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

April 13, 2017

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, zero-inflated Poisson and zero-inflated negative binomial and Beta-Poisson distribution to the empirical RNA-seq read counts [2, 6, 3]. We used the following statistics to evaluate which distribution fits best:

- goodness-of-fit 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 negative binomial model is appropiate.
- Vuong Test 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.

For Kolodziejczk et al. (2015) [7], we found that the negative binomial distribution is an adequate fit (Figure S1): The Chi-Square Test indicates an acceptable fit of the negative binomial for the majority of genes (Figure S1 A). Moreover, the AIC suggests that the negative binomial fits in 55% better than the Poisson, zero-inflated Poisson, Beta-Poisson and zero-inflated negative binomial (Figure S1 B). The zero-inflated negative binomial is the only of the commonly used distributions that comes close, providing the best fit for 45% of all compared genes, however this difference is only significant for 22% (Figure S1D).

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 negative binomial distribution is relatively small (Figure S1 C). The zero-inflated negative binomial only gives a small advantage with respect to dropouts. We thus refrain from using a mixture distribution. The comparison of models by LRT and Vuong illustrates the small improvement of the model fit by assuming a zero-inflated negative binomial distribution (22%) (FigureS1 D).

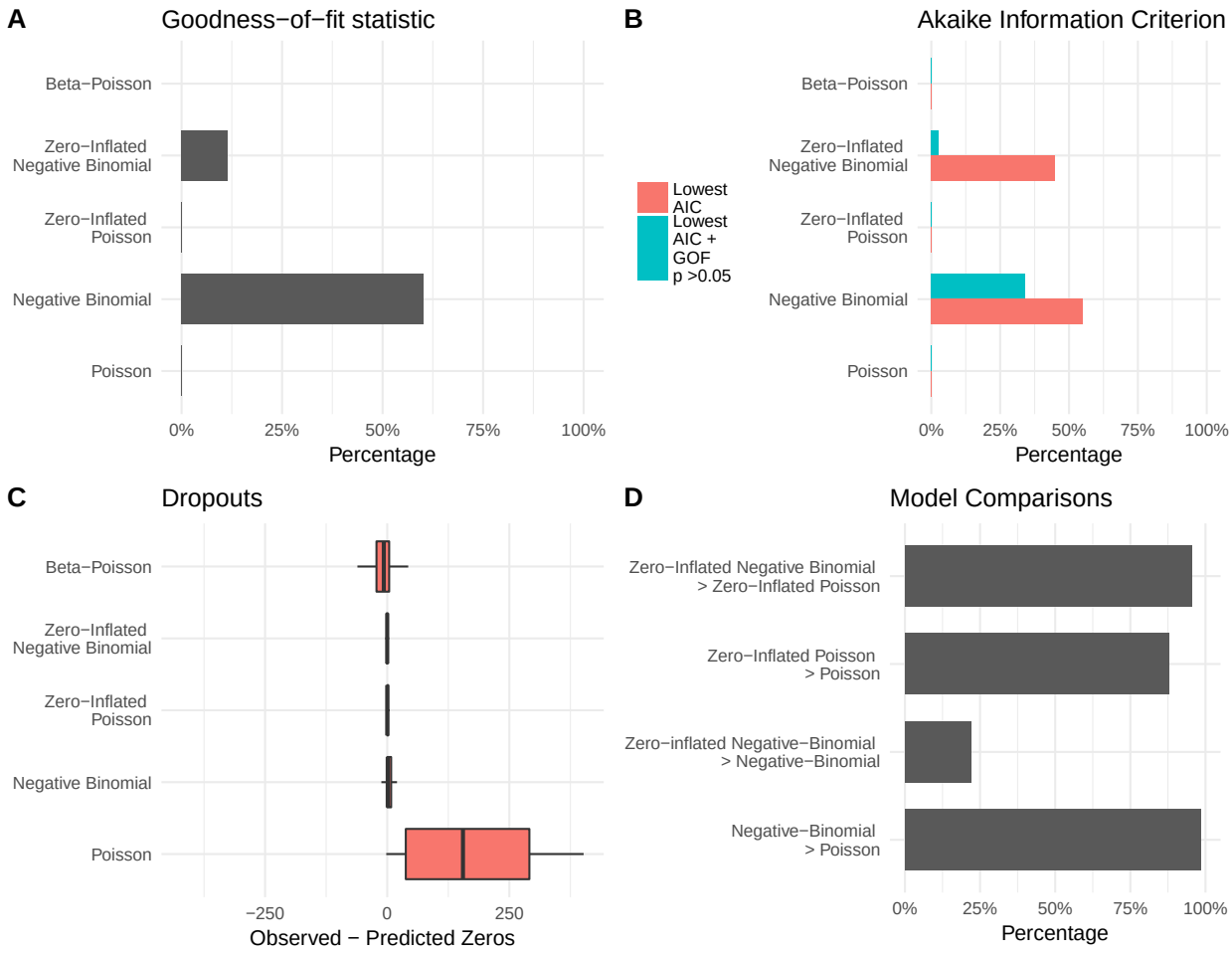

Figure S1: A) Goodness-of-fit of the model per gene assessed with a chi-square test based on residual deviance and degrees of freedom. B) Akaike Information Criterion per gene: Model with the lowest AIC (red). Model with the lowest AIC and passed goodness-of-fit statistic test (teal). C) Observed versus predicted dropouts per model and gene. D) Model assessment per gene based on likelihood ratio test for nested models and Vuong test for non-nested models.

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 [9]). 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 S2).

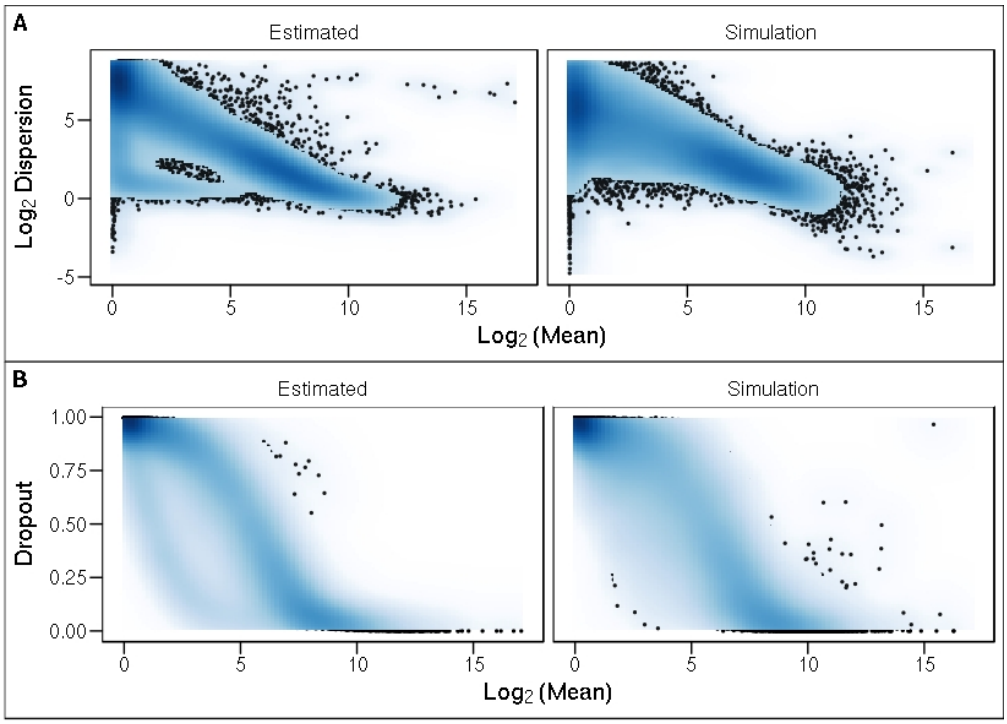

Figure S2: A) Dispersion versus mean. B) Dropout versus mean.

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 decrease constrained B-splines regression (CRAN R package cobs [8]) 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 $\mu_i < \mu_{DP5}$ 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:

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

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

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 realsitic log_2-fold change distribution, which we determined for the Kolodziejczyk et al. (2015) [7] as a narrow $\Gamma(\alpha, \beta)$- distribution plus $-1 \times \Gamma(\alpha, \beta)$ (Figure S3).

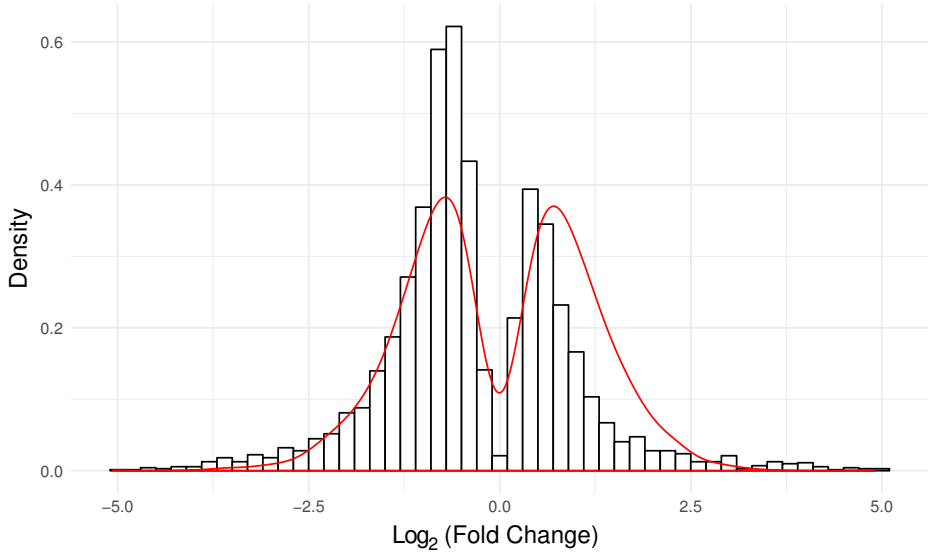

Figure S3: Log2 fold changes between serum+LiF and 2i+LiF cultured cells (Kolodziejczk et al.). Red line indicates the density of a theoretical narrow gamma distribution (shape and rate equal to 3).

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).

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

	Study	Accession	Species	No. Cells	Cell-type*	Library preparation	UMI	Special treatment
1	Kolodziejczk et al. (2015) [7]	E-MTAB-2600	Mouse	869	ESC	Smart-seq C1	no	different growth media
2	Islam et al. (2011) [4]	GSE29087	Mouse	48	ESC	STRT-seq	no	-
3	Islam et al. (2014) [5]	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) [10]	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] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.