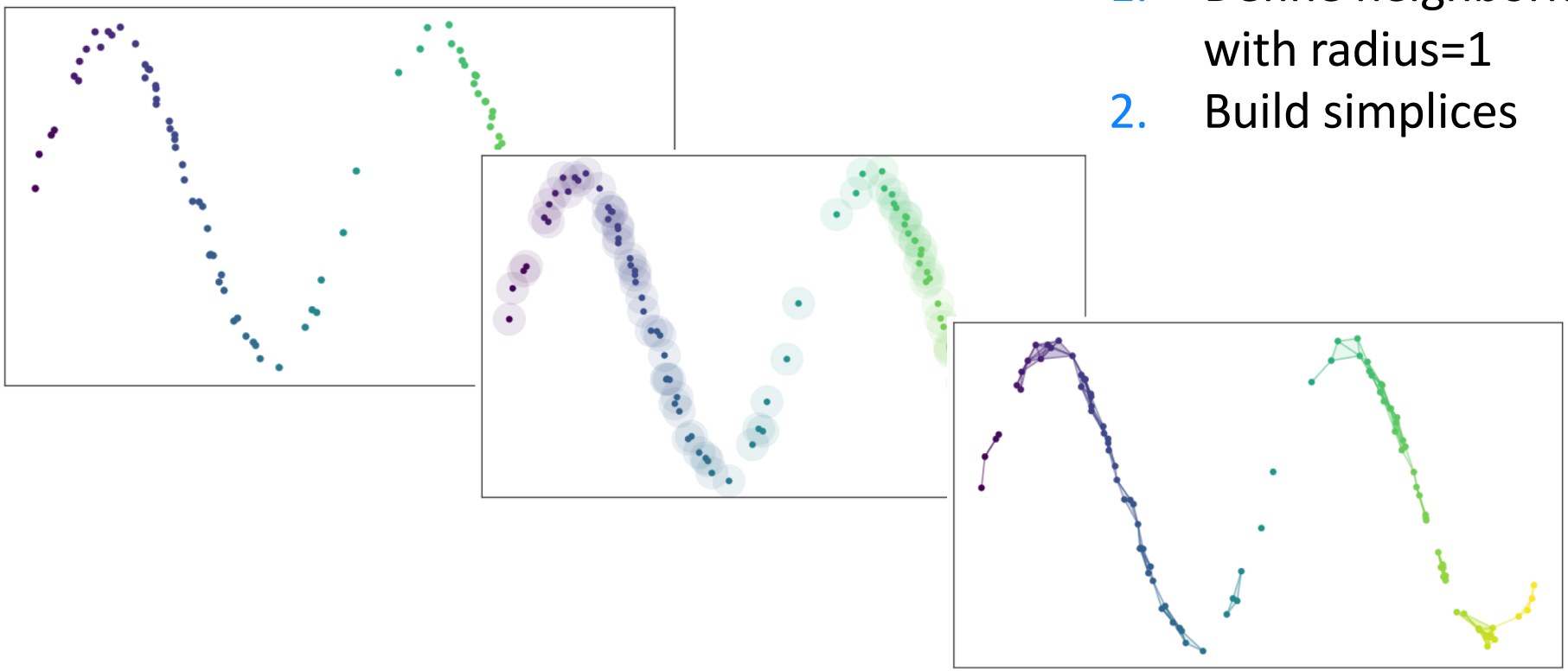
UMAP

- Uniform Manifold Approximation and Projection

- Approach: Find for each point the neighbors and build simplices



UMAP steps

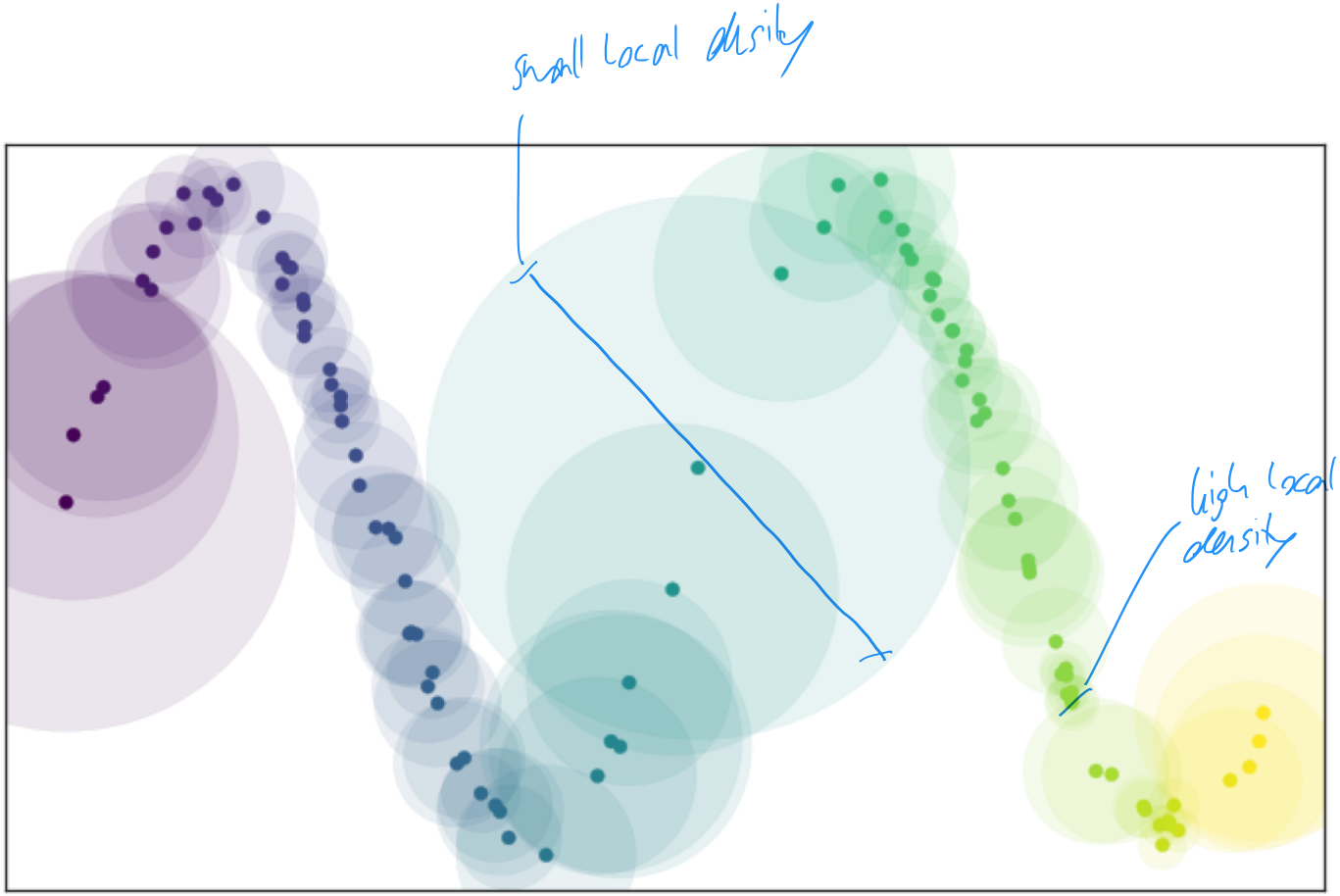


1. Define neighborhood with radius=1
2. Build simplices

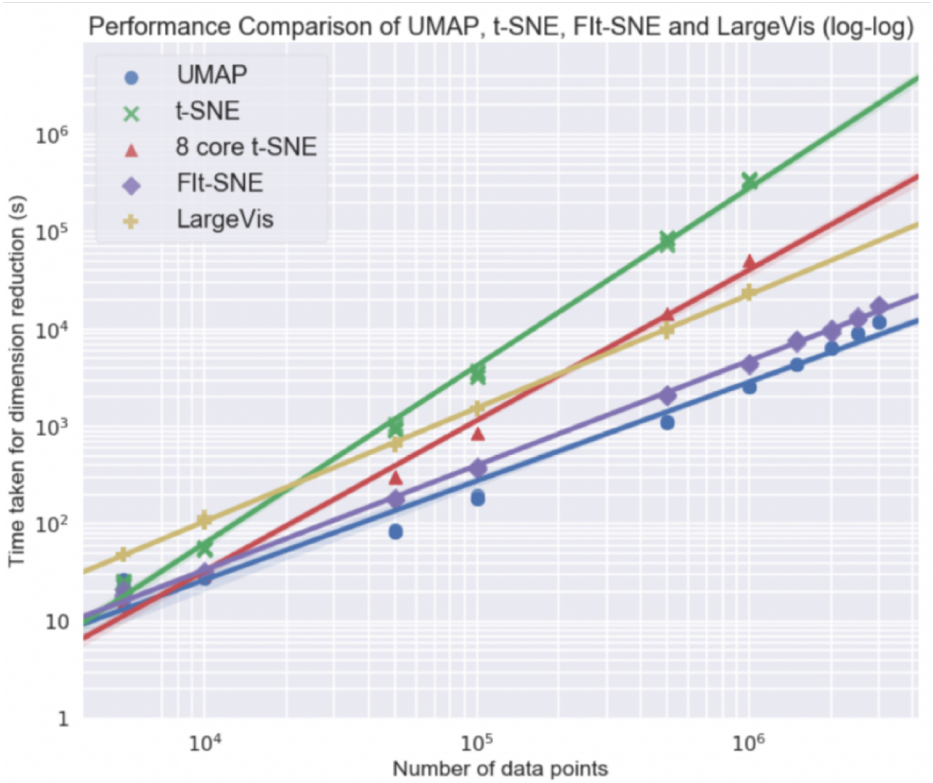
https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP

- All balls have radius=1 using a locally varying metric
- Avoids Clumping



UMAP computing time



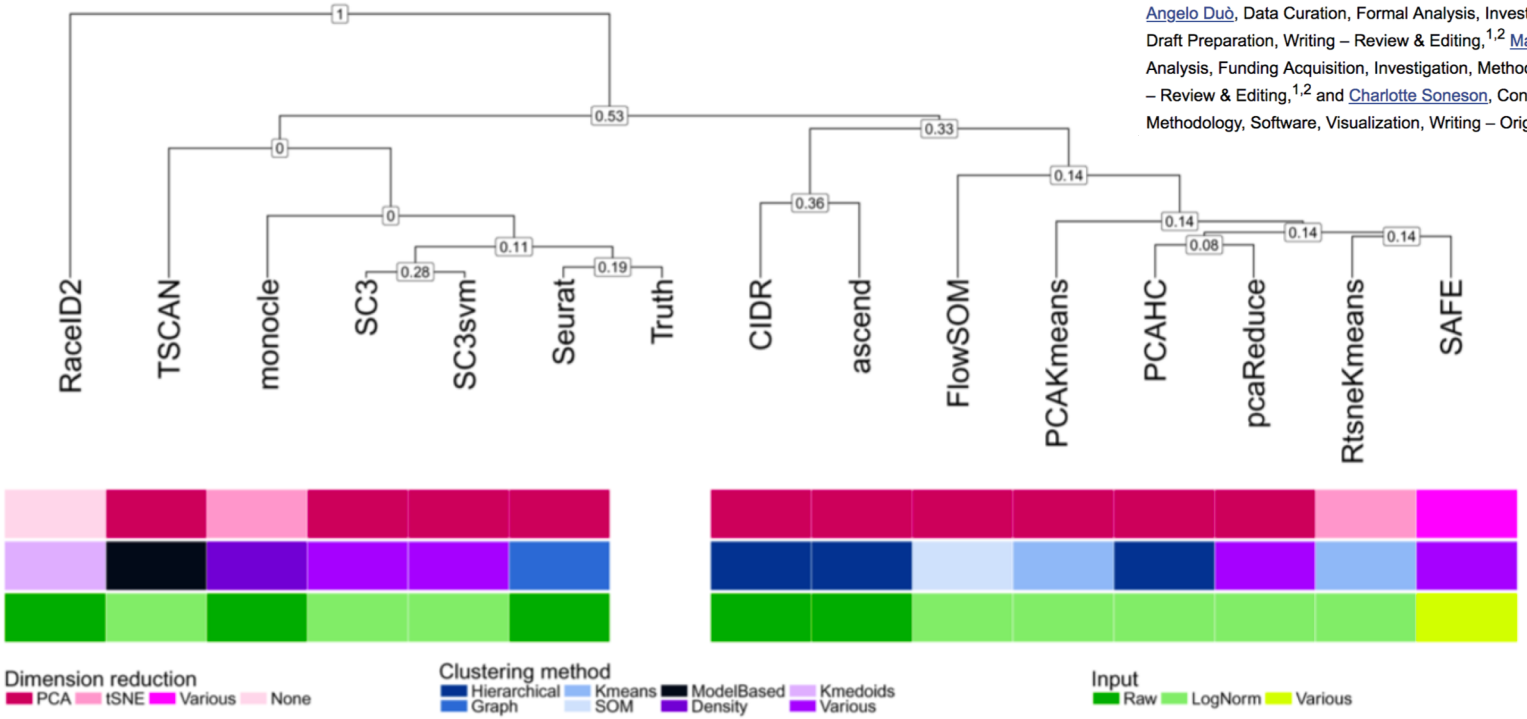
- UMAP is faster and scales better than classical t-SNE
- comparable to FIt-SNE

Clustering

Version 2. F1000Res. 2018; 7: 1141.
Published online 2018 Sep 10. doi: 10.12688/f1000research.15666.2
PMCID: PMC6134335
Other versions
PMID: 30271584

A systematic performance evaluation of clustering methods for single-cell RNA-seq data

Angelo Duò, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing,[1,2] Mark D. Robinson, Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Writing – Original Draft Preparation, Writing – Review & Editing,[1,2] and Charlotte Soneson, Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing[a,1,2]



⇒ scRNA seq matrices are sparse. Depending on the analysis you may have to transform this representation in a dense representation

Figure 4. Clustering of the methods based on the average similarity of their partitions across data sets, for the true number of clusters.
Numbers on internal nodes indicate the fraction of dendrograms from individual data sets where a particular subcluster was found.