Zamani Dadaneh *et al. BMC Genomics* 2020, **21**(Suppl 9):585
https://doi.org/10.1186/s12864-020-06938-8

BMC Genomics

RESEARCH Open Access

Bayesian gamma-negative binomial modeling of single-cell RNA sequencing data

Siamak Zamani Dadaneh[1], Paul de Figueiredo[2,3,4], Sing-Hoi Sze[5], Mingyuan Zhou[6] and Xiaoning Qian[1,7*]

From The Sixth International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2019)
Niagara Falls, NY, USA. 07 September 2019

Abstract

Background: Single-cell RNA sequencing (scRNA-seq) is a powerful profiling technique at the single-cell resolution. Appropriate analysis of scRNA-seq data can characterize molecular heterogeneity and shed light into the underlying cellular process to better understand development and disease mechanisms. The unique analytic challenge is to appropriately model highly over-dispersed scRNA-seq count data with prevalent dropouts (zero counts), making zero-inflated dimensionality reduction techniques popular for scRNA-seq data analyses. Employing zero-inflated distributions, however, may place extra emphasis on zero counts, leading to potential bias when identifying the latent structure of the data.

Results: In this paper, we propose a fully generative hierarchical gamma-negative binomial (hGNB) model of scRNA-seq data, obviating the need for explicitly modeling zero inflation. At the same time, hGNB can naturally account for covariate effects at both the gene and cell levels to identify complex latent representations of scRNA-seq data, without the need for commonly adopted pre-processing steps such as normalization. Efficient Bayesian model inference is derived by exploiting conditional conjugacy via novel data augmentation techniques.

Conclusion: Experimental results on both simulated data and several real-world scRNA-seq datasets suggest that hGNB is a powerful tool for cell cluster discovery as well as cell lineage inference.

Keywords: Single-cell RNA sequencing, Bayesian, Hierarchical modeling

Background

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool for unbiased identification of previously uncharacterized molecular heterogeneity at the cellular level [1]. This is in contrast to standard bulk RNA-seq techniques [2], which measures average gene expression levels within a cell population, and thus ignore tissue heterogeneity. Consideration of cell-level variability of gene expressions is essential for extracting signals from complex heterogeneous tissues [3], and also for understanding dynamic biological processes, such as embryo development [4] and cancer [5].

A large body of statistical tools developed for scRNA-seq data analysis include a dimensionality reduction step. This leads to more tractable data, from both statistical and computational point of views. Moreover, the noise in the data can be decreased, while retaining the often intrinsically low-dimensional signal of interest. Dimensionality reduction of scRNA-seq data is challenging. In addition to high gene expression variability due to cell heterogeneity, the excessive amount of zeros in scRNA-seq hinders

*Correspondence: xqian@ece.tamu.edu
[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA
[7]TEES-AgriLife Center for Bioinformatics & Genomic Systems Engineering, Texas A&M University, College Station, Texas, USA
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

the application of classical dimensionality reduction techniques such as principal component analysis (PCA). For instance, in real-world datasets, it has been reported that the first or second principal components often depend more on the proportion of detected genes per cell (i.e., genes with at least one read) than on the actual biological signal [6].

Several existing computational tools adopt explicit zero-inflation modeling to infer the latent representation of scRNA-seq data. Zero-inflated factor analysis (ZIFA) [7] extends the framework of probabilistic PCA [8] to the zero-inflated setting, by modeling the excessive zeros using Bernoulli distributed random variables which indicate the dropout event. Zero-inflated negative binomial-based wanted variation extraction (ZINB-WaVE) [9] directly models the scRNA-seq counts using a zero-inflated negative binomial distribution, while accounting for both gene- and cell-level covariates. It infers the model parameters using a penalized maximum likelihood procedure.

Despite its popularity, using an explicit zero-inflation term may place unnecessary emphasis on the zero counts, leading to complication in discovering the latent representation of scRNA-seq data. In this paper, we propose a hierarchical gamma-negative binomial (hGNB) model to both perform dimensionality reduction and adjust for the effects of the gene- and cell-level confounding factors simultaneously. Exploiting the hierarchical structure, the proposed hGNB model is capable of capturing the high over-dispersion present in the scRNA-seq data. More precisely, we factorize the logit of the negative-binomial (NB) distribution probability parameter to identify latent representation of the data. In addition to factorization, linear regression terms are also included in that logit function to adjust for the impact of covariates.

In hGNB, a gamma distribution with varying rate parameter is used to model the cell dependent dispersion parameter of the NB distribution. The cell-level dispersion serves as a means of representing the prevalence of the dropout events. For instance, cells that are sequenced deeply will naturally include less dropped-out genes with zero counts, and thus this will be reflected in the cell specific dispersion parameter of NB distribution.

We follow a Bayesian framework, similar to bulk RNA-seq setting [10–14], and derive closed-form Gibbs sampling update equations for the model parameters of hGNB, by exploiting sophisticated data augmentation techniques. More specifically, we apply the data augmentation technique of [15] (2015) for the NB distribution, and the Polya-Gamma distributed auxiliary variable technique of [16] (2013) for the closed-form inference of regression coefficients and also latent factor parameters, removing the need for non-trivial Metropolis-Hastings correction steps [17]. Experimental results on several real-world scRNA-seq datasets demonstrate the superior performance of hGNB to identify cell clusters, especially in complex settings, and also its potential application in cell lineages inference.

Methods

hGNB model

In this section we present the hierarchical gamma-negative binomial (hGNB) model for factor analysis of scRNA-seq data. The graphical representation of hGNB is shown in Fig. 1. The parameters of the hGNB model with their interpretations in the context of scRNA-seq experiments are presented in Table 1. Let n_{vj} denote the number of sequencing reads mapped to gene $v \in \{1, ..., V\}$ in the cell $j \in \{1, ..., J\}$. Under the hGNB model, gene counts are distributed according to a negative binomial (NB) distribution:

$$n_{vj} \sim \text{NB}\left(r_j, p_{vj}\right), \tag{1}$$

where r_j and p_{vj} are dispersion and probability parameters of NB distribution, respectively. The probability mass function (PMF) of this distribution can be expressed as $f_N\left(n_{vj}\right) = \frac{\Gamma(n_{vj}+r_j)}{n_{vj}!\Gamma(r_j)} p_{vj}^{n_{vj}} \left(1-p_{vj}\right)^{r_j}$, where $\Gamma(\cdot)$ is the gamma function.

Data from scRNA-seq experiments exhibit high variability between different cells, even for genes with medium or high levels of expression. To capture this variability, we impose a gamma prior on the cell-level dispersion parameters as

$$r_j \sim \text{Gamma}(e_0, 1/h), \tag{2}$$

where for simplification, the hyper-parameter e_0 is set to 0.01 in our experiments, and the rate h is learned during the Gibbs sampling inference, presented in the following section. This hierarchical prior on the dispersion parameter, enhances the flexibility of NB distribution to capture the high over-dispersion of scRNA-seq counts, without the need for explicit zero-inflation modeling.

To account for various technical and biological effects common in scRNA-seq technologies, we impose a regression model on the logit of NB probability parameter as

$$\psi_{vj} = \text{logit}\left(p_{vj}\right) = \boldsymbol{\beta}_v^T \boldsymbol{x}_j + \boldsymbol{\delta}_j^T \boldsymbol{z}_v + \boldsymbol{\phi}_v^T \boldsymbol{\theta}_j. \tag{3}$$

With this particular formulation, the expected count for gene v in cell j can be written as

$$\mathbb{E}\left[n_{vj}\right] = r_j e^{\psi_{vj}}. \tag{4}$$

Thus the regression and factorization terms in (3) can directly adjust the value of expected gene counts in different samples.

The three terms in (3) are integrated to capture different expression variability sources. In the first term, $\boldsymbol{x}_j$ is a known vector of P covariates for cell j and $\boldsymbol{\beta}_v$ is

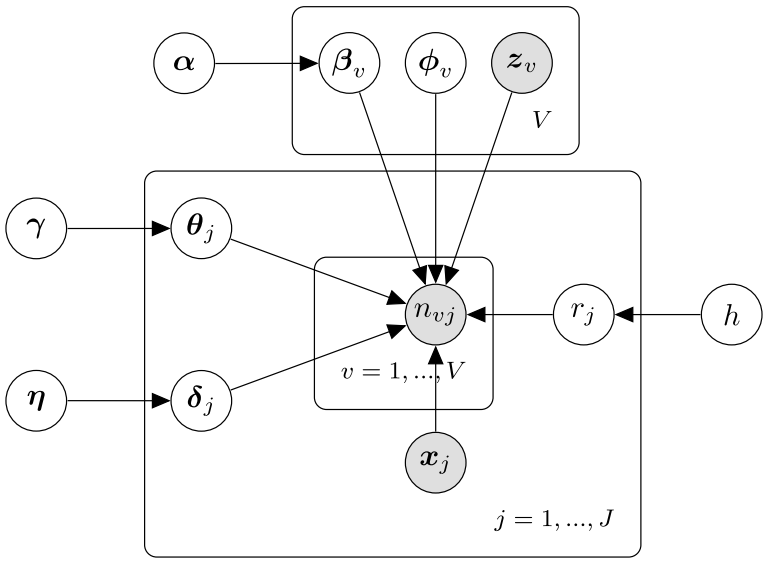

Fig. 1 Graphical representation of the hierarchical gamma-negative binomial (hGNB) model

the regression-coefficient vector adjusting the effect of covariates on gene v. The covariate vector $\boldsymbol{x}_j$ can represent variations of interest, such as cell types, or unwanted variations, such as batch effects or quality control measures. An intercept term can also be included in these cell-level covariates to account for gene dependent baseline expressions.

In the second term, $\boldsymbol{z}_v$ is a vector of Q covariates for gene v, representing gene length or GC-content for example [18], and $\boldsymbol{\delta}_j$ is its associated regression-coefficient vector. We also include a fixed intercept element in $\boldsymbol{z}_v$ to account for cell-specific expressions, such as the size factors representing differences in sequencing depth.

In the third term, $\boldsymbol{\phi}_v^T \boldsymbol{\theta}_j$ corresponds to the latent factor representation of the count n_{vj}, after accounting for the effects of gene- and cell-level covariates. More precisely, the unknown $K \times 1$ vector $\boldsymbol{\phi}_v$ contains the factor loading parameters which determine the association between genes and latent factors. Moreover, the unknown $K \times 1$ vector $\boldsymbol{\theta}_j$ encodes the popularity of the K factors in the expression of cell j.

Table 1 Parameters of the hierarchical gamma-negative binomial (hGNB) model and their interpretations in the context of scRNA-seq data

Parameter	Constraint	Interpretation
r_j	$r_j > 0$	Expression heterogeneity of genes in sample j
ϕ_{vk}	$\sum_{v=1}^{V} \phi_{vk} = 1, \phi_{vk} > 0$	Gene-latent factor association
θ_{jk}	$\theta_{kj} > 0$	Popularity of factor k in sample j
β_{vp}	$\beta_{vp} \in \mathbb{R}$	Impact of cell covariate p on expression of gene v
δ_{jq}	$\beta_{vp} \in \mathbb{R}$	Impact of gene covariate q on expression of cell j

The inputs of hGNB are gene counts n_{vj} and vector of cell- and gene-level covariates $\mathbf{x}_j$ and $\mathbf{z}_v$

We place independent zero-mean normal distributions on the components of the regression coefficient parameters $\boldsymbol{\beta}_v$ and $\boldsymbol{\delta}_j$ as

$$\begin{aligned} \boldsymbol{\beta}_v &\sim \prod_{p=1}^{P} \mathrm{N}\left(\beta_{vp}; 0, \alpha_p^{-1}\right), \\ \boldsymbol{\delta}_j &\sim \prod_{q=1}^{Q} \mathrm{N}\left(\delta_{jq}; 0, \eta_q^{-1}\right), \end{aligned} \tag{5}$$

where α_p and η_q are precision parameters of the normal distributions and gamma priors are imposed on them. These priors are known as automatic relevance determination (ARD), which are effective tools for pruning large numbers of irrelevant covariates [19, 20]. In addition, by assuming identical precision for components of the regression coefficients across all genes or samples, hGNB borrows statistical strengths to infer these precision parameters.

We impose independent normal priors on latent factor loading and score parameters $\boldsymbol{\phi}_v$ and $\boldsymbol{\theta}_j$:

$$\begin{aligned} \boldsymbol{\phi}_v &\sim \mathrm{N}\left(\boldsymbol{\phi}_v; 0, I_K\right), \\ \boldsymbol{\theta}_j &\sim \prod_{k=1}^{K} \mathrm{N}\left(\theta_{jk}; 0, \gamma_k^{-1}\right). \end{aligned} \tag{6}$$

Note that the posterior for these terms is not generally independent or normal, but accounts for the statistical dependence as reflected in the data.

We complete the model by imposing a gamma prior on the precision parameters of normal distributions, and also the rate parameter of gamma distributions. Specifically, throughout the experiments, we set both the shape and rate of these gamma priors to 0.01.

Inference via Gibbs sampling

In this section, we provide an efficient inference algorithm that adopts data augmentation techniques tailored to our hGNB model. Algorithm 1 summarizes all the steps in the Gibbs sampling algorithm.

Sample dispersion parameter We start with the data augmentation technique developed for inferring the NB dispersion parameter [15]. More precisely, the negative binomial random variable $n \sim \text{NB}(r,p)$ can be generated from a compound Poisson distribution as

$$n = \sum_{t=1}^{\ell} u_t, \quad u_t \sim \text{Log}(p), \quad \ell \sim \text{Pois}\left(-r\ln(1-p)\right),$$

where $u \sim \text{Log}(p)$ corresponds to the logarithmic random variable [21], with the PMF $f_U(u) = -\frac{p^u}{u\ln(1-p)}$, $u \in \{1,2,...\}$. As shown in [15], given n and r, the distribution of ℓ is a Chinese Restaurant Table (CRT) distribution, $(\ell|n,r) \sim \text{CRT}(n,r)$, which can be generated as $\ell = \sum_{t=1}^{n} b_t, b_t \sim \text{Bernoulli}\left(\frac{r}{r+t-1}\right)$.

Utilizing this augmentation technique, for each observed count n_{vj}, an auxiliary count is sampled as

$$\left(\ell_{vj}|-\right) \sim \text{CRT}\left(n_{vj}, r_j\right). \tag{7}$$

Using gamma-Poisson conjugacy, the cell-dependent dispersion parameters are updated as

$$\left(r_j|-\right) \sim \text{Gamma}\left(e_0 + \sum_v \ell_{vj}, \frac{1}{h - \sum_v \ln\left(1-p_{vj}\right)}\right). \tag{8}$$

Sample regression coefficients For the regression coefficients modeling potential covariate effects, the lack of conditional conjugacy precludes immediate closed-form inference. Therefore we adopt another data augmentation technique, specifically designed for hGNB, to infer the regression coefficients $\boldsymbol{\beta}_v$ and $\boldsymbol{\delta}_j$, relying on the Polya-Gamma (PG) data augmentation [16, 22].

Denote ω_{vj} as a random variable drawn from the PG distribution as $\omega_{vj} \sim \text{PG}(n_{vj}+r_j, 0)$. Since $\mathbb{E}_{\omega_{vj}}\left[\exp\left(-\omega_{vj}\psi_{vj}^2/2\right)\right] = \cosh^{(n_{vj}+r_j)}\left(\psi_{vj}^2/2\right)$, the likelihood of ψ_{vj} in (3) can be expressed as

$$\begin{aligned}\mathcal{L}\left(\psi_{vj}\right) &\propto \frac{\left(e^{\psi_{vj}}\right)^{n_{vj}}}{\left(1+e^{\psi_{vj}}\right)^{n_{vj}+r_j}} \\ &\propto \exp\left(\frac{n_{vj}-r_j}{2}\psi_{vj}\right)\mathbb{E}_{\omega_{vj}}\left[\exp\left(-\omega_{vj}\psi_{vj}^2/2\right)\right].\end{aligned} \tag{9}$$

Exploiting the exponential tilting of the PG distribution in [16], we draw ω_{vj} as

$$\left(\omega_{vj}|-\right) \sim \text{PG}\left(n_{vj}+r_j, \psi_{vj}\right). \tag{10}$$

Given the values of the auxiliary variables ω_{vj} for $j = 1,...,J$ and the prior in (5), the conditional posterior of $\boldsymbol{\beta}_v$ can be updated as

$$\left(\boldsymbol{\beta}_v|-\right) \sim \text{N}\left(\mu_v^{(\beta)}, \Sigma_v^{(\beta)}\right), \tag{11}$$

where $\Sigma_v^{(\beta)} = \left(\text{diag}\left(\alpha_1,...,\alpha_P\right) + \sum_j \omega_{vj}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)^{-1}$ and $\mu_v^{(\beta)} = \Sigma_v^{(\beta)}\left[\sum_j\left(\frac{n_{vj}-r_j}{2} - \omega_{vj}\left(\boldsymbol{\delta}_j^T\boldsymbol{z}_v + \boldsymbol{\phi}_v^T\boldsymbol{\theta}_j\right)\right)\boldsymbol{x}_j\right]$.

A similar procedure can be followed to derive the conditional updates for cell-level regression coefficients as

$$\left(\boldsymbol{\delta}_j|-\right) \sim \text{N}\left(\mu_j^{(\delta)}, \Sigma_j^{(\delta)}\right), \tag{12}$$

where $\Sigma_j^{(\delta)} = \left(\text{diag}\left(\eta_1,...,\eta_Q\right) + \sum_v \omega_{vj}\boldsymbol{z}_v\boldsymbol{z}_v^T\right)^{-1}$ and $\mu_j^{(\delta)} = \Sigma_j^{(\delta)}\left[\sum_v\left(\frac{n_{vj}-r_j}{2} - \omega_{vj}\left(\boldsymbol{\beta}_v^T\boldsymbol{x}_j + \boldsymbol{\phi}_v^T\boldsymbol{\theta}_j\right)\right)\boldsymbol{z}_v\right]$.

Sample latent factor parameters Using the likelihood function in (9) and the priors in (6), we can derive closed-form update steps for factor loading and score parameters. More specifically, the full conditional for factor loading $\boldsymbol{\phi}_v$ is a normal distribution:

$$\left(\boldsymbol{\phi}_v|-\right) \sim \text{N}\left(\mu_v^{(\phi)}, \Sigma_v^{(\phi)}\right), \tag{13}$$

where $\Sigma_v^{(\phi)} = \left(I_K + \sum_j \omega_{vj}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^T\right)^{-1}$ and $\mu_v^{(\phi)} = \Sigma_v^{(\phi)}\left[\sum_j\left(\frac{n_{vj}-r_j}{2} - \omega_{vj}\left(\boldsymbol{\beta}_v^T\boldsymbol{x}_j + \boldsymbol{\delta}_j^T\boldsymbol{z}_v\right)\right)\boldsymbol{\theta}_j\right]$.

The full conditional for factor score $\boldsymbol{\theta}_j$ is also a normal distribution:

$$\left(\boldsymbol{\theta}_j|-\right) \sim \text{N}\left(\mu_j^{(\theta)}, \Sigma_j^{(\theta)}\right), \tag{14}$$

where $\Sigma_j^{(\theta)} = \left(\text{diag}\left(\gamma_1,...,\gamma_K\right) + \sum_v \omega_{vj}\boldsymbol{\phi}_v\boldsymbol{\phi}_v^T\right)^{-1}$ and $\mu_j^{(\theta)} = \Sigma_j^{(\theta)}\left[\sum_v\left(\frac{n_{vj}-r_j}{2} - \omega_{vj}\left(\boldsymbol{\beta}_v^T\boldsymbol{x}_j + \boldsymbol{\delta}_j^T\boldsymbol{z}_v\right)\right)\boldsymbol{\phi}_v\right]$.

Sample precision and rate The precision parameters of normal distributions in (5) and (6) can be updated using the normal-gamma conjugacy:

$$\alpha_p \sim \text{Gamma}\left(e_0 + V/2, \frac{1}{f_0 + \sum_{v=1}^{V} \beta_{vp}/2}\right),$$
$$\eta_q \sim \text{Gamma}\left(e_0 + J/2, \frac{1}{f_0 + \sum_{v=1}^{V} \delta_{jq}/2}\right),$$
$$\gamma_k \sim \text{Gamma}\left(e_0 + J/2, \frac{1}{f_0 + \sum_{v=1}^{V} \theta_{jk}/2}\right). \quad (15)$$

Finally, the rate of gamma distribution in (2) can be updated using the gamma-gamma conjugacy with respect to the rate parameter:

Algorithm 1 hGNB model inference

Inputs: scRNA-seq counts, design matrix of covariate effects, N

Output: gene module membership matrix

Initialize model parameters
# Do Gibbs sampling:
for *iter* $= 1$ to N **do**
Sample ℓ_{vj} using the CRT distribution (Eq. (7))
Update r_j using the gamma-Poisson conjugacy (Eq. (8))
Sample auxiliary variables ω_{vj}, using the PG distribution (Eq. (10))
Update cell- and gene-level regression coefficients (Eq. (12),(11))
Update factor loadings and scores (Eq. (13),(14))
Update α_p, η_q and γ_k (Eq. (15))
end for

$$h \sim \text{Gamma}\left(e_0(1+J), \frac{1}{f_0 + \sum_{j=1}^{J} r_j}\right). \quad (16)$$

Results

We evaluate our hGNB model on four different sets of real-world scRNA-seq data from different platforms, and compare its performance to those of principal component analysis (PCA), ZIFA [7], and ZINB-WaVE [9]. In the following, We briefly describe these scRNA-seq datasets. To pre-process these datasets when needed, we followed the same procedures as in [9].

V1 dataset. This dataset characterizes more than 1600 cells from the primary visual cortex (V1) in adult male mice, using a set of established Cre lines [23]. A subset of three Cre lines, including Ntsr1-Cre, Rbp4-Cre, and Scnn1a-Tg3-Cre, that respectively label layer 4, layer 5, and layer 6 excitatory neurons were selected. We only retained 285 cells that passed the authors' quality control (QC) filters. The dimensionality reduction methods were only applied to the 1000 most variable genes.

S1/CA1 dataset. This dataset characterizes 3005 cells from the primary somatosensory cortex (S1) and the hippocampal CA1 region, using the Fluidigm C1 microfluidics cell capture platform followed by Illumina sequencing [24]. Gene expression is quantified by UMI counts.

mESC dataset. This dataset includes the transcriptome measurement of 704 mouse embryonic stem cells (mESCs), across three culture conditions (serum, 2i, and a2i), using the Fluidigm C1 microfluidics cell capture platform followed by Illumina sequencing [25]. We excluded the samples that did not pass the authors's QC filters, resulting in a total of 169 serum cells, 141 2i cells, and 159 a2i cells. The dimensionality reduction methods were only applied to the 1000 most variable genes.

OE dataset. This data characterizes 849 FACS-purified cells from the mouse OE, using the Fluidigm C1 microfluidics cell capture platform followed by Illumina sequencing [26]. We followed the filtering procedure of [27], and filtered the cells that exhibited poor sample quality, retaining a total of 747 cells.

For all datasets, hGNB was run using 2000 MCMC iterations, where after the first 1000 burn-in iterations, the posterior samples with the highest likelihood were collected as the point estimates of model parameters corresponding to latent factors. In the dimensionality reduction analysis below, following [9], for S1/CA1 dataset we set the number of latent factors $K = 3$, and for V1 and mESC we set $K = 2$. The average run time for hGNB with 2000 MCMC iterations on a cluster compute node with Intel Xeon 2.5GHz processor was approximately 4 hours.

Goodness-of-fit of hGNB model

We have examined the goodness-of-fit of hGNB model on V1, S1/CA1 and mESC datasets, using the mean-difference (MD) plots. Figure 2 shows the MD plot for the S1/CA1 dataset, where the y-axis is the difference between observed counts and the expected counts under hGNB, and x-axis is the average of these two sets of counts. The solid red line in this figure, which represents the local regression fit [28] to the data, resides near zero for various average levels. This supports the good fit of hGNB model to the highly over-dispersed scRNA-seq data. Similar trends are observed for V1 and mESC datasets (Supplementary materials).

Capturing zero-inflation

Next we evaluate the performance of hGNB on simulated data based on the zero-inflated NB distribution of [9] (ZINB-WaVE) to show that hGNB faithfully captures zero inflation without the need of explicit zero-inflation

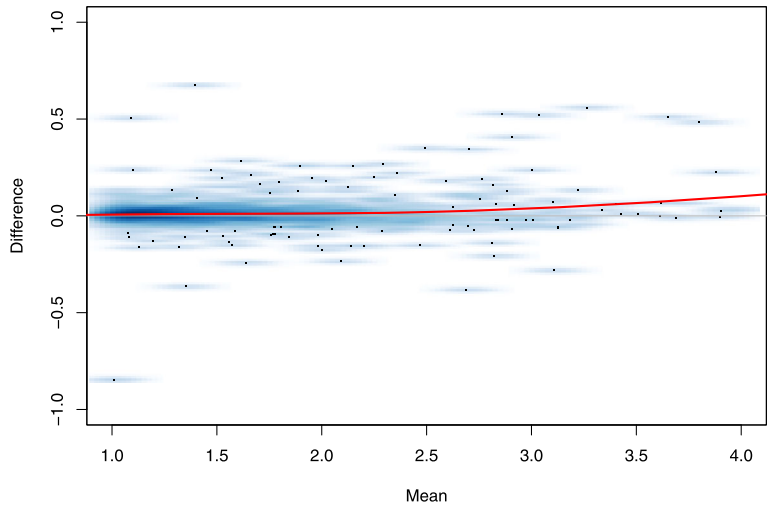

Fig. 2 Mean-difference (MD) plot for S1/CA1 dataset. The solid red line represents the local regression fit to the data

modeling. Specifically, the capability of hGNB to recover true clustering structure of cells under three zero-count prevalence levels with two different total numbers of cells.

ZINB-WaVE models the gene count n as

$$n \sim \pi\delta_0 + (1-\pi)\mathrm{NB}\left(n;\mu,\sigma^2\right),$$

where π is the zero-inflation probability and μ and σ^2 are mean and variance of the NB distribution. For each gene and cell, the zero-inflation probability π and the NB mean μ are linked to regression and factorization as in (3). In our simulations, the parameters were learned based on the S1/CA1 dataset. Genes that did not have at least five reads in at least five cells were filtered out and 1000 genes were then sampled at random for each dataset. The number of latent factors was set to $K = 2$. To simulate cell clustering, a K-variate Gaussian mixture distribution with three components was fitted to the inferred factor score parameters, and then for each simulated dataset, factor scores were generated from K-variate Gaussian distributions. By adjusting the value of regression coefficients in the zero-inflation term of ZINB-WaVE model, we generated synthetic datasets with three levels of zero-count percentages as 40%, 60% and 80% (for details refer to [9]). The number of cells were set to $J = 100$ and $J = 1000$. For each scenario, including cell numbers and zero-count prevalence (sparsity) levels, we simulated 10 datasets.

We evaluate the performance of our method for the clustering task based on the average silhouette width measure. The silhouette width s_j of sample j is defined as

$$s_j = \frac{b_j - a_j}{\max\left\{a_j, b_j\right\}},$$

where a_j is the average distance between sample j and all samples in the cluster that it belongs to, and b_j is the minimum average distance between sample j and samples in other clusters.

Tables 2 and 3 show the mean and standard deviation of clustering average silhouette width based on multiple runs of the above simulation setup, for different zero-count prevalence levels and cell numbers. In the setting with small sample size, for 40% and 60% zero fractions, hGNB has the best clustering silhouette width, and for the 80% zero fraction its performance is identical to that of ZINB-WaVE. In the setting with moderate sample size,

Table 2 Clustering performance based on synthetic data ($J = 100$)

Zero-Inflation	40%	60%	80%
hGNB	**0.3645** ±0.011	**0.1698** ±0.007	0.0905 ±0.009
PCA	0.2929 ±0.012	0.1265 ±0.012	0.0631 ±0.013
ZIFA	0.2642 ±0.019	0.1314 ±0.011	0.0728 ±0.010
ZINB	0.3501 ±0.022	0.1641 ±0.011	**0.0911** ±0.010
Monocle	0.2453 ±0.015	0.1155 ±0.017	0.0613 ±0.010
scVI	0.3122 ±0.029	0.1476 ±0.023	0.0593 ±0.005

Table 3 Clustering performance based on synthetic data ($J = 1000$)

Zero-Inflation	40%	60%	80%
hGNB	**0.3697** ±0.007	0.1470 ±0.010	0.0669 ±0.008
PCA	0.2594±0.009	0.0964 ±0.018	0.0349 ±0.018
ZIFA	0.3189 ±0.011	0.1191 ±0.004	0.0475 ±0.002
ZINB	0.3574 ±0.019	**0.1534** ±0.013	**0.0770** ±0.009
Monocle	0.2316 ±0.015	0.0995 ±0.011	0.0490 ±0.001
scVI	0.2590 ±0.046	0.1025 ±0.014	0.0351 ±0.006

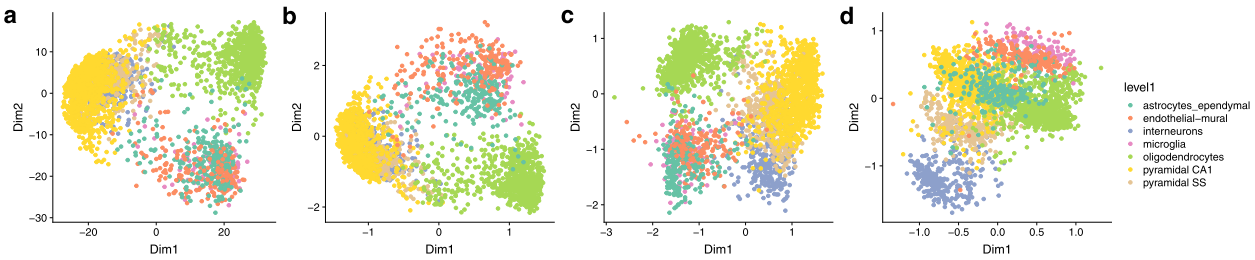

Fig. 3 Low-dimensional representations of the S1/CA1 dataset. Panels correspond to (**a**) PCA (on total-count normalized data), (**b**) ZIFA (on total-count normalized data), (**c**) ZINB-WaVE, and (**d**) hGNB

hGNB has the best clustering silhouette width for 40% zero fraction, and for 60% and 80% zero fractions it closely follows the performance of ZINB-WaVE. This suggests that the hierarchical structure of hGNB equips it with the capacity to capture highly over-dispersed count data, even though an explicit zero-inflation term is not included in its model. Also, ZINB-WaVE requires large enough samples to have robust inference results due to the introduction of zero-inflation terms in its model. ZIFA and PCA have a less competitive performance, as they normalize the data before learning its latent representation. Furthermore, in these tables, we have included the performance of Monocle [29] and scVI [30]. Despite exploiting independent component analysis (Monocle) or a deep generative framework (scVI), these two methods also fail to compete with hGNB in terms of cell clustering quality.

Dimensionality reduction

We applied hGNB to the three scRNA-seq datasets, V1, S1/CA1 and mESC, to assess its power to separate cell clusters in the low dimensional space, and compared it to PCA, ZIFA, and ZINB-WaVE methods. Figure 3 illustrates the projected scRNA-seq expression of profiled cells in the two-dimensional space for S1/CA1 dataset. The proposed hGNB model provides more biologically meaningful latent representations of scRNA-seq gene expressions for S1/CA1 cells, especially compared to PCA and ZIFA that do not model the counts directly. Furthermore, hGNB leads to more separated clusters of cells in the two-dimensional space, compared to ZINB-WaVE. Specifically, hGNB distinguishes `microglia` from `endothelial-mural` cells, while ZINB-WaVE fails to accomplish this task.

To examine the dimensionality reduction results more carefully, we used the average silhouette width as a measure of goodness for clustering.

Figure 4 shows the average silhouette width of different methods on V1, S1/CA1, and mESC datasets. For PCA and ZIFA, the results on both raw counts and normalized counts are included in this figure. For S1/CA1 dataset, which has the highest number of clusters, the proposed hGNB method outperforms all other methods in terms of clustering average silhouette. For mESC dataset, performance of hGNB is comparable to ZINB-WaVE, and it is significantly better than PCA and ZIFA. For V1 dataset, however, we observe that hGNB, besides PCA applied to raw counts, possess the lowest average silhouette. By further examination of the latent representations of cells for this dataset (Supplementary materials), we observe that all methods split the `Rbp4-Cre_KL100` cells into two

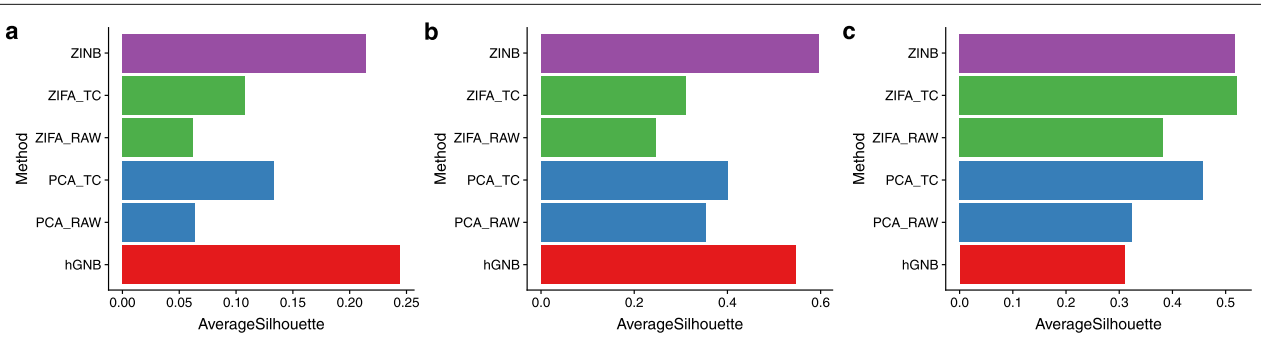

Fig. 4 Average silhouette width in scRNA-seq datasets (**a**) S1/CA1, (**b**) mESC, and (**c**) V1. Silhouette widths were computed in the low-dimensional space, using the groupings provided by the authors of the original publications. PCA and ZIFA were applied with both unnormalized (RAW) data and after total count (TC) normalization

clusters, one of them located near `Scnn1a-Tg3-Cre` cells, suggesting the presence of batch effects, which have led to confounding of latent representations [9].

Identification of developmental lineages

In addition to characterization of cell types, we further demonstrate the capability of hGNB to derive novel biological insights, by analyzing a set of cells from the mouse olfactory epithelium (OE). The samples were collected to identify the developmental trajectories that generate olfactory neurons (mOSN), sustentacular cells (mSUS), and microvillous cells (MV) [26].

We first performed dimensionality reduction on the OE dataset by applying hGNB with $K = 50$. Next, we clustered the cells using the low-dimensional factor score parameters θ_{kj}. More specifically, the resampling-based sequential ensemble clustering (RSEC) framework implemented in the `RSEC` function from the Bioconductor R package `clusterExperiment` [31] was applied to factor scores, leading to identification of 14 cell clusters. The correspondence between the detected clusters and the underlying biological cell types is presented in Table 4. In addition to these already known cell clusters in OE, hGNB is able to detect new clusters, potentially offering novel biological insights.

We further investigated the potential benefit of using the learned latent representation by our proposed hGNB model to infer branching cell lineages and order cells by developmental progression along each lineage. To infer the global lineage structure (i.e., the number of lineages and where they branch), a minimum spanning tree (MST) was constructed on the clusters identified above by RSEC. We used the R package `slingshot` [32]. Figure 5 illustrates the inferred lineages for the OE dataset, in a two-dimensional space obtained by applying multi-dimensional scaling (MDS) algorithm to the factor scores learned by hGNB. There are three branches in the inferred lineages, with endpoints located in microvillous (MV), mature olfactory sensory neurons (mOSN), and mature sustentacular (mSUS) cells.

Table 4 Correspondence between identified clusters and cell types in OE dataset

Cell Type	Clusters
GBC	cl4,cl9
mSUS	cl2,cl3,cl5,cl11
mOSN	cl8,cl12,cl13
Immature Neurons	cl10
MV	cl14

Conclusions

We propose a hierarchical Bayesian gamma-negative binomial (hGNB) model for extracting low dimensional representations from single-cell RNA sequencing (scRNA-seq) data. hGNB obviates the need for explicit modeling of the zero-inflation prevalent in scRNA-seq count data. Our hGNB can naturally account for covariate effects at both the gene and cell levels, and does not require the commonly adopted pre-processing steps such as normalization. By taking advantage of sophisticated data augmentation techniques, hGNB possesses efficient closed-form Gibbs sampling update equations. Our experimental results on real-world scRNA-seq data demonstrates that hGNB is capable of identifying insightful cell clusters, especially in complex settings.

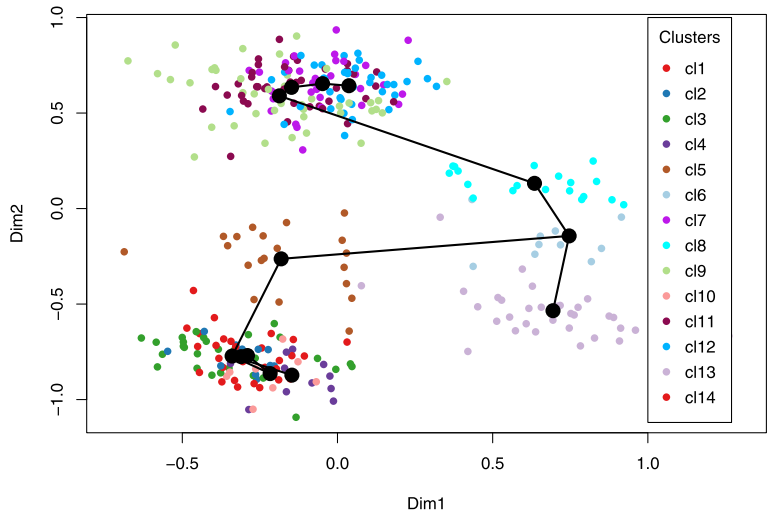

Fig. 5 Lineage inference on the OE dataset. The low dimensional data representation derived by hGNB were used to cluster cells by `RSEC`. The minimum spanning tree (MST) of the derived clusters constructed by `slingshot` is also displayed

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06938-8.

Additional file: Additional figures.

Abbreviations

hGNB: Hierarchical gamma-negative binomial; NB: Negative binomial; CRT: Chinese restaurant table; ZIFA: Zero-inflated factor analysis; ZINB-WaVE: Zero-inflated negative binomial-based wanted variation extraction; PCA: Principal component analysis; ARD: Automatic relevance determination; PG: Polya-Gamma; RSEC: Resampling-based sequential ensemble clustering; OE: Olfactory epithelium; mSUS: Sustentacular cells; MV: Microvillous cells; mOSN: Olfactory neurons

Acknowledgements

The authors would like to thank Texas A&M High Performance Research Computing and Texas Advanced Computing Center for providing computational resources to perform experiments in this paper. The authors also would like to thank Ehsan Hajiramezanali and Seyed Nami Niyakan for their help in running scVI and Monocle on the synthetic data.

About this supplement

This article has been published as part of *BMC Genomics Volume 21 Supplement 9, 2020: Selected original articles from the Sixth International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2019): genomics*. The full contents of the supplement are available online https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-9.

Authors' contributions

Conceived the method: SZD, MZ, XQ. Developed the algorithm: SZD, MZ, XQ. Performed the simulations: SZD. Analyzed the results and wrote the paper: SZD, PdF, SHS, MZ, XQ. All authors read and approved the final manuscript.

Funding

This work was supported by the National Science Foundation (NSF) Grants 1553281 and 1812641, as well as the United States Department of Agriculture National Institute of Food and Agriculture competitive grant USDA-NIFASCRI-2017-51181-26834 through the National Center of Excellence for Melon at the Vegetable and Fruit Improvement Center of Texas A&M University to XQ; the NSF Grant 1812699 to MZ; the NSF Grant 1532188 and the funding support from QNRF (NPRP9-001-2-001, NPRP7-1634-2-604), CSTR (2017-01), and CONACYT to PdF. Publication costs are funded by NSF Grant 1553281. The funding agencies had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA. [2]Department of Microbial Pathogenesis and Immunology, Texas A&M Health Science Center, Bryan, Texas, USA. [3]Department of Veterinary Pathobiology, Texas A&M University, College Station, Texas, USA. [4]Norman Borlaug Center, Texas A&M University, College Station, Texas, USA. [5]Department of Computer Science and Engineering, Texas A&M University, College Station, Texas, USA. [6]McCombs School of Business, The University of Texas at Austin, Austin, Texas, USA. [7]TEES-AgriLife Center for Bioinformatics & Genomic Systems Engineering, Texas A&M University, College Station, Texas, USA.

Published: 9 September 2020

References

1. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14(9):618–30.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9.
3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
4. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343(6167):193–6.
5. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–401.
6. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):1–13.
7. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1):243.
8. Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Stat Soc Ser B Stat Methodol. 1999;61(3):611–22.
9. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9(1):1–17.
10. Dadaneh SZ, Qian X, Zhou M. BNP-Seq: Bayesian nonparametric differential expression analysis of sequencing count data. J Am Stat Assoc. 2018;113(521):81–94. https://doi.org/10.1080/01621459.2017.1328358.
11. Dadaneh SZ, Zhou M, Qian X. Bayesian negative binomial regression for differential expression with confounding factors. Bioinformatics. 2018;34(19):3349–56.
12. Zamani Dadaneh S, Zhou M, Qian X. Covariate-dependent negative binomial factor analysis of RNA sequencing data. Bioinformatics. 2018;34(13):61–9.
13. Boluki S, Dadaneh SZ, Qian X, Dougherty ER. Optimal clustering with missing values. BMC Bioinformatics. 2019;20(12):321.
14. Boluki S, Esfahani MS, Qian X, Dougherty ER. Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. BMC Bioinformatics. 2017;18(14):552.
15. Zhou M, Carin L. Negative binomial process count and mixture modeling. IEEE Trans Pattern Anal Mach Intell. 2013;37(2):307–20.
16. Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya–Gamma latent variables. J Am Stat Assoc. 2013;108(504):1339–49.
17. Chib S, Greenberg E. Understanding the metropolis-hastings algorithm. Am Stat. 1995;49(4):327–35.
18. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011;12(1):480.
19. Wipf DP, Nagarajan SS. A new view of automatic relevance determination. In: Advances in Neural Information Processing Systems; 2008. p. 1625–32.
20. Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1(Jun):211–44.
21. Johnson NL, Kemp AW, Kotz S. Univariate Discrete Distributions, vol. 444. Hoboken: Wiley; 2005.
22. Zhou M, Li L, Dunson D, Carin L. Lognormal and gamma mixed negative binomial regression. In: Proceedings of the International Conference on Machine Learning; 2012. p. 1343–50.
23. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19(2):335–46.
24. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138–42.

25. Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, Marioni JC. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 2015;17(4):471–85.
26. Fletcher RB, Das D, Gadye L, Street KN, Baudhuin A, Wagner A, Cole MB, Flores Q, Choi YG, Yosef N, Purdom E. Deconstructing olfactory stem cell trajectories at single-cell resolution. Cell Stem Cell. 2017;20(6):817–30.
27. Perraudeau F, Risso D, Street K, Purdom E, Dudoit S. Bioconductor workflow for single-cell RNA sequencing: Normalization, dimensionality reduction, clustering, and lineage inference. F1000Research. 2017;6(1158):1158.
28. Shyu WM, Grosse E, Cleveland WS. Local regression models. In: Statistical Models in S. Routledge; 2017. p. 309–76.
29. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979.
30. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12): 1053–8.
31. Purdom E, Risso D. clusterexperiment: Compare clusterings for single-cell sequencing. R package version. 2017;1(0):.
32. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions