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

Supplementary Information

by

Beate Vieth[1], Christoph Ziegenhain[1], Swati Parekh[1], Wolfgang Enard[1] and Ines Hellmann[1]

[1]Anthropology & Human Genomics, Department of Biology II,
Ludwig-Maximilians University, Munich, Germany

1 Determining the best fitting distribution per gene

To determine the best fitting distribution to the observed RNA-seq count data, we compare the theoretical fit of the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) and Beta-Poisson (BP) distribution to the empirical RNA-seq read counts [2, 8, 3]. We used the following statistics to evaluate which distribution fits best:

- goodness of fit (GOF) statistics based on Chi-square statistic using residual deviances and degrees of freedom (Chi-square test).
- Akaike Information Criterium (AIC).
- Likelihood Ratio Test (LRT) for nested models, i.e. testing whether estimating a dispersion parameter in the NB models is appropriate.
- Vuong Test (VT) for non-nested models, i.e. testing whether assuming zero-inflation results in a better fit.
- Comparing the observed dropouts to the zero count prediction of the models.

Note that the goodness of fit statistics could not be calculated for the BP, however, since it already the AIC statistic suggested that the BP fit worse than the other distributions and could neither predict the dropouts correctly (Figure S1, Supplementary File S2), we did not follow this further.

We analyzed 8 published single cell RNA-seq studies ([1, 9, 11, 6, 7, 14, 13, 15]) produced using 9 different RNA-seq library preparation methods (Smart-seq/C1, Smart-seq2, MARS-seq, SCRB-seq, STRT, STRT-UMI, Drop-seq, 10XGenomics, CEL-seq2). For illustrative purposes, we focus on Kolodziejczk et al. (2015) [9], but the distribution analysis for all can be found in Supplementary File S2.

For the Kolodziejczk et al. (2015) data, we found that the NB distribution is an adequate fit (Figure S1): The Chi-Square test indicates that the NB is appropriate for at least 40 % of the genes (Figure S1 A). Moreover, the AIC suggests that the NB is in 60% of the cases better than the Poisson, ZIP, ZINB and BP (Figure S1 B). The ZINB is the only of the commonly used distributions that comes close, providing the best fit for 40% of all compared genes, however this difference is only significant for 6% (Figure S1D).

One of the major differences between the methods is the use of Unique Molecular Identifiers (UMIs) that allow for confident removal of PCR-duplicates [5, 15]. For all protocols considered, we evaluated the fit of the 5 different distributions, and for the vast majority the NB would be the distribution of choice (Figure S2). This is especially true for the UMI-methods: Here no zero-inflation is needed for modeling the gene expression distribution. On the contrary, also a simple Poisson often provides the best fit (Figure S4).

Next, we assess the fit of the dropout rate by comparing expected and predicted zero counts per gene. Interestingly, even though the negative binomial does not model dropouts explicitly, the deviation of predicted zero counts from the expected under the NB distribution is relatively small (Figure S1 C). The ZINB only gives a small advantage with respect to dropouts. The comparison of models by LRT and VT illustrates the small improvement of the model fit by assuming a ZINB distribution (10%) (FigureS1 D) for the Kolodziejczk data, which is comparable to the average for non-UMI methods, and much lower for the UMI-methods (<5%)(Figure S4 and Figure S3).

We thus refrain from using a mixture distribution, however for some of the protocols that do not utilize UMIs, such as e.g. Smart-Seq2, the ZINB might provide a better fit and should be used as a sampling distribution in the power simulations.

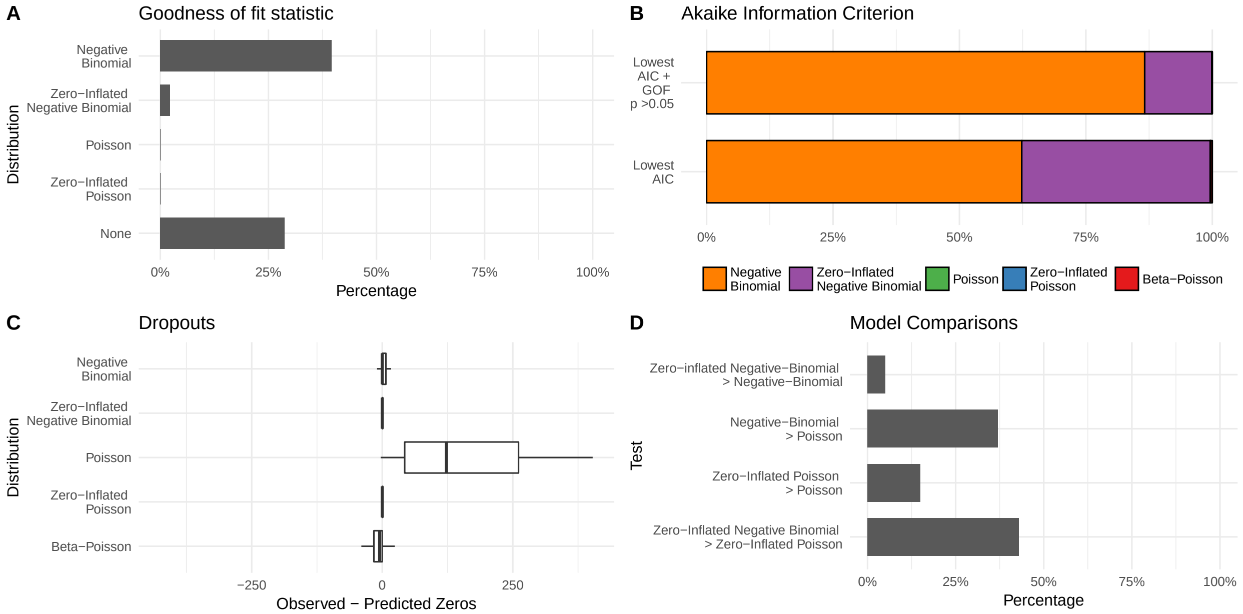

Figure S1: A) Goodness of fit of the model per gene assessed with a Chi-square test based on residual deviance and degrees of freedom. B) The fraction of genes for which the respective distribution has the lowest AIC and additionally the distribution with the lowest AIC as well as not rejected by the goodness of fit statistic. C) Observed versus predicted dropouts per distributional model and gene. D) Model assessment per gene based on Likelihood Ratio Test for nested models and Vung Test for non-nested models.
The same plot representing other datasets can be found in Supplementary File S2.

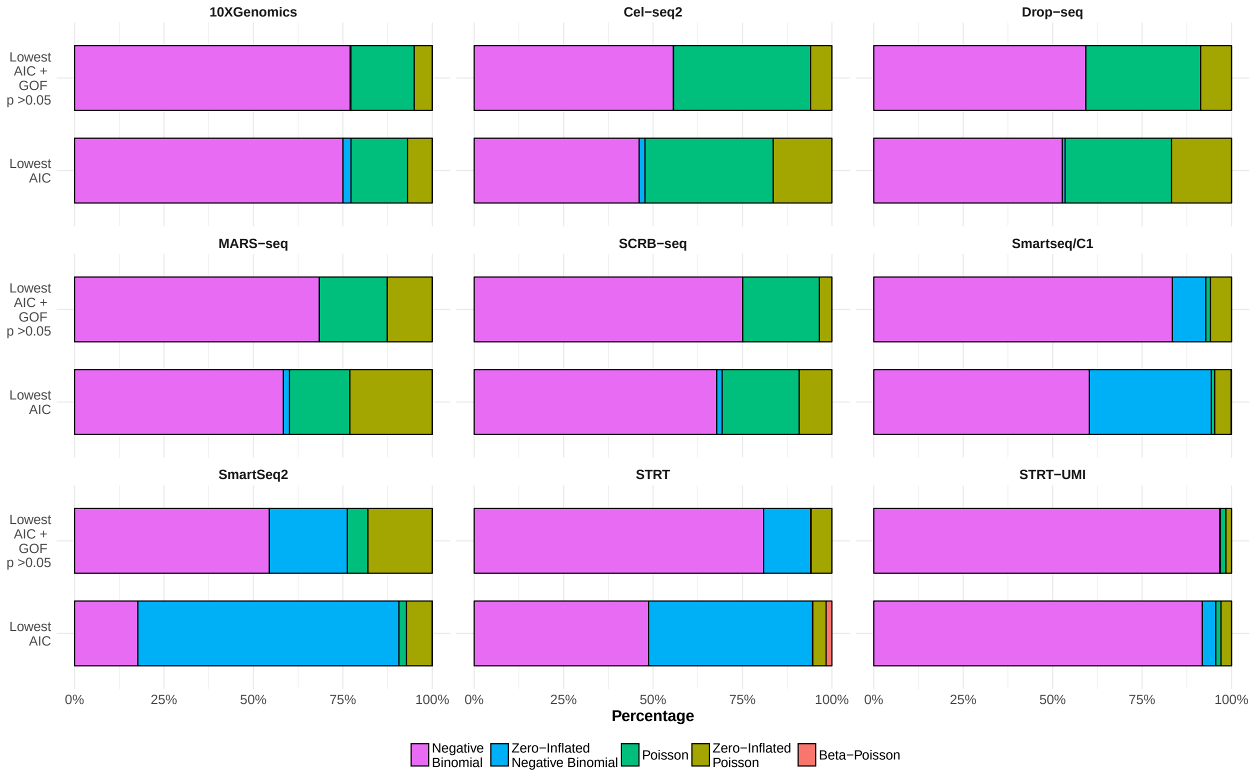

Figure S2: The negative binomial gives the best fit for the majority of genes (i.e. lowest AIC) for all UMI datasets. For protocols that do not account for PCR duplicates, the zero-inflated negative binomial often has a lower AIC, however this is mainly due to genes that cannot be fitted very well in general (GOF p-value<=0.05).

2 Read Count Simulation Framework

We have implemented a read count simulation framework assuming an underlying negative binomial distribution. To predict the dispersion θ given a random draw of an observed mean expression value μ, we apply a locally weighted polynomial regression fit. Furthermore, to capture the variability of the observed dispersion estimates, a local variability prediction band is applied (R package msir [12]). The read count for gene i in sample j is then given by:

$$X_{ij} \sim NB(\mu, \theta) \tag{1}$$

The mean, dispersion and dropout rates of an example read count simulation closely resembles the observed estimates for the Kolodziejczk data set (Figure S5).

For bulk RNA-seq experiments, the negative binomial alone is not able to capture the observed number of dropouts appropriately. Here, we predict the dropout probability (p_0) using a decreasing constrained B-splines regression (CRAN R package cobs [10]) of dropout rate against mean expression to determine the mean expression value μ_{DP5}, where the dropout probability is expected to fall below 5%. For all genes with

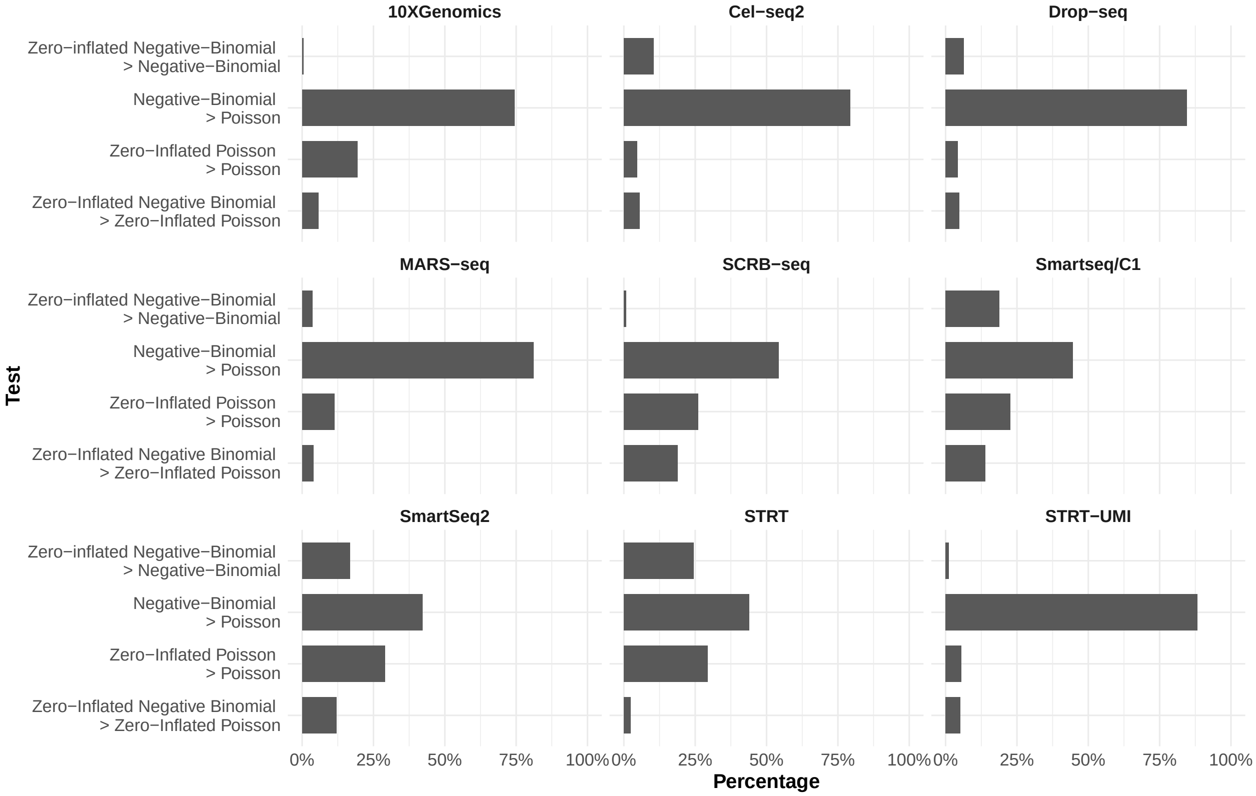

Figure S3: Model assessment per gene based on likelihood ratio test for nested models and Vuong test for non-nested models shows that zero-inflated negative binomial significantly improves the fit for maximally 25% of the genes (STRT protocol).

$\mu_i < \mu_{DP5}$ we do not estimate a gene specific dropout probability, but sample the dropout probability from all genes with $< \mu_{DP5}$. With these parameters, the read count for a gene i in a sample j is modeled as a product of a negative binomial multiplied with an indicator whether that sample was a dropout or not, which is determined using binomial sampling:

$$X_{ij} \sim I * NB(\mu, \theta), \text{where } I \in \{0, 1\} \tag{2}$$

$$P(I = 0) = B(1 - p_0) \tag{3}$$

The necessity of this apparently unintuitive zero inflation for bulk data is illustrated by the dataset from Eizirik et al. 2012 [4]. Note that dropouts occur across genes with different mean expression levels so that there is only a very weak relationship between mean expression and dropout probabilities (Figure S6).

For the simulations of expression changes, the user can freely define a distribution, a list of log_2-fold changes or simply a constant. We recommend to simulate with a realistic log_2-fold change distribution, which we determined for the Kolodziejczyk et al. (2015) [9] as a narrow $\Gamma(\alpha, \beta)$- distribution plus $-1 \times \Gamma(\alpha, \beta)$ (Figure S7).

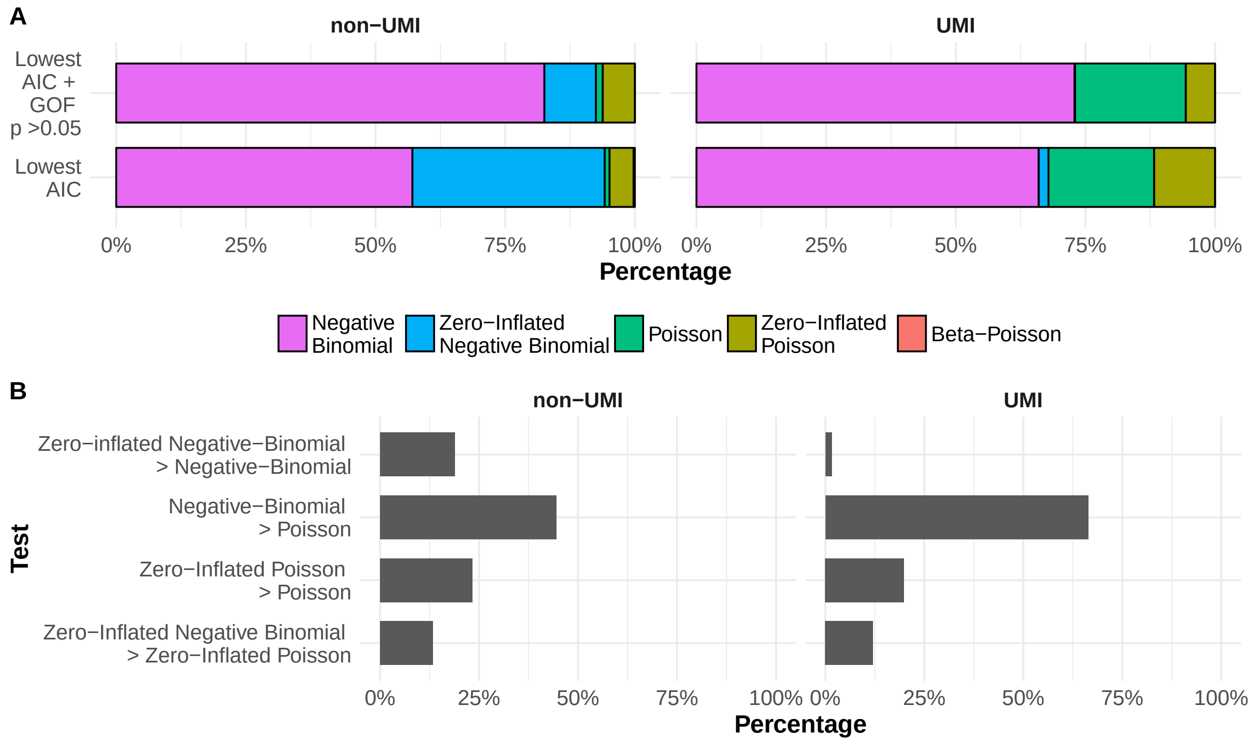

Figure S4: 6 UMI-protocols (STRT-UMI, Cel-SEq2, Drop-seq, MARS-seq, SCRB-seq,10XGenomics) are compared to 3 protocols not using UMIs (Smartseq/C1, SmartSeq2, STRT), showing that zero-inflation is only relevant for non-UMI-methods. A) The fraction of genes for which the respective distribution has the lowest AIC and additionally the distribution with the lowest AIC is not to rejected by the goodness of fit statistic. D) Model assessment per gene based on likelihood ratio test for nested models and Vuong test for non-nested models.

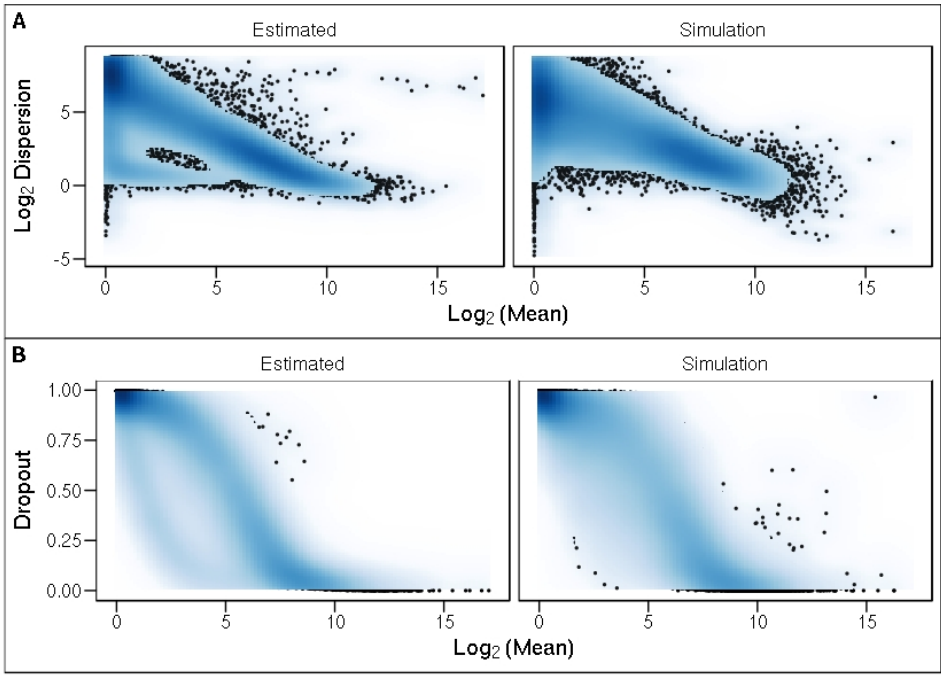

Figure S5: A) Dispersion versus mean. B) Dropout versus mean.

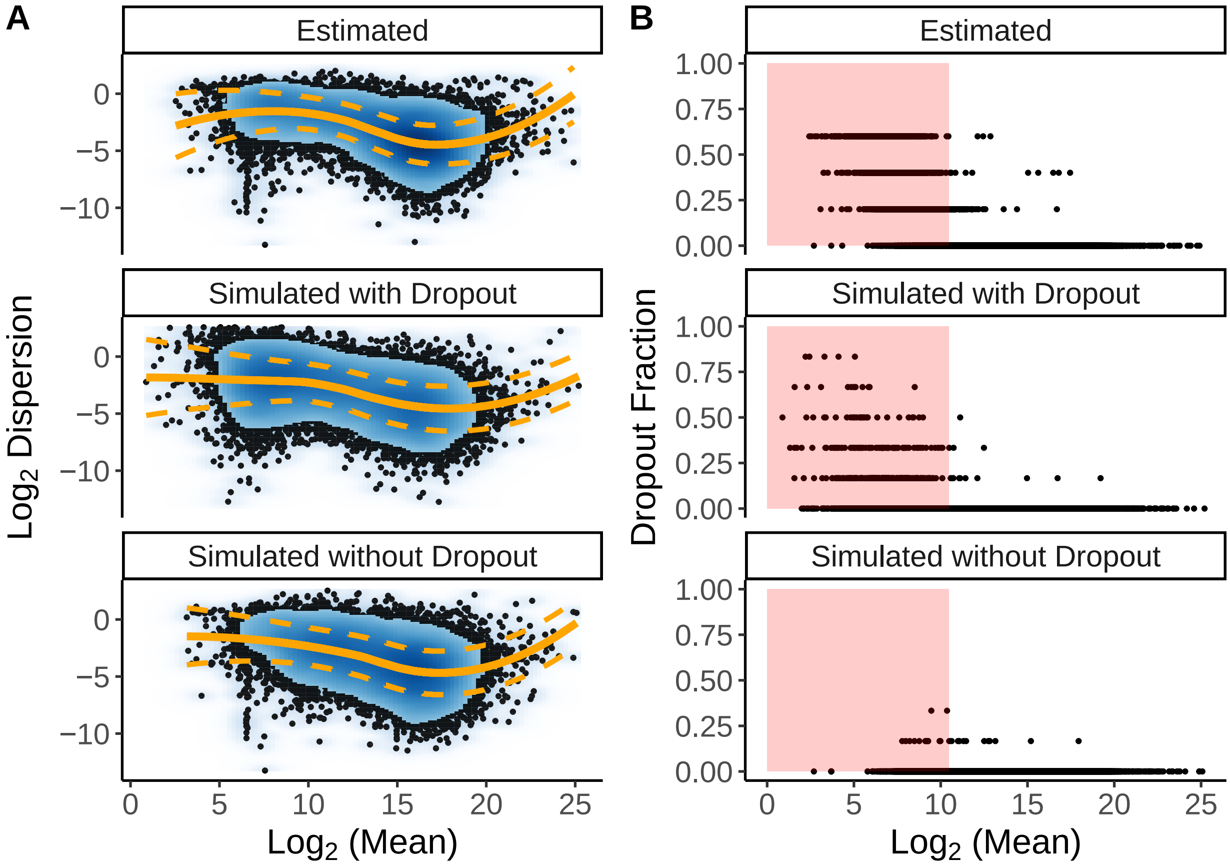

Figure S6: For bulk RNA-seq, the simulations include dropout sampling to better mimic the observed mean-dropout relation. A) Dispersion versus mean with locally weighted polynomial regression fit (orange line) and variability prediction band (dashed orange line). B) Dropout versus mean with red box indicating genes with $< \mu_{DP5}$ from which the dropout probability will be sampled from.

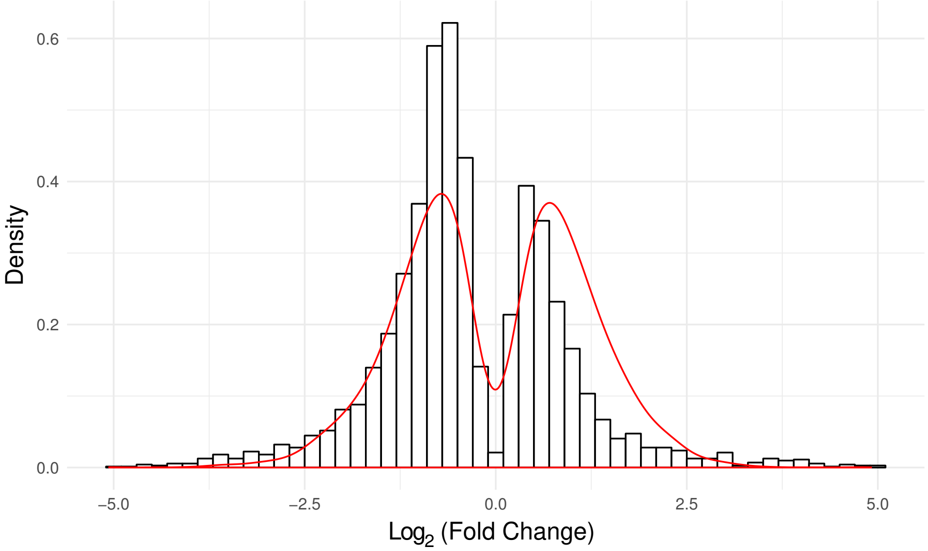

Figure S7: Log2 fold changes between serum+LiF and 2i+LiF cultured cells (Kolodziejczk et al. 2015). Red line indicates the density of a theoretical narrow gamma distribution (shape and rate equal to 3).

3 Included RNA-seq Experiments

We provide raw count matrices for several published single cell data sets (Table S1 on github (https://github.com/bvieth/powsimRData). Furthermore, the vignette gives an example on how to access RNA-seq datasets in online repositories such as recount (https://jhubiostatistics.shinyapps.io/recount/).

Table S1: Key properties of the example data-sets included in powsimR.

	Study	Accession	Species	No. Cells	Cell-type*	Library preparation	UMI	Remarks
1	Kolodziejczk et al. (2015) [9]	E-MTAB-2600	Mouse	869	ESC	Smart-seq C1	no	different growth media
2	Islam et al. (2011) [6]	GSE29087	Mouse	48	ESC	STRT-seq	no	-
3	Islam et al. (2014) [7]	GSE46980	Mouse	96	ESC	STRT-seq C1	yes	-
4	Buettner et al. (2015) [1]	E-MTAB-2805	Mouse	288	ESC	Smart-seq C1	no	FACs-sorted for cell-cycle
5	Soumillon et al. (2014) [13]	GSE53638	Human	12,000	adipocytes	SCRB-seq	yes	time-series

* ESC - embryonic stem cells

References

[1] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, advance online publication, 19 January 2015.

[2] A Colin Cameron and Pravin K Trivedi. *Regression Analysis of Count Data (Econometric Society Monographs)*. Cambridge University Press, 2 edition edition, 27 May 2013.

[3] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17:110, 29 February 2016.

[4] Décio L Eizirik, Michael Sammeth, Thomas Bouckenooghe, Guy Bottu, Giorgia Sisino, Mariana Igoillo-Esteve, Fernanda Ortis, Izortze Santin, Maikel L Colli, Jenny Barthson, Luc Bouwens, Linda Hughes, Lorna Gregory, Gerton Lunter, Lorella Marselli, Piero Marchetti, Mark I McCarthy, and Miriam Cnop. The human pancreatic islet transcriptome: Expression of candidate genes for type 1 diabetes and the impact of Pro-Inflammatory cytokines. *PLoS Genet.*, 8(3):e1002552, 2012.

[5] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, June 2014.

[6] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167, 1 July 2011.

[7] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, February 2014.

[8] Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.*, 14(1):R7, 28 January 2013.

[9] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason C H Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C Marioni, and Sarah A Teichmann. Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485, 1 October 2015.

[10] Pin Ng and Martin Maechler. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7(4):315–328, 2007.

[11] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, Naveen Ramalingam, Gang Sun, Myo Thu, Michael

Norris, Ronald Lebofsky, Dominique Toppani, Darnell W Kemp, Ii, Michael Wong, Barry Clerkson, Brittnee N Jones, Shiquan Wu, Lawrence Knutsson, Beatriz Alvarado, Jing Wang, Lesley S Weaver, Andrew P May, Robert C Jones, Marc A Unger, Arnold R Kriegstein, and Jay A A West. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, 32(10):1053–1058, October 2014.

[12] Luca Scrucca. Model-based sir for dimension reduction. *Computational Statistics & Data Analysis*, 5(11):3010–3026, 2011.

[13] Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, page 003236, 5 March 2014.

[14] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 16 January 2017.

[15] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, 16 February 2017.