

Are scRNA-seq data zero-inflated?

Discussion on

“Droplet scRNA-seq is not zero-inflated” by Svensson 2020

Hunyong Cho

Biostatistics Department

UNC, Chapel Hill

Papers

“Droplet scRNA-seq is not zero-inflated” - Svensson 2020 Nature

“Separating measurement and expression models clarifies confusion in scRNA-seq analysis” - Sarkar & Stephens 2020 bioRxiv

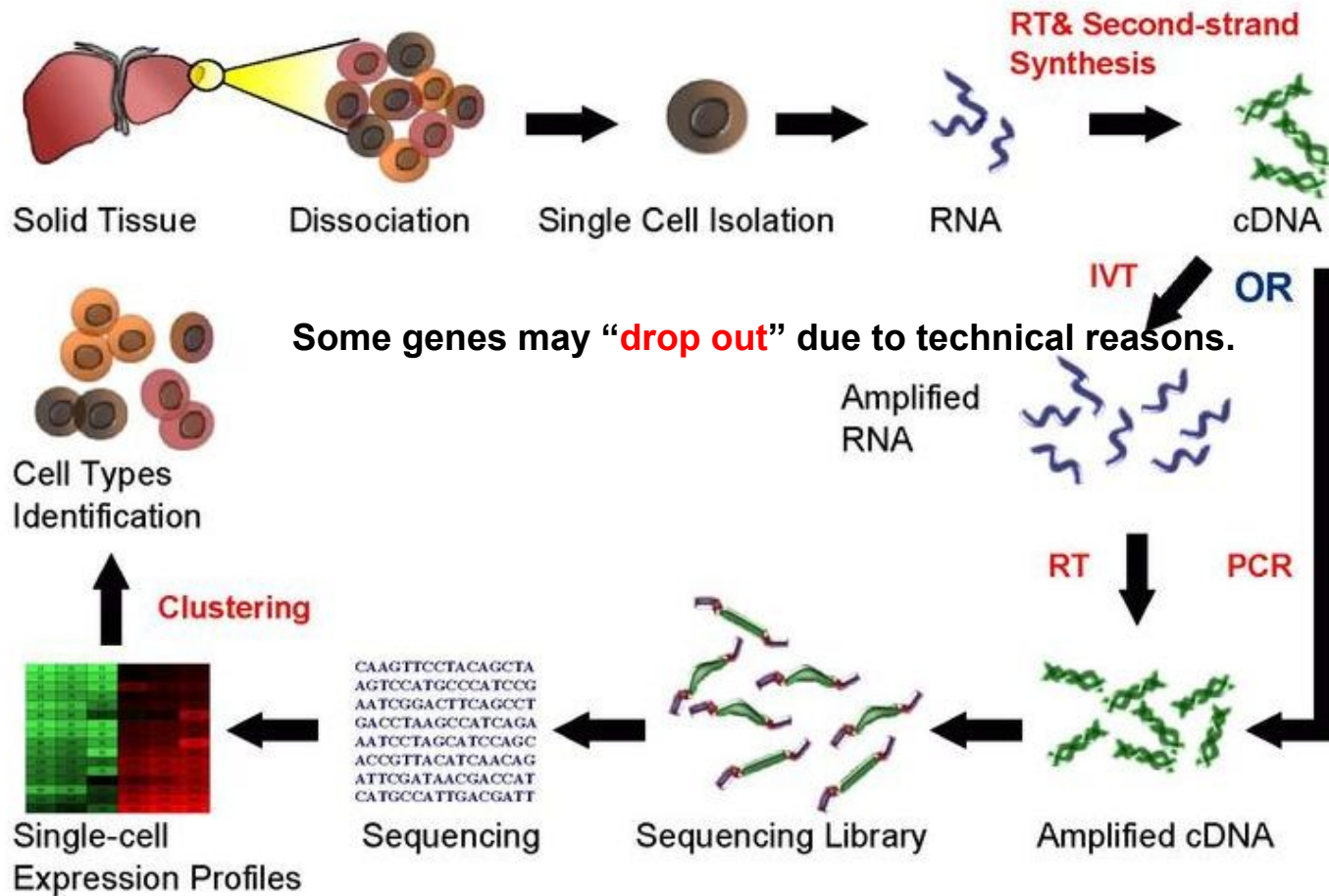
“M3Drop: Dropout-based feature selection for scRNASeq”
- Andrews & Hemberg 2019 Bioinformatics

“Feature selection and dimension reduction for scRNA-Seq based on a multinomial model”
- Townes et al. 2020 GenomeBiology

“Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics”
- Choi et al. 2020 bioRxiv

Background

scRNA-seq data



scRNA-seq data example

Gene	Cell1	Cell2	Cell3	Cell4	Cell5	Cell6	Cell7	Cell8	Cell9
1110038F14Rik	0	0	0	0	0	0	0	0	0
1110046J04Rik	0	0	0	0	0	0	0	0	0
1110051M20Rik	0	0	0	0	0	0	0	0	0
1110054M08Rik	0	0	0	0	0	0	0	0	0
1110057K04Rik	0	0	0	0	0	0	0	0	0
1110058L19Rik	0	0	0	0	0	0	0	0	0
1110059E24Rik	1	0	0	0	0	0	0	0	1
1110059G10Rik	0	0	0	0	0	0	0	0	0
1110059M19Rik	0	0	0	0	0	0	0	0	0
1110065P20Rik	0	0	0	0	0	0	0	0	18
1190002F15Rik	0	0	0	0	0	0	0	0	0
1190002N15Rik	0	0	0	0	0	0	0	0	0
1190003J15Rik	0	0	0	0	0	0	0	0	0
1190003K10Rik	0	0	0	0	0	0	0	0	0
1190005I06Rik	0	0	0	0	0	0	0	0	0
1190007I07Rik	0	0	0	0	0	0	0	0	0
1200011I18Rik	0	0	0	0	0	0	0	0	0
1200014J11Rik	0	0	0	0	0	0	0	0	0
1300002E11Rik	0	0	0	0	0	0	0	0	0
1300002K09Rik	1	0	0	0	0	0	0	0	0
1300015D01Rik	0	0	0	0	0	0	0	0	0
1300017J02Rik	0	0	0	0	0	0	0	0	0
1300018J18Rik	0	0	0	0	0	0	0	0	0
1500002010Rik	0	0	0	0	0	0	0	0	0
1500004A13Rik	0	0	0	0	0	0	0	0	0
1500009C09Rik	0	0	0	0	0	0	0	0	0
1500009L16Rik	0	0	0	0	0	0	0	0	1
1500011B03Rik	0	0	0	0	0	0	0	0	0

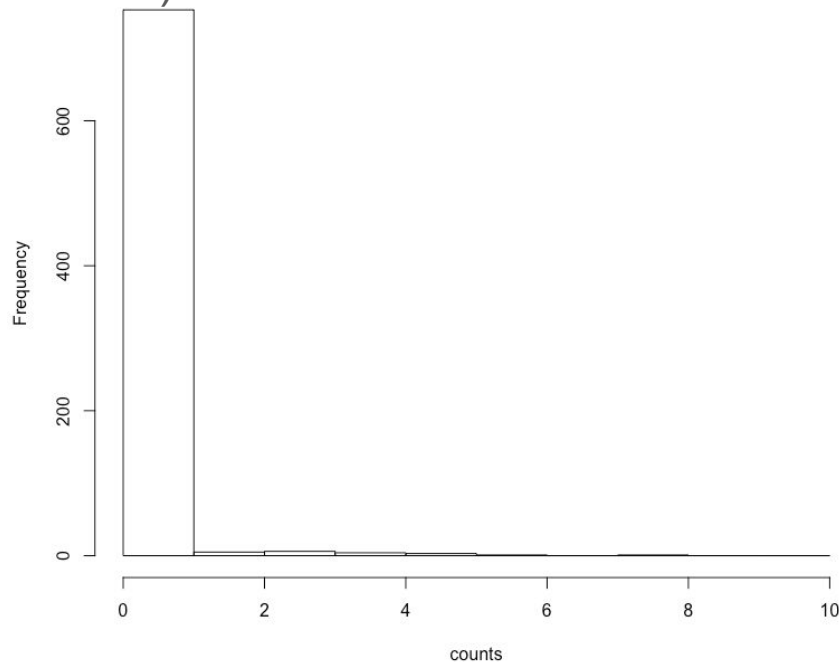
Motivating example

A typical scRNA-seq data (a specific gene count).

How would you model this?

Negative binomial?

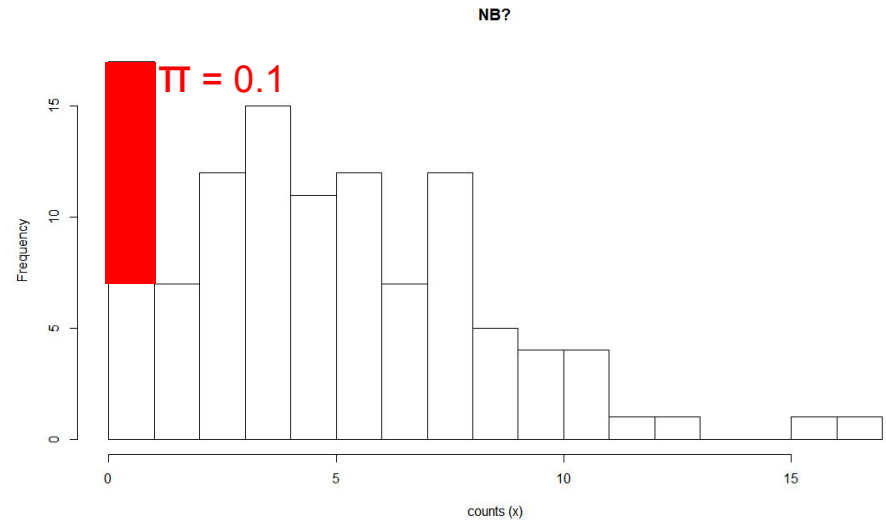
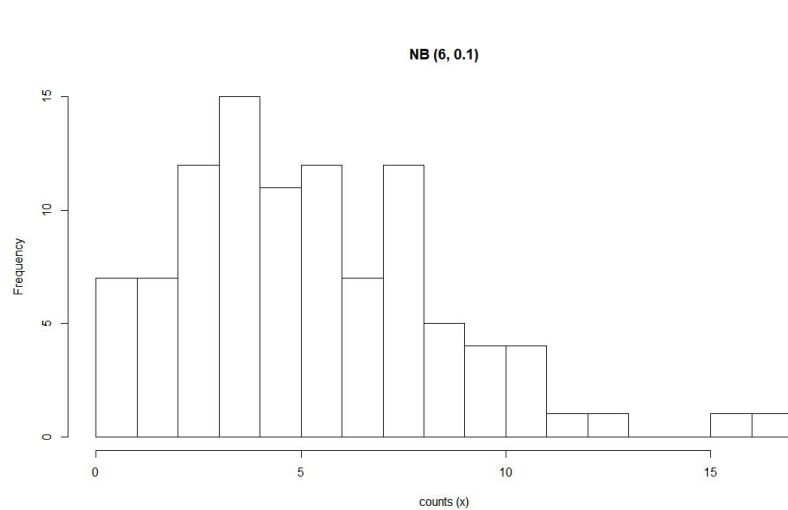
Zero-inflated Negative binomial?



Models - NB and ZINB

Negative binomial (NB) - the mean (μ) and the overdispersion (φ)

Zero-inflated Negative binomial (ZINB) - μ , φ , and the zero-inflation (π)



The nature of over-dispersion in NB

$X \sim \text{Poisson}(\mu)$

Mean = μ , variance = μ

What if *the mean parameter* is random?

Then, variance is higher.

$X \sim \text{Poisson}(R)$ where

$R \sim \text{Gamma}(1/\varphi, \mu\varphi)$.

Then, mean = μ , variance = $\mu + \mu^2\varphi$

X is fuzzier by $\mu^2\varphi$.

We reparametrize X as

$X \sim \text{Negative Binomial}(\mu, \varphi)$

A method paper - BZINB model

A bivariate zero-inflated negative binomial model for identifying underlying dependence with application to single cell RNA sequencing data

Hun Yong Cho^{1*}, Chuwen Liu¹, John S. Preisser¹, and Di Wu^{1,2**}

¹ *Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599,*

² *Department of Periodontology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599*

**email:* hunycho@live.unc.edu

***email:* did@email.unc.edu

SUMMARY: Measuring gene-gene dependence in single cell RNA sequencing (scRNA-seq) count data is often of interest and remains challenging, because an unidentified portion of the zero counts represent non-detected RNA due

Valentine Svensson (2020)

Nature

Droplet scRNA-seq is not zero-inflated

To the Editor — Potential users of single-cell RNA-sequencing (scRNA-seq)¹ often encounter a choice between high-throughput droplet-based methods and high-sensitivity plate-based methods. There is a widespread belief that scRNA-seq will often fail to generate measurements for some genes from some cells owing to technical molecular inefficiencies. It is believed that this causes data to have an overabundance of zero values compared to what is expected from random sampling and that this effect is particularly pronounced in droplet-based

dimensionality reduction methods with special handling of zero values had been introduced^{8,9}. All these approaches share two common themes (which deviate from SCDE): first, expression data are considered continuous with additional zero values, and second, a proportional relation is identified between the number of zero values and the average expression level of a gene¹⁰.

In the field of computational methods for scRNA-seq analysis, many methods have been designed to correct zero values in data, with the aim of allowing users to

control data with no biological variation. This will answer whether technical shortcomings in scRNA-seq methods produce an excess of technical zeros compared to expectations.

Negative-control datasets have been generated by adding a solution of RNA to the fluid in microfluidic systems, making the RNA content in each droplet identical. Five such datasets have been published: one to benchmark Drop-seq²⁴, one to benchmark InDrops²⁵, one to benchmark an early version of the commercial scRNA-

Key message:

- “ZINB is not necessary; NB is enough”
- “Zero-inflation is not technical but mostly biological”

1. [Literature 1](#): NB has good fit for the UMI data, ZINB not necessary.
2. [Literature 2](#): The zero inflation seems to reflect the biological variation.
3. [Literature 3](#): zero-inflation appears to be an artifact of log transformation.
4. [An experiment](#) comparing the *biological data* vs. the *negative control data*

Literature 1: NB has good fit for the UMI data, ZINB not necessary.

powsimR: power analysis for bulk and single cell RNA-seq experiments

Beate Vieth*, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard and Ines Hellmann*

Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, 82152 Munich, Germany

*To whom correspondence should be addressed.

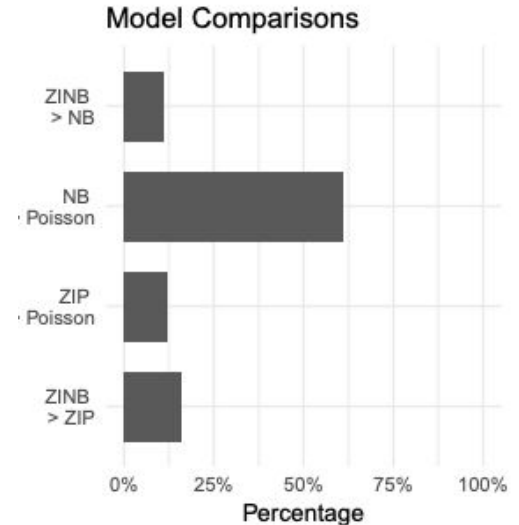
Associate Editor: Ivo Hofacker

Received on March 15, 2017; revised on June 29, 2017; editorial decision on July 2, 2017; accepted on July 4, 2017

Abstract

Summary: Power analysis is essential to optimize the design of RNA-seq experiments and to assess and compare the power to detect differentially expressed genes in RNA-seq data. PowsimR is a flexible tool to simulate and evaluate differential expression from bulk and especially single-cell RNA-seq data making it suitable for a priori and posterior power analyses.

Likelihood ratio tests say NB is good in most cases.



Literature 2: The zero inflation seems to reflect the biological variation.

M3Drop: dropout-based feature selection for scRNASeq

Tallulah S. Andrews  and Martin Hemberg  *

Department of Cellular Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on October 8, 2018; revised on November 29, 2018; editorial decision on December 18, 2018; accepted on December 19, 2018

Abstract

Motivation: Most genomes contain thousands of genes, but for most functional responses, only a subset of those genes are relevant. To facilitate many single-cell RNASeq (scRNASeq) analyses the set of genes is often reduced through feature selection, i.e. by removing genes only subject to technical noise.

Results: We present M3Drop, an R package that implements popular existing feature selection methods and two novel methods which take advantage of the prevalence of zeros (dropouts) in scRNASeq data to identify features. We show these new methods outperform existing methods on simulated and real datasets.

“These novel methods exploit the observation that dropout-rates per gene are strongly correlated with gene expression level (Pierson & Yau, 2015; Kharchenko et al, 2015)”

Literature 2B: The zero inflation seems to reflect the biological variation.

ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis



Emma Pierson¹ and Christopher Yau^{1,2*}

Abstract

Single-cell RNA-seq data allows insight into normal cellular function and various disease states through molecular characterization of gene expression on the single cell level. Dimensionality reduction of such high-dimensional data sets is essential for visualization and analysis, but single-cell RNA-seq data are challenging for classical dimensionality-reduction methods because of the prevalence of dropout events, which lead to zero-inflated data. Here, we develop a dimensionality-reduction method, (Z)ero (I)nfated (F)actor (A)nalysis (ZIFA), which explicitly models the dropout characteristics, and show that it improves modeling accuracy on simulated and biological data sets.

“The fundamental empirical observation ... is that the dropout rate for a gene depends on the expected expression level of that gene in the population ”

Literature 2B: The zero inflation seems to reflect the biological variation.

ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis



Emma Pierson¹ and Christopher Yau^{1,2*}

Abstract

Single-cell RNA-seq data allows insight into normal cellular function and various disease states through molecular characterization of gene expression on the single cell level. Dimensionality reduction of such high-dimensional data sets is essential for visualization and analysis, but single-cell RNA-seq data are challenging for classical dimensionality-reduction methods because of the prevalence of dropout events, which lead to zero-inflated data. Here, we develop a dimensionality-reduction method, (Z)ero (I)nfated (F)actor (A)nalysis (ZIFA), which explicitly models the dropout characteristics, and show that it improves modeling accuracy on simulated and biological data sets.

$$p_0 = \exp(-\lambda\mu^2),$$

$$\mathbf{z}_i \sim \text{Normal}(0, \mathbf{I}),$$

$$\mathbf{x}_i | \mathbf{z}_i \sim \text{Normal}(\mathbf{A}\mathbf{z}_i + \boldsymbol{\mu}, \mathbf{W}),$$

$$h_{ij} | x_{ij} \sim \text{Bernoulli}(p_0),$$

$$y_{ij} = \begin{cases} x_{ij}, & \text{if } h_{ij} = 0, \\ 0, & \text{if } h_{ij} = 1, \end{cases}$$

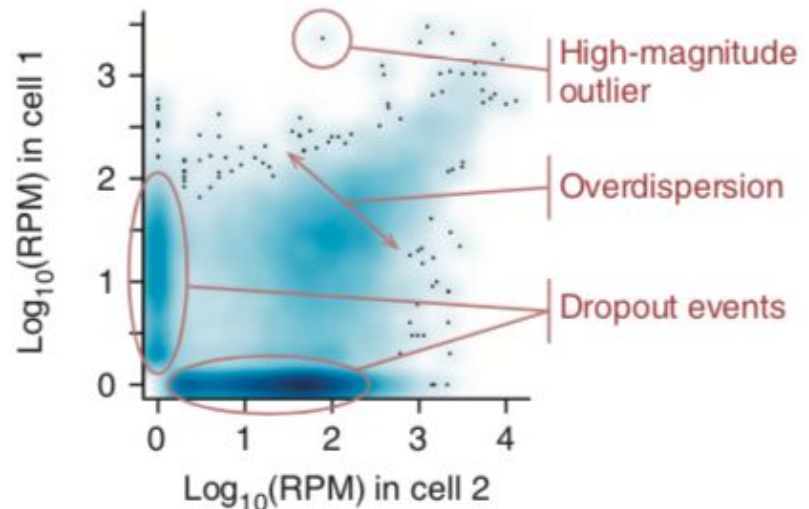
1. 'dropout rate (p_0)' here is simply the zero-fraction - not really a dropout nor zero-inflation.
2. As p_0 contains the true zeros, it of course is related with the non-zero mean (μ) even under the presence of zero-inflation.

Literature 2C: The zero inflation seems to reflect the biological variation.

Bayesian approach to single-cell differential expression analysis

Peter V Kharchenko¹⁻³, Lev Silberstein³⁻⁵ & David T Scadden³⁻⁵

Single-cell data provide a means to dissect the composition of complex tissues and specialized cellular environments. However, the analysis of such measurements is complicated by high levels of technical noise and intrinsic biological variability. We describe a probabilistic model of expression-magnitude distortions typical of single-cell RNA-sequencing measurements, which enables detection of differential expression signatures and identification of subpopulations of cells in a way that is more tolerant of noise.



Again, dropout is defined by zeros.

“the dropout rate for a given cell depends on the average expression magnitude of a gene in a population with dropouts being more frequent for genes with lower expression magnitude”

Literature 2: The zero inflation seems to reflect the biological variation.

Svenson 2020

One group proposed that genes with higher fractions of zero values than suggested by the negative binomial distribution might be good candidates for further analysis because this seems to reflect biological variation²².

“higher fractions of zero values than suggested by the NB distribution” = zero-inflation


≠ dropout

≠ zero proportion

Literature 3: zero-inflation appears to be an effect of log transformation

Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model



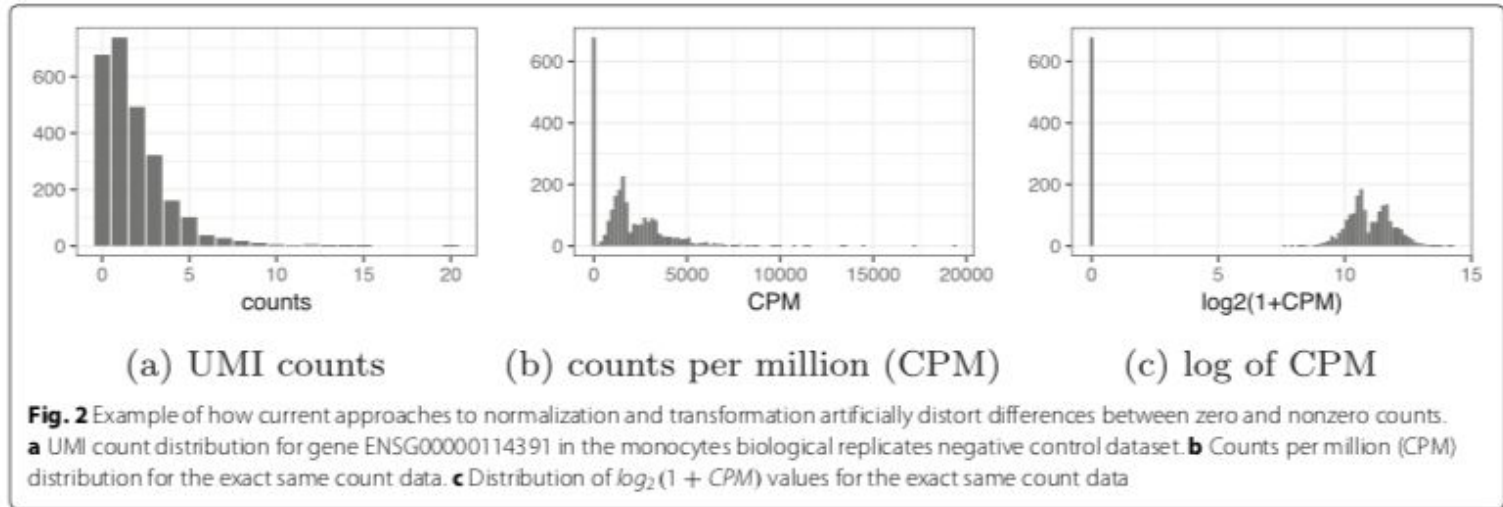
F. William Townes^{1,2} , Stephanie C. Hicks³, Martin J. Aryee^{1,4,5,6} and Rafael A. Irizarry^{1,7*}

Abstract

Single-cell RNA-Seq (scRNA-Seq) profiles gene expression of individual cells. Recent scRNA-Seq datasets have incorporated unique molecular identifiers (UMIs). Using negative controls, we show UMI counts follow multinomial sampling with no zero inflation. Current normalization procedures such as log of counts per million and feature selection by highly variable genes produce false variability in dimension reduction. We propose simple multinomial methods, including generalized principal component analysis (GLM-PCA) for non-normal distributions, and feature selection using deviance. These methods outperform the current practice in a downstream clustering assessment using ground truth datasets.

Actually, it is not from the “*log-transformation*,” but is from the “*normalization*” e.g. RPK

Literature 3: zero-inflation appears to be an effect of log transformation



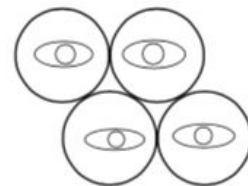
By counts → CPM transformation,

Zeros are mapped to zeros keeping its probability mass,
while ones (or larger values) are mapped to multiple values.

Log-transformation is not the key here.

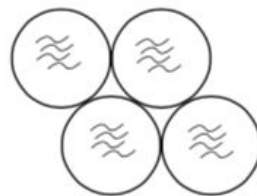
4. More about the experiment

The biological data: Each droplet is made from a cell
(Droplets are heterogeneous.)



The negative control data: All droplets are made from the same RNA solution.
(Every droplet is homogeneous)

Droplets with RNA solution



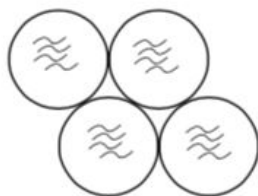
The negative control data were used to
remove the biological variability.

Their hypothesis:

“If zero-inflation can be explained by the biological variation,
then the zero-inflation, if any, is not technical zeros but are real zeros.”

Negative control data

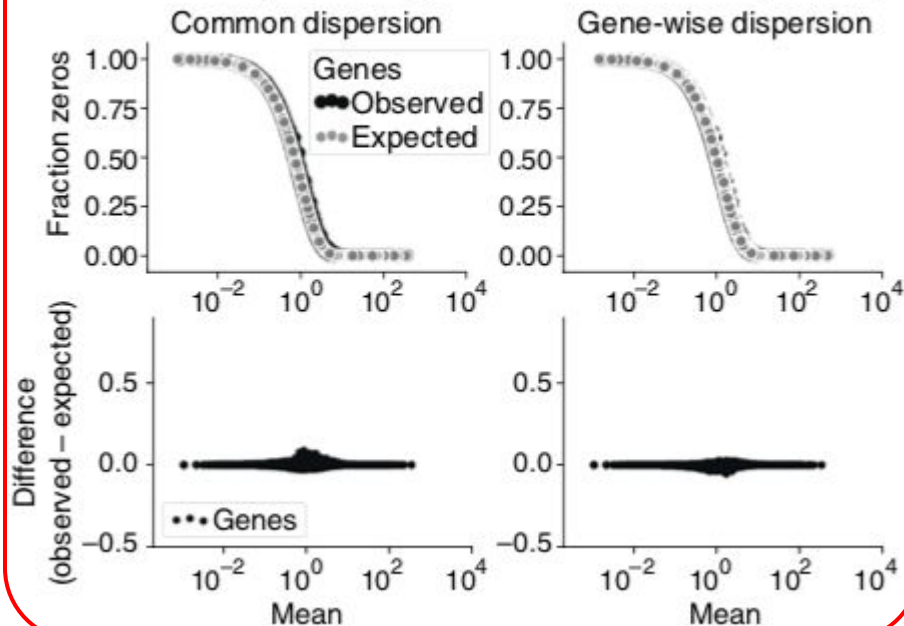
Droplets with RNA solution



a

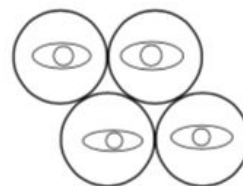
Klein et al., 2015

Solution (K562 endogenous RNA and ERCC spike-ins)



Biological data

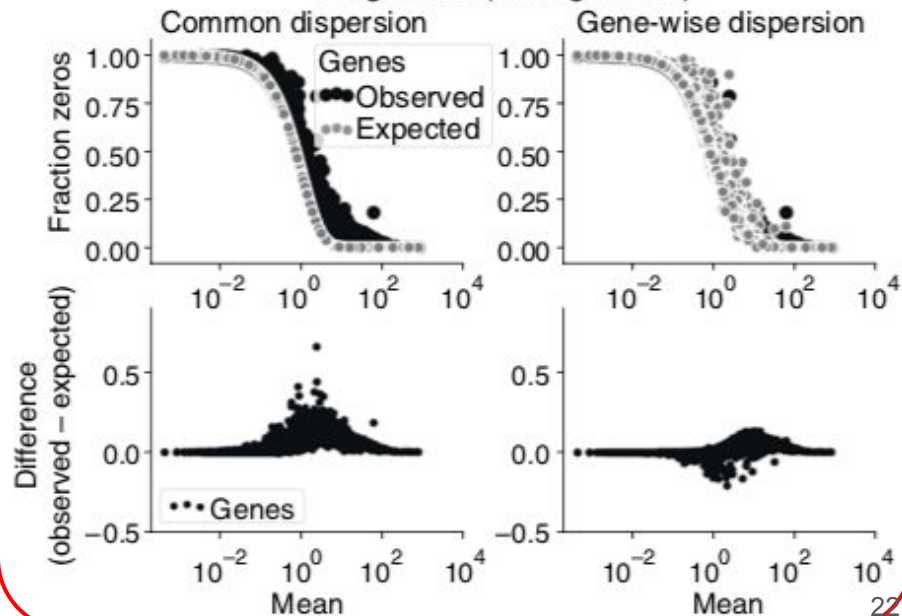
Droplets with homogeneous single cells



f

10x v3 HEK293T

Single cells (homogeneous)



More discussion on
“the experiment”

1. Confusion between dropouts and zero-inflation
2. Is the gap between two data sets purely due to biological variation?
3. Why common φ ?

1. Confusion between dropouts and zero-inflation

- Distinction between dropouts (technical zeros) and zero-inflation
- They define both “zero-inflation” and “dropouts”
“to observe more ‘technical zeros’ than expected”
- This can be a definition of zero-inflation, but cannot be of dropouts.
- zero-inflation should be understood in terms of model fit,
while dropouts are purely mechanical results.

(Relevant paper 1)

“Separating measurement and expression models clarifies confusion in scRNA-seq analysis”

- Sarkar & Stephens 2020 bioRxiv

dropouts = missing data = zeros?

No! Not all zeros are dropouts.

Imputation = filling in values for zeros?

No! Not all zeros are fake.

Zero-inflation = extra zeros that cannot be explained by simpler models

Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis

(say, NB)

 Abhishek Sarkar,  Matthew Stephens

doi: <https://doi.org/10.1101/2020.04.07.030007>

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

Full Text

Info/History

Metrics

 Preview PDF

ABSTRACT

How to model and analyze scRNA-seq data has been the subject of considerable

(Relevant paper 2)

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

Kwangbom Choi¹, Yang Chen², Daniel A. Skelly¹, Gary A. Churchill^{1,*}

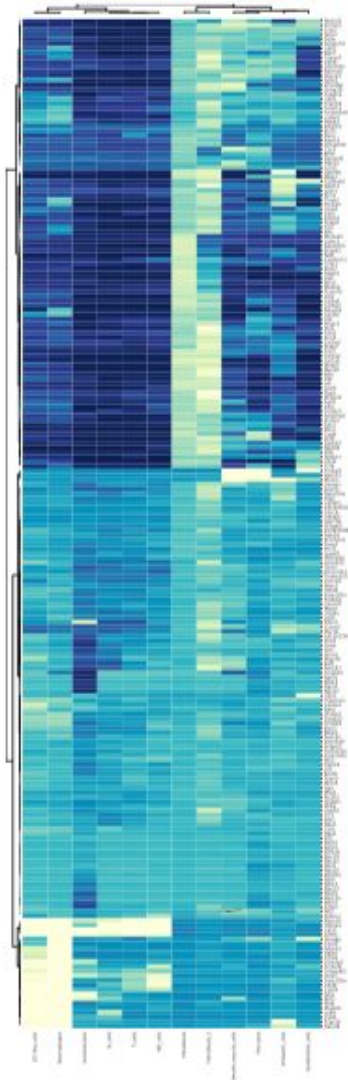
¹The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine, 04609

²University of Michigan, 500 South State Street, Ann Arbor, Michigan, 48109

Abstract. Single-cell RNA sequencing is a powerful tool for characterizing cellular heterogeneity in gene expression.

However, high variability and a large number of zero counts present challenges for analysis and interpretation. There is

- (a) “... the primary causes of zero inflation are not technical but rather biological in nature.”
- (b) “the parameter estimates of the ZINB distribution are an unreliable indicator of zero-inflation”



Hypothesis:

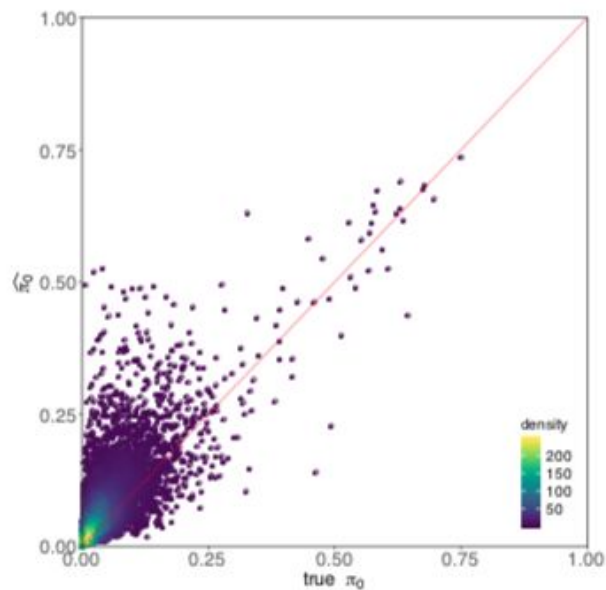
“If zero-inflation is primarily due to dropouts, we would expect to see zeros evenly distributed across cell-types”

The heatmap (genes x cell types)

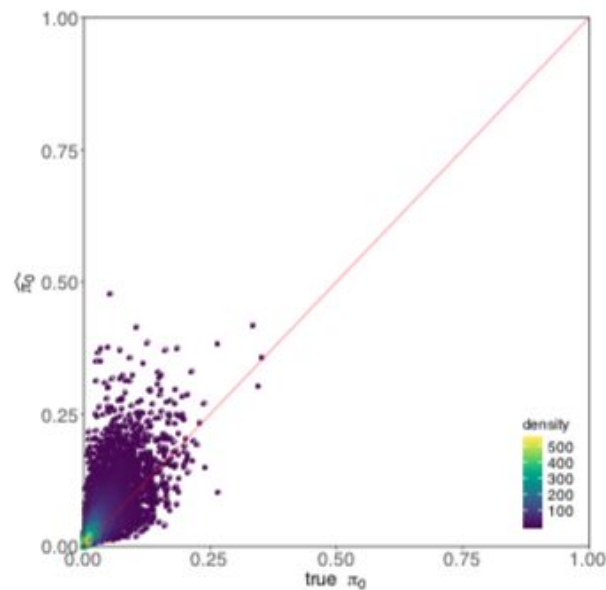
dark = excess zeros

light = fewer zeros than expected

- 1) *Dropout = zero-inflation?*
- 2) *May dropout rates not vary by cell types?*



(c)



(d)

Simulated ZINB data. Left: cell-type not adjusted. Right: cell-type adjusted.

2. Is the gap between two data sets purely due to biological variation?

1. Negative control data: an RNA solution -> scRNAseq
2. Biological data: Homogeneous “cells” -> scRNAseq

The difference between (1) & (2) can be decomposed into

- A. heterogeneity between cells. (ϕ can be different by genes in (2), and we should look at the gene-wise dispersion. See the next discussion point.)
- B. Possibly higher dropouts for (2).

The difference from A is considered biological, while B is technical.

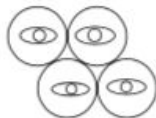
3. Why common ϕ ?

Negative control data - common ϕ (okay),

Biological data can have **different ϕ** due to cell heterogeneity.

They compare the more-than-expected fraction of zeros based on common ϕ .

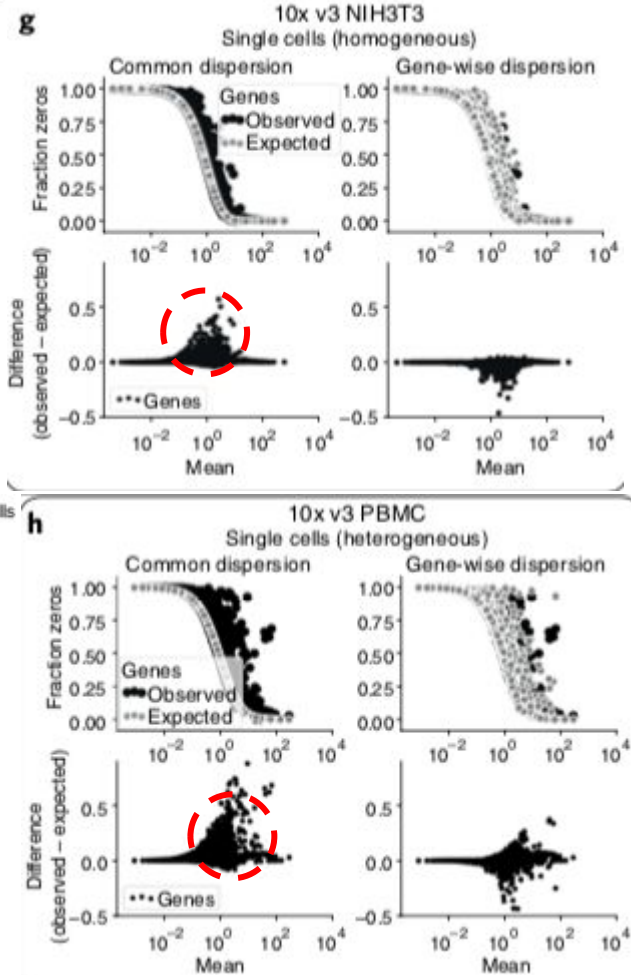
Droplets with homogeneous single cells



Droplets with heterogeneous single cells



Biological data



3. Why common ϕ ? - a thought experiment

Consider only two genes: Genes A and B with about-equal means.

It can be that gene A count is more variable than Gene B in **biological data**:

<i>Cell</i>	1	2	3	4	5	...
<i>Gene A</i>	3	1	0	1	7	...
<i>Gene B</i>	3	3	2	3	2	...

Genes could have different dispersion across cells.

But this is **hardly true for the negative control data**.

3. Why common ϕ ?

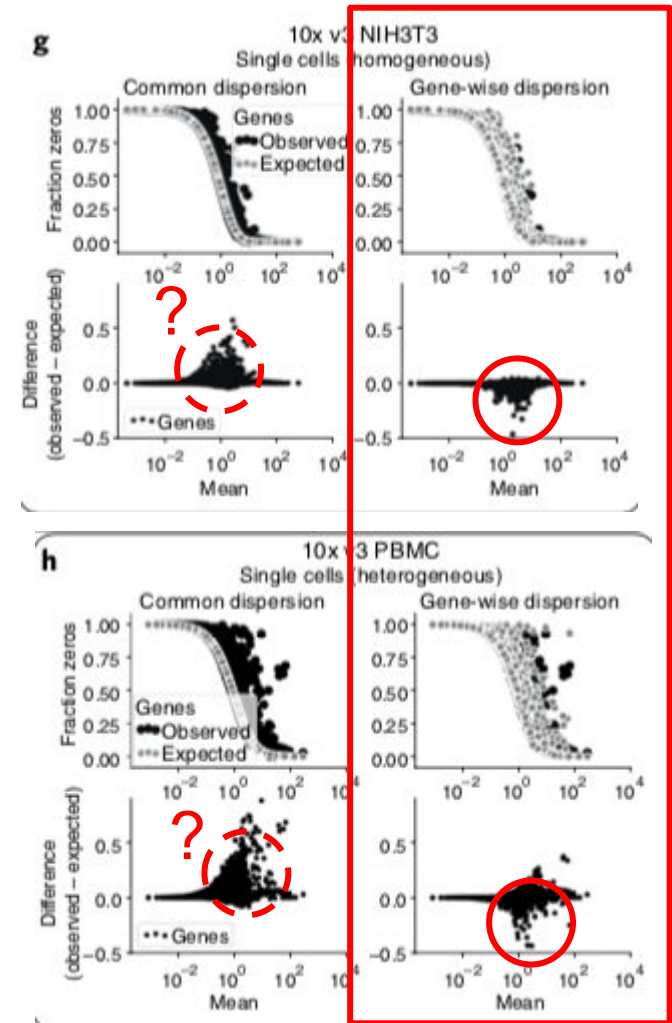
Look at the right panels!

The right panels rather imply that in biological data, genes may have different overdispersion.

There are less-than-expected zero fractions!

So their claim that extra zeros are from biological variability is questionable.

This figure can only show that negative binomial distribution is good enough.



2. Is the gap between two data sets purely due to biological variation?

Alternative approach for detecting dropouts.

To compare the zero proportion between the biological data and the negative control data after controlling for the nonzero-means.

Conclusion

1. Lots of confusion between zero-inflation and dropouts

Modeling dropouts with zero-inflation models is another thing!

2. Although ZINB may not be needed in many data, that does not imply there is no dropout.
3. Biological variability can be well explained by varying overdispersion.

References

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 1-4.

Vieth, B., Ziegenhain, C., Parekh, S., Enard, W., & Hellmann, I. (2017). powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21), 3486-3488.

Andrews, T. S., & Hemberg, M. (2019). M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics*, 35(16), 2865-2867.

Pierson, E., & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1), 241.

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7), 740.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome biology*, 20(1), 1-16.

Sarkar, A. K., & Stephens, M. (2020). Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. *BioRxiv*.

Choi, K., Chen, Y., Skelly, D. A., & Churchill, G. A. (2020). Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *bioRxiv*.