UMI or not UMI, that is the question for scRNA-seq zero-inflation

Yingying Cao[1], *Simo Kitanovski*[1], *Ralf Küppers*[2] *& Daniel Hoffmann*[1]

[1]*Bioinformatics and Computational Biophysics, Faculty of Biology and Center for Medical Biotechnology, University of Duisburg-Essen, Essen, Germany*

[2]*Institute of Cell Biology (Cancer Research), University Hospital Essen, Essen, Germany*

Arising from Svensson, V. *Nature Biotechnology* https://doi.org/10.1038/s41587-019-0379-5 (2020).

In your January 2020 issue, Svensson[1] addressed the problem of zero-inflation in single-cell RNA-sequencing (scRNA-seq) data—that is, the observation that more genes in more cells than expected appear to have zero expression. Using examples, the Correspondence demonstrates that droplet-based methods that make use of Unique Molecular Identifier (UMI)[3,4] counts to quantify gene expression are adequately modeled with negative binomial distributions without zero-inflation. We agree with this, and we also share the concern that there is confusion about the validity of zero-inflation and the necessity of computational methods to eliminate it.

Even so, we find Svensson's subsequent discussion of plate-based scRNA-seq methods misleading because a reader not deeply immersed in the subject may be tempted to draw the conclusion that zero-inflation is a matter of the technical platform; specifically, that droplet-based scRNA-seq are not zero-inflated, whereas plate-based scRNA-seq are zero-inflated. Such a conclusion would be misguided and potentially could damage prospects for important technological developments and applications in the highly dynamic scRNA-seq field. We therefore felt the need for a clarifying response.

Our response can be stated crisply as follows: what matters most for zero-inflation with current scRNA-seq is not the technical platform (droplet versus plate), but whether gene

expression is measured in terms of UMI counts or read counts—suppressed zero-inflation with UMIs, stronger zero-inflation with read counts. Because for UMI count experiments, we typically have the raw read counts as well, this point can be made by direct comparison within the same experiment. Figure 1 shows exemplary cases from published data[7] for all four combinations of plate-based (Fig. 1a,c) and droplet-based (Fig. 1.b,d) scRNA-seq, with read counts (Fig. 1a,b) and UMI counts (Fig. 1c,d), demonstrating that even for heterogeneous samples not the platform but the use of UMIs makes the difference for zero-inflation. Supplementary Table 1 shows this for more data sets in a form similar to Table 1 of the correspondence[1].

The point that we are making here has been made already by others for data across technical platforms[5] and also specifically for droplet-based data[6], but it has apparently not received the necessary attention. Curiously, the Correspondence[1] itself mentions possible reasons for zero-inflation, including that the use of UMI counts deflates amplification bias, though the Correspondence[1] ignores that the use of UMI counts is not limited to droplet-based methods[2]. The main reason for suppressed zero-inflation with UMI counts is likely that UMI counts collapse multiple reads from the same original RNA molecule to a single read, thus also collapsing for many lowly expressed genes the gap between zero and non-zero expression that had been artificially widened by amplification. After this collapsing, the non-zero-inflated negative binomial is again the appropriate distribution.

Although measuring UMI counts is a good way to avoid zero-inflation problems, UMI counting is not a panacea. For instance, if accurate mapping of reads or detection of isoforms is a major objective of a scRNA-seq study, a tag-based protocol with UMIs could be less useful than a full-length sequencing protocol without UMIs.

Acknowledgements

We thank Deutsche Forschungsgemeinschaft for funding (grants KU1315/14-1 and HO1582/12-1).

Competing interests

The authors have no competing interests.

Figure 1. Comparison of importance of UMI and sequencing platform on zero inflation. Read counts for (**a**) plate-based CEL-Seq2 and (**b**) droplet-based Drop-seq versus UMI counts for (**c**) CEL-Seq2 and (**d**) Drop-seq on a sample of heterogeneous cells (peripheral blood monocytes; PBMCs)[7]. In the left hand plot of each panel (a–d), the solid curve is a least-squares fit (var = $\mu + \varphi\mu^2$, valid for negative binomial distribution with mean μ and dispersion φ, as in correspondence[1]) used to determine φ. In the right-hand plot of each panel (a–d) the solid curve is the predicted fraction of zeros with that φ. Density of actual scRNA-seq data is represented from low (light grey) to high (black). The data are clearly zero-inflated for both plate-based and droplet-based scRNA-seq with read count quantification (a,b), whereas for UMI count quantification zero-inflation is suppressed for both plate-based and droplet-based scRNA-seq (c,d).

1 Svensson, V. *Nature Biotechnology* 1–4 (2020).

2 Hashimshony, T. *et al. Genome biology* **17**, 77 (2016).

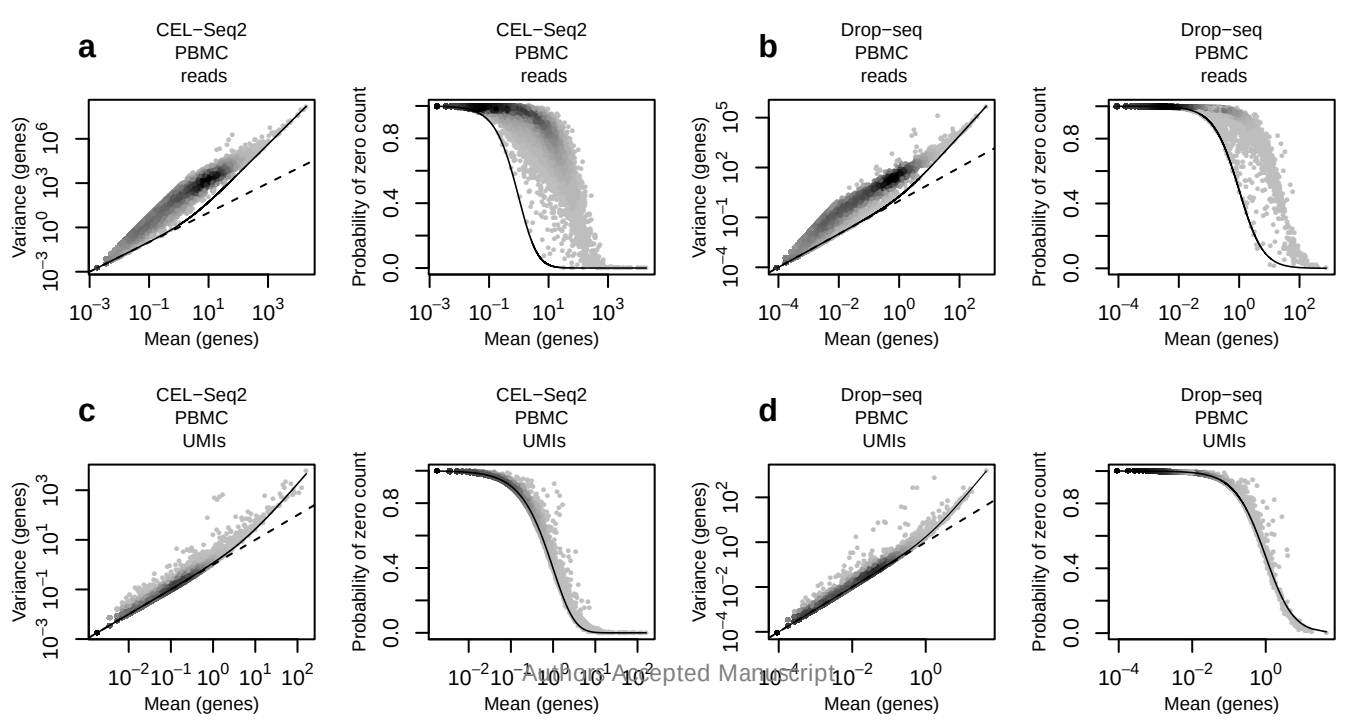
3 Kivioja, T. *et al. Nature methods* **9**, 72–74 (2012).

4 Islam, S. *et al. Nature methods* **11**, 163 (2014).

5 Chen, W. *et al. Genome biology* **19**, 70 (2018).

6 Townes, F. W. *et al. Genome biology* **20**, 295 (2019).

7 Ding, J. *et al. BioRxiv* 632216 (2019).



DOI: 10.1038/s41587-020-00810-6
URN: urn:nbn:de:hbz:465-20230912-163113-2