BNP-Seq: Bayesian Nonparametric Differential Expression Analysis of Sequencing Count Data

Siamak Zamani Dadaneh[*], Xiaoning Qian[*#], and Mingyuan Zhou[†#]

August 16, 2016

Abstract

We perform differential expression analysis of high-throughput sequencing count data under a Bayesian nonparametric framework, removing sophisticated *ad-hoc* pre-processing steps commonly required in existing algorithms. We propose to use the gamma (beta) negative binomial process, which takes into account different sequencing depths using sample-specific negative binomial probability (dispersion) parameters, to detect differentially expressed genes by comparing the posterior distributions of gene-specific negative binomial dispersion (probability) parameters. These model parameters are inferred by borrowing statistical strength across both the genes and samples. Extensive experiments on both simulated and real RNA sequencing count data show that the proposed differential expression analysis algorithms clearly outperform previously proposed ones in terms of the areas under both the receiver operating characteristic and precision-recall curves.

Keywords: Markov chain Monte Carlo, negative binomial processes, over-dispersion, RNA-Seq, symmetric Kullback-Leibler divergence.

[*] Department of Electrical & Computer Engineering, Center for Bioinformatics & Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA.

[†] Department of Information, Risk, & Operations Management and Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX 78712, USA.

[#] Correspondence should be addressed to Xiaoning Qian (`xqian@ece.tamu.edu`) or Mingyuan Zhou (`mingyuan.zhou@mccombs.utexas.edu`).

1

1   Introduction

There has been significant recent interest in analyzing RNA sequencing (RNA-Seq) count data for studying life systems [Wang et al., 2009, Metzker, 2010]. It is challenging to model RNA-Seq data, not only because it is typically a large-p-small-n problem [West, 2003] where the data dimension is high while the sample size is small, but also because the sequencing counts are nonnegative, skewed, having large dynamical ranges, and highly over-dispersed [Anders and Huber, 2010, Datta and Nettleton, 2014]. A key task in RNA-Seq analysis is to identify the genes that are differentially expressed between different groups of samples (*e.g.*, samples measured under different medical conditions) [Wang et al., 2010, Robinson and Oshlack, 2010, Anders and Huber, 2010, Oshlack et al., 2010, Li et al., 2011, Li and Tibshirani, 2013]. The expression level of each RNA locus, here the gene, is determined by the number of sequenced reads to the transcript [Mortazavi et al., 2008]. Unlike a gene probe based method such as microarrays [Schena et al., 1995], the abundance of genes in RNA-Seq is restricted by the sequencing depth and there often exist dependencies between the expressions of different transcripts [Roberts et al., 2011].

Modeling the sequencing counts using an over-dispersed count distribution, such as the negative binomial (NB) distribution [Greenwood and Yule, 1920, Bliss and Fisher, 1953], is one of the most popular approaches for differential expression analysis [Robinson et al., 2010, Anders and Huber, 2010]. In the null hypothesis that a gene is not differently expressed, it is common to assume that the expectations of the counts of that gene are the same across different groups, after making adjustments to account for both technical and biological variations. In particular, almost all existing comparative analysis algorithms, before downstream analyses, require normalizing the sequencing counts to compensate the variations of sequencing depths across samples [Soneson and Delorenzi, 2013, Dillies et al., 2013, Zyprych-Walczak et al., 2015]. For instance, edgeR and DESeq, two widely used differential expression analysis R software packages adopt different *ad-hoc* normalization procedures: edgeR either calculates a trimmed mean of M-values [Robinson and Oshlack,

2010] between each pair of samples or uses an upper quantile of samples [Bullard et al., 2010] for normalization [Robinson et al., 2010], while DESeq takes the median of the ratios of observed sample's counts to the geometric mean across samples as a scaling factor for that specific sample [Anders and Huber, 2010, Love et al., 2014].

Normalizing the sequencing counts, however, inevitably destroys the discrete nature of the raw data and makes the performance clearly depend on whether the introduced normalization is suitable for the structure of the RNA-Seq data under study [Soneson and Delorenzi, 2013, Dillies et al., 2013, Zyprych-Walczak et al., 2015]. If the normalization procedure extracts normalization constants from the data under study to parameterize the distributions of the gene counts, the discrete nature of the raw data is preserved, but the model can no longer be considered as a generative model. In addition, almost all existing normalization procedures assume that most of the genes are not differentially expressed, and the differentially expressed genes are equally likely to be up- and down-regulated [Lovén et al., 2012, Lorenz et al., 2014, Risso et al., 2014a,b]. The violation of the assumption may potentially be addressed by using external RNA control consortium (ERCC) spike-in sequences for controls; however, it is shown in Risso et al. [2014a,b] that the read counts for ERCC spike-ins alone are usually not stable enough to be used for normalization. Moreover, despite that a wide array of methods have been proposed to adjust the counts to account for technical and biological variations, there is not a single one that clearly outperforms the others under various scenarios [Soneson and Delorenzi, 2013, Rapaport et al., 2013, Dillies et al., 2013, Risso et al., 2014a, Zyprych-Walczak et al., 2015, Zhang et al., 2014].

In this paper, rather than heuristically modifying the data and estimating the model parameters based on an empirical Bayes procedure, we introduce a generative model to analyze differential expression directly on the raw sequencing counts, without the need to preprocess the data by normalization. Instead of using parametric count distributions to describe the counts, we use a stochastic process to model the observed sample-gene

random count matrix in each group, whose model parameters are estimated by sharing statistical strength across both the genes and samples. The stochastic process can be used to explain not only the counts and the total number of expressed genes in the observed random count matrix, but also the number of newly expressed genes and the counts on both existing and newly expressed genes to be brought by a new sample. Such flexible random-process-based models lift the need of *ad-hoc* data normalization and strict parametric assumptions, allowing heterogeneity across samples and gene expression variations across different conditions to be well captured.

More specifically, moving beyond existing algorithms that model over-dispersed counts with the NB distribution, our Bayesian non-parametric (BNP) algorithms model the gene counts using the gamma-negative binomial process (GNBP) [Zhou et al., 2015], which mixes the NB shape parameter for each gene with the distribution of the weight of an atom of a gamma process [Ferguson, 1973], or beta-negative binomial process (BNBP) [Zhou et al., 2012, Zhou and Carin, 2015, Broderick et al., 2015], which mixes the NB probability parameter of each gene with the distribution of the weight of an atom of a beta process [Hjort, 1990]. In addition to the GNBP and BNBP, for comparison, we have extended the negative binomial process (NBP) of Zhou et al. [2015] by multiplying the gene-specific Poisson rates with gamma distributed sample-specific scaling parameters, and refer to it as the scaled NBP. While the NBP of Zhou et al. [2015] is not expected to work well since it does not explicitly model the variation of a sample's total count, the scaled NBP, even with a scaling parameter for each sample to capture that variation, is found to provide poor performance, indicating a clear limitation of the Poisson distribution assumption. We will show that while the variations of the gene counts across samples are well captured by neither the Poisson rates of the scaled NBP nor the normalized Poisson rates of the NBP, they are well modeled by both the GNBP and BNBP, using the NB shape and probability parameters, respectively.

Unlike previous algorithms for differential expression analysis, the proposed Bayesian

nonparametric (BNP) algorithms require no normalization pre-processing steps and they infer the posterior distributions, instead of point estimates, of their model parameters, using Gibbs sampling with closed-form update equations, achieving state-of-the-art performance in detecting truly differentially expressed genes for both synthetic and real data. To our knowledge, we are the first in constructing BNP algorithms with these distinct advantages to analyze the sequencing counts to detect differential expression in genomics.

The remainder of the paper is organized as follows. After reviewing existing differential expression analysis algorithms in Section 2, we introduce the NBP-Seq, GNBP-Seq, and BNBP-Seq for differential expression analysis in Section 3. We present experimental results on both synthetic and real benchmark RNA-Seq data in Section 4 and show that both the proposed GNBP-Seq and BNBP-Seq clearly outperform previously proposed differential express analysis algorithms. We conclude the paper in Section 5.

2   Differential expression analysis

For J RNA-Seq samples organized into the same group, let us denote n_{jk} as the number of reads in sequencing sample $j \in \{1, \ldots, J\}$ that are assigned to gene $k \in \{1, \ldots, K\}$, where K is the number of genes in the genome. Since the counts of a gene across samples are often over-dispersed, it is natural to model them using a NB distribution, where its variance σ^2 is related to its mean μ as $\sigma^2 = \mu + \phi\mu^2$, where ϕ is the dispersion parameter. As it is also common to refer to $r = \phi^{-1}$ as the dispersion parameter, to avoid ambiguity, we will refer to $r = \phi^{-1}$ as the NB shape parameter.

Methods such as edgeR and DESeq propose different ways to estimate ϕ. EdgeR models the gene count n_{jk} as a NB distribution with mean $n_j\lambda_{jk}$ and dispersion ϕ_k, where n_j is the observed total count (or the sum of adjusted counts) for sample j, λ_{jk} represents the abundance of gene k in sample j, and ϕ_k is considered as the coefficient of biological variation that is estimated by conditional maximum likelihood [Smyth and

Verbyla, 1996]. Furthermore, an empirical Bayes procedure is applied to shrink the dispersion parameters ϕ_k towards a common value [Robinson and Smyth, 2007].

DESeq also models the gene counts with the NB distribution. It considers two terms to estimate the variance σ_{jk}^2 for gene k in sample j, where the first term (shot noise) is associated with the mean expression of the gene, and the second one (raw variance) takes into account the biological variations between replicates. More specifically, it lets $\sigma_{jk}^2 = \mu_{jk} + n_j^2 v_{k,\rho(j)}$. Here, $\rho(j)$ is the group to which sample j belongs, and $v_{k,\rho(j)}$ is the per-gene raw variance, which is a smooth function of λ and ρ, an assumption that allows pooling data from different genes to estimate their variances.

Another widely used tool, baySeq [Hardcastle and Kelly, 2010], takes an empirical Bayesian approach to estimate the posterior probabilities of a set of models that define different patterns of differential expression for each gene. For instance, in the simplest case of a pairwise comparison between conditions A and B, with two biological replicates for each condition, the model for no differential expression is defined by the set of samples {A1, A2, B1, B2}, while differential expression between conditions A and B is defined by the sets {A1, A2} and {B1, B2}. The method then assumes that the counts follow the NB distribution and derives an empirically determined prior distribution from the data.

The final component of these methods is the test for gene differential expression. Both edgeR and DESeq use variations of Fisher's exact test, adjusted for the NB distribution, to compute exact p-values for the null hypothesis that the mean expressions of the genes are equal in both conditions under comparison. EdgeR also considers the generalized linear model approach to identify differentially expressed genes in its later versions, nevertheless, it has been shown to have similar performance to the method based on Fisher's exact test [Schurch et al., 2016]. Different from edgeR and DESeq, baySeq ranks the genes based on the inferred posterior probabilities of differential expression.

3 Bayesian nonparametric differential expression analysis for RNA-Seq

We consider a family of NB processes, each of which can be used to describe the row-by-row sequential construction of a sample-gene sequencing count matrix, where the addition of a new sample (row) brings counts at not only previously expressed genes (columns), but also previously unexpressed ones. We also describe the equivalent construction that draws a Poisson random number of independent, and identically distributed (i.i.d.) columns simultaneously, where each column corresponds to the counts of a gene that is expressed at least once across all the observed samples of a group. Showing these two equivalent constructions helps clearly understand the underlying statistical assumption made on the RNA-Seq data by a BNP prior, and how the statistical strength is shared across both the genes and samples to estimate both the sample-specific model parameters, which account for the variations in sequencing depths, and the gene-specific model parameters, whose posterior distributions are used to detect differentially expressed genes.

We explore a family of NB processes, specifically NBP, GNBP, and BNBP in this paper. To model the gene counts in each group with the GNBP, we consider the null hypothesis that with sample-specific NB probability parameters, the posterior distributions of the gene-specific NB shape parameters, regularized by the gamma process in the prior, are the same across different groups; whereas with the BNBP, we consider the null hypothesis that with sample-specific NB shape parameters, the posterior distributions of the gene-specific NB probability parameters, regularized by the beta process in the prior, are the same across different groups. For comparison, we also include NBP and its scaled version, and consider the null hypothesis that the posterior distributions of the gene-specific Poisson rate parameters of the scaled NBP, regularized by the gamma process in the prior, or the corresponding normalized Poisson rates of the NBP, are the same across different groups. After fitting the sequencing counts with a BNP prior, we assess the

7

significance of gene expression changes across groups by measuring the distances between the gene-specific posterior distributions of the NB shape parameters for the GNBP, NB probability parameters for the BNBP, Poisson rates for the scaled NBP, and normalized Poisson rates for the NBP, using symmetric Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951].

Below we show how a stochastic process can be used to model the counts in each group, where the group index is omitted for brevity. We represent the counts of all expressed genes in a group as a random count matrix $\mathbf{N}_J \in \mathbb{Z}^{J \times K_J}$, where $\mathbb{Z} = \{0, 1, \ldots\}$ represents the set of nonnegative integers, K_J denotes the random number of genes that are expressed at least once in the J samples of the group, and the element n_{jk} represents the number of reads in sequencing sample $j \in \{1, \ldots, J\}$ that are assigned to gene $k \in \{1, \ldots, K_J\}$. Note that K_J can potentially increase without bound as J increases.

3.1 NBP-Seq: Negative binomial process for RNA-Seq

Let us denote G_0 as a finite and continuous base measure over a complete and separable metric space Ω, $c \in \mathbb{R}_+$ as a scale parameter, and $q_j \in \mathbb{R}_+$ as sample-specific scaling parameters, where $\mathbb{R}_+ := \{x : x > 0\}$. We define the scaled negative binomial process (NBP) that has sample-specific scaling parameters as

$$(X_1, \ldots, X_J) \,|\, c, G_0, \{q_j\}_{1,J} \sim \text{NBP}(G_0, c_0, q_1, \ldots, q_J),$$

which is obtained by marginalizing out a gamma process [Ferguson, 1973] $G \sim \Gamma\text{P}(G_0, 1/c)$ from J conditionally independent Poisson processes [Kingman, 1993] $X_j \,|\, q_j, G \sim \text{PP}(q_j G)$, where for disjoint Borel sets $A_j \subset \Omega$, the gamma process G is defined such that $G(A_i) \sim \text{Gamma}[G_0(A_i), 1/c]$ are independent gamma random variables, and the Poisson process X_j is defined such that $X_j(A_i) \sim \text{Pois}[q_j G(A_i)]$ are independent Poisson random variables. With a draw from the gamma process G expressed as $G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}$, where ω_k

and r_k are the atoms and their associated weights, respectively, a draw from X_j can be expressed as

$$X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k}, \ n_{jk} \sim \text{Pois}(q_j r_k). \tag{1}$$

Note that if we fix $q_j = 1$ for all j, then the proposed NBP with sample-specific scaling parameters reduces to the NBP in Zhou and Carin [2015] and Zhou et al. [2015].

The conditional likelihood of the observed J samples of a group can be written as

$$p(\{X_j\}_{j=1}^{J} \mid G) = e^{-q.G(\Omega\backslash\mathcal{D}_J)} \left[\prod_{k=1}^{K_J} \frac{r_k^{n.k}e^{-q.r_k}}{\prod_{j=1}^{J} n_{jk}!} \right] \left[\prod_{j=1}^{J} q_j^{n_j} \right], \tag{2}$$

where $\mathcal{D}_J = \{\omega_k\}_{k:n.k>0}$ is the set of points of discontinuity, $K_J = |\mathcal{D}_J| = \sum_k \delta(n.k > 0)$ is the number of genes that are expressed at least once, $q. = \sum_{j=1}^{J} q_j$, and $n.k = \sum_{j=1}^{J} n_{jk}$. We map the counts associated with the elements of $\mathcal{D}_J$ to the random count matrix $\mathbf{N}_J$. While the labelings of the atoms in $\mathcal{D}_J$ are arbitrary, they are mapped in one of the $K_J!$ possible ways to the columns of $\mathbf{N}_J$. Similar to the derivation in Zhou et al. [2015], using a marginalization procedure shown in Caron et al. [2014], one may marginalize out the gamma process G, leading to the distribution of the random count matrix as

$$\begin{aligned} f(\mathbf{N}_J \mid \gamma_0, c, q_1, \dots, q_J) &= \frac{p(\{X_j\}_{1,J} \mid \gamma_0, c, q_1, \dots, q_J)}{K_J!} \\ &= \frac{\gamma_0^{K_J} \exp\left[-\gamma_0 \ln(\frac{q.+c}{c})\right]}{K_J!} \left[\prod_{k=1}^{K_J} \frac{\frac{\Gamma(n.k)}{(q.+c)^{n.k}}}{\prod_{j=1}^{J} n_{jk}!} \right] \left[\prod_{j=1}^{J} q_j^{n_j} \right]. \end{aligned} \tag{3}$$

One may verify by direct calculation that a scaled NBP random count matrix with the probability mass function (PMF) shown in (3) can be generated column by column

as i.i.d. count vectors:

$$\boldsymbol{n}_{:k} \sim \text{Multinomial}(n_{\cdot k}, q_1/q_\cdot, \ldots, q_J/q_\cdot),$$

$$n_{\cdot k} \sim \text{Logarithmic}[q_\cdot/(c + q_\cdot)],$$

$$K_J \sim \text{Pois}\left\{\gamma_0 \left[\ln(c + q_\cdot) - \ln(c)\right]\right\}. \tag{4}$$

It is clear from (4) that the columns of $\mathbf{N}_J$ are i.i.d. multivariate count vectors, which all follow the same logarithmic-multinomial (mixture) distribution. Thus the scaled NBP random count matrix $\mathbf{N}_J$ is column exchangeable. It is also row exchangeable if and only if the q_j are the same for all $j \in \{1, \ldots, J\}$.

Now consider the row-wise sequential construction of the scaled NBP random matrix. With the prior on $\mathbf{N}_J \in \mathbb{Z}^{J \times K_J}$ well defined, direct calculations using (4) yield the following form for this prediction rule, expressed in terms of familiar PMFs:

$$
\begin{aligned}
\frac{f(\mathbf{N}_{J+1}|\boldsymbol{\theta})}{f(\mathbf{N}_J|\boldsymbol{\theta})} &= \frac{K_J! K_{J+1}^+!}{K_{J+1}!} \prod_{k=1}^{K_J} \text{NB}\left(n_{(J+1)k}; n_{\cdot k}, \frac{q_{J+1}}{c + q_\cdot + q_{J+1}}\right) \\
&\quad \times \prod_{k=K_J+1}^{K_{J+1}} \text{Logarithmic}\left(n_{(J+1)k}; \frac{q_{J+1}}{c + q_\cdot + q_{J+1}}\right) \\
&\quad \times \text{Pois}\left\{K_{J+1}^+; \gamma_0 \left[\ln(c + q_\cdot + q_{J+1}) - \ln(c + q_\cdot)\right]\right\},
\end{aligned} \tag{5}
$$

where $\boldsymbol{\theta} := \{\gamma_0, c, q_1, \ldots, q_J\}$. This formula indicates that to add a new row to $\mathbf{N}_J \in \mathbb{Z}^{J \times K_J}$, we first draw count $\text{NB}[n_{\cdot k}, q_{J+1}/(c + q_\cdot + q_{J+1})]$ at each existing column. We then draw K_{J+1}^+ new columns as $K_{J+1}^+ \sim \text{Pois}\{\gamma_0 [\ln(c + q_\cdot + q_{J+1}) - \ln(c + q_\cdot)]\}$. Finally, each entry in the new columns has a Logarithmic $[n_{(J+1)k}; q_{J+1}/(c + q_\cdot + q_{J+1})]$ distributed random count.

It is clear in the sequential construction of the scaled NBP random count matrix, for a point of discontinuity $\omega_k \in \mathcal{D}_J$, the variance and mean are related as

$$\text{var}[n_{(J+1)k}] = \mathbb{E}[n_{(J+1)k}] + \frac{\mathbb{E}^2[n_{(J+1)k}]}{n_{\cdot k}}. \tag{6}$$

Since $n_{\cdot k}$, the total count of gene k of all the J samples of the group, is fixed, the above equation indicates a variance and mean relationship that does not change.

3.1.1 Inference for the scaled NBP

The parameters of the scaled NBP can be inferred using Gibbs sampling with closed-form update equations. Using likelihoods (2) and (3), with $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$, $c \sim \text{Gamma}(c_0, 1/d_0)$, and $q_j \sim \text{Gamma}(a_0, 1/b_0)$ in the prior, each Gibbs sampling iteration proceeds as

$$(\gamma_0 \,|\, -) \sim \text{Gamma}\left(e_0 + K_J, \frac{1}{f_0 - \ln(\frac{c}{c+q_\cdot})}\right),$$

$$(r_k \,|\, -) \sim \text{Gamma}[n_{\cdot k}, 1/(c + q_\cdot)],$$

$$[G(\Omega \setminus \mathcal{D}_J) \,|\, -] \sim \text{Gamma}[\gamma_0, 1/(c + q_\cdot)],$$

$$(q_j \,|\, -) \sim \text{Gamma}\{a_0 + n_j, 1/[b_0 + G(\Omega)]\},$$

$$(c \,|\, -) \sim \text{Gamma}\{c_0 + \gamma_0, 1/[d_0 + G(\Omega)]\}, \tag{7}$$

where $G(\Omega) := G(\Omega \backslash \mathcal{D}_J) + \sum_{k=1}^{K_J} r_k$.

3.1.2 NBP-Seq differential expression analysis

To detect differentially expressed genes using the scaled NBP, we notice in the prior that

$$\mathbb{E}[n_{jk} \,|\, q_j, G] = \text{var}[n_{jk} \,|\, q_j, G] = q_j r_k$$

and in the conditional posterior that

$$\mathbb{E}[r_k \,|\, -] = n_{\cdot k}/(c + q_\cdot), \qquad \mathbb{E}[q_j \,|\, -] = (a_0 + n_j)/[b_0 + G(\Omega)]. \tag{8}$$

11

Thus one may consider r_k as a gene-specific Poisson rate parameter that indicates the expression level of gene k, whose conditional posterior is related to both $n_{.k}$, the total count of gene k across all the J samples of the group, and $q_.$, the total sum of the sample-specific gamma distributed scaling parameters; one may consider q_j as a scaling factor to be inferred from the data, whose conditional posterior is determined not only by n_j, the total count of all genes in sample j that indicates the sequencing depth of sample j, but also by $G(\Omega)$, the total sum of all countably infinite gene-specific Poisson rate parameters; and the conditional posterior of γ_0 is clearly related to K_J, the total number of expressed genes in the group. Therefore, the scaled NBP borrows statistical strength across both the genes and samples to infer the conditional posterior of r_k.

To assess whether the difference between the expressions of the same gene at different sample groups is statistically significant, we collect posterior Markov chain Monte Carlo (MCMC) samples for each r_k in each group, and use these MCMC samples to measure the distance between the posterior distributions of the r_k of the same gene across different groups. Note that for a gene whose total count across all samples in a group is zero, the posterior values of its r_k would be fixed at 0.

Instead of using the scaled NBP that introduces q_j to model sample-specific sequencing depths, we also consider the original NBP of Zhou et al. [2015] with all q_j fixed at one. To compensate for the variations of sequencing depths between samples, for the original NBP, we normalize the inferred Poisson rates r_k and use them to evaluate the significance of differential gene expressions.

3.2 GNBP-Seq: Gamma-negative binomial process for RNA-Seq

To generate the random count matrix $\mathbf{N}_J$ in a group, we construct a gamma-negative binomial process (GNBP) [Zhou et al., 2015] as

$$X_j \,|\, G \sim \text{NBP}(G, p_j), \quad G \sim \Gamma\text{P}(G_0, 1/c), \tag{9}$$

where $j \in \{1,..,J\}$ and $X_j \,|\, G \sim \text{NBP}(G, p_j)$ is defined as a NBP such that $X_j(A) \sim$ $\text{NB}[G(A), p_j]$ for each Borel subset $A \subset \Omega$. Note that $X_j \,|\, G \sim \text{NBP}(G, p_j)$ can also be augmented as a gamma process mixed sum-logarithmic process (SumLogP) as

$$X_j \,|\, L_j \sim \text{SumLogP}(L_j, p_j), \ \ L_j \,|\, G \sim \text{PP}(q_j G), \tag{10}$$

where $q_j := -\ln(1 - p_j)$, i.e., $p_j = 1 - e^{-q_j}$, and the SumLogP is defined in Zhou et al. [2015] such that $X_j(A) = \sum_{t=1}^{L_j(A)} u_t, \ u_t \sim \text{Logarithmic}(p_j)$ for each Borel subset $A \subset \Omega$. Thus the GNBP also can be expressed as a NBP mixed SumLogP as

$$X_j \,|\, L_j \sim \text{SumLogP}(L_j, p_j), \ \ (L_1, \ldots, L_J) \sim \text{NBP}(G_0, c, q_1, \ldots, q_J). \tag{11}$$

With a draw from the gamma process G expressed as $G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}$, a draw from X_j can be expressed as

$$X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\omega_k}, \ n_{jk} \sim \text{NB}(r_k, p_j). \tag{12}$$

The GNBP employs sample-specific NB probability parameters p_j to model row heterogeneity. In the context of RNA-Seq data, the variations of p_j can be used to account for those of sequencing depths.

Both the row-wise and column-wise constructions of the GNBP random count matrix mimic these of the NBP random count matrix. They are described in detail in Zhou et al. [2015] and hence omitted here for brevity. We mention that the two key differences in their row-wise sequential constructions are that the GNBP uses the gamma-NB instead of NB distributions to model the counts at previously expressed genes brought by a new sample, and the GNBP uses the logarithmic mixed sum-logarithmic instead of logarithmic distributions to model the counts at newly expressed genes brought by a new sample.

As shown in Zhou et al. [2015], in the sequential construction of the GNBP random

count matrix, for a point of discontinuity $\omega_k \in \mathcal{D}_J$, the variance and mean are related as

$$\mathrm{var}[n_{(J+1)k}] = \frac{\mathbb{E}[n_{(J+1)k}]}{1 - p_{J+1}} + \frac{\mathbb{E}^2[n_{(J+1)k}]}{l_{\cdot k}}, \tag{13}$$

which depends on both p_{J+1} and $l_{\cdot k}$ that are random, where $l_{\cdot k} := \sum_{j=1}^{J} l_{jk}$, $l_{jk} \sim$ CRT(n_{jk}, r_k), with the Chinese Restaurant Table (CRT) distribution defined as in Zhou and Carin [2015]. Comparing (6) and (13), it is clear that since $p_{J+1} < 1$ and $l_{\cdot k} \leq n_{\cdot k}$, the GNBP can model much more over-dispersed counts than the NBP.

3.2.1 Inference for the GNBP

Letting $\gamma_0 \sim \mathrm{Gamma}(e_0, 1/f_0)$, $p_j \sim \mathrm{Beta}(a_0, b_0)$, and $c \sim \mathrm{Gamma}(c_0, 1/d_0)$ in the prior, as in [Zhou et al., 2015], a Gibbs sampling iteration for the GNBP proceeds as

$$(\gamma_0 \,|\, -) \sim \mathrm{Gamma}\left(e_0 + K_J, \frac{1}{f_0 - \ln(\frac{c}{c+q_\cdot})}\right),$$

$$(l_{jk} \,|\, -) \sim \mathrm{CRT}(n_{jk}, r_k), \ \ (r_k \,|\, -) \sim \mathrm{Gamma}\big[l_{\cdot k}, 1/(c + q_\cdot)\big],$$

$$\{G(\Omega \backslash \mathcal{D}_J) \,|\, -\} \sim \mathrm{Gamma}\big[\gamma_0, 1/(c + q_\cdot)\big],$$

$$(p_j \,|\, -) \sim \mathrm{Beta}\big[a_0 + n_j, b_0 + G(\Omega)\big],$$

$$(c \,|\, -) \sim \mathrm{Gamma}\big\{c_0 + \gamma_0, 1/[d_0 + G(\Omega)]\big\}. \tag{14}$$

3.2.2 GNBP-Seq differential expression analysis

In the GNBP, since in the prior we have

$$\mathbb{E}[n_{jk} \,|\, G, p_j] = r_k \frac{p_j}{1 - p_j},$$

$$\mathrm{var}[n_{jk} \,|\, G, p_j] = r_k \frac{p_j}{(1 - p_j)^2} = \mathbb{E}[n_{jk} \,|\, G, p_j] + r_k^{-1} \mathbb{E}^2[n_{jk} \,|\, G, p_j],$$

and in the conditional posterior, if $b_0 + G(\Omega) > 1$, we have

14

$$\mathbb{E}[r_k \,|\, -] = l_{\cdot k}/(c + q.), \quad \mathbb{E}[p_j/(1 - p_j) \,|\, -] = (a_0 + n_j)/(b_0 + G(\Omega) - 1). \qquad (15)$$

Thus one may interpret $p_j/(1 - p_j)$ as a term that accounts for the sequencing depth of sample j, and may compare the posterior distributions of the NB shape parameter r_k of the same gene at different groups to assess differential expression of that gene. For an unexpressed gene, whose total count across all samples in a group is zero, the posterior values of its r_k would be fixed at 0.

Comparing (8) and (15) shows that both the GNBP and scaled NBP have similar sample-specific scaling parameters, but, as in (14), since $\mathbb{E}[l_{jk} \,|\, -] = \sum_{t=1}^{n_{jk}} r_k/(r_k + t - 1)$ and hence $\mathbb{E}[l_{jk} \,|\, -] \approx r_k \ln(n_{jk} + r_k)$ for large n_{jk}, the posteriors of the gene-specific parameters r_k in the GNBP would be impacted much less by some genes whose expressions n_{jk} are significantly larger than their mean expression levels, which are commonly observed in genomic studies.

3.3 BNBP-Seq: Beta-negative binomial process for RNA-Seq

Similar to the GNBP, the BNBP can be used to model the RNA-Seq samples. The BNBP can be constructed by sharing NB probability parameters across the J sequencing samples of the same group as

$$X_j \,|\, r_j, B \sim \text{NBP}(r_j, B), \quad B \sim \text{BP}(c, B_0), \qquad (16)$$

where $j \in \{1, \ldots, J\}$ and $B \sim \text{BP}(c, B_0)$ is a beta process with a finite and continuous base measure B_0 over Ω and a concentration parameter c, with Lévy measure

$$\nu(dpd\omega) = p^{-1}(1 - p)^{c-1}dpB_0(d\omega). \qquad (17)$$

With a draw from the beta process B expressed as $B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}$, where ω_k and p_k are atoms and their associated probability weights, respectively, a draw from X_j given B

can be expressed as

$$X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k}, \ \ n_{jk} \sim \text{NB}(r_j, p_k). \tag{18}$$

In the BNBP, different r_j's are used to model the sequencing depth variations.

Both the row-wise and column-wise constructions of the BNBP random count matrix, as described in detail in Zhou et al. [2015] and hence omitted here for brevity, mimic these of the scaled NBP random count matrix. We mention that the two key differences in their row-wise sequential constructions are that the BNBP uses the beta-NB instead of NB distributions to model the counts at previously expressed genes brought by a new sample, and the BNBP uses the digamma instead of logarithmic distributions to model the counts at newly expressed genes brought by a new sample.

As shown in Zhou et al. [2015], in the sequential construction of the BNBP random count matrix, for a point of discontinuity ω_k, the variance and mean are related as

$$\text{var}[n_{(J+1)k}] = \frac{\mathbb{E}[n_{(J+1)k}]}{\frac{c+r.-2}{n_{\cdot k}+c+r.-1}} + \frac{\mathbb{E}^2[n_{(J+1)k}]}{\frac{n_{\cdot k}(c+r.-2)}{n_{\cdot k}+c+r.-1}}, \tag{19}$$

which depends on both c and $r.$ that are random. Comparing (6) and (19), it is clear that since $\frac{c+r.-2}{n_{\cdot k}+c+r.-1} \leq 1$ and $\frac{n_{\cdot k}(c+r.-2)}{n_{\cdot k}+c+r.-1} < n_{\cdot k}$ for $c + r. > 2$, similar to the GNBP, the BNBP can also model much more over-dispersed counts than the scaled NBP.

The variance-mean relationships expressed by (6), (13), and (19) show that the GNBP and BNBP can model much more over-dispersed counts than the (scaled) NBP, and as shown in Figure 1 of Zhou et al. [2015], given the same expected total count, while the counts in NBP random count matrices usually have small dynamic ranges, the counts in both the GNBP and BNBP matrices can contain values that are significantly above the average. In RNA-Seq, it is common to have large dynamical range for highly over-dispersed gene counts, which are likely to be better modeled by both the GNBP and BNBP than by the (scaled) NBP, as confirmed by our experiments in Section 4.

3.3.1 Inference for the BNBP

Letting $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$, $p_j \sim \text{Beta}(a_0, b_0)$, and $c \sim \text{Gamma}(c_0, 1/d_0)$, as in Zhou et al. [2015], a Gibbs sampling iteration for the BNBP proceeds as

$$(\gamma_0 \,|\, -) \sim \text{Gamma}\left(e_0 + K_J, \frac{1}{f_0 + \psi(c + r.) - \psi(c)}\right),$$

$$(p_k \,|\, -) \sim \text{Beta}(n_{.k}, c + r.), \quad (p_* \,|\, -) \sim \text{logBeta}(\gamma_0, c + r.),$$

$$(l_{jk} | -) \sim \text{CRT}(n_{jk}, r_j),$$

$$(r_j \,|\, -) \sim \text{Gamma}\left(a_0 + l_{j.}, \frac{1}{b_0 + p_* - \sum_{k=1}^{K_J} \ln(1 - p_k)}\right). \tag{20}$$

Inside each Gibbs sampling iteration, as in Zhou et al. [2015], an independence chain Metropolis-Hastings sampling step can be used to update the concentration parameter c.

3.3.2 BNBP-Seq differential expression analysis

In the BNBP, since in the prior we have

$$\mathbb{E}[n_{jk} \,|\, r_j, B] = r_j \frac{p_k}{1 - p_k},$$

$$\text{var}[n_{jk} \,|\, r_j, B] = r_j \frac{p_k}{(1 - p_k)^2} = (1 - p_k)^{-1} \mathbb{E}[n_{jk} \,|\, r_j, B], \tag{21}$$

and in the conditional posterior, if $c + r. > 1$, we have

$$\mathbb{E}[p_k/(1 - p_k) \,|\, -] = n_{.k}/(c + r. - 1),$$

$$\mathbb{E}[r_j \,|\, -] = \frac{a_0 + l_{j.}}{b_0 + p_* - \sum_{k=1}^{K_J} \ln(1 - p_k)}. \tag{22}$$

Thus one may consider that the NB sample-specific shape parameter r_j accounts for the sequencing depth of sample j, and may compare the posterior distributions of $p_k/(1 - p_k)$ to evaluate differential expression of gene k between different groups. Note that for an unexpressed gene, whose total count across all samples in a group is zero, the posterior

17

values of its p_k would be fixed at 0.

Comparing (8) and (22) shows that the BNBP and scaled NBP have similar gene-specific parameters, but, as in (20), since $\mathbb{E}[l_{jk}\,|-] \approx r_j \ln(n_{jk}+r_j)$ for large n_{jk}, for some genes whose expressions n_{jk} are significantly larger than the mean expression levels, the posteriors of the sample-specific parameters r_j in the BNBP also would be impacted much less than these of the sample-specific parameters q_j in the scaled NBP.

3.4 Distance between posterior distributions

In order to compare the posterior distributions, we use the symmetric Kullback-Leibler (KL) divergence defined between two discrete distributions P and Q as

$$KL(P,Q) = \sum_x \left[p(x) - q(x)\right] \log\left[p(x)/q(x)\right].$$

Supposing r is the parameter to be compared between two different groups, we estimate the symmetric KL-divergence between the posterior distributions of $r^{(1)}$ and $r^{(2)}$, the values of r of the first and second groups, respectively, using collected MCMC samples. We first find both the minimum and maximum values of the MCMC samples of r across both groups to define an interval for r. After adjusting the lower- and upper-limits of the interval as $[\max(0, Q_1 - 1.5 * Q_\triangle), \;\; Q_3 + 1.5 * Q_\triangle]$, where Q_1 and Q_3 are 25% and 75% quantiles and $Q_\triangle = Q_3 - Q_1$, we equally divide the adjusted interval into $N = 100$ bins. For each group, we count the number of MCMC samples falling into each bin, and then normalize these bin counts to a 100 dimensional discrete probability vector, referred to as $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ for the first and second groups, respectively. Finally, with a small constant set as $\epsilon = 10^{-10}$, we calculate the symmetric KL-divergence as

$$KL(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}) = \sum_{i=1}^{N} \left(\pi_i^{(1)} - \pi_i^{(2)}\right) \log\left(\frac{\pi_i^{(1)} + \epsilon}{\pi_i^{(2)} + \epsilon}\right). \tag{23}$$

4 Experimental Results

To evaluate the proposed BNP differential expression analysis algorithms, we compare their performance on both synthetic and real benchmark RNA-Seq data with those of edgeR [Robinson et al., 2010], DESeq [Anders and Huber, 2010], and baySeq [Hardcastle and Kelly, 2010], three widely used algorithms in biomedical studies. We first consider synthetic RNA-Seq data generated under different models, and we show that the proposed GNBP and BNBP differential expression analysis algorithms consistently provide outstanding performance. We then consider the real benchmark RNA-Seq data extracted from the SEquencing Quality Control (SEQC) project [Xu et al., 2013, SEQC/MAQC-III Consortium, 2014]. We consider the RNA-Seq data from both Beijing Genomics Institute (BGI) and the Pennsylvania State University (PSU) provided in the SEQC project [Xu et al., 2013, SEQC/MAQC-III Consortium, 2014], available in the R package SEQC on Bioconductor [Gentleman et al., 2004]. Both the BGI and PSU datasets, which are the transcriptomic expression measurements of the RNA samples prepared at the same biological conditions but sequenced at different sequencing sites, contain the counts for approximately 26,000 genes. In our experiments, we employ sample groups A and B, which are derived from Agilent's Universal Human Reference RNA and Life Technologies' Human Brain Reference RNA cell lines, respectively. We collect the counts from the first flow cells of the sequencing machines on five replicates for each group (condition).

On both synthetic and real RNA-Seq count data, comparison of both the area under the receiver operating characteristic (ROC) curve (AUC-ROC) and area under the precision-recall (PR) curve (AUC-PR) shows that the proposed GNBP and BNBP algorithms clearly outperform the (scaled) NBP and previously proposed differential expression analysis algorithms, as described below in detail.

4.1 Synthetic data

We first generate synthetic RNA-Seq data with the GNBP generative model, the BNBP generative model, or the NB distribution based procedure adopted in baySeq [Hardcastle and Kelly, 2010]. For each setting, to make the synthetic data closely resemble real RNA-Seq data, we first infer the parameters of the corresponding model on the BGI or PSU datasets from SEQC, and then generate synthetic sequencing counts using these inferred model parameters. To simulate samples from two different groups (conditions), each of which has five replicates, we randomly select 10% of the genes and set them to be differentially expressed between the two groups, with the fold change of differentially expressed genes chosen as an adjustable parameter. For quality control, we discard the bottom 10% of genes with low expressions across groups in data generation. In order to produce both up- and down-regulated differentially expressed genes, each differentially expressed gene is randomly set to be either up- or down-regulated. Below we denote $b > 1$ as the fold change to be set. We use the PSU dataset for the baySeq setting and the BGI dataset for both the GNBP and BNBP settings. Using different datasets to infer model parameters and different models to generate synthetic datasets allows us to assess the robustness of various methods in different practical settings.

In the GNBP setting, if gene k is up-regulated, then we generate its counts using $\text{NB}(r_k, p_j)$ and $\text{NB}(b\, r_k, p_j)$ for the samples in the first and second groups, respectively; whereas if gene k is down-regulated, then we generate its counts using $\text{NB}(r_k/b, p_j)$ and $\text{NB}(r_k, p_j)$ for the five samples in the first and second groups, respectively.

In the BNBP setting, if gene k is up-regulated, then we generate its counts using $\text{NB}(r_j, p_k)$ and $\text{NB}(r_j, p_k')$, where p_k' is selected to satisfy $bp_k/(1 - p_k) = p_k'/(1 - p_k')$, for the samples in the first and second groups, respectively; whereas if gene k is down-regulated, then we generate its counts using $\text{NB}(r_j, \tilde{p}_k)$ and $\text{NB}(r_k, p_k)$, where where $\tilde{p}_k$ is selected to satisfy $p_k/(1 - p_k) = b\tilde{p}_k/(1 - \tilde{p}_k)$, for the five samples in the first and second groups, respectively.

In the baySeq setting of Hardcastle and Kelly [2010] that generates a count from a NB distribution given its mean and dispersion parameter, if gene k is up-regulated, then we generate its counts using $b\mu_k$ and μ_k as the means for the first and second groups, respectively; whereas if gene k is down-regulated, then we generate its counts using μ_k and μ_k/b as the means for the first and second groups, respectively.

We infer the model parameters via Gibbs sampling for the proposed BNP differential expression analysis algorithms. For each algorithm, we collect 1,000 MCMC samples after 1,000 burn-in iterations. We use the collected MCMC samples to calculate the symmetric KL divergence, as in (23), between two groups for each gene, and rank the genes according to these values. For edgeR and DESeq, we follow the standard analysis pipelines and rank the genes using the computed p-values; and for baySeq, we rank the genes using model likelihoods. We set the fold change b as 1.4, 1.6, 1.8, or 2 in simulating synthetic data to assess how sensitive the algorithms under study are to different levels of differential expression. For each fold change, we report the results of each algorithm based on ten independent random trials.

For these three different types of synthetic data, as shown in Figure 1, measured by both AUC-ROC and AUC-PR, baySeq has the worst overall performance even when the synthetic data are generated based on its model assumption, followed by the scaled NBP; the NBP, DESeq, and edgeR all have similar performance; and the GNBP and BNBP clearly outperform all the other differential expression analysis algorithms. To further compare the operating characteristics of different algorithms, we show in Figure 5 in the Appendix the full ROC and PR curves for the fold change of $b = 1.8$. We provide in the Appendix the detailed numerical values used to plot Figure 1, as shown in Tables 1-6, where for each setting the best result and the ones that are less than one standard deviation away from the best are highlighted in bold.
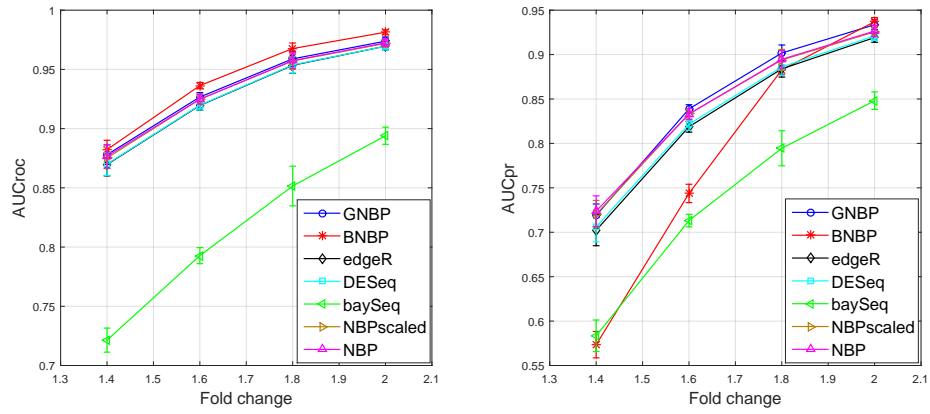
More carefully examining Figures 1 and 5, it is interesting to notice that for the synthetic data generated with either the GNBP or BNBP, the scaled NBP, which extends
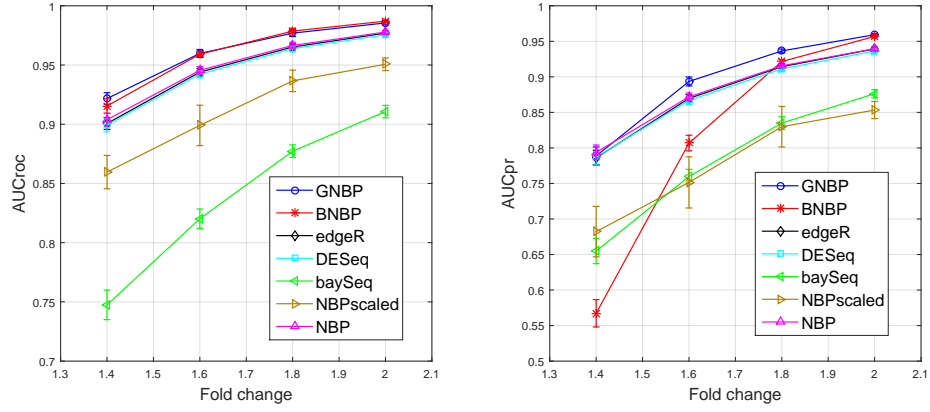
(a) GNBP synthetic data
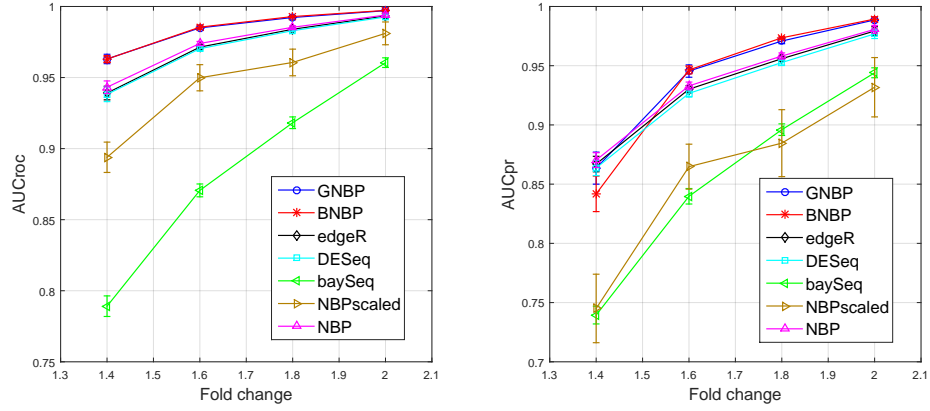
(b) BNBP synthetic data

(c) baySeq synthetic data

Figure 1: **left column**: AUC-ROC values, **right column**: AUC-PR values. Performance comparison of different methods in detecting differentially expressed genes under various fold changes, using synthetic data generated under three different negative binomial distribution based models.

the original NBP with sample-specific scaling parameters q_j to model sample sequencing depth variations, in fact clearly underperforms the original NBP. Suggesting that explicitly modeling the sample sequencing depths, using the gamma-Poisson construction of the scaled NBP, is insufficient to model the over-dispersed gene counts generated using the gamma- or beta-NB constructions.

While the original NBP fixes $q_j = 1$ and hence does not explicitly model the sample sequencing depth variations, it performs as well as both DESeq and edgeR in all three settings, which may be explained by the fact that it normalizes the posterior Poisson rates before applying them to compare the gene expression levels between two groups, a post-processing step that plays a similar role as the pre-processing normalization steps used in both DESeq and edgeR to account for different sequencing depths.

It is also interesting to notice that while the GNBP consistently ranks the best or very close to the best, in terms of both AUC-ROC and AUC-PR, the BNBP does so only in terms of AUC-ROC. For synthetic data generated using the GNBP and baySeq, the performance of the BNBP in terms of AUC-PR quickly deteriorates as the fold change reduces from 1.8 to 1.4, suggesting a large number of false positives among the top ranked genes of the BNBP when the fold change is not sufficiently large for the GNBP synthetic data. The disparity between the performance measured by AUC-ROC and that measured by AUC-PR, which only happens for the BNBP, indicates that the BNBP employs a distinct mechanism to detect differentially expressed genes, as carefully discussed below.

To compare the expression levels of the kth gene between two groups, the GNBP compares the posterior NB shape parameters r_k, whereas the BNBP compares the posterior NB probability parameters p_k. The difference between the posterior NB shape parameters r_k explains the differences between both the means and dispersions, but does not explain that of the variance-to-mean ratios (VMR), of the counts of gene k at different groups, since if $n_{jk} \sim \text{NB}(r_k, p_j)$, then $\mathbb{E}[n_{jk}] = r_k p_j / (1 - p_j)$, $\text{var}[n_{jk}] = \mathbb{E}[n_{jk}] + (\mathbb{E}[n_{jk}])^2 / r_k$, and $\text{VMR}[n_{jk}] = 1 + \mathbb{E}[n_{jk}] / r_k$; whereas the difference between the

posterior NB probability parameters p_k explains the differences between both the means and VMRs of the counts of gene k at different groups, since if $n_{jk} \sim \text{NB}(r_j, p_k)$, then $\mathbb{E}[n_{jk}] = r_j p_k/(1 - p_k)$, $\text{var}[n_{jk}] = \mathbb{E}[n_{jk}] + (\mathbb{E}[n_{jk}])^2/r_j$, and $\text{VMR}[n_{jk}] = 1/(1 - p_k)$. Therefore, for the counts of a gene generated with the GNBP, if the r_k is small, a small change in its value may lead to a significant change of $\text{VMR}[n_{jk}] = 1 + (\mathbb{E}[n_{jk}])/r_k$, which implies that a large difference in a gene's VMRs between two groups may not be taken by the GNBP as a strong evidence for differential expression. By contrast, since the gene-specific parameter p_k in the BNBP is explicitly responsible for the VMR, a large difference in a gene's VMRs between two groups may encourage the BNBP to rank that gene as strongly differentially expressed, which may be used to explain why the BNBP has good AUC-ROC but poor AUC-PR if the fold change is small for the GNBP synthetic data. In practice, however, it is often unclear whether it is the change of the quadratic relationship between the variance and mean, as captured by the NB dispersion parameter, or the VMR, as captured by the NB probability parameter, that is responsible for the change of a gene's expression level. Thus it is often unclear whether the GNBP or BNBP would be a better choice for a real dataset, and it seems promising to combine the advantages of both for differential expression analysis, a research topic beyond the scope of the paper that is to be investigated in our future study.
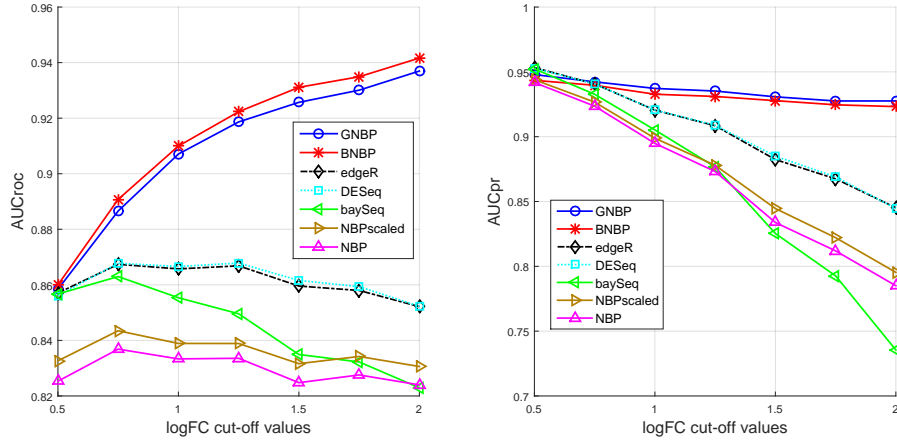
4.2   SEQC benchmark RNA-Seq data

In order to characterize various RNA-Seq technologies and quantification pipelines in the SEQC project [Xu et al., 2013, SEQC/MAQC-III Consortium, 2014], the same RNA samples for a comprehensive group of control genes were analyzed based on quantitative Reverse Transcription Polymerase Chain Reaction (qRT-PCR) using TaqMan assays [Joyce, 2002], which is referred as the TaqMan benchmark data [Shi et al., 2006, MAQC Consortium, 2010]. For sample groups A and B, the expression intensity values of 955 selected control genes have been derived in the TaqMan qPT-PCR analysis for se-

quencing benchmarking. Without knowing in practice which genes are truly differentially expressed between different conditions, we consider thresholding the qRT-PCR expression ratios between different conditions at a certain value to define the ground-truth set of differentially expressed genes. Based on these 955 genes in the TaqMan data, we evaluate the performance of different differential expression analysis pipelines.

While it is unknown which genes are truly differentially expressed for both the BGI and PSU RNA-Seq data, we rely on the qRT-PCR expression intensity of the 955 genes in the TaqMan data and set different cut-offs for the binary logarithm (log2) of the qRT-PCR expression ratio to define "truly" differentially expressed genes. We increase this log2 cut-off value gradually from 0.5 to 2, and calculate both AUC-ROC and AUC-PR. The symmetric KL divergence is used to assess differential expression. As shown in Figure 2 for both the BGI and PSU datasets, the GNBP and BNBP outperform all the other methods in both ROC and PR analyses with significant margins. Note that the performance gains of the GNBP and BNBP over the other methods become more significant as one increases the log2 cut-off for the qRT-PCR expression ratio, which reduces the number of genes that are considered as truly differentially expressed.

Comparing Figure 1 for synthetic data with Figure 2 for real data, one may notice that while both the AUC-ROC and AUC-PR curves in Figure 1 seem to monotonically increase as the fold change increases, the AUC-ROC and AUC-PR curves in Figure 2 do not necessarily share similar trends. It is not surprising to observe these seemly distinct behaviors, since for the synthetic data in Figure 1, the set of truly differentially expressed genes are fixed and known exactly, remaining unchanged regardless of how one sets the fold change that is used to detect differentially expressed genes, whereas for the real data in Figure 2, the number of genes considered as truly expressed reduces as the cut-off value of the qRT-PCR expression ratio increases. In addition, we note that the results of edgeR, DESeq, and baySeq on both the BGI and PSU real datasets reported in this paper are similar to those reported in [Rapaport et al., 2013].

(a) BGI dataset
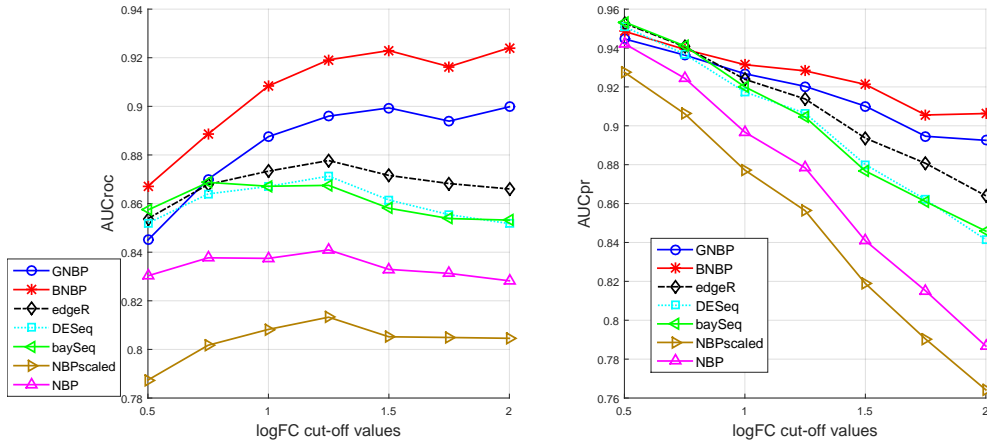
(b) PSU dataset

Figure 2: **left column**: AUC-ROC values, **right column**: AUC-PR values. Performance comparison of different methods in detecting differentially expressed genes on real benchmark RNA-Seq data from the SEQC project.

To investigate the experimental results more thoroughly, we fix the true positives and negatives at the log2 cut-off value of 2 and illustrate the ROC and PR curves for BGI dataset in Figure 3. It is clear that both the GNBP and BNBP not only have higher AUC-ROC and AUC-PR, but also outperform all the other methods in almost all regions of the ROC and PR curves.

In addition to showing the ROC and PR curves, we also plot the number of false discoveries to highlight the performance on the top ranked genes. Since there are 400
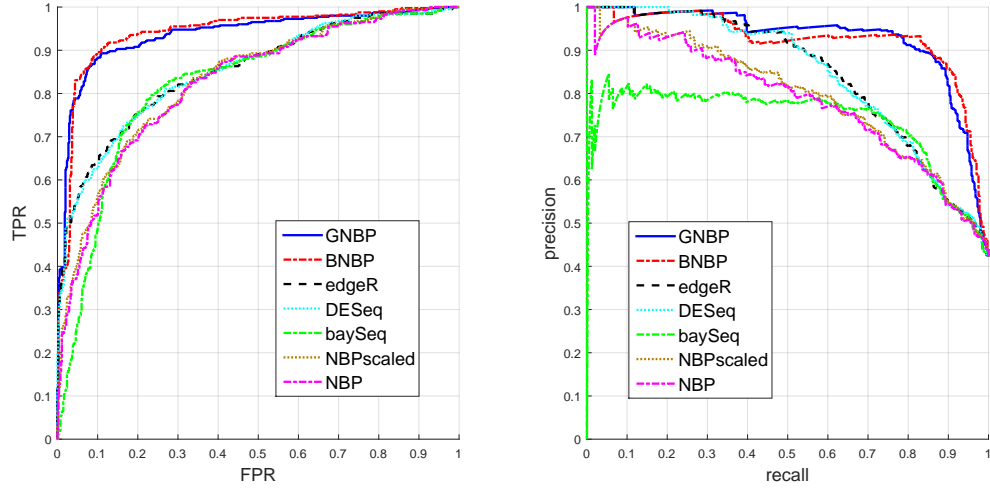
Figure 3: **left**: ROC curves, **right**: Precision-Recall (PR) curves. Performance comparison of different methods with the log2 cut-off value fixed at 2 for the BGI dataset from the SEQC project.
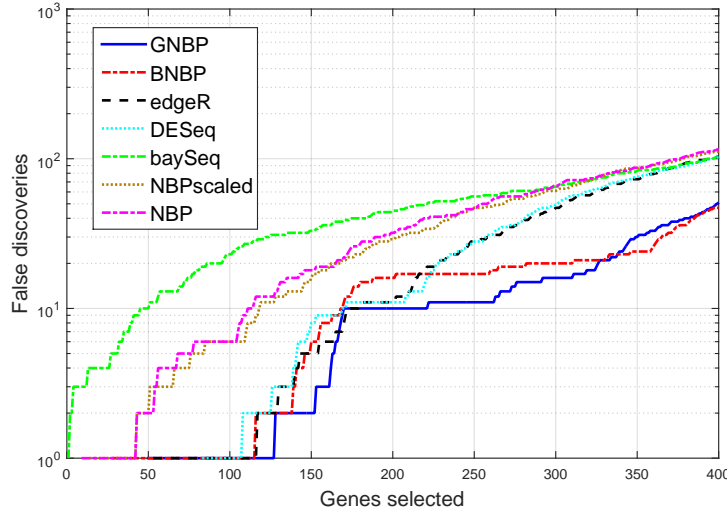

Figure 4: False discovery plots for different methods on the BGI dataset from the SEQC project, with the log2 cut-off value fixed at 2. The x-axis shows the number of genes selected, in order of their detected differential expression levels, and the y-axis shows the number of selected genes that are false positives.

truly differentially expressed genes based on the log2 cut-off value of 2, the top 400 genes detected by each approach are selected and the number of false discoveries are plotted. It is clear from Figure 4 that the GNBP and BNBP return much smaller number of false positives in comparison to all the other differential expression analysis algorithms.

5 Conclusions

We exploit Bayesian nonparametric priors, including the gamma-Poisson, gamma-negative binomial, and beta-negative binomial processes, to model RNA sequencing count matrices. With different sequencing depths captured by sample-specific model parameters, the posterior distributions of certain gene-specific model parameters are used to detect the genes that are differentially expressed between different conditions. With the model parameters inferred by borrowing statistical strength across both the genes and samples, there is no need to adjust the raw counts using heuristics before downstream analyses, an important pre-processing step that is often required in previously proposed algorithms. Example results on both synthetic and real RNA-Seq data demonstrate the state-of-the-art performance of both the gamma- and beta-negative binomial processes based differential expression analysis algorithms. Given the success of the proposed random-process-based algorithms in differential expression analysis, it is of interest to investigate Bayesian nonparametric algorithms for many other real-world applications in biomedicine that require analyzing next-generation sequencing data.

6 Acknowledgments

References

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

C. I. Bliss and R. A. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200, 1953.

T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

F. Caron, Y. W. Teh, and B. T. Murphy. Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data. *Annal of Applied Statistics*, 2014.

S. Datta and D. Nettleton. *Statistical Analysis of Next Generation Sequencing Data*. Frontiers in Probability and the Statistical Sciences. Springer International Publishing, 2014.

M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14 (6):671–683, 2013.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2): 209–230, 1973.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

M. Greenwood and G. U. Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. R. Stat. Soc.*, 1920.

T. J. Hardcastle and K. A. Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, 2010.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 1990.

C. Joyce. Quantitative rt-pcr. *RT-PCR Protocols*, pages 83–92, 2002.

J. F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

J. Li and R. Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research*, 22(5):519–536, 2013.

J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, page kxr031, 2011.

D. J. Lorenz, R. S. Gill, R. Mitra, and S. Datta. Using RNA-seq data to detect differentially expressed genes. In *Statistical Analysis of Next Generation Sequencing Data*, pages 25–49. Springer, 2014.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

J. Lovén, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, 2012.

MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827–838, 2010.

M. L. Metzker. Sequencing technologiesthe next generation. *Nature reviews genetics*, 11(1): 31–46, 2010.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

A. Oshlack, M. D. Robinson, and M. D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):1–10, 2010.

F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95, 2013.

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902, 2014a.

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. The role of spike-in standards in the normalization of RNA-seq. In *Statistical Analysis of Next Generation Sequencing Data*, pages 169–190. Springer, 2014b.

A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):1, 2011.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1–9, 2010.

M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467, 1995.

N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, 2016.

SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9):903–914, 2014.

L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. De Longueville, E. S. Kawasaki, K. Y. Lee, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

G. Smyth and A. Verbyla. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Stat. Soc: Series B*, 58(3):565–572, 1996.

C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:91, 2013.

L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138, 2010.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 2003.

J. Xu, Z. Su, H. Hong, J. Thierry-Mieg, D. Thierry-Mieg, D. P. Kreil, C. E. Mason, W. Tong, and L. Shi. Cross-platform ultradeep transcriptomic profiling of human reference RNA samples by RNA-Seq. *Scientific Data*, 1:140020–140020, 2013.

Z. H. Zhang, D. J. Jhaveri, V. M. Marshall, D. C. Bauer, J. Edson, R. K. Narayanan, G. J. Robinson, A. E. Lundberg, P. F. Bartlett, N. R. Wray, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. *PloS one*, 9(8):e103207, 2014.

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):307–320, 2015.

M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, pages 1462–1471, 2012.

M. Zhou, O. H. M. Padilla, and J. G. Scott. Priors for random count matrices derived from a family of negative binomial processes. *To appear in J. Amer. Statist. Assoc.*, 2015.

J. Zyprych-Walczak, A. Szabelska, L. Handschuh, K. Górczak, K. Klamecka, M. Figlerowicz, and I. Siatkowski. The impact of normalization methods on RNA-Seq data analysis. *BioMed Research International*, 2015, 2015.

A Additional tables and figures

Table 1: AUC-ROC in the GNBP simulation setup for different true fold changes.

Method	Fold change			
	1.4	1.6	1.8	2
GNBP	**0.9226** ± 0.006	**0.9625** ± 0.003	0.9777 ± 0.003	0.9864 ± 0.002
BNBP	0.9156 ± 0.005	0.9610 ± 0.003	**0.9783** ± 0.002	**0.9875** ± 0.002
edgeR	0.9004 ± 0.007	0.9463 ± 0.004	0.9653 ± 0.003	0.9778 ± 0.003
DESeq	0.8986 ± 0.008	0.9444 ± 0.004	0.9634 ± 0.003	0.9764 ± 0.003
baySeq	0.7542 ± 0.008	0.8247 ± 0.012	0.8752 ± 0.003	0.9114 ± 0.008
NBP	0.9035 ± 0.007	0.9476 ± 0.004	0.9665 ± 0.003	0.9786 ± 0.003
NBPscaled	0.8596 ± 0.014	0.8990 ± 0.017	0.9366 ± 0.009	0.9506 ± 0.0053

Table 2: AUC-PR in the GNBP simulation setup for different true fold changes.

Method	Fold change			
	1.4	1.6	1.8	2
GNBP	0.7873 ± 0.011	**0.8998** ± 0.006	**0.9382** ± 0.003	**0.9607** ± 0.003
BNBP	0.5660 ± 0.011	0.8189 ± 0.008	0.9213 ± 0.005	0.9563 ± 0.002
edgeR	0.7857 ± 0.015	0.8742 ± 0.007	0.9136 ± 0.003	0.9403 ± 0.003
DESeq	0.7848 ± 0.014	0.8714 ± 0.007	0.9107 ± 0.002	0.9369 ± 0.004
baySeq	0.6517 ± 0.012	0.7655 ± 0.015	0.8329 ± 0.003	0.8756 ± 0.004
NBP	**0.7934** ± 0.014	0.8770 ± 0.007	0.9156 ± 0.003	0.9399 ± 0.003
NBPscaled	0.6822 ± 0.035	0.7515 ± 0.036	0.8298 ± 0.028	0.8533 ± 0.012

Table 3: AUC-ROC in the BNBP simulation setup for different true fold changes.

Method	Fold change			
	1.4	1.6	1.8	2
GNBP	**0.9648** ± 0.001	**0.9847** ± 0.001	**0.9914** ± 0.0014	**0.9968** ± 0.001
BNBP	0.9635 ± 0.001	**0.9848** ± 0.002	**0.9922** ± 0.0009	**0.9971** ± 0.0009
edgeR	0.9399 ± 0.001	0.9706 ± 0.003	0.9829 ± 0.0017	0.9929 ± 0.00189
DESeq	0.9383 ± 0.002	0.9694 ± 0.003	0.9818 ± 0.0016	0.9920 ± 0.0018
baySeq	0.7919 ± 0.007	0.8699 ± 0.07	0.9167 ± 0.007	0.9590 ± 0.0041
NBP	0.9438 ± 0.001	0.9729 ± 0.003	0.9844 ± 0.002	0.9935 ± 0.0016
NBPscaled	0.8939 ± 0.0107	0.9499 ± 0.0092	0.9606 ± 0.0094	0.9811 ± 0.008

Table 4: AUC-PR in the BNBP simulation setup for different true fold changes.

Method	Fold change			
	1.4	1.6	1.8	2
GNBP	0.8632 ± 0.011	$\mathbf{0.9431} \pm 0.005$	$\mathbf{0.9703} \pm 0.003$	$\mathbf{0.9881} \pm 0.002$
BNBP	0.8356 ± 0.012	$\mathbf{0.9432} \pm 0.003$	$\mathbf{0.9725} \pm 0.002$	$\mathbf{0.9889} \pm 0.003$
edgeR	0.8674 ± 0.006	0.9275 ± 0.005	0.9557 ± 0.003	0.9783 ± 0.004
DESeq	0.8634 ± 0.004	0.9240 ± 0.005	0.9523 ± 0.003	0.9759 ± 0.003
baySeq	0.7413 ± 0.015	0.8408 ± 0.01	0.8963 ± 0.007	0.9434 ± 0.003
NBP	$\mathbf{0.8708} \pm 0.006$	0.9302 ± 0.005	0.9577 ± 0.003	0.9798 ± 0.003
NBPscaled	0.7450 ± 0.03	0.8648 ± 0.019	0.8846 ± 0.028	0.9318 ± 0.025

Table 5: AUC-ROC in the baySeq simulation setup for different true fold changes.

Method	Fold change			
	1.4	1.6	1.8	2
GNBP	0.8772 ± 0.009	0.9286 ± 0.005	0.9585 ± 0.004	0.9738 ± 0.001
BNBP	$\mathbf{0.8823} \pm 0.005$	$\mathbf{0.9382} \pm 0.004$	$\mathbf{0.9674} \pm 0.003$	$\mathbf{0.9812} \pm 0.0015$
edgeR	0.8702 ± 0.008	0.9216 ± 0.0042	0.9518 ± 0.004	0.9687 ± 0.003
DESeq	0.8705 ± 0.0083	0.9220 ± 0.004	0.9520 ± 0.0036	0.9688 ± 0.003
baySeq	0.7222 ± 0.0089	0.7887 ± 0.0067	0.8489 ± 0.012	0.8911 ± 0.0099
NBP	0.8769 ± 0.0075	0.9270 ± 0.0045	0.9567 ± 0.0031	0.9725 ± 0.0026
NBPscaled	0.8752 ± 0.009	0.9248 ± 0.0044	0.9571 ± 0.0071	0.9719 ± 0.0031

Table 6: AUC-PR in the baySeq simulation setup for different true fold changes.

Method	Fold change			
	1.4	1.6	1.8	2
GNBP	$\mathbf{0.7194} \pm 0.015$	$\mathbf{0.8372}\ 0.0095 \pm$	$\mathbf{0.8984} \pm 0.0074$	$\mathbf{0.9332} \pm 0.0041$
BNBP	0.5733 ± 0.012	0.7448 ± 0.014	0.8826 ± 0.0055	$\mathbf{0.9337} \pm 0.0058$
edgeR	0.7004 ± 0.013	0.8152 ± 0.008	0.8787 ± 0.0066	0.9173 ± 0.0054
DESeq	0.7042 ± 0.013	0.8180 ± 0.0082	0.8813 ± 0.0058	0.9194 ± 0.005
baySeq	0.5806 ± 0.0096	0.7034 ± 0.0057	0.7877 ± 0.0104	0.8482 ± 0.0106
NBP	$\mathbf{0.7223} \pm 0.0129$	0.8312 ± 0.0075	0.8913 ± 0.0058	0.9248 ± 0.0055
NBPscaled	0.7203 ± 0.0155	0.8333 ± 0.006	$\mathbf{0.8940} \pm 0.012$	0.9256 ± 0.0047

(a) GNBP setup
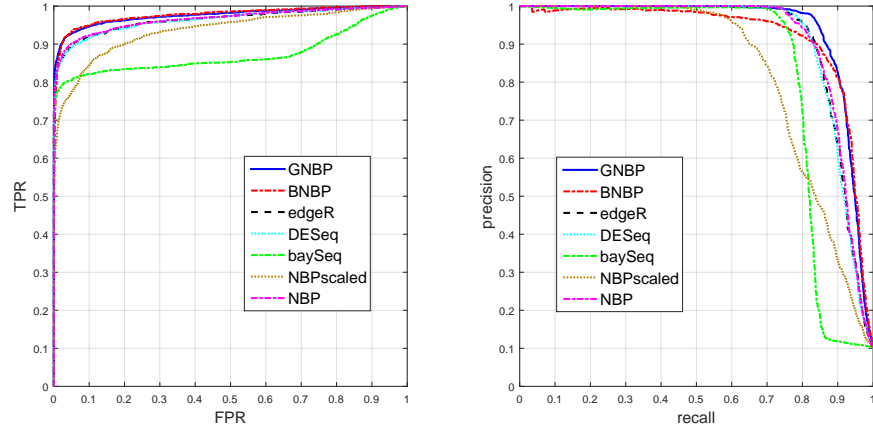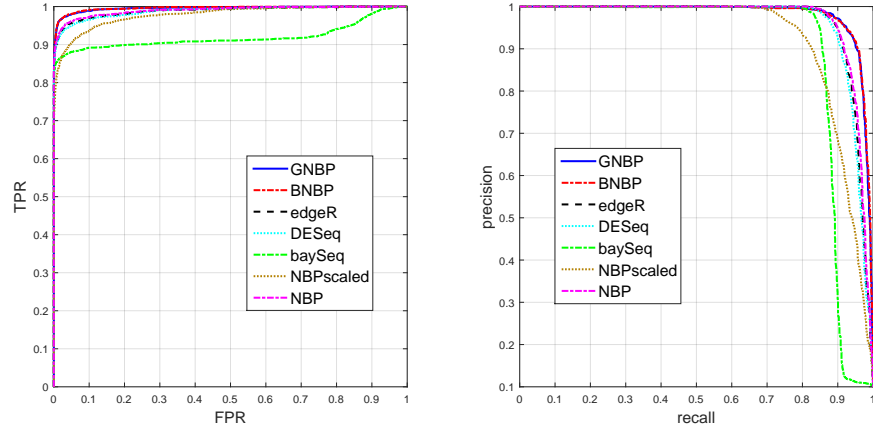
(b) BNBP setup
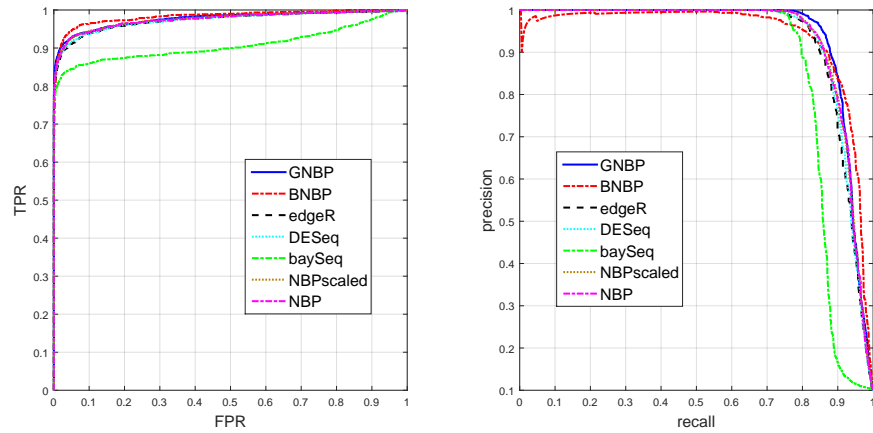
(c) baySeq setup

Figure 5: **left column**: ROC curve, **right column**: PR curve. Performance of different methods in detecting the differential expression of simulated data generated from different setups with a fold change of 1.8 for truly differentially expressed genes.