Understanding drop-outs in single-cell UMI:
two papers with different approaches

Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

K Choi, Y Chen, DA Skelly, GA Churchill

Demystifying "drop-outs" in single-cell UMI data

TH Kim, X Zhou, M Chen

CSE 590C Fall 2020

October 19th, 2020

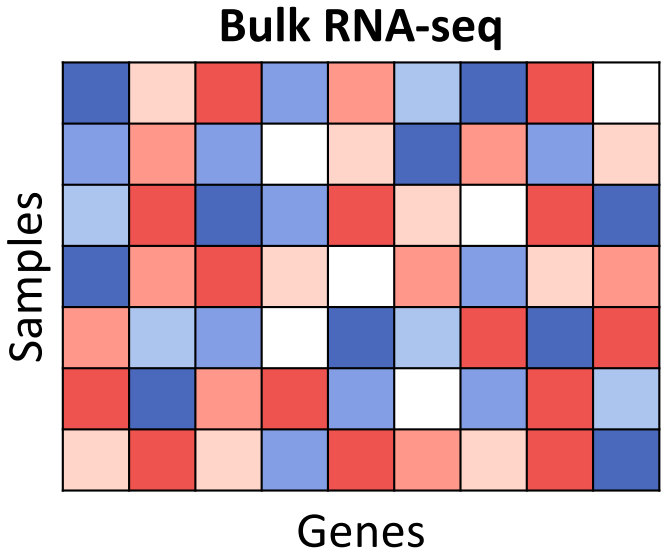
Ayse Dincer & Walter L. Ruzzo

Single-cell RNA sequencing (scRNA-seq)

Genotype ⟷ Phenotype

A challenge in biology and medicine

Transcriptomes can be informative

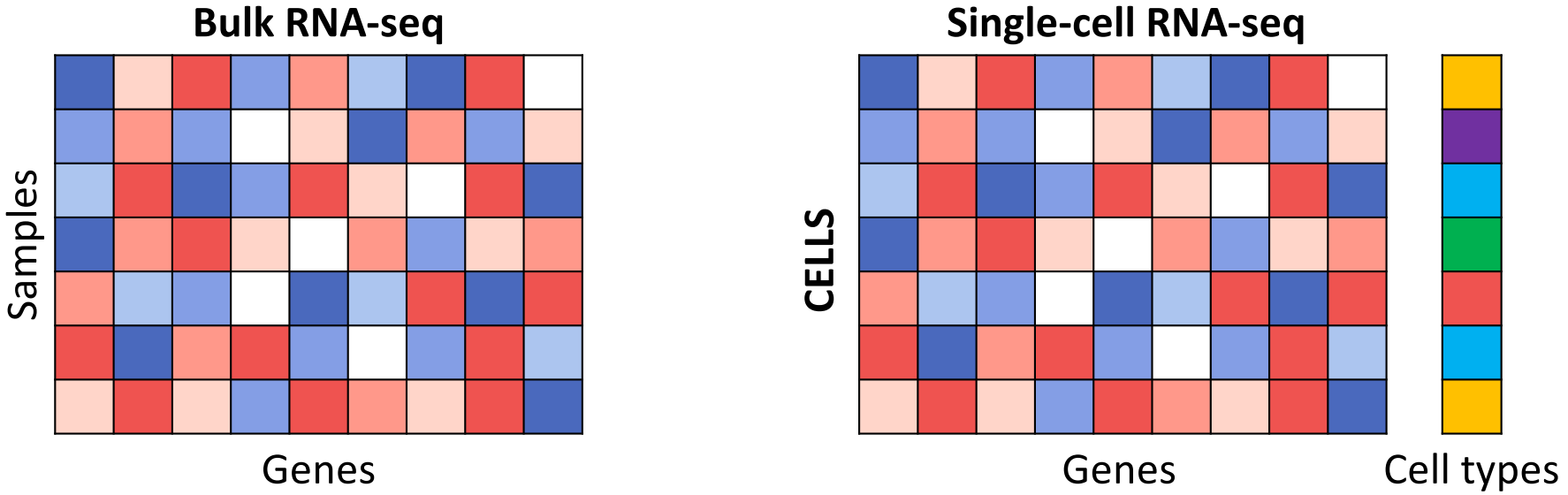


- Bulk population sequencing can provide only the average expression signal for an ensemble of cells
- However, diverse cell types in our body each express a unique transcriptome

Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

Single-cell RNA sequencing (scRNA-seq)

We need a more precise understanding of the transcriptome in individual cells



Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

Single-cell RNA sequencing (scRNA-seq)

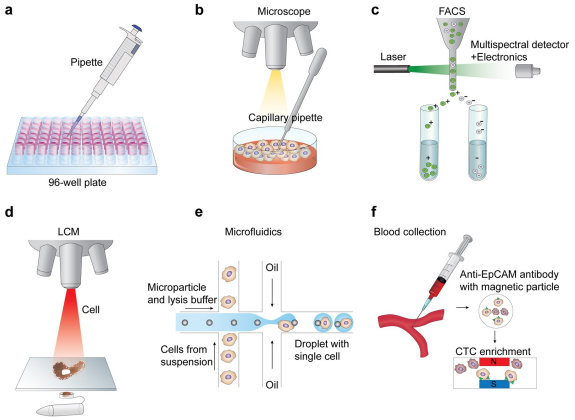
- Pioneered by James Eberwine et al. and Iscove et al.
- First analysis in 2009 by Tang et al.
  - characterization of cells from early developmental stages

- Many studies followed:
  - Identify rare cell populations
  - Characterize outlier cells to understand drug resistance and relapse in cancer treatment
  - Detect diverse immune cell populations
  - Understand cell lineage relationships in early development

Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

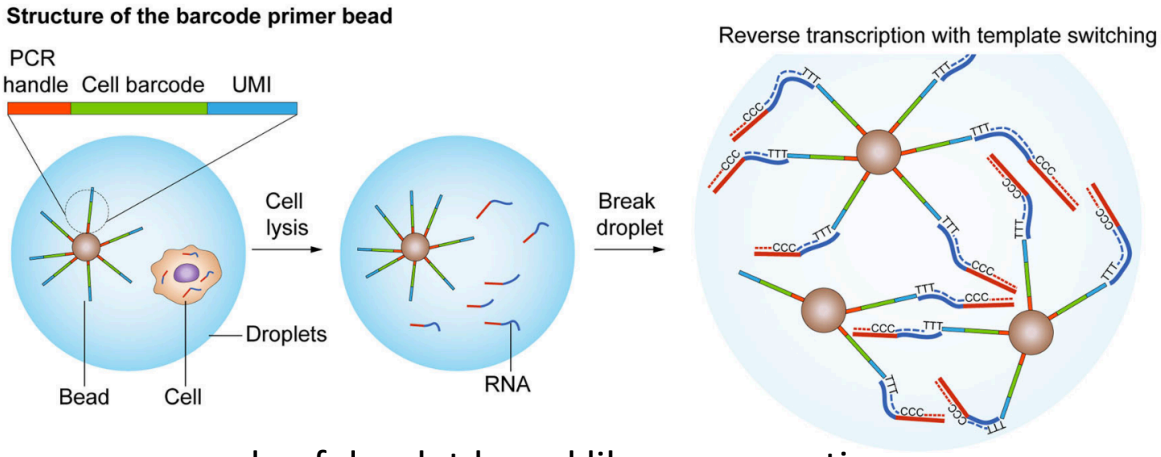
scRNA-seq Technology

First step: single-cell isolation



Many techniques exist to isolate cells

Second step: generation of scRNA-seq libraries



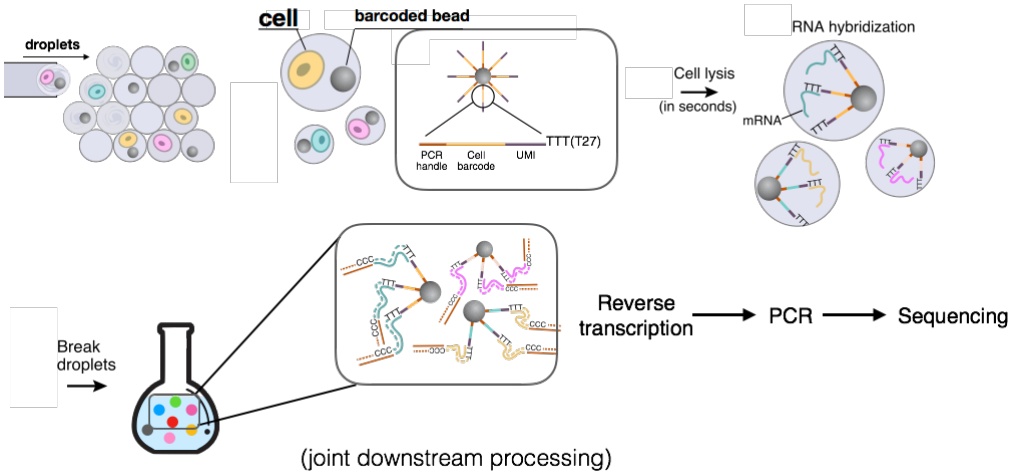
example of droplet-based library generation

Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

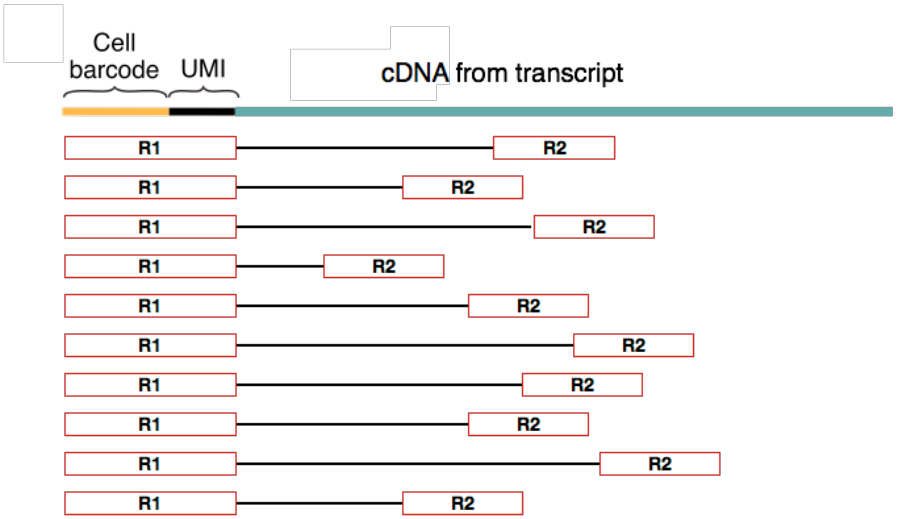
scRNA-seq Technology: What is UMI?

"Unique molecular identifiers (UMI) are molecular tags that are used to detect and quantify unique mRNA transcripts"

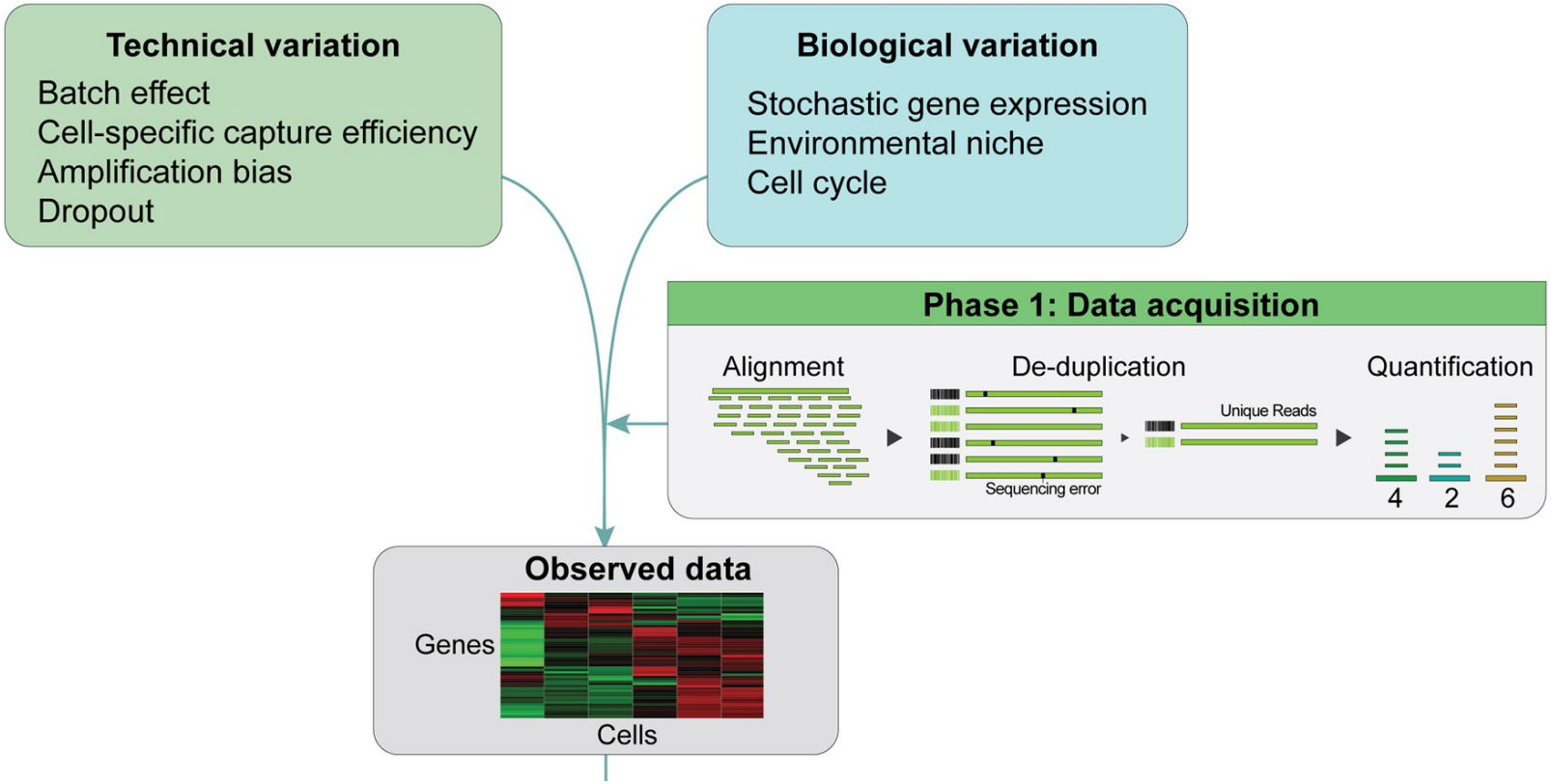
Drop-Seq workflow



Paired-end reads

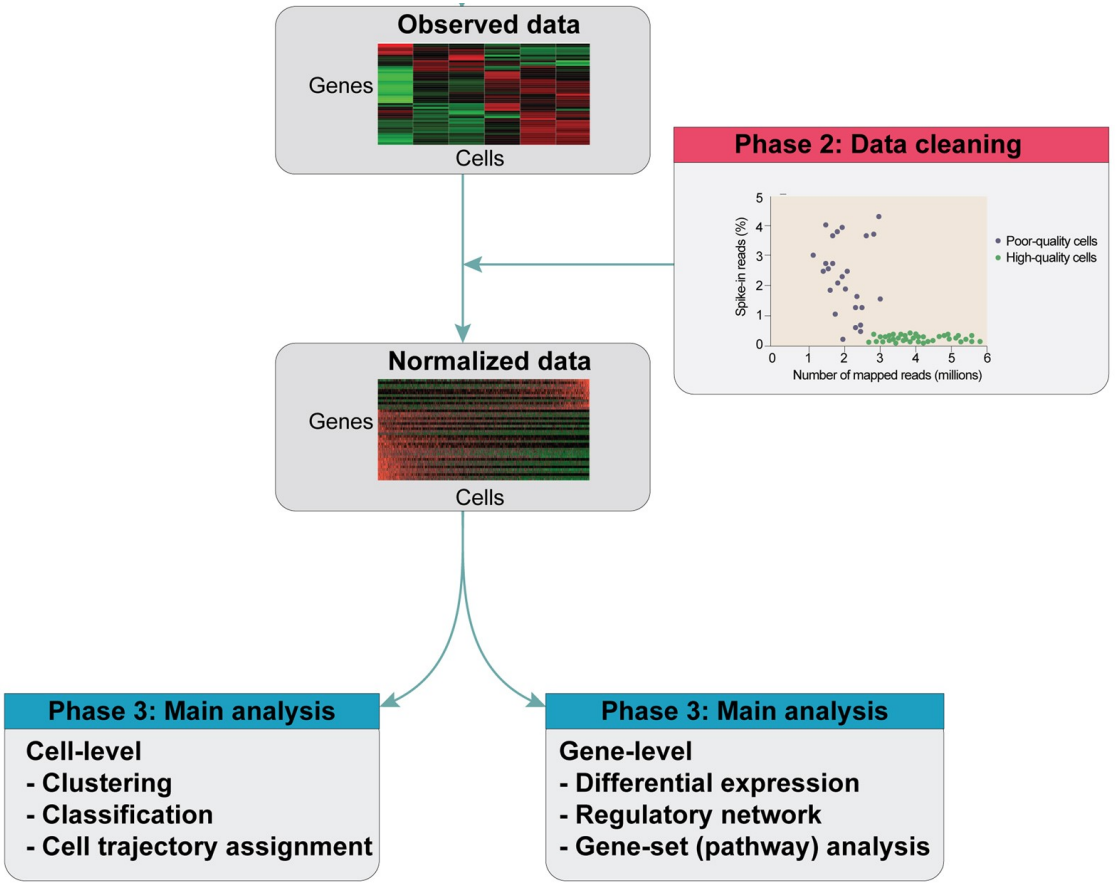


scRNA-seq: Computational pipeline



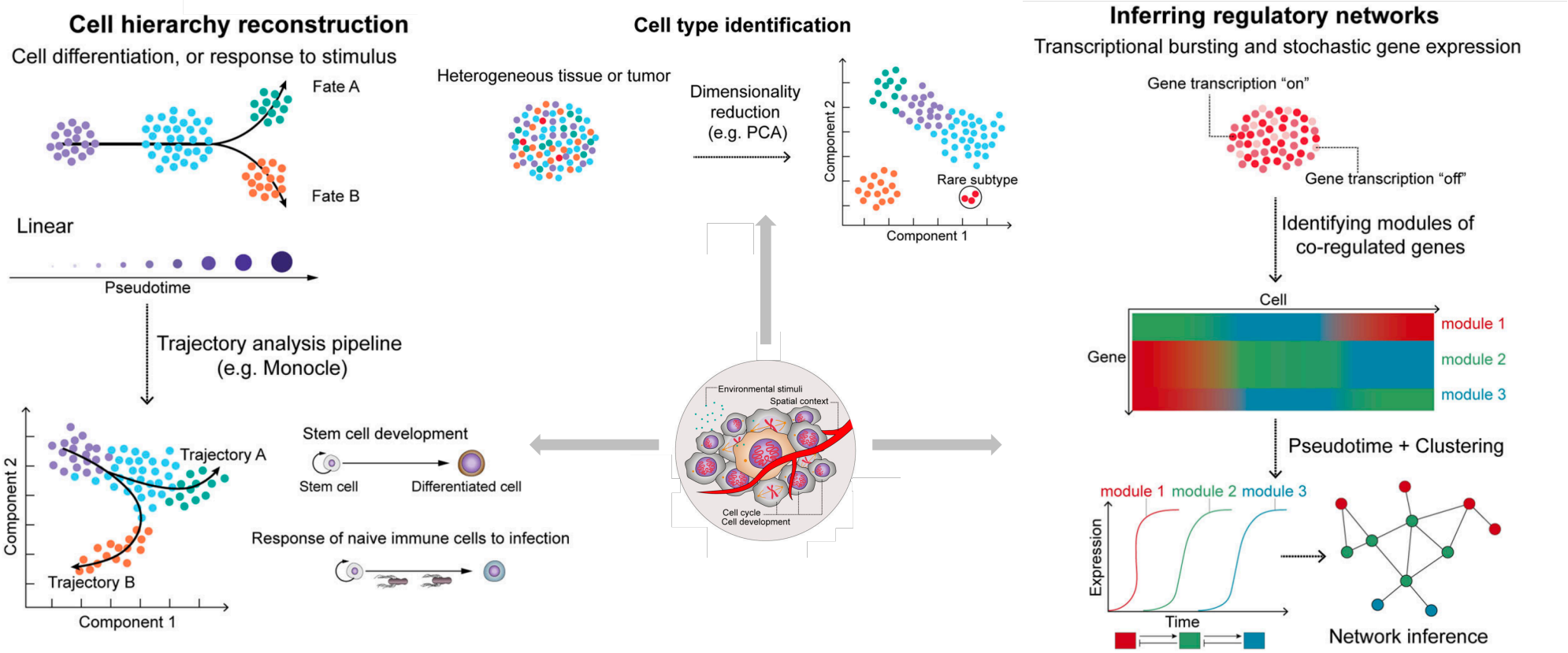
Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

scRNA-seq: Computational pipeline



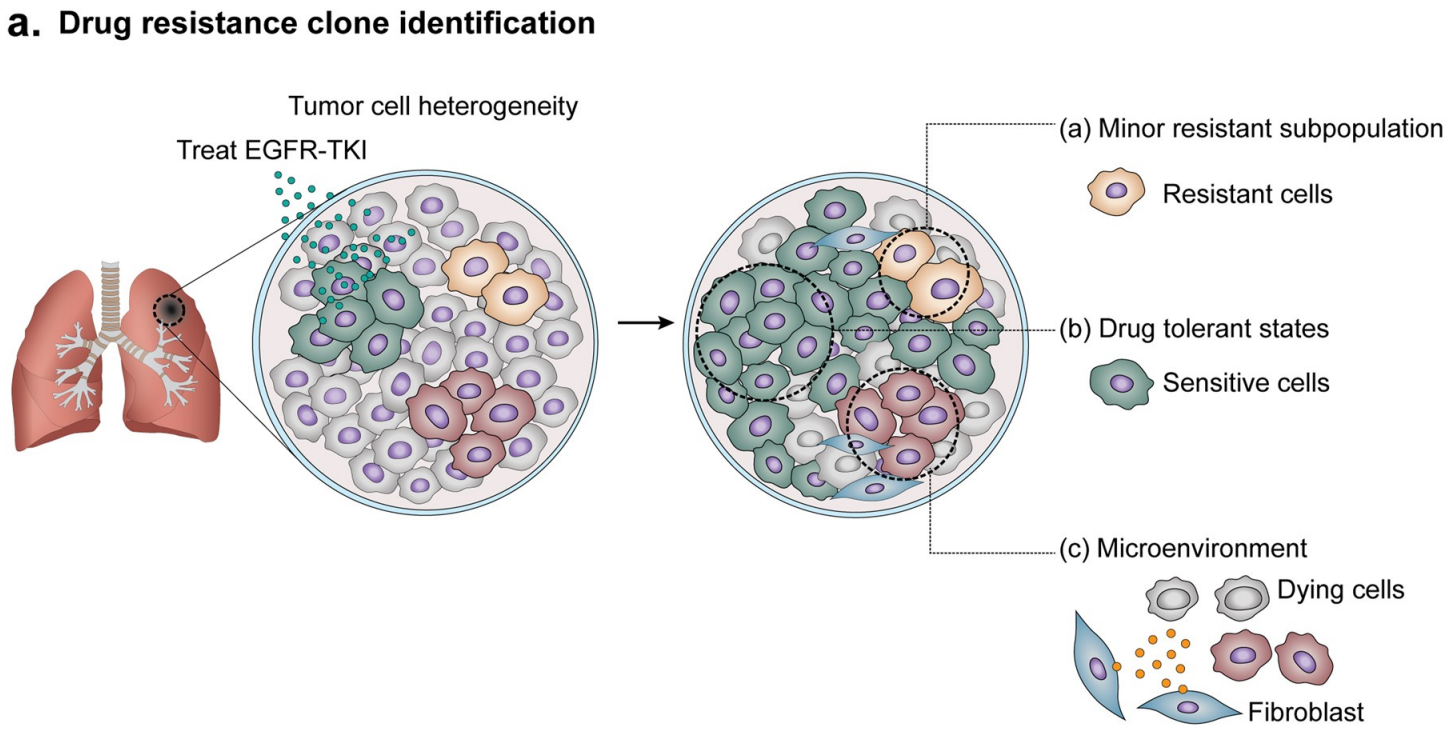
Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

scRNA-seq Applications



Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

scRNA-seq Applications



Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50,** 96 (2018).

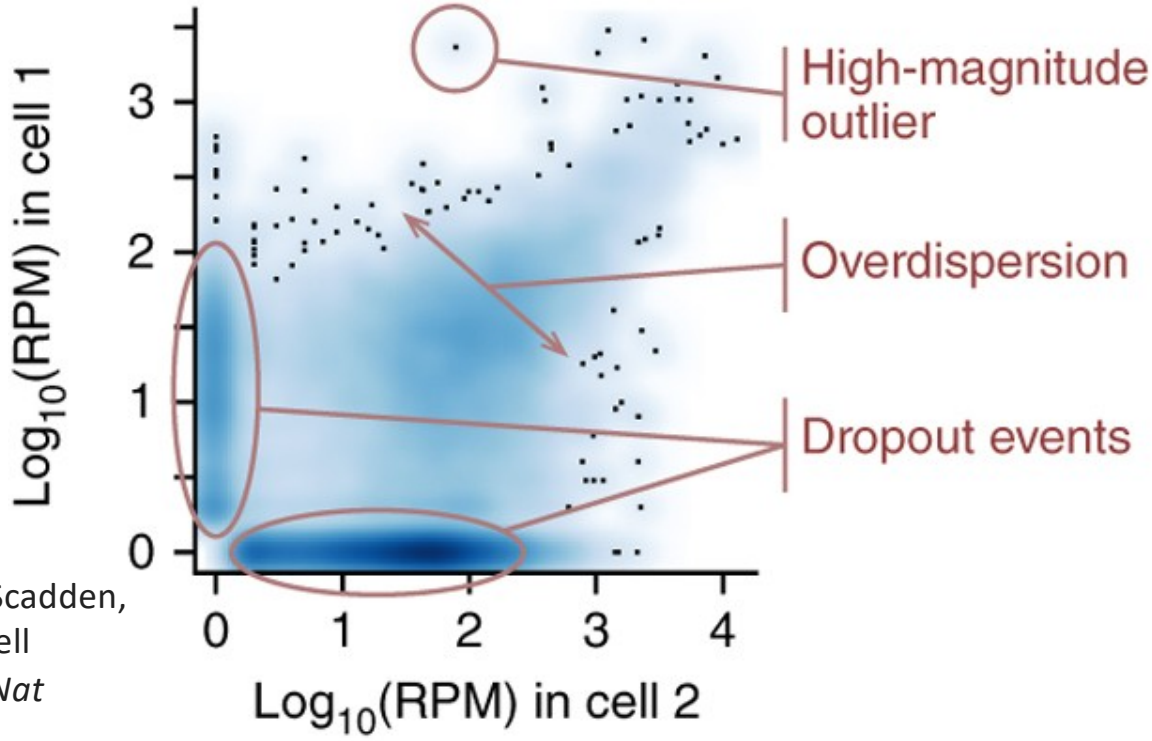
Single-cell RNA sequencing (scRNA-seq)

- Single-cell RNA sequencing is a very promising technology
- It can allow new biological insights

- Yet it also presents many technical and computation challenges
- One problem we will focus on today is **drop-out or zero-inflation**

What is dropout in single cell?

a gene is observed at a moderate or high expression level in one cell but is not detected in another cell



Kharchenko, P., Silberstein, L. & Scadden, D. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11,** 740–742 (2014).

There are many many different approaches

scDoc: correcting drop-o
RNA-seq data

Di Ran, Shanshan Zhang, Nicholas Lytal, Li

Bioinformatics, Volume 36, Issue 15, 1 Augu

nature COMMUNICATIONS

atics

ARTICLE

DOI: 10.1038/s41467-018-03405-7 OPEN

ccess

An accurate and robust imputation method scImpute for single-cell RNA-seq data

Check for updates

Wei Vivian Li [1] & Jingyi Jessica Li [1,2]

Droplet scRNA-seq is not zero-inflated

Gong *et al. BMC Bioinformatics* (2018) 19:220
https://doi.org/10.1186/s12859-018-2226-y

BMC Bioinformatics

chenko[1–3], Lev Siberstein[3–5] &

mputation method to reduce
-cell RNA sequencing

METHODOLOGY ARTICLE Open Access

CrossMark

DrImpute: imputing dropout events in single cell RNA sequencing data

Wuming Gong[†], Il-Youp Kwak[†], Pruthvi Pota, Naoko Koyano-Nakagawa and Daniel J. Garry[*]

2020, Pages 4021–4029,
taa278
ory ▾

Why do dropouts occur in single cell?

There are different views

Why do we observe dropouts?

- technical artifacts
- statistical sampling
- cell type differences
- biological factors

What should we do about them?

- impute before learning
- preprocess/cluster/reduce dimensions
- incorporate technical variates
- incorporate biological variates
  - model zero inflation
  - ignore zero inflation

Today we are going to examine 2 papers

There are two main views

Drop-outs are technical artefacts

Choi *et al. Genome Biology* (2020) 21:183
https://doi.org/10.1186/s13059-020-02103-2

Genome Biology

RESEARCH Open Access

Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

Kwangbom Choi[1], Yang Chen[2], Daniel A. Skelly[1] and Gary A. Churchill[1*]

To solve drop-outs ->
Take cell type heterogeneity and biological covariates into account

Drop-outs are related to biological signals

Kim *et al. Genome Biology* (2020) 21:196
https://doi.org/10.1186/s13059-020-02096-y

Genome Biology

RESEARCH Open Access

Demystifying "drop-outs" in single-cell UMI data

Tae Hyun Kim[1], Xiang Zhou[2*] and Mengjie Chen[3*]

To detect cell type heterogeneity ->
Use drop-out rates

Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

Paper 1

Short summary of paper 1

- They apply a Bayesian model selection approach to demonstrate zero inflation in multiple biologically realistic scRNA-seq datasets
- They show that the primary causes of zero inflation are not technical but rather biological in nature
- They recommend the negative binomial count distribution, not zero-inflated, as a suitable reference model for scRNA-seq analysis

Outline for paper 1

Problem: Potential reasons for zero inflation/dropout

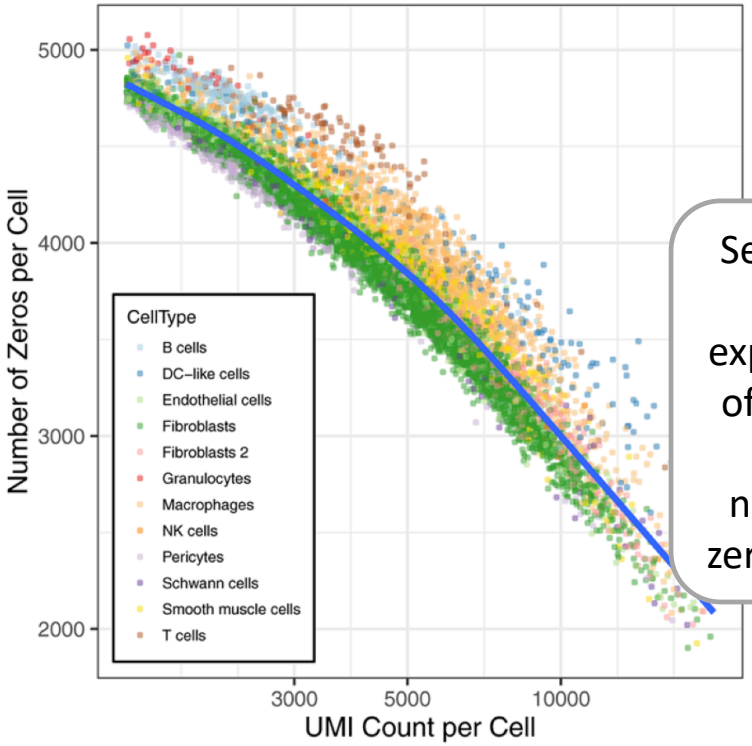
Method: Bayesian model selection approach to identify genes with zero inflation

Results #1: scRATE can identify genes with zero inflation

Results #2: Zero-inflation of genes is highly associated with cell types

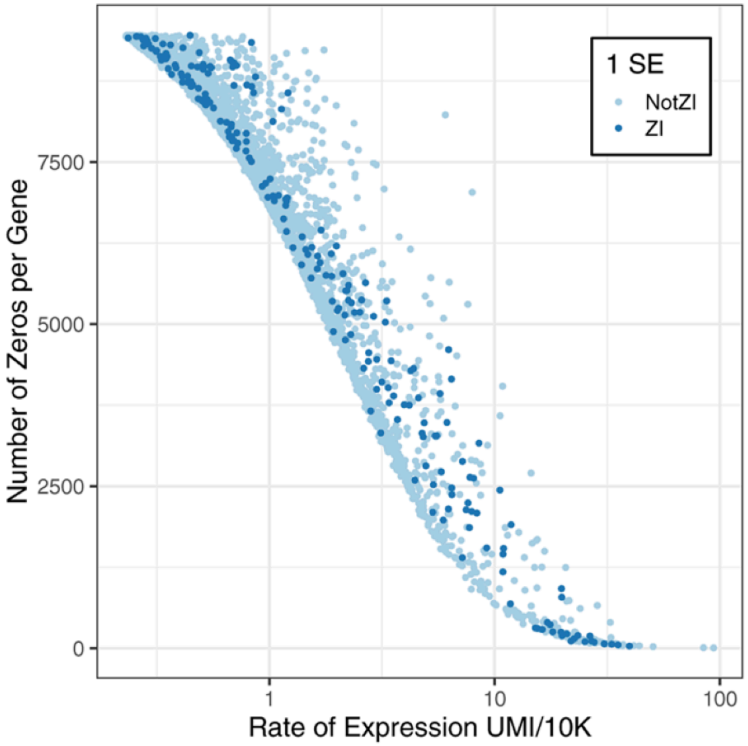
Problem: Why are there so many zeros?

1. Sequencing Depth



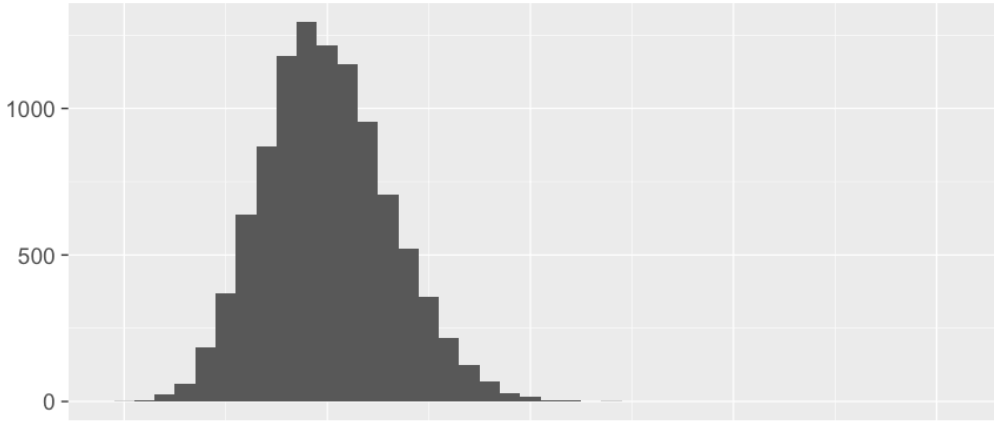
Sequencing depth explains 95% of variation in the number of zeros per cell

2. Per-gene average rate of expression

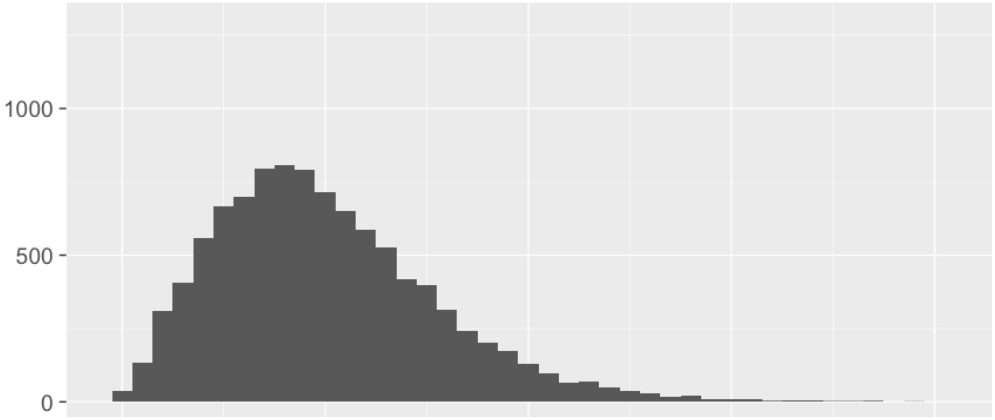


Background: Statistical Models

1. Poisson (P)



2. Negative Binomial (NB)



3. Zero-inflated Poisson (ZIP)

4. Zero-inflated Negative Binomial (ZINB)

Method: Bayesian model selection to identify genes exhibiting zero inflation

What is Bayesian model selection?

- The goal is to select the model that maximizes the likelihood of the observed data
- The probability of the data given the model is computed by integrating over the unknown parameter values in that model:

$$p(D|M) = \int_{\theta} p(D|\theta)p(\theta|M)d\theta$$

http://alumni.media.mit.edu/~tpminka/statlearn/demo/

Method: Bayesian model selection to identify genes exhibiting zero inflation

- Is based on generalized linear models (GLMs)
- Implemented a Bayesian model selection criterion **the expected log predictive density (ELPD)**

$$\mathrm{ELPD} = \sum_{c=1}^{C} \log p(y_c | y_{-c}).$$

denotes LOOCV value for each cell vs. all the other cells

- ELPD score is calculated for four statistical models (P, ZIP, NB, or ZINB)
- scRATE examines all the data, including non-zero counts
- Uses leave-one-out cross-validation, which provides a standard error (SE) to quantify uncertainty in the estimated ELPD scores
- Penalizes both underfitting and overfitting models, a more complex model is selected only when the ELPD is substantially better

Results #1: Model selection can identify genes exhibiting zero inflation

(a) False Positive rates

(b) True Positive rates

Table 1 Error rates and power of `scRATE` classification

Sequencing depth	Threshold		
	0 SE	1 SE	2 SE
(a)			
10k	$0.2349_{\pm 0.0695}$	$0.0325_{\pm 0.0174}$	$0.0014_{\pm 0.0016}$
50k	$0.1837_{\pm 0.0557}$	$0.0206_{\pm 0.0159}$	$0.0009_{\pm 0.0016}$
(b)			
10k	$0.8116_{\pm 0.0365}$	$0.6152_{\pm 0.0312}$	$0.4641_{\pm 0.0160}$
50k	$0.8955_{\pm 0.0158}$	$0.7934_{\pm 0.0176}$	$0.7062_{\pm 0.0165}$

Results #2: Most zero-inflated genes are due to variable expression rates across cell types

Table 2 `scRATE` classification of genes in the heart data

Threshold	Selected model			
	P	NB	ZIP	ZINB
(a)				
0 SE	1111	2930	525	949
1 SE	2112	3183	81	139
2 SE	2930	2509	5	71
3 SE	3445	2035	1	34

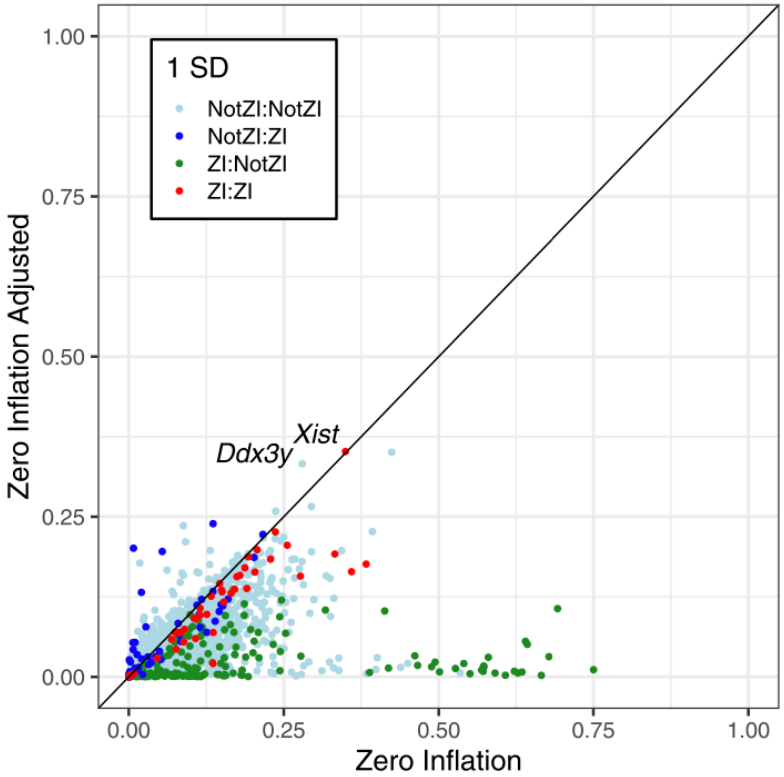
Applied scRATE directly

After accounting for cell type, the number of zero-inflated genes drops

Genes that are no longer ZI vary across cell types

Examples: Col1a2 -> fibroblasts, Ptpn18 -> immune cells

Results #2: Most zero-inflated genes are due to variable expression rates across cell types



Majority of genes were originally classified as ZI are no longer ZI after accounting for cell type

A few of genes remain or become ZI:
female-specific *Xist*
Y-chromosome gene *Ddx3y*

After accounting for sex as an explanatory variable, these genes are no longer ZI

Paper 1

Their conclusions:

- High frequency of zeros does not necessarily imply technical dropout
- Instead, zero inflation is largely explained by biological factors, such as cell type and sex
- Recommend against the practice of replacing zeros in data with imputed non-zero values, could mask biological signals
- Recommend the generalized linear model with negative binomial error, and taking cell types and biological factors as explanatory variables

Paper 1

- Do you think simulation tests make sense?
  - What other simulation experiments can be carried?
  - Do you think simulated data can reflect true patterns?

- Do you prefer to see more real-data experiments and biological covariate examples?

- What are the advantages/disadvantages of this model?
  - Does it make sense that cell type is a determinant of zero-inflation?

Demystifying “drop-outs” in single-cell UMI data

Paper 2

Short summary of paper 2

- Proposed a novel framework HIPPO (Heterogeneity-Inspired Pre-Processing tOol) that leverages zero proportions to explain cellular heterogeneity and integrates feature selection with iterative clustering
- Showed that clustering should be the foremost step of the workflow
- Showed that cell-type heterogeneity can resolve drop-outs, while imputing or normalizing heterogeneous data can introduce unwanted noise

Outline for paper 2

Problem: Potential reasons for zero inflation/dropout

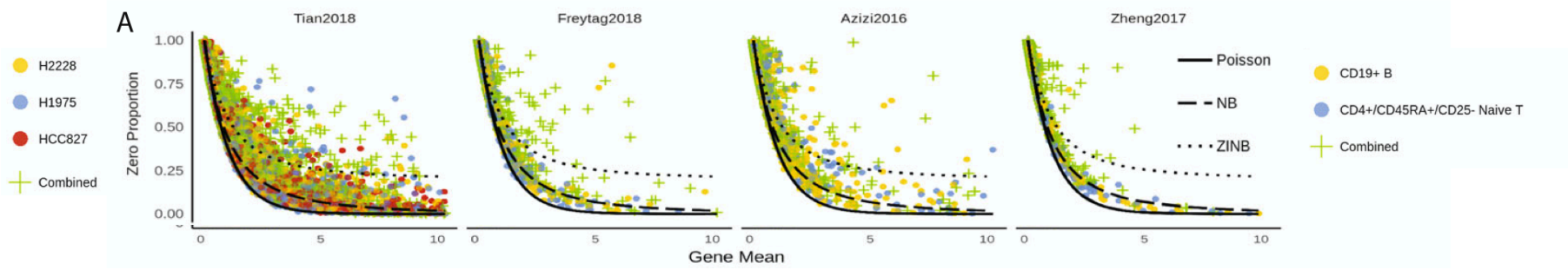
Method: Zero inflation test to detect cellular heterogeneity and HIPPO

Results #1: Zero inflation test is successful at detecting cellular heterogeneity

Results #2: Appropriate pre-processing introduces unwanted noise in the downstream analysis

Results #3: HIPPO can identify cell types

Problem: Demystifying drop-outs

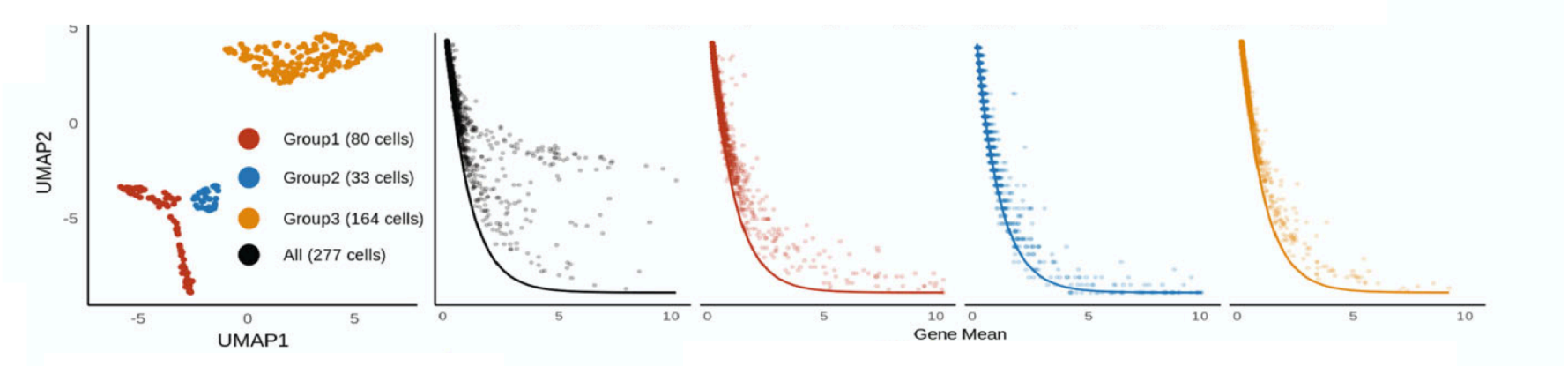


1. For a homogeneous cell population, zero proportions in most genes can be modeled by the Poisson distribution (more than 95% of absolute z values are below 2)

2. For mixed cell types, zero proportions considerably deviate from expected values under the Poisson model (less than 30% of the genes have z values below 2)

Conclusion: Zero-inflation test is an effective way to find genes that contribute to cellular heterogeneity

Problem: Demystifying drop-outs



Conclusion: Zero proportions can be a metric to evaluate cellular heterogeneity and can discern cell types

Method: Zero inflation test for cellular heterogeneity

They developed a new feature selection strategy that uses detected zero proportion of a given gene as the statistic to test for cellular heterogeneity

Framework:

- **Null hypothesis =** assumes complete cellular homogeneity = the proportion of zeros is equal to the expected zero proportion under Poisson distribution
- **Alternative hypothesis** = zero proportion is inflated, as if the count data follows mixture of Poisson distributions

$$H_0: \quad p_g = e^{-\lambda_g},$$
$$H_A: \quad p_g > e^{-\lambda_g}$$

p_g = true zero proportion

Advantages of the framework:

1. Only the proportion of zeros is used
2. Allows each gene to have different grouping structure across cells
3. No complicated modeling

Results #1: Zero inflation test is successful at detecting cellular heterogeneity

Cell population	Gene mean	*z* score
CD34+	25.89	1838203
Subtype 1	0.5625	6.19
Subtype 2	22.36	0
Subtype 3	38.96	0

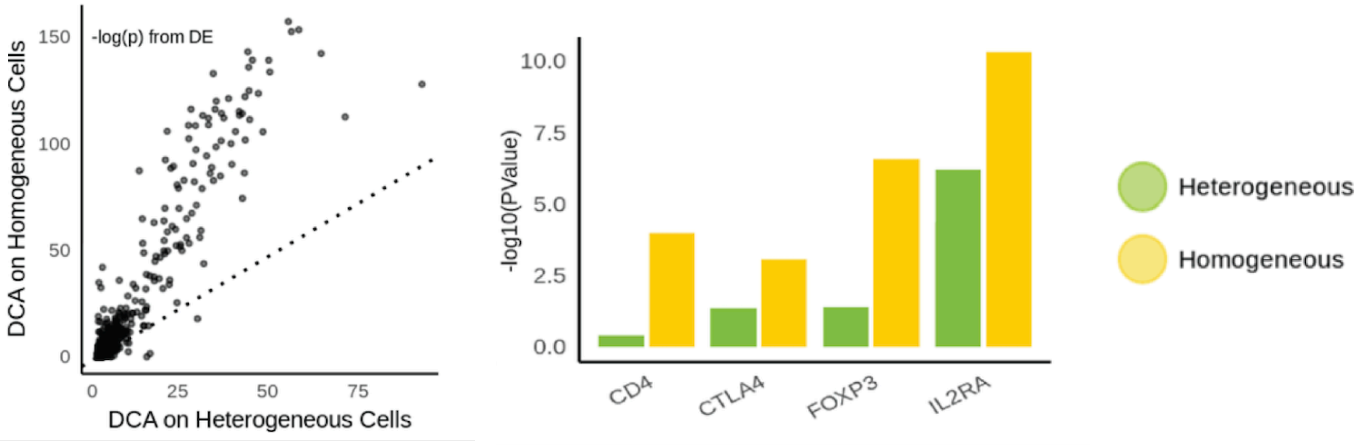
Zero inflation test statistics for PPBP gene in CD34+ cells

- PPBP was identified with a high zero proportion of 26% within CD34+ cells, indicating very high zero inflation
- After they separated CD34+ cells into three subtypes, the test within each subtype is no longer statistically significant

Conclusion: cellular heterogeneity can drive excessive zeros and zero proportions can be used to discern cell types

Results #2: Inappropriate pre-processing introduces unwanted noise in the downstream analysis

A popular pre-processing step is to apply deep learning based de-noising tools (e.g. Deep Count Autoencoder (DCA)) which de-convolute the technical effects from biological effects and impute zero accounts due to drop- outs



Conclusion: imputing the UMI data without resolving cell heterogeneity can lead to loss of important biological information

Method: HIPPO: Heterogeneity-Inspired Pre-Processing tOol

HIPPO integrates the proposed zero inflation test into a hierarchical clustering framework

Step 1 Feature Selection:

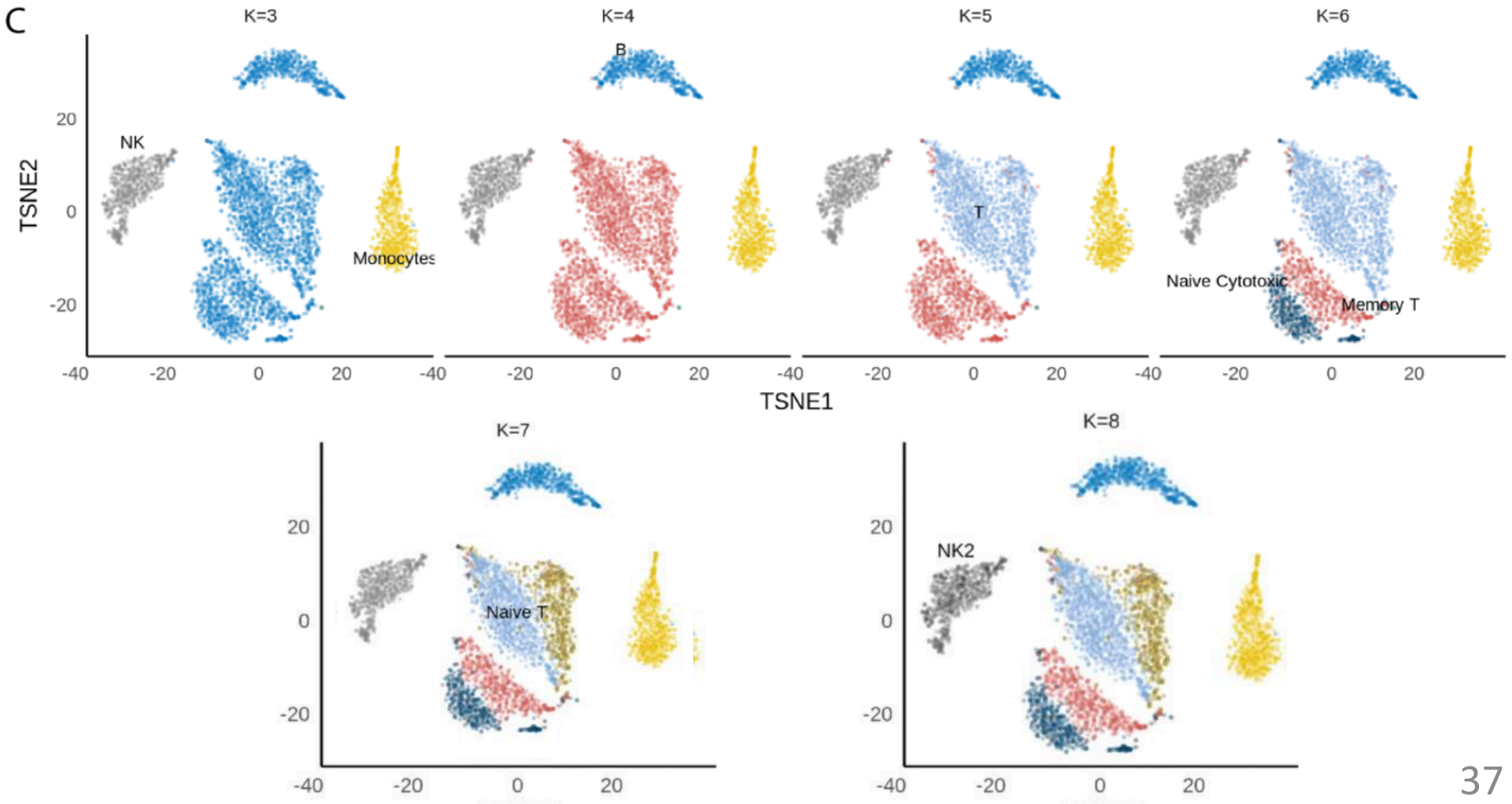
- Select genes with strong indication for cellular heterogeneity (cutoff of 2 on *z* score)

Step 2 Cluster:

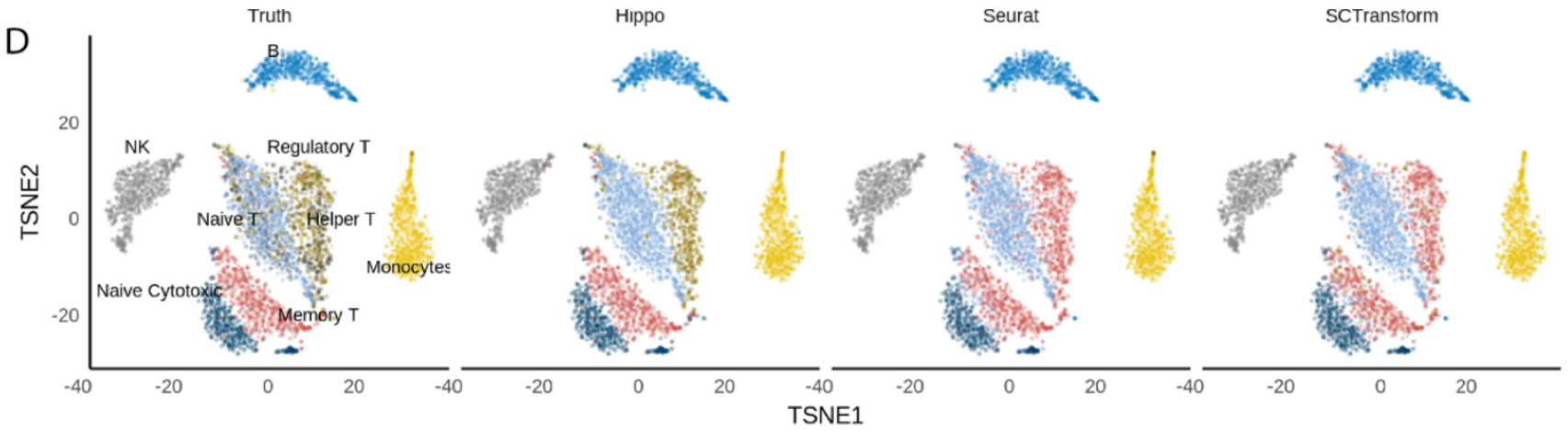
- With the selected features, cluster the cells into 2 groups using PCA + K-means
- Each cluster is evaluated with their intra-variability using the mean Euclidean distance from the centers of K-mean algorithm. The group with the highest intra-variability is selected and assigned for next round of clustering.

Computationally cheap because fewer and fewer features will be left for the next round of clustering

Results #3: HIPPO can successfully identify cell types



Results #3: HIPPO can identify cell types



Seurat and Sctransform fails to separate the memory T cells, regulatory T cells, and helper T cells, grouping them as one cluster

Paper 2

Their conclusions:

- Cell-type heterogeneity must be tackled as the first step of analysis for more reliable downstream analysis
- They introduced computationally and mathematically simple analysis tool for feature selection with great interpretability
- This pre-processing tool can resolve cellular heterogeneity and help avoid unnecessary normalizing steps that can introduce unwanted bias and noise

Paper 2

- What are the advantages of this model?
  - Do you think having a simple model can be helpful?
  - What are the advantages/disadvantages of not taking non-zero counts into account?
- What can be potential limitations of predicting cell type heterogeneity from drop-out rates?
  - Do you think more datasets are required to support conclusions?
- What are the advantages/disadvantages of inferring cell types from zero-inflation?
  - How can we solve the circular dependence of cell type heterogeneity and dropout?

Summary and comparison of 2 papers

PAPER 1: Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

PAPER 2: Demystifying "drop-outs" in single-cell UMI data

Common Points

- Drop-out rates in scRNA-Seq is determined by cell types
- Drop-out rates are not technical problems that should be eliminated but provide important biological information
- Zero-inflated distributions are not good fits for scRNA-Seq especially after taking cell type into account

Summary and comparison of 2 papers

PAPER 1: Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

PAPER 2: Demystifying "drop-outs" in single-cell UMI data

Differences

- To solve drop-outs -> uses cell type heterogeneity and biological covariates
- The goal is to select the best distribution for each gene
- Negative binomial distribution should be used to model scRNA-Seq

- To detect cell type heterogeneity -> uses drop-out rates
- The goal is to cluster the cells using drop-out rates
- Poisson distribution should be used to model scRNA-Seq