Annales Mathematicae et Informaticae
53 (2021) pp. 283–298
DOI: https://doi.org/10.33039/ami.2021.03.003
URL: https://ami.uni-eszterhazy.hu

SimBPDD: Simulating differential distributions in Beta-Poisson models, in particular for single-cell RNA sequencing data

Roman Schefzik

German Cancer Research Center (DKFZ),
Heidelberg, Germany
Current affiliation: Medical Faculty Mannheim,
Heidelberg University, Germany
Roman.Schefzik@medma.uni-heidelberg.de

Submitted: December 19, 2020
Accepted: March 8, 2021
Published online: May 18, 2021

Abstract

Beta-Poisson (BP) models employ Poisson distributions, where the corresponding rate parameter itself is a Beta-distributed random variable. They have been shown to appropriately mimic gene expression distributions in the context of single-cell ribonucleic acid sequencing (scRNA-seq), a breakthrough technology allowing to sequence information from individual biological cells and facilitating fundamental insights into numerous fields of biology. A prominent scRNA-seq data analysis task is to identify differences in gene expression distributions across two conditions. To validate new statistical approaches in this context, one typically has to rely on accurate simulations, as usually no ground truth for an assessment is available. We introduce several simulation procedures that allow to generate differential distributions (DDs) based on BP models. In particular, we describe how to create different types of DDs, mirroring various sources or origins of a difference, and different degrees of DDs, from a weak to a strong difference. The soundness of the simulation procedures is shown in a validation study in which theoretically expected model properties of the DD simulations are confirmed. The findings are in principle not restricted to the scRNA-seq context and may be gener-

ally applicable also to other application areas. The simulation approaches are implemented in the publicly available R package `SimBPDD`.

Keywords: Beta-Poisson model, differential distributions, single-cell RNA sequencing, Wasserstein distance

AMS Subject Classification: 62P10, 62-04, 62-08, 92-08

1. Introduction

Beta-Poisson (BP) models, sometimes also referred to as Poisson-Beta models, employ Poisson distributions, where the corresponding rate parameter itself is a Beta-distributed random variable [3, 4]. Thus, the BP distribution is an example of a mixed Poisson distribution [6] and a discrete compound distribution, respectively. It is used in various theoretical and practical applications [8, 13, 15, 17].

Specifically, the BP distribution has been recently used in the biological context to model single-cell ribonucleic acid sequencing (scRNA-seq) data [15, 17]. Due to major technological advances, it is nowadays possible to sequence information from individual biological cells. Such single-cell sequencing, and in particular the scRNA-seq, enables the quantification of cellular heterogeneity and provides new fundamental insights into various biological fields [16], thus being highly relevant. Along with the ever increasing amount of produced scRNA-seq data, there is a need to develop statistical methods for the analysis of such data [1]. The most striking difference compared to previous data obtained by bulk experiments is that gene expression in scRNA-seq data is available over multiple cells and not only as an average single point value. Consequently, models for scRNA-seq gene expression should take the form of distributions. Moreover, they should take account of the specific nature of scRNA-seq data (e.g. abundance of zero expression or increased variability). Besides other approaches, the BP distribution considered in this paper has been shown to model scRNA-seq data appropriately, where there are different procedures for model fitting and estimation of model parameters [2, 15].

To evaluate novel statistical methods in scRNA-seq data analysis, simulations play a very important role, as typically no ground truth is available for real data. For instance, to adequately test and validate differential expression methods for scRNA-seq data [2], it is important to simulate differential distributions (DDs) in a reliable way. While there are already methods to do so [18], we here explicitly focus on a specific procedure to generate DDs in the scRNA-seq context using BP models. In particular, we describe how to create different types of DDs, mirroring various sources or origins of a difference, and different degrees of DDs, from a weak to a strong difference.

While the focus of this paper is on using BP models in the context of scRNA-seq data, the generation of DDs for the BP models is generally applicable also to other application areas, for both theoretical and practical considerations.

2. Beta-Poisson models in scRNA-seq data

We here consider Beta-Poisson (BP) models as introduced in [15], which have been found to appropriately mimic scRNA-seq data. Precisely, in [15], three different BP models are considered: a three-parameter BP model (BP_3), a four-parameter BP model (BP_4) and a five-parameter BP model (BP_5).

The BP_3 model is a mixture of Poisson distributions $\mathrm{Poi}(\lambda_1 u)$ with mean $\lambda_1 u$, where $\lambda_1 \in (0, \infty)$ denotes a scaling parameter and $u \sim \mathrm{Beta}(\alpha, \beta)$ has a Beta distribution with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$:

$$X \sim \mathrm{BP}_3(x|\alpha, \beta, \lambda_1) := \mathrm{Poi}(x|\lambda_1 \,\mathrm{Beta}(\alpha, \beta)).$$

Here, α is a shape parameter, where a large α indicates a high burst frequency (i.e. transcription rate, where transcription bursts correspond to an "on" state), and β is a scale parameter, with a large β indicating a high burst size [15, 17]. We can interpret this in the way that α may reflect among others the number of zero expression values (i.e. the proportion of zero expression), while β may mirror the size or magnitude of the non-zero expression values.

According to [15], the mean and the variance of the BP_3 model are given by

$$\mathbb{E}(X) = \lambda_1 \frac{\alpha}{\alpha + \beta}$$

and

$$\mathrm{Var}(X) = \lambda_1 \frac{\alpha}{\alpha + \beta} + \lambda_1^2 \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

respectively.

As the BP_3 model can only account for count data (i.e. non-negative integers), the BP_4 model is proposed in [15], which employs an additional parameter $\lambda_2 \in (0, \infty)$ to allow for modeling non-negative real-valued data, i.e., the usual data format we have to deal with after normalization of the raw scRNA-seq count data:

$$Y \sim \mathrm{BP}_4(x|\alpha, \beta, \lambda_1, \lambda_2) := \lambda_2 \,\mathrm{BP}_3(x|\alpha, \beta, \lambda_1).$$

In addition to what has been done in [15], straightforward calculations yield

$$\mathbb{E}(Y) = \lambda_2 \lambda_1 \frac{\alpha}{\alpha + \beta}$$

and

$$\mathrm{Var}(Y) = \lambda_2^2 \left(\lambda_1 \frac{\alpha}{\alpha + \beta} + \lambda_1^2 \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \right).$$

Finally, the BP_5 model has an additional parameter $p_0 \in [0, 1]$ explicitly capturing the proportion of cells with zero expression (besides the parameter α reflecting the burst frequency, as discussed before):

$$Z \sim \mathrm{BP}_5(x|\alpha, \beta, \lambda_1, \lambda_2, p_0) := p_0 \mathbb{1}_{\{x=0\}} + (1 - p_0)\,\mathrm{BP}_4(x|\alpha, \beta, \lambda_1, \lambda_2)\mathbb{1}_{\{x>0\}},$$

with $\mathbb{1}$ denoting the indicator function. In addition to what has been outlined in [15], by applying corresponding formulas for mixture distributions, it can be computed that

$$\mathbb{E}(Z) = (1 - p_0)\lambda_2\lambda_1\frac{\alpha}{\alpha + \beta}$$

and

$$\mathrm{Var}(Z) = (1 - p_0)(\mathbb{E}(Y)^2 + \mathrm{Var}(Y)) - \mathbb{E}(Z)^2.$$

Note that for $\lambda_2 := 1$ and $p_0 := 0$, the BP_4 and the BP_5 models actually reduce to the BP_3 model.

3. Simulating differential distributions for Beta-Poisson models

The starting point of our procedure is a pre-processed (including quality control and normalization) real-experiment scRNA-seq data set in form of a $(G \times C)$ expression matrix, with G denoting the number of genes and C the number of cells. We first fit a BP_5 model to the expression data for each gene separately using the method provided by [15] in the R [12] package `BPSC` and obtain corresponding parameter estimates $\alpha, \beta, \lambda_1, \lambda_2$ and p_0. Further, we test for each gene whether its distribution is indeed fitted well by the corresponding BP_5 model, using the procedure proposed in Section 3.2 in [15]. While filtering out low-quality fits, the cases (genes) that show a good fit, together with their corresponding parameter estimates, are kept in our pipeline and will be referred to as the controls in what follows.

We then simulate differential distributions (DDs) for each control Z separately by manipulating the corresponding parameters α, β and λ_1. We do not explicitly consider a manipulation of the parameter λ_2 here, as λ_2 only controls the transformation from (a discrete spectrum of) non-negative integers (expression counts) to (a discrete spectrum of) non-negative real values (expression after normalization). Moreover, we do not consider a manipulation of the parameter p_0 at this point; however, this will be discussed at the end of the section, when we explicitly describe how to construct differential proportions of zero expression (DPZ) in the context of BP_5 models.

Here, we consider multiplicative manipulations of the parameters α, β and λ_1 as follows, where we set $\lambda := \lambda_1$ for simplicity (as λ_2 is anyway not explicitly considered): $\lambda \mapsto \Delta_\lambda\lambda$, $\alpha \mapsto \Delta_\alpha\alpha$, $\beta \mapsto \Delta_\beta\beta$. The parameters obtained by (one or multiple of) these transformations are then the corresponding parameters in the manipulated BP_5 model $\tilde{Z}$. As $\lambda \in (0, \infty)$, $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$, it must hold that $\Delta_\lambda \in (0, \infty)$, $\Delta_\alpha \in (0, \infty)$ and $\Delta_\beta \in (0, \infty)$, respectively, to get a reasonable model.

We not only want to create DDs, but also to incorporate different degrees θ of DD that range from weak to strong differences. Here, a degree θ of DD between a control BP_5 model Z and a manipulated BP_5 model $\tilde{Z}$ is first introduced using a

multiplicative change (i.e., a fold change) with respect to the expected value:

$$\mathbb{E}(\tilde{Z}) = \theta\,\mathbb{E}(Z).$$

Inserting the corresponding expressions for the expected values and some algebra yields

$$\theta = \Delta_\lambda \cdot \frac{\Delta_\alpha\alpha + \Delta_\alpha\beta}{\Delta_\alpha\alpha + \Delta_\beta\beta}. \tag{3.1}$$

Hence, $\theta = \theta(\Delta_\lambda, \Delta_\alpha, \Delta_\beta)$ can be viewed as a function of $\Delta_\lambda, \Delta_\alpha$ and Δ_β, and the degree of DD can be specified by varying $\Delta_\lambda, \Delta_\alpha$ and Δ_β, respectively. Vice versa, we may consider Δ_λ as a function of θ in case Δ_α and Δ_β are fixed, i.e., $\Delta_\lambda = \Delta_\lambda(\theta)$. Analogously, $\Delta_\alpha = \Delta_\alpha(\theta)$ in case Δ_λ and Δ_β are fixed, and $\Delta_\beta = \Delta_\beta(\theta)$ in case Δ_λ and Δ_α are fixed. Note here again that θ generally refers to the degree of the DD (weak to strong), while $\Delta_\lambda, \Delta_\alpha$ or Δ_β represents the model manipulation that is necessary in order to achieve a DD of degree θ.

To get an understanding of the influence of the single parameter manipulations and to facilitate interpretability, we here first only consider those cases in which only one of the three original BP_5 model parameters is changed.

Case DLambda:

Here, the model is changed by manipulating the parameter λ only: $\lambda \mapsto \Delta_\lambda\lambda$, and $\Delta_\alpha = \Delta_\beta = 1$.

By inserting the corresponding expressions in (3.1), we get

$$\theta = \theta(\Delta_\lambda) = \mathbb{E}(\tilde{Z})/\mathbb{E}(Z) = \Delta_\lambda,$$

i.e.,

$$\Delta_\lambda = \Delta_\lambda(\theta) = \theta.$$

As $\theta(0) = 0$ and $\theta(\Delta_\lambda) \to \infty$ for $\Delta_\lambda \to \infty$, θ is bounded from below by zero, but has no upper bound. Hence, the range for possible values of θ is $\theta \in (0, \infty)$. DDs of arbitrary degree can be created using either positively oriented or negatively oriented fold changes. Here, a negatively oriented fold change means that $\mathbb{E}(\tilde{Z}) < \mathbb{E}(Z)$, hence, $\theta \in (0, 1)$. Conversely, a positively oriented fold change means that $\mathbb{E}(\tilde{Z}) > \mathbb{E}(Z)$, hence, $\theta \in (1, \infty)$. For instance, a positively oriented fold change of 3 in our setting here practically has the same effect as a negatively oriented fold change of 1/3, as we are only interested in the magnitude (i.e. the degree) of the difference here, and not in the direction of the change.

A manipulation of the scaling parameter λ in the BP model, while keeping $\text{Beta}(\alpha, \beta)$ unmodified, changes location (mean) and size (variance). In contrast, the shape should be affected only to a minor extent by a manipulation as here, if at all [15, 17]. Moreover, a change of λ should not affect the proportion of zero expression too much.

Case DAlpha:

Here, the model is changed by manipulating the parameter α only: $\alpha \mapsto \Delta_\alpha \alpha$, and $\Delta_\lambda = \Delta_\beta = 1$.

By inserting the corresponding expressions in (3.1), we get

$$\theta = \theta(\Delta_\alpha) = \mathbb{E}(\tilde{Z})/\mathbb{E}(Z) = \frac{\Delta_\alpha(\alpha+\beta)}{\Delta_\alpha \alpha + \beta},$$

i.e.,

$$\Delta_\alpha = \Delta_\alpha(\theta) = \frac{\beta\theta}{(\alpha+\beta) - \alpha\theta}.$$

As $\theta(0) = 0$ and $\lim_{\Delta_\alpha \to \infty} \theta(\Delta_\alpha) = 1 + \frac{\beta}{\alpha}$, θ is bounded from below by zero and has an upper bound $1 + \frac{\beta}{\alpha}$. Hence, the range for possible values of θ is $\theta \in (0, 1 + \frac{\beta}{\alpha})$. DDs can be generated using positively oriented (i.e. $\theta \in (1, 1 + \frac{\beta}{\alpha})$) or negatively oriented (i.e. $\theta \in (0, 1)$) fold changes. However, DDs of arbitrary degree can thus be created using negatively oriented fold changes only.

Here, location, size and shape can change. Also, a manipulation of α can affect the proportion of zero expression.

Case DBeta:

Here, the model is changed by manipulating the parameter β only: $\beta \mapsto \Delta_\beta \beta$, and $\Delta_\lambda = \Delta_\alpha = 1$.

By inserting the corresponding expressions in (3.1), we get

$$\theta = \theta(\Delta_\beta) = \mathbb{E}(\tilde{Z})/\mathbb{E}(Z) = \frac{\alpha+\beta}{\alpha + \Delta_\beta \beta},$$

i.e.,

$$\Delta_\beta = \Delta_\beta(\theta) = \frac{(\alpha+\beta) - \alpha\theta}{\beta\theta}.$$

As $\theta(0) = 1 + \frac{\beta}{\alpha}$ and $\lim_{\Delta_\beta \to \infty} \theta(\Delta_\beta) = 0$, θ is bounded from below by zero and has an upper bound $1 + \frac{\beta}{\alpha}$. Hence, the range for possible values of θ is $\theta \in (0, 1 + \frac{\beta}{\alpha})$. DDs can be generated using positively oriented (i.e. $\theta \in (1, 1 + \frac{\beta}{\alpha})$) or negatively oriented (i.e. $\theta \in (0, 1)$) fold changes. However, DDs of arbitrary degree can thus be created using negatively oriented fold changes only.

Here, location, size and shape can change. However, a manipulation of β should in principle not affect the proportion of zero expression too much.

We now consider a specific scenario, in which the expected value of the control BP_5 model is the same as that of the manipulated BP_5 model. The construction of such a type of DD may be relevant in case one wants to check whether an scRNA-seq differential expression analysis method is able to detect differences that are not caused by differences with respect to means [7].

Case DAlphaBeta:

Here, the model is changed by manipulating both the parameters α and β using a *common* parameter $\Delta := \Delta_\alpha = \Delta_\beta$: $\alpha \mapsto \Delta\alpha$, $\beta \mapsto \Delta\beta$, and $\Delta_\lambda = 1$.

As $\alpha, \beta \in (0, \infty)$, it must hold that $\Delta \in (0, \infty)$ to get a reasonable model. As discussed before, the expected value of the manipulated model $\tilde{Z}$ in this setting is the same as the expected value of the control model Z: $\mathbb{E}(\tilde{Z}) = \mathbb{E}(Z)$. We therefore introduce DDs by considering a multiplicative manipulation (i.e., a fold change) θ of the variance instead of the expected value, with somewhat more complex formulas involved:

$$\mathrm{Var}(\tilde{Z}) = \theta \ \mathrm{Var}(Z).$$

Thus,

$$\begin{aligned}
&\theta = \theta(\Delta) = \mathrm{Var}(\tilde{Z}) / \mathrm{Var}(Z) \\
&= \frac{1}{\mathrm{Var}(Z)} \left[(1-p_0) \left(\left(\lambda_1 \lambda_2 \frac{\alpha}{\alpha+\beta} \right)^2 + \lambda_2^2 \left(\lambda_1 \frac{\alpha}{\alpha+\beta} + \lambda_1^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\Delta(\alpha+\beta)+1)} \right) \right) \right. \\
&\qquad \left. -(1-p_0)^2 \left(\lambda_1 \lambda_2 \frac{\alpha}{\alpha+\beta} \right)^2 \right] \\
&= \frac{1}{\mathrm{Var}(Z)} \left[(1-p_0) \left(\mathbb{E}(Y)^2 + \lambda_2 \mathbb{E}(Y) + \frac{\lambda_1 \lambda_2 \beta \mathbb{E}(Y)}{(\alpha+\beta)(\Delta(\alpha+\beta)+1)} \right) - \mathbb{E}(Z)^2 \right],
\end{aligned}$$

i.e., after some tedious calculations,

$$\begin{aligned}
\Delta = \Delta(\theta) &= \frac{1}{\alpha+\beta} \left(\frac{\lambda_1^2 \lambda_2^2 \alpha\beta}{(\alpha+\beta)^2 \left(\frac{\mathrm{Var}(Z)\theta + \mathbb{E}(Z)^2}{1-p_0} - \mathbb{E}(Y)^2 - \lambda_2^2 \lambda_1 \frac{\alpha}{\alpha+\beta} \right)} - 1 \right) \\
&= \frac{1}{\alpha+\beta} \left(\frac{\lambda_1 \lambda_2 \beta \mathbb{E}(Y)}{(\alpha+\beta) \left(\frac{\mathrm{Var}(Z)\theta + \mathbb{E}(Z)^2}{1-p_0} - \mathbb{E}(Y)^2 - \lambda_2 \mathbb{E}(Y) \right)} - 1 \right).
\end{aligned}$$

For the degree of DD θ, we have the upper bound

$$\begin{aligned}
&L_{\mathrm{up}} := \theta(0) \\
&= \frac{1}{\mathrm{Var}(Z)} \left[(1-p_0) \left(\left(\lambda_1 \lambda_2 \frac{\alpha}{\alpha+\beta} \right)^2 + \lambda_2^2 \left(\lambda_1 \frac{\alpha}{\alpha+\beta} + \lambda_1^2 \frac{\alpha\beta}{(\alpha+\beta)^2} \right) \right) \right. \\
&\qquad \left. - \ (1-p_0)^2 \left(\lambda_1 \lambda_2 \frac{\alpha}{\alpha+\beta} \right)^2 \right] \\
&= \frac{1}{\mathrm{Var}(Z)} \left[\mathbb{E}(Z) \left(\mathbb{E}(Y) + \lambda_2 \left(1 + \frac{\lambda_1 \beta}{\alpha+\beta} \right) - \mathbb{E}(Z) \right) \right]
\end{aligned}$$

and the lower bound

$$L_{\mathrm{low}} := \lim_{\Delta \to \infty} \theta(\Delta)$$

$$= \frac{1}{\mathrm{Var}(Z)} \left[(1-p_0) \left(\left(\lambda_1 \lambda_2 \frac{\alpha}{\alpha+\beta} \right)^2 + \lambda_2^2 \lambda_1 \frac{\alpha}{\alpha+\beta} \right) - (1-p_0)^2 \left(\lambda_1 \lambda_2 \frac{\alpha}{\alpha+\beta} \right)^2 \right]$$
$$= \frac{1}{\mathrm{Var}(Z)} \left[\mathbb{E}(Z)(\mathbb{E}(Y) + \lambda_2 - \mathbb{E}(Z)) \right].$$

Hence, the range for possible values of θ is $\theta \in (L_{\mathrm{low}}, L_{\mathrm{up}})$, where $0 < L_{\mathrm{low}} < 1 < L_{\mathrm{up}}$. It is therefore not possible to create arbitrary degrees of DD in each case, be it for positively or negatively oriented fold changes with respect to the variance.

Note again that here, only size and shape change, but not the location. However, also the proportion of zero expression can change, as α varies, even though a variation of β should have no effect on this.

Finally, we consider the construction of manipulated BP_5 models with an explicit difference with respect to the proportion of zero expression, compared to the control model.

Table 1. Settings for the DD simulations based on BP models, where ✗ corresponds to "no" and ✓to "yes".

case	differential distributions	changed parameter(s)	same location	same size	same shape
DLambda	$Z \sim \mathrm{BP}_5(x\|\alpha,\beta,\lambda_1,\lambda_2,p_0)$ vs. $\tilde{Z} \sim \mathrm{BP}_5(x\|\alpha,\beta,\Delta_\lambda\lambda_1,\lambda_2,p_0)$	λ_1	✗	✗	✓
DAlpha	$Z \sim \mathrm{BP}_5(x\|\alpha,\beta,\lambda_1,\lambda_2,p_0)$ vs. $\tilde{Z} \sim \mathrm{BP}_5(x\|\Delta_\alpha\alpha,\beta,\lambda_1,\lambda_2,p_0)$	α	✗	✗	✗
DBeta	$Z \sim \mathrm{BP}_5(x\|\alpha,\beta,\lambda_1,\lambda_2,p_0)$ vs. $\tilde{Z} \sim \mathrm{BP}_5(x\|\alpha,\Delta_\beta\beta,\lambda_1,\lambda_2,p_0)$	β	✗	✗	✗
DAlphaBeta	$Z \sim \mathrm{BP}_5(x\|\alpha,\beta,\lambda_1,\lambda_2,p_0)$ vs. $\tilde{Z} \sim \mathrm{BP}_5(x\|\Delta\alpha,\Delta\beta,\lambda_1,\lambda_2,p_0)$	α,β	✓	✗	✗
DPZ	$Z \sim \mathrm{BP}_5(x\|\alpha,\beta,\lambda_1,\lambda_2,p_0)$ vs. $\tilde{Z} \sim \mathrm{BP}_5(x\|\alpha,\beta,\lambda_1,\lambda_2,p_0+\Delta_{p_0})$	p_0	✗	✗	✗

Case DPZ:

Here, the model is changed by manipulating the parameter p_0 of the control BP_5 model only: $p_0 \mapsto \tilde{p}_0 := p_0 + \Delta_{p_0}$, leading to differential proportions of zero expression (DPZ). While there is no intuitive feeling for the parameter ranges of the parameters λ, α and β, which is the reason why we used the models described above to construct different degrees of DDs for the other cases, we have an immediate and clear interpretability of the parameter p_0.
As it has to hold that $p_0 + \Delta_{p_0} \in [0,1]$ (since $p_0 \in [0,1]$), we choose Δ_{p_0} as follows:

$$\Delta_{p_0} = \begin{cases} \theta, & \theta \leq 1 - p_0, \\ -\theta, & \theta < p_0, \end{cases}$$

where $\theta \in (0, 0.5]$. A change of p_0 should obviously affect the proportion of zero expression.

For an overview of all the considered settings described before, which are partly similar to those in [14], see the summaries in Tables 1 and 2. For the cases DLambda, DAlpha and DBeta, it is recommended to only consider negatively oriented fold changes θ (i.e. $\theta \in (0, 1)$), as all possible degrees of DDs can be achieved only using them. For the case DAlphaBeta, all possible degrees of DDs indeed cannot be achieved with negatively oriented fold changes, but neither this works for positively oriented fold changes. Hence, for reasons of consistency, we by default also focus on negatively oriented fold changes then.

Table 2. General overview of the different manipulated models $\tilde{Z}$ of the control BP_5 models Z. Note that for the case DAlphaBeta, $L_{\mathrm{low}} := \frac{1}{\mathrm{Var}(Z)}\left[\mathbb{E}(Z)(\mathbb{E}(Y) + \lambda_2 - \mathbb{E}(Z))\right]$ and $L_{\mathrm{up}} := \frac{1}{\mathrm{Var}(Z)}\left[\mathbb{E}(Z)\left(\mathbb{E}(Y) + \lambda_2\left(1 + \frac{\lambda_1\beta}{\alpha+\beta}\right) - \mathbb{E}(Z)\right)\right]$. Here, Δ may refer to $\Delta_\lambda, \Delta_\alpha, \Delta_\beta$ or Δ_{p_0}, according to the descriptions of the corresponding cases in the main text.

case	manipulation	choice of Δ	possible values for θ
DLambda	$\mathbb{E}(\tilde{Z}) = \theta\,\mathbb{E}(Z)$	$\Delta = \theta$	$\theta \in (0, \infty)$
DAlpha	$\mathbb{E}(\tilde{Z}) = \theta\,\mathbb{E}(Z)$	$\Delta = \frac{\beta\theta}{(\alpha+\beta)-\alpha\theta}$	$\theta \in (0, 1 + \frac{\beta}{\alpha})$
DBeta	$\mathbb{E}(\tilde{Z}) = \theta\,\mathbb{E}(Z)$	$\Delta = \frac{(\alpha+\beta)-\alpha\theta}{\beta\theta}$	$\theta \in (0, 1 + \frac{\beta}{\alpha})$
DAlphaBeta	$\mathrm{Var}(\tilde{Z}) = \theta\,\mathrm{Var}(Z)$	$\Delta = \frac{1}{\alpha+\beta} \times \left(\frac{\lambda_1\lambda_2\beta\mathbb{E}(Y)}{(\alpha+\beta)\left(\frac{\mathrm{Var}(Z)\theta+\mathbb{E}(Z)^2}{1-p_0} - \mathbb{E}(Y)^2 - \lambda_2\mathbb{E}(Y)\right)} - 1\right)$	$\theta \in (L_{\mathrm{low}}, L_{\mathrm{up}})$
DPZ	$\tilde{p}_0 = p_0 + \Delta$	$\Delta = \begin{cases} \theta, & \theta \leq 1 - p_0 \\ -\theta, & \theta < p_0 \end{cases}$	$\theta \in (0, 0.5]$

4. Validation study

4.1. Evaluation tools

To validate the soundness of our simulation procedures, we employ the `waddR` tool available at `https://github.com/goncalves-lab/waddR`. Specifically, a semi-parametric, permutation-based test using the 2-Wasserstein distance is applied to compare two distributions F_A and F_B [10].

In our validation study, for each instance (here, each gene), information about F_A (the control model) and F_B (the manipulated model) is available in the form of a sample $x_{A,1}, \ldots, x_{A,C_A}$ from F_A, and $x_{B,1}, \ldots, x_{B,C_B}$ from F_B, respectively, where in general, C_A does not need to equal C_B. In the context of scRNA-seq data, the sample sizes C_A and C_B correspond to the respective numbers of cells. Using the corresponding empirical cumulative distribution functions $\hat{F}_A$ and $\hat{F}_B$ as

approximations, the (squared) 2-Wasserstein distance d is then computed by

$$\begin{aligned} d(\hat{F}_A, \hat{F}_B) &\approx \frac{1}{K}\sum_{k=1}^{K}(Q_A^{\alpha_k} - Q_B^{\alpha_k})^2 \\ &\approx \underbrace{(\hat{\mu}_A - \hat{\mu}_B)^2}_{\text{location}} + \underbrace{\underbrace{(\hat{\sigma}_A - \hat{\sigma}_B)^2}_{\text{size}} + \underbrace{2\hat{\sigma}_A\hat{\sigma}_B(1-\hat{\rho}_{A,B})}_{\text{shape}}}_{\text{variability}}, \end{aligned} \tag{4.1}$$

with $(Q_A^{\alpha_k})_{k=1,\ldots,K}$ and $(Q_B^{\alpha_k})_{k=1,\ldots,K}$ denoting the α_k-quantiles of $\hat{F}_A$ and $\hat{F}_B$, respectively, where we use equidistant levels $\alpha_k = \frac{k-0.5}{K}, k = 1, \ldots, K$. Here, $\hat{\mu}_A, \hat{\mu}_B$ denote the corresponding empirical means, $\hat{\sigma}_A, \hat{\sigma}_B$ the corresponding empirical standard deviations, and

$$\hat{\rho}_{A,B} := \mathrm{Cor}((Q_A^{\alpha_1}, \ldots, Q_A^{\alpha_K}), (Q_B^{\alpha_1}, \ldots, Q_B^{\alpha_K}))$$

the sample Pearson correlation coefficient between $(Q_A^{\alpha_k})_{k=1,\ldots,K}$ and $(Q_B^{\alpha_k})_{k=1,\ldots,K}$. For the calculations, we use $K := 1000$ here.

For each instance separately, we calculate the corresponding 2-Wasserstein distance as a test statistic and obtain a p-value using a semi-parametric, permutation-based testing procedure involving a generalized Pareto distribution approximation to estimate very small p-values accurately. Along with the p-value, the decomposition of the 2-Wasserstein distance in (4.1) may help to judge whether overall differences between two distributions (i.e. BP models) are mainly due to differences with respect to location (referring to differences with respect to the expected values), size (referring to differences with respect to the standard deviations) and/or shape (referring to differences not mainly caused by differences with respect to expected values and/or standard deviations) [5, 9].

To explicitly test for DPZ, we use Fisher's exact test, applied to each instance separately.

4.2. Setting

In the following validation study, we start with the real-experiment scRNA-seq data set in [11], downloaded from `https://hemberg-lab.github.io/scRNA.seq.datasets/human/tissues`. The data set consists of log2-transformed TPM (transcripts per million) expression values, normalized for both gene length and sequencing depth, for 301 cells, where we only keep those genes from the original data set that are expressed in at least three cells. A BP_5 model is fitted to each gene in the data set using the R package `BPSC` [15]. Specifically, maximum likelihood estimation combined with a binning approach to reduce computation time is employed to estimate the model parameters, using the standard R function `optim` for optimization. For more details, in particular the choice of initial values for the BP model optimization, see Section 3 in [15]. To assess the quality of the BP_5 model fits, a goodness-of-fit test statistic comparing the observed and expected frequencies

from the model is considered, where a Monte-Carlo method is used to generate a suitable null distribution that is employed to derive a corresponding p-value P. A gene is then declared to be fitted well by the BP model if $P \geq 0.05$. For details, see Section 3.2 in [15]. In our study, 8773 genes are declared to be fitted well by the corresponding BP_5 model. For the further analyses, we keep only these well-fitted genes as controls, from which manipulated BP_5 models are then constructed according to the procedures discussed before. At this point, we emphasize that for our purposes here, the real data set in [11] is only used for obtaining reasonable BP model parameters in the scRNA-seq context, from which the control and manipulated BP models in our purely numerical experiments are constructed. However, no specific biological investigations or aspects are considered in our validation study.

For each case (see Tables 1 and 2), we here consider five different degrees θ of DD, ranging from weak to strong, where the explicit choices for θ are shown in Table 3. Note that for the cases DLambda, DAlpha, DBeta and DPZ, these degrees of DD can be achieved for all 8773 genes in the simulation study, and all these genes are used in the studies. In contrast, for the case DAlphaBeta, due to the existence of the lower bound L_{low}, we only keep those genes for the study for which the corresponding degree of DD can be achieved.

As representatives of the corresponding control and manipulated BP_5 models, for each gene, we draw a sample from each model. In this context, the samples from a BP distribution are independent random draws. Specifically, the function `rBP` from the `BPSC` package is used for drawing the samples from the BP distributions, which combines the classical `rpois` and `rbeta` functions for randomly drawing from Poisson and Beta distributions, respectively, in R. In our study, we for convenience consider situations in which the sample size (i.e. the number of cells) $C := C_A = C_B$ in both conditions (control and manipulated) is equal and cover a range $C \in \{25, 50, 75, 100, 500\}$ of examples for small to large sample sizes.

Table 3. Different degrees of DDs specifically chosen in the simulation study.

case \ degree	D1	D2	D3	D4	D5
DLambda, DAlpha, DBeta, DAlphaBeta	$\theta = 10/11$	$\theta = 2/3$	$\theta = 1/2$	$\theta = 2/5$	$\theta = 1/3$
DPZ	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.25$	$\theta = 0.4$	$\theta = 0.5$
	weak	$\longrightarrow$			strong

4.3. Results

We now discuss the results for the validation study in terms of detection power and the decomposition of the 2-Wasserstein distance in the `waddR` test. In this context, for each fixed case, degree of DD and number of cells, detection power is defined as

$$\text{detection power} = \frac{\#\ \text{p-values} \leq \alpha}{\#\ \text{tests (genes)}},$$

Table 4. Detection powers (in %), based on p-values at a 5% significance level, with varying degrees of DD (D1: weak to D5: strong), numbers of cells $C \in \{25, 50, 75, 100, 500\}$ and cases from Table 1.

	case \ degree		D1	D2	D3	D4	D5
$C = 25$	DLambda	`waddR` DD	1.39	8.48	21.86	32.06	38.31
		Fisher DPZ	0.25	0.40	0.48	1.03	1.87
	DAlpha	`waddR` DD	1.57	7.67	18.03	26.73	32.73
		Fisher DPZ	0.30	1.45	5.11	10.48	16.95
	DBeta	`waddR` DD	1.61	8.36	20.43	30.42	36.24
		Fisher DPZ	0.25	0.42	0.96	1.74	2.74
	DAlphaBeta	`waddR` DD	1.35	2.50	5.50	0.87	12.66
		Fisher DPZ	0.20	1.64	7.55	16.53	25.67
	DPZ	`waddR` DD	1.37	2.09	23.50	54.21	65.61
		Fisher DPZ	0.26	0.39	12.71	68.97	77.03
$C = 50$	DLambda	`waddR` DD	2.17	17.91	40.61	51.61	58.00
		Fisher DPZ	0.40	0.52	1.58	3.21	5.49
	DAlpha	`waddR` DD	2.14	14.89	34.22	44.88	50.95
		Fisher DPZ	0.59	3.65	14.97	29.02	38.32
	DBeta	`waddR` DD	2.47	16.77	37.25	48.93	55.08
		Fisher DPZ	0.34	0.80	2.43	5.32	8.96
	DAlphaBeta	`waddR` DD	1.62	4.09	10.92	19.49	23.96
		Fisher DPZ	0.51	4.29	19.48	37.09	53.11
	DPZ	`waddR` DD	1.99	6.62	50.83	75.56	81.33
		Fisher DPZ	0.57	1.30	67.11	82.75	86.11
$C = 75$	DLambda	`waddR` DD	2.83	27.62	53.06	63.75	69.63
		Fisher DPZ	0.54	0.97	2.58	5.04	8.36
	DAlpha	`waddR` DD	2.72	22.61	44.74	56.35	62.93
		Fisher DPZ	0.68	6.38	26.38	42.00	51.76
	DBeta	`waddR` DD	2.36	25.35	49.29	60.31	67.05
		Fisher DPZ	0.51	1.42	4.34	9.18	15.22
	DAlphaBeta	`waddR` DD	1.83	6.01	18.10	26.30	32.86
		Fisher DPZ	0.83	7.54	33.13	56.49	70.34
	DPZ	`waddR` DD	2.62	12.19	66.05	81.98	84.94
		Fisher DPZ	0.81	7.61	79.41	86.78	88.93
$C = 100$	DLambda	`waddR` DD	3.00	34.41	60.21	71.21	76.56
		Fisher DPZ	0.51	0.95	3.08	6.84	11.23
	DAlpha	`waddR` DD	2.90	28.78	52.10	63.48	69.96
		Fisher DPZ	0.88	9.80	34.42	50.88	60.58
	DBeta	`waddR` DD	3.05	31.63	56.41	67.54	74.75
		Fisher DPZ	0.59	1.89	6.19	12.90	20.81
	DAlphaBeta	`waddR` DD	1.62	9.05	26.51	35.93	40.39
		Fisher DPZ	0.98	11.42	45.10	69.32	81.52
	DPZ	`waddR` DD	3.64	17.85	74.09	83.98	86.91
		Fisher DPZ	1.09	15.76	83.11	88.62	90.74
$C = 500$	DLambda	`waddR` DD	11.67	76.78	93.35	96.82	97.61
		Fisher DPZ	0.81	5.30	15.05	30.35	46.37
	DAlpha	`waddR` DD	9.86	68.94	87.43	92.57	94.85
		Fisher DPZ	2.68	53.03	78.10	88.13	92.83
	DBeta	`waddR` DD	10.40	72.12	90.07	94.94	96.90
		Fisher DPZ	0.93	11.66	38.47	58.55	68.68
	DAlphaBeta	`waddR` DD	3.04	62.78	78.12	85.17	89.33
		Fisher DPZ	2.87	59.94	92.03	96.65	98.57
	DPZ	`waddR` DD	23.74	70.61	87.99	91.51	93.25
		Fisher DPZ	27.50	82.00	92.10	96.57	98.34

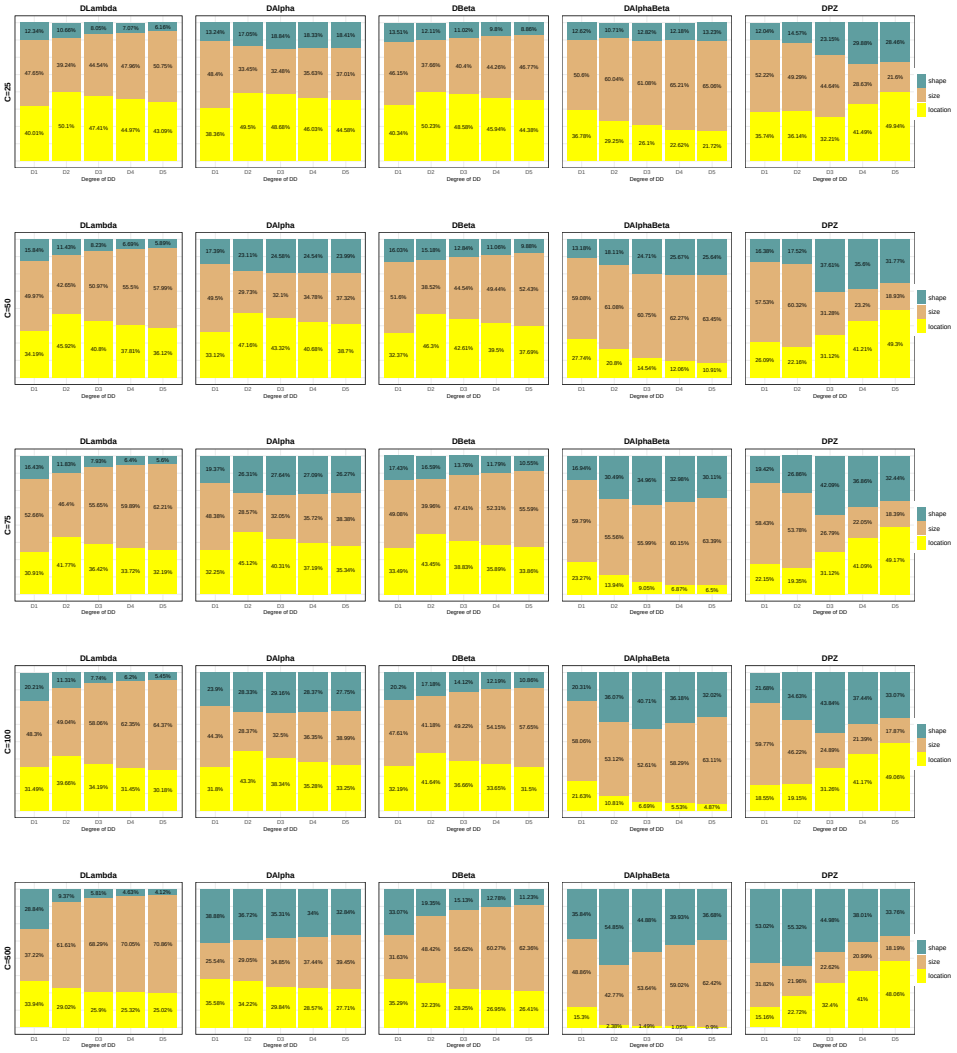

Figure 1. Decomposition results for the `waddR` test for varying degrees of DD (D1: weak to D5: strong) and numbers of cells $C \in \{25, 50, 75, 100, 500\}$, based on averages over those of the runs that are considered to show significant DDs in that the corresponding p-value is ≤ 0.05, with cases according to Table 1.

with $\alpha \in (0, 1)$. Detection powers for the different numbers of cells for the standard level of $\alpha = 5\%$ are listed in Table 4. In general, for all cases, detection powers meaningfully increase with increasing numbers of cells. Moreover, they increase with increasing strength of the difference between the distributions (weak to strong degree of DD; D1 to D5). While only very little detection power can be

observed for the very weak degree of DD D1, the detection powers get bigger for the stronger degrees of DD, for which the differences become more and more obvious. This intuitively makes sense and confirms in particular that the implementation of the varying degrees of DD from weak to strong in our simulation procedure is valid. When comparing the p-values of the `waddR` test and the separate DPZ test, we observe that DPZ can mainly be detected when the parameters α or p_0 are changed (i.e. in the cases DAlpha, DAlphaBeta and DPZ). In contrast, DPZ typically plays only a minor role when the parameters λ or β are changed (i.e. in the cases DLambda and DBeta). This is in line with the theoretical properties and the interpretation of the parameters of the BP model.

A further confirmation that the simulation procedures are able to reflect what is to be expected from the underlying theory (Table 1) of the BP model is given by the decomposition of the 2-Wasserstein distance into location, size and shape parts within the `waddR` test. For the different degrees of DDs and numbers of cells, Figure 1 shows for all cases the average fractions of the location, size and shape parts with respect to the overall 2-Wasserstein distance for the `waddR` test based on those runs with a p-value less than or equal to 5%. Again, the respective decomposition patterns meaningfully become more and more obvious the larger the number of cells is and the stronger the degree of DD is. In particular, the shape and location component in the cases DLambda and DAlphaBeta, respectively, are minor to negligible compared to the corresponding other components, in line with the theoretical models according to Table 1. Moreover, for instance, the shape component is more pronounced in cases in which the shape parameter α is changed (i.e. in the cases DAlpha and DAlphaBeta) than in those where α is not changed (i.e. in the cases DLambda and DBeta). An explicit change of the proportion of zero expression by manipulating the parameter p_0 (case DPZ) can obviously also affect the shape.

5. Discussion

We have discussed how to create DDs of varying degrees, ranging from weak to strong differences, for BP models, using various manipulations of the BP model parameters. The soundness of our approaches has been shown in a validation study, in which theoretically expected properties of our procedures have been confirmed.

In particular, based on the construction of our simulations and their validation in the study, we can provide some guidance on how to generate DDs between two BP models when the difference shall be of a specific type. For instance, when no difference with respect to shape is desired, the DLambda simulation, in which only the BP model parameter λ is changed, can be used. Similarly, in case no difference with respect to location is desired, one can employ the DAlphaBeta simulation, in which the BP model parameters α and β are changed using a common manipulation parameter. In case there shall be no DPZ, one may rely on the DLambda or DBeta simulations, in which only the BP model parameters λ and β, respectively, are changed.

Despite the focus of this paper is on the application field of scRNA-seq data, the introduced procedures can in principle be applied also to settings in other research areas.

While we have presented first attempts to simulate DDs for BP models here, we by far did not consider all possible combinations of BP parameter manipulations. This provides opportunities for future work, in which in particular interactions of changes when multiple BP parameters are manipulated simultaneously could be investigated in more detail. Moreover, up to now, only univariate BP distributions, that allow for individual (gene-wise) modeling, have been considered in the models here. However, certain variables may be correlated (such as genes in the scRNA-seq context), and taking account of specific correlation structures is an important issue that could be addressed in future extensions of the models.

Software availability. The simulations of DDs based on BP models presented in this paper are implemented in the R package `SimBPDD`, which is publicly available at `https://github.com/RomanSchefzik/SimBPDD`, along with documentation of the functions.

Acknowledgements. Angela Goncalves and Marc Schwering are thanked for helpful discussions and useful comments. Moreover, thanks are given to two anonymous reviewers for valuable comments and suggestions. The work was funded by project VH-NG-1010 of the HGF.

References

[1] R. Bacher, C. Kendziorski: *Design and computational analysis of single-cell RNA-sequencing experiments*, Genome Biology 17 (2016), art. 63, doi: https://doi.org/10.1186/s13059-016-0927-y.

[2] M. Delmans, M. Hemberg: *Discrete distributional differential expression (D^3E) – a tool for gene expression analysis of single-cell RNA-seq data*, BMC Bioinformatics 17 (2016), art. 110, doi: https://doi.org/10.1186/s12859-016-0944-6.

[3] J. Gurland: *A generalized class of contagious distributions*, Biometrics 14 (1958), pp. 229–249, doi: https://doi.org/10.2307/2527787.

[4] M. S. Holla, S. K. Bhattacharya: *On a discrete compound distribution*, Annals of the Institute of Statistical Mathematics 17 (1965), pp. 377–384, doi: https://doi.org/10.1007/BF02868181.

[5] A. Irpino, R. Verde: *Basic statistics for distributional symbolic variables: a new metric-based approach*, Advances in Data Analysis and Classification 9 (2015), pp. 143–175, doi: https://doi.org/10.1007/s11634-014-0176-4.

[6] D. Karlis, E. Xekalaki: *Mixed Poisson distributions*, International Statistical Review 73 (2005), pp. 35–58.

[7] K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, C. Kendziorski: *A statistical approach for identifying differential distributions in single-cell RNA-seq experiments*, Genome Biology 17 (2016), art. 222, doi: https://doi.org/10.1186/s13059-016-1077-y.

[8] K. L. Leask, L. M. Haines: *The beta-Poisson distribution in Wadley's problem*, Communications in Statistics–Theory and Methods 43 (2014), pp. 4962–4971, doi: https://doi.org/10.1080/03610926.2012.744047.

[9] Y. Matsui, M. Mizuta, S. Ito, S. Miyano, T. Shimamura: *D^3M: detection of differential distributions of methylation levels*, Bioinformatics 32 (2016), pp. 2248–2255, doi: https://doi.org/10.1093/bioinformatics/btw138.

[10] V. M. Panaretos, Y. Zemel: *Statistical aspects of Wasserstein distances*, Annual Review of Statistics and Its Application 6 (2019), pp. 405–431, doi: https://doi.org/10.1146/annurev-statistics-030718-104938.

[11] A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp II, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, J. A. A. West: *Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex*, Nature Biotechnology 32 (2014), pp. 1053–1058, doi: https://doi.org/10.1038/nbt.2967.

[12] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, url: https://www.R-project.org/.

[13] J. M. Sarabia, E. Gómez-Déniz: *Multivariate Poisson-Beta distributions with applications*, Communications in Statistics–Theory and Methods 40 (2011), pp. 1093–1108, doi: https://doi.org/10.1080/03610920903537269.

[14] M. Schwering: *Batch effects in single cell RNA sequencing analysis*, MA thesis, Heidelberg University, 2017.

[15] T. N. Vu, Q. F. Wills, K. R. Kalari, N. Niu, L. Wang, M. Rantalainen, Y. Pawitan: *Beta-Poisson model for single-cell RNA seq data analyses*, Bioinformatics 32 (2016), pp. 2128–2135, doi: https://doi.org/10.1093/bioinformatics/btw202.

[16] Y. Wang, N. E. Nevin: *Advances and applications of single-cell sequencing technologies*, Molecular Cell 58 (2015), pp. 598–609, doi: https://doi.org/10.1016/j.molcel.2015.05.005.

[17] Q. F. Wills, K. J. Livak, A. J. Tipping, T. Enver, A. J. Goldson, D. W. Sexton, C. Holmes: *Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments*, Nature Biotechnology 31 (2013), pp. 748–752, doi: https://doi.org/10.1038/nbt.2642.

[18] L. Zappia, B. Phipson, A. Oshlack: *Splatter: simulation of single-cell RNA sequencing data*, Genome Biology 18 (2017), art. 174, doi: https://doi.org/10.1186/s13059-017-1305-0.