Bioinformatics, 32(14), 2016, 2128–2135
doi: 10.1093/bioinformatics/btw202
Advance Access Publication Date: 19 April 2016
Original Paper

OXFORD

Gene expression

Beta-Poisson model for single-cell RNA-seq data analyses

Trung Nghia Vu[1], Quin F. Wills[2,3], Krishna R. Kalari[4], Nifang Niu[5], Liewei Wang[5], Mattias Rantalainen[1,†,*] and Yudi Pawitan[1,†,*]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK, [4]Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA and [5]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN 55905, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: Ziv Bar-Joseph

Received on December 1, 2015; revised on March 2, 2016; accepted on April 9, 2016

Abstract

Motivation: Single-cell RNA-sequencing technology allows detection of gene expression at the single-cell level. One typical feature of the data is a bimodality in the cellular distribution even for highly expressed genes, primarily caused by a proportion of non-expressing cells. The standard and the over-dispersed gamma-Poisson models that are commonly used in bulk-cell RNA-sequencing are not able to capture this property.

Results: We introduce a beta-Poisson mixture model that can capture the bimodality of the single-cell gene expression distribution. We further integrate the model into the generalized linear model framework in order to perform differential expression analyses. The whole analytical procedure is called BPSC. The results from several real single-cell RNA-seq datasets indicate that ~90% of the transcripts are well characterized by the beta-Poisson model; the model-fit from BPSC is better than the fit of the standard gamma-Poisson model in > 80% of the transcripts. Moreover, in differential expression analyses of simulated and real datasets, BPSC performs well against edgeR, a conventional method widely used in bulk-cell RNA-sequencing data, and against scde and MAST, two recent methods specifically designed for single-cell RNA-seq data.

Availability and Implementation: An R package BPSC for model fitting and differential expression analyses of single-cell RNA-seq data is available under GPL-3 license at https://github.com/nghiavtr/BPSC.

Contact: yudi.pawitan@ki.se or mattias.rantalainen@ki.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression at single-cell level is a stochastic process affected by extrinsic and intrinsic noise. The extrinsic noise comes from the environment outside the cell such as hormones or drugs. The intrinsic noise involves intra-cellular factors such as transcription rate (burst frequency), the number of mRNAs (burst size) and degradation rate (Wills *et al.*, 2013). Previous studies (Daigle *et al.*, 2015; Sanchez *et al.*, 2013; Shahrezaei and Swain, 2008) have introduced different models to capture the stochastic process due to the intrinsic noise. In general there are two main approaches to measure expression of genes in a single cell (i) targeted and (ii) whole transcriptome-based. The former—e.g. using qPCR or immunohistochemistry—is used to

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

investigate a limited number of genes, while the latter exploits the single-cell RNA sequencing (scRNA-seq) technology to estimate abundances of genes/transcripts of the whole transcriptome. In this study, we focus on modeling gene expression from scRNA-seq studies.

Gene expression of traditional bulk-cell RNA sequencing is the integration of gene expression from multiple single cells; the distribution is commonly modelled by the gamma-Poisson mixture model, or equivalently, the negative-binomial distribution. This model cannot capture the bimodality of gene expression in single-cell data (Shalek *et al.*, 2013). Recently Wills *et al.* (2013) briefly introduced the beta-Poisson model to capture the bimodality as well as the long-tailed behavior in the distribution. As an attractive feature, unlike the purely empirical gamma-Poisson model for bulk-RNA, the beta-Poisson model parameters introduce one biological interpretation in terms of burst size and burst frequency of the cell-level expression. However, due to computational problems with the beta-Poisson model, Wills *et al.* used an alternative discrete Poisson mixture, a less-parameterized version of the beta-Poisson model, to extract the information of the burst size and frequency. The model was applied for modelling of gene-level expression in qPCR-based datasets. Furthermore, no details of evaluation of modeling performance were reported. Another attempt of using the beta distribution to capture bimodality property of single-cell gene expression was introduced in the BATBayes model (Velten *et al.*, 2015). However, in this approach, instead of integrating it with the Poisson distribution, the beta distribution was combined with binomial distribution to mimic the partitioning process of RNA molecules into transcripts.

Here we develop BPSC, an analysis tool based on the beta-Poisson model for the single-cell gene expression data, and implement and apply it to several scRNA-seq datasets. BPSC addresses practical and realistic issues such as non-integer expression values or low expression values. Theoretically it is suitable for both transcript-level and gene-level expression, which is usually higher than the transcript-level expression. It is worth noting that the term 'transcript' herein includes both isoform and other types of splicing variants. The beta-Poisson model allows for control of the expression drop-off caused by technical noises or sequencing sensitivity. BPSC includes a generalized linear model (GLM) based on the beta-Poisson model to perform differential expression analyses of single-cell RNA-seq data. Experiments using simulated and real datasets show that BPSC performs well against edgeR, an established method for bulk-cell RNA-seq, and against MAST and scde, two recent differential-expression analysis methods designed for single-cell RNA-seq.

2 Beta-Poisson model

The beta-Poisson model captures the burst frequency and burst size through the shape and scale parameters α and β, respectively. Large α indicates high burst frequency; large β means large burst size (Wills *et al.*, 2013). Instead of using a range $[a, b]$, we consider the beta distribution in [0, 1] and scale it by λ in order to avoid optimizing two boundary parameters a and b. We then extend the model with more parameters to increase its flexibility. We start with the simplest model with three parameters, then describe its extensions with four and five parameters.

2.1 Three-parameter beta-Poisson model

The simplest beta-Poisson model is a mixture of Poisson distributions with mean $v = \lambda u$, where λ is a scale parameter, and u has a beta distribution with parameters (α,β). Thus, the probability distribution function (pdf) can be computed as follows:

$$P(X \leq x) = \int_0^{\lambda} P(X \leq x|v) f(v) dv. \tag{1}$$

Using $f(v)dv = f(u)du$, we first replace v by u to get

$$\begin{aligned} P(X \leq x) &= \int_0^1 P(X \leq x|\lambda u) f(u) du \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 P(X \leq x|\lambda u) u^{\alpha-1}(1-u)^{\beta-1} du \end{aligned} \tag{2}$$

Changing the variable $t \equiv 2u - 1$, so that $u = (t+1)/2$ and $dt = 2du$, with some algebra we have:

$$\begin{aligned} P(X \leq x) = \frac{1}{B(\alpha,\beta)} \frac{1}{2^{\alpha+\beta-1}} \int_{-1}^1 P(X \\ \leq x|\frac{\lambda(t+1)}{2})(1+t)^{\alpha-1}(1-t)^{\beta-1} dt \end{aligned} \tag{3}$$

The integral (3) can be easily and rapidly computed by using the Gauss-Jacobi quadrature method (Hildebrand, 2013). Thus, similar to the discrete Poisson mixture model (Wills *et al.*, 2013), our first model requires only three parameters. However, in practice, this model proves inadequate for modeling single-cell transcript expression data. Therefore, we extended the model in the following sections.

2.2 Four-parameter beta-Poisson model

The simple beta-Poisson model has a serious weakness in that its sample space contains non-negative integers only. Because of various preprocessing steps of the sequence data, expression values are generally not discrete counts. Transcript expression is typically estimated in the form of normalized indices such as fragments per kilo-base per million reads (FPKM) or counts-per-million reads (cpm). Figure 1a shows an example that the expression values (in FPKM) of a transcript are not discrete and small (mostly < 2.0). As displayed in the middle plot, this makes the three-parameter beta-Poisson model fail to capture the distribution. To address this problem, we introduce an extra parameter as follows. First we denote the three-parameter beta-Poisson model in a simpler form as

$$\mathrm{BP}_3(x|\alpha,\beta,\lambda_1) = \mathrm{Poisson}(x|\lambda_1 \mathrm{Beta}(\alpha,\beta)). \tag{4}$$

In this formula, we use λ_1 in place of λ in formula (1). Then, define the four-parameter beta-Poisson model with the extra parameter λ_2 as

$$\mathrm{BP}_4(x|\alpha,\beta,\lambda_1,\lambda_2) \equiv \lambda_2 \mathrm{BP}_3(x|\alpha,\beta,\lambda_1). \tag{5}$$

In practice we expect the parameter λ_2 to be a small positive value less than one, so that the BP_4 model allows fractional values in its sample space. The pdf of the four-parameter beta-Poisson model can be computed using the following connection:

$$\mathrm{BP}_4(x|\alpha,\beta,\lambda_1,\lambda_2) = \mathrm{BP}_3(\frac{x}{\lambda_2}|\alpha,\beta,\lambda_1) \tag{6}$$

The transcript expression in Figure 1a is now successfully modelled by BP_4 as presented in top-right plot of the panel. In this BP_4 model, the goodness-of-fit χ^2 is 5.79 with 4 degrees of freedom (P-value = 0.21), indicating the BP_4 model fits well and better than the BP_3 model.

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

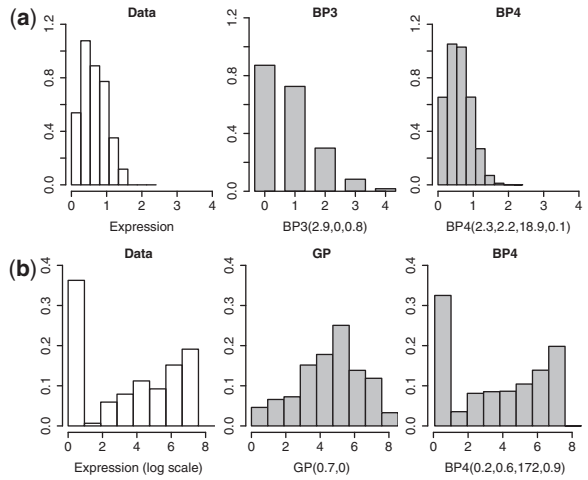

Fig. 1. (**a**) displays the expression distribution (in FPKM scale) of a low expression transcript (the left-most and white distributions) and its beta-Poisson models (the grey distributions) BP_3 and BP_4. Due to low and non-integer expression values, the BP_3 cannot capture the expression distribution of the transcript, but the BP_4 model can. The grey plots of (**b**) compare the performance of the gamma-Poisson (GP) model and the BP_4 model in capturing a distribution with a large proportion of unexpressed cells. The real data are from the MDA-MB-231 dataset

2.3 Five-parameter beta-Poisson model

Interestingly, even without any explicit parameter, the four-parameter model is able to some extent capture the bimodality in the expression distribution due to a large proportion of unexpressed cells. This is illustrated in Figure 1b, where the standard gamma-Poisson fails. To show more formally that the four-parameter model is often adequate for the data, we will compare it to a final extension that explicitly models the proportion of cells with zero expression. Zero expression is expected to represent a true underlying biology for a fraction of the cells, although it might also be due to technical reasons such as the detection limit of sequencing machine. Figure S1 in Supplementary file presents an example when the four-parameter model does not capture the distribution of expression data because of the large proportion of cells with zero expression.

The five-parameter beta-Poisson model adds an extra parameter p_0 capturing the proportion of cells with zero expression. The pdf of this model is given by a mixture model

$$\mathrm{BP}_5(x|\alpha,\beta,\lambda_1,\lambda_2,p_0) = p_0 I(x=0) + (1-p_0)\mathrm{BP}_4(x|\alpha,\beta,\lambda_1,\lambda_2)), \tag{7}$$

where $I(x=0)$ is the indicator function. Transcripts that fail the BP_4 model can be successfully modelled by the five-parameter model, as shown in Figure S1 in Supplementary file.

To summarize, in the original beta-Poisson model (Wills *et al.*, 2013), the two parameters of the beta distribution α and β represent the burst frequency and burst size in the transcription process. In our models, these burst parameters can be estimated by scaling α and β with λ_1 for the three-parameter model, and further modified by λ_2 for the four-parameter and five-parameter models. The fifth parameter p_0 controls the proportion of non-expressing cells.

3 Parameter estimation

For each transcript, our objective is to estimate the parameters of the beta-Poisson model given the expression data from n cells. From the pdf of the model, in principle we can compute the log-likelihood given the data

$$\log L(\theta) = \sum_x n(x)\log(p(x|\theta)),$$

where θ is the vector of parameters, $n(x)$ is the number of cells with expression equal to x, and $p(x|\theta)$ is the probability of x. For the three-parameter model $\theta = (\alpha, \beta, \lambda_1)$. To speed up the computation, we partition the data into K bins, so the log-likelihood is

$$\log L(\theta) = \sum_{k=1}^{K} n_k \log(p(k|\theta)),$$

where n_k is the number of cells in the kth bin, and $p(k|\theta)$ is the probability of the kth bin. The latter is computed from the pdf of the model. To assess the estimated model we compute a goodness-of-fit test, described in more detail in Section 3.2.

In practice we need to fit the model separately to each of >20 000 transcripts, thus requiring a fast, robust and reliable procedure. We select $K = 10$ to obtain a fast performance, while flexible enough to capture a variety of shapes in the data distribution. To make the procedure robust to outliers, we exclude cells from the top 2.5% expression from each transcript. The break points are selected to create approximately equal-sized bins. If the data are highly skewed or have a very long tail, we get the break points with equal-sized bins in a $\log_2$ scale; this is the case if most of the expression values are small but there exists a few high values so that the default break points in the original scale cannot properly capture information of the distribution. The optimization is implemented using the optim(.) function in the R software.

3.1 Initial values for model optimization

One important step to achieve a fast and reliable estimation is the selection of starting values. Here we use the method-of-moments estimates from the observed data. We first focus on the starting value estimation for the three-parameter model, then extend it to the other models. Let $E(X)$ and $Var(X)$ be the marginal mean and variance of transcript-expression X. Then we have the following results:

$$\begin{aligned} Var(X|\nu) &= E(X|\nu) = \lambda_1 u \\ E(X) &= E(E(X|\nu)) = \lambda_1 \frac{\alpha}{\alpha+\beta} \end{aligned} \tag{8}$$

$$\begin{aligned} Var(X) &= E(Var(X|\nu)) + Var(E(X|\nu)) \\ &= \lambda_1 \frac{\alpha}{\alpha+\beta} + \lambda_1^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned} \tag{9}$$

Given λ_1, we can estimate α and β by inverting (8) and (9) to get

$$\begin{aligned} \alpha &= \frac{E(X)}{\lambda_1}\left(\frac{E(X)(\lambda_1 - E(X))}{Var(X) - E(X)} - 1\right) \\ \beta &= \alpha\frac{\lambda_1 - E(X)}{E(X)}, \end{aligned}$$

and use the observed sample mean and sample variance for $E(X)$ and $Var(X)$, respectively.

We select the initial estimate of λ_1 to be the maximum of the data points. The starting value of λ_2 in the four-parameter model is set to be $0.1 \times$ median if the median expression value <1.0 (thus at least half of the cells have fractional expression value), otherwise $\lambda_2 = 1.0$. Then, we scale the expression values by dividing them with λ_2 before estimating the initial values of the other parameters. The initial value of mixture parameter p_0 of the five-parameter

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

beta-Poisson model is set to zero. Alternatively, we could fit a model to non-zero expressed values only; this approach leads to longer overall time and generally produces similar final estimates.

3.2 Model evaluation

In order to evaluate how well the model fits the data, the observed and the expected frequencies from the model are compared using the goodness-of-fit statistic

$$T = \sum_{k=1}^{K} n_k \log \frac{n_k}{e_k},$$

where n_k is the observed frequency of the kth bin and e_k is the expected frequency. Asymptotically, under the null hypothesis that the model fits the data, T has a χ^2 distribution with $(K - p - 1)$ degrees of freedom, where p is the number of parameters. However, because of small samples, many bins at the tail of the distribution always have low expected frequencies, so the asymptotic distribution is not appropriate.

We use a Monte-Carlo method to generate a more appropriate null distribution as follows: first 1000 random datasets are created, each containing n values from the estimated beta-Poisson model. Then we apply the estimation procedure to each random dataset and compute T^* as the goodness-of-fit statistic above. The collection of 1000 T^*s represents the Monte-Carlo null distribution, and the P-value is computed as the proportion of T^*s greater than the observed T. We declare a transcript is well-fitted by the beta-Poisson model if its P-value ≥ 0.05. An example of the Monte-Carlo null distribution is shown in Figure S2 in Supplementary report.

4 Differential expression analysis

In this section we integrate the beta-Poisson model into the generalized linear model (GLM) framework, so we can easily perform differential expression analyses. In general, given response variable $Y = y_i, i = 1, .., n$ and explanatory variable $X = x_j, j = 1, .., k$, a GLM consists of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \tag{10}$$

where the dependence of the expectation of the response variable $\mu_i = E(y_i)$ on the linear predictor is set by a link function $g(\mu_i) = \eta_i$. In this study, the distribution of the response variables is a beta-Poisson model and we use the log-link function. In order to estimate the model parameters we use the iterative weighted least-squares (IWLS) algorithm (Pawitan, 2013), which only requires a variance function that specifies the relationship between the mean and variance.

For the BP_3 model, using $\phi_1 = \frac{\alpha}{\alpha+\beta}$ and $\phi_2 = \frac{\beta}{\alpha(\alpha+\beta+1)}$, from formula (8) and (9), with some algebra we have:

$$\begin{aligned} \mu &\equiv E(X) = \lambda_1 \phi_1 \\ Var(X) &= \lambda_1 \phi_1 + \lambda_1^2 \phi_1^2 \phi_2 \\ &= \mu + \mu^2 \phi_2. \end{aligned} \tag{11}$$

Similarly, we extend the variance function for BP_4 with the additional parameter λ_2 as follows:

$$\begin{aligned} \mu &\equiv E(X) = \lambda_1 \lambda_2 \phi_1 \\ Var(X) &= \lambda_1 \lambda_2^2 \phi_1 + \lambda_1^2 \lambda_2^2 \phi_1^2 \phi_2 \\ &= \mu \lambda_2 + \mu^2 \phi_2. \end{aligned} \tag{12}$$

In group comparisons, we assume the mean μ varies across groups, but the shape parameters λ_2 and ϕ_2 are fixed and estimated from control group. In our implementation we utilize the glm(.) function in *R* software for fitting the GLM based on the BP_4 model with variance function defined by (12). The whole analytic procedure including the differential-expression analysis will be called the BPSC, which is also the name of the resulting R package made available at https://github.com/nghiavtr/BPSC.

5 Datasets

5.1 Real datasets

The first dataset in this study including 384 cells from a triple-negative breast cancer cell line (MDA-MB-231), half of which were treated with metformin. The cells were captured using the Fluidigm C1 system on a 96-well format. Two independent cell culture batches were used from which 2×96 untreated cells were captured, and similarly 2×96 treated cells. Sequencing libraries were prepared using the standard Fluidigm protocol based on SMARTer chemistry and Illumina Nextera XT kit. Library sizes of samples are in a range of 1–10 million.

The second dataset consists of 96 cells from HTC116 cell-line extracted from a public dataset (Wu *et al.*, 2014). These single-cell RNA-seq libraries were also prepared with SMARTer cDNA synthesis using the C1 microfluidic system (Fluidigm), based on Nextera library construction (Illumina). The 96 libraries, divided into two pooled samples of 48 libraries, were put in two lanes for a Illumina HiSeq sequencing. More details about this dataset are given in the original paper (Wu *et al.*, 2014). Library sizes in this dataset vary from 0.1 to 4 million.

There are various available transcript expression estimation methods that can be used for RNA-seq datasets. In this study, we selected a widely-used method Cufflinks (Trapnell *et al.*, 2010) to estimate transcript expression in the MDA-MB-231 dataset, and Sailfish (Patro *et al.*, 2014) to quickly extract transcript expression in the HTC116 dataset. The reference annotation was obtained from the most recent hg19 Homo sapiens reference built from UCSC data sources and extracted from igenomes website http://support.illumina.com/sequencing/sequencing_software/igenome.html (in folder archive-2014-06-02-13-47-56).

For the distribution analysis, the MDA-MB-231 dataset contains 21 728 transcripts from 165 cells from the control group and the HTC116 dataset consists of 23 889 transcripts from 96 cells. The cells from treated group were excluded in order to avoid the confounding effects of the metformin treatment.

For differential expression analysis, the MDA-MB-231 dataset includes also transcript-level expression of 162 cells from the treated group. Moreover, we use Htseq software (Anders *et al.*, 2015) to generate gene-level expression of the MDA-MB-231 dataset. This dataset consists of 12 079 genes in total.

5.2 Simulated datasets

We use two simulation settings to evaluate and compare the performance of BPSC in differential expression analyses. In each setting, the data consist of two equal-sized groups with 100 samples in each group and a total of 10 000 genes measured per sample. Five percent of the genes are set to be differentially expressed (DE) with fold-change 4.0, half of which are upregulated and the rest downregulated.

The first setup *bcSim* simulates bulk-cell RNA-seq data. We use the simulation design of a recent study (Law *et al.*, 2014) to create a

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

bulk-cell RNA-seq dataset. In brief, a baseline distribution from the real dataset is used to generate relative proportion of expected counts of genes. To create biological variation of DE genes, we multiply their proportions by the fold-change. Then the proportion multiplies by library size to obtain expected count of gene. In this simulation, the library sizes of the samples are randomly selected in a range from 2 to 20 million. An inverse-chi-squared distribution is used to create dispersions in the simulation (Law *et al.*, 2014). After that, the counts of genes are generated from the gamma-Poisson model with the specified mean and dispersion.

The second simulation setup *scSim* is designed to mimic a single-cell RNA-seq dataset. We use the well fitted four-parameter beta-Poisson models from HTC116 (Monte-Carlo *P*-value $\geq$5%) as baseline distributions for the gene expression of the dataset. For each gene, the expression across samples in the control group and the treated group is generated from the same beta-Poisson model. To create biological effects for the DE genes, we multiply the parameter λ_1 of one group by the fold-change, while the other parameters are kept fixed. We also scale the gene expression to the predefined library sizes of the cells, which are randomly sampled from a range of 1–3 million. This range is taken from the HTC116 dataset.

In the analyses, we filter out genes whose total reads across all samples <10. We generate 100 replications for each setting, and the final results are based on these replications.

6 Results and discussion

6.1 Model fitting

The analysis results for the two real datasets are summarized in Table 1. In general, these two datasets have similar performances, where ~90% of transcripts are successfully modeled (Monte-Carlo *P*-value $\geq$5%) by the BP_4 and BP_5 models. This proportion increases to 94% if we set the *P*-value threshold to 1% (data not shown). It is worth noting that the BP_4 model is only slightly worse than the BP_5 model. As illustrated in Figure 1b, the BP_4 model is already able to capture the proportion of unexpressed cells, so it appears sufficient for practical applications. The BP_5 model may need to be utilized only for the failed BP_4 models to improve overall performances. The poor performances of the BP_3 model in both datasets demonstrate the important effect of the λ_2 fractional-scaling parameter in the BP_4 model.

To demonstrate the value of the beta-Poisson model, we also compare its performance to the traditional gamma-Poisson (GP) model through their Akaike information criterion (AIC) scores. As shown in Figure 1b the beta-Poisson model performs much better than the gamma-Poisson model for transcripts with many unexpressed cells. To have a more even comparison, hence harder for the beta-Poisson model, we compare the simplest model BP_3 and GP on a subset of high-abundance transcripts which are defined by the transcripts that have top 10% sums of expression across cells. From Table 2, for more than 80% of the transcripts in both datasets, the AIC of the BP model (AIC_{BP}) is less than the AIC of the GP model (AIC_{GP}), indicating the BP model has better performance than the GP model. Figure 2 shows the scatterplot of the AIC scores from the two models applied to the MDA-MB-231 dataset. A great majority of the points falls below the line of identity, indicating that the BP model has lower AIC than the GP model.

To capture the improvement, we also compare the AIC of the BP_3 model with the AIC of the GP model up to a constant difference, i.e. we count the transcripts satisfying the condition $AIC_{BP} + c < AIC_{GP}$, where $c = 0, 2, 5$ in Table 2. The proportion is ~60% when we allow an AIC difference of 5.

To investigate why the BP model still sometimes fails to fit the data, we bin the expression values of the MDA-MB-231 data into a high resolution histogram with 100 equal-size breaks. Then a hierarchical clustering is applied to investigate transcript similarities, and we identify 4 groups categorizing the failed models. These groups are shown in Figure S3 in Supplementary report. Most of the failed models are in groups 1 and 3, where the right tail of the distribution is marked by several extreme values. These values will be investigated further in our future study.

To assess the sensitivity of the beta-Poisson model (BP_4 model) to sample size, we simulate datasets from *scSim* setup with different numbers of cells 100, 50, 25 and 20. The results (Fig. S4 in Supplementary report) show that the proportion of genes that fits the beta-Poisson model (*P*-value $\geq$0.05) is higher than 85% for the dataset with at least 25 cells. Therefore, we recommend that BPSC should be applied for datasets with at least 25 cells. Besides, due to the requirements for computing the goodness-of-fit statistic, the

Table 1. Proportion of transcripts that fits the beta-Poisson model (*P*-value $\geq$ 0.05)

	BP_3	BP_4	BP_5
MDA-MB-231	0.31	0.87	0.90
HTC116	0.37	0.88	0.91

Table 2. Comparisons between BP_3 and the standard gamma-Poisson model , where *c* represents the level of improvement in the AIC score.

c	0	2	5
MDA-MB-231	0.82	0.73	0.59
HTC116	0.88	0.78	0.60

The entries in the table are the proportion of transcripts where $AIC_{BP} + c < AIC_{GP}$, where *c* represents the level of improvement in the AIC score.

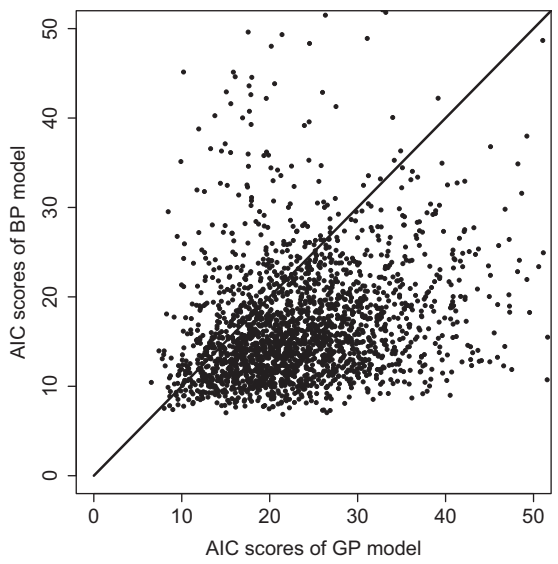

Fig. 2. BP3 model versus GP model in AIC scores from MDA-MB-231 dataset. In more than 80% of the transcripts, the AIC of the BP model (AIC_{BP}) is less than the AIC of the GP model (AIC_{GP})

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

number of bins K also should be greater than $p+1$. For example, number of bins K of the BP_4 model is at least 6 for practical use.

Differential expression analysis

To evaluate the performance of BPSC for differential expression analysis, we use the four-parameter beta-Poisson model combined with the generalized linear model. We compare the results to edgeR-glm of the edgeR software version 3.10.2 (McCarthy *et al.*, 2012; Robinson *et al.*, 2010) as a representative of methods for bulk-cell RNA-seq data analyses based on the negative-binomial distribution or equivalently the gamma-Poisson distribution. The edgeR software is applied on the read-count dataset as the input requirements of the software, and a normalization step is integrated inside the software. We use the protocol in a recent study (Anders *et al.*, 2013) for edgeR-glm implementation. We also compare BPSC to MAST version 1.0.1 (Finak *et al.*, 2015), a recent method for differential expression analysis in scRNA-seq data. In brief, MAST proposes to use a hurdle model (Finak *et al.*, 2015) to capture bimodality of gene expression, then apply likelihood ratio to test for differential expression.

BPSC works on the data normalized to library sizes from cpm function of the edgeR software (in the simulated datasets) or directly on the FPKM data (in the real datasets). In the original study, MAST was input from $\log_2(\text{TPM}+1)$ expression matrix. Herein, TPM (or transcripts per million) indicates normalized expression data. In our implementation, we replace TPM by FPKM for the real dataset and cpm for simulated datasets. Default parameter settings of the software are applied. Finally, we extract the P-values from each method and report the false discovery rates (Pawitan *et al.*, 2005), called estimated FDR. In simulated datasets, we are also able to calculate the true FDR by comparing the top genes ranked by the P-values to the true status of DE genes. These FDR values are ranked in increasing order and used for method comparison in Figure 3. In this figure, the method with a lower FDR at a certain number of top genes (transcripts) is better.

We first compare BPSC, MAST and edgeR in the two simulation setups: bcSim and scSim. The bcSim uses the gamma-Poisson model to generate a traditional bulk-cell RNA-seq dataset, while scSim simulate a single-cell RNA-seq dataset from the beta-Poisson model. We report the average FDR values of 100 simulations, and for visual purposes we present the top 1000 genes for comparison. For the bcSim setting, Figure 3a shows that BPSC, MAST and edgeR have similar performances, both in terms of estimated FDR (left) and true FDR (right). The estimated FDRs from BPSC and MAST are close to the true FDR when the true value <0.2, and they slightly under-estimate when the value FDR >0.2. These results indicate that BPSC is also able to work well on the conventional bulk-cell RNA-seq data in this experiment.

Figure 3b shows that BPSC outperforms MAST and edgeR in both the true FDR (right) and estimated FDR (left). Moreover, Figure S5 of Supplementary report, a zoomed-out of Figure 3 with more genes and a larger FDR range, shows BPSC and MAST are more reliable than edgeR in the scSim setting: there is a large discrepancy between the estimated FDR and the true FDR curves of the edgeR method, which is in contrast to the curves from BPSC. Since we have shown above that the beta-Poisson model fits the single-cell data well, this is an indication that edgeR may not be suitable for single-cell RNA-seq data.

For the MDA-MB-231 dataset, in Figure 3c (left plot) BPSC shows lower FDR than both MAST and edgeR. Moreover, in the zoomed-out version (Fig. S5c of Supplementary report), the pattern

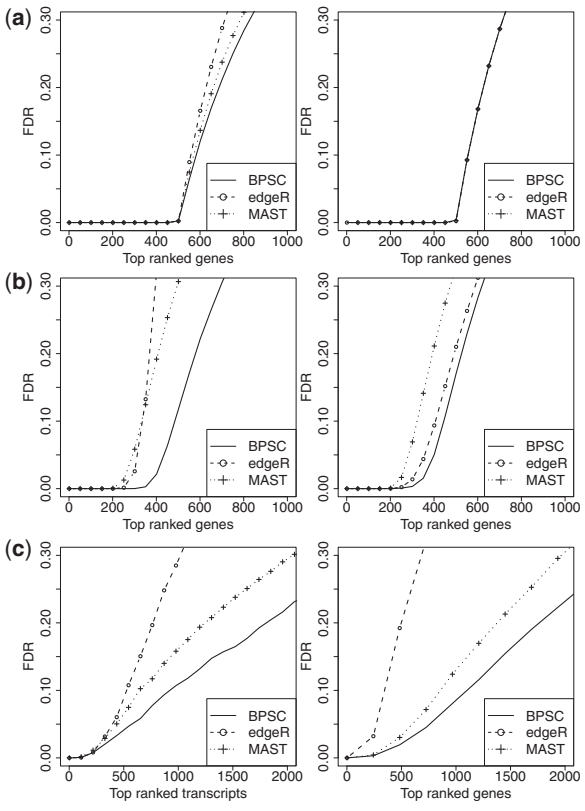

Fig. 3. False discovery rates evaluated in (**a**) bcSim, (**b**) scSim and (**c**) MDA-MB-231 datasets using BPSC, MAST and edgeR methods. In each plot, the horizontal axis indicates the number of top ranked genes/transcripts obtained by the methods. The vertical axis presents corresponding FDR values of gene/transcripts in the horizontal axis. In simulated datasets, the FDR value is the average from 100 simulations. In (a) and (b), the left plots present estimated FDR, while the right ones display the true FDR of the simulated datasets. (c) The estimated FDR of the transcript-level (left) and the gene-level (right) expression datasets

is quite similar to the scSim simulation, so the most likely explanation is that the beta-poisson model is more appropriate than the gamma-poisson model in this single-cell RNA-seq dataset. Figure 3c (right plot) shows that the analyses at transcript-level and at gene-level expression produce similar results.

Table 3 presents the top 10 DE transcripts discovered by BPSC. The FDR values of these transcripts from the edgeR method in the last column expresses that they are also well differentially expressed by the method. Next we investigate the affected biological pathways using the Reactome tool (Croft *et al.*, 2014) and the website of human gene database http://www.genecards.org. Many of the top DE genes (RPS17, RPL28, EIF6 and EIF5AL1) are involved in the mRNA translation processes, such as the eukaryotic translation initiation, eukaryotic translation elongation and eukaryotic translation termination. Since it is known that metformin can induce the down-regulation of translation of some mRNAs (Larsson *et al.*, 2012), the perturbation of metformin in this study may have a similar effect. We also discover two DE genes PRDX2 and PRDX3. These genes encode for peroxiredoxin (Prdx) proteins and are involved in the detoxification of reactive-oxygen-species pathway, which is previously shown to protect a breast cancer cell-line against doxorubicin-mediated toxicity (McDonald *et al.*, 2014). A full biological analysis of the results is out of the scope of this current methodological study and will be done separately.

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

Table 3. List of top 10 DE transcripts in the MDA-MB-231 dataset discovered by the BPSC method. The gene symbols and FDR values of these transcripts are displayed in column *Gene* and *BPSC*, respectively

Rank	Transcript	Gene	BPSC	edgeR
1	NM_001190470	MTRNR2L2	1.365439e-84	6.094011e-98
2	NM_000146	FTL	1.134770e-23	6.180201e-36
3	NM_001021	RPS17	2.842266e-20	1.560132e-13
4	NM_000991	RPL28	1.722874e-12	1.052418e-15
5	NM_005782	ALYREF	6.952639e-11	1.006880e-14
6	NM_001267810	EIF6	2.069525e-10	1.166470e-05
7	NM_001099692	EIF5AL1	6.553389e-10	4.268580e-09
8	NM_001165415	LDHA	3.359049e-09	2.806511e-06
9	NM_006793	PRDX3	1.182270e-08	9.044264e-08
10	NM_005809	PRDX2	1.684913e-08	1.320194e-07

The last column presents the corresponding FDR values of the transcripts discovered by edgeR.

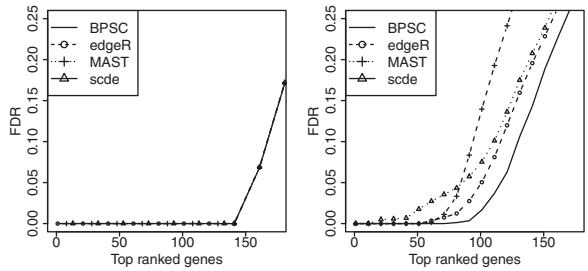

Fig. 4. True FDR evaluated in the bcSim (left) and scSim (right) simulation setting using the scde, edgeR and BPSC methods. The horizontal axis indicates the number of top-ranked genes obtained by the methods; the vertical axis presents the corresponding FDR values. The FDR value is the average from 10 simulations

We also compare BPSC to scde (Kharchenko *et al.*, 2014) version 1.99.0, a recent method for differential expression analysis of single-cell RNA-seq data using a Bayesian approach. Due to the heavy computation of the scde software (see more details in Fig. S6 of Supplementary report), we keep the same number of cells in each group, but reduce the number of genes down to 3000 genes in both simulated bulk-cell and single-cell RNAseq datasets. Similar to edgeR, the scde uses read counts as input. Default parameter settings of the software are applied. Since the scde does not report *P*-values but use *Z*-scores for the differential expression analysis, we do not compare the methods in terms of estimated FDR. Instead, as suggested by the software tutorials, we use the top genes ranked by the absolute of their *Z*-scores, then compute the true FDR for comparison.

We run the simulation 10 times (low number due to computational demands of scde), and the average results are shown in Figure 4. The left panel shows that all methods are successful in the bulk-cell simulation setting (bcSim) and achieve the same FDR. However, for the single-cell setting (scSim), the BPSC method has better performance as compared to the scde, as shown in right panel of figure. Table 4 briefly summarizes the performance of the methods on differential expression analysis for different benchmarks.

7 Conclusions

We have presented a model for gene expression of single-cell RNA-seq data based on the beta-Poisson mixture model. Experiments with real datasets show that our approach is able to correctly model a great majority of the transcripts. We also introduce an application of the model for differential expression analysis through a combination with the generalized linear model. The results of differential expression analysis in both simulated and real single-cell RNA-seq datasets demonstrate that the proposed method performs well against a traditional method commonly used in bulk-cell RNA-seq data and recent methods designed for differential expression analysis in single-cell RNA-seq dataset. The success of the model helps understand further the transcription mechanism of gene expression in single-cell level and opens opportunities for better analyses of single-cell RNA-seq data.

Table 4. Performance of BPSC, scde, MAST and edgeR on differential expression analysis using different benchmarks

Setting	edgeR	MAST	scde	BPSC
Bulk-cell (GP model, true FDR)	+	+	+	+
Bulk-cell (GP model, estimated FDR)	+	+	NA	+
Single-cell (BP model, true FDR)	+	−	−	++
Single-cell (BP model, estimated FDR)	−	+	NA	++
Single-cell (real data, estimated FDR)	−	+	NA	++
Computational time	++	++	−	+

The methods are assessed as very positive (++), positive (+), not positive (−) or result not available (NA).

Funding

This work was supported by grants from the Swedish Science Council and the Swedish Cancer Foundation.

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, **8**, 1765–1786.

Anders,S. *et al.* (2015) HTSeqa Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Croft,D. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.

Daigle,B.J. *et al.* (2015) Inferring single-cell gene expression mechanisms using stochastic simulation. *Bioinformatics*, btv007.

Finak,G. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.

Hildebrand,F.B. (1974). *Introduction to Numerical Analysis*. 2nd edition. McGraw-Hill, New York.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Larsson,O. *et al.* (2012) Distinct perturbation of the translatome by the antidiabetic drug metformin. *Proc. Natl. Acad. Sci. USA*, **109**, 8977–8982.

Law,C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.

McCarthy,D.J. *et al.* (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

McDonald,C. *et al.* (2014) Peroxiredoxin proteins protect MCF-7 breast cancer cells from doxorubicin-induced toxicity. *Int. J. Oncol.*, **45**, 219–226.

Patro,R. *et al.* (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**, 462–464.

Pawitan,Y. (2013). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford.

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024

Pawitan,Y. *et al*. (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sanchez,A. *et al*. (2013) Stochastic models of transcription: from single molecules to single cells. *Methods*, **62**, 13–25.

Shahrezaei,V. and Swain,P.S. (2008) Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. USA*, **105**, 17256–17261.

Shalek,A.K. *et al*. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**, 236–240.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–515.

Velten,L. *et al*. (2015) Single-cell polyadenylation site mapping reveals 3 isoform choice variability. *Mol. Syst. Biol*., **11**, 812.

Wills,Q.F. *et al*. (2013) Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol*., **31**, 748–752.

Wu,A.R. *et al*. (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, **11**, 41–46.

Downloaded from https://academic.oup.com/bioinformatics/article/32/14/2128/2288270 by guest on 25 April 2024