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 methods. Here I present an investigation of published data for technical controls in droplet-based scRNA-seq experiments that demonstrates that the number of zero values in the data is consistent with common distributional models of molecule sampling counts. Thus, any additional zero values in biological data likely result from biological variation or may reflect variation in gene abundance among cell types or cell states.

As scRNA-seq started to gain popularity, users expressed concern about the number of zero values among gene expression measures. That is, for any given gene, several cells showed no expression, even if expression was relatively high in other cells[2–4]. These 'technical zeros' are considered distinct from 'biological zeros', which are zero because the gene is not expressed[5].

The original reference to this 'dropout' problem of yielding an inflation of zero values in scRNA-seq data was in the description of single-cell differential expression (SCDE)[6]. It was reported that genes in scRNA-seq data generated with the low-throughput plate-based methods single-cell tagged reverse transcription (STRT)-seq and Smart-seq resulted in more zero values when comparing two single-cell samples than was typically observed when comparing two traditional bulk RNA-seq samples. To solve the task of reliably finding shifts in mean between conditions, a mixture model was used that could model the unexpected zeros.

There is no clear record in the literature from where the idea of 'zero inflation' in droplet-based data arose. The conceptual origin seems to have been the combination of the model in SCDE attempting to model counts in scRNA-seq data as well as expectations based on intuition about continuous distributions.

Over time, several statistical tests, such as model-based analysis of single-cell transcriptomics (MAST)[7] and eventually 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 gene[10].

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 predict what the expression level of a gene in a cell would have been, had there been no zero-inflation or dropouts[11–17], referred to as 'imputation'.

High-throughput droplet-based scRNA-seq methods measure discrete counts of unique molecular identifiers (UMIs). The stochastic sampling of counts can be modeled using a Gamma-Poisson distribution, more colloquially known as the 'negative binomial' distribution. Under stochastic sampling of counts, zero values are expected. When researchers discuss 'zero-inflation' or 'dropouts', they refer to observing more 'technical zeros' than expected, which would pose challenges when estimating molecular abundance. This is not a concern with the expected 'sampling zeros'[5].

It has recently become popular to make statistical models for scRNA-seq counts using the 'zero-inflated negative binomial' distribution[18–20]. This distribution adds a probability of observing a zero value in any given draw from the distribution of each gene.

Closer investigations seem to counter the assumption that zero inflation is an inherent property of scRNA-seq data. A methods paper on simulations for scRNA-seq data reported negative binomial to be sufficient for UMI count data, with little added benefit from a zero-inflation component[21]. Similarly, the authors of the method bayNorm found that zero-inflation was not necessary[16]. 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 variation[22]. Finally, another investigation suggested that what appears as zero inflation is an effect of log transformation of the counts, and factor models aware of count distributions alleviate the problem[23].

To investigate the number of zero values in droplet-based high-throughput scRNA-seq platforms, it is possible to use negative 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-seq[24], one to benchmark InDrops[25], one to benchmark an early version of the commercial scRNA-seq platform from 10x Genomics[26] and two to benchmark a later version of the commercial platform from 10x Genomics[27] (Supplementary Note 1). The negative control data differ in the RNA solution added: bulk RNA from cells or tissues or artificial External RNA Control Consortium (ERCC) RNA (Table 1). For simplicity, here I refer to each RNA species, regardless of source, as a 'gene'. All negative control experiments aimed to make an RNA dilution that would fill each droplet with an RNA amount similar to that in a typical mammalian cell.

The negative binomial distribution has two parameters: a 'mean' (sometimes called 'rate') parameter μ and a 'dispersion' parameter ϕ. Historically, the negative binomial distribution has been used to model data in which there is unknown random variation in the exposure compared with a Poisson distribution[28], where the dispersion ϕ determines the increased amount of variation.

For negative control data, it is reasonable to expect ϕ to be common for all genes as they are all affected by the same sampling process. In bulk RNA-seq data, gene-wise ϕ parameters are often used to account for biological variation in gene expression between samples[29]. A negative binomial distribution with mean μ and dispersion ϕ will have the probability of observing a count of 0 determined by

$$P(0|\mu,\ \phi) = \left(\frac{\phi^{-1}}{\mu + \phi^{-1}}\right)^{\phi^{-1}}$$

This probability can be compared with the fraction of droplets where a gene had zero counts. In general, the fraction of zero values can be perfectly predicted for negative control data using a common dispersion parameter (Fig. 1a–e).

Table 1 | Datasets analyzed in this study, including scRNA-seq method, input material, number of droplets and number of genes observed[a]

Data source	Method	Input material	Droplets	Genes	Model	Number and percentage of genes over *x* percentage points away from expected fraction							
						1 percentage point		5 percentage points		10 percentage points		20 percentage points	
10x v3 PBMC	Chromium v3	Single PBMCs	1,222	33,538	Poisson	3,750	11.18%	1,088	3.24%	445	1.33%	182	0.54%
					NB common	3,236	9.65%	800	2.39%	353	1.05%	151	0.45%
					NB gene-wise	1,281	3.82%	288	0.86%	86	0.26%	25	0.07%
10x v3 HEK293T	Chromium v3	Single HEK293 cells	2,378	57,905	Poisson	6,958	12.02%	2,306	3.98%	360	0.62%	16	0.03%
					NB common	8,957	15.47%	4,903	8.47%	2412	4.17%	66	0.11%
					NB gene-wise	4,311	7.44%	977	1.69%	123	0.21%	1	0.00%
10x v3 NIH3T3	Chromium v3	Single NIH3T3 cells	2,458	54,232	Poisson	3,864	7.12%	541	1.00%	186	0.34%	46	0.08%
					NB common	2,925	5.39%	392	0.72%	134	0.25%	40	0.07%
					NB gene-wise	1,178	2.17%	120	0.22%	41	0.08%	8	0.01%
Klein et al. 2015	InDrops	Endogenous RNA from K562 cell line and ERCC spike-ins	953	25,435	Poisson	3,515	13.82%	4	0.02%	0	0.00%	0	0.00%
					NB common	3,705	14.57%	4	0.02%	0	0.00%	0	0.00%
					NB gene-wise	2,036	8.00%	1	0.00%	0	0.00%	0	0.00%
Svensson et al.[27] dataset 1	Chromium v1	Endogenous RNA from human brain and ERCC spike-ins	2,000	24,116	Poisson	103	0.43%	1	0.00%	0	0.00%	0	0.00%
					NB common	101	0.42%	1	0.00%	0	0.00%	0	0.00%
					NB gene-wise	6	0.02%	0	0.00%	0	0.00%	0	0.00%
Svensson et al.[27] dataset 2	Chromium v1	Endogenous RNA from human brain and ERCC spike-ins	2,000	24,116	Poisson	1,134	4.70%	16	0.07%	1	0.00%	0	0.00%
					NB common	1,134	4.70%	16	0.07%	1	0.00%	0	0.00%
					NB gene-wise	117	0.49%	3	0.01%	1	0.00%	0	0.00%
Zheng et al.[26]	GemCode	ERCC spike-ins	1,015	92	Poisson	17	18.48%	3	3.26%	1	1.09%	0	0.00%
					NB common	19	20.65%	3	3.26%	1	1.09%	0	0.00%
					NB gene-wise	28	30.43%	12	13.04%	7	7.61%	1	1.09%
Macosko et al.[24]	Drop-seq	ERCC spike-ins	84	80	Poisson	54	67.50%	35	43.75%	28	35.00%	16	20.00%
					NB common	54	67.50%	37	46.25%	28	35.00%	16	20.00%
					NB gene-wise	26	32.50%	4	5.00%	0	0.00%	0	0.00%
Padovan-Merhar et al.[30]	SMARTer (C1)	Single fibroblast cells	96	25,590	Poisson	15,342	59.95%	13,841	54.09%	12,697	49.62%	10,787	42.15%
					NB common	15,174	59.30%	13,188	51.54%	10,852	42.41%	6,392	24.98%
					NB gene-wise	12,865	50.27%	7,539	29.46%	4,091	15.99%	912	3.56%

[a]For each dataset and counting model, the number of genes passing different thresholds of difference in percentage points is reported. Because different datasets consider different numbers of genes, the percentage of genes is also reported. Chromium v1, 10x Genomics Chromium Single Cell 3′ Reagents kit (version 1 chemistry) on Chromium controller instrument; Chromium v3, 10x Genomics Chromium Single Cell 3′ Reagents kit (version 3 chemistry) on Chromium controller instrument; InDrops, indexing droplets; GemCode, GemCode Single-Cell 3′ Gel Bead and Library Kit on GemCode Single-Cell instrument; PBMC, peripheral blood mononuclear cell; NB, negative binomial; HEK, human embryonic kidney; NIH3T3, mouse fibroblast-derived cells; K562, chronic myelogenous leukemia lymphoblast; SMARTer (C1), Clontech SMARTer Ultra Low RNA Kit for the Fluidigm C1 System.

As context, the same statistics have been calculated for data with single cells in droplets that include biological heterogeneity. Biological data display a larger number of zero values than expected for a homogenous populations of cells (Fig. 1f,g) and, to an even larger degree, for heterogeneous populations of cells (Fig. 1h). When allowing independent dispersion parameters for each gene, many of the zero values in biological data can be accounted for, whereas negative control data only show marginal improvements.

The results can be quantified by counting the number of genes passing different thresholds of percentage points in the absolute difference between the expected fraction and the observed fraction of zero values (Table 1). This makes it clear that genes with unexpected zero values are due to biological variation. Biological data with single cells in droplets have hundreds of genes 10 percentage points away from expectation. In contrast, technical data with RNA solution in droplets only have at most one gene as far as 10 percentage points away from expectation. When scaling the counts in expression matrices to account for differences in total counts between cells (referred to as 'exposure' in statistical nomenclature), even a Poisson distribution can explain the fraction of zero values observed in negative control data (Supplementary Fig. 1).

The issue of zero values and 'dropouts' in scRNA-seq data has been discussed in the literature since the earliest scRNA-seq studies; it continues to be a topic of concern. Although not recorded in literature, various personal interchanges indicate that it is a particular worry among researchers choosing between high-throughput droplet-based scRNA-seq methods and lower-throughput plate-based methods.

The negative-control data for droplet-based methods presented here indicate that the number of zero values observed is consistent with what is expected from count

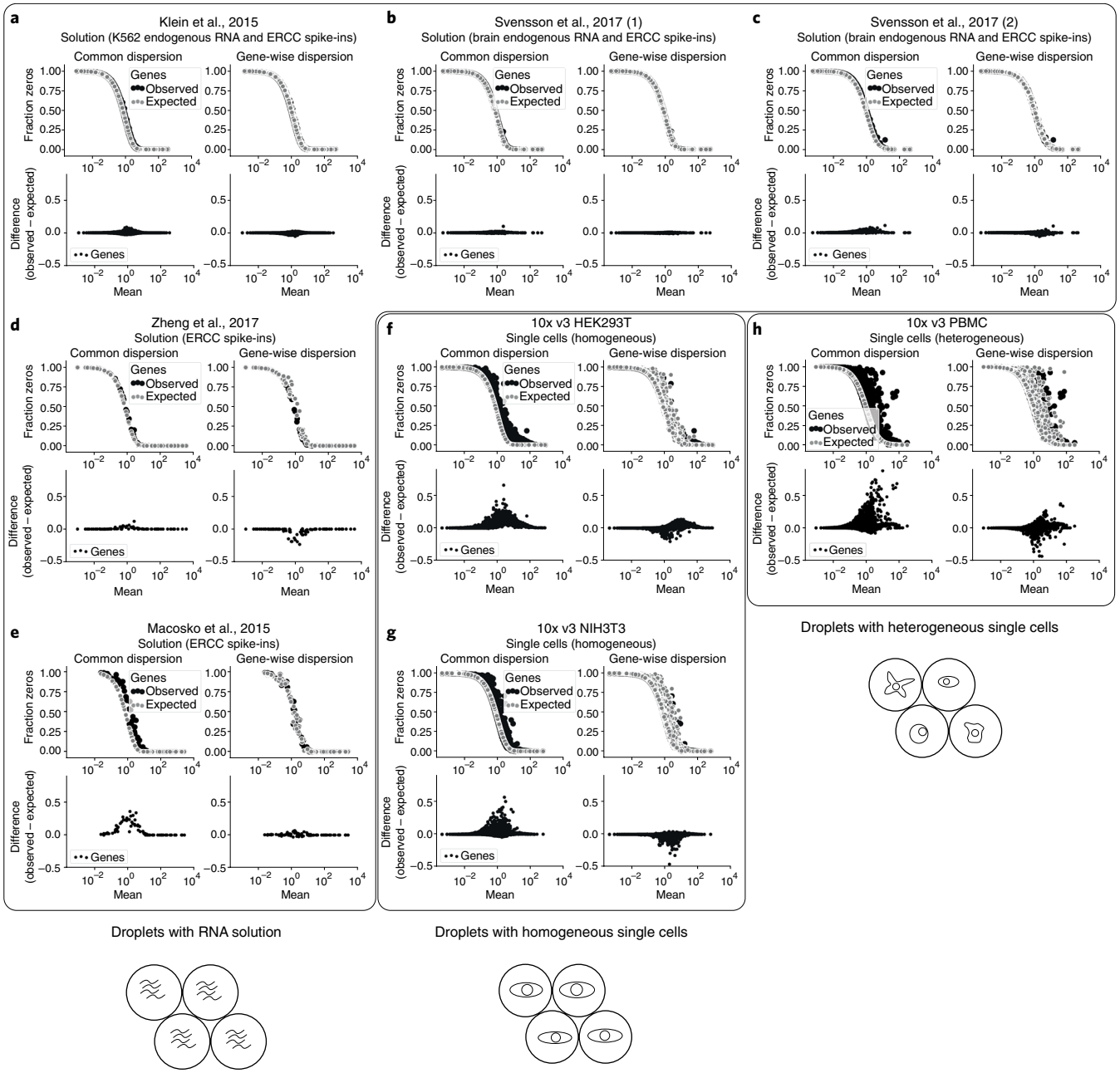

Fig. 1 | Comparing observed with expected zeros in scRNA-seq data. a–e, Technical control datasets, where an RNA solution is encapsulated in droplets: solution of K562 endogenous RNA and ERCC spike-ins measured with InDrops[25] (**a**); solution of endogenous brain RNA and ERCC spike-ins measured with Chromium v1 at high concentration (200,000 molecules per droplet)[27] (**b**); solution of endogenous brain RNA and ERCC spike-ins measured with Chromium v1 at low concentration (100,000 molecules per droplet)[27] (**c**); solution of ERCC spike-ins measured with GemCode[26] (**d**); and solution of ERCC spike-ins measured with Drop-seq[24] (**e**). **f**,**g**, Biological datasets where single homogeneous cells from cell cultures are encapsulated in droplets: single HEK293T cells with 10x Chromium v3 (**f**) and single NIH3T3 cells with 10x Chromium v3 (**g**). **h**, A biological dataset where single heterogeneous cells from human peripheral blood are encapsulated in droplets. For each dataset the fraction of droplets with zero count for each gene was compared to the mean UMI count of the gene. In each panel, shown are the expected fraction with common dispersion (left), with gene-wise dispersion (right), and the difference between observed and expected fractions (bottom). Panels are grouped by their frames into whether they represent droplets with RNA solution (**a**–**e**), droplets with homogeneous single cells (**f**, **g**), or droplets with heterogeneous single cells (**h**), illustrated by sketches below the groups of panels.

data. Additional zero values in biological data are likely due to biological variation. Unfortunately, no comparable negative control data exist for plate-based methods. However, the analysis presented here was applied to a dataset of read counts from a relatively homogeneous population of cells for which transcript full-length scRNA-seq had been performed using the Fluidigm system[30]. With common dispersion, a very large amount of genes had unexpected fractions of zero values (Supplementary Fig. 2). Over 40%

of genes were more than 10 percentage points away from the expected fraction (Table 1). With gene-wise dispersion this was reduced to 16% of genes, which is much larger than in biological droplet-based method data (0.1–0.3% of genes).

In a study on simulation of scRNA-seq data, the investigators found droplet-based data to be sufficiently modeled using a negative binomial distribution, whereas plate-based data needed zero inflation to be accurately simulated[21]. It is possible that UMIs in droplet-based methods deflate counts of genes with particularly high PCR duplication. Another possibility is that uneven sampling of fragments from gene bodies in plate-based methods introduces an additional layer of count noise that gives rise to gene- and cell-specific overdispersion in addition to global overdispersion, manifesting as additional zero values that can be difficult to tell apart from biological variation of interest.

Nevertheless, statistical analysis of data in which noise follows the negative binomial distribution is much simpler than analysis of zero-inflated distributions, which have problems with identifiability[5]. When working with count data, there is nothing particularly special about observing a zero value, and just as the formula above describes the expected number of zero values, similar formulas can be written for expected numbers of other values. What is important to remember is that even when count data are transformed or 'normalized', they still exhibit properties that are different from naturally continuous data. This also implies the number of zero values observed can be decreased by counting more molecules through global increases in capture efficiency or increased sequencing depths per droplet. ❐

Valentine Svensson
Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA.
e-mail: v@nxn.se

Published online: 14 January 2020
https://doi.org/10.1038/s41587-019-0379-5

References

1. Chen, X., Teichmann, S. A. & Meyer, K. B. *Annu. Rev. Biomed. Data Sci.* **1**, 29–51 (2018).
2. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. *Nat. Methods* **14**, 565–571 (2017).
3. Bacher, R. & Kendziorski, C. *Genome Biol.* **17**, 63 (2016).
4. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. *Genome Med.* **9**, 75 (2017).
5. Silverman, J.D., Roche, K., Mukherjee, S. & David, L.A. Preprint at *bioRxiv* https://doi.org/10.1101/477794 (2018).
6. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. *Nat. Methods* **11**, 740–742 (2014).
7. Finak, G. et al. *Genome Biol.* **16**, 278 (2015).
8. Pierson, E. & Yau, C. *Genome Biol.* **16**, 241 (2015).
9. Lin, P., Troup, M. & Ho, J. W. K. *Genome Biol.* **18**, 59 (2017).
10. Tung, P.-Y. et al. *Sci. Rep.* **7**, 39921 (2017).
11. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. *BMC Bioinforma.* **19**, 220 (2018).
12. Zhu, L., Lei, J., Devlin, B. & Roeder, K. *Ann. Appl. Stat.* **12**, 609–632 (2018).
13. Azizi, E., Prabhakaran, S., Carr, A. & Pe'er, D. *Genomics Computational. Biol.* **3**, e46 (2017).
14. Li, W. V. & Li, J. J. *Nat. Commun.* **9**, 997 (2018).
15. van Dijk, D. et al. *Cell* **174**, 716–729.e27 (2018).
16. Tang, W. et al. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz726 (2019).
17. Huang, M. et al. *Nat. Methods* **15**, 539–542 (2018).
18. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. *Nat. Commun.* **9**, 284 (2018).
19. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. *Nat. Methods* **15**, 1053–1058 (2018).
20. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. *Nat. Commun.* **10**, 390 (2019).
21. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. *Bioinformatics* **33**, 3486–3488 (2017).
22. Andrews, T.S. & Hemberg, M. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty1044 (2018).
23. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. *Genome Biol.* **20**, 295 (2019).
24. Macosko, E. Z. et al. *Cell* **161**, 1202–1214 (2015).
25. Klein, A. M. et al. *Cell* **161**, 1187–1201 (2015).
26. Zheng, G. X. Y. et al. *Nat. Commun.* **8**, 14049 (2017).
27. Svensson, V. et al. *Nat. Methods* **14**, 381–387 (2017).
28. McCullagh, P. & Nelder, J.A. *Generalized Linear Models, Second Edition* (CRC Press, 1989).
29. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. *Bioinformatics* **26**, 139–140 (2010).
30. Padovan-Merhar, O. et al. *Mol. Cell* **58**, 339–352 (2015).

Acknowledgements

I thank E. da Veiga Beltrame for feedback on the manuscript, G. Eraslan for making fast code for fitting negative binomial models available, and the scientific community on Twitter for suggesting writing up this analysis as a manuscript. V.S. was funded in part by the EMBL International PhD Programme and NIH U19MH114830.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0379-5.