Lecture 19-20: Single-cell genomics

- Introduction, Missing value imputation, Dimensionality reduction
- Trajectory inference, Spatial reconstruction

Single-cell RNA-seq

BREAKTHROUGH OF THE YEAR

Development cell by cell

With a trio of techniques, scientists are tracking embryo development in stunning detail

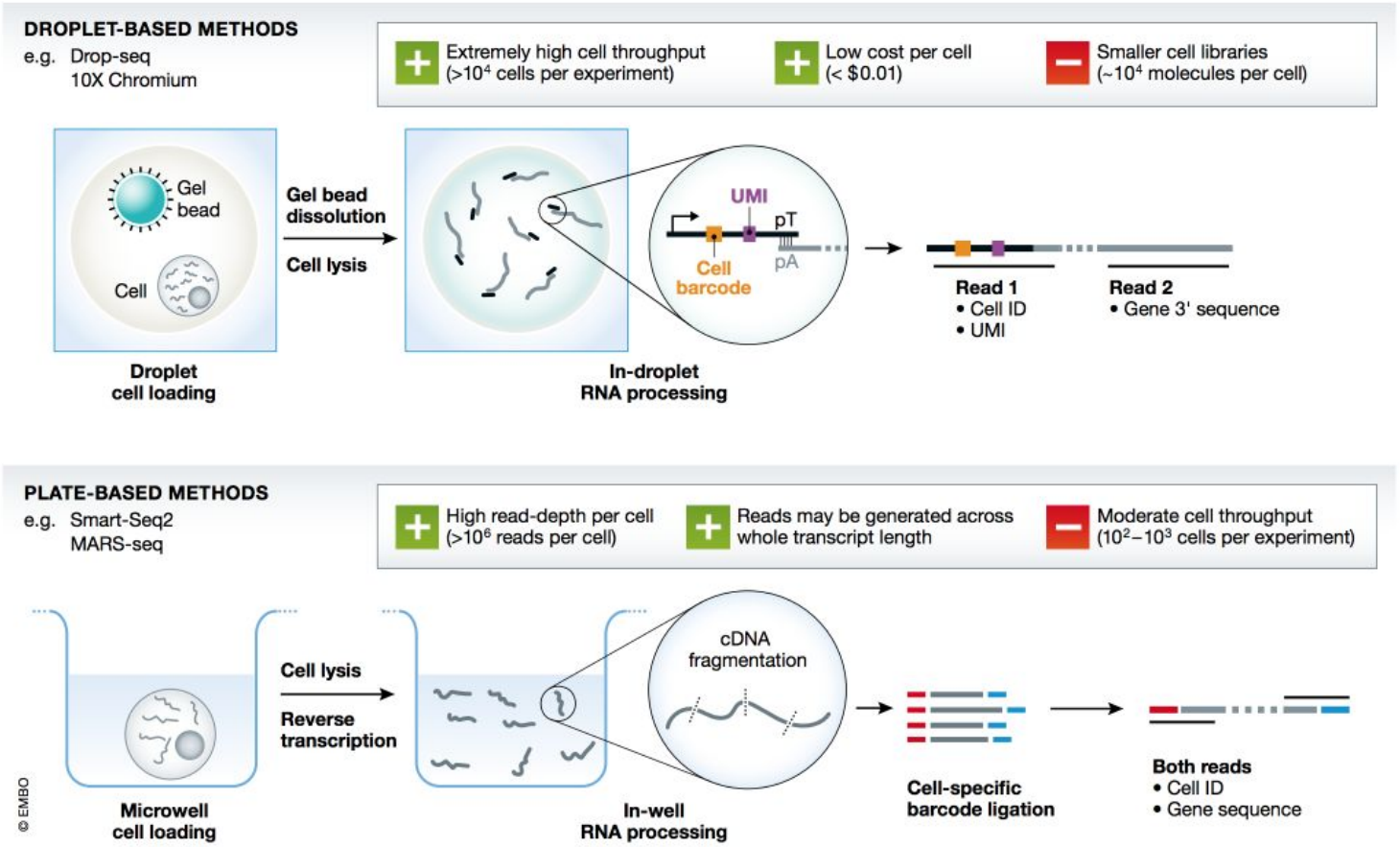
A zebrafish embryo at an early stage of development. Fluorescent markers highlight cells expressing genes that help determine the type of cell they will become. (JEFFREY FARRELL, SCHIER LAB/HARVARD UNIVERSITY)

The single-cell revolution is just starting.

— Elizabeth Pennisi

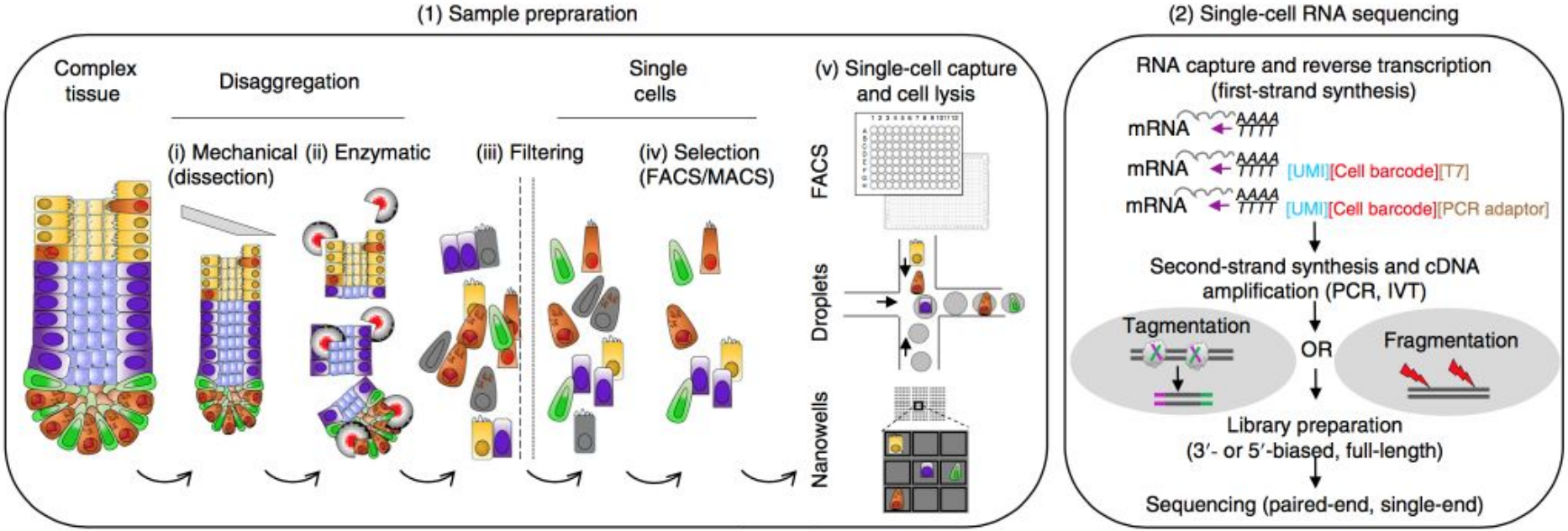
https://vis.sciencemag.org/breakthrough2018/finalists/#cell-development

Single-cell RNA-seq



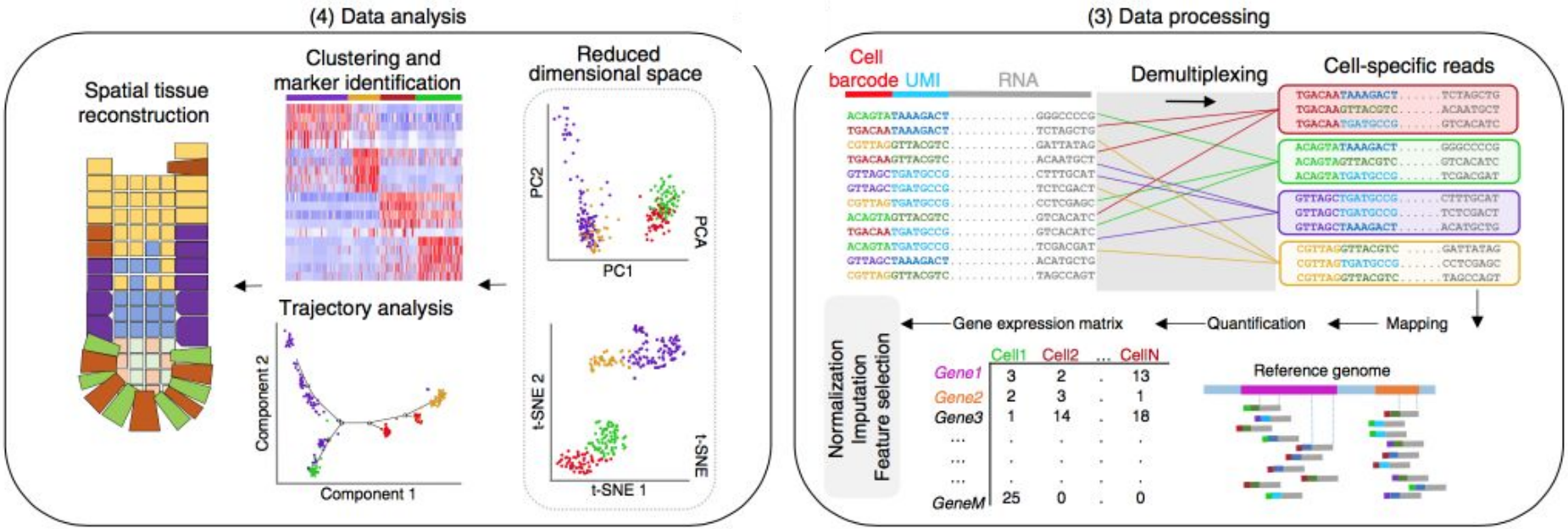
Griffiths (2018) Mol. Sys. Biol.

Single-cell RNA-seq



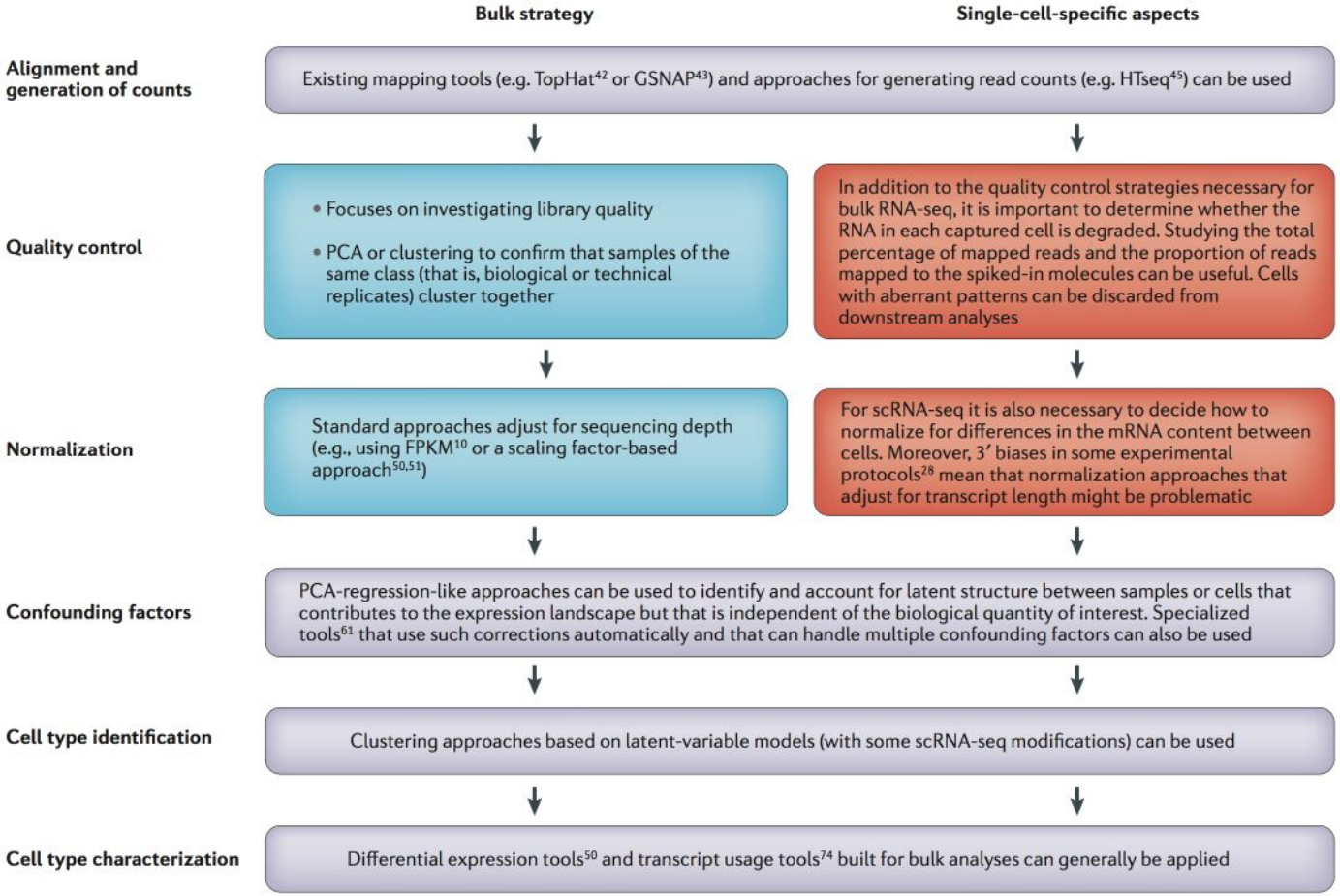
Lafzi (2018) Nat. Prot.

Single-cell RNA-seq



Lafzi (2018) Nat. Prot.

Single-cell RNA-seq

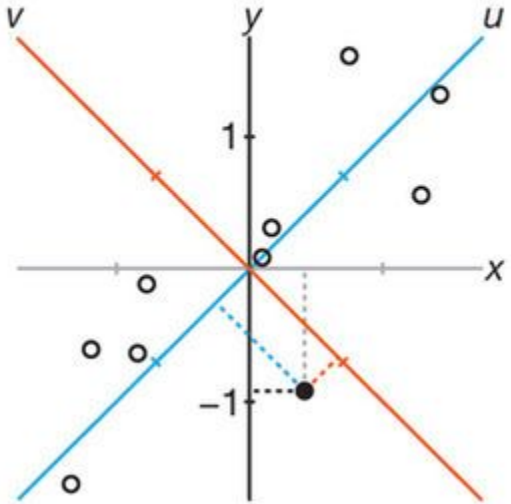


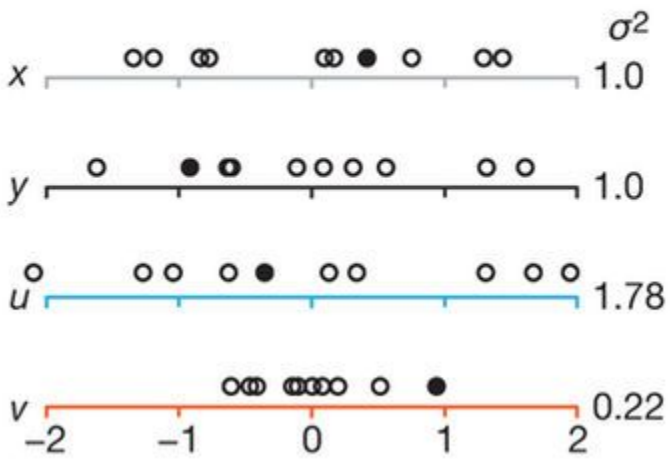
Stegle (2015) Nat. Rev. Genet.

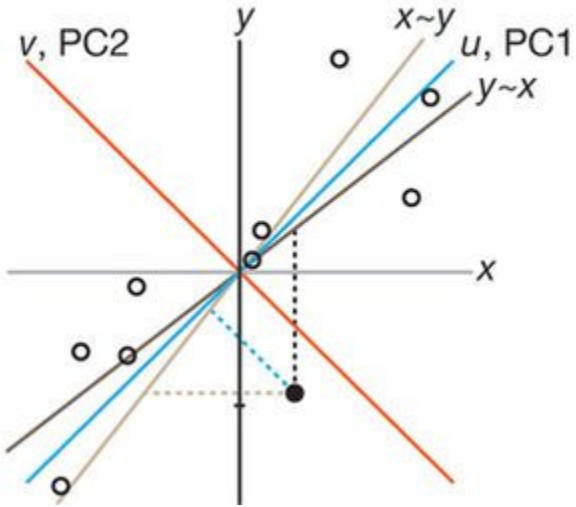
Dimensionality reduction by PCA

PCA geometrically projects data onto a lower-dimensional space

- Each lower dimension is a 'linear' combination of correlated original dimensions.
- The principal components (PCs) represent the directions of maximum variation.

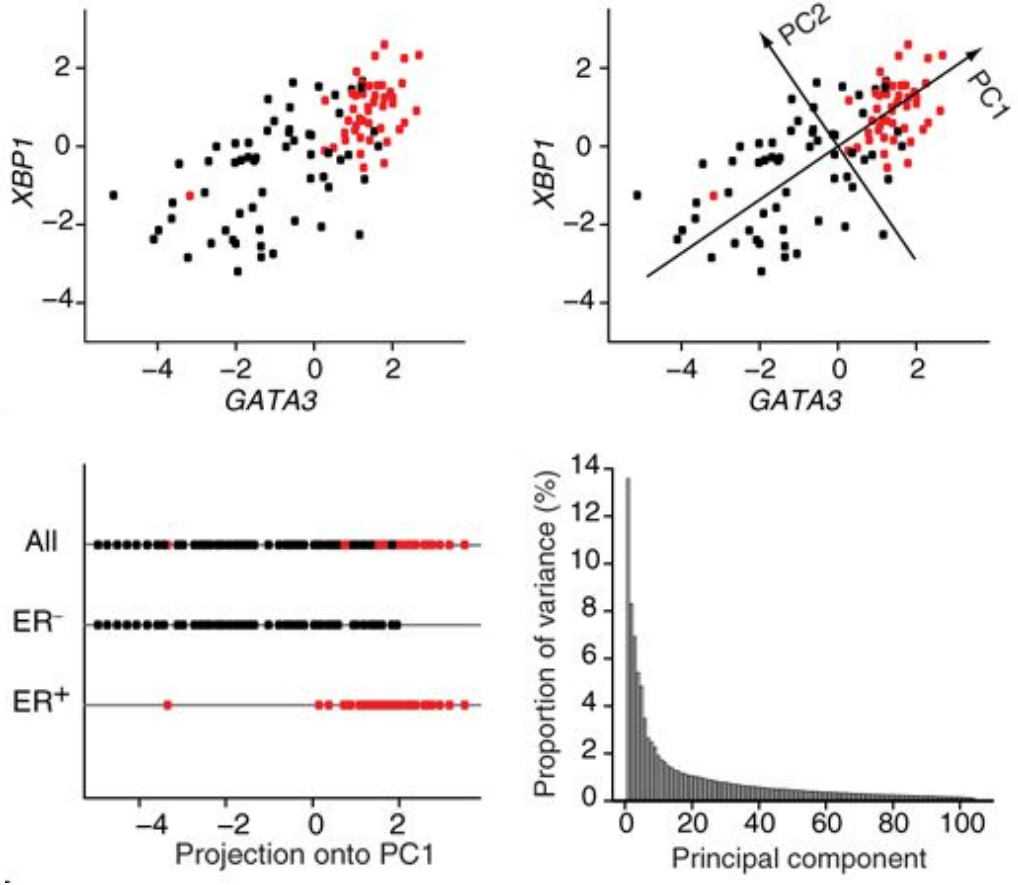






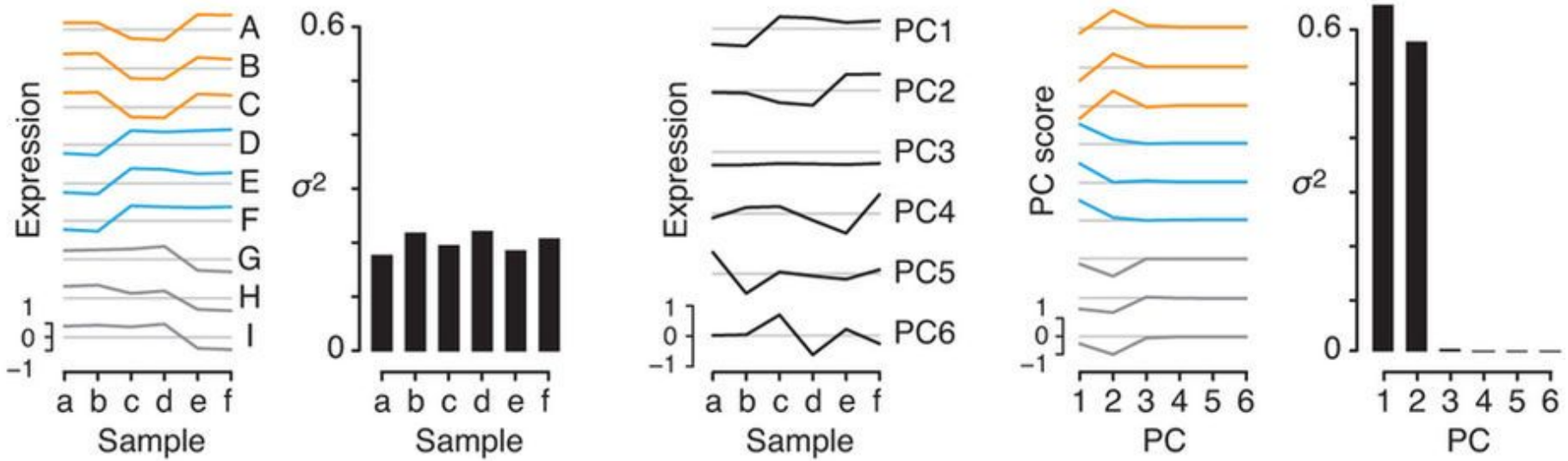
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



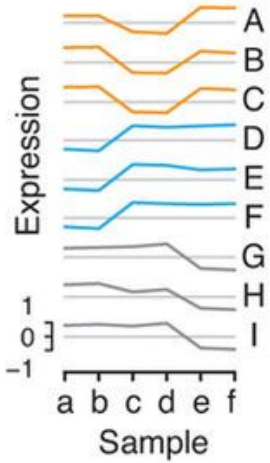
Ringnér (2008) Nat. Biotech

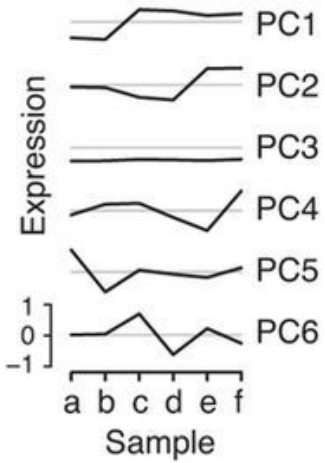
Dimensionality reduction by PCA

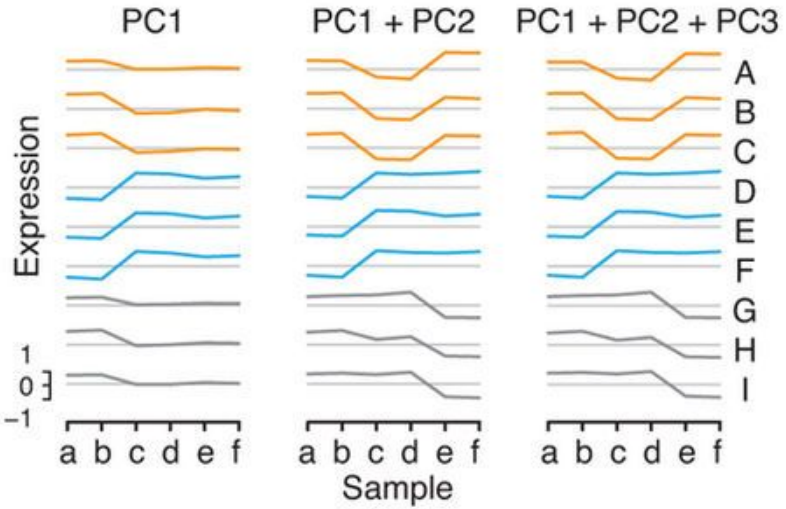


Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA

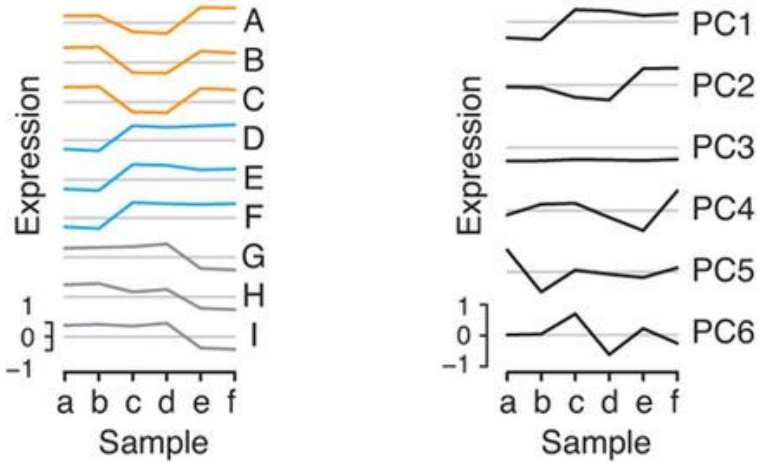


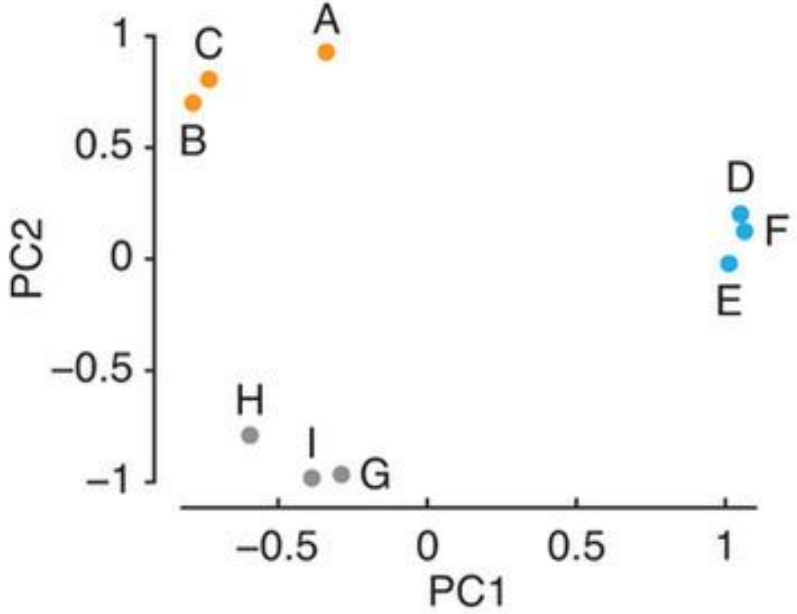




Lever, Krzywinski, & Altman (2017) Nat. Meth.

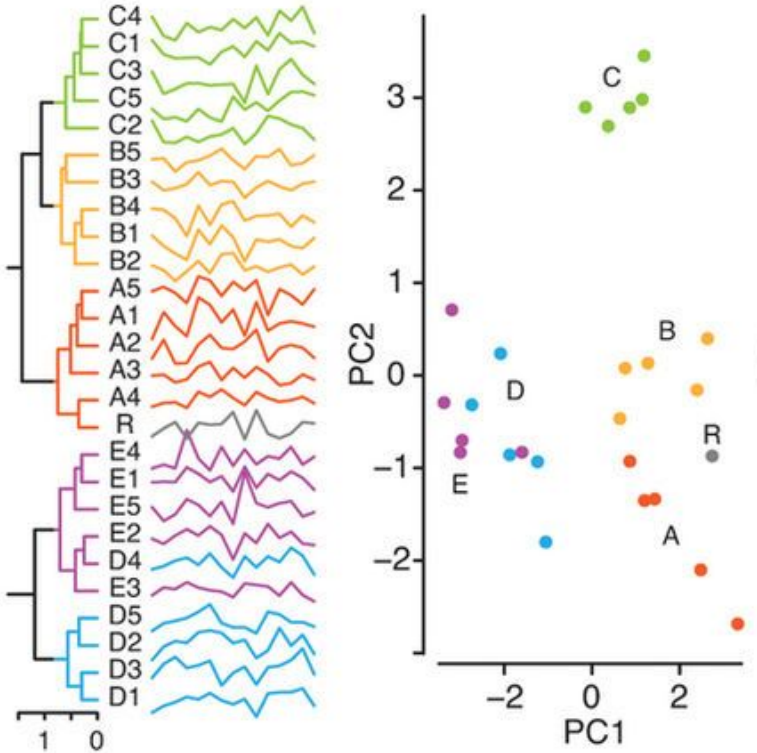
Dimensionality reduction by PCA





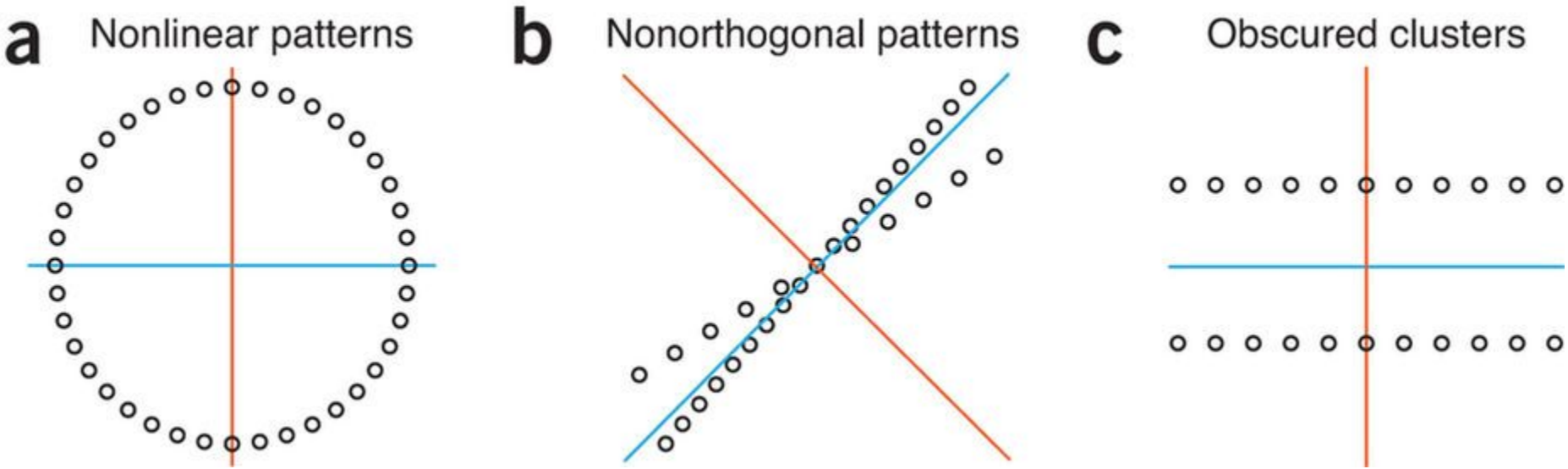
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



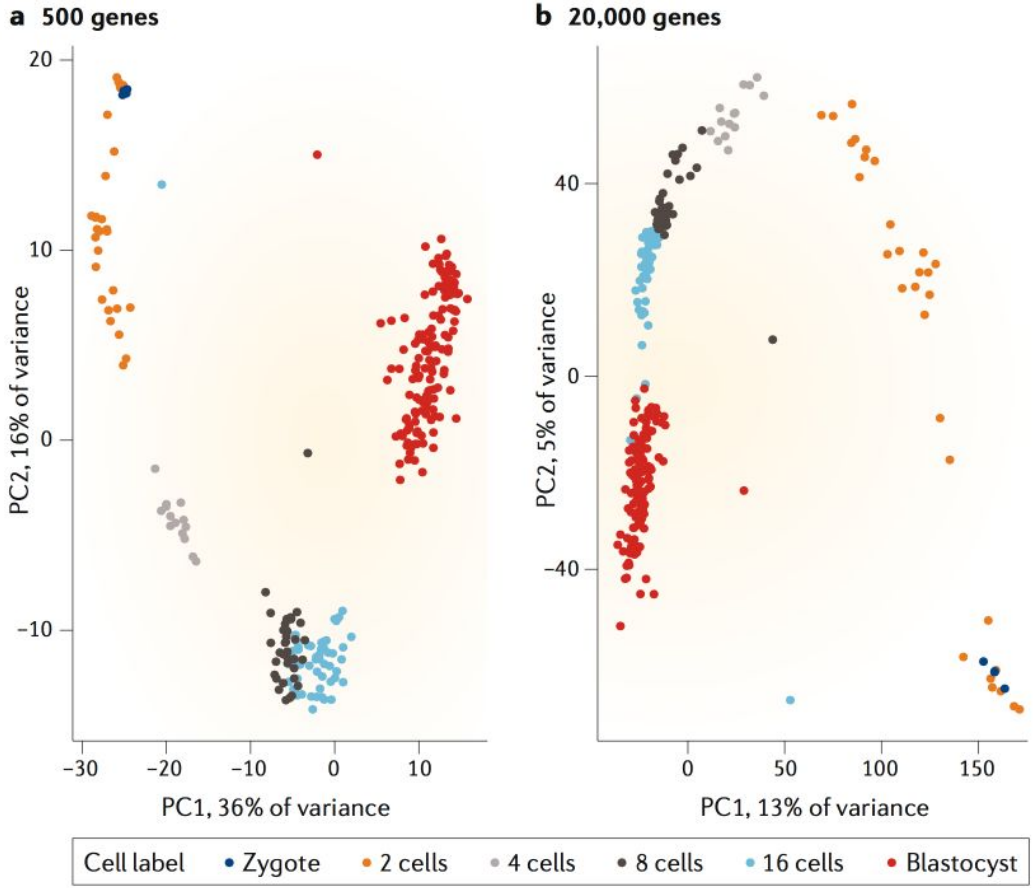
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Imputing missing values

scRNA-seq data has:

- **a high frequency of zero values**, often referred to as dropouts, and
- **high levels of noise** due to the low amounts of input RNA obtained from individual cells.

Zero values in scRNA-seq may arise due to:

- low experimental sensitivity, e.g. sequencing sampling noise, technical dropouts during library preparation, or
- biologically the gene is not expressed in the particular cell.

Imputing missing values

Zero values in scRNA-seq may arise due to:

- low experimental sensitivity, e.g. sequencing sampling noise, technical dropouts during library preparation, or
- because biologically the gene is not expressed in the particular cell.

Imputation is a common approach when dealing with sparse genomics data: predict missing values from the rest of the measured values.

One challenge when imputing expression values is to **distinguish true zeros from missing values**.

scRNA-seq data imputation methods use information internal to the dataset to be imputed.

- Some degree of circularity → false positive results when identifying marker genes, gene-gene correlations, or testing differential expression.

Imputing missing values

Many imputation methods:

- **SAVER, DrImpute & scImpute**: use models of the expected gene expression distribution to distinguish true biological zeros from zeros originating from technical noise.
    - Assume homogenous cell populations → identify clusters of similar cells to which an appropriate mixture model is fitted.
    - Values falling above a given probability threshold to originate from technical effects are subsequently imputed.
- **MAGIC & knn-smooth**: perform data smoothing.
    - Infer values of missing data + reduces noise present in observed values (using information from neighbouring data points).
    - Use each cell's k nearest neighbours either through the application of diffusion models or weighted sums respectively.

Imputing missing values

Many imputation methods:

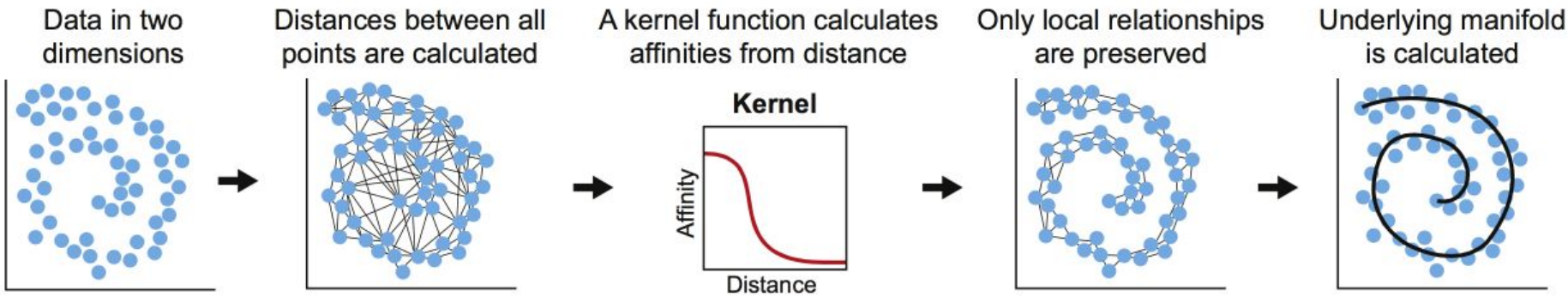
	Designed for single cell	Local or global	Beyesian method	Need other information	Imputation strategy
LLSimpute	N	local	N	No. of nearest genes	1
Low-rank	N	global	N	error tolerance δ	2
BISCUIT	Y	global	Y	dispersion parameter	1 and 2
scUnif	Y	global	Y	cell labels	2
MAGIC	Y	global	N	diffusion time	2
scImpute	Y	local	N	dropout rate cutoff	2
DrImpute	Y	local	N	cluster numbers	2
SAVER	Y	global	Y	size factor	1

Strategy 1 represents imputing dropout based on co-expressed or similar genes, while strategy 2 denotes imputing dropout by borrowing information from similar cells.

Zhang & Zhang (2017) bioRxiv

Imputing missing values

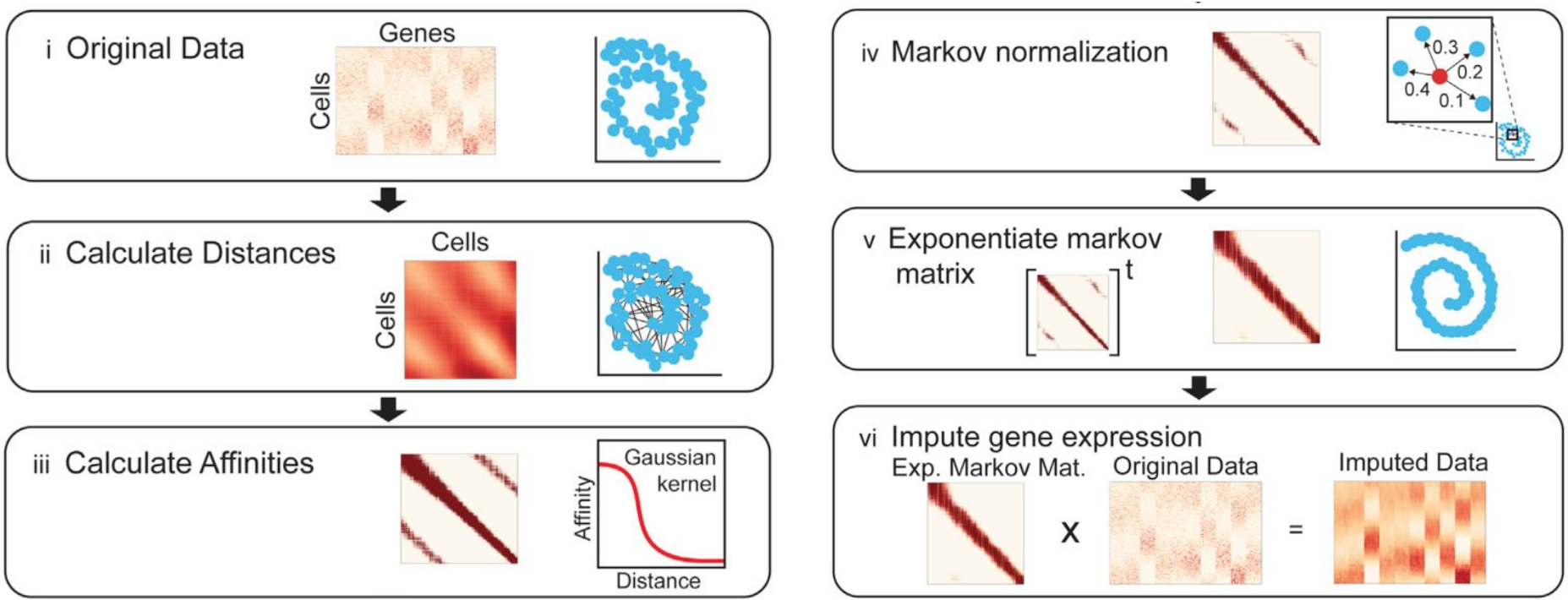
Manifold learning



Moon (2018) Curr. Opin. Sys. Biol.

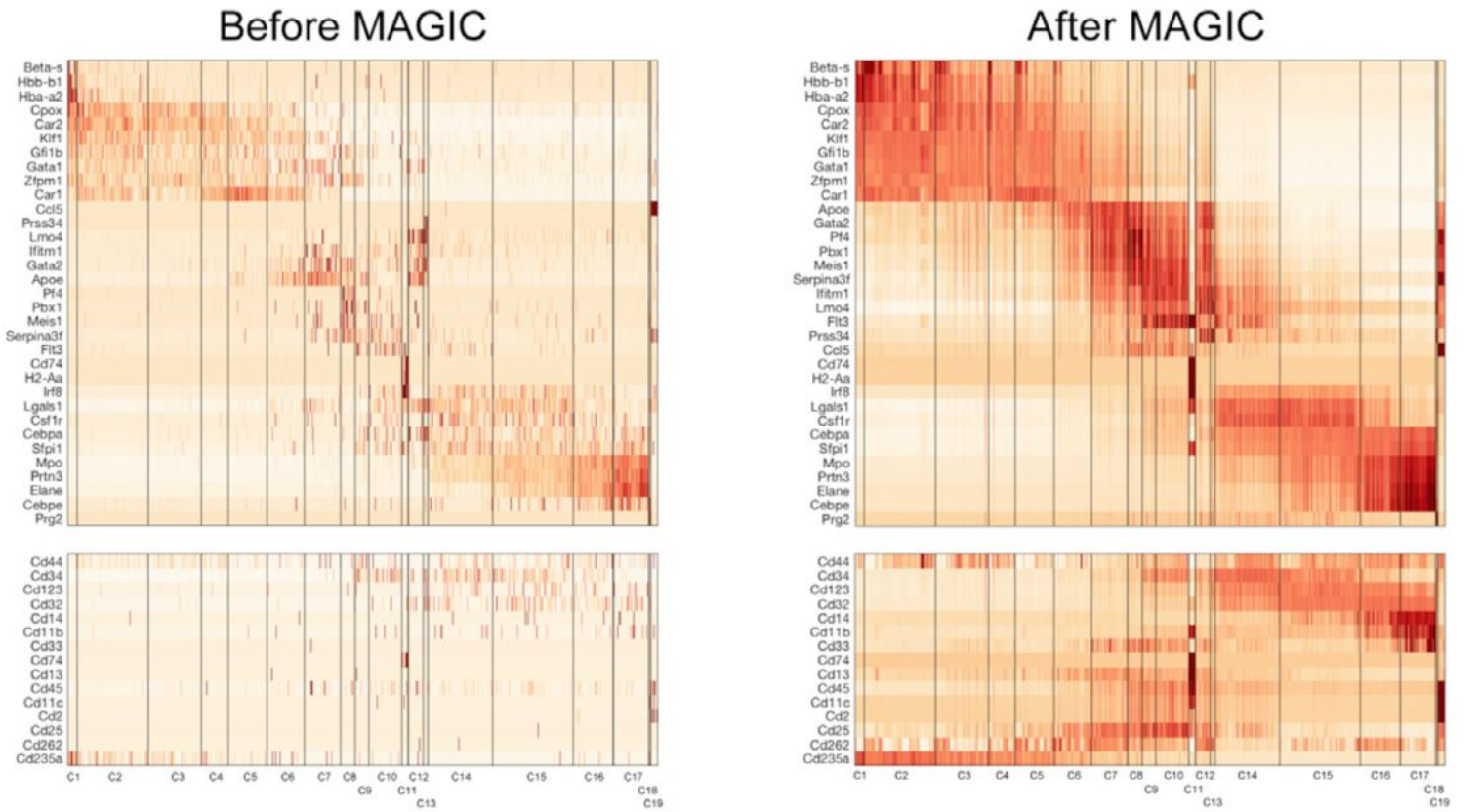
Imputing missing values

MAGIC



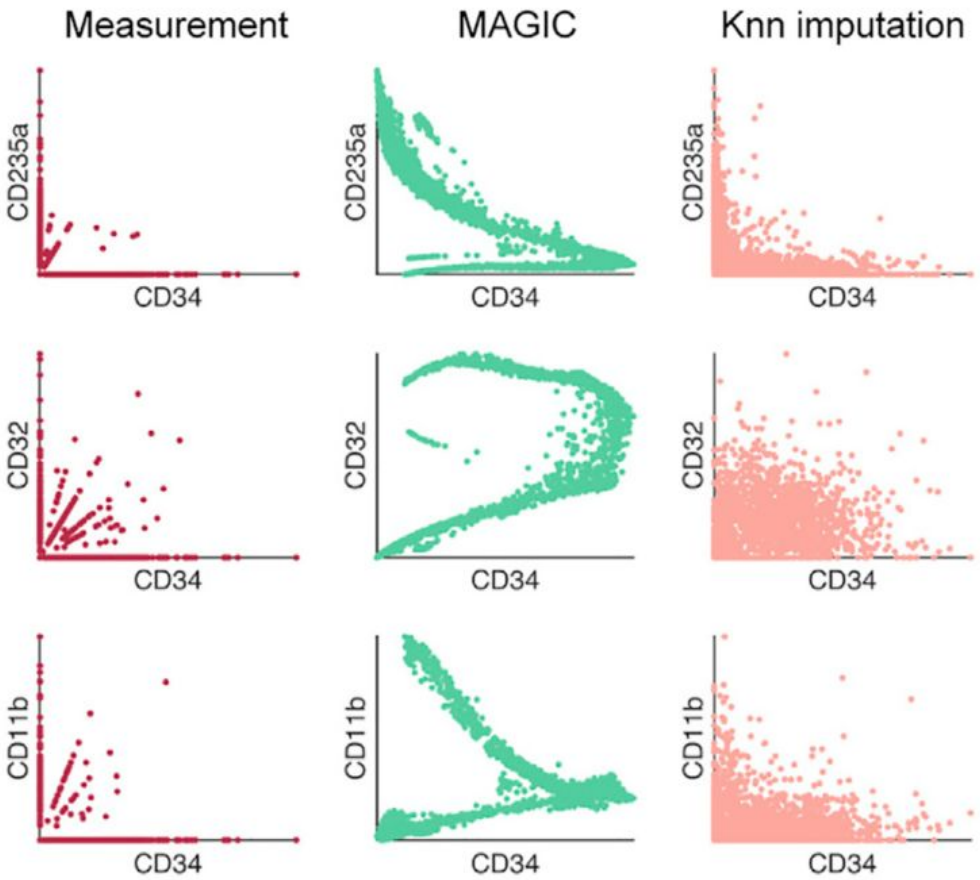
Vandijk (2018) Cell

Imputing missing values



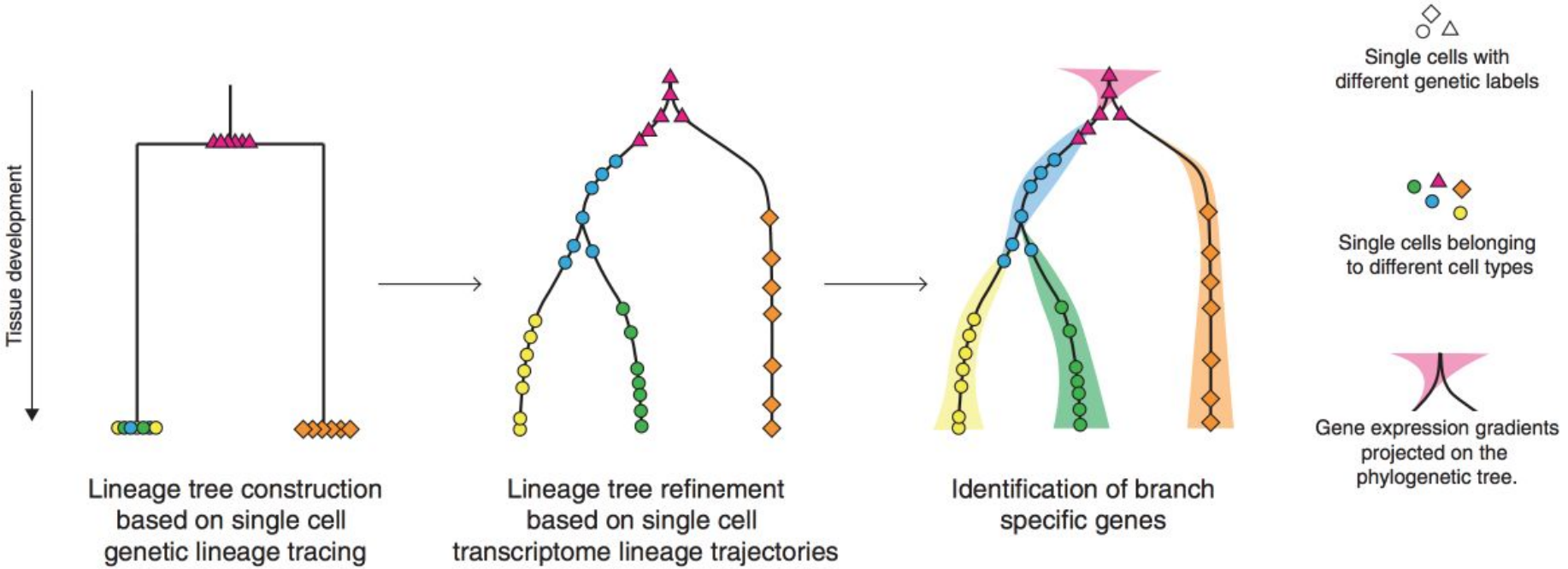
Vandijk (2018) Cell

Imputing missing values



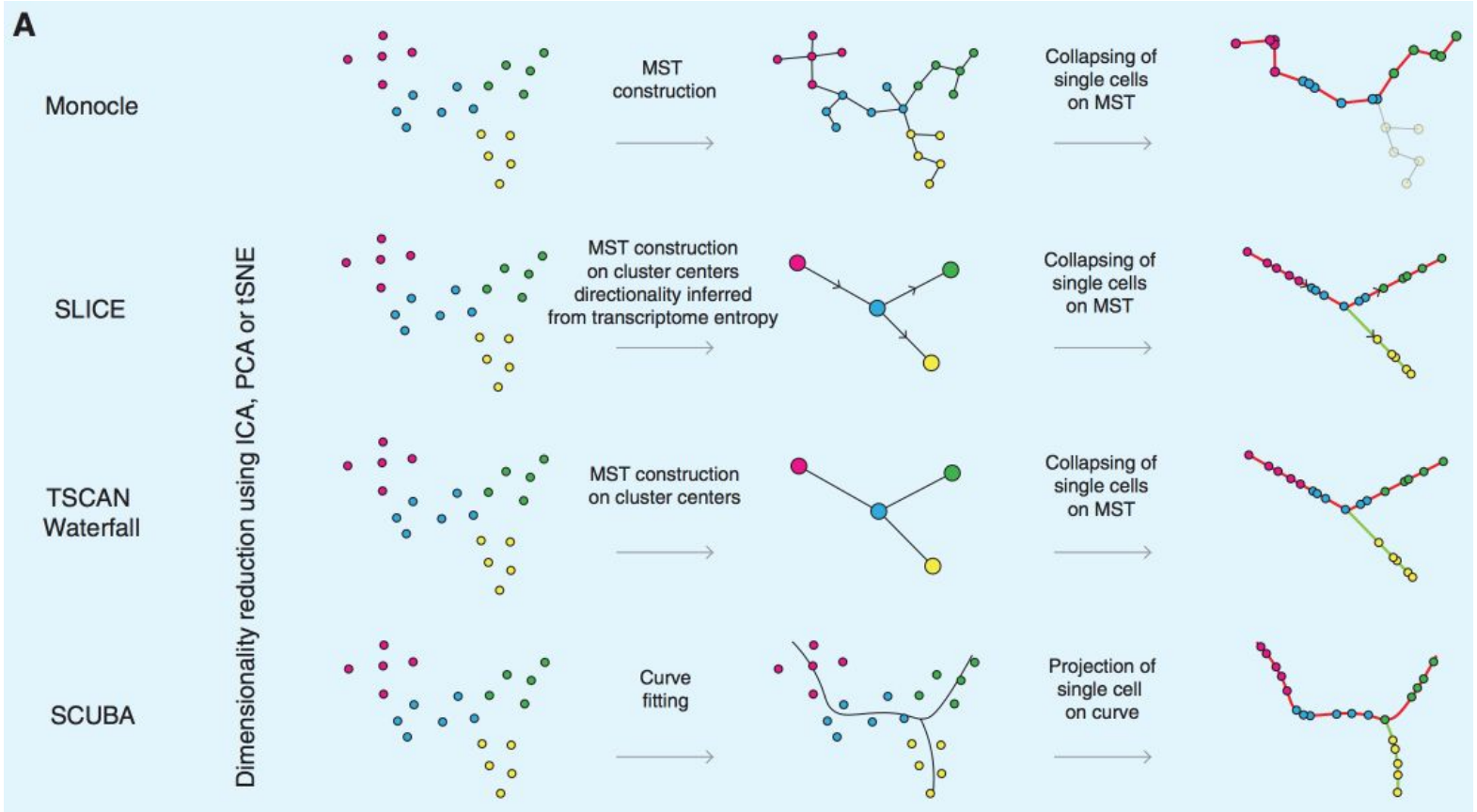
Vandijk (2018) Cell

Lineage tracing



Vandijk (2018) Cell

Lineage tracing



Kester (2018) Cell Stem Cell