BayCount: A Bayesian Decomposition Method for Inferring Tumor Heterogeneity using RNA-Seq Counts

Fangzheng Xie[*] Mingyuan Zhou[†] Yanxun Xu[*‡]

Abstract

Tumors are heterogeneous – a tumor sample usually consists of a set of subclones with distinct transcriptional profiles and potentially different degrees of aggressiveness and responses to drugs. Understanding tumor heterogeneity is therefore critical for precise cancer prognosis and treatment. In this paper, we introduce BayCount, a Bayesian decomposition method to infer tumor heterogeneity with highly over-dispersed RNA sequencing count data. Using negative binomial factor analysis, BayCount takes into account both the between-sample and gene-specific random effects on raw counts of sequencing reads mapped to each gene. For the posterior inference, we develop an efficient compound Poisson based blocked Gibbs sampler. Simulation studies show that BayCount is able to accurately estimate the subclonal inference, including number of subclones, the proportions of these subclones in each tumor sample, and the gene expression profiles in each subclone. For real-world data examples, we apply BayCount to The Cancer Genome Atlas lung cancer and kidney cancer RNA sequencing count

[*] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA.

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

[‡] Correspondence should be addressed to Yanxun Xu (yanxun.xu@jhu.edu)

data and obtain biologically interpretable results. Our method represents the first effort in characterizing tumor heterogeneity using RNA sequencing count data that simultaneously removes the need of normalizing the counts, achieves statistical robustness, and obtains biologically/clinically meaningful insights. The R package BayCount implementing our model and algorithm is available for download.

KEY WORDS: Cancer genomics, compound Poisson, Markov chain Monte Carlo, negative binomial, over-dispersion

1 Introduction

Tumor heterogeneity (TH) is a phenomenon that describes distinct molecular profiles of different cells in one or more tumor samples. TH arises during the formation of a tumor as a fraction of cells acquire and accumulate different somatic events (*e.g.*, mutations in different cancer genes), resulting in heterogeneity within the same biological tissue sample and between different ones, spatially and temporally (Russnes et al., 2011; Ding et al., 2012). As a result, tumor cell populations are composed of different subclones (subpopulations) of cells, characterized by distinct genomes, transcriptional profiles (Kim et al., 2015), as well as other molecular profiles, such as copy number alterations. Understanding TH is critical for precise cancer prognosis and treatment. Heterogenetic tumors may exhibit different degrees of aggressiveness and responses to drugs among different samples due to genetic or gene expression differences. The level of heterogeneity itself can be used as a biomarker to predict treatment response or prognosis since more heterogeneous tumors are more likely to contain treatment-resistant subclones (Marusyk et al., 2012). This will ultimately facilitate the rational design of combination treatments, with each distinct compound targeting a specific tumor subclone based on its transcriptional profile.

Large-scale sequencing techniques provide valuable information for understanding tumor complexity and open a door for the desired statistical inference on TH. Previous studies have focused on reconstructing the subclonal composition by quantifying the structural subclonal copy number variations (Carter et al., 2012; Oesper et al., 2013), somatic mutations (Nik-Zainal et al., 2012; Roth et al., 2014; Xu et al., 2015), or both (Deshwar et al., 2015; Lee et al., 2016). In this paper, we aim to learn tumor transcriptional heterogeneity using RNA sequencing (RNA-Seq) data.

In the analysis of gene expression data, matrix decomposition models have been extensively studied in the context of microarray and *normalized* RNA-Seq data (Venet et al., 2001; Lähdesmäki et al., 2005; Wang et al., 2006; Abbas et al., 2009; Repsilber et al., 2010; Shen-Orr et al., 2010; Gong et al., 2011; Hore et al., 2016; Wang et al., 2016). Generally, given gene expression data matrix $X = (x_{ij})_{G\times S}$, where the (i, j)th element records the expression value of the ith gene in the jth sample, they decompose X by modeling x_{ij} with $\sum_{k=1}^{K} \phi_{ik}\theta_{kj}$, where ϕ_{ik} encodes the expression level of the ith gene in the kth subclone, θ_{kj} represents the mixing weight of the kth subclone in the jth sample, and K is the number of subclones. The decomposition can be solved by either optimization algorithms (Venet et al., 2001; Wang et al., 2016) or statistical inference by assuming a normal distribution on x_{ij}. While it is reasonable to assume normality for microarray gene expression data, it is often inappropriate to adopt such an assumption for directly modeling RNA-Seq data, which involve nonnegative integer observations. If a model based on normal distribution is used, one often needs to first normalize RNA-Seq data before performing any downstream analysis. See Dillies et al. (2013) for a review on normalization methods. Although normalization often destroys the nonnegative and discrete nature of the RNA-Seq data, it remains the predominant way for data preprocessing due to not only the computational convenience in modeling normalized data, but also the lack of appropriate count data models. Distinct from previously proposed methods, in this paper, we propose an attractive class of count data models in decomposing RNA-Seq count matrices.

There are, nevertheless, statistical challenges with RNA-Seq count data. First, the distributions of the RNA-Seq count data are typically over-dispersed and sparse. Second, the scales of the read counts in sequencing data across samples can be enormously different due to the mechanism of the sequencing experiment such as the variations in technical lane capacities. The larger the library sizes (*i.e.*, sequencing depths) are, the larger the read counts tend to be. In addition, the differences in gene lengths or GC-content (Pickrell et al., 2010) can bias gene differential expression analysis, particularly for lowly expressed genes (Oshlack and Wakefield, 2009). A number of count data models have been developed for RNA-Seq data (Lee et al., 2013; Kharchenko et al., 2014; Fan et al., 2016). For example, Lee et al. (2013) proposed a Poisson factor model on microRNA to reduce the dimension of count data and identify low-dimensional features, followed by a clustering procedure over tumor samples. Kharchenko et al. (2014) developed a method using a mixture of negative binomial and Poisson distributions to model single cell RNA-Seq data for gene differential expression analysis. None of these methods, however, address the problem of TH.

To this end, we propose BayCount, a Bayesian matrix decomposition model built upon the negative binomial model (Zhou, 2016), to infer tumor transcriptional heterogeneity using RNA-Seq count data. BayCount accounts for both the between-sample and gene-specific random effects and infers the number of latent subclones, the proportions of these subclones in each sample, and subclone-specific gene expression simultaneously. The R package `BayCount` implementing our model and algorithm is available at `http://pages.jh.edu/~fxie5/Research/BayCount_0.1.0.tar.gz` with the installation script at `http://pages.jh.edu/~fxie5/Research/Installation_script.R`.

The remainder of the paper is organized as follows. In Section 2, we introduce BayCount, a hierarchical Bayesian model for RNA-Seq count data, and develop an efficient compound Poisson based blocked Gibbs sampler. We investigate the performance of the posterior inference and robustness of BayCount through extensive simulation studies in Section 3, and apply BayCount to analyze two real-world RNA-Seq datasets from The Cancer Genome

Atlas (TCGA) (Cancer Genome Atlas Research Network, 2012) in Section 4. We conclude the paper in Section 5.

2 Hierarchical Bayesian Model and Inference

In this section we present the proposed hierarchical model for RNA-Seq count data, develop the corresponding posterior inference, and discuss how to determine the number of subclones.

2.1 BayCount Model

We assume that S tumor samples are available from the same or different patients. Consider a $G \times S$ count matrix $Y = (y_{ij})_{G\times S}$, where each row represents a gene, each column represents a tumor sample, and the element y_{ij} records the read count of the ith gene from the jth tumor sample. The Poisson distribution $\text{Pois}(\lambda)$ with mean $\lambda > 0$ is commonly used for modeling count data. Poisson factor analysis (PFA) (Zhou et al., 2012) factorizes the count matrix Y as $y_{ij} \sim \text{Pois}\left(\sum_{k=1}^{K} \phi_{ik}\theta_{kj}\right)$, where $\Phi = (\phi_{ik})_{G\times K} \in \mathbb{R}_+^{G\times K}$ is the factor loading matrix and $\Theta = (\theta_{kj})_{K\times S} \in \mathbb{R}_+^{K\times S}$ is the factor score matrix. Here K is an integer indicating the number of latent factors, and each column of Φ is subject to the constraint that $\sum_{i=1}^{G} \phi_{ik} = 1$ and $\phi_{ij} \geq 0$. However, the restrictive equidispersion property of the Poisson distribution that the variance and mean are the same limits the application of PFA in modeling sequencing data, which are often highly over-dispersed. For this reason, one may consider negative binomial factor analysis (NBFA) of Zhou (2016) that factorizes Y as $y_{ij} \sim \text{NB}\left(\sum_{k=1}^{K} \phi_{ik}\theta_{kj}, p_j\right)$, where $p_j \in (0, 1)$. We denote $y \sim \text{NB}(r, p)$ as a negative binomial distribution with shape parameter $r > 0$ and success probability $p \in (0, 1)$, whose mean and variance are $rp/(1-p)$ and $rp/(1-p)^2$, respectively, with the variance-to-mean ratio as $1/(1-p)$.

Denote the jth column of $\boldsymbol{Y}$ as $\boldsymbol{y}_j = (y_{1j}, y_{2j}, \ldots, y_{Gj})^T$, the count profile of the jth tumor sample. To account for both the between-sample and gene-specific random effects when modeling RNA-Seq count data, we propose

$$y_{ij} \mid \lambda, \alpha_i, \zeta_j, p_j, \Phi, \Theta \;\sim\; \mathrm{NB}\left(\lambda\alpha_i + \sum_{k=1}^{K} \phi_{ik}\theta_{kj}\zeta_j,\; p_j\right), \tag{2.1}$$

where α_i accounts for the gene-specific random effect of the ith gene, λ and p_j control the overall scale of the gene-specific effects and between-sample effect of the jth sample, respectively, and $\sum_{k=1}^{K} \phi_{ik}\theta_{kj}\zeta_j$ represents the average effect of the K subclones on the expression of the ith gene in the jth sample.

To see this, recall that the mean of y_{ij} based on (2.1) is

$$\mathbb{E}[y_{ij}] = \left(\lambda\alpha_i + \sum_{k=1}^{K} \phi_{ik}\theta_{kj}\zeta_j\right) \frac{p_j}{1-p_j}. \tag{2.2}$$

Since p_j is sample-specific, the term $p_j/(1-p_j)$ describes the effect of sample j on read counts due to technical or biological reasons (*e.g.*, different library sizes, biopsy sites, etc). We assume the relative expression of the ith gene in the kth subclone is described by ϕ_{ik}, where $\phi_{ik} \geq 0$. Since the sample-specific effect has already been captured by p_j, for modeling convenience, we normalize the gene expression so that the expression levels sum to one for each subclone. Namely, $\sum_{i=1}^{G} \phi_{ik} = 1$ for all $k = 1, \cdots, K$. Furthermore, we assume that θ_{kj} represents the proportion of the kth subclone in the jth sample, where $\theta_{kj} \geq 0$ and $\sum_{k=1}^{K} \theta_{kj} = 1$. We can interpret $\theta_{kj}\zeta_j$ as the population frequency of the kth subclone in the jth sample, where parameter ζ_j controls the scale. Together, the summation $\sum_{k=1}^{K} \phi_{ik}\theta_{kj}\zeta_j$ represents the aggregated expression level of the ith gene across all K subclones in the jth sample. To further account for the gene-specific random effects that are independent of the samples and subclones, we introduce an additional term $\lambda\alpha_i$ to describe the random effect of the ith gene on the read counts such as GC-content and gene length. We assume $\sum_{i=1}^{G} \alpha_i = 1$ so that α_i represents the relative gene-specific random effect of the ith gene with respect to all the genes and λ controls the overall scale of the gene-specific random effects.

Following Zhou (2016), the model in (2.1) has an augmented representation as

$$\begin{aligned} y_{ij} &= x_{ij} + z_{ij}, \\ x_{ij} &= \sum_{k=1}^{K} x_{ijk}, \\ z_{ij} \mid \lambda, \alpha_i, p_j &\sim \mathrm{NB}(\lambda\alpha_i, p_j), \\ x_{ijk} \mid \boldsymbol{\phi}_k, \boldsymbol{\theta}_j, \zeta_j, p_j &\sim \mathrm{NB}\left(\phi_{ik}\theta_{kj}\zeta_j, p_j\right). \end{aligned} \tag{2.3}$$

From (2.3), the raw count y_{ij} of the ith gene in the jth sample can be interpreted as coming from multiple sources: x_{ijk} represents the count of the ith gene contributed by the kth subclone in the jth sample, where $k = 1, \dots, K$, while z_{ij} is the count contributed by the gene-specific random effect of the ith gene in the jth sample. Note that for the auxiliary count matrix $(x_{ij})_{G\times S}$, we factorize the negative binomial shape parameter matrix into the product of Φ and Θ under the negative binomial likelihood. This is different from the exponential family formulation of non-negative matrix factorization in Ghahramani et al. (2014), as the negative binomial distribution $\mathrm{NB}(r, p)$ belongs to the exponential family only if the shape parameter r is fixed.

Denote $y_{\cdot j} = \sum_{i=1}^{G} y_{ij}$. Since $\sum_{i=1}^{G} \phi_{ik} = 1$ and $\sum_{k=1}^{K} \theta_{kj} = 1$ by construction, under (2.3), by the additive property of independent negative binomial random variables with the same success probability, we have

$$y_{\cdot j} \mid \lambda, \alpha_i, \zeta_j, p_j, \Phi, \Theta \sim \mathrm{NB}\left(\lambda + \zeta_j,\ p_j\right),$$

and, in particular, the mean as $\mathbb{E}[y_{\cdot j}] = (\lambda + \zeta_j)p_j/(1 - p_j)$ and the variance as $\mathrm{Var}(y_{\cdot j}) = \mathbb{E}[y_{\cdot j}] + \mathbb{E}^2[y_{\cdot j}]/(\lambda + \zeta_j)$. It is clear that p_j, the between-sample random effect of the jth sample, governs the variance-to-mean ratio of $y_{\cdot j}$, whereas $\lambda + \zeta_j$, the sum of the scale λ of the gene-specific random effects and the scale ζ_j for the jth sample, controls the quadratic relationship between $\mathrm{Var}(y_{\cdot j})$ and $\mathbb{E}[y_{\cdot j}]$.

We complete the model by setting the following priors that will be shown to be amenable to the posterior inference:

$$\begin{aligned} \boldsymbol{\phi}_k &\sim \mathrm{Dirichlet}(\eta, \cdots, \eta), & \boldsymbol{\alpha} &\sim \mathrm{Dirichlet}(\delta, \cdots, \delta), \\ \boldsymbol{\theta}_j \mid r_1, \cdots, r_K &\sim \mathrm{Dirichlet}\left(r_1, \cdots, r_K\right), & p_j &\sim \mathrm{Beta}(a_0, b_0), \\ \zeta_j \mid r_1, \cdots, r_K, c_j &\sim \mathrm{Gamma}\left(\sum_{k=1}^K r_k, c_j^{-1}\right), & \lambda &\sim \mathrm{Gamma}\left(u_0, v_0^{-1}\right), \end{aligned}$$

where $\boldsymbol{\phi}_k = (\phi_{1k}, \cdots, \phi_{Gk})^T$, $\boldsymbol{\theta}_j = (\theta_{1j}, \cdots, \theta_{Kj})^T$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_G)^T$, $\mathrm{Gamma}(a, b)$ denotes a gamma distribution with mean ab and variance ab^2, and $\mathrm{Dirichlet}(\eta_1, \cdots, \eta_d)$ denotes a d-dimensional Dirichlet distribution with parameter vector $(\eta_1, \cdots, \eta_d)$. We further impose the hyperpriors, expressed as $r_k \mid \gamma_0,\ c_0 \sim \mathrm{Gamma}\left(\gamma_0/K, c_0^{-1}\right)$, $c_j \sim \mathrm{Gamma}\left(e_0, f_0^{-1}\right)$, $\gamma_0 \sim \mathrm{Gamma}\left(g_0, h_0^{-1}\right)$, and $c_0 \sim \mathrm{Gamma}\left(e_0, f_0^{-1}\right)$ to construct a more flexible model.

Shown in Figure 1 is the graphical representation of BayCount.

2.2 Gibbs Sampling via Data Augmentation

For BayCount, while the full conditional posterior distributions of p_j, c_j and c_0 are straightforward to derive due to conjugacy, a variety of data augmentation techniques are used to derive the closed-form Gibbs sampling update equations for all the other model parameters. Rather than going into the details here, let us first assume that we have already sampled the latent counts x_{ijk} given the observations y_{ij} and model parameters, which, according to Theorem 1 of Zhou (2016), can be realized by sampling from the Dirichlet-multinomial distribution. Given x_{ijk}, we derive the Gibbs sampling update equations for Φ and Θ via data augmentation. Then we describe in Section A of the Supplementary Material a compound Poisson based blocked Gibbs sampler that completely removes the need of sampling x_{ijk}.

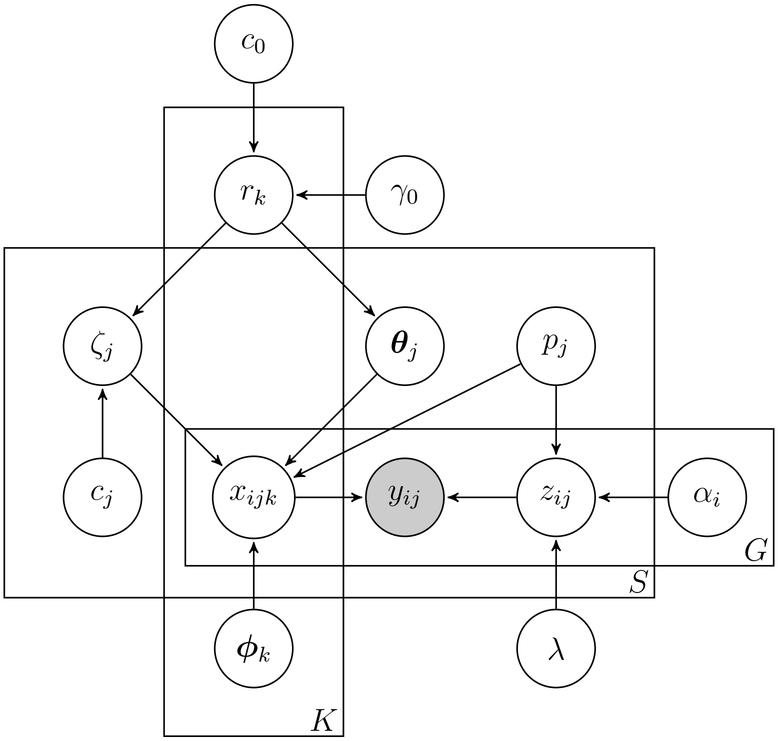


Figure 1: Graphical representation of BayCount. The boxes represent replicates. For example, the box containing r_k, x_{ijk} and $\boldsymbol{\phi}_k$, with K in its bottom right corner, indicates that there are K "copies" of r_k, x_{ijk} and $\boldsymbol{\phi}_k$ with $k = 1, \cdots, K$. Shaded nodes represent observations.

Sampling Φ and Θ

We introduce an auxiliary variable ℓ_{ijk} that follows a Chinese restaurant table (CRT) distribution, denoted by $\ell_{ijk} \mid x_{ijk}, \phi_{ik}\theta_{kj}\zeta_j \sim \mathrm{CRT}(x_{ijk}, \phi_{ik}\theta_{kj}\zeta_j)$, with probability mass function

$$p(\ell_{ijk} \mid x_{ijk}, \phi_{ik}\theta_{kj}\zeta_j) = \frac{\Gamma(\phi_{ik}\theta_{kj}\zeta_j)}{\Gamma\left(x_{ijk} + \phi_{ik}\theta_{kj}\zeta_j\right)} |s(x_{ijk}, \ell_{ijk})| \left(\phi_{ik}\theta_{kj}\zeta_j\right)^{\ell_{ijk}},$$

supported on $\{0, 1, 2, \cdots, x_{ijk}\}$, where $s(x_{ijk}, \ell_{ijk})$ are Stirling numbers of the first kind (Johnson et al., 1997). Sampling $\ell \sim \mathrm{CRT}(x, r)$ can be realized by taking the summation of m independent Bernoulli random variables: $\ell = \sum_{t=1}^{x} b_t$, where $b_t \sim \mathrm{Bernoulli}\left(r/(r + t - 1)\right)$ independently. Following Zhou and Carin (2012), the joint distribution of ℓ_{ij} and x_{ij} described by

$$\begin{aligned} \ell_{ijk} \mid x_{ijk}, \phi_{ik}, \theta_{kj}, \zeta_j &\sim \mathrm{CRT}\left(x_{ijk}, \phi_{ik}\theta_{kj}\zeta_j\right), \\ x_{ijk} \mid \phi_{ik}, \theta_{kj}, \zeta_j, p_j &\sim \mathrm{NB}\left(\phi_{ik}\theta_{kj}\zeta_j, p_j\right), \end{aligned}$$

can be equivalently characterized under the compound Poisson representation

$$\begin{aligned} x_{ijk} \mid \ell_{ijk}, p_j &\sim \mathrm{SumLog}\left(\ell_{ijk}, p_j\right), \\ \ell_{ijk} \mid \phi_{ik}, \theta_{kj}, \zeta_j, p_j &\sim \mathrm{Pois}\left(-\phi_{ik}\theta_{kj}\zeta_j \log(1-p_j)\right), \end{aligned}$$

where $x \sim \mathrm{SumLog}\,(\ell, p)$ denotes the sum-logarithmic distribution generated as $x = \sum_{t=1}^{\ell} u_t$, where $(u_t)_{t=1}^{\ell}$ are independent, and identically distributed (i.i.d.) according to the logarithmic distribution (Quenouille, 1949) with probability mass function $p(u) = -p^u/[u\log(1-p)]$, supported on $\{1, 2, \cdots\}$.

Under this augmentation, the likelihood of ϕ_{ik}, θ_{kj} and ζ_j becomes

$$\mathcal{L}(\phi_{ik}, \theta_{kj}, \zeta_j) \propto \mathrm{Pois}\left(\ell_{ijk} \mid -\phi_{ik}\theta_{kj}\zeta_j \log(1-p_j)\right),$$

where $\mathrm{Pois}(\cdot \mid \lambda)$ denotes the probability mass function of the Poisson distribution with mean λ . It follows immediately that the full conditional posterior distributions for $\boldsymbol{\phi}_k$ and $\boldsymbol{\theta}_j$ are

$$\begin{aligned} (\boldsymbol{\phi}_k \mid -) &\sim \mathrm{Dirichlet}\left(\eta + \sum_{j=1}^{S} \ell_{1jk}, \cdots, \eta + \sum_{j=1}^{S} \ell_{Gjk}\right), \\ (\boldsymbol{\theta}_j \mid -) &\sim \mathrm{Dirichlet}\left(r_1 + \sum_{i=1}^{G} \ell_{ij1}, \cdots, r_K + \sum_{i=1}^{G} \ell_{ijK}\right). \end{aligned}$$

Using data augmentation, we can similarly derive the full conditional posterior distributions for ζ_j, $\boldsymbol{\alpha}$, r_k and γ_0, as described in detail in Section A of the Supplementary Material.

2.3 Determining the Number of Subclones K

We have so far assumed *a priori* that K is fixed. Determining the number of factors in factor analysis is, in general, challenging. Zhou (2016) suggested adaptively truncating K during Gibbs sampling iterations. This adaptive truncation procedure, which is designed to fit the data well, may tend to choose a large number of factors, some of which may be highly correlated to each other and hence are potentially redundant. To facilitate the interpretation of the model output, we seek a model selection procedure that estimates K in a more conservative manner. It is critical to select a moderate K that is large enough to fit the data reasonably well, but at the same time is small enough for the sake of interpretation. As is suggested by Shen and Huang (2008) and Ghahramani et al. (2014), one way of determining K is cross-validation. This method is computationally expensive since it requires repeated leave-out testing procedure for each fixed K.

Alternatively, one can generalize the idea of "finding the elbow of scree plots", an ad-hoc method for selecting the latent dimension in principal component analysis (Zhu and Ghodsi, 2006). In the scree plots, the reductions of the residual sum of squares, a measurement of goodness-of-fit to the data, are plotted against the latent dimension. An "elbow" is the point that maximizes the difference of the slopes of the two adjacent line segments. Generalizing to BayCount, we calculate the estimated log-likelihood of the model under different numbers of subclones (using post-burn-in MCMC samples) as the measurement of the goodness-of-fit to the data. These samples are obtained by running the compound Poisson based blocked Gibbs sampler for different K's. The estimate of K is the point at which an apparent decrease in the slopes of segments that connect the log-likelihood $\log\mathcal{L}(K)$ evaluated at two consecutive K values is detected. Formally, we denote the log-likelihood function $\log\mathcal{L}(K)$ as a function of K, and define the second-order finite difference $\Delta^2\log\mathcal{L}(K)$ of the log-likelihood function by $\Delta^2\log\mathcal{L}(K) := 2\log\mathcal{L}(K) - \log\mathcal{L}(K-1) - \log\mathcal{L}(K+1)$, for $K = K_{\min}+1, \cdots, K_{\max}-1$, where $K_{\min}$ and $K_{\max}$ are the lower and upper limits of K, respectively. Then an estimate of K is given by $\widehat{K} = \arg\max_K \Delta^2 \log \mathcal{L}(K)$. Notice that similar approaches are adopted to detect the number of latent factors in the context of time series of inhomogeneous Poisson processes (Shen and Huang, 2008) and Poisson factor models (Lee et al., 2013).

3 Simulation Study

In this section, we evaluate BayCount through simulation studies. Two different scenarios are considered.

- **Scenario I.** We simulate the data according to BayCount itself in (2.1). In particular, we generate the subclone-specific gene expression data matrix $\Phi = (\phi_{ik})_{G\times K^o} \in \mathbb{R}_+^{G\times K^o}$ by i.i.d. draws of $\boldsymbol{\phi}_k \sim \text{Dirichlet}(0.05, \cdots, 0.05)$, the proportion matrix $\Theta = (\theta_{kj})_{K^o\times S}$ by i.i.d. draws of $\boldsymbol{\theta}_j \sim \text{Dirichlet}(0.5, \cdots, 0.5)$, and ζ_j by i.i.d. draws of $\zeta_j \sim \text{Gamma}(0.5K^o, 1)$, where $i = 1, \cdots, G$, $j = 1, \cdots, S$, and $k = 1, \cdots, K^o$. Here G is the number of genes, S is the number of samples, and K^o is the simulated number of subclones. We set $\lambda = 1$, draw $\boldsymbol{\alpha}$ from $\text{Dirichlet}(0.5, \cdots, 0.5)$, and generate p_j from a uniform distribution such that the variance-to-mean ratio $p_j/(1-p_j)$ of $y_{\cdot j}$ ranges from 100 to 10^6, encouraging the simulated data to be over-dispersed.
- **Scenario II.** To evaluate the robustness of BayCount, we simulate the data from a model that is different from BayCount. We generate the subclone-specific gene expression data matrix $W = (w_{ik})_{G\times K^o} \in \mathbb{R}_+^{G\times K^o}$ by i.i.d. draws of $w_{ik} \sim \text{Gamma}(0.05, 10)$, and the proportion matrix $\Theta = (\theta_{kj})_{K^o\times S}$ by i.i.d. draws of $\boldsymbol{\theta}_j \sim \text{Dirichlet}(0.5, \cdots, 0.5)$. We set $\lambda = 1$, draw $\boldsymbol{\alpha}$ from $\text{Dirichlet}(0.5, \cdots, 0.5)$, and generate p_j from a uniform distribution such that the variance-to-mean ratio $p_j/(1-p_j)$ of $y_{\cdot j}$ ranges from 100 to 10^6. The count matrix $Y = (y_{ij})_{G\times S}$ is generated from $y_{ij} \sim \text{NB}\left(\lambda\alpha_i + \sum_{k=1}^{K^o} w_{ik}\theta_{kj}, p_j\right)$. Note that in scenario II the scales of $W = (w_{ik})_{G\times K^o}$ are not subject to the constraint

$\sum_{i=1}^{G} w_{ik} = 1$.

We will show that BayCount can accurately recover both the subclone-specific gene expression patterns and subclonal proportions. The hyperparameters are set to be $\eta = 0.1$, $a_0 = b_0 = 0.01$, $e_0 = f_0 = 1$, $g_0 = h_0 = 1$, and $u_0 = v_0 = 100$. We consider $K \in \{2, 3, \cdots, 10\}$. The compound Poisson based blocked Gibbs sampler is implemented with an initial burn-in of $B = 1000$ iterations, followed by $n = 1000$ post-burn-in iterations. Notice that for any permutation matrix $\Pi \in \{0,1\}^{K\times K}$, $\Phi\Pi^{\mathrm{T}}\Pi\Theta = \Phi\Theta$, leading to potential label switching phenomenon during the MCMC. The following procedure is implemented in practice to address this issue.

- **Step 1:** Collect the n post-burn-in MCMC samples $\Phi^{(t)} = \left[\phi_{ik}^{(t)}\right]_{G\times K}$, $\Theta^{(t)} = \left[\theta_{kj}^{(t)}\right]_{K\times S}$, $\left[\alpha_i^{(t)}\right]_{i=1}^{G}$, $\left[p_j^{(t)}\right]_{j=1}^{S}$, $\lambda^{(t)}$, and $\left[\zeta_j^{(t)}\right]_{j=1}^{S}$, where $t = 1, \cdots, n$.
- **Step 2:** Find the posterior MCMC sample that maximizes the log-likelihood: $t^\star = \arg\max_{t\in\{1,\cdots,n\}} \sum_{i=1}^{G}\sum_{j=1}^{S} \log p\left(y_{ij}\middle|\lambda^{(t)}\alpha_i^{(t)} + \sum_{k=1}^{K}\phi_{ik}^{(t)}\theta_{kj}^{(t)}, p_j^{(t)}\right)$.
- **Step 3:** For $t = 1, 2, \cdots, n$, find $\Pi^{(t)} = \arg\min_{\Pi} \left\|\Theta^{(t^\star)} - \Pi\Theta^{(t)}\right\|_F^2$, where the arg min is taken over all $K \times K$ permutation matrices and $\|\cdot\|_F$ is the matrix Frobenius norm.
- **Step 4:** For $t = 1, 2, \cdots, n$, replace $\Phi^{(t)}$ by $\Phi^{(t)}\Pi^{(t)\mathrm{T}}$ and $\Theta^{(t)}$ by $\Pi^{(t)}\Theta^{(t)}$.

After implementing the procedure above for the posterior samples of Φ and Θ, we compute the posterior means and 95% credible intervals for all parameters using the post-burn-in MCMC samples.

3.1 Synthetic data with $K^o = 3$

We first simulate two datasets with $G = 100$, $S = 20$, and $K^o = 3$ under both scenario I and scenario II. Under scenario I, the data generation scheme is the same as BayCount. Figure S1 in the Supplementary Material plots $\Delta^2 \log \mathcal{L}(K)$ versus K, indicating $\widehat{K} = 3$, which is the same as the simulation truth. In terms of estimating K, we also compare BayCount with three alternative competitors: Bayesian information criterion (BIC), deviance information criterion (DIC), and the logarithmic conditional predictive ordinate (log-CPO). See Section C of the Supplementary Material for the detailed results and comparisons of estimating K. The estimated subclone-specific gene expression matrix $\widehat{\Phi}$ and subclonal proportions $\widehat{\Theta}$ are computed as the posterior means of the post-burn-in MCMC samples. Figure S2 and S3 compare the simulated true Φ and Θ with the estimate $\widehat{\Phi}$ and $\widehat{\Theta}$, respectively. We can see that both the subclone-specific gene expression patterns and the subclonal proportions are successfully recovered.

A competitive alternative for RNA-seq decomposition is the non-negative matrix factorization (NMF) (Lee and Seung, 1999) to the normalized expression data. The normalized expression data are obtained by taking the Anscombe transformation (Anscombe, 1948) to the original count matrix: $y_{ij} \mapsto \operatorname{arcsinh}\left(\sqrt{\frac{y_{ij}+c}{r-2c}}\right)$ for some constants r and c, $i = 1, \cdots, G$, $j = 1, \cdots, S$. The detailed results and comparison between BayCount and the NMF on the normalized expression data are provided in Section E of the Supplementary Material. As shown in Figures S20-22, BayCount outperforms the NMF in terms of estimating the number of subclones, the subclonal expression, and the subclonal proportions.

The analysis under scenario II is of greater interest, since the focus is to evaluate the robustness of BayCount. BayCount yields an estimate of $\widehat{K} = 3$, as shown in Figure S4. We then focus on the posterior inference based on $\widehat{K} = 3$. Figure 2 compares the estimated subclonal proportions $\widehat{\Theta}$ with the simulated true subclonal proportions across samples, along with the posterior 95% credible intervals. The results show that the estimate $\widehat{\Theta}$ approximates the simulated true Θ well. We then report the posterior inference on the subclone-specific gene expression Φ. Notice that under BayCount , $\sum_{i=1}^{G} \phi_{ik} = 1$, and hence the estimate $\widehat{\Phi}$ by BayCount and the unnormalized gene expression profile matrix W used in generating the simulated data are not directly comparable. To see whether the gene expression patterns are recovered, we first normalize W by its column sums as $\widehat{W} = W\Lambda^{-1}$, where

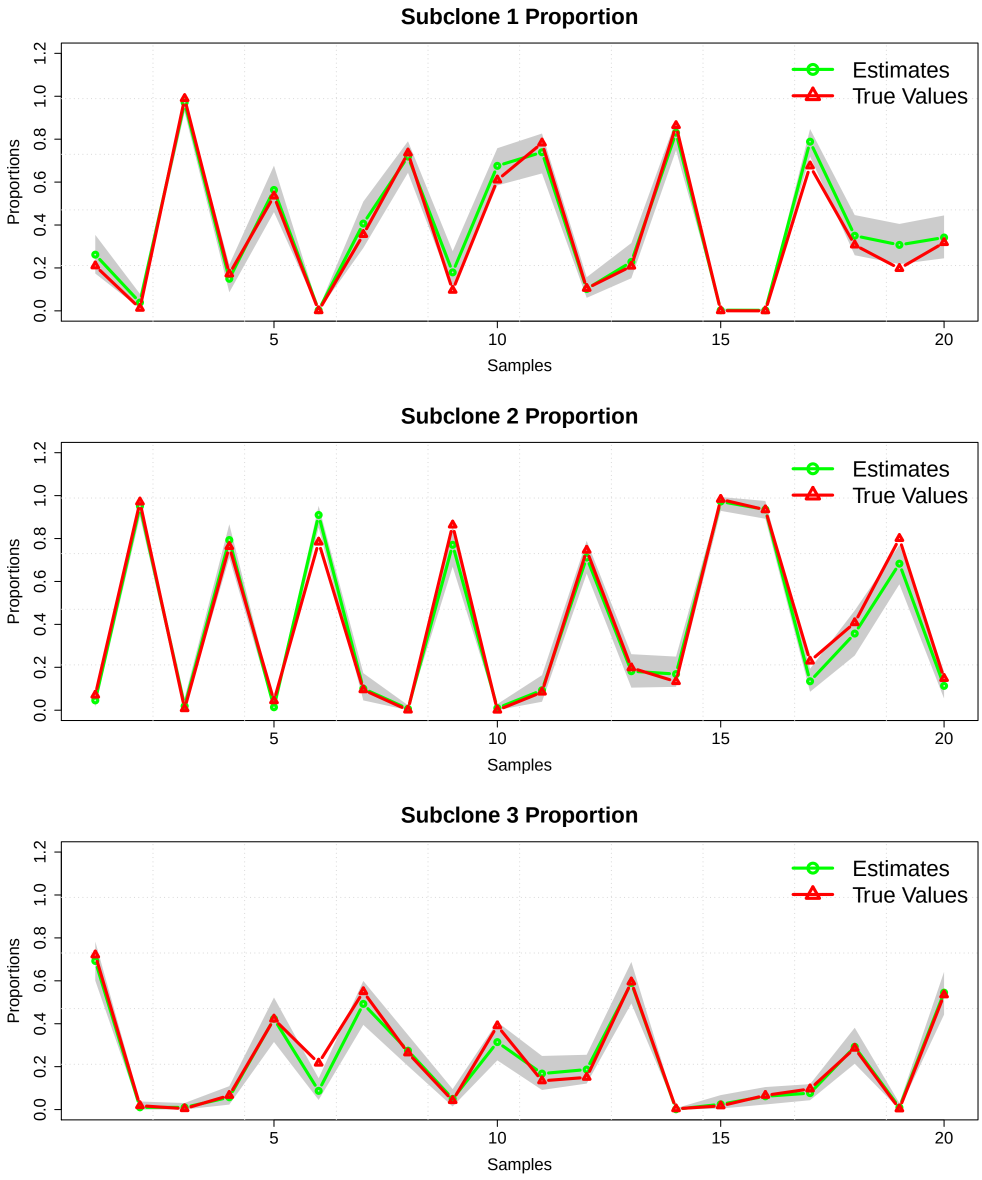


Figure 2: The estimated subclonal proportions $\widehat{\Theta}$ across samples $j = 1, \cdots, 20$ for the synthetic dataset with $K^o = 3$ under scenario II. Horizontal axis is the index $j = 1, \cdots, 20$ of tumor samples, and vertical axis is the proportion. The green lines represent the estimate $\widehat{\Theta}$, and the red lines represent the simulated true subclonal proportions. The shaded area represents the posterior 95% credible bands.

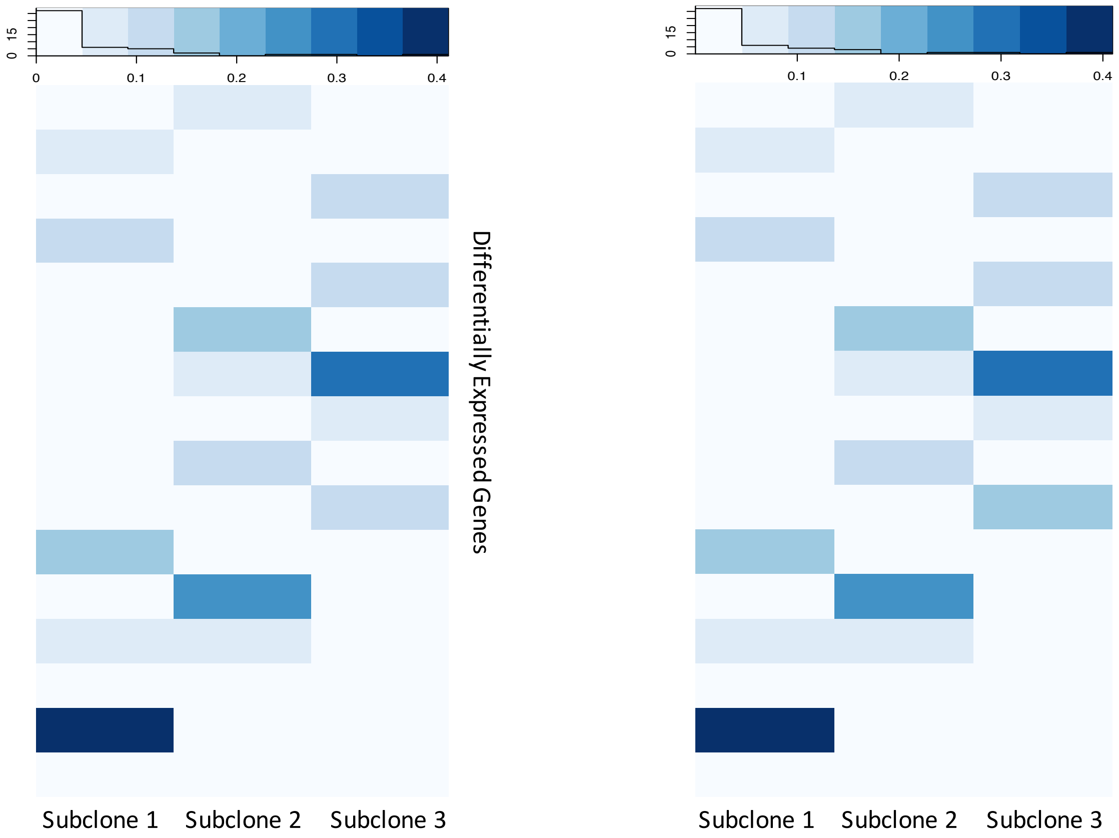
$\Lambda = \text{diag}\left(\sum_{i=1}^{G} w_{i1}, \cdots, \sum_{i=1}^{G} w_{iK}\right)$, so that $\widehat{w}_{ik}$ represents the relative expression level of the ith gene in the kth subclone, and then compare $\widehat{\Phi}$ with $\widehat{W}$. For visualization, the genes with small standard deviations (less than 0.01) are filtered out due to their indistinguishable expressions across different subclones. Figure 3 compares the heatmap of $\widehat{\Phi}$, with the heatmap of the simulated true (normalized) subclone-specific gene expression $\widehat{W}$ on selected differentially expressed genes. It is clear that the patterns of subclone-specific gene expression estimated under BayCount closely match the simulation truth. We also evaluate the stability of BayCount by adding independent Pois(10) noisy counts to the original count matrix as perturbations and then analyze the perturbed count matrix using BayCount. Figures S18 and S19 in Section D.2 of the Supplementary Material indicate that BayCount is stable in the presence of noisy perturbations.

3.2 Synthetic data with $K^o = 5$

Similarly as in Section 3.1, we simulate two datasets with $G = 1000$, $S = 40$, and $K^o = 5$ under scenarios I and II, respectively. Under scenario I, BayCount yields an estimate of $\widehat{K} = 5$ (Figure S5), and from Figures S6 and S7, both the subclone-specific gene expression patterns and the subclonal proportions are successfully captured.

Under scenario II, BayCount yields an estimate of $\widehat{K} = 5$ (Figure S8). For the subclonal proportions $\Theta = (\theta_{kj})_{K\times S}$, Figure 4 shows that the estimate $\widehat{\Theta}$ successfully recovers the simulated true proportions. Notice that the credible bands are narrower than those in Figure 2, implying relatively smaller uncertainty in estimating subclonal proportions for larger dataset. Figure S9 presents the autocorrelation plots of the posterior samples of some randomly selected θ_{kj}'s generated by the compound Poisson based blocked Gibbs sampler, indicating that the Markov chains mix well. The Markov chains also mix well when the sample size S varies over $\{40, 80, 120, 200\}$, where the synthetic dataset is simulated with $G = 1000$ and $K = 5$ under scenario II. See Section F Figure S25 in the Supplementary Material for the trace plots of some randomly selected θ_{kj}'s when S varies.

Figure 5 compares the simulated true (normalized) subclone-specific gene expression $\widehat{W}$ with the estimate $\widehat{\Theta}$ under BayCount. For this dataset we pre-screen $\widehat{W}$ with a threshold



(a) $\widehat{W}$

(b) $\widehat{\Phi}$

Figure 3: Comparison of subclone-specific gene expression patterns for the synthetic dataset with $K^o = 3$ under scenario II. Panel (a) is the heatmap of $\widehat{W}$, computed by normalizing the simulated true expression data W by its column sums, and panel (b) is the heatmap of the estimate $\widehat{\Phi}$.

0.008 on the across-subclone standard deviations for all genes for visualization. The high concordance between the heatmaps of the estimated and true expression patterns of the differentially expressed genes indicates that the subclone-specific gene expression patterns have been successfully recovered as well.

In summary, BayCount can accurately identify the number of subclones, estimate the subclonal proportions in each sample, and recover the subclone-specific gene expression patterns of the differentially expressed genes.

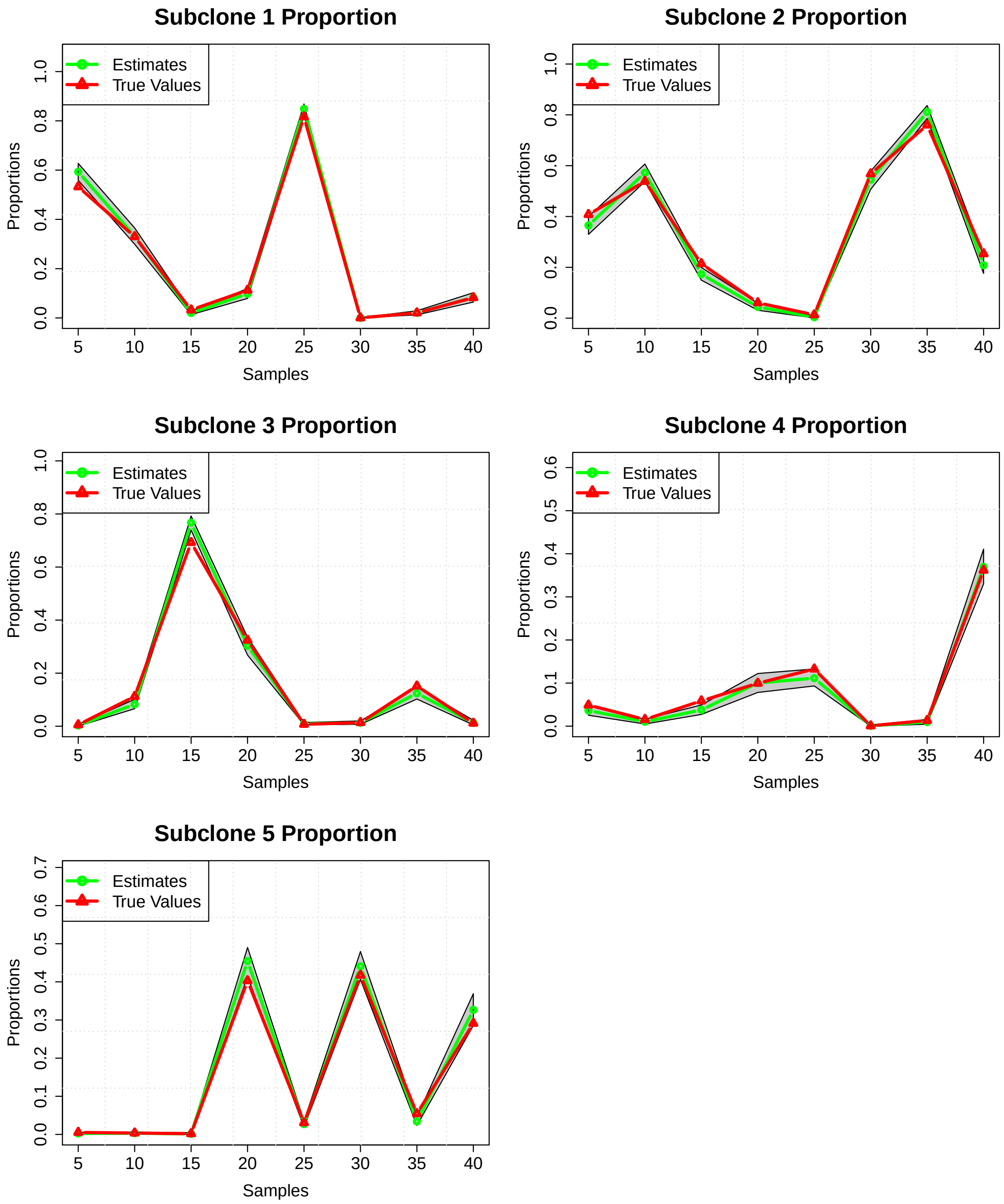
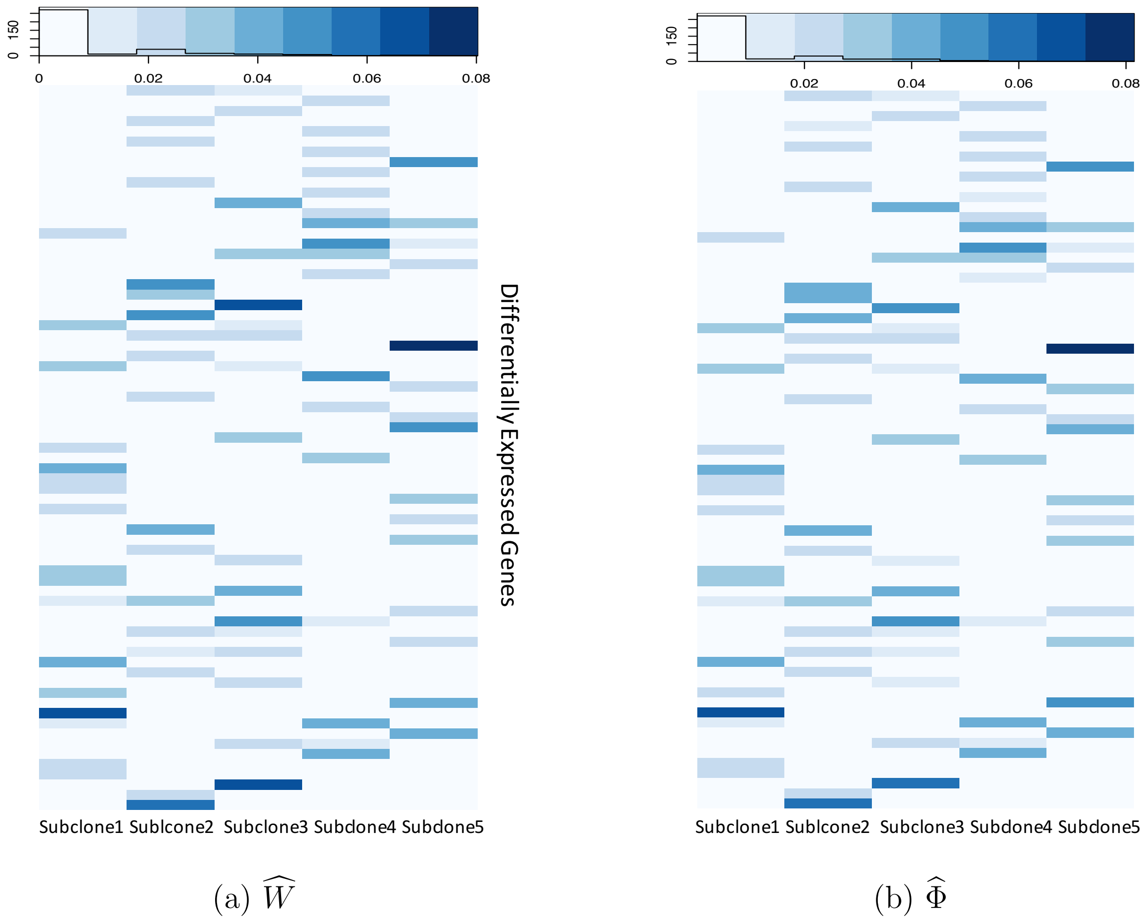


Figure 4: Subclonal proportions across samples $j = 5, 10, \cdots, 35, 40$ for the synthetic dataset with $K^o = 5$ under scenario II. Horizontal axis is the index of tumor samples, and vertical axis is the proportion. The green lines represent $\widehat{\Theta}$, and red lines represent the simulated true subclonal proportions. The shaded area represents the posterior 95% credible bands.



(a) $\widehat{W}$

(b) $\widehat{\Phi}$

Figure 5: Comparison of subclone-specific gene expression patterns for the synthetic dataset with $K^o = 5$ under scenario II. Panel (a) is the heatmap of $\widehat{W}$, computed by normalizing the simulated true expression data W by its column sums, and panel (b) is the heatmap of the estimate $\widehat{\Phi}$.

4 Real-world Data Analysis

We implement and evaluate BayCount on the RNA-Seq data from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, 2012) to study tumor heterogeneity (TH) in both lung squamous cell carcinoma (LUSC) and kidney renal clear cell carcinoma (KIRC). We first run the proposed Gibbs sampler for each fixed $K \in \{2, 3, \cdots, 10\}$, compute both the posterior mean $\log \mathcal{L}(K)$ of the log-likelihood for each fixed K, and estimate K by maximizing

$\Delta^2 \log \mathcal{L}(K)$ over K. Next, based on the estimate $\widehat{K}$ and the posterior samples generated by the proposed Gibbs sampler, we estimate the proportions of the identified subclones in each tumor sample and the subclone-specific gene expression, which in turn can be used for a variety of downstream analyses.

4.1 TCGA LUSC Data Analysis

We apply BayCount to the TCGA RNA-Seq data in lung squamous cell carcinoma (LUSC), which is a common type of lung cancer that causes nearly one million deaths worldwide every year. We downloaded FASTQ formatted files for LUSC tumor samples via the National Cancer Institute's Cancer Genomics Hub (Wilks et al., 2014) and then used the `featureCounts` function in the `Rsubread` package (Liao et al., 2014; Rahman et al., 2015) to obtain integer-based, gene-level read counts. We select 200 primary tumor samples and 382 previously reported important lung cancer genes (Wilkerson et al., 2010; Cancer Genome Atlas Research Network, 2012) for analysis of LUSC, such as KRAS, STK11, BRAF, and RIT1.

BayCount yields an estimate of five subclones (Figure S10) and their proportions in each tumor sample are shown in Figure 6. To identify the dominant subclone for each sample, we compare the estimate $\widehat{\Theta}$ of the five subclones in each tumor sample, and use them to cluster the patients. Formally, for each patient $j = 1, \cdots, S$, we compute the dominant subclone $k_j = \arg\max_{k=1,\cdots,K} \widehat{\theta}_{kj}$, and then cluster patients according to $\{j : k_j = k\}$, $k = 1, \ldots, \widehat{K}$. That is to say, the patients with the same dominant subclone belong to the same cluster. We next check if the identified subclones have any clinical utility, *e.g.*, stratification of patients in terms of overall survival. Figure 7a shows the Kaplan-Meier plots of the overall survival of the patients among the five clusters identified by their dominant subclones. Indeed, patients stratified by these five BayCount-identified groups exhibit very distinct survival patterns (log-rank test p value $= 0.0194$).

For comparison, we implement the NMF to the normalized expression data after Anscombe transformation. The NMF yields an estimate of $\widehat{K} = 2$ (Figure S23), and patients stratified by the two NMF-identified groups do not exhibit distinct survival patterns (log-rank test p value $= 0.125$, see Figure S24a). The detailed results and comparisons are provided in Section E of the Supplementary Material.

Figure 7b shows the expression levels of the top 30 differentially expressed genes (ranked by the standard deviations of the subclone-specific gene expression levels ϕ_{ik}'s in an increasing order) in these five subclones. Distinct expression patterns are observed among different subclones. For example, the FTL level is elevated in subclone 1; the expression levels of several genes encoding keratins (KRT5, KRT6A, etc.) are elevated in subclone 3; and the COL1A1 and COL1A2 expression levels are elevated in subclone 4. Interestingly, the patients with these dominant subclones also show the expected survival patterns. The subclone-1 dominated patients have better overall survival. Previous studies show that the expression of FTL is decreased in lung tumors compared to normal tissues (Kudriavtseva et al., 2009), and one plausible explanation is that subclone 1 may descend from less malignant cells and therefore resemble (or consist of) normal cells. Keratins and collagen I (encoded by COL1A1 and COL1A2) are known to play key roles in epithelial-to-mesenchymal transition (EMT), which subsequently initiates metastasis and promotes tumor progression (DePianto et al., 2010; Karantza, 2011; Shintani et al., 2008). This agrees with our observation of worse prognosis in patients who have either subclone 3 (with elevated Keratin-coding genes) or subclone 4 (with elevated collagen I coding genes) as their dominant subclone.

With the inferred 5 subclones and the corresponding parameters of the LUSC dataset under BayCount, we perform an additional simulation study that is realistic: we simulate a synthetic dataset using the parameters inferred from the LUSC dataset with $G = 382$, $S = 200$, and ground true $K = 5$. Figures S15-17 show that BayCount successfully recovers the ground true K, the subclonal expression patterns, and the subclonal proportions. See Section D.1 of the Supplementary Material for the detailed results.

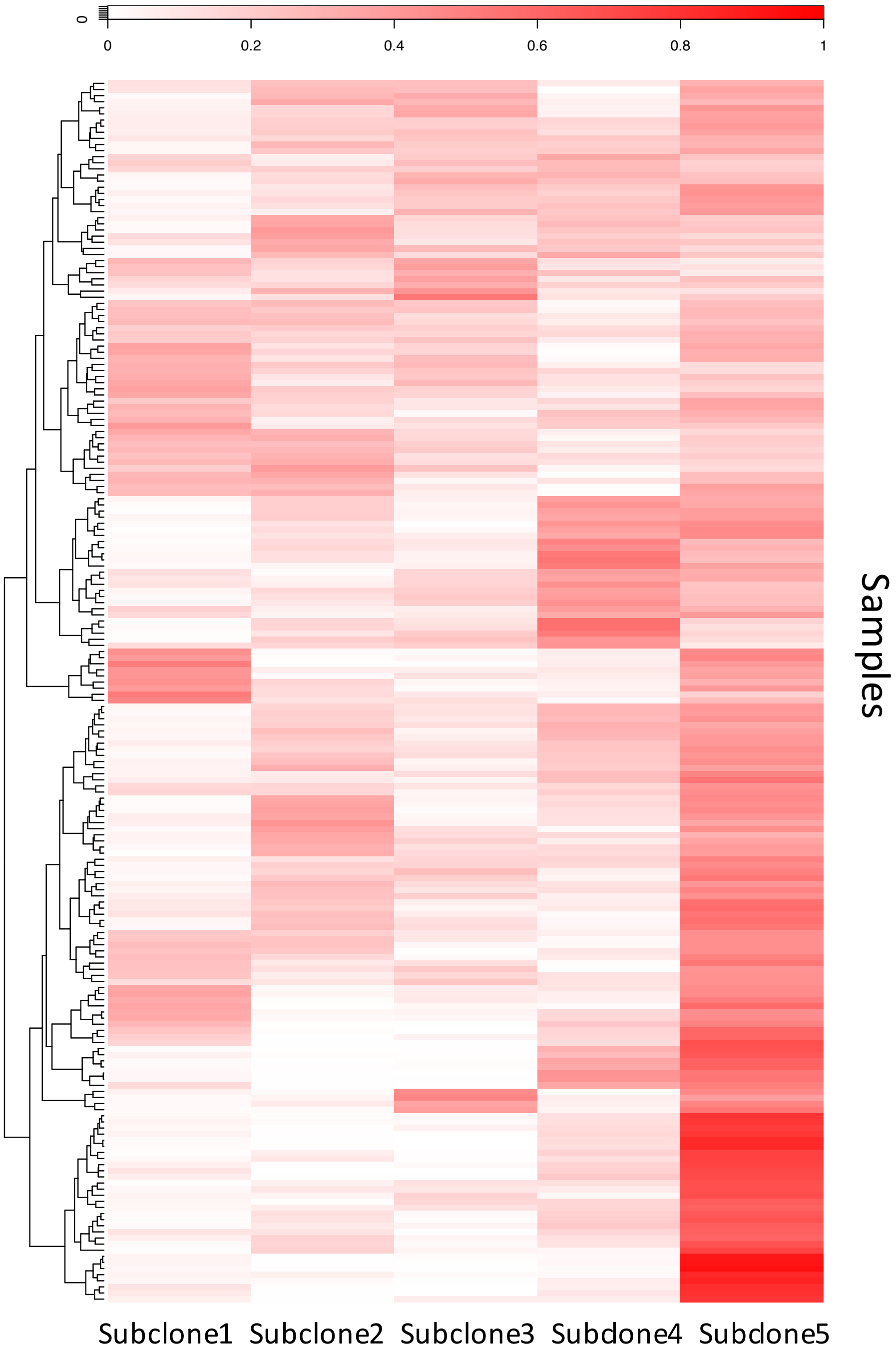
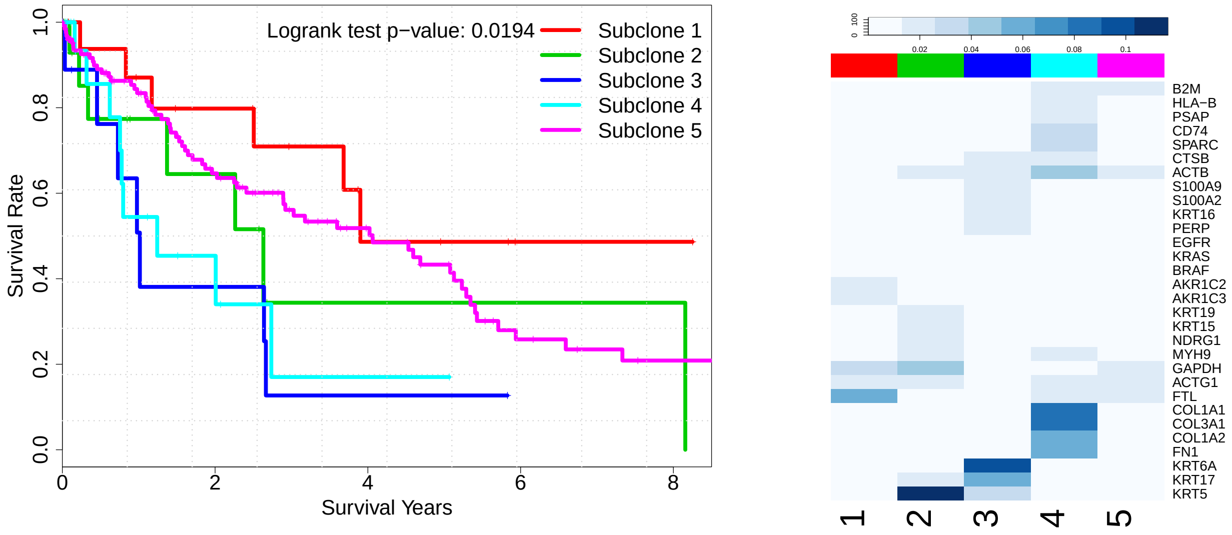


Figure 6: Heatmap of the subclonal proportions across LUSC tumor samples $j = 1, \cdots, 200$. From the heatmap it is clear that subclone 5 occupies relatively larger proportions for a large number of patients than the other 4 subclones.



(a)

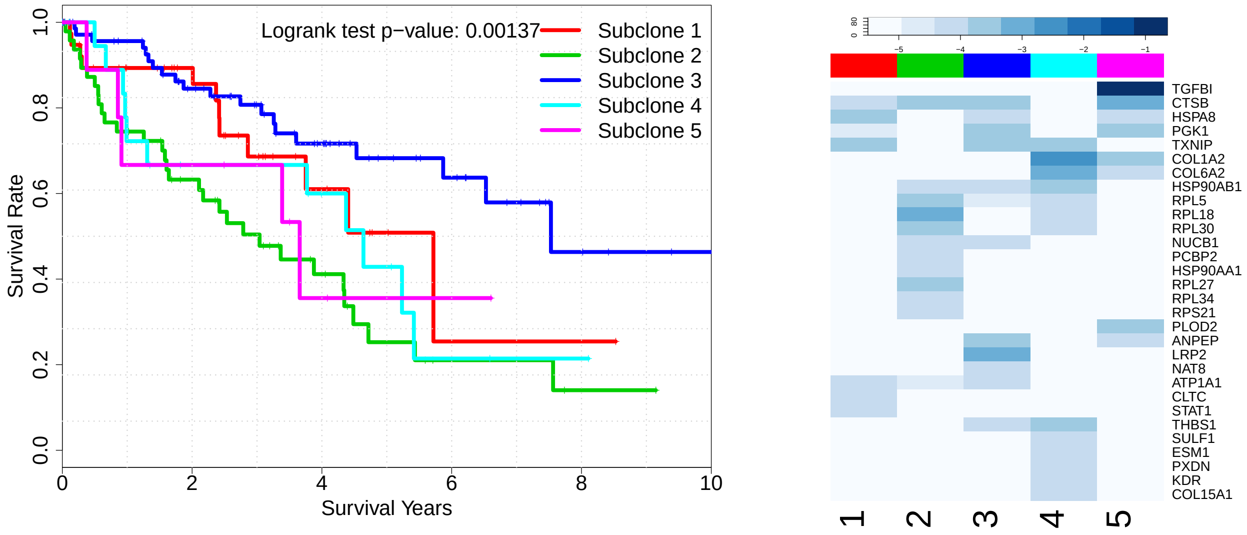
(b)

Figure 7: Panel (a) shows the Kaplan-Meier plots of overall survival in the LUSC dataset, where the patients are stratified by five clusters identified by subclone domination under BayCount. Panel (b) shows the subclone-specific gene expression of the top differentially expressed genes among five subclones.

4.2 Kidney Cancer (KIRC) Data Analysis

Similarly, we obtain gene level read counts (Liao et al., 2014) for 200 TCGA kidney renal clear cell carcinoma (KIRC) tumor RNA-seq samples and analyze them under BayCount. Among a total of 23,368 genes, 966 significantly mutated genes (Cancer Genome Atlas Research Network, 2013) in KIRC patients are selected, including VHL, PTEN, MTOR, etc.

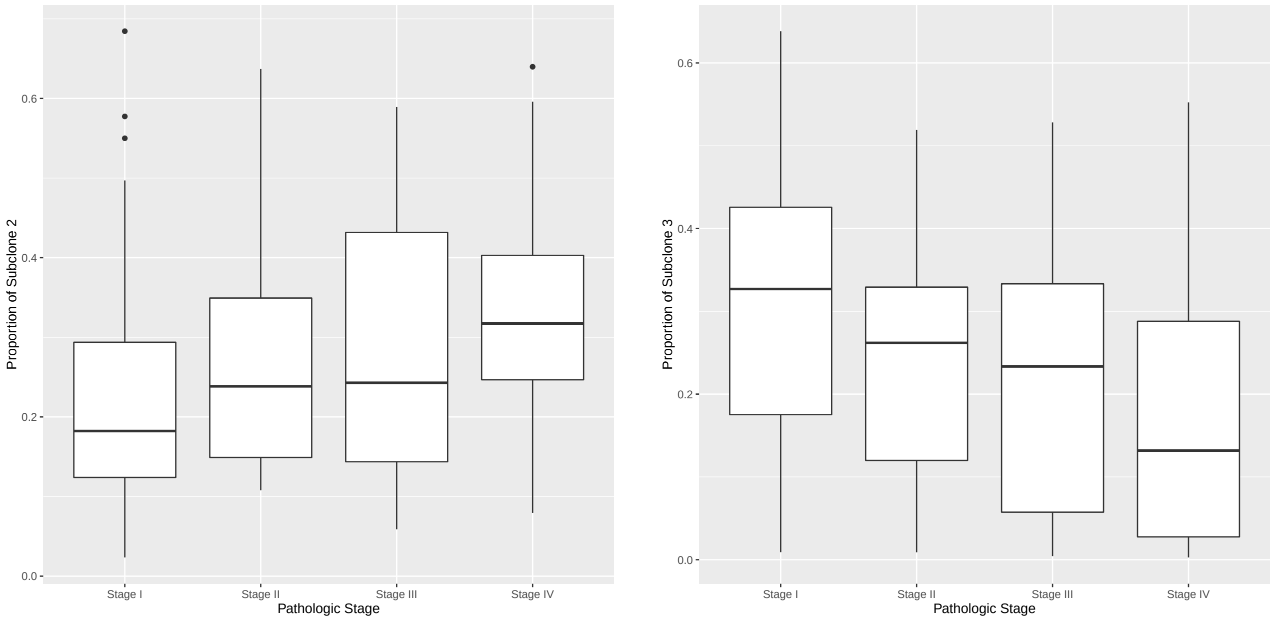
BayCount yields an estimate of five subclones in KIRC (Figure S11). Figure 8 shows the Kaplan-Meier plots of the overall survival of the patients grouped by their dominant subclones (panel a) and the heatmap of the subclone-specific gene expression corresponding to the top 30 differentially expressed genes (panel b). Since we have a large number of genes to begin with, whereas $\sum_{i=1}^{G} \phi_{ik} = 1$ for all $k = 1, \cdots, K$, the subclone-specific gene expression estimates $\widehat{\Phi}$ will be small. For better visualization, we plot $\widehat{\Phi}$ in the logarithmic scale. The subclonal proportions across 200 KIRC tumor samples are shown in Figure S12. As shown in Figure 8, the patients with these dominant subclones again show distinct survival patterns. One of the poor survival groups (dominated by subclone 5) is characterized by elevated expression of TGFBI, which is known to be associated with poor prognosis (Zhu et al., 2015) and matches our observation here.



(a) (b)

Figure 8: Panel (a) shows the Kaplan-Meier plots of overall survival in the KIRC dataset, where the patients are stratified by five clusters identified by subclone domination under BayCount. Panel (b) shows the subclone-specific gene expression (in the logarithmic scale) of the top differentially expressed genes among the five inferred subclones.

One distinction of our method from conventional subgroup analysis methods is that we focus on characterizing the underlying subclones (*i.e.*, biologically meaningful subpopulations), by not only their individual molecular profiles but also their proportions. Instead of grouping the patients by their dominant subclones, we examine the proportion itself in terms of clinical utility. Interestingly, as shown in Figure 9a, the proportion of subclone 2 increases with tumor stage: *i.e.*, as subclone 2 expands and eventually outgrows other subclones, the tumor becomes more aggressive. In contrast, the proportion of subclone 3 decreases with tumor stage (Figure 9b). Subclone 3 might be characterized by the less malignant (or normal-like) cells and takes more proportion in the beginning of the tumor life cycle. As tumor progresses to more advanced stages, subclone 3 could be suppressed by more aggressive subclones (*e.g.*, subclone 2) and takes a decreasing proportion. Unsurprisingly, the survival patterns agree with our speculations about subclones 2 and 3, with the patients dominated by subclone 2 (the more aggressive subclone) and subcolone 3 (the less aggressive subclone) showing the worst and best survivals, respectively.



(a) Subclone 2

(b) Subclone 3

Figure 9: Panel (a): the proportions of subclone 2 in each tumor sample versus their pathologic stages (p-value = 0.00173). Panel (b): the proportions of subclone 3 in each tumor sample versus their pathologic stages (p-value = 0.00299).

More excitingly, we find that the proportions of these two subclones can complement clinical variables in further stratifying patients. For patients at early stage where the event rate is low and clinical information is relatively limited, the proportions of subclones 2 and 3 serve as a potent factor in further stratifying patients (Figure S13) when dichotomizing at a natural cutoff. Combining our observations above, subclone proportions may provide additional insights into the progression course of tumors, assistance in biological interpretation, and potentially more accurate clinical prognosis.

5 Conclusion

The emerging high-throughput sequencing technology provides us with massive information for understanding tumors' complex microenvironment and allows us to develop novel statistical models for inferring tumor heterogeneity. Instead of normalizing RNA-Seq data that may bias downstream analysis, we propose BayCount to directly analyze the raw RNA-Seq count data. Overcoming the natural challenges of analyzing raw RNA-seq count data, BayCount is able to factorize them while adjusting for both the between-sample and gene-specific random effects. Simulation studies show that BayCount can accurately recover the subclonal inference used to generate the simulated data. We apply BayCount to the TCGA LUSC and KIRC datasets, followed by correlating the subclonal inferences with their clinical utilities for comparison. In particular, by grouping patients according to their dominant subclones, we observe distinct and biologically sensible overall survival patterns for both LUSC and KIRC patients. Moreover, the proportions of the subclones may complement clinical variables in further stratifying patients. In addition to prognosis value, tumor heterogeneity may be used as a biomarker to predict treatment response. For example, tumor samples with large proportions of cells bearing higher expressions on clinically actionable genes should be treated differently from those that have no or a small proportion of such cells. In addition, metastatic or recurrent tumors may possess very different compositions of subclones and should be treated differently.

BayCount provides a general framework for inference on latent structures arising naturally in many other biomedical applications involving count data. For example, analyzing single-cell data is a potential further application of BayCount due to their sparsity and overdispersion nature. Macosko et al. (2015) describe Drop-Seq, a technology for profiling more than 40,000 single cells at one time. The unique characteristic of dropped-out events (Fan et al., 2016) in single cell sequencing limits the applicability of normalization methods in bulk RNA-Seq data. Also, such huge amount number of single-cells and high levels of sparsity pose difficulties for dimensionality reduction methods such as principal component analysis. Inferring distinct cell populations in single-cell RNA count data will be an interesting extension of BayCount.

Acknowledgement

Yanxun Xu's research is partly supported by Johns Hopkins *in*Health and Booz Allen Hamilton.

References

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLOS ONE*, 4(7):e6098.

Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254.

Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519.

Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421.

DePianto, D., Kerns, M. L., Dlugosz, A. A., and Coulombe, P. A. (2010). Keratin 17 promotes epithelial proliferation and tumor growth by polarizing the immune response in skin. *Nature Genetics*, 42(10):910–914.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):1.

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

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510.

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, 13(3):241–244.

Ghahramani, Z., Mohamed, S., and Heller, K. A. (2014). *Partial Membership and Factor Analysis.* Chapman and Hall/CRC.

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and Szustakowski, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLOS ONE*, 6(11):e27156.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, 48(9):1094–1100.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete multivariate distributions*, volume 165. Wiley New York.

Karantza, V. (2011). Keratins in health and cancer: more than mere epithelial cell markers. *Oncogene*, 30(2):127–138.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742.

Kim, K.-T., Lee, H. W., Lee, H.-O., Kim, S. C., Seo, Y. J., Chung, W., Eum, H. H., Nam, D.-H., Kim, J., Joo, K. M., et al. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol*, 16(1):127.

Kudriavtseva, A., Anedchenko, E., Oparina, N. Y., Krasnov, G., Kashkin, K., Dmitriev, A., Zborovskaya, I., Kondratjeva, T., Vinogradova, E., Zinovyeva, M., et al. (2009). Expression of FTL and FTH genes encoding ferritin subunits in lung and renal carcinomas. *Molecular Biology*, 43(6):972–981.

Lähdesmäki, H., Dunmire, V., Yli-Harja, O., Zhang, W., et al. (2005). In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, 6(1):1.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, J., Müller, P., Sengupta, S., Gulukota, K., and Ji, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(4):547–563.

Lee, S., Chugh, P. E., Shen, H., Eberle, R., and Dittmer, D. P. (2013). Poisson factor models with applications to non-normalized microRNA profiling. *Bioinformatics*, 29(9):1105–1111.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. *Cell*, 149(5):994–1007.

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol*, 14(7):R80.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(1):1.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772.

Quenouille, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, 5(2):162–164.

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., and Piccolo, S. R. (2015). Alternative preprocessing of RNA-sequencing data in the cancer genome atlas leads to improved analysis results. *Bioinformatics*, 31(22):3666–3672.

Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J., Parida, S. K., Kaufmann, S. H., and Jacobsen, M. (2010). Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinformatics*, 11(1):1.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398.

Russnes, H. G., Navin, N., Hicks, J., and Borresen-Dale, A.-L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of Clinical Investigation*, 121(10):3810.

Shen, H. and Huang, J. Z. (2008). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *The Annals of Applied Statistics*, pages 601–623.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type–specific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289.

Shintani, Y., Maeda, M., Chaika, N., Johnson, K. R., and Wheelock, M. J. (2008). Collagen I promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming growth factor–β signaling. *American Journal of Respiratory Cell and Molecular Biology*, 38(1):95–104.

Venet, D., Pecasse, F., Maenhaut, C., and Bersini, H. (2001). Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(suppl 1):S279–S287.

Wang, M., Master, S. R., and Chodosh, L. A. (2006). Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, 7(1):1.

Wang, N., Hoffman, E. P., Chen, L., Chen, L., Zhang, Z., Liu, C., Yu, G., Herrington, D. M., Clarke, R., and Wang, Y. (2016). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific Reports*, 6.

Wilkerson, M. D., Yin, X., Hoadley, K. A., Liu, Y., Hayward, M. C., Cabanski, C. R., Muldrew, K., Miller, C. R., Randell, S. H., Socinski, M. A., et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical Cancer Research*, 16(19):4864–4875.

Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., et al. (2014). The cancer genomics hub (CGHub): overcoming cancer through the power of torrential data. *Database*, 2014.

Xu, Y., Müller, P., Yuan, Y., Gulukota, K., and Ji, Y. (2015). MAD Bayes for tumor heterogeneityfeature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514.

Zhou, M. (2016). Nonparametric Bayesian negative binomial factor analysis. *arXiv preprint arXiv:1604.07464*.

Zhou, M. and Carin, L. (2012). Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems*, pages 2546–2554.

Zhou, M., Hannah, L., Dunson, D. B., and Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, volume 22, pages 1462–1471.

Zhu, J., Chen, X., Liao, Z., He, C., and Hu, X. (2015). TGFBI protein high expression predicts poor prognosis in colorectal cancer patients. *International Journal of Clinical and Experimental Pathology*, 8(1):702.

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930.