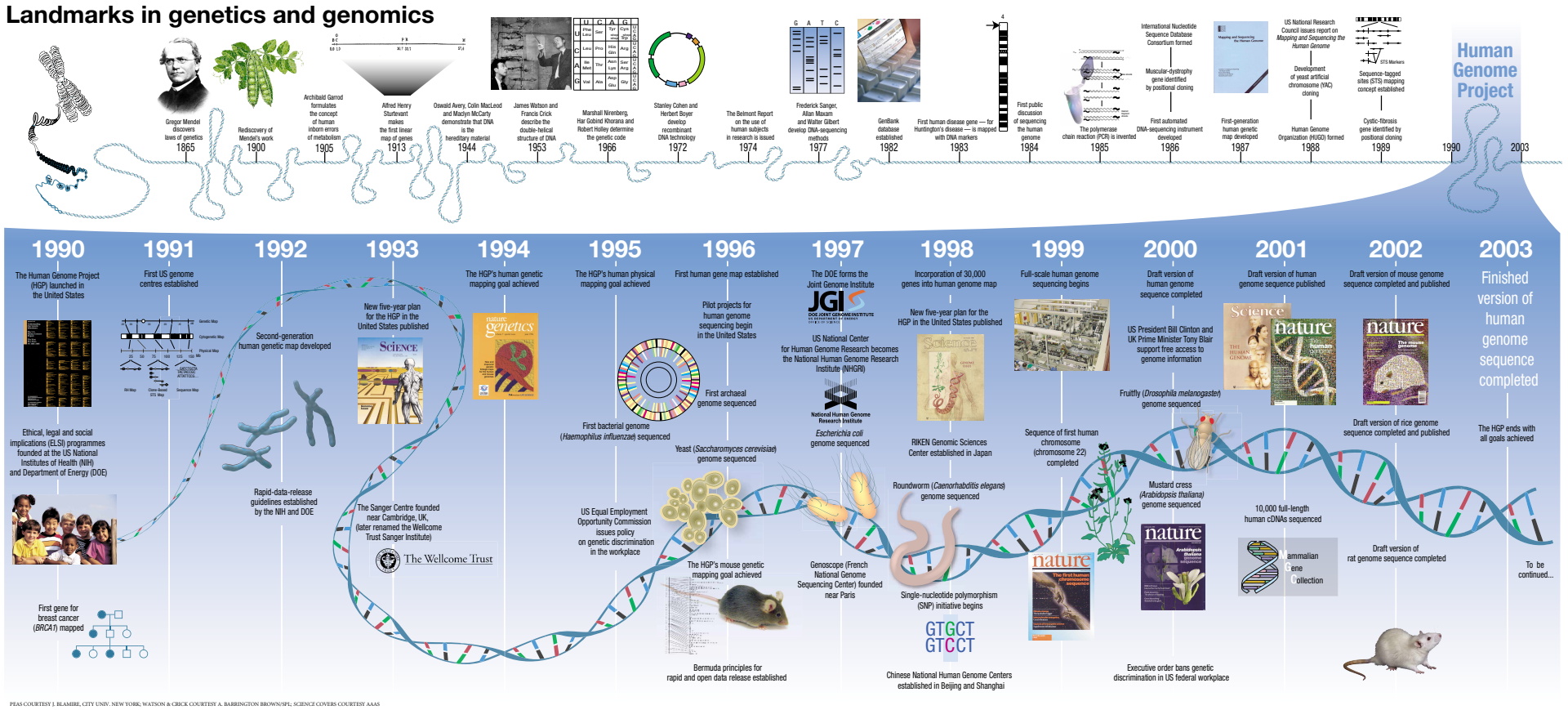
CCMI data driven modelling week

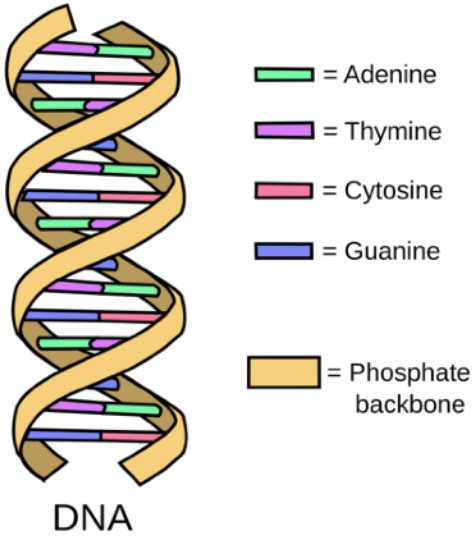
Data analysis and modelling: Case study of single-cell RNA-sequencing (transcription) data

Vahid Shahrezaei
Department of Mathematics
IMPERIAL

Development in genomics: Human Genome Project
Relating genotype to phenotype is not easy!



https://www.genome.gov/human-genome-project



Genome size
(A species comparison)

1 billion base pairs

100 thousand base pairs

1 thousand base pairs

10 million base pairs

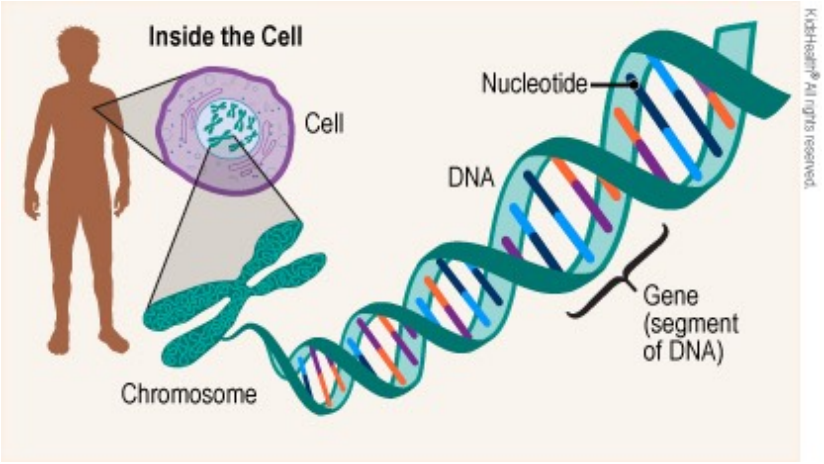
500 million base pairs

10 billion base pairs

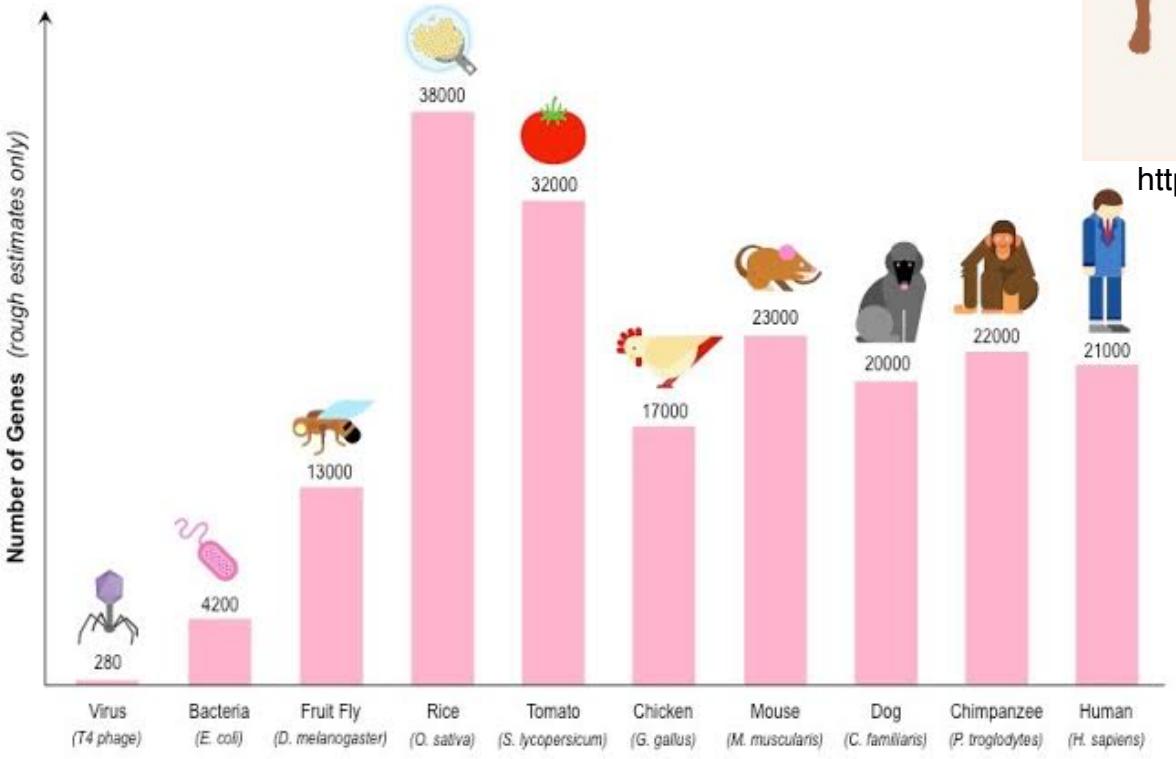
150 billion base pairs

Platypus 2.4 Gb
Round-eared bat 1.6 Gb
King Cobra 1.59 Gb
Ostrich 1.3 Gb
Cane Toad 2.55 Gb
Cheetah 2.5 Gb
Humpback Whale 2.7 Gb
Zebra fish 1.6 Gb
Breviolum minutum (dinoflagellate algae) 1.5 Gb
Quinoa 1.5 Gb
Soy Bean 1.15 Gb
Tobacco Plant 2.46 Gb
Northern Spotted Owl 1.26 Gb
Humminbird 0.91 Gb
Orangutan 3.2 Gb
Human 3.2 Gb
Wolf 2.75 Gb
Chicken 1.05 Gb
Tomato 900 Mb
Sea snail 916 Mb
Sea Urchin 814 Mb
Pyrococcus furiosus (Archaea) 1.9 Mb
Scorpion 880 Mb
Bluebottle (Man-of-war) 3.25 Gb
Dechloromonas aromatic (Bacteria) 4.5 Mb
Influenza A (virus) 14 Kb
Candidatus Nasuia (Bacteria) 112 Kb
Box jellyfish 753 Mb
Koala 3.4 Gb
Encephalitozoon intestinalis (Fungi) 2.25 Mb
Sugar beet 714 Mb
Sunflower 3.6 Gb
Pseudomonas putida (Bacteria) 6.2 Mb
African Elephant 4.01 Gb
Streptomyces avermitilis (Bacteria) 9 Mb
Hepatitis D (virus) 1.7 Kb
Aardvark 4.13 Gb
Brown Algae Saccharina japonica 580 Mb
Eucalyptus (Rose Gum) 640 Mb
Great white shark 4.63 Gb
Yeast (Fungi) 12.16 Mb
Sea anemone (Anthopleura japonica) 470 Mb
Golden Mole 6.18 Gb
Lyric cicada 6.86 Gb
Sorangium cellulosum (Bacteria) 14.8 Mb
Silkworm 530 Mb
Red vizcacha rat 8.2 Gb
Ladybird 420 Mb
Coral (Acropora digitifera) 420 Mb
Green Algae (Chlorella vulgaris) 37.3 Mb
Fly 132 Mb
Honey bee 231 Mb
Starfish 16.19 Mb
Maidenhair Tree (Ginko biloba) 11 Gb
Puffer Fish 365 Mb
Roundworm 95 Mb
Monarch Butterfly 273 Mb
Genlisea aurea (a type of orchid) 63.6 Mb
Grasshopper 18 Gb
Salamander 120 Gb
Mammals
Birds
Plants, Algae
Insects, Molluscs, Nematode
Fish, Reptiles
Amphibians, Cnidaria, Echinodermata
Fungi, Bacteria, Archae, Viruses
Marbled lungfish 130 Gb
Paris Japonica lily 149 Gb (149 billion base pairs)

An organism's complexity, its genome size and its total number of (protein-coding) genes are only loosely correlated!

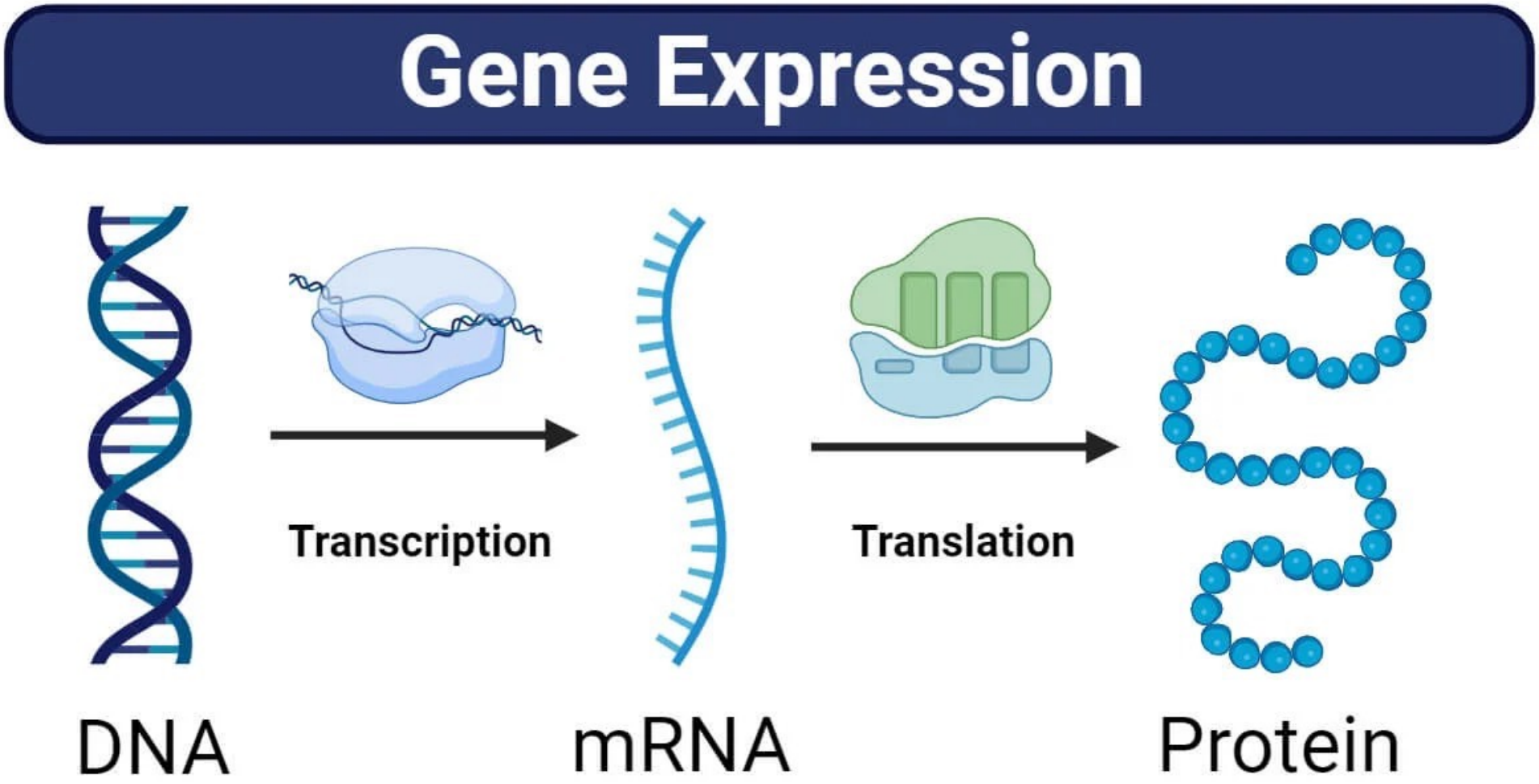


https://kidshealth.org/en/parents/about-genetics.html



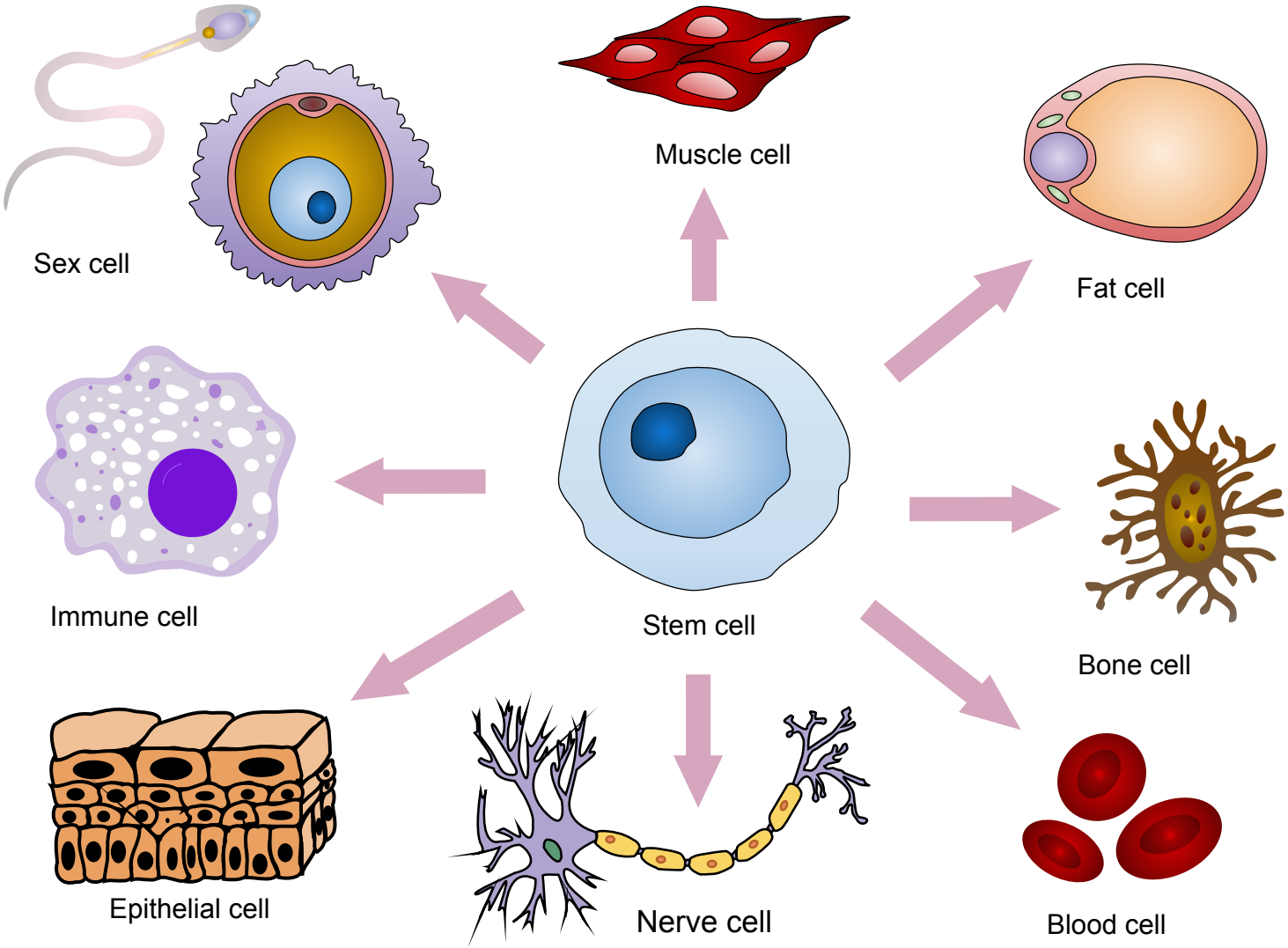
https://old-ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/chromosome-number.html

Central dogma of Molecular Biology:



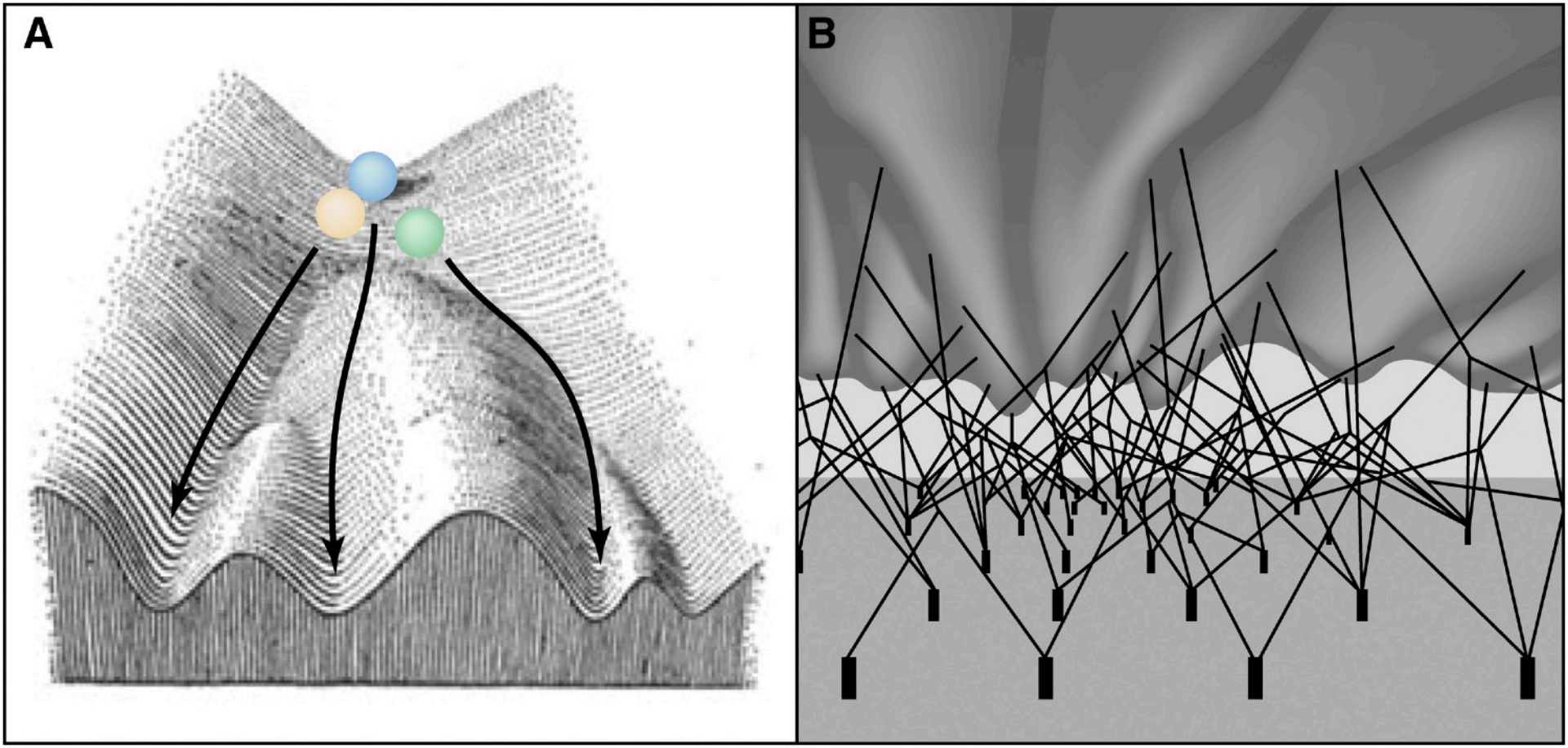
https://microbenotes.com/gene-expression/

All the cell-types contain the same genome: Each cell-type expresses a different subset of genes



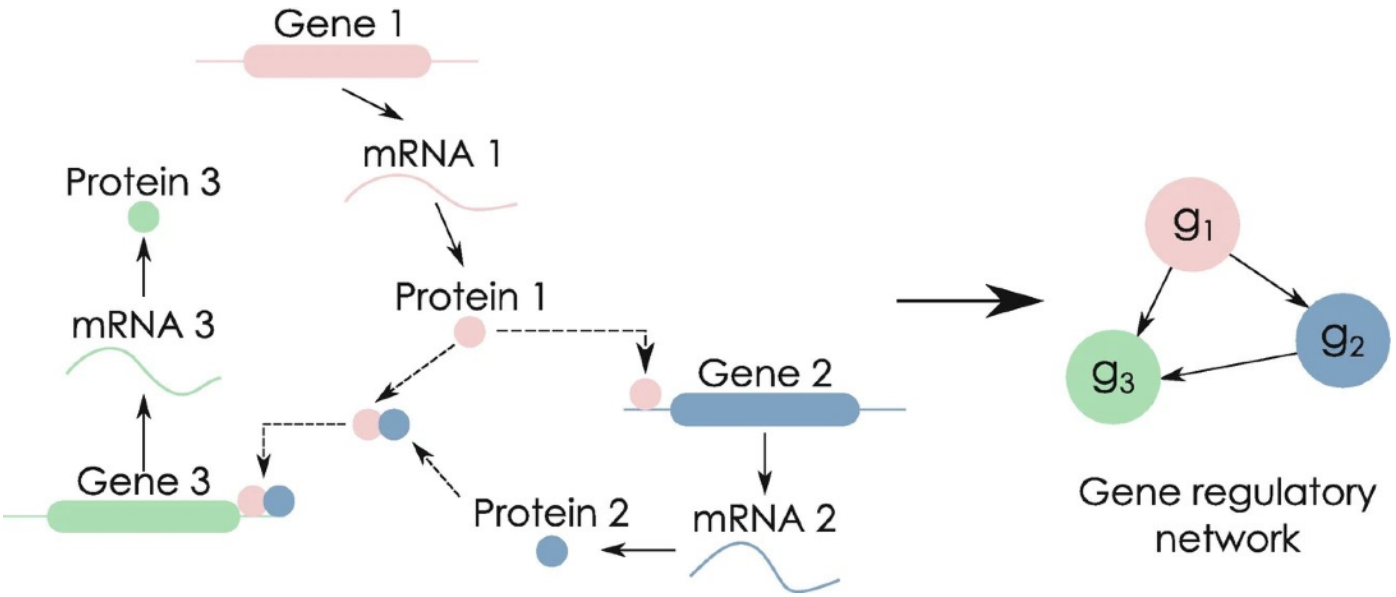
Human genome has about 20K-30K genes. Image from Wikipedia

Waddington development landscape and cellular differentiation



Adapted from The Strategy of the Genes (Waddington, 1957).
From Rajapogal & Stanger Developmental Cell 2016

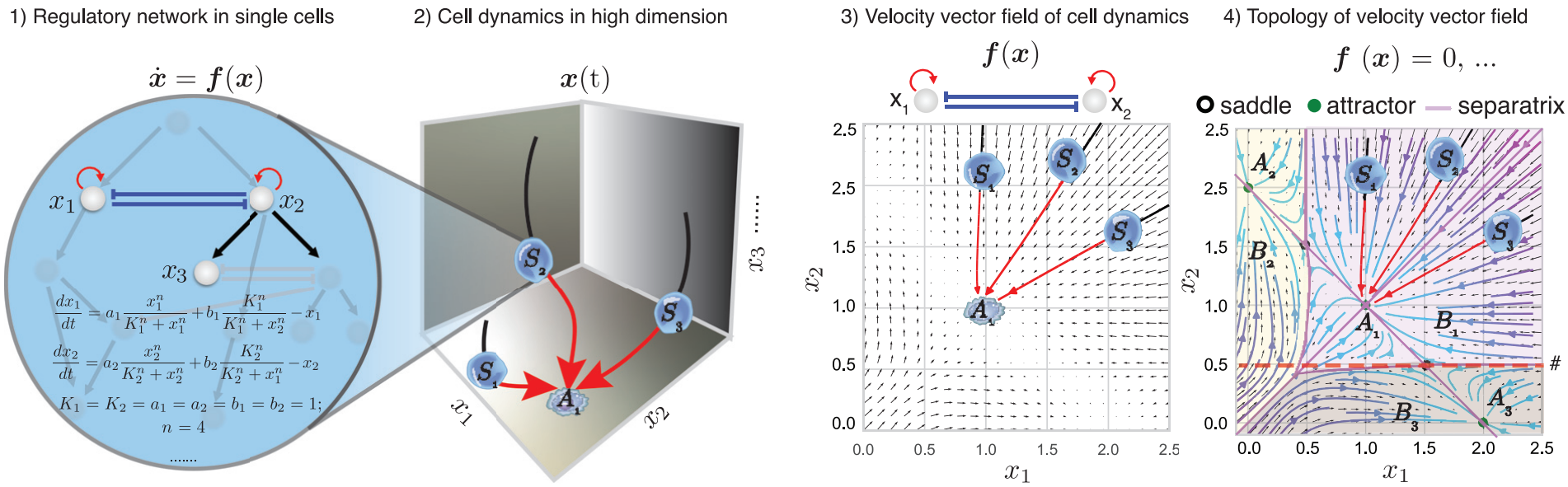
Gene Regulatory Networks (GRNs)



Huynh-Thu & Sanguinetti Gene Regulatory Network Inference, Methods in Molecular Biology Springer Protocols 2019

https://doi.org/10.1007/978-1-4939-8882-2

GRNs can be modelled as (stochastic, nonlinear) dynamical systems and cell states can be seen as (stable) fixed points of these systems



Qiu et al Cell 185, 650, 2022 (Dynamo)

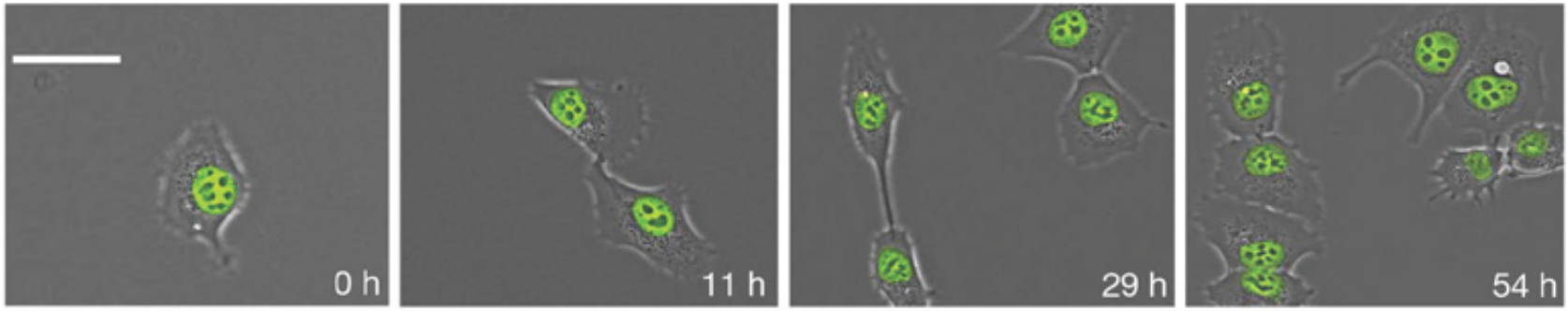
Stochastic influences on phenotype

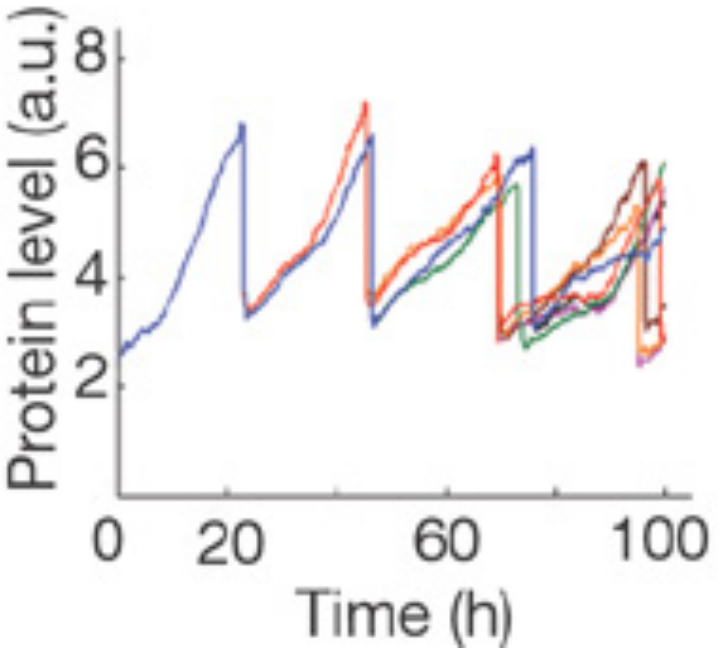
Fingerprints of identical twins

Cc the first clone cat & Rainbow, Cc's genetic mother

Raser and O'shea, Science 309, 5743 (2006)

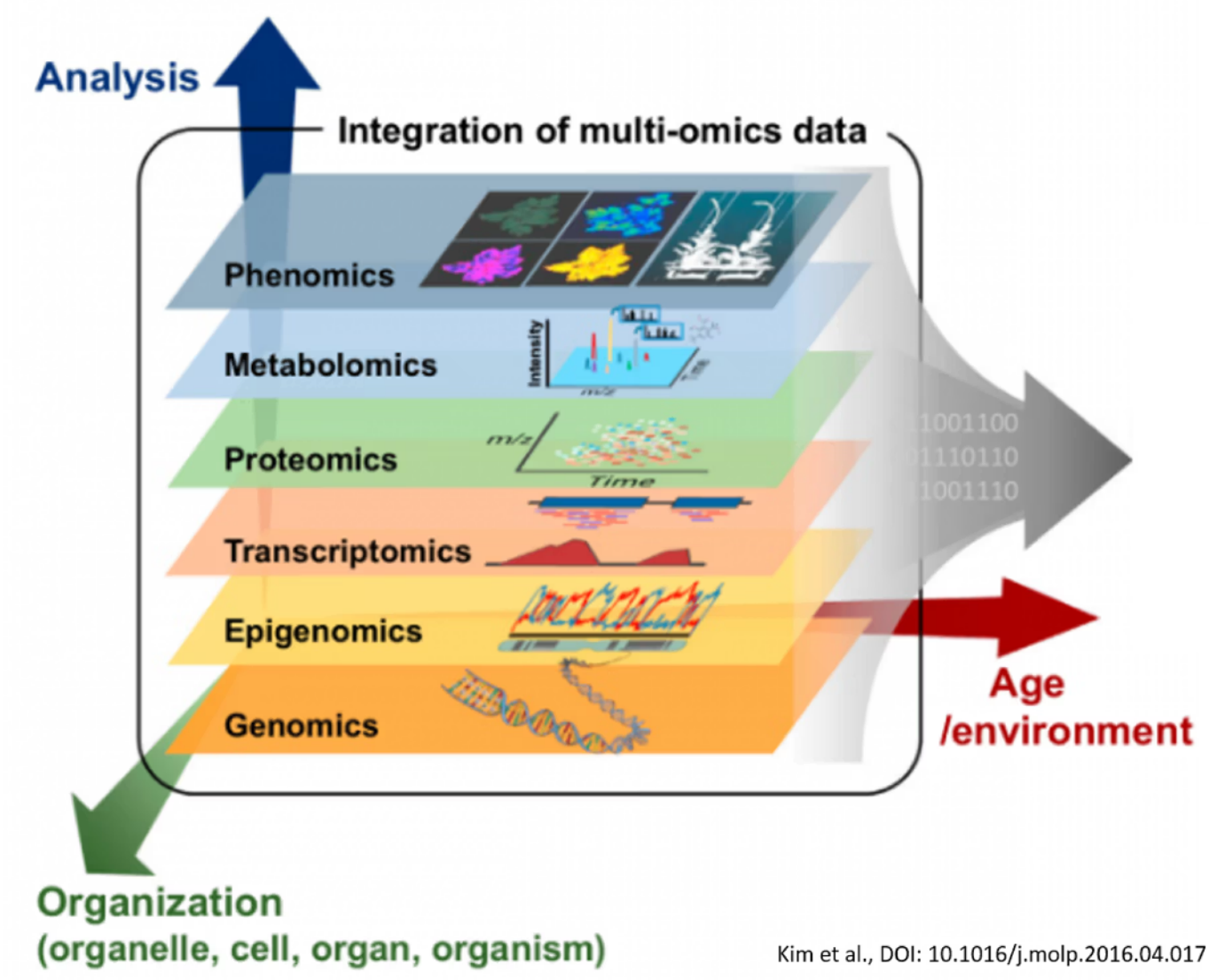
Gene expression variability in single cells
Fluorescent protein reporters





Sigal *et al.* Nature 444, 643 (2006)

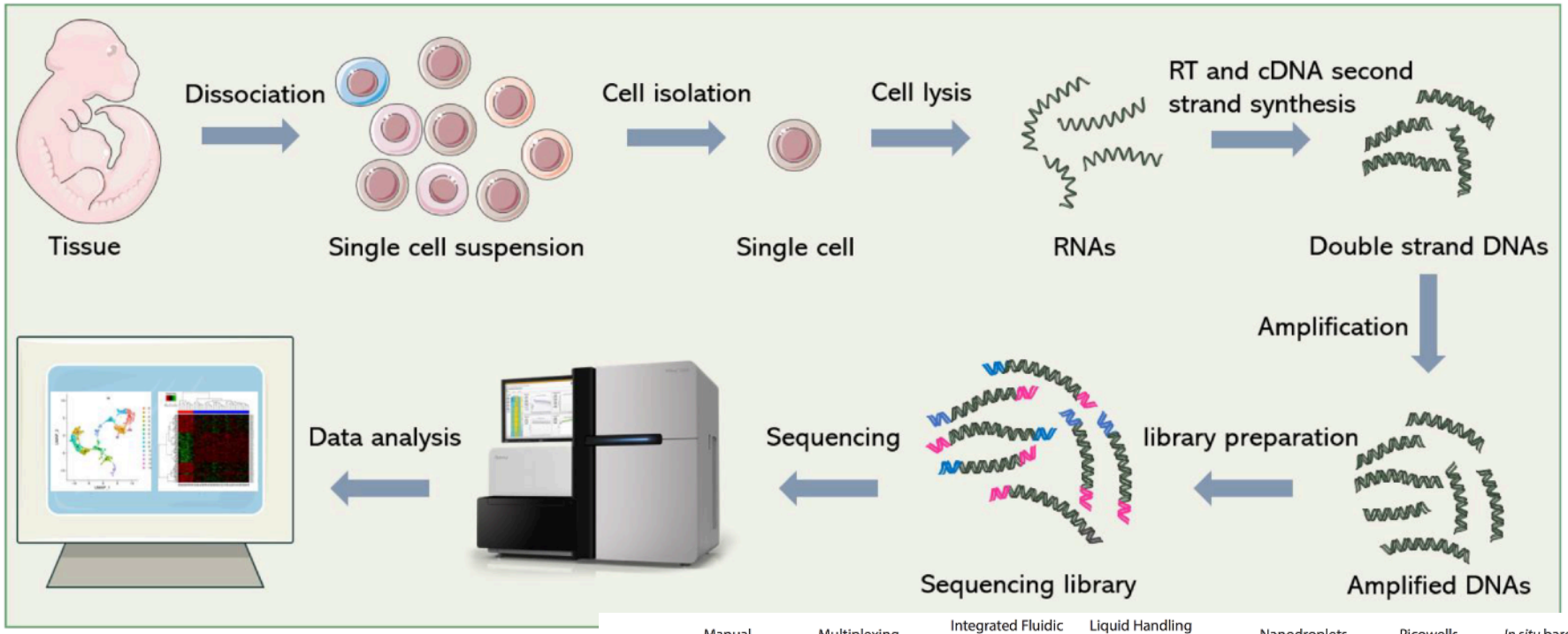
High-throughput global biological data: Different types of omics data



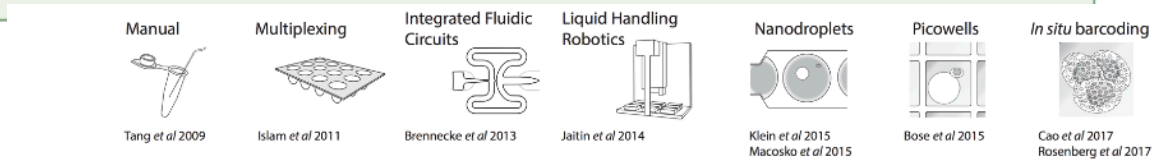
Kim et al., DOI: 10.1016/j.molp.2016.04.017

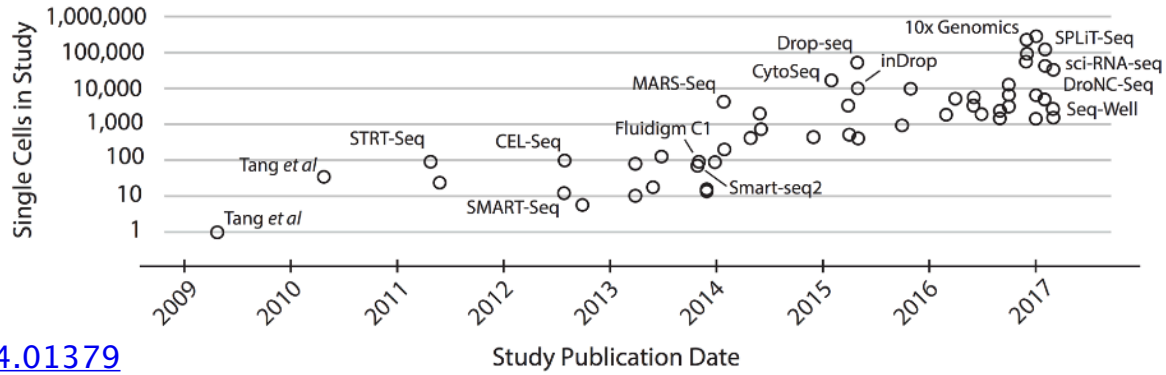
Development of single-cell RNA-sequencing

A method for global high-throughput single cell transcriptomics



https://encyclopedia.pub/entry/24618

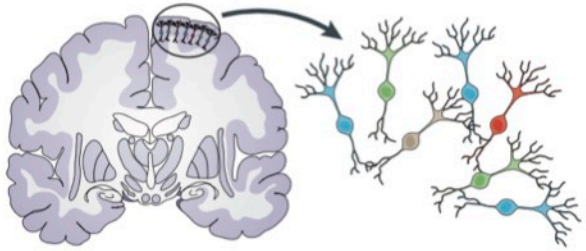




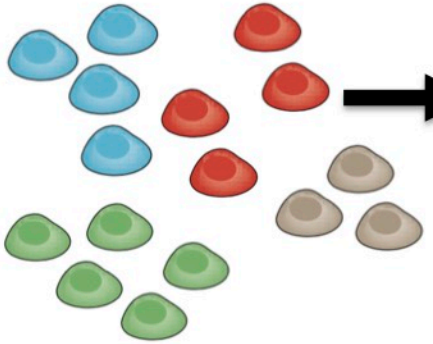
https://doi.org/10.48550/arXiv.1704.01379

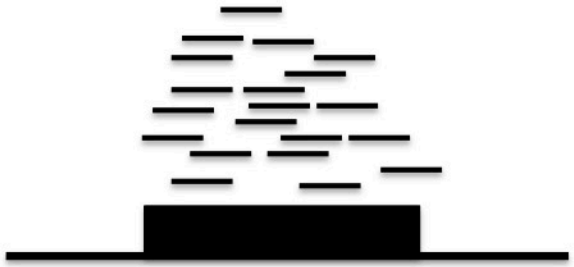
Single-cell RNA-Seq (scRNA-Seq)

Tissue (e.g. tumor)



Isolate and sequence individual cells



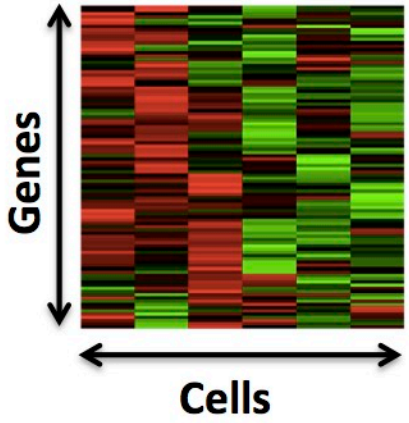


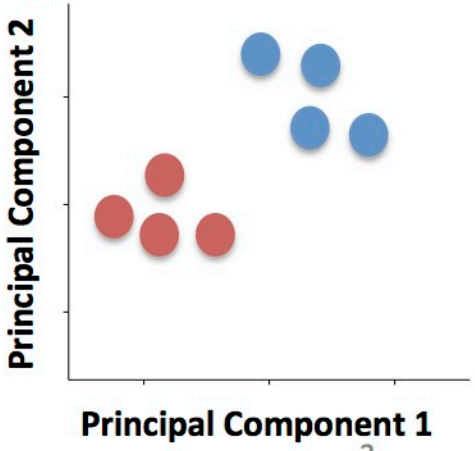
Gene 1
Cell 1

Read Counts

	Cell 1	Cell 2	…
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
…			

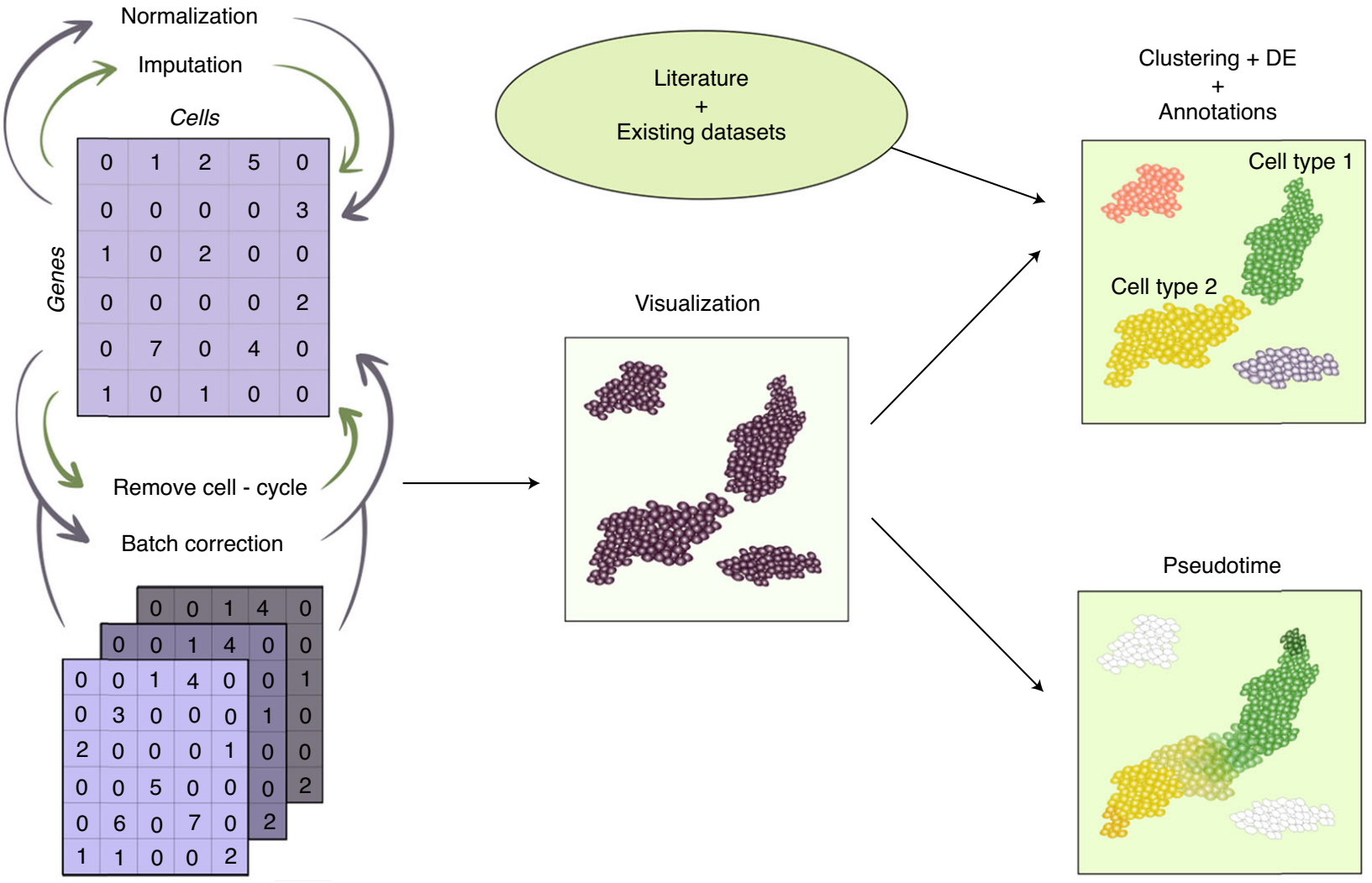
Compare gene expression profiles of single cells





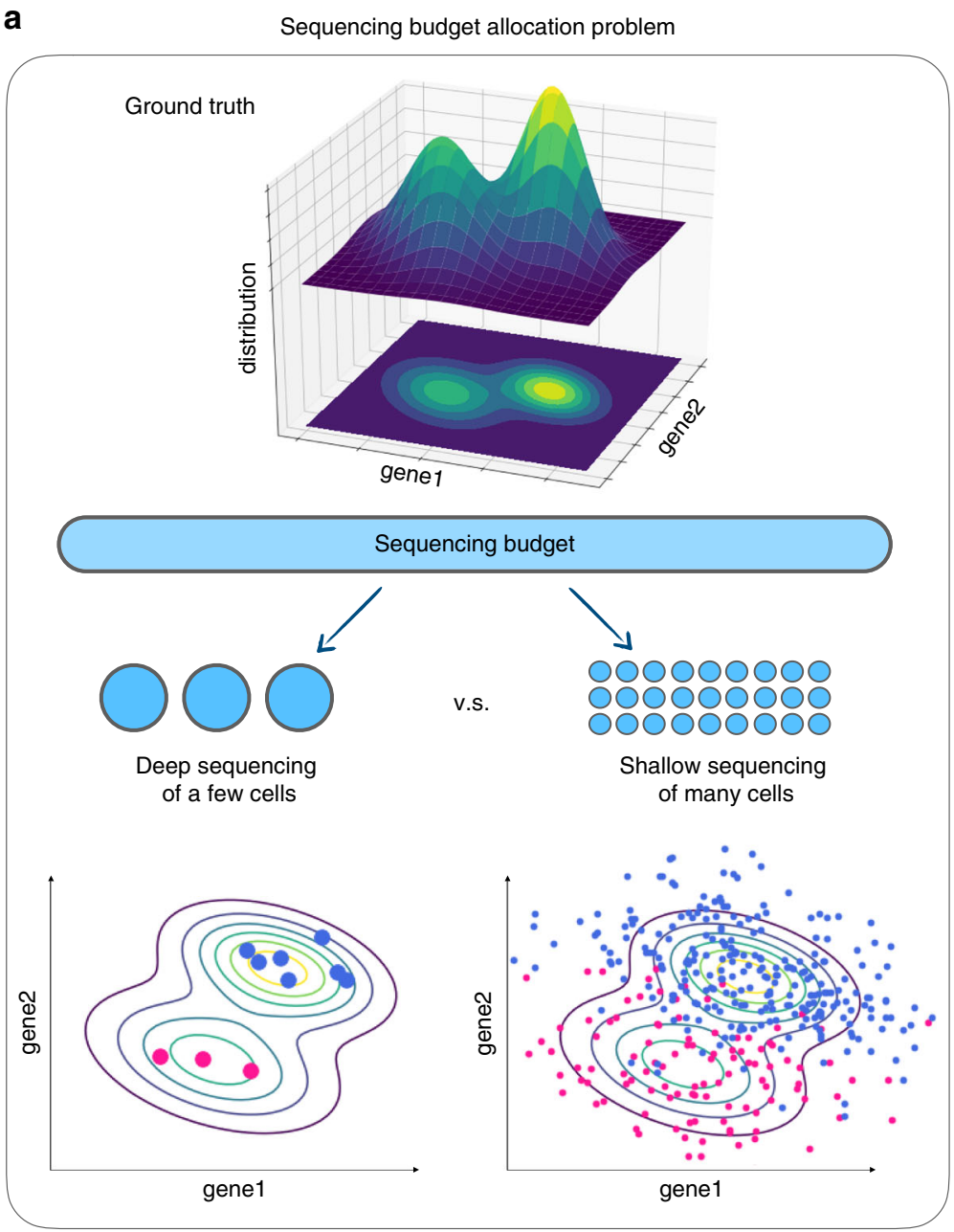
https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

Mathematical, statistical and computational analysis of scRNA-Seq data

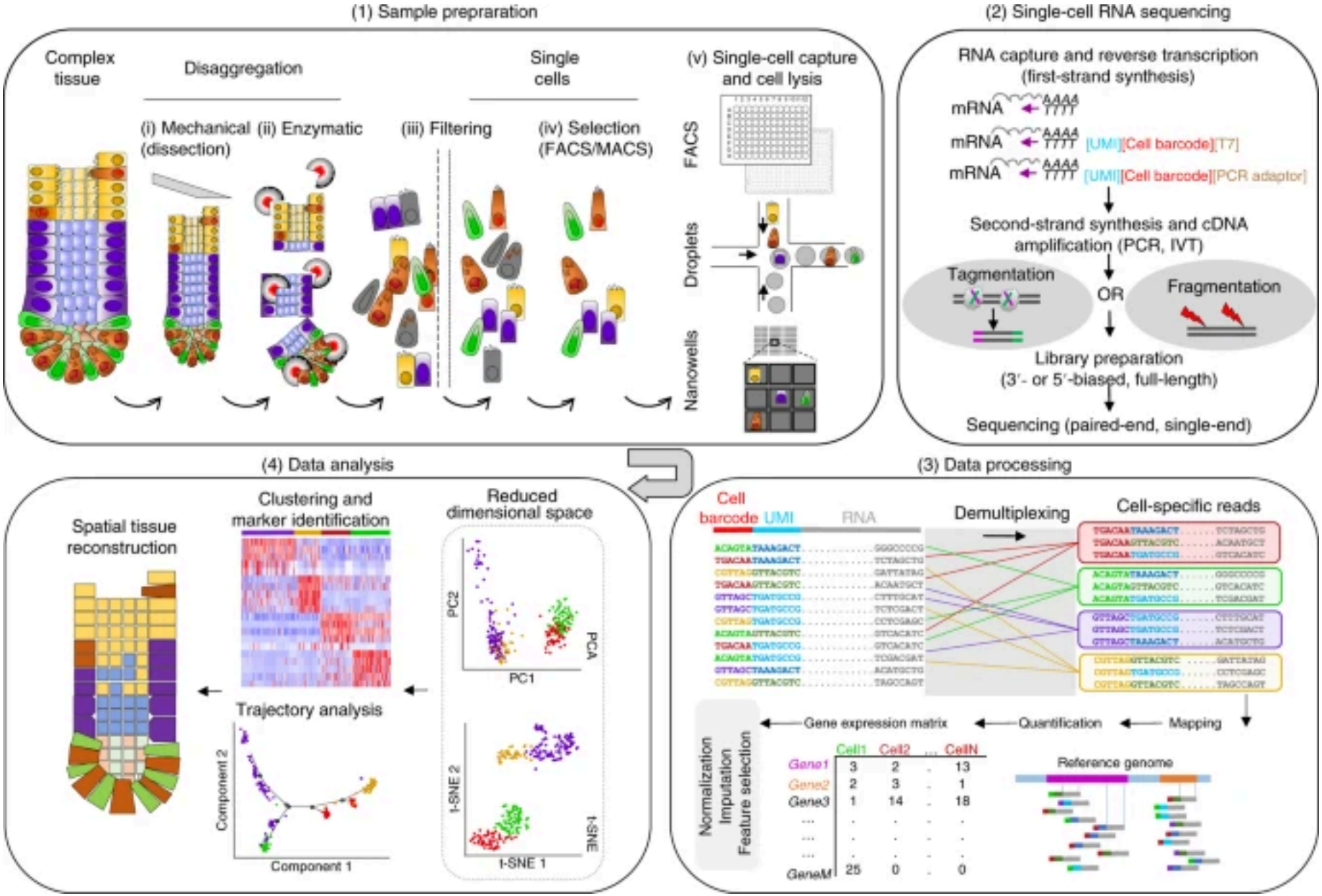


https://doi.org/10.1038/s41596-020-00409-w

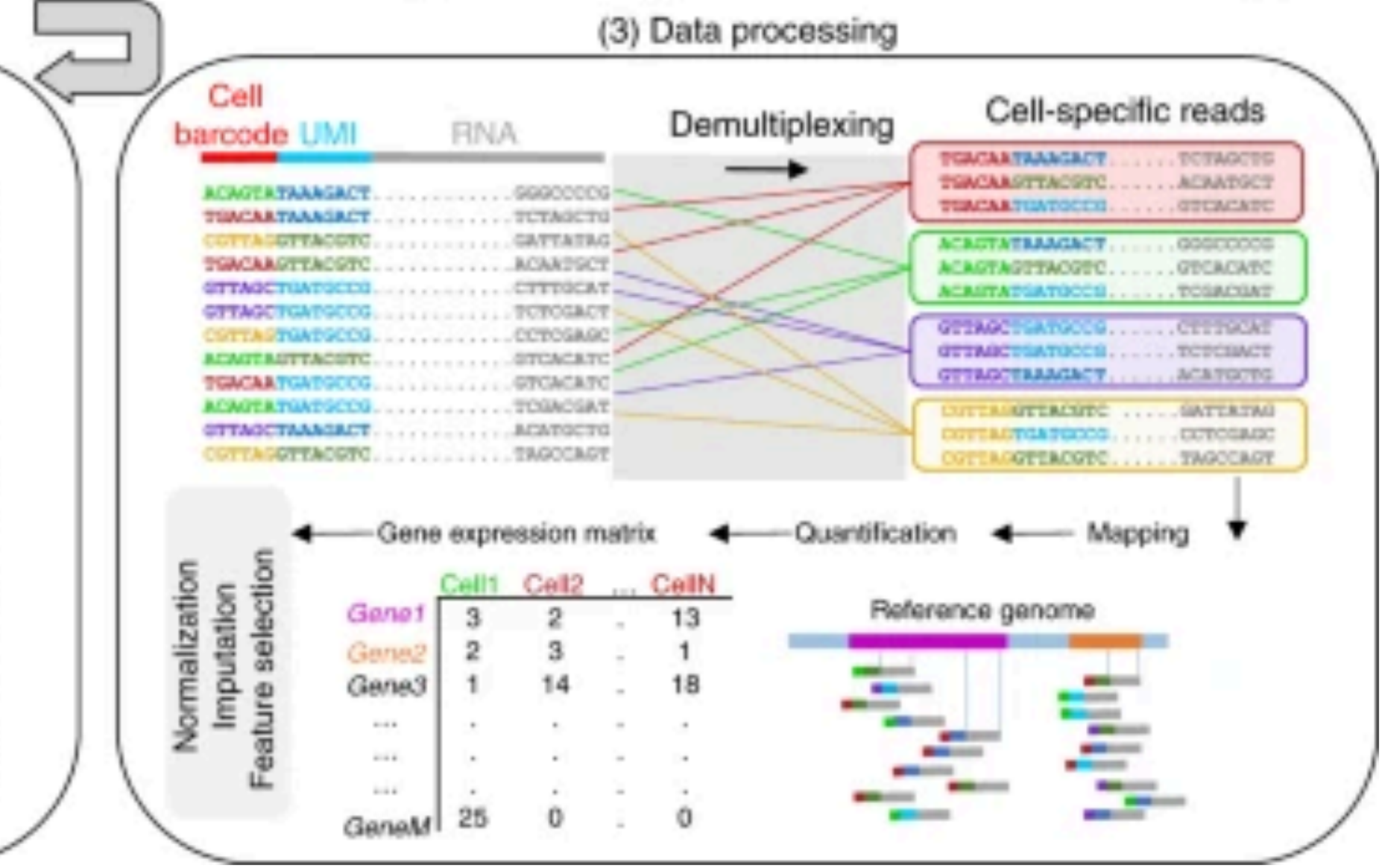
1. Experimental design



2. From raw reads to a count matrix



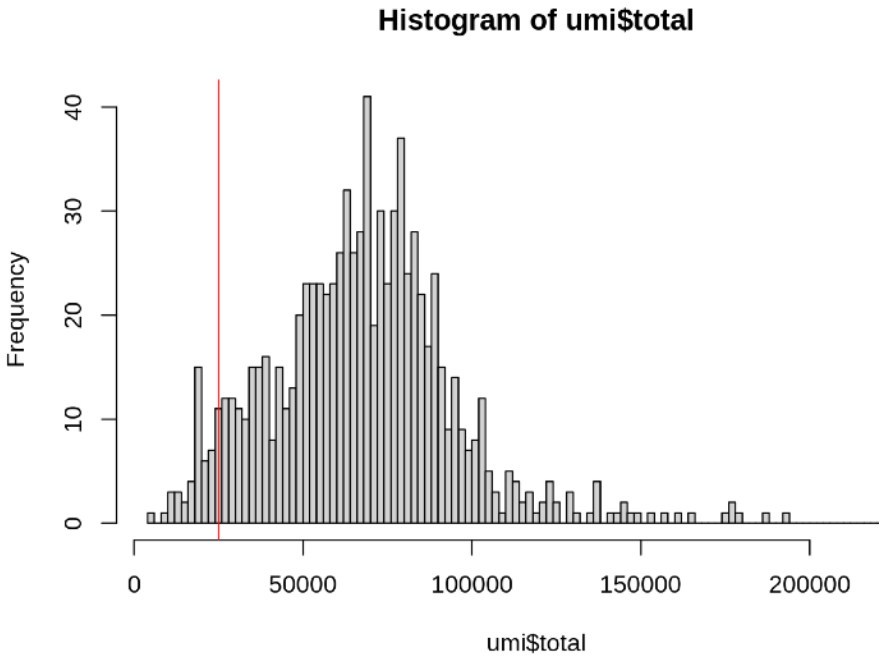
https://doi.org/10.1038/s41596-018-0073-y

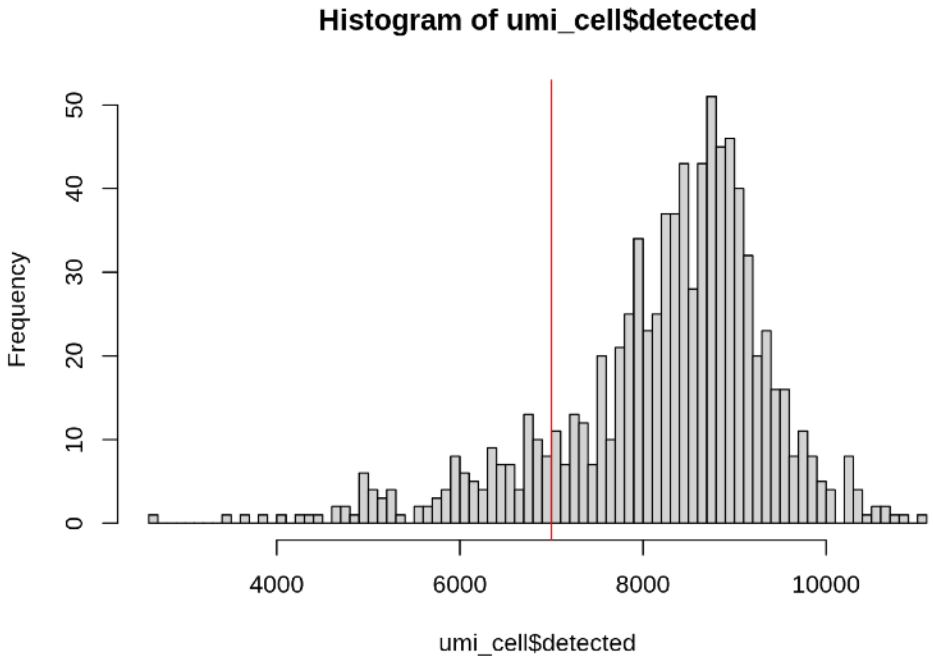


Sequence alignment to reference genome and demultiplexing unique molecular identifiers (UMIs) and cell barcodes to remove amplification biases and map reads to specific genes and cells.

3. Basic quality control

Filtering cells based on total number of transcripts per cell and detected genes

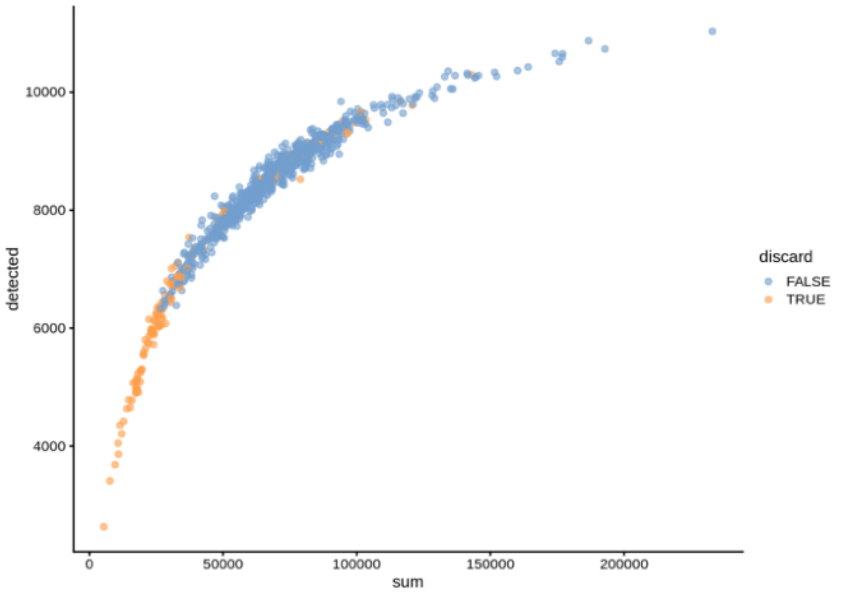


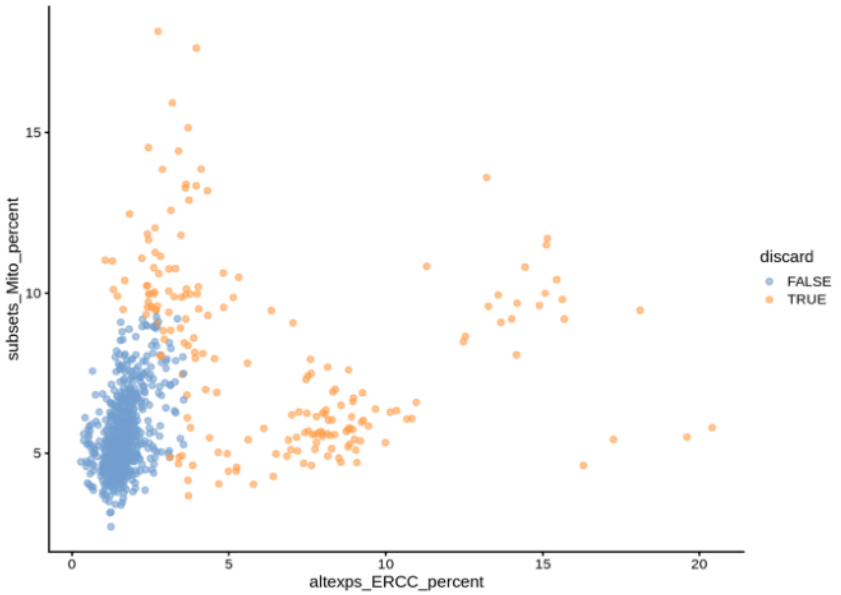


https://www.singlecellcourse.org/

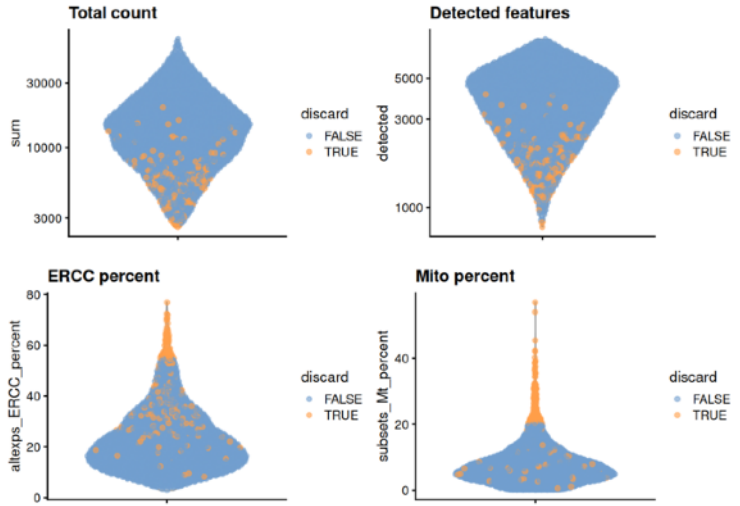
3. Basic quality control

Additional information such as percentage of mitochondrial genes and use of control spike-in (ERCCs) are common for filtering low-quality cells





https://www.singlecellcourse.org/

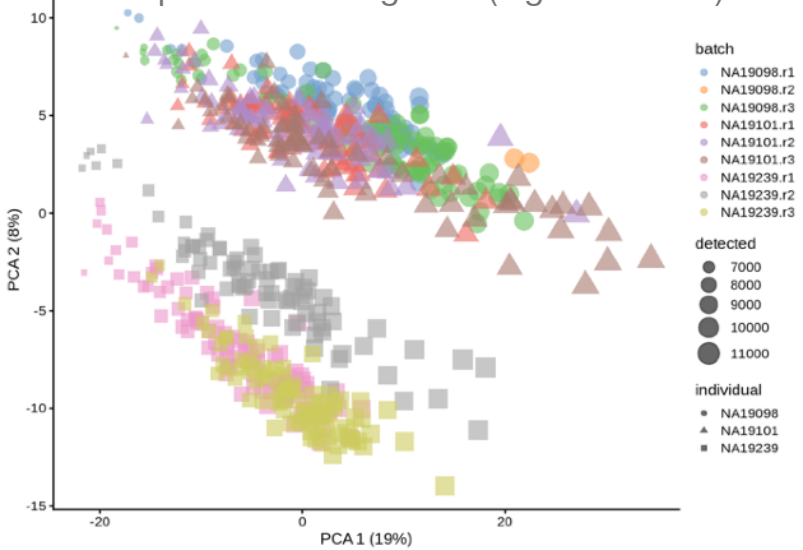


OSCA Workflows chapter 2

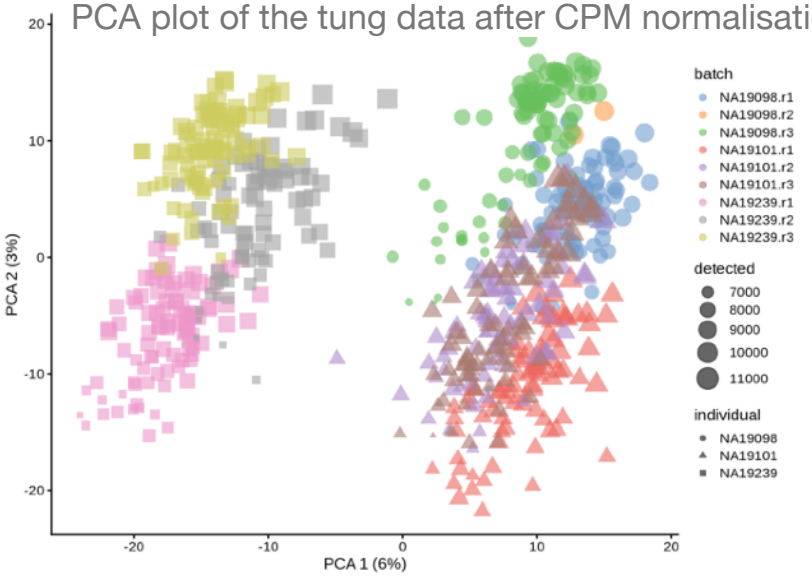
4. Normalisation

The simplest approach would be scaling normalisation by cell-specific size factor such as total transcripts per cell.

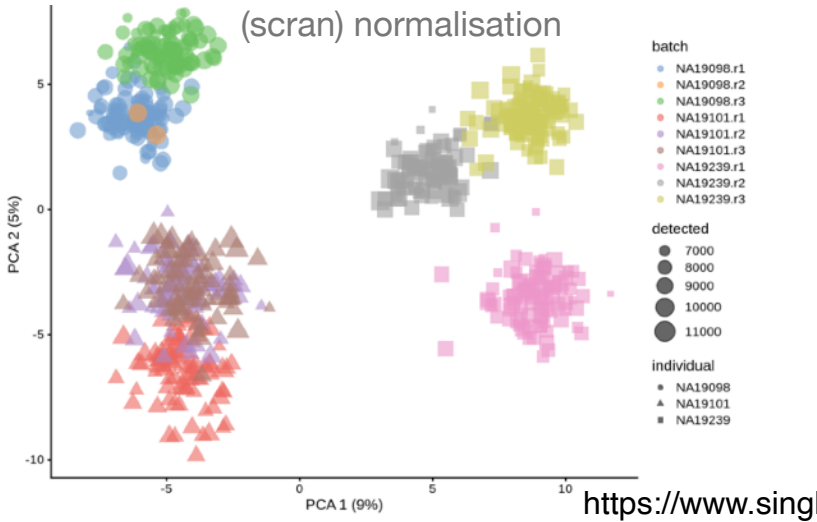
PCA plot of the Tung data (logcounts raw)



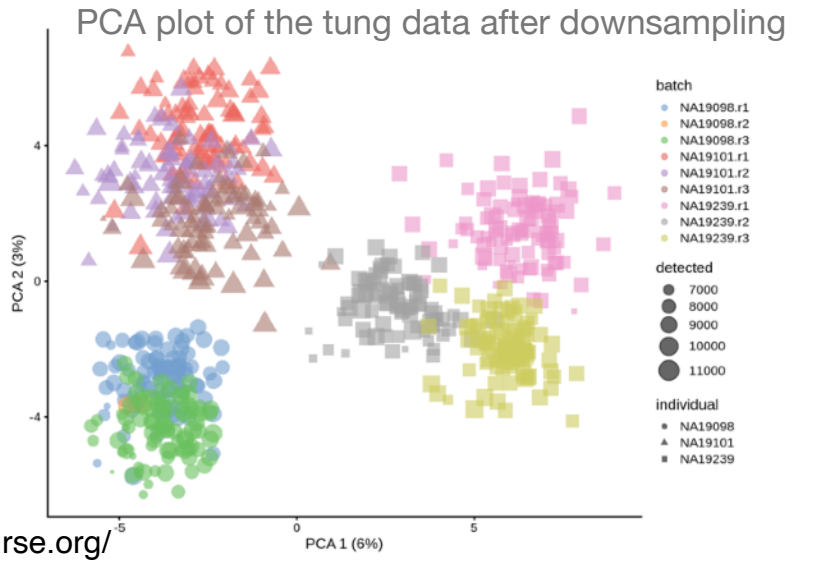
PCA plot of the tung data after CPM normalisation



PCA plot of the tung data after deconvolution-based (scran) normalisation



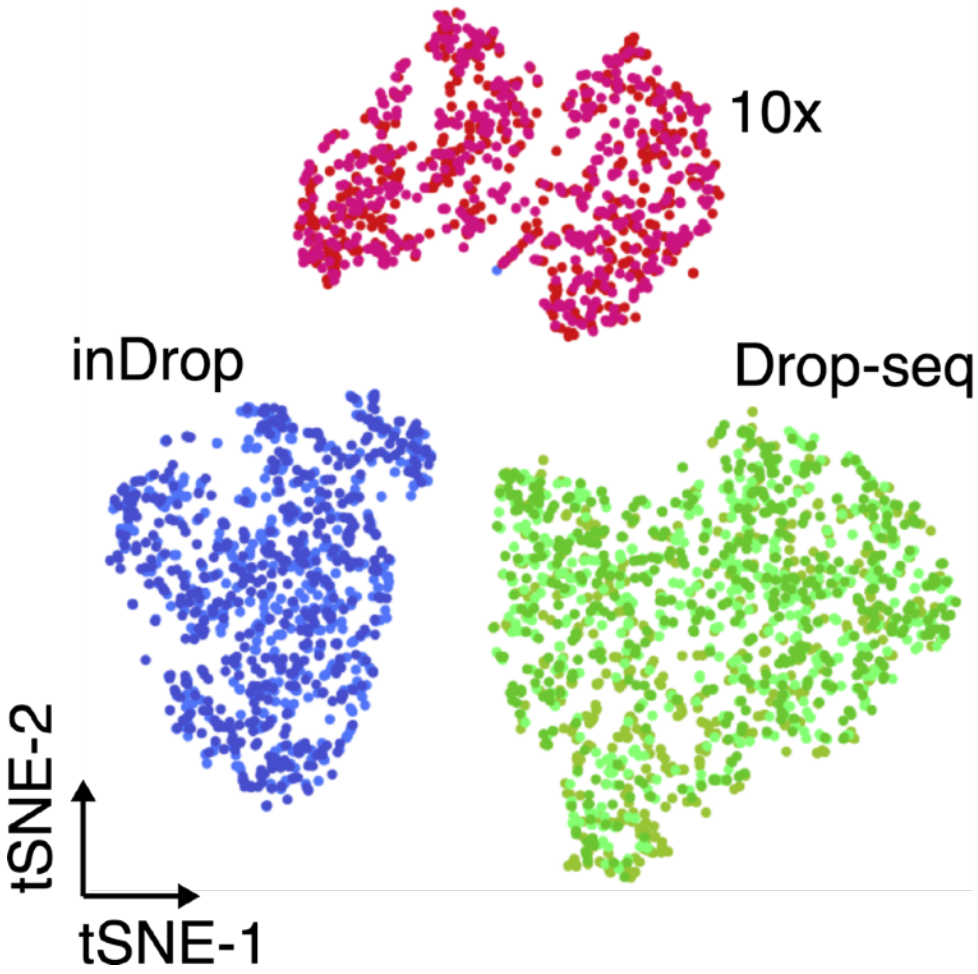
PCA plot of the tung data after downsampling



https://www.singlecellcourse.org/

5. Batch Correction

Related to normalisation and dealing with sparsity and dropouts

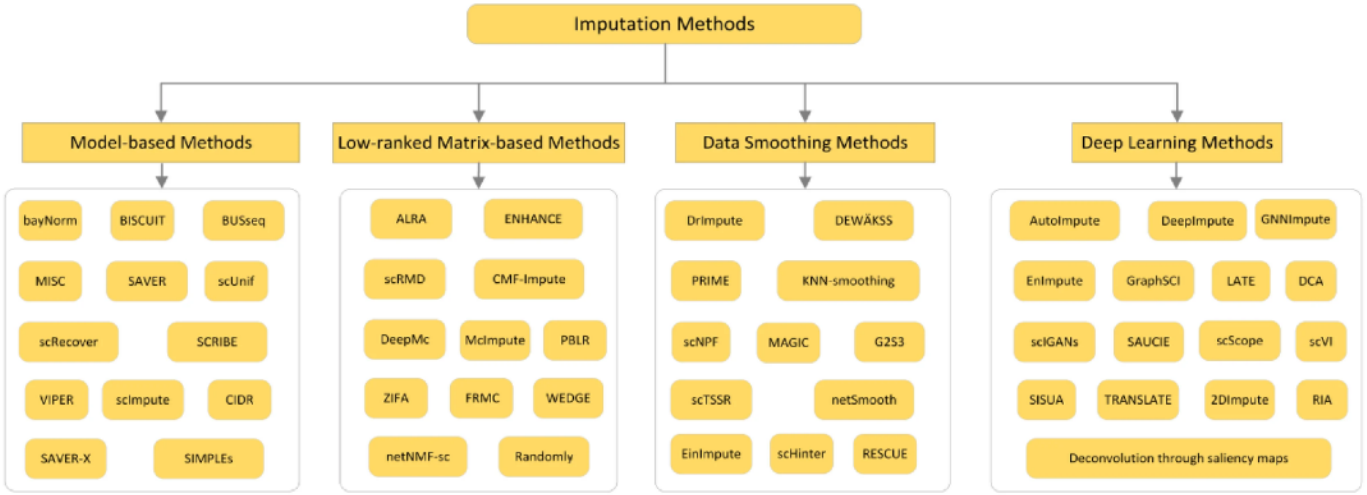


The same cell population was sequenced with three different single-cell protocols (colours).

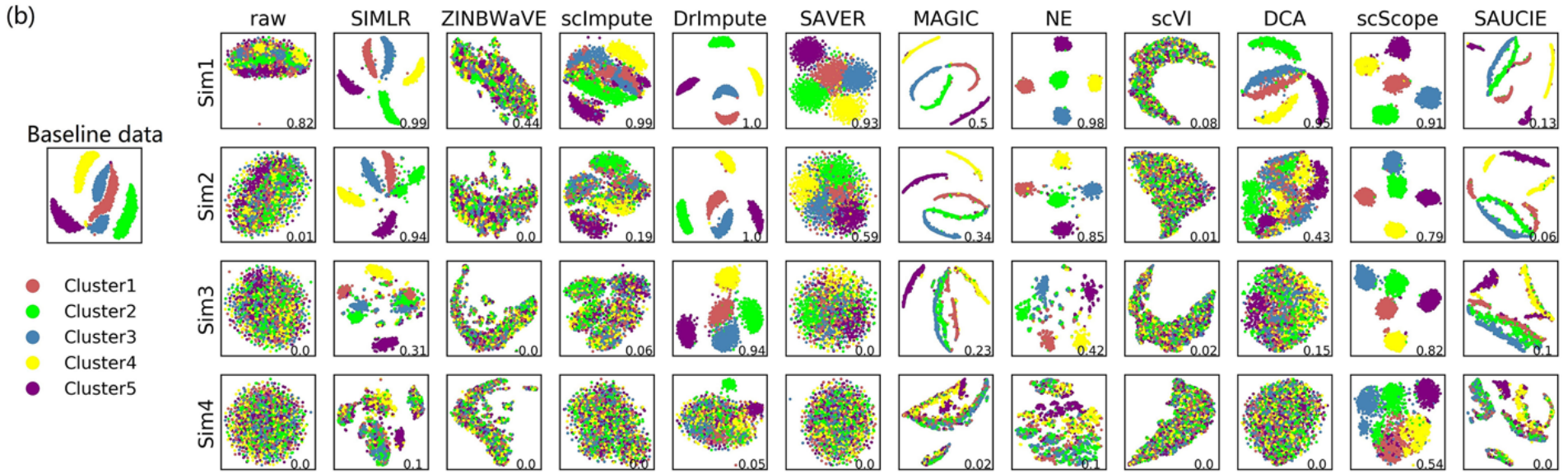
Adapted from Zhang et al.. Molecular Cell 73, 130 (2019)

6. Imputation

Related to normalisation and dealing with sparsity and dropout



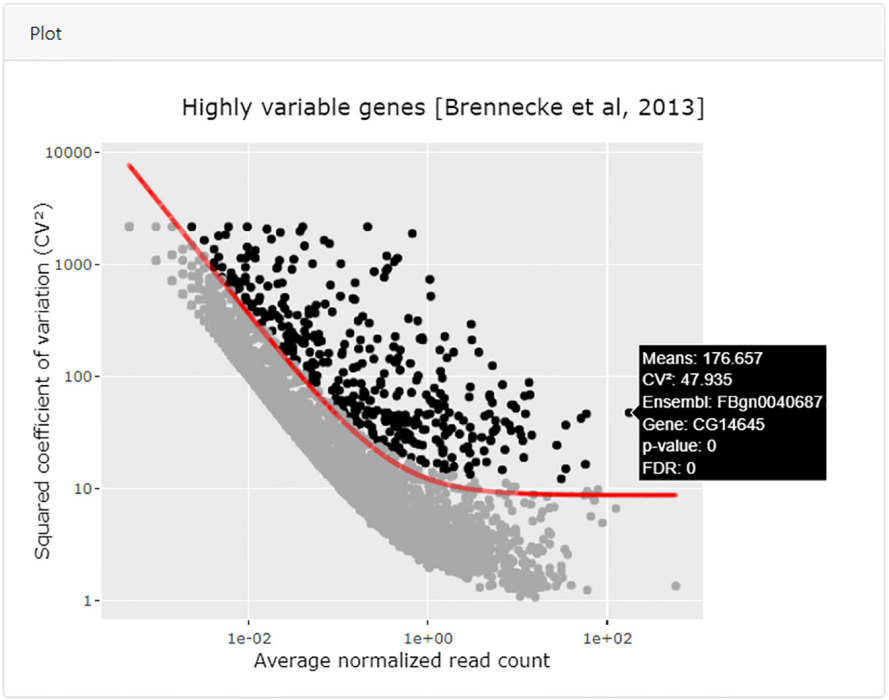
https://doi.org/10.3390/app122010684

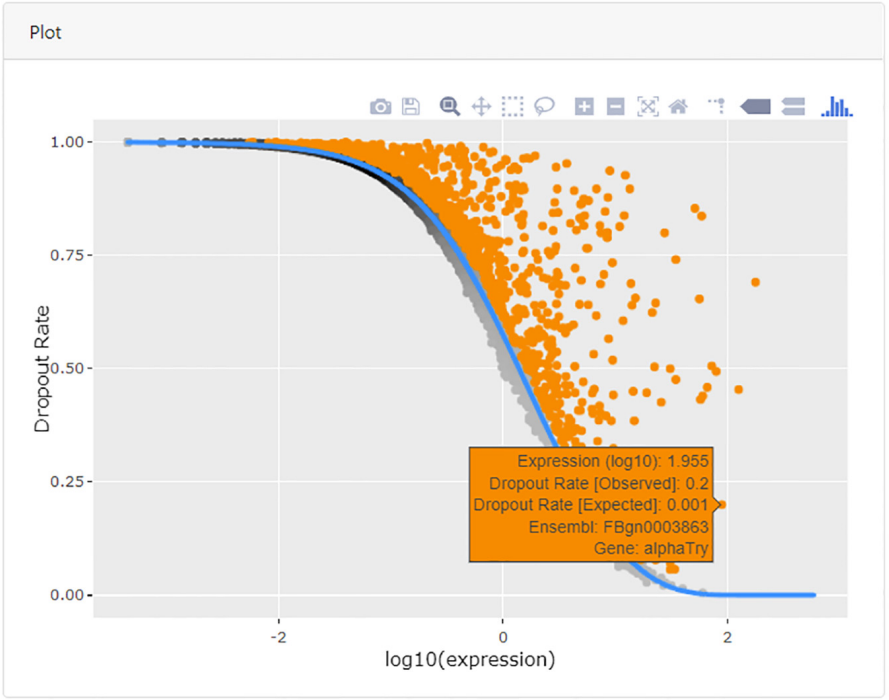


https://doi.org/10.1186/s12859-023-05417-7

7. Future selection

e.g. using Highly variable genes

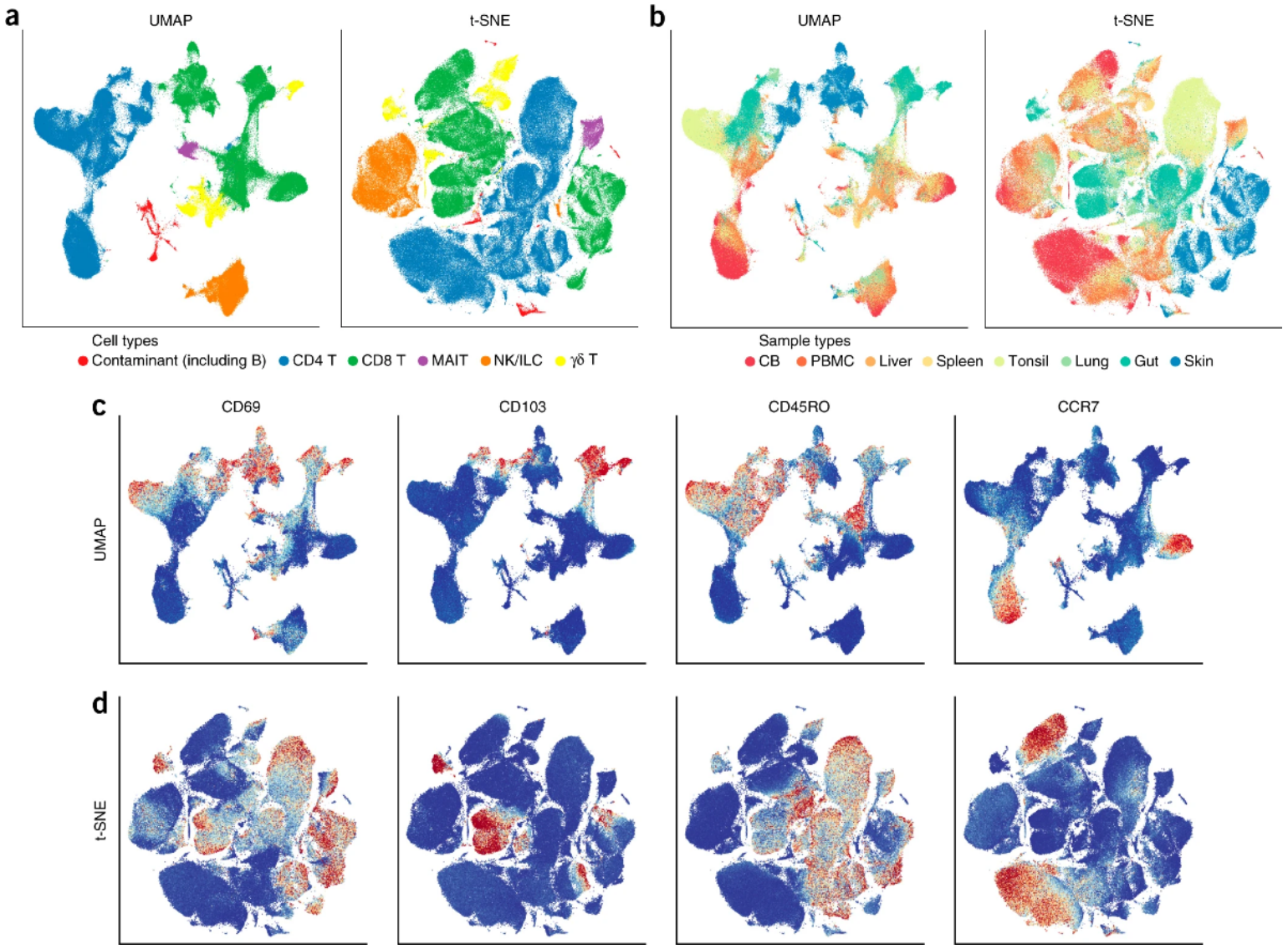




doi: 10.1093/nar/gkaa412

8. Dimensionality reduction

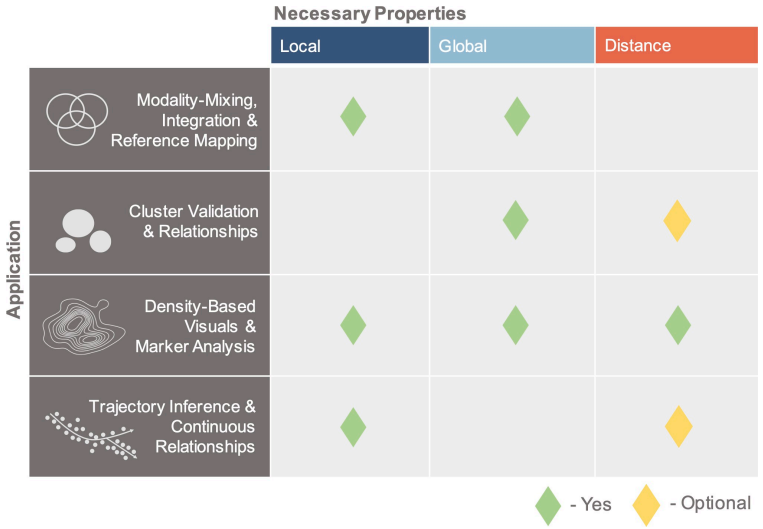
Promise of UMAP and t-SNE

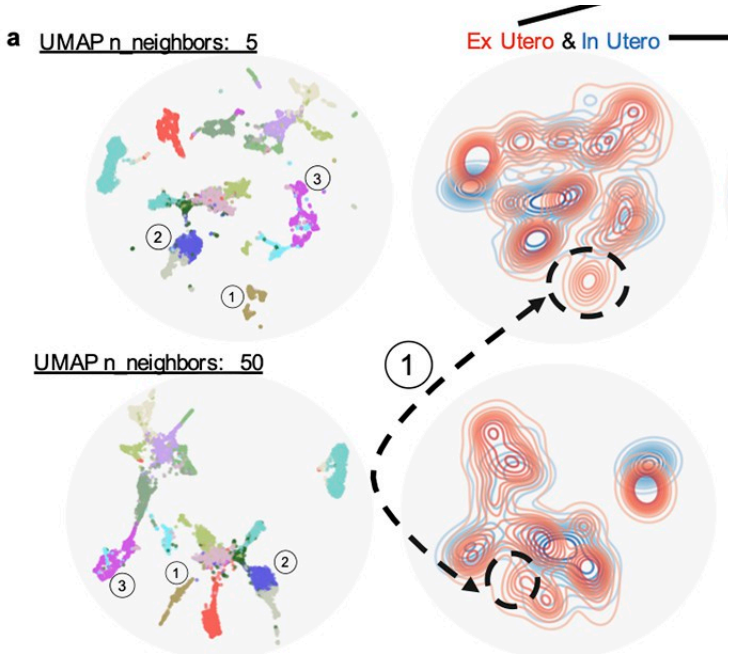


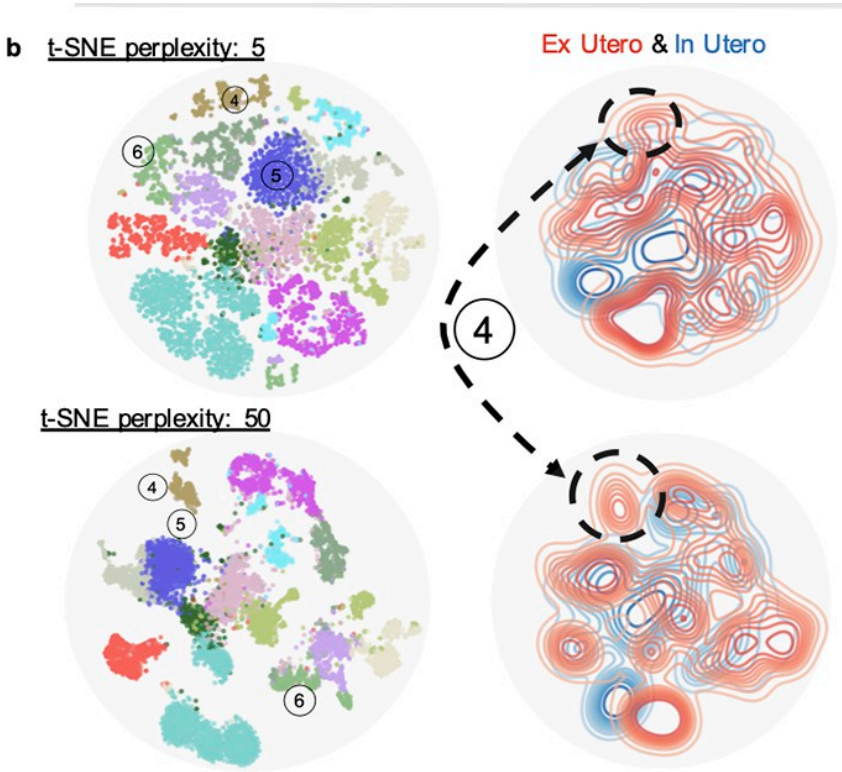
https://doi.org/10.1038/nbt.4314

8. Dimensionality reduction

Problems with UMAP and t-SNE



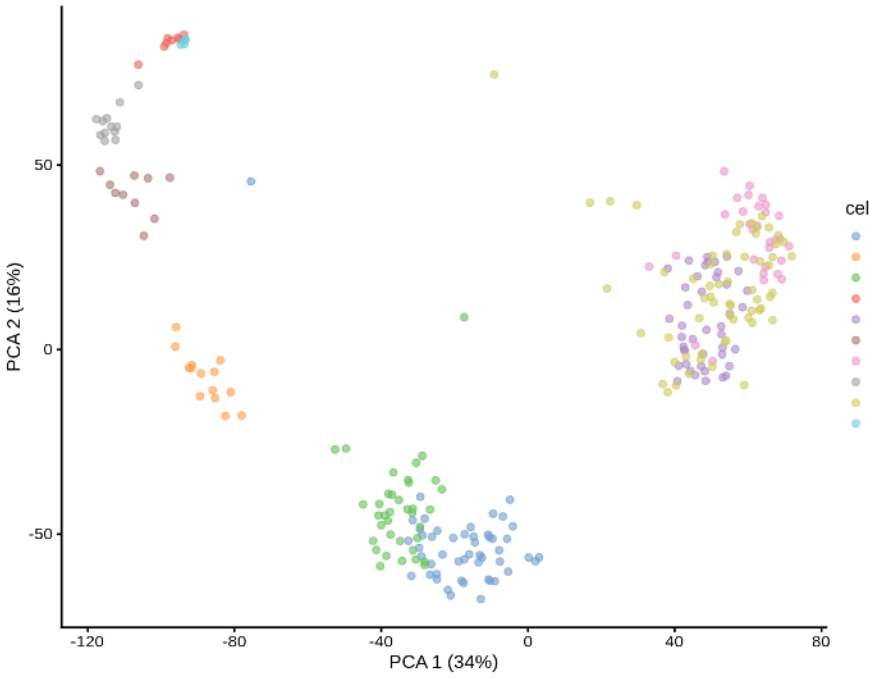


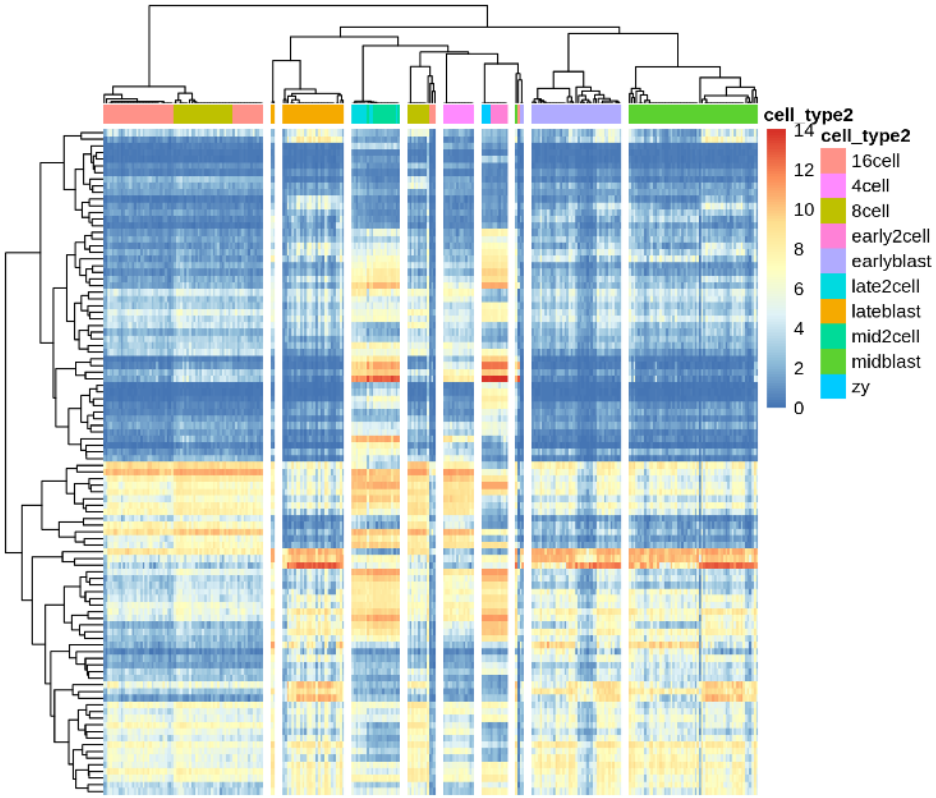


https://doi.org/10.1371/journal.pcbi.1011288

9. Clustering

Classification: cell type allocation-based on marker gene expressions

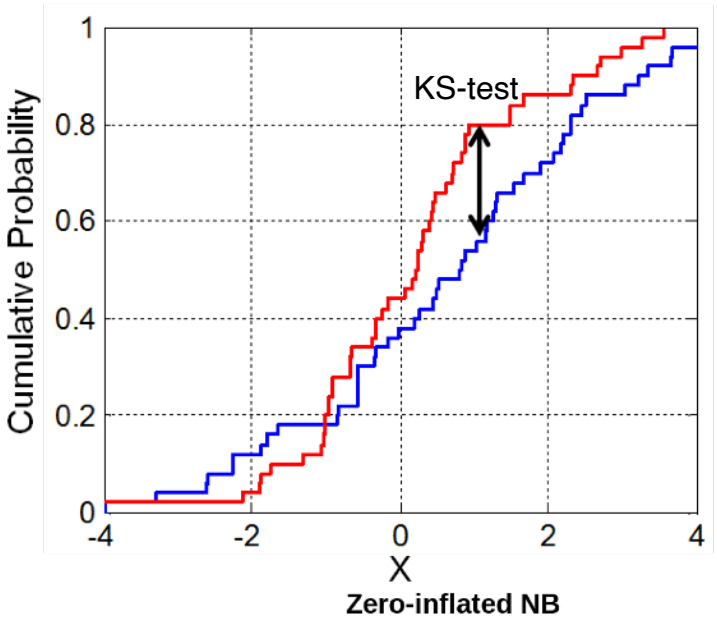




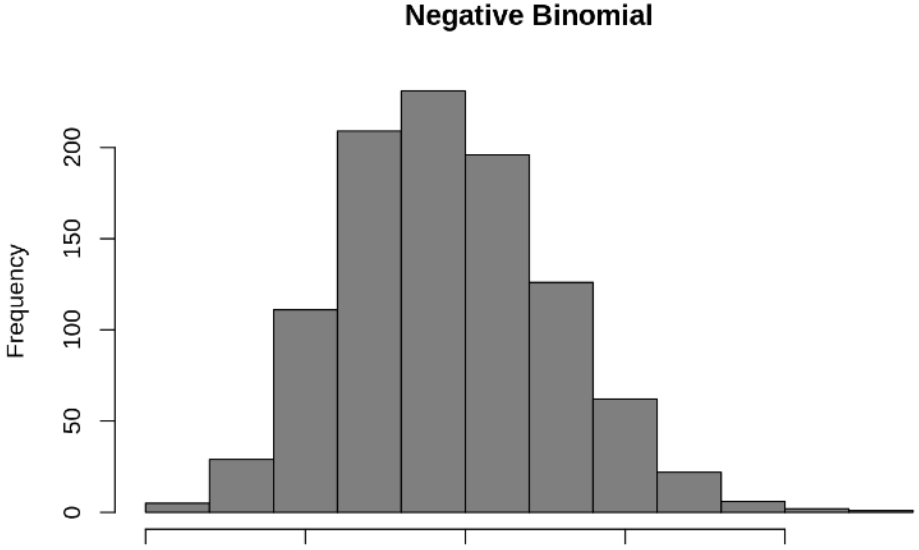
https://www.singlecellcourse.org/

10. Differential Expression Analysis

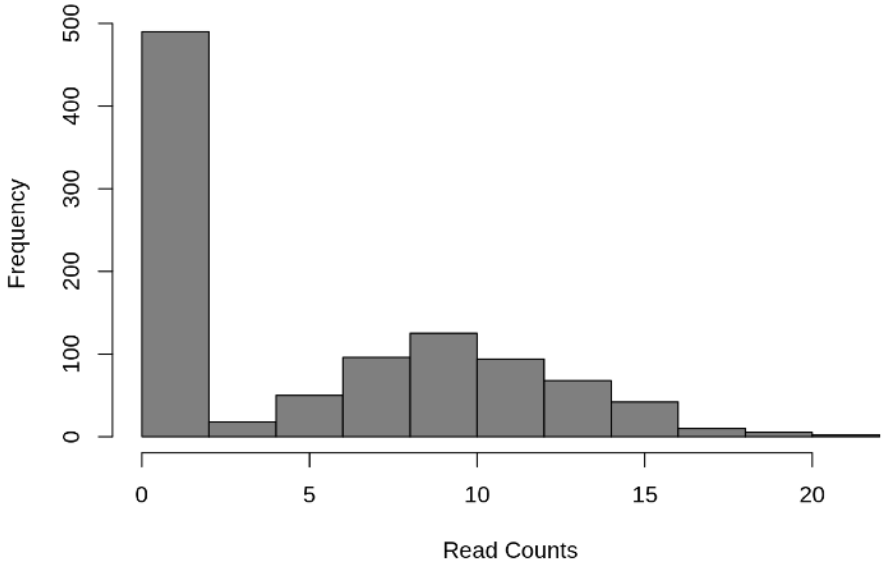
Parametric and non-parametric (+ counting for FDR)

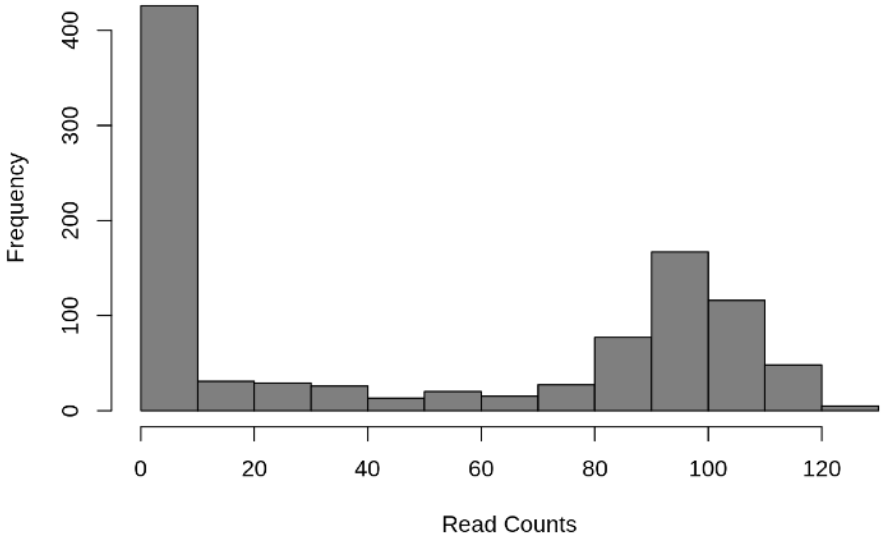


Zero-inflated NB

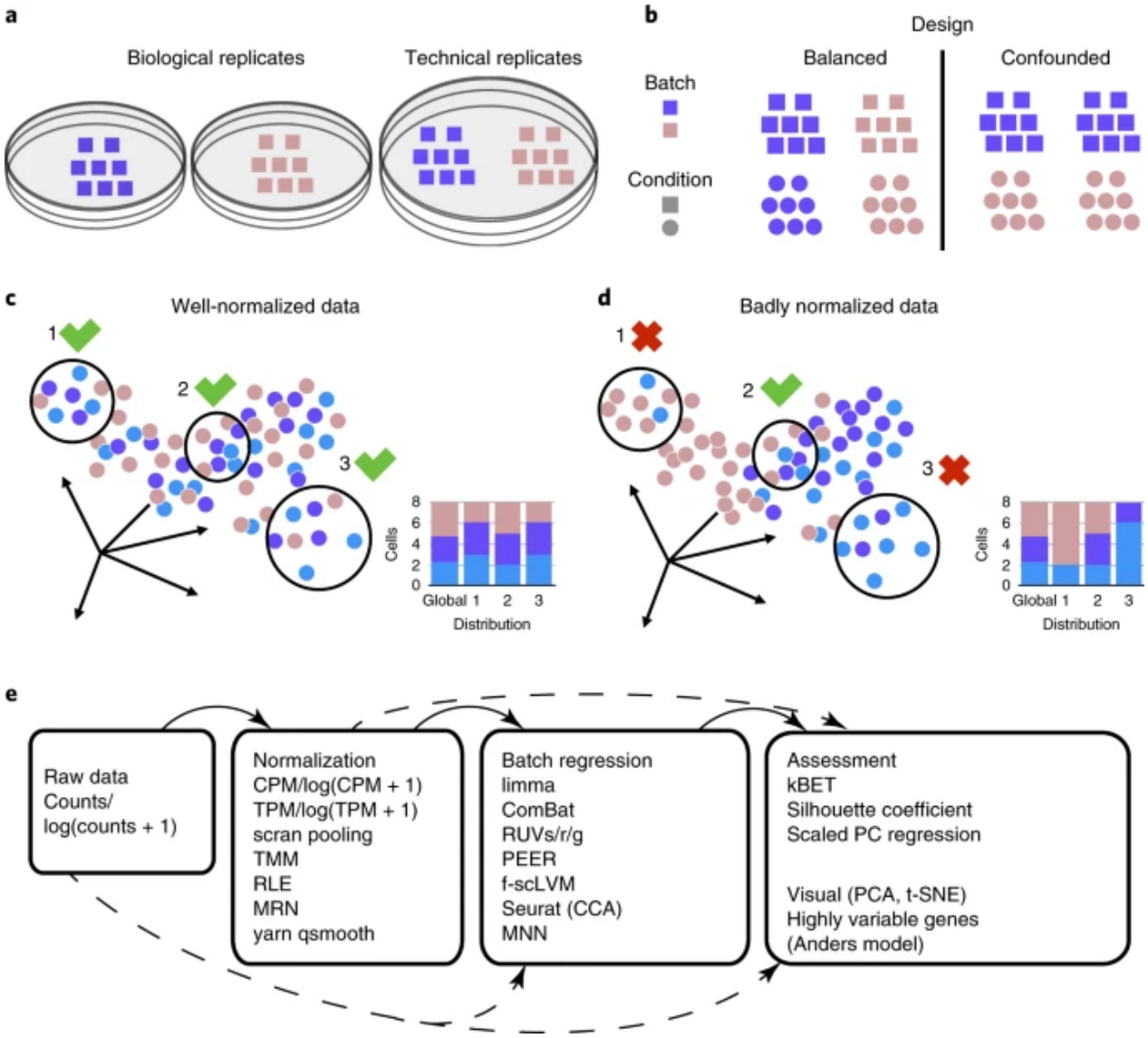


Poisson-Beta



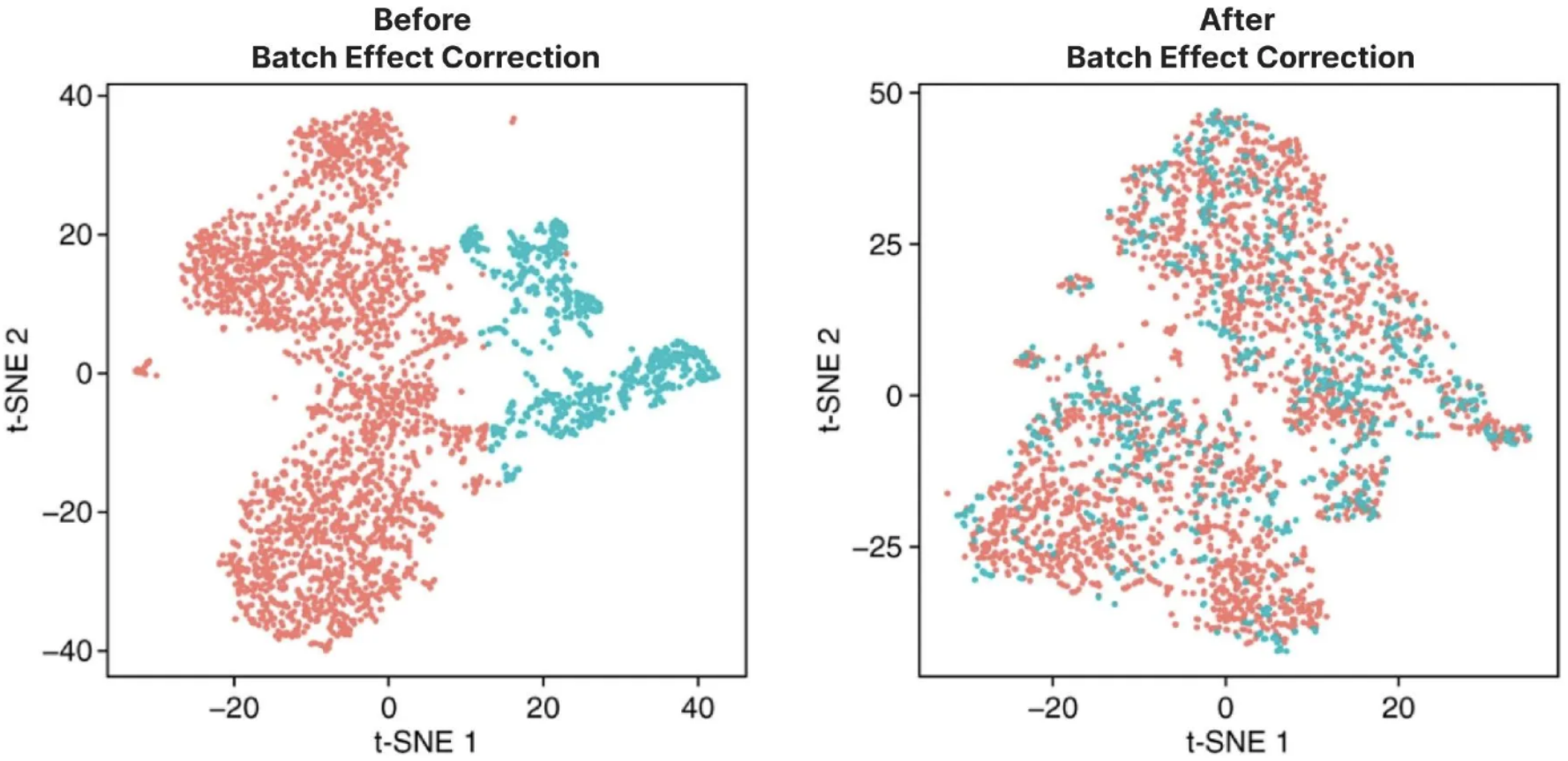


11. Data integration (and batch correction)

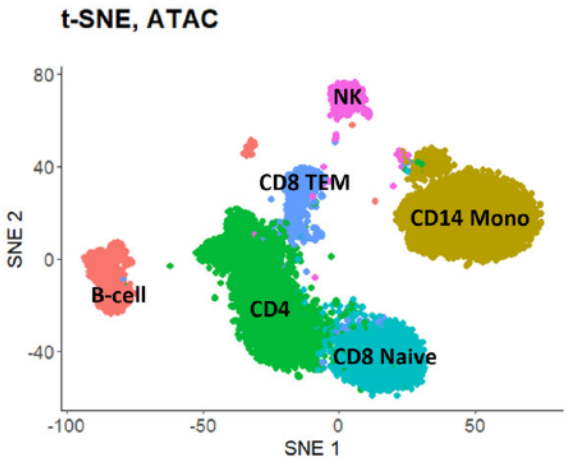


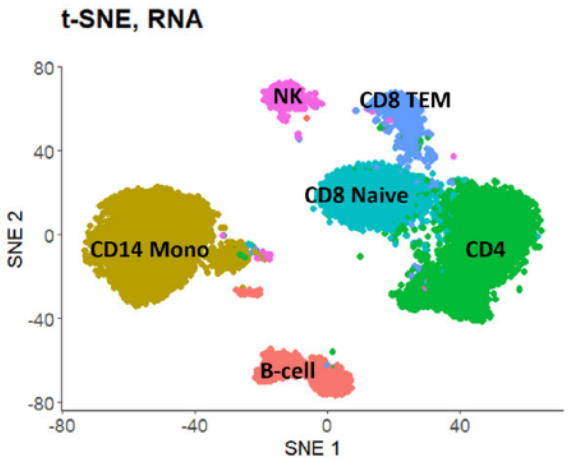
https://doi.org/10.1038/s41592-018-0254-1

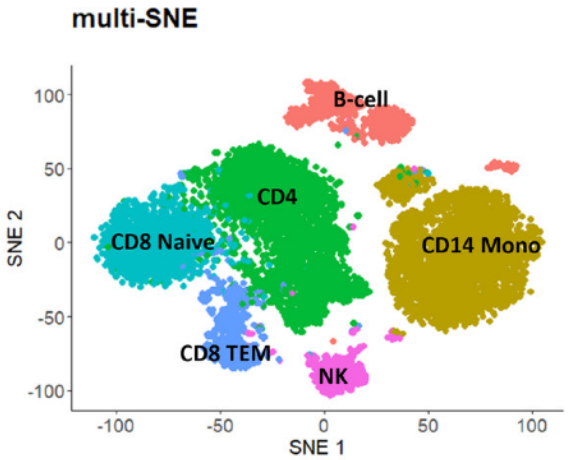
11. Data integration (and batch correction)



12. Multi-modal data integration

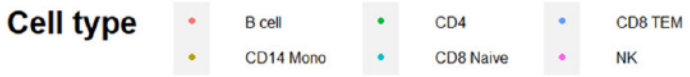






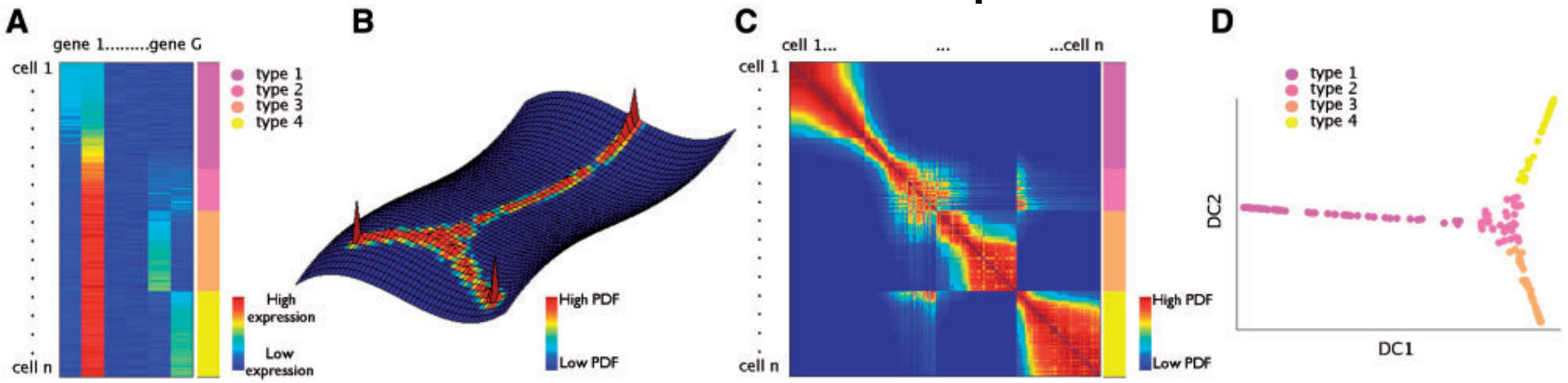
Multi-view clustering performance on single-cell data set

	ACC	NMI
Liu et al. (2013)	0.798	**0.751**
t-SNE, RNA	0.840	0.720
t-SNE, ATAC	0.805	0.704
multi-SNE	**0.872**	0.749

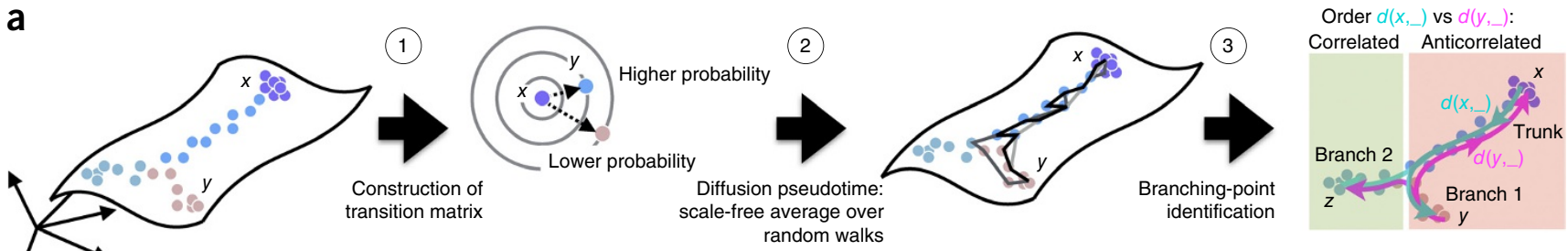


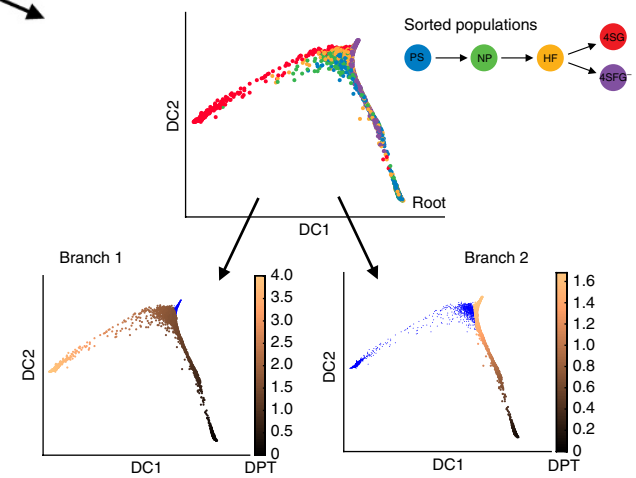
https://doi.org/10.7717/peerj-cs.1993

13. Trajectory-inference and pseudo-time ordering
Diffusion maps



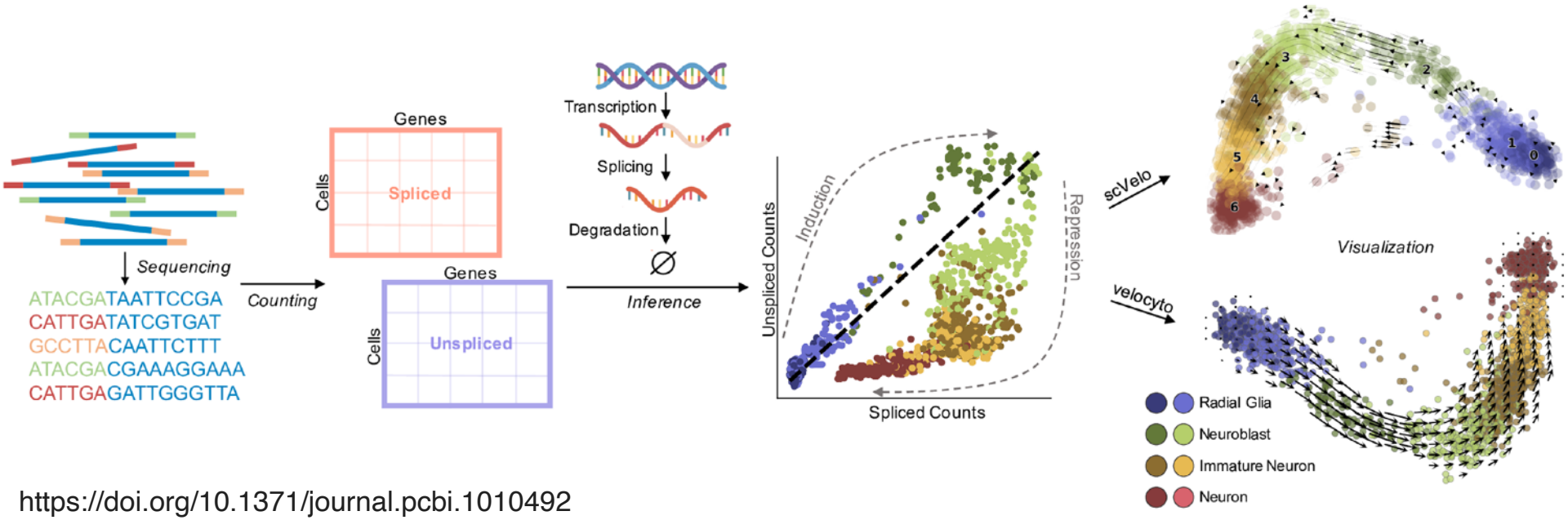
doi: 10.1093/bioinformatics/btv325



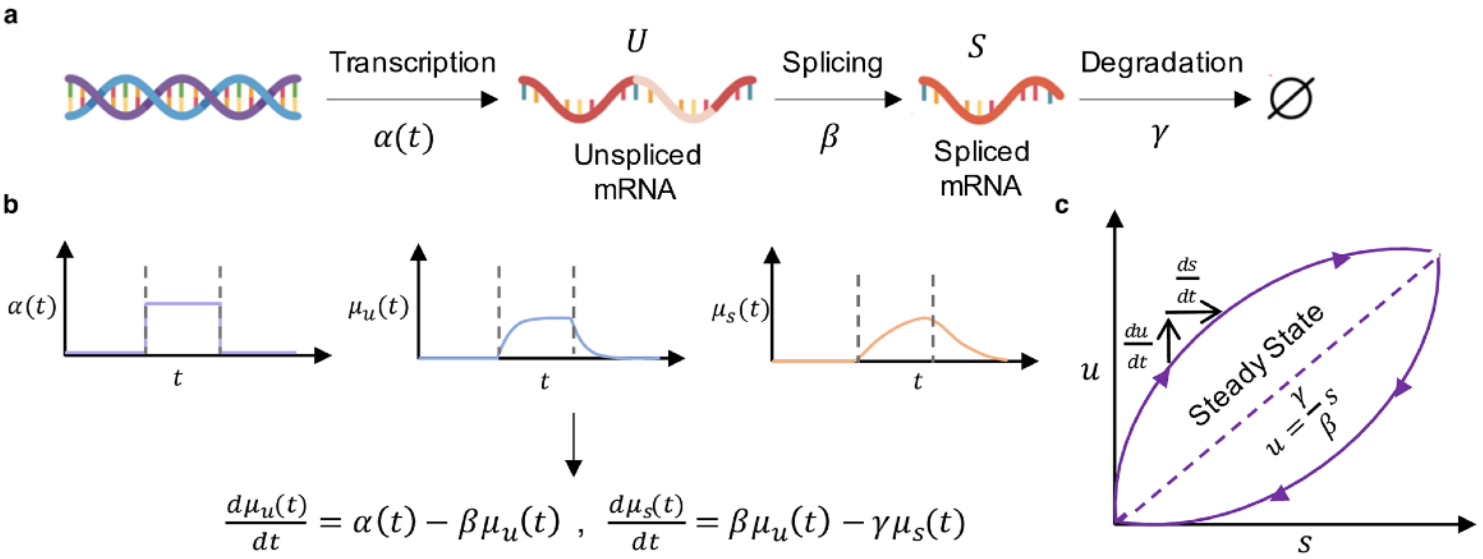


doi:10.1038/nmeth.3971

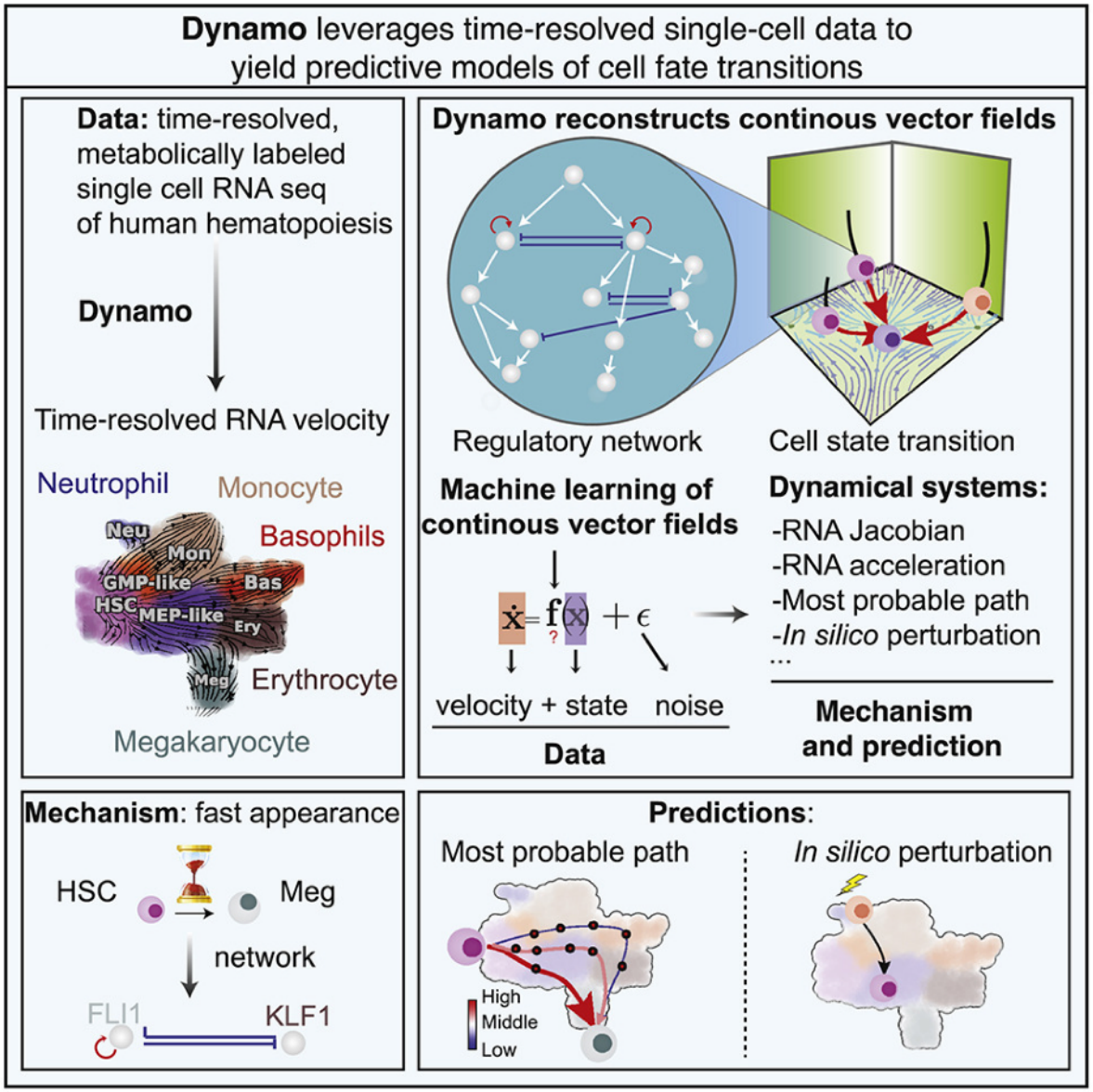
14. RNA-velocity analysis



https://doi.org/10.1371/journal.pcbi.1010492



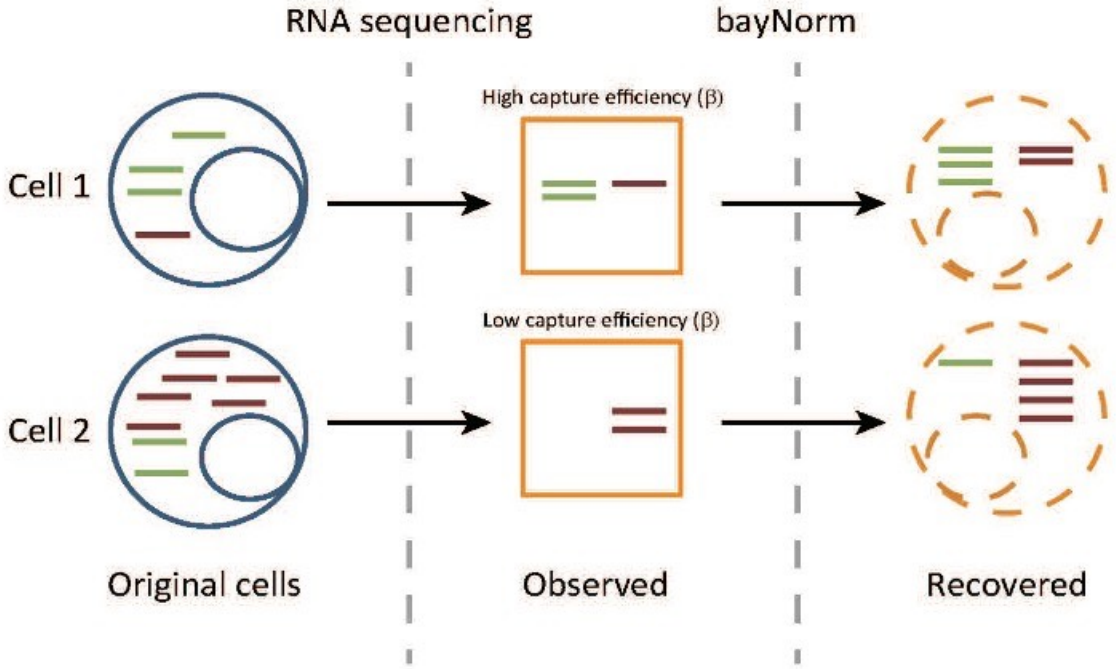
15. Gene-regulatory network inference



https://doi.org/10.1016/j.cell.2021.12.045

Some research examples from the work in my group

bayNorm: A method for normalisation, imputation and true gene expression recovery



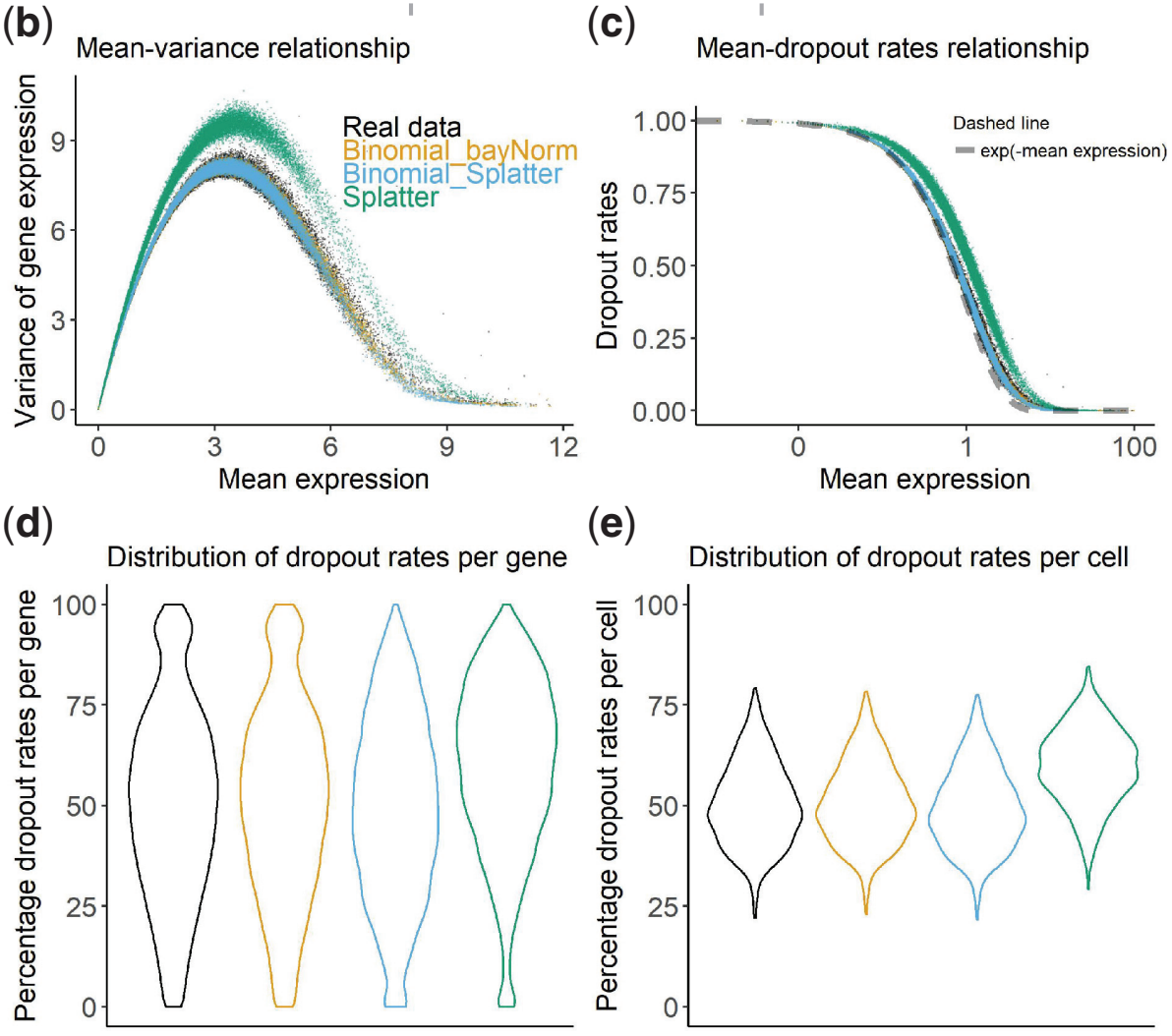
$$\underbrace{\Pr(x_{ij}^0|x_{ij},\beta_j,\mu_i,\phi_i)}_{\text{Posterior}} = \frac{\overbrace{\Pr(x_{ij}|x_{ij}^0,\beta_j)}^{\text{Likelihood}} \times \overbrace{\Pr(x_{ij}^0|\mu_i,\phi_i)}^{\text{Prior}}}{\underbrace{\Pr(x_{ij}|\mu_i,\phi_i,\beta_j)}_{\text{Marginal likelihood}}}$$

$$x_{ij}|x_{ij}^0 \sim \text{Binom}(x_{ij}^0,\ \text{prob} = \beta_j),$$

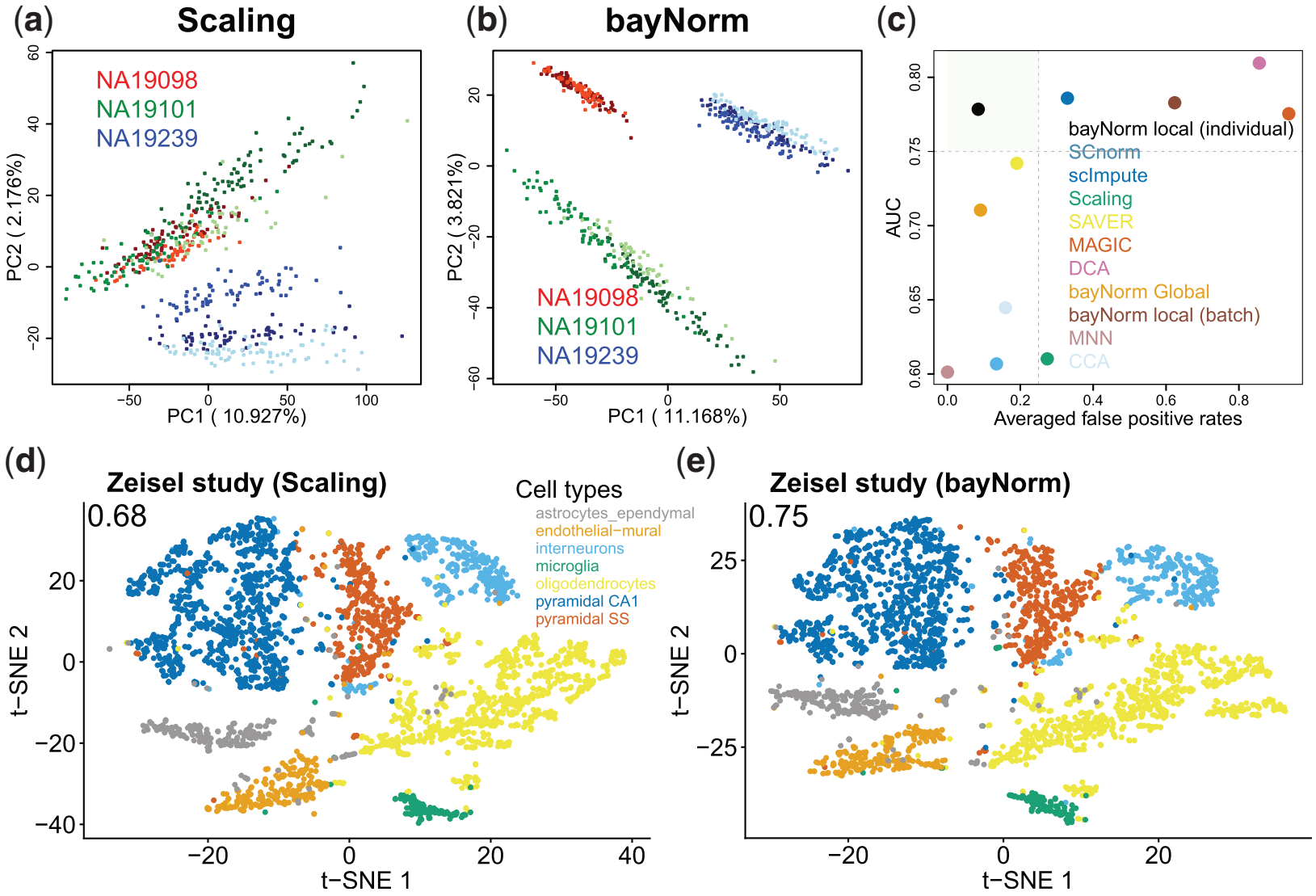
$$x_{ij}^0 \sim \text{NB}(\text{mean} = \mu_i,\ \text{size} = \phi_i).$$

doi: 10.1093/bioinformatics/btz726

bayNorm posterior predictive explains drop-out (No need for zero-inflated distributions)



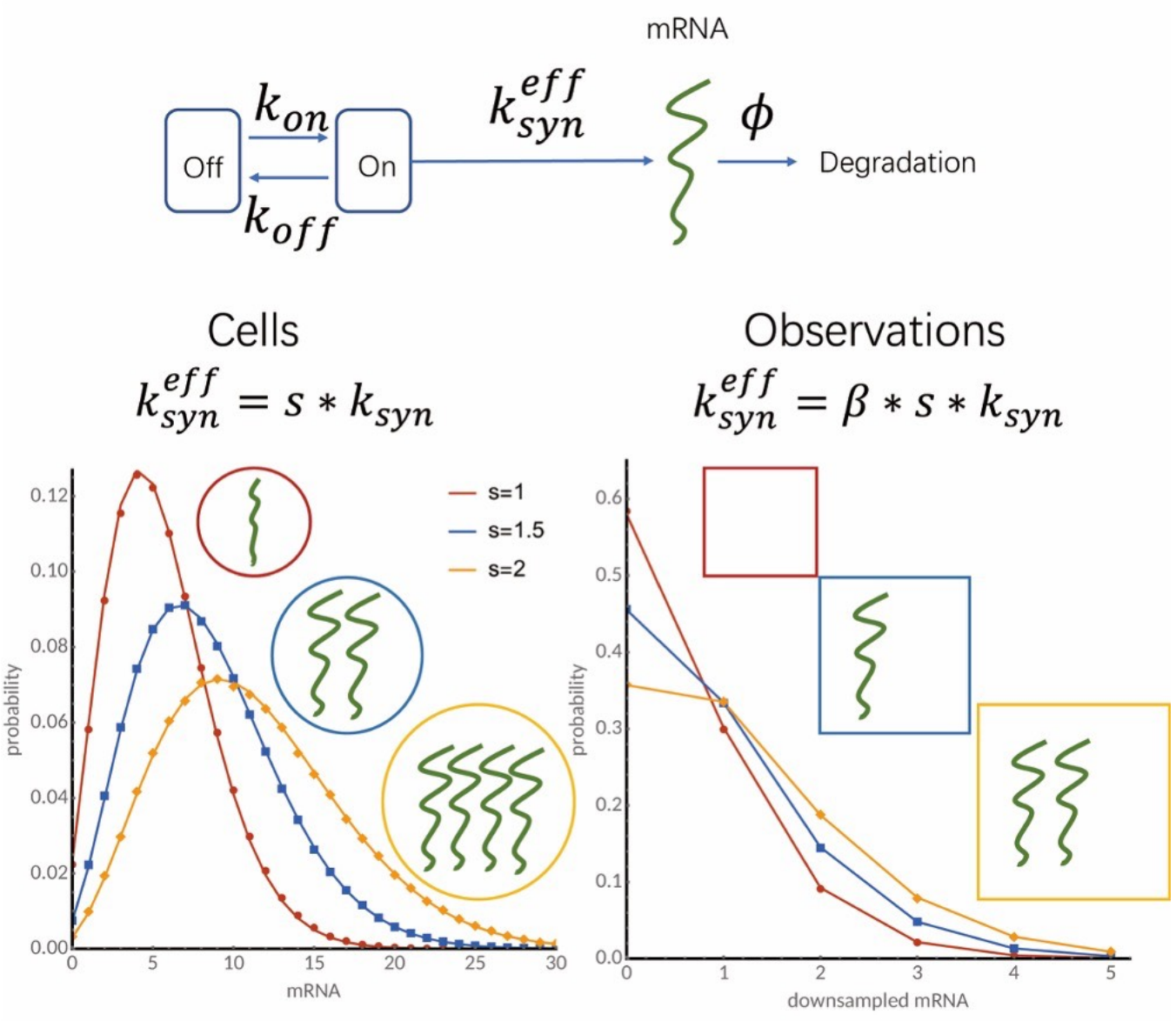
bayNorm can remove batch affect and cell type identification



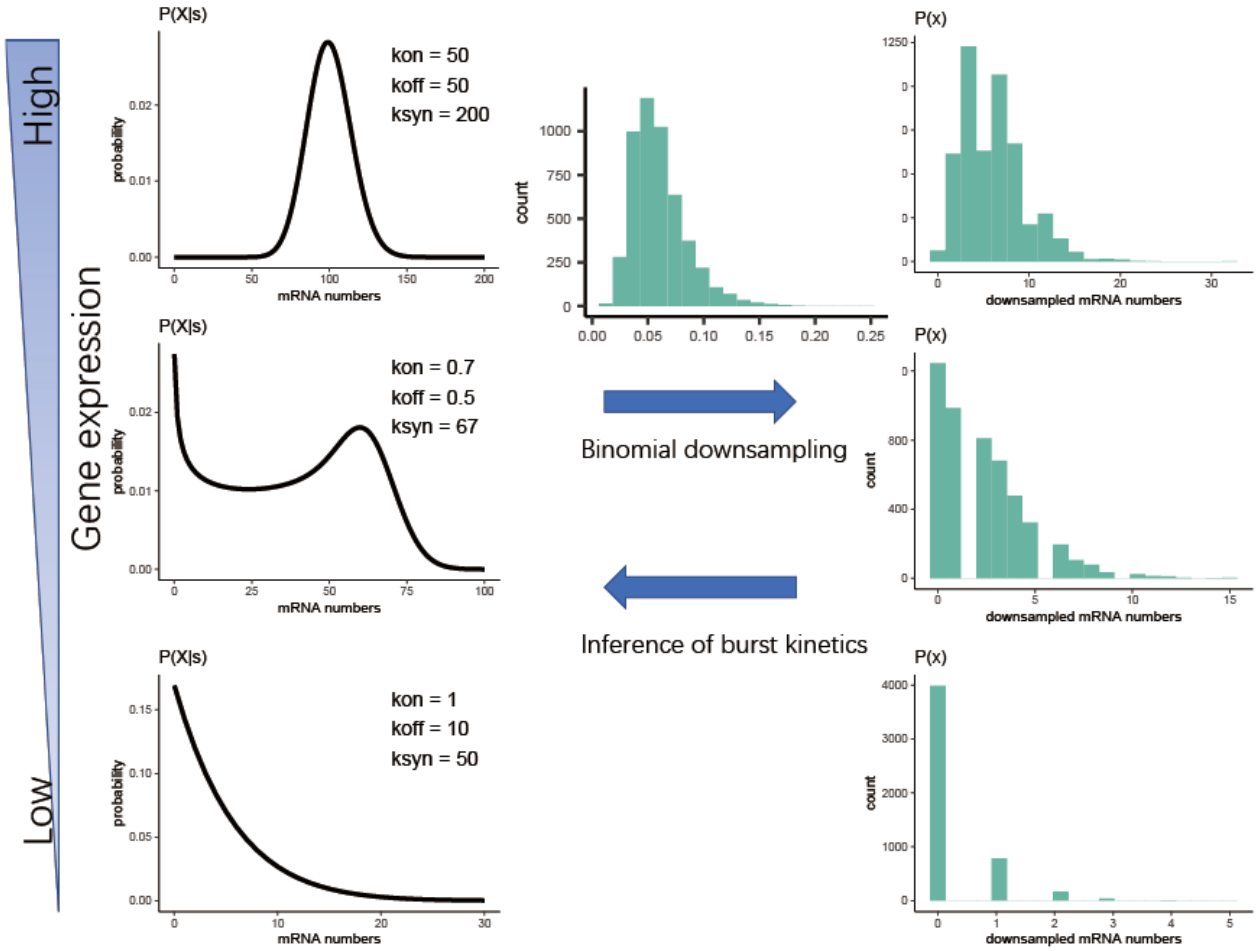
Inference of burst kinetics from scRNA-seq data using bayNorm model

$$X \sim \text{Pois}(k_{\text{syn}}^{\text{eff}} p)$$

$$p \sim \text{Beta}(k_{\text{on}}, k_{\text{off}})$$

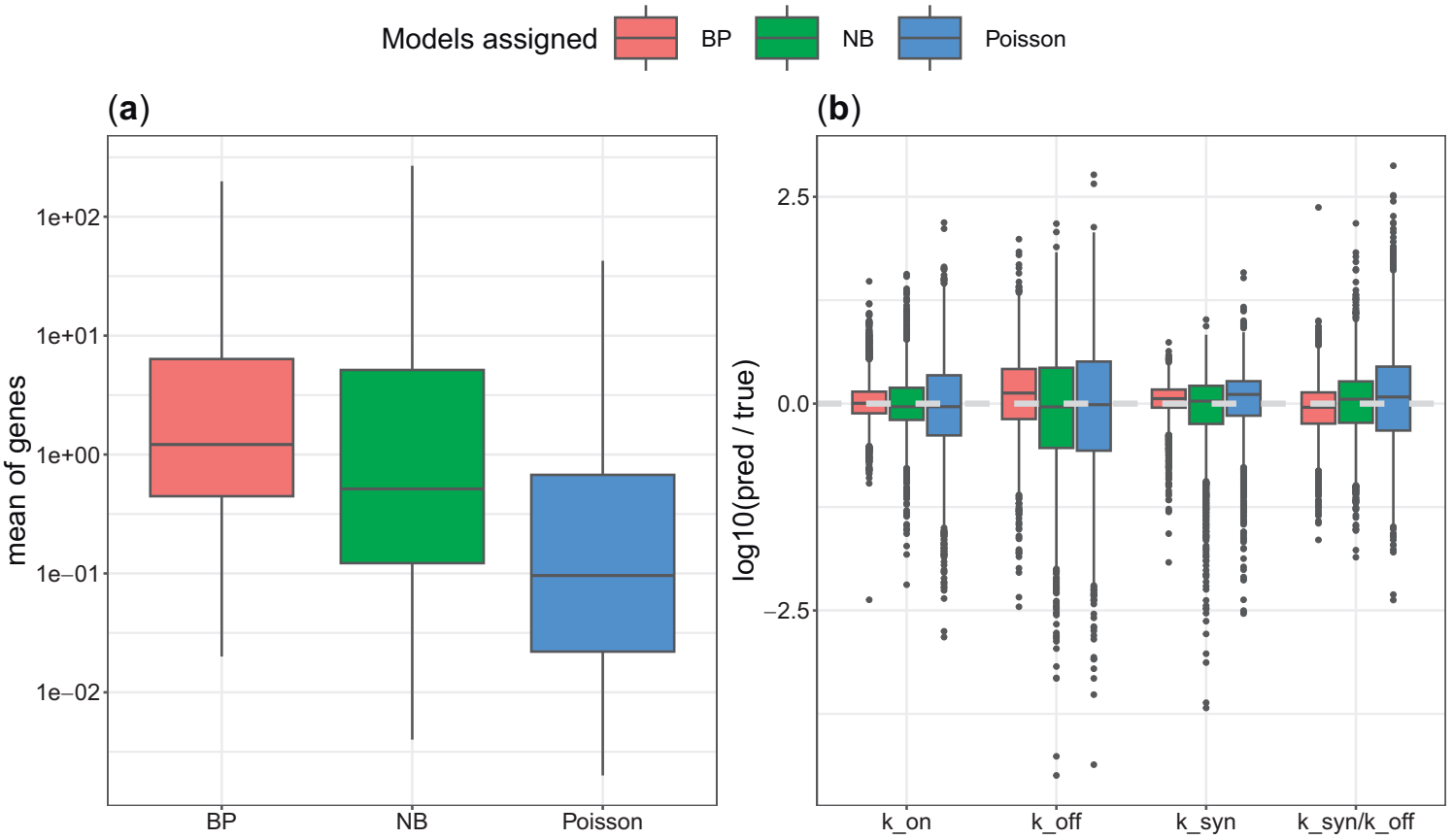


Burst kinetics inference using simulation-based inference based on neural networks

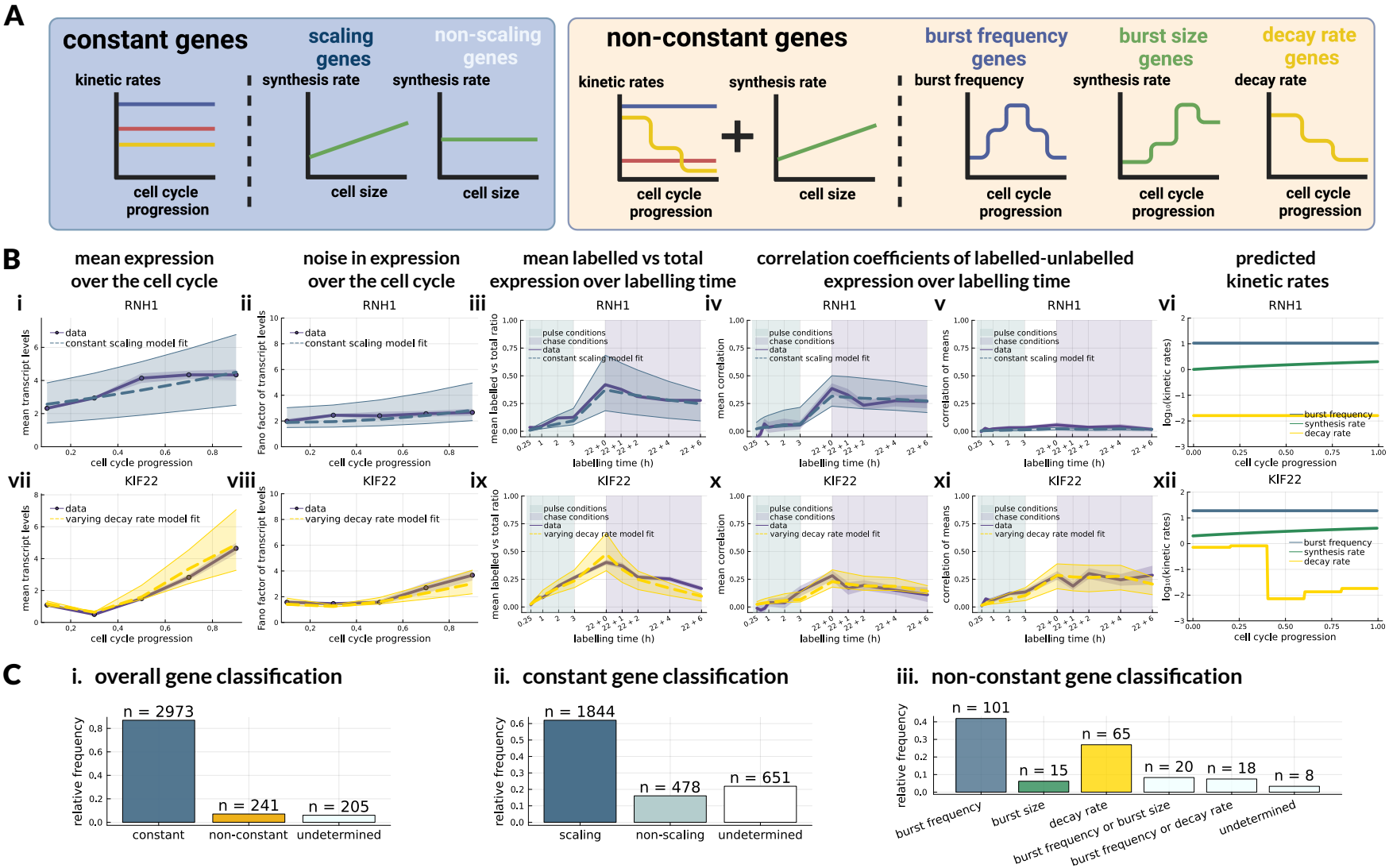


Tang et al, Bioinformatics 2020
Tang et al, Bioinformatics 2023

Model selection results chooses simpler models when there is less data



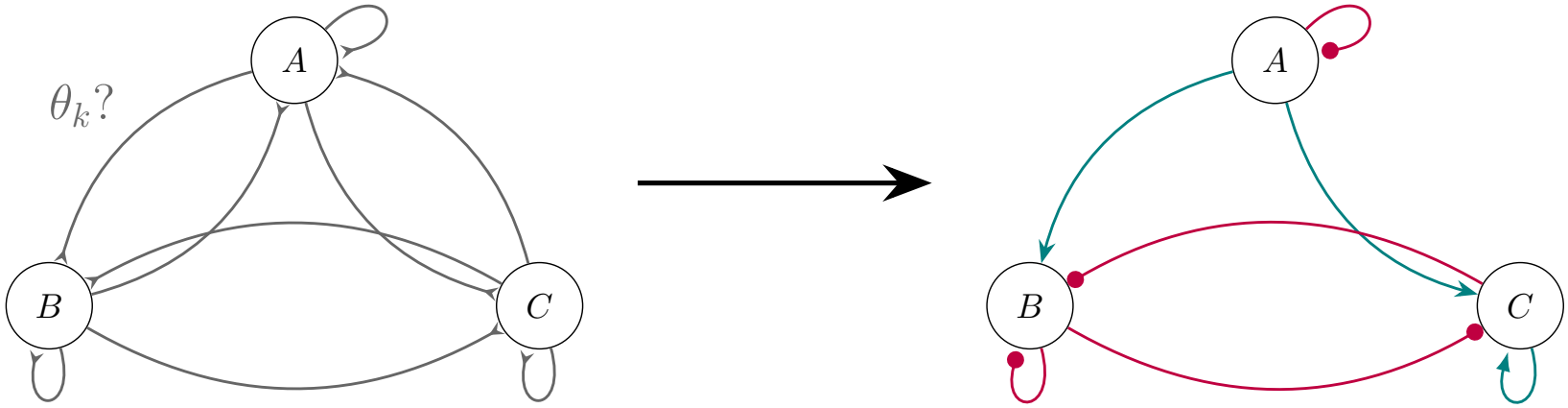
Genome-wide inference of cell size scaling and cell cycle dependence of gene expression Using metabolic labelling time-resolved transcription data (scEU-RNA-seq)



Volteras, Shahrezaei & Thomas Cell Systems 2024

seeDimitris Volteras talk on Friday

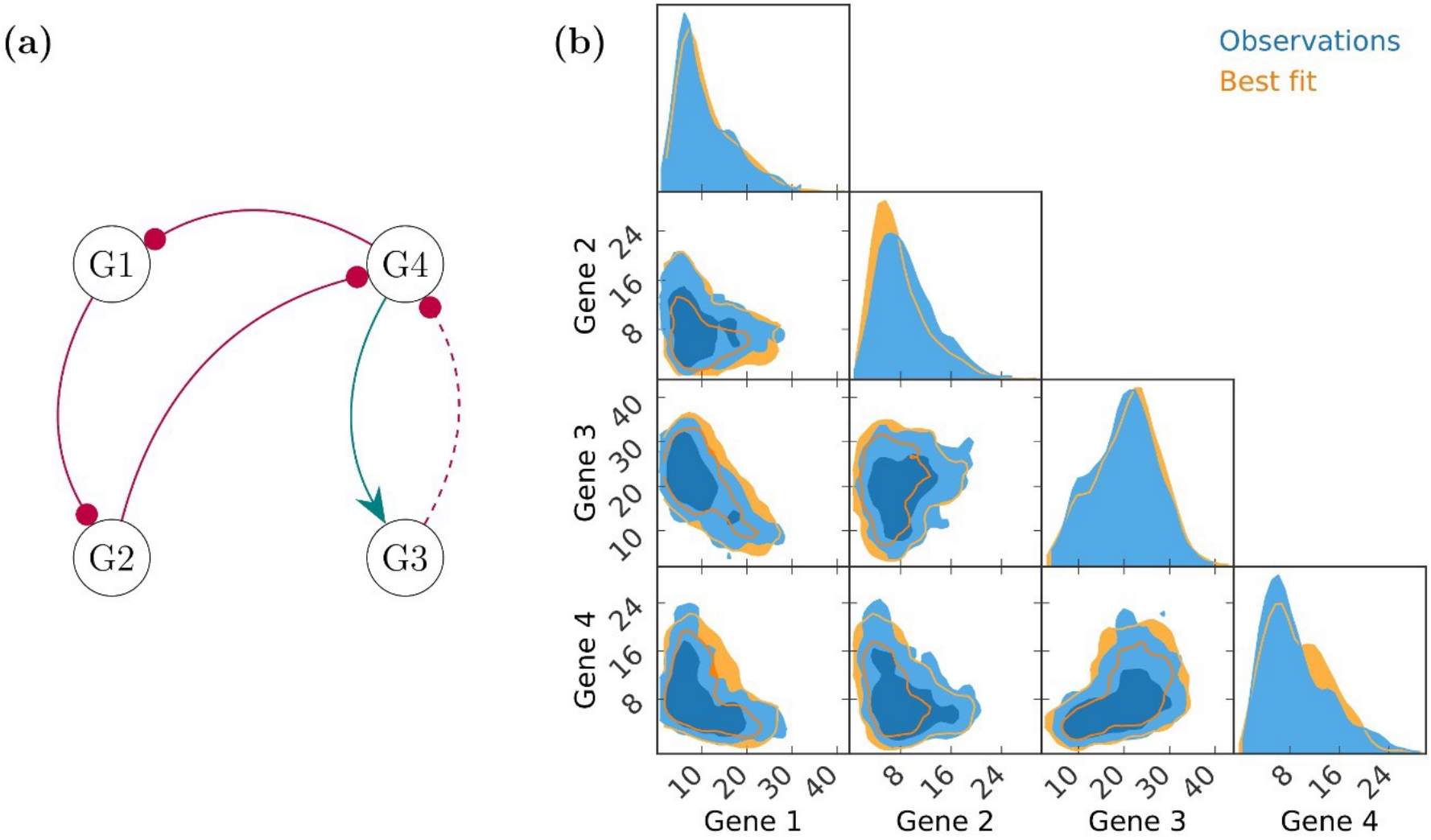
Inferring model structure
Model discovery (or equation discovery) for biochemical networks



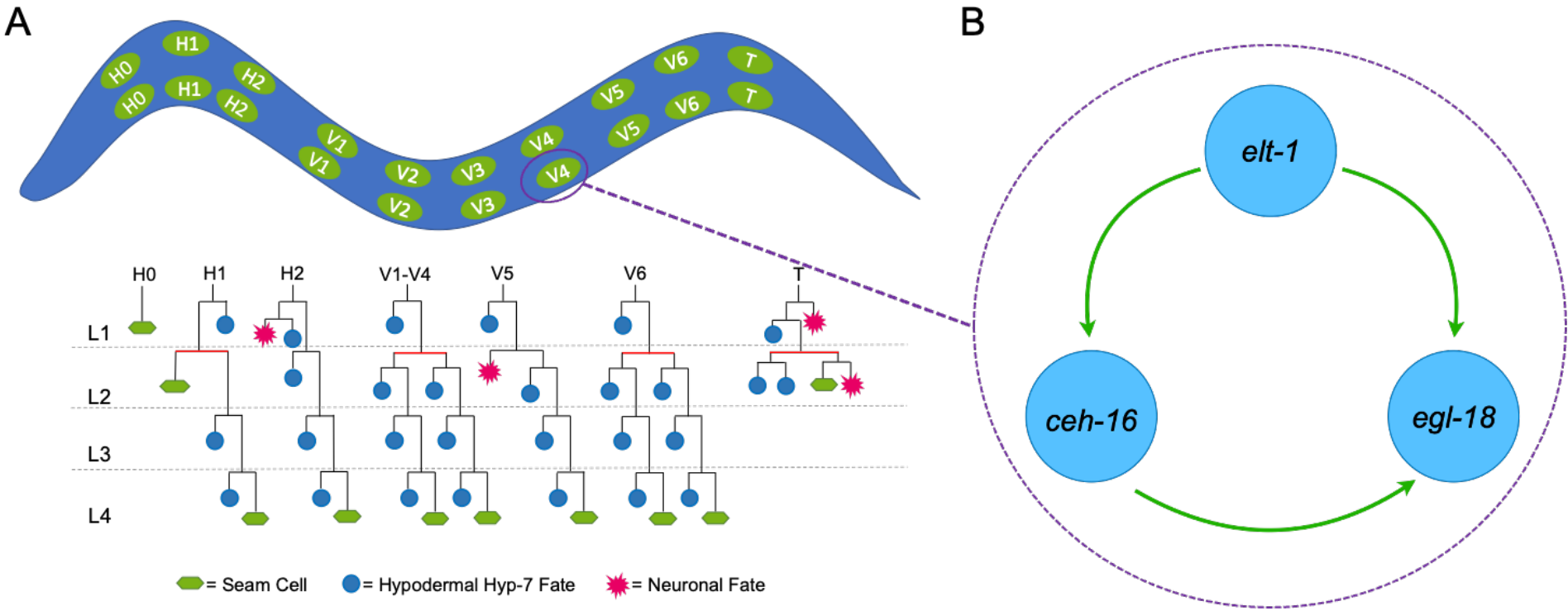
Large scale model selection can be reduced to variable selection

Bayesian Lasso, ABC with sparsity inducing parameters, SINDy

Inference of stochastic genetic networks from single cell data

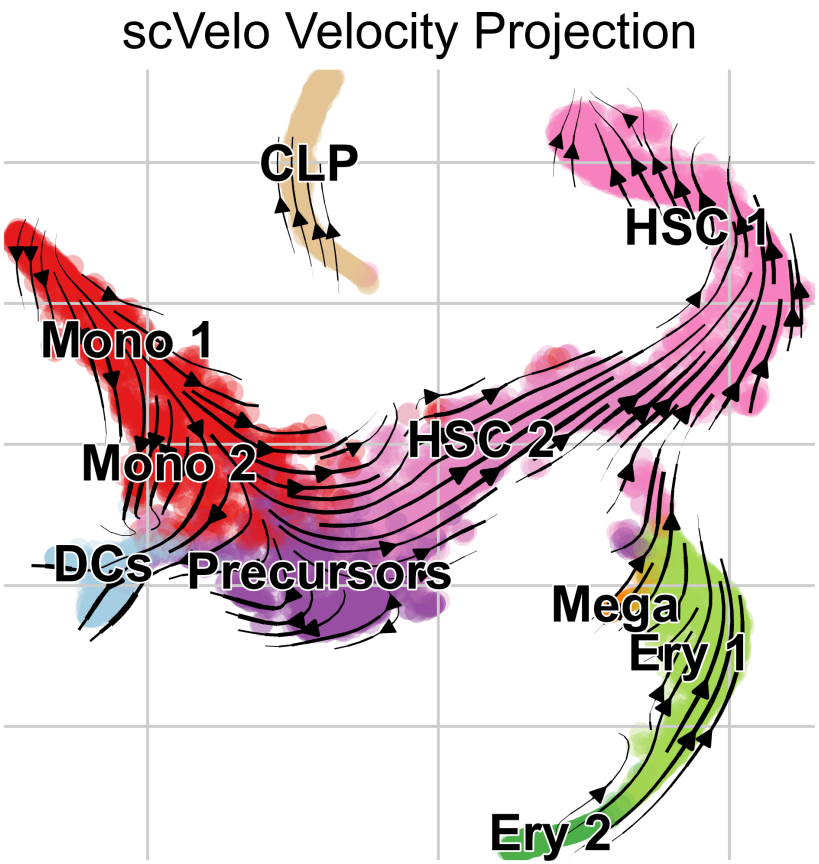


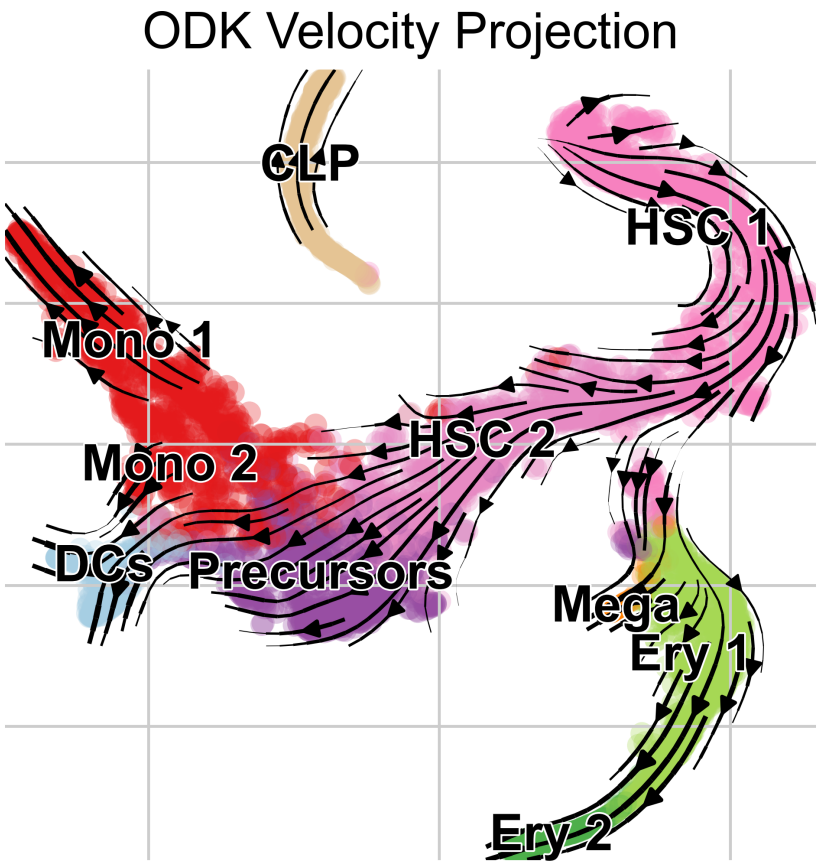
Application to the inference of the genetic network controlling stem cell fate decision



Brozek et al, iScience 2025

Inferring the low-dimensional dynamics of underlying stochastic process from the high-dimensional scRNA-seq data using a novel ordered diffusion kernel approach



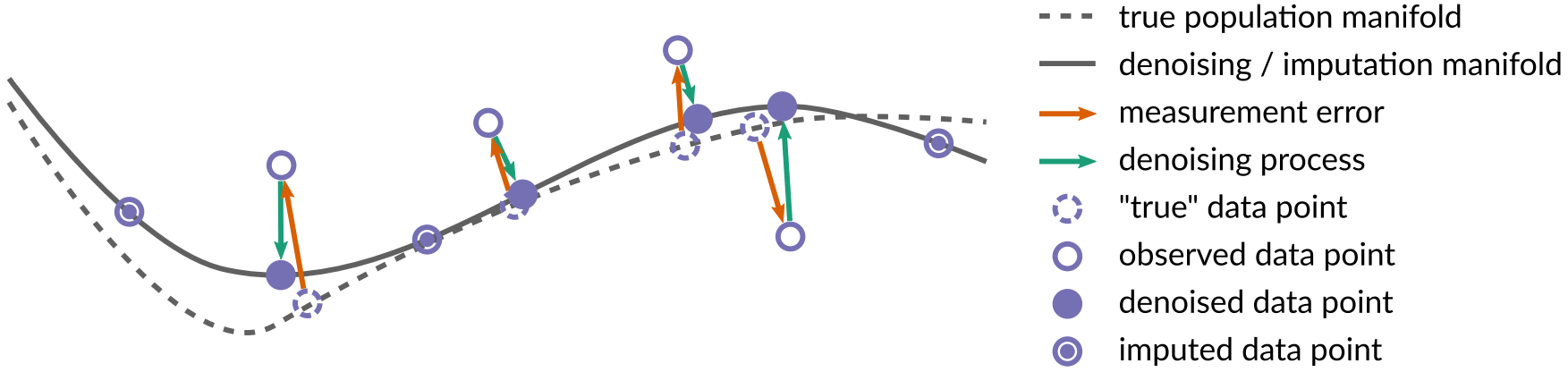


Jack Soulsby, unpublished work of current PhD student

See talk on Friday

There are long-standing grand challenges for single-cell data science:

Technical errors, sparcity, high-dimensionality and data integration across batches and modalities
Inferring patterns, trajectories and dynamics



Lähnemann *et al*. *Genome Biology* (2020) 21:31

https://doi.org/10.1186/s13059-020-1926-6