Single-cell RNA-seq

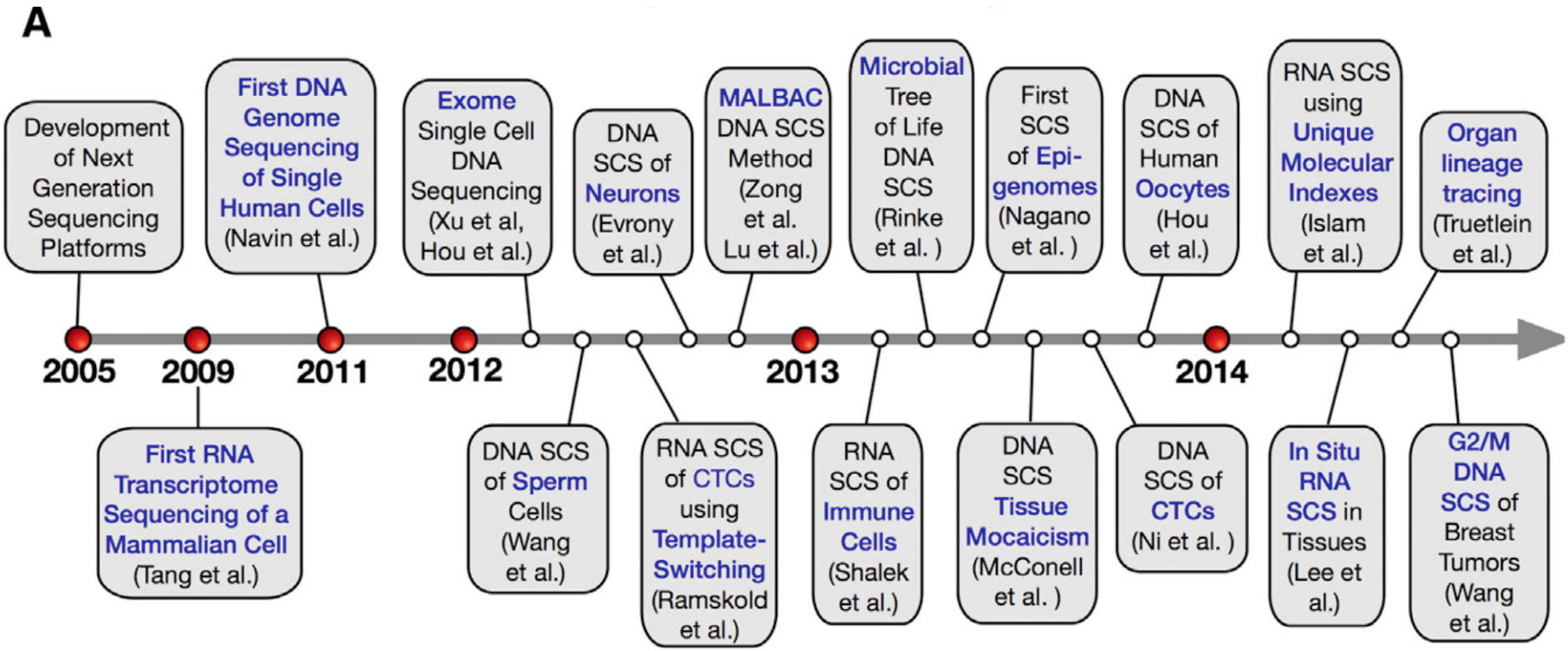
Mikhail Dozmorov

Spring 2018

Single cell sequencing applications

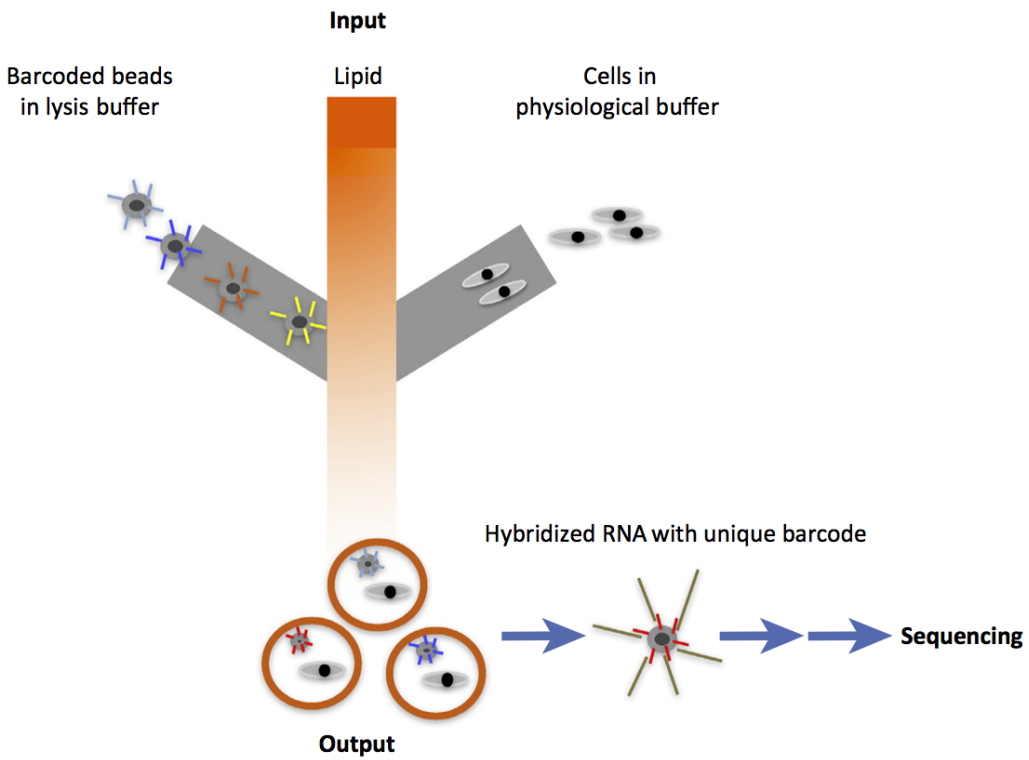
- Infer cell lineages
- Identify subpopulations
- Outline temporal evolution
- Define cell-specific biological characteristics, e.g., differentially expressed genes

Single cell timeline



http://www.cell.com/molecular-cell/fulltext/S1097-2765(15)00341-X

Single-cell Sequencing Technology

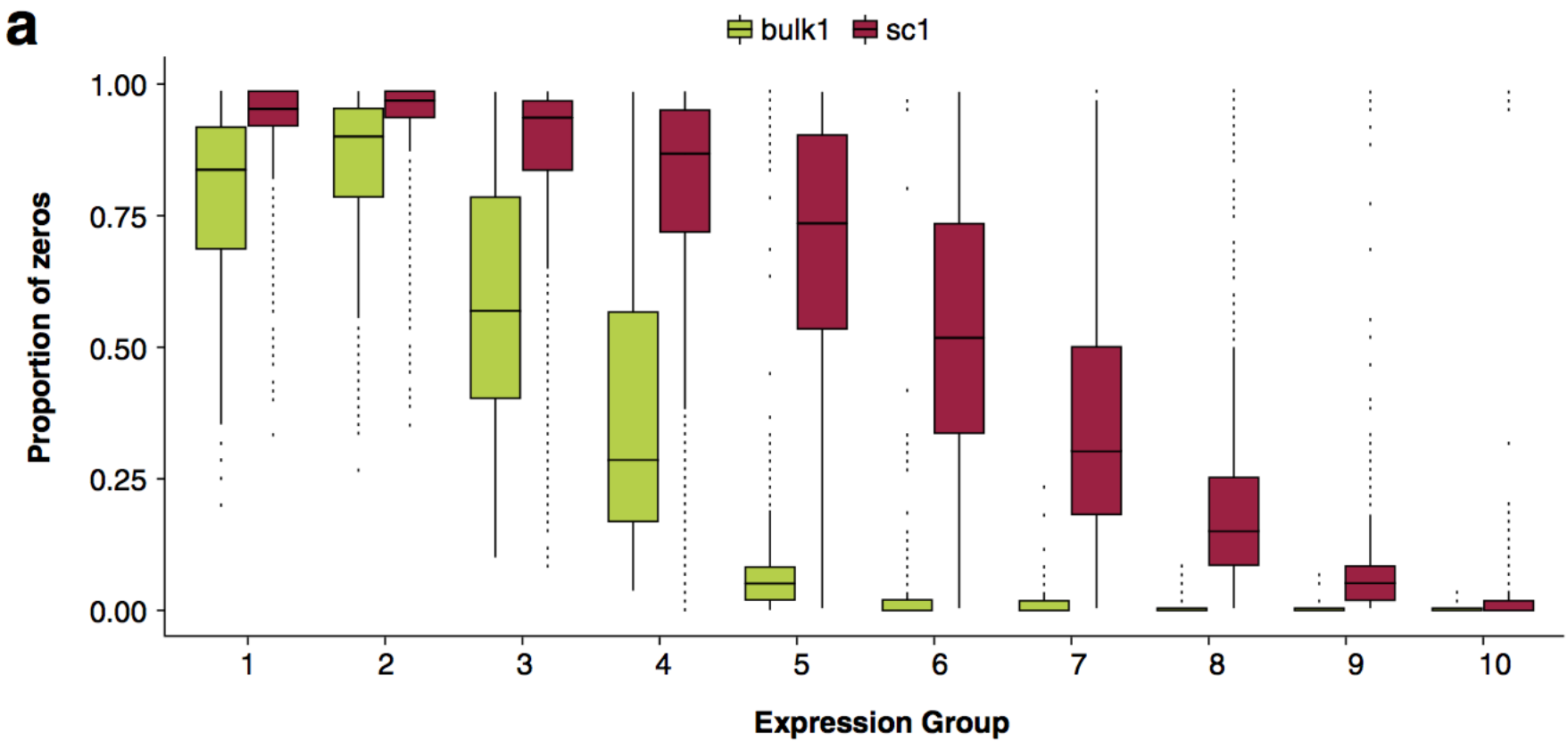


A single device has three input ports (oil, barcoded beads in lysis buffer, and cells of interest) and a single output port used for collecting bead–cell-containing lipid droplets. Then each cell (or RNA in the cell) is marked by the unique barcode and processed on the bead for sequencing

How does single-cell data differ from bulk RNA-seq

- Even with the most sensitive platforms, the data are relatively sparse owing to a high frequency of dropout events (lack of detection of specific transcripts)
- The numbers of expressed genes detected from single cells are typically lower compared with population-level ensemble measurements
- The commonly used 'reads per kilobase per million' (RPKM) transcript quantification is biased on a single-cell level, at the very least the 'transcripts per million' (TPM) should be used

Abundance of zeros

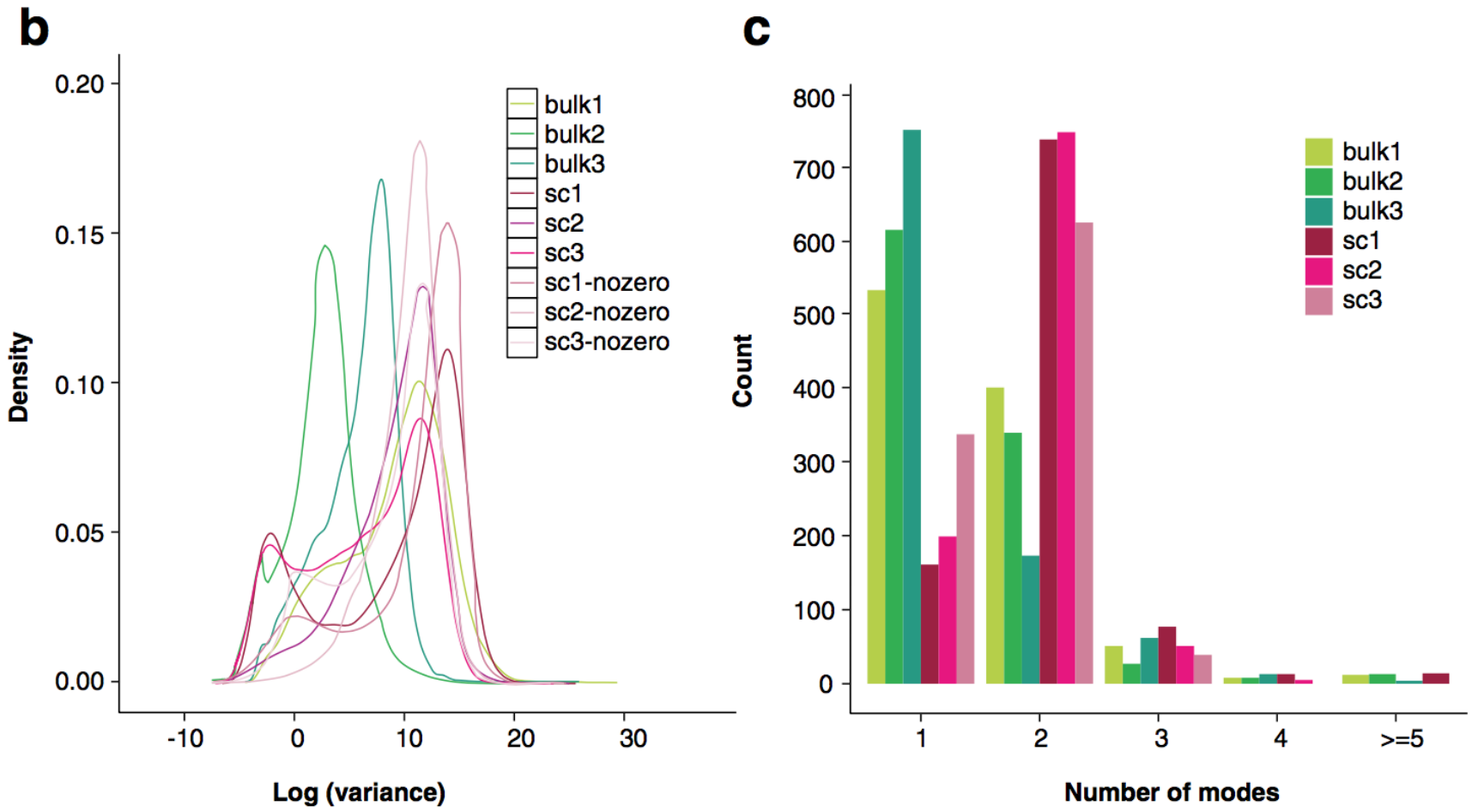


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0927-y

How does single-cell data differ from bulk RNA-seq

- scRNA-seq data, in general, are much more variable than bulk data
- Distributions of transcript quantities are often more complex in single-cell datasets than in bulk RNA-seq - negative binomial or multimodal distributions

Multimodal distribution of variance

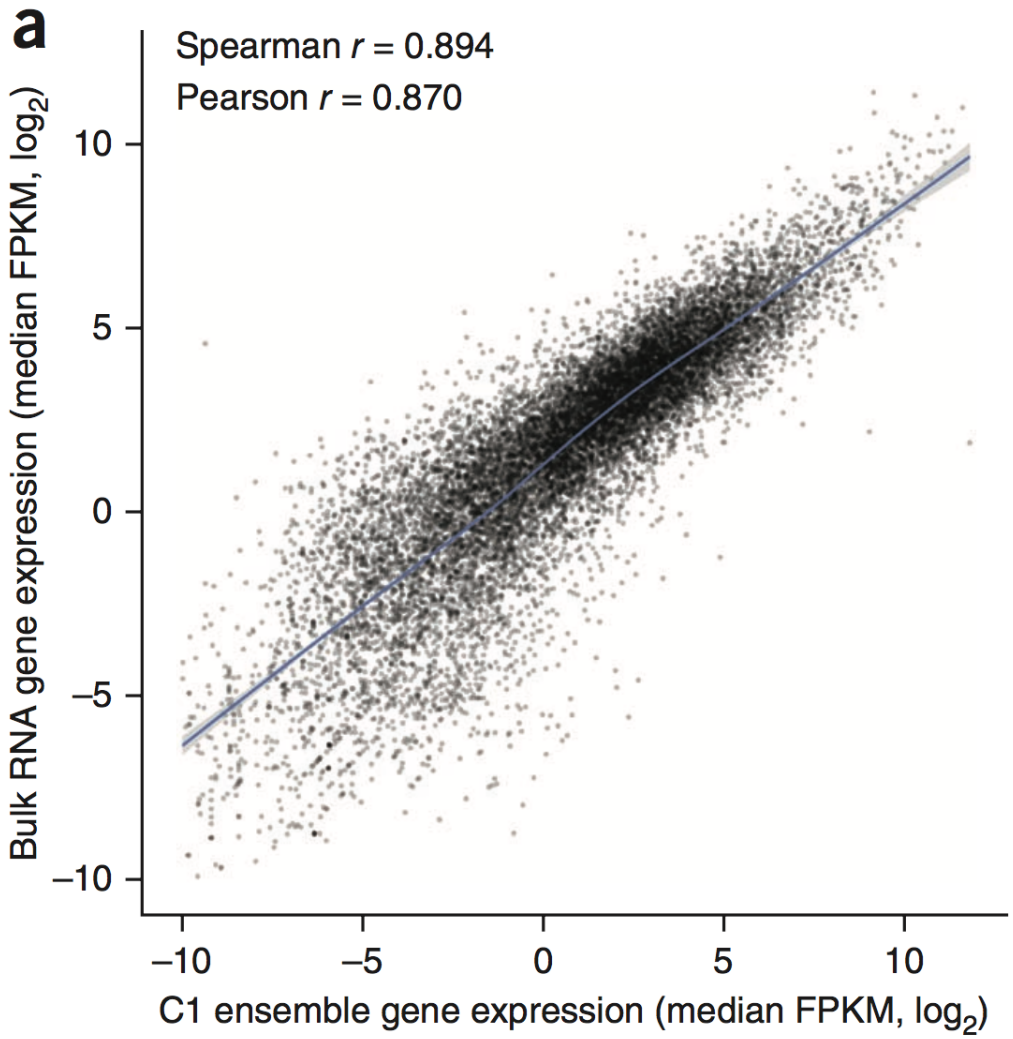


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0927-y

Filtering

- Filter cells and/or genes
- No single consensus, frequently used criteria include:
  - relative library size
  - number of detected genes
  - fraction of reads mapping to mitochondria-encoded genes or synthetic spike-in RNAs

Correlation with regular RNA-seq data



https://www.nature.com/nmeth/journal/v11/n1/full/nmeth.2694.html

scRNA-seq design considerations

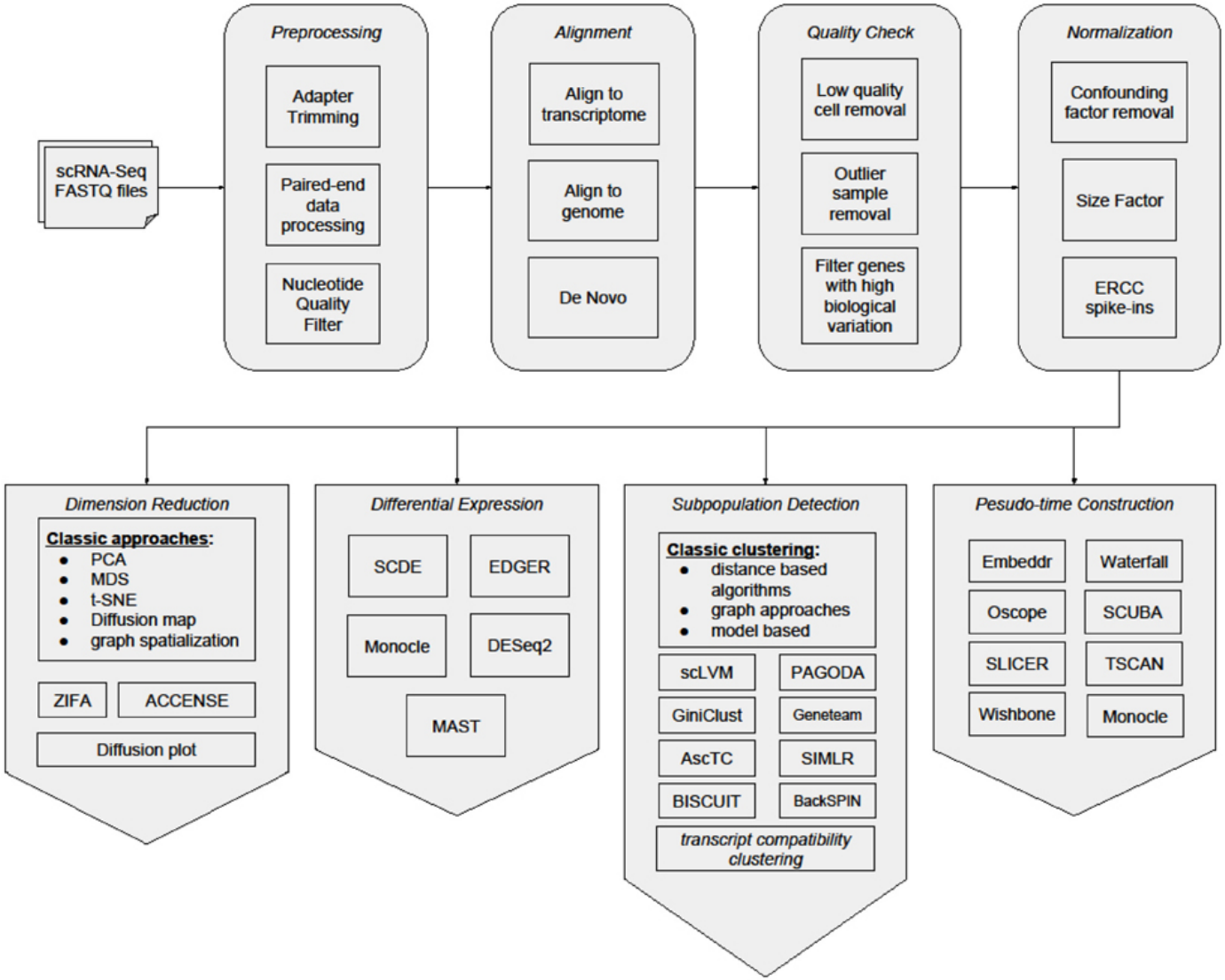
Same as for RNA-seq:

- Randomize batch effects
- Spike-ins (debatable), or unique molecular identifiers (UMIs)
- Record all sources of variability, check for confounding with the main effect

Low amount of starting material

- ~500,000 to 1M reads per cell (sometimes less (~50,000) reads is sufficient for cell classification [Pollen AA et.al. Nat. Biotechnol. 2014]) vs. 20-30M reads in bulk RNA-seq

Single cell workflow



Noise in scRNA-seq

- Technical noise can be approximated with Poisson distribution
- Low–read count genes show strong noise and high–read count genes show weak noise

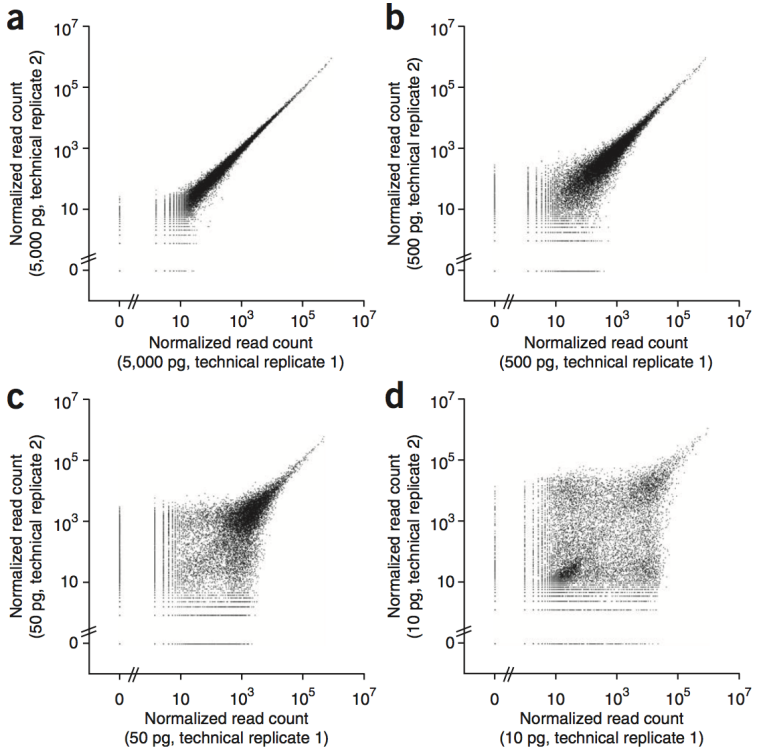


Figure 1 | Dilution series of total *A. thaliana* RNA. (**a**–**d**) Experiments with 5,000 pg (**a**), 500 pg (**b**), 50 pg (**c**) and 10 pg (**d**) of total RNA.

Drop-out rate

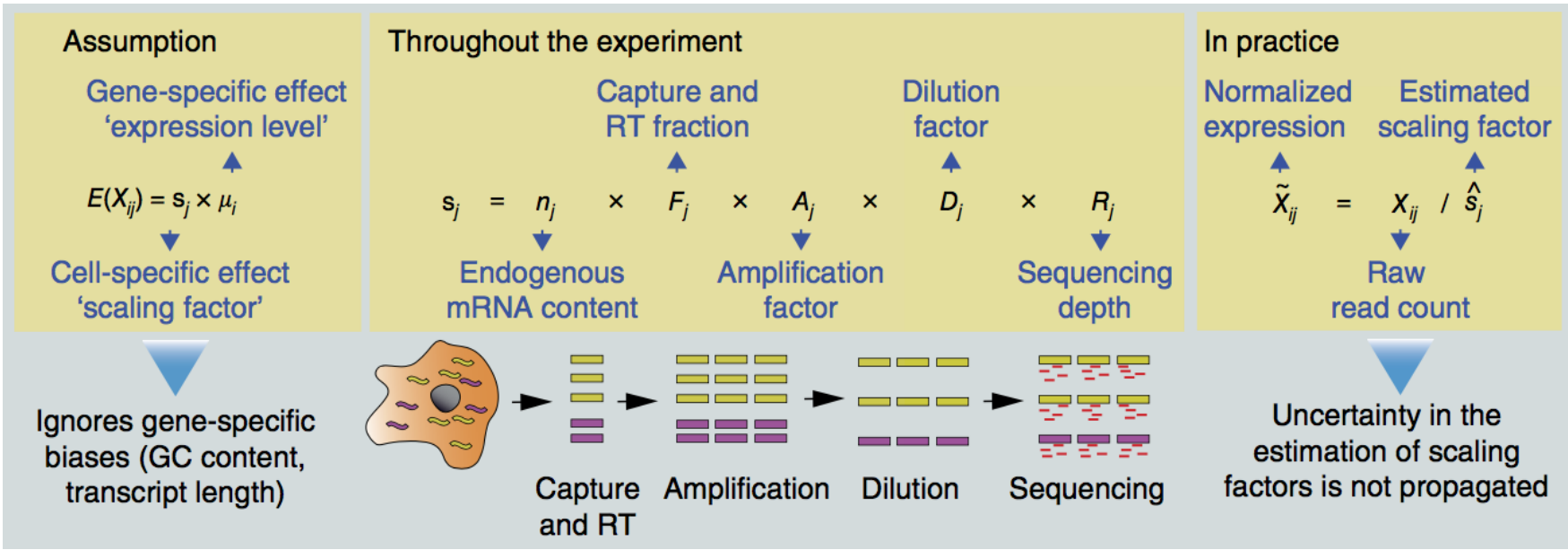
- Depends on the expected expression magnitude
- Genes with lower expression magnitude are more likely to be affected by dropout than genes that are expressed with greater magnitude

Normalization

Ideally, normalize for

- capture efficiency
- amplification biases
- GC content
- Differences in total RNA content
- sequencing depth (that's what is done in reality)

Global-scaling normalization



Vallejos, Catalina A, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. "Normalizing Single-Cell RNA Sequencing Data: Challenges and Opportunities." Nature Methods 14, no. 6 (May 15, 2017): 565–71. https://doi.org/10.1038/nmeth.4292.

Between-sample normalization

TPM or RPKM/FPKM (within-cell normalization) is insufficient - between-sample normalization is needed

- **Median normalization** - identify relatively stable genes to calculate global scaling factors (one for each cell, common across genes in the cell)
- **Spike-in based normalization** - estimate global rescaling factors from known spike-in concentration

Spike-in sequences and normalization

- A set of RNA standards for RNA-seq
  - 92 polyadenylated transcripts that mimic natural eukaryotic mRNAs
  - Designed to have a wide range of lengths (250–2,000 nucleotides) and GC-contents (5–51%) and can be spiked into RNA samples before library preparation at various concentrations (106-fold range)
- External RNA Control Consortium (ERCC) spike-in controls can be used for normalization in the context of a global expression shift
  - Count the number of cells in each sample
  - Add the ERCC spike-in sequences to each sample in proportion to the number of cells
  - Normalize read counts based on cyclic loess robust local regression on the spike-in counts

Baker, S.C. et al. The external RNA controls consortium: a progress report. Nat. Methods 2, 731–734 (2005).

Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011).

Loven, J. et al. Revisiting global gene expression analysis. Cell 151, 476–482 (2012).

SCnorm - normalization for single-cell data

- Quantile regression to estimate the dependence of transcript expression on sequencing depth for every gene
- Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group
- Within-group adjustment for sequencing depth is then performed using the estimated scale factors to provide normalized estimates of expression

https://www.biostat.wisc.edu/~kendzior/SCNORM/

Bacher, Rhonda, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. "SCnorm: Robust Normalization of Single-Cell RNA-Seq Data." Nature Methods 14, no. 6 (April 17, 2017): 584–86. https://doi.org/10.1038/nmeth.4263.

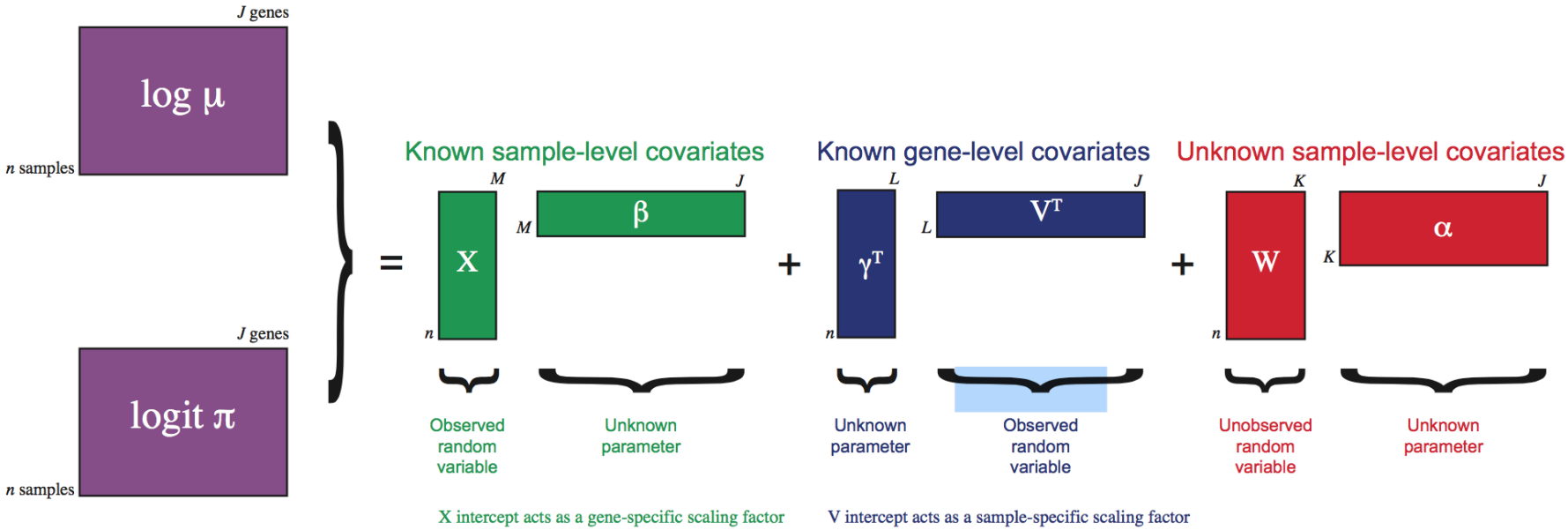
ZINB-WaVE

- Zero-inflated negative binomial model for normalization, batch removal and dimensionality reduction
- Extends the RUV model with more careful definition of "unwanted" variation as it may be biological

https://bioconductor.org/packages/release/bioc/html/zinbwave.html

Davide Risso et al., "ZINB-WaVE: A General and Flexible Method for Signal Extraction from Single-Cell RNA-Seq Data," BioRxiv, January 1, 2017, https://doi.org/10.1101/125112.

ZINB-WaVE



- $\mu_{ij} = E[Y_{ij}|Z_{ij} = 0, X, V, W]$
  - Y_{ij} is the count of gene j $(j = 1, ..., J)$ for cell i $(i = 1, ..., n)$
  - Z_{ij} an unobserved indicator variable, equal to one if gene j is a dropout in cell i and zero otherwise
- $\pi_{ij} = Pr(Z_{ij} = 1|X, V, W)$
- Model $ln(\mu)$ and $logit(\pi)$ with the regression as shown. Both models allow for covariate inclusion

ZINB-WaVE

- PMF of the negative binomial distribution with mean μ and inverse dispersion parameter θ

$$f_{NB}(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}^{\theta}\right) \left(\frac{\mu}{\mu+\theta}^{y}\right)$$

- Its variance $\sigma^2 = \mu + \frac{\mu^2}{\theta} = \mu + \phi\mu^2$, given the dispersion parameter $\phi = \theta^{-1}$ (when $\phi = 0$, NB = Poisson)
- The PMF fo the zero-inflated negative binomial. For any $\pi \in [0,1]$ - the probability that a 0 is ovserved instead of the actual counts - we have an inflation of zeros compared to the NB distribution

$$f_{ZINB}(y;\mu,\theta,\pi) = \pi\delta_0(y) + (1-\pi)f_{NB}(y;\mu,\theta)$$

$\delta_0(.)$ is the Dirac function

ZINB-WaVE

Estimate the parameters from the following regression models:

$$ln(\mu_{i,j}) = (X\beta_\mu + (V\Gamma_\mu))^T + W\alpha_\mu + O_\mu)_{i,j}$$

$$logit(\pi_{i,j}) = ln(\frac{\pi_{i,j}}{1-\pi_{i,j}}) = (X\beta_\mu + (V\Gamma_\mu))^T + W\alpha_\mu + O_\mu)_{i,j}$$

$$ln(\theta_{i,j}) = \zeta_j$$

ζ is a vector of gene-specific dispersion parameters

https://bioconductor.org/packages/release/bioc/html/zinbwave.html

Davide Risso et al., "ZINB-WaVE: A General and Flexible Method for Signal Extraction from Single-Cell RNA-Seq Data," BioRxiv, January 1, 2017, https://doi.org/10.1101/125112.

Sub-population identification

Standard methods used in RNA-Seq

- **Hierarchical clustering**, **PCA**, **tSNE** of highly variable, or differentially expressed, genes. Zeros can be a problem
- **ZIFA** - Zero-inflated dimensionality reduction algorithm for single-cell data
- **SNN-Cliq** - A clustering method for high dimensional dataset. Rank-based (not expression) similarity

https://github.com/epierson9/ZIFA

http://bioinfo.uncc.edu/SNNCliq/

Many more at https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0927-y

Differentially expressed genes

- Need to accomodate unobserved dropouts, bimodality in expression levels due to abundance of zero or low values (**MAST**, **SCDE**)
- **scDD** - Distinguishes four types of differential expression changes to increase power:
  - shifts in unimodal distribution
  - differences in the number of modes
  - differences in the proportion of cells within modes
  - combination of the previous two

https://github.com/kdkorthauer/scDD

SCDE - a Bayesian approach to single-cell differential expression detection

- A two-component mixture model to capture drop-out events (modeled by low-magnitude Poisson) and events where a transcript is faithfully amplified (Negative Binomial)
- Incorporates evidence from other cells to estimate both the likelihood of a gene being expressed in each subpopulation of cells and the likelihood of expression fold change between them

https://hms-dbmi.github.io/scde/index.html

Kharchenko, Peter V., Lev Silberstein, and David T. Scadden. "Bayesian Approach to Single-Cell Differential Expression Analysis." Nature Methods 11, no. 7 (July 2014): 740–42. https://doi.org/10.1038/nmeth.2967.

SCDE - a Bayesian approach to single-cell differential expression detection

The posterior probability of a gene being expressed at an average level x in a subpopulation of cells S is determined as an expected value (E) as:

$$p_S(x) = E\left[\prod_{c \in B} p(x|r_c, \Omega_c)\right]$$

where B is a bootstrap sample of S, and $p(x|r_c, \Omega_c)$ is the posterior probability for a given cell c, as:

$$p(x|r_c, \Omega_c) = p_d(x)p_{Poisson}(x) + (1 - p_d(x))p_{NB}(x|r_c)$$

where p_d is the probability of obseerving a dropout event, $p_{Poisson}(x)$ and $p_{NB}(x|r_c)$ are the probabilities of observing expression magnitude of r_c in case of a dropout (Poisson) or successful amplification (NB) for a gene expressed at an average level x in S in cell c with the parameters of the

SCDE - a Bayesian approach to single-cell differential expression detection

- For the differential expression analysis, the posterior probability that the gene shows a fold expression difference of f between subpopulations S and G was evaluated as:

$$p(f) = \sum_{x \in X} p_S(x) p_G(fx)$$

where x is the valid range of expression levels

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

- A two-part generalized linear model (hurdle model) explicitly parameterizing expressed and non-detectable gene distributions
- Includes as a covariate the fraction of genes that are detectably expressed in each cell as a proxy for both technical and biological sources of variation (CDR). For cell i, $CRD_i = 1/N \sum_{g=1}^{N} z_{ig}$, where z_{ig} is an indicator if gene g in cell i is expressed above background
- The expression measure of a detected gene is modeled by linear regression and the probability of detection by logistic regression

https://github.com/RGLab/MAST

Pseudotemporal ordering

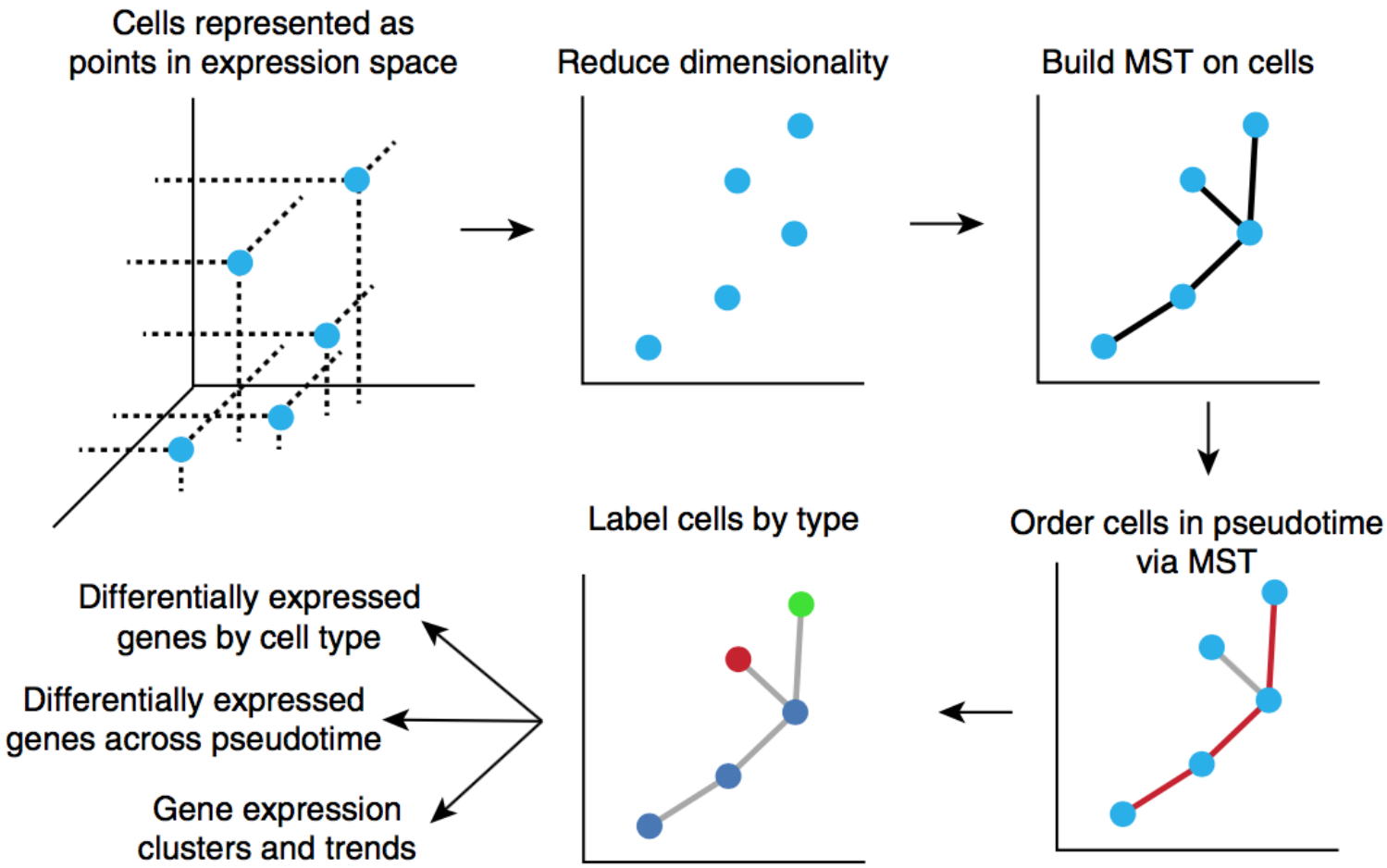
- Idea - cells at different differentiation (or other biological process) stage are presented with different expression profiles
- Dynamics of cellular processes can be reconstructed from expression profiles
- Key assumption: genes do not change direction very often, thus samples with similar transcriptional profiles should be close in order
- Most approaches are dimensionality reduction-based, and apply graph theory designed to traverse nodes in a graph efficiently
- **Monocle** - Independent component analysis, then a minimum spanning three through the dimension-reduced data

https://cole-trapnell-lab.github.io/monocle-release/

Many more at https://github.com/agitter/single-cell-pseudotime

Monocle, An analysis toolkit for single-cell RNA-seq

Single-cell trajectories, clustering, visualization, differential expression



https://cole-trapnell-lab.github.io/monocle-release/

Slingshot

- Inferring multiple developmental lineages from single-cell gene expression
- Clustering by gene expression, then inferring cell lineage as an ordered set of clusters - minimum spanning tree through the clusters using Mahalanobis distance
- Initial state and terminal state specification
- Principal curves to draw a path through the gene expression space of each lineage

https://github.com/kstreet13/slingshot

Single-cell network analysis

- `SCENIC` R package - single-cell network reconstruction and cell-state identification. Three modules:
  1. `GENIE3` - Connect co-expressed genes and TFs using random forest regression;
  2. `RcisTarget` - Refine them using cis-motif enrichment;
  3. `AUCell` - assign activity scores for each network in each cell type.

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. "SCENIC: Single-Cell Regulatory Network Inference and Clustering." Nature Methods 14, no. 11 (November 2017): 1083–86. https://doi.org/10.1038/nmeth.4463.

https://gbiomed.kuleuven.be/english/research/50000622/lcb/tools/scenic

https://github.com/aertslab/SCENIC

https://github.com/aertslab/GENIE3

https://github.com/aertslab/AUCell

ZIFA - dimensionality reduction for zero-inflated

- Given the mean level of non-zero expression (log read count) μ and the dropout rate for that gene p_0, model the dropout as $p_0 = exp(-\lambda\mu^2)$, where λ is a fitted parameter, based on a double exponential function
- EM algorithm that incorporates imputation step for the expected gene expression level of drop-outs

https://github.com/epierson9/ZIFA

Pierson, Emma, and Christopher Yau. "ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis." Genome Biology 16 (November 2, 2015): 241. https://doi.org/10.1186/s13059-015-0805-z.