BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Latent Process Decomposition Of High-Dimensional Count Data

Sanvesh Srivastava and R.W. Doerge*

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Next-generation sequencing (NGS) technologies have become the preferred way of exploring a genome. These data are high-dimensional discrete counts with correlated variables (e.g., genes). We present a novel latent factor model for high-dimensional count data, Latent Process Decomposition (LPD-C), that accounts for the correlations among genes and models the biological hypothesis that genes work in groups (e.g., pathways), which are referred to as processes. LPD-C is a two stage unsupervised approach for grouping genes into a pre-specified number of clusters, and for selecting genes that belong to these clusters with high probability. The first stage of LPD-C uses a variational Bayes approach for efficient estimation of its parameters. The second stage of LPD-C selects genes grouped as gene-subsets using empirical Bayes hypothesis testing.

Results: The performance of LPD-C is explored using simulated and publicly available NGS data, compared with existing approaches, and shown to be a useful and extensible framework for identifying genes suitable for further exploration. Although we apply LPD-C in a genomic context, it can be used for any high-dimensional count data.

Availability: R code for fitting LPD-C is available from the authors on request.

Contact: doerge@purdue.edu

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have enabled measurements of complex biological and genomic activities at an extremely high resolution (Marioni *et al.*, 2008). These technologies are widely used for measuring expression of genes under different treatments. Gene expression data from NGS technologies are in form of discrete counts, and the number of genes is typically larger than the number of samples; therefore, these data are high-dimensional count data. Extensive literature exists for identifying differentially expressed genes and for clustering genes in NGS gene expression data. Most of these approaches assume that gene expressions are mutually independent. This assumption is relaxed by Latent Process Decomposition of high-dimensional count data (LPD-C) presented in this work. LPD-C is an unsupervised approach for NGS data analysis that models the dependence among genes through latent variables called processes that are assumed to correspond to pathways (Rogers *et al.*, 2005). It also selects genes that belong to these latent processes with high probabilities. We explore LPD-C's application in the context of NGS gene expression data and select genes that have a high probability of belonging to their respective processes.

NGS technologies are preferred for exploring a genome over conventional technologies, such as microarrays, because NGS data are highly replicable with little technical variation (Marioni *et al.*, 2008). Data analysis methods that incorporate features specific to NGS data, without any restrictive distributional assumptions, are essential for deriving reliable conclusions from these data. For example, statistical models for NGS data must account for the biological hypothesis that genes in the same pathway are more likely to be dependent on each other than those in different pathways, and that a significant amount of biological or phenotypic variation can be explained by a small fraction of genes. Restrictions such as these, reduce computational complexity of statistical algorithms, facilitate interpretation, and exploit known structures in the data.

Limited finite-sample results exist for the analysis of high-dimensional non-Gaussian data. The issues in NGS data analysis are magnified simply due to the non-Gaussian, discrete, and over-dispersed nature of the data (Marioni *et al.*, 2008). Further, interacting genes only complicate the analysis (Efron, 2010). In NGS data analysis, two major themes arise for selecting candidate genes from samples with different treatments. The first theme frames the problem as a gene-wise multiple hypotheses testing problem, with the rejected hypotheses corresponding to the candidate genes; most of these approaches assume a negative binomial model for NGS data. Because there are no finite sample equivalents of the t- or F-test statistics, these hypotheses tests rely either on asymptotic test statistics based on Gaussian or Chi-square distribution or on modified versions of Fisher's exact test based on the sampling model (Robinson and Smyth, 2007, 2008; Anders and Huber, 2010; Hardcastle and Kelly, 2010; Robinson *et al.*, 2010). Asymptotic tests are unreliable in the current small-sample setting of NGS data, and none of these tests model the dependence among genes. Young *et al.* (2012) provide an excellent overview of existing hypothesis testing based methods for NGS data analysis. The second theme proposes modeling the exchangeability of genes either using two level generative Bayesian models or using penalized likelihood approaches (Hastie *et al.*, 2009; Friedman *et al.*, 2010). The Bayesian approach uses posterior distributions and the penalized likelihood approach chooses appropriate tuning parameters to select candidate genes. Because most posterior distributions are analytically intractable, Bayesian inference uses Markov chain Monte Carlo (MCMC) for

*to whom correspondence should be addressed

© Oxford University Press 2005.

sampling from the posterior. MCMC is of limited use in high-dimensions due to its computational intractability. Addressing statistical significance in penalized likelihood approaches is still an active area of research. Witten (2011) proposes sparse Poisson Linear Discriminant Analysis (SPLDA), a penalized likelihood approach using Lasso penalty, that performs much better than many existing methods for classifying and clustering NGS data. Currently, no Bayesian approach exists for NGS gene expression data analysis that selects genes while modeling the dependence among genes.

LPD-C is a latent factor model (FM), which is widely used for modeling multivariate Gaussian data, including microarrays. Rogers *et al.* (2005) proposed Latent Process Decomposition (LPD) framework for high-dimensional Gaussian data (LPD-G) in the context of microarray data. LPD-G is a more flexible approach than classical unsupervised approaches (hierarchical or K-means clustering) to model the biological hypothesis that genes work in groups or networks. Because their main objective was to find clusters of genes in microarray data, Rogers *et al.* (2005) do not select candidate genes suitable for further exploration. The extended LPD framework, of which LPD-G and LPD-C are special cases, amends the original generative Bayesian model via a second stage that selects candidate gene-subsets. Selected genes have two properties: they are a small fraction of the total number of genes, and they are associated to their respective subsets with high probabilities. The generative model of LPD is an example of Bayesian FM, the processes in LPD correspond to factors in FM. West (2003) and Carvalho *et al.* (2008) present applications of FMs to microarray data. Their model is similar to LPD-G's modeling approach. Dunson and Herring (2005) model discrete outcomes, including count data, using a FM. An extension of their model to high-dimensional count data, which accounts for the fact that genes act in networks, is similar to LPD-C's generative Bayesian model. That said, there is a key difference between FMs and LPD. In FMs, latent factors (processes) are of main interest, whereas in LPD the focus is on estimating mean genomic effects.

Motivated by the need for a computationally efficient and unsupervised Bayesian approach for NGS data analysis that groups and selects genes into clusters, we develop the methodology and associated computations for LPD-C. When applied to NGS data, LPD-C's first stage is an unsupervised approach to model the biological hypothesis that genes work in groups, or processes, and is a special case of the mixed membership modeling framework (Airoldi *et al.*, 2005). This stage adapts and extends the variational Bayes algorithm of Latent Dirichlet Allocation algorithm (LDA) (Blei *et al.*, 2003) for computationally efficient estimation of the parameters and hyperparameters. The second stage uses the parameter estimates from the first stage to select candidate genes, organized as gene-subsets, using empirical Bayes hypothesis testing framework (Efron, 2010). The second stage has few assumptions and controls the number of false discoveries. In real data analysis, LPD-C's results agree closely with those of hypothesis testing and penalized likelihood based approaches. LPD-C's distinguishing feature is that it selects a small fraction of genes, grouped as gene-subsets, in NGS data. Being an unsupervised approach, LPD-C cannot model the effects of covariates. Specifically, it cannot be used for differential gene expression analysis; however, it can be easily modified to yield its supervised extensions.

2 METHODS

NGS gene expression data can be represented as a matrix N of gene counts with S rows and G columns that represent samples and genes, respectively. The gene counts for s-th sample are denoted as $\mathbf{n}_s$ (i.e., the s-th row of N), and n_{sg} is the count for gene g in sample s. There are K latent processes (hereafter processes) associated with each sample. Any gene in a sample can belong to one of the K processes. Due to the unsupervised nature of the analysis, we ignore any covariate information associated with the samples.

2.1 First stage of LPD-C: Hierarchical Bayesian model

Consider a three level generative Bayesian model for $\mathbf{n}_s$. The first (population) level of the sampling model generates the probability vector $\boldsymbol{\pi}_s = (\pi_{s1}, \ldots, \pi_{sK})$ of process memberships for genes in sample s from a Dirichlet distribution with parameters, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, such that

$$\boldsymbol{\pi}_s \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad \text{(Level 1)} \tag{1}$$

and $\boldsymbol{\pi}_s$ is a latent variable specific to sample s. Model (1) implies that the genes in sample s belong to K sub-populations called processes, and the probability of a gene belonging to a process depends on the sample. The processes represent the dimensions of latent dependence among genes in sample s. For LPD-C to be practically useful, K is assumed to be of the order $\log G$ (Bhattacharya and Dunson, 2011).

In the second level, the model generates the process membership k of gene g in sample s as the latent Multinomial random vector $\mathbf{z}_{sg}$ of length K, with all zeros except 1 at the k-th position

$$\mathbf{z}_{sg} \mid \boldsymbol{\pi}_s \sim \text{Multinomial}(1; \boldsymbol{\pi}_s), \text{ for } g = 1, \ldots, G, \quad \text{(Level 2)} \tag{2}$$

$$\mathbf{z}_{sg} = (z_{sg1}, \ldots, z_{sgk}),\ k \text{ is such that } z_{sgk} = 1 \text{ and } z_{sgj} = 0 \text{ for } j \neq k.$$

Z_s is a latent indicator matrix specific to sample s with $\mathbf{z}_{sg}$ as its rows. It has G rows and K columns representing genes and processes, respectively. The column with the non-zero entry in the g-th row of Z_s indicates the latent process membership of gene g; therefore, Z_s represents the latent dependence structure in sample s, and genes in the same process (or "pathway") behave similarly.

Finally, the third level generates the count n_{sg} for gene g in sample s based on its process membership k as

$$n_{sg} \mid \lambda_{gk} \sim \text{Poisson}(\lambda_{gk}), \quad \text{(Level 3)} \tag{3}$$

where λ_{gk} is an element of the gene- and process-specific mean ("loadings") matrix Λ with G rows and K columns that represent genes and processes, respectively. For ease of presentation, we assume that the loadings matrix also includes the effect of library size. The gene counts for all the samples are generated following (1) – (3). Marginalizing k's in (3) imposes a K-dimensional covariance structure among genes that depends on Λ and $\boldsymbol{\alpha}$. These parameters are estimated based on the NGS data, which makes our approach empirical Bayesian. The generative model of LPD-C adapts the sampling models of LDA (Blei *et al.*, 2003) and LPD-G (Rogers *et al.*, 2005) for NGS data. All of these models are examples of FMs that have been successfully used for analyzing high-dimensional multivariate data, including microarray data (West, 2003; Bishop, 2006; Carvalho *et al.*, 2008).

The generative model (1) – (3) makes LPD-C more flexible than classical unsupervised approaches, such as hierarchical and K-means clustering (Blei *et al.*, 2003; Rogers *et al.*, 2005). Specifically, (2) associates genes in sample s to different processes chosen from the K processes using Multinomial(1; $\boldsymbol{\pi}_s$). This level gives rise to two major advantages of LPD-C. First, (2) enables LPD-C to both model the biological hypothesis that genes work in groups (processes). Second, due to its greater flexibility than classical clustering models, LPD-C can better adapt to the latent structure of NGS data.

2.1.1 Estimation of posterior distributions of LPD-C's parameters

LPD-C selects K gene-subsets using test statistics obtained from the posterior density of Z_s's because each gene-subset corresponds to a process and Z_s's relate samples, genes, and processes. The joint density of the

latent variables $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S$ (hereafter $\boldsymbol{\pi}_{1:S}$) and $Z_1, \ldots, Z_S$ (hereafter $Z_{1:S}$) and NGS data $\mathbf{n}_1, \ldots, \mathbf{n}_S$ (hereafter $\mathbf{n}_{1:S}$) given $\boldsymbol{\alpha}$ and Λ, $p(\boldsymbol{\pi}_{1:S}, Z_{1:S}, \mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$ is analytically intractable (Blei *et al.*, 2003); therefore, the posterior density $p(Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ is also analytically intractable. There are a host of techniques that can be used to approximate $p(Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$, including MCMC.

We employ Poisson variational Bayes methods from machine learning and obtain analytically tractable variational density $q(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ that approximates analytically intractable $p(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ (Bishop, 2006). This choice is important for the computational efficiency and practical applicability of LPD-C. Assuming that latent variables $\boldsymbol{\pi}_{1:S}$ and $Z_{1:S}$ are independent under the variational posterior density, so that

$$q(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda) = \prod_{s=1}^{S} q(\boldsymbol{\pi}_s) \Big(\prod_{g=1}^{G} q(\mathbf{z}_{sg}) \Big), \qquad (4)$$

the variational approach minimizes the Kullback-Liebler (KL) divergence between $q(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$ and $p(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$. The variational posterior densities of $\boldsymbol{\pi}_s$ and Z_s are $q(\boldsymbol{\pi}_s) = q(\boldsymbol{\pi}_s \,|\, \mathbf{n}_s, \boldsymbol{\alpha}, \Lambda)$ and $q(Z_s) = q(Z_s|\, \mathbf{n}_s, \boldsymbol{\alpha}, \Lambda)$. The factorization (4) alone guarantees the analytic tractability of $q(\boldsymbol{\pi}_{1:S}, Z_{1:S}|\, \mathbf{n}_{1:S}, \boldsymbol{\alpha}, \Lambda)$, and there are no further distributional assumptions for q's. Using Section 1 of Supplementary Material, the variational approximation introduces variational parameters, $\boldsymbol{\gamma}_s = (\gamma_{s1}, \ldots, \gamma_{sK})$ and $\{\Phi_{sg} = (\phi_{sg1}, \ldots, \phi_{sgK})\}_{g=1}^{G}$, which are estimated using $\mathbf{n}_s, \boldsymbol{\alpha}$, and Λ, so that

$$\begin{aligned} q(\boldsymbol{\pi}_s \,|\, \boldsymbol{\gamma}_s) &= \text{Dirichlet}(\gamma_{s1}, \ldots, \gamma_{sK}), \\ q(\mathbf{z}_{sg} \,|\, \Phi_{sg}) &= \text{Multinomial}(1; \phi_{sg1}, \ldots, \phi_{sgK}), \end{aligned} \qquad (5)$$

$\gamma_{sk} = \alpha_k + \sum_{g=1}^{G} \phi_{sgk}$, $\phi_{sgk} = \frac{\mathcal{P}(n_{sg}|\lambda_{gk}) \exp[\Psi(\gamma_{sk})]}{\sum_{k'=1}^{K} \mathcal{P}(n_{sg}|\lambda_{gk'}) \exp[\Psi(\gamma_{sk'})]}$, and $\mathcal{P}(n_{sg}|\lambda_{gk})$ denotes the Poisson density with mean λ_{gk} evaluated at n_{sg}.

Because in real data analysis $\boldsymbol{\alpha}$ and Λ are rarely known, we choose an empirical Bayesian approach and estimate $\boldsymbol{\alpha}$ and Λ based on $\mathbf{n}_{1:S}$. Following Blei *et al.* (2003), instead of maximizing analytically intractable $\log p(\mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$, its evidence lower bound (ELBO) from variational inference $\log q(\mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$, which is analytically tractable, is maximized for estimating $\boldsymbol{\alpha}$ and Λ. ELBO is obtained by replacing the functions of latent variables $\boldsymbol{\pi}_{1:S}$ and $Z_{1:S}$ in $\log p(\boldsymbol{\pi}_{1:S}, Z_{1:S}, \mathbf{n}_{1:S} \,|\, \boldsymbol{\alpha}, \Lambda)$ by their conditional expectations with respect to $q(\boldsymbol{\pi}_s \,|\, \boldsymbol{\gamma}_s)$ and $q(Z_s|\, \Phi_s)$ for $s = 1, \ldots, S$. This observation motivates simultaneous iterative estimation of $\boldsymbol{\alpha}$, Λ, $\gamma_{1:S}$, and $\Phi_{1:S}$ based on $\mathbf{n}_{1:S}$ similar to EM algorithm (Dempster *et al.*, 1977). Specifically, if $\phi_{sgk}^{(t)}, \gamma_{sk}^{(t)}, \lambda_{gk}^{(t)}, \alpha_k^{(t)}$ represent the parameter estimates at the t-th iteration, then variational E step at the $(t+1)$-th iteration updates ϕ_{sgk}'s and γ_{sk}'s as

$$\phi_{sgk}^{(t+1)} = \frac{\mathcal{P}(n_{sg}|\lambda_{gk}^{(t)}) \exp[\Psi(\gamma_{sk}^{(t)})]}{\sum_{k=1}^{K} \mathcal{P}(n_{sg}|\lambda_{gk}^{(t)}) \exp[\Psi(\gamma_{sk}^{(t)})]}, \gamma_{sk}^{(t+1)} = \alpha_k^{(t)} + \sum_{g=1}^{G} \phi_{sgk}^{(t+1)},$$

and variational M step at the $(t+1)$-th iteration updates λ_{gk}'s and α_k's as

$$\lambda_{gk}^{(t+1)} = \frac{\sum_{s=1}^{S} \phi_{sgk}^{(t+1)} n_{sg}}{\sum_{s=1}^{S} \phi_{sgk}^{(t+1)}}, \boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \mathbf{H}(\boldsymbol{\alpha}^{(t)})^{-1}\, \mathbf{g}(\boldsymbol{\alpha}^{(t)}), \qquad (6)$$

where $\Psi(.)$ is the digamma function and $\mathbf{H}$ and $\mathbf{g}$ are the Hessian and gradient for $\boldsymbol{\alpha}$ update. We start the iterations using $\phi_{sgk} = \frac{1}{K}$ for all samples, genes, and processes, and $\alpha_k = 1$ for all processes. Later, we recommend two practical approaches for choosing K (see Sections 1 and 2 of Supplementary Material for details).

2.1.2 Interpretation Of Parameter Estimates The interpretation of parameters in Section 2.1.1, and the relation between them are described using (6). The probability that gene g in sample s belongs to the process k is ϕ_{sgk}; therefore, $\sum_{k=1}^{K} \phi_{sgk} = 1$ and $\sum_{g=1}^{G} \phi_{sgk}$ is the expected number of genes in sample s that belong to process k. The probability that sample s belongs to the process k is proportional to γ_{sk}. The prior probability that a gene in any NGS experiment belongs to process k is proportional to α_k. The expected number of genes in process k in sample s is $\gamma_{sk} - \alpha_k$, which equals $\sum_{g=1}^{G} \phi_{gsk}$. This relation can be used for checking the convergence of iterative updates in (6). The expected value of the count for gene g when it belongs to the process k is λ_{gk}.

2.2 Second stage of LPD-C: Selection of gene-subsets

LPD-C's second stage selects genes in subset k based on the posterior means of $q(Z_{1:S})$, $\Phi_{1:S}$, which are estimated in the variational E step (6). The selected genes are a small fraction of the total number of genes and are associated to their respective subsets with high probabilities. Most importantly, this stage extends the original LPD framework of Rogers *et al.* (2005) and makes it more useful for genomic data analysis by selecting genes, grouped in subsets, while controlling the number of false discoveries using a local false discovery rate (locfdr) cutoff (Efron, 2007, 2010). Based on the locfdr procedure, the second stage of LPD-C has these advantages: it does not require modeling of full error structure of the original data set, has few assumptions, and is easy to implement (Efron, 2007). The trade-off for these advantages is the loss of statistical efficiency (Efron, 2007).

Because gene-subsets correspond to processes, genes in subset k are selected using test statistics based on the approximate posterior means of z_{sgk}'s, ϕ_{sgk}'s, for G genes across S samples. If z_{sgk}'s are known for all the samples and genes, then $p_{gk} = \frac{\sum_{s=1}^{S} z_{sgk}}{S}$ represents the probability that gene g belongs to process k. Motivated from EM algorithm (Dempster *et al.*, 1977), modified test statistics $\widehat{p}_{gk}$ are defined by replacing the latent variables z_{sgk}'s in p_{gk} by their conditional expectations with respect to $q(Z_s|\Phi_s)$ for $s = 1, \ldots, S$ and

$$\widehat{p}_{gk} = \frac{\sum_{s=1}^{S} \mathbb{E}[z_{sgk} \,|\, \mathbf{n}_s]}{S} \approx \frac{\sum_{s=1}^{S} \mathbb{E}_{\phi_{sgk}}[z_{sgk}]}{S} = \frac{\sum_{s=1}^{S} \phi_{sgk}}{S}. \qquad (7)$$

The test statistic (7) represents the approximate posterior probability of gene g belonging to process k. Instead of directly using $q(z_{sgk})$'s for quantifying uncertainty in $\widehat{p}_{gk}$, we used their posterior means ϕ_{sgk}'s due to two main reasons. First, it is well-known that variational posterior density under-represents the true variability (Bishop, 2006; Ormerod and Wand, 2010); therefore, using $q(z_{sgk})$'s for uncertainty quantification of genes selection in subset k could possibly lead to greater number of false positives. Second, the variational updates are guaranteed to converge to a local mode of the true posterior density; therefore, $\Phi_{1:S}$ are good approximations of a posterior mode of $Z_{1:S}$. Because Efron (2007) recommends using the test statistics for genes that have the same range as the normal distribution, $\widehat{p}_{gk}$ is transformed to the corresponding quantile of the central t-distribution with ν degrees of freedom, t_{gk}, using its cumulative distribution function $\mathcal{F}_{t_\nu}$, and $t_{gk} = \mathcal{F}_{t_\nu}^{-1}(\widehat{p}_{gk})$. The t-distribution is chosen due to its heavy tails; in real data analysis, we choose $\nu = 3$. Assuming that T represents the matrix of test statistics with G rows and K columns, genes with high posterior probabilities of belonging to process k are in the right tail of $\mathbf{t}_k$, k-th column of T; therefore, $\mathbf{t}_k$ is used as the vector of test statistics in an empirical Bayes testing framework to select genes in subset k that are non-null, that lie in the right tail of $\mathbf{t}_k$, and when locfdr is controlled at a small pre-specified value. This procedure selects a small fraction of genes that are associated with subset or process k with high probabilities. We select K gene-subsets based on the columns of T and separately control locfdr for each column. For the NGS data applications presented later, the R package `locfdr` (Efron *et al.*, 2008) is employed.

3 APPLICATIONS OF LPD-C

We apply LPD-C to simulated and real NGS data, and compare its performance to both SPLDA (Witten, 2011) and a negative binomial model (EdgeR; Robinson *et al.* (2010)). These methods are chosen because Witten (2011) shows that SPLDA performs significantly better than current approaches (except EdgeR) for classifying and clustering NGS data. The simulated data are generated using the hierarchical model (1) – (3). Two publicly available NGS datasets are used: human cervical cancer data (hereafter cervical cancer data;

Witten *et al.* (2010)) and human gene expression data from liver and kidney (hereafter human data; Marioni *et al.* (2008)). These real data are chosen because Witten (2011) shows that both EdgeR and SPLDA perform well for the human data, but that the cervical cancer data are challenging for both of these methods. It is important to remember that LPD-C is a Bayesian latent factor model, and that it is fundamentally different from the hypothesis testing based approach of EdgeR, and from the penalized likelihood based approach of SPLDA. However, the comparisons illustrate the similarities and differences in these methods. The novel feature that distinguishes LPD-C from existing approaches for NGS data analysis is that it groups selected genes into a pre-specified number of gene-subsets. Further, Rogers *et al.* (2005) show that LPD is more flexible than hierarchical and k-means clustering, so we do not compare LPD-C with these classical unsupervised methods.

3.1 Simulation

We simulated 10 NGS datasets such that each dataset contains 12 samples (S) with 2 processes (K) for different settings of G and λ_{gk}'s. For each setting, the simulated data have the following process membership for the genes. In samples 1 to 10, the first 100 genes (hereafter group 1 genes) belong to the first process and the last 100 genes (hereafter group 2 genes) belong to the second process. For a particular G, the first 10 samples have the following five settings of gene- and process-specific means λ_{gk}'s depending on Δ,

$$\lambda_{g1} = \begin{cases} \exp(z_{g1}),\ z_{g1} \sim \text{Normal}(\Delta, 1) & \text{for } g = 1, \ldots, 100, \\ \exp(z_{g1}),\ z_{g1} \sim \text{Normal}(0, 0.25) & \text{for } g = 101, \ldots, G, \end{cases}$$

$$\lambda_{g2} = \begin{cases} \exp(z_{g2}),\ z_{g2} \sim \text{Normal}(0, 0.25) & \text{for } g = 1, \ldots, G-100, \\ \exp(z_{g2}),\ z_{g2} \sim \text{Normal}(\Delta, 1) & \text{for } g = G-99, \ldots, G, \end{cases} \quad (8)$$

where Δ is varied as $1, 2, 3, 4$, and 5. These values of Δ represent the difference between the log-means of the "null" (i.e., genes that are not in group 1 and 2) and "non-null" genes (i.e., group 1 and 2 genes) in the two processes. The number of genes (G) is varied as 2000 and 20,000 genes, respectively, while the number of non-null genes is 200 in both cases. For samples 11 and 12, the process memberships of group 1 and 2 genes are reversed. The remaining genes belong to the two processes with 0.5 probability across all samples; therefore, group 1 and 2 genes belong to processes 1 and 2, respectively, with high probability (10/12 $\sim$ 80%). These parameter values are motivated from Efron *et al.* (2008) and Witten (2011). NGS data are simulated using these parameter values and LPD-C's generative model (1) – (3). The simulated data are similar to those observed in practice, with a large fraction of small counts and a small fraction of large counts.

3.1.1 Application of EdgeR, LPD-C, and SPLDA We applied the first stage of LPD-C to 50 replications of the simulated data. For each application of LPD-C, we chose $K = 2$ to facilitate comparison with the truth and estimated $\boldsymbol{\alpha}$, Λ, Φ's, and $\boldsymbol{\gamma}$'s (see (6) for their definition). The results of variational approximation are known to be sensitive to the starting points, which in LPD-C's case depend on $\boldsymbol{\alpha}$ and Φ's (Bishop, 2006). We used multiple starting points until convergence to the posterior mode was stable. We observed that the final parameter estimates were most sensitive to the starting values of Φ's and were fairly robust to the starting values of $\boldsymbol{\alpha}$. The process numbers are identified based on the ascending order of α_k's such that $\alpha_{(1)}$ and $\alpha_{(2)}$ correspond to processes 1 and 2, respectively. Although LPD-C uses an approximate estimation method, its estimates of λ_{gk}'s agree closely with their true values, even at low values of Δ. After estimating $\widehat{p}_{gk}$'s from ϕ_{sgk}'s, we obtain $t_{gk} = \mathcal{F}_{t_3}^{-1}(\widehat{p}_{gk})$, where $\mathcal{F}_{t_3}$ is the cumulative distribution function of the standard t-distribution with 3 degrees of freedom (see Section 2.2). We apply empirical Bayes hypothesis testing to the columns of T, which correspond to processes, and select genes that are non-null, that are in the right tail, and that have a locfdr of 0. The selected genes belong to their processes, and hence to the corresponding gene-subsets, with high probability.

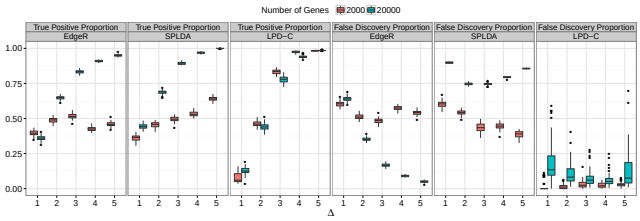


Fig. 1: Comparison of true positives and false discoveries in the genes selected by EdgeR, LPD-C, and SPLDA in 50 replications of simulated data analysis. The x-axis represents the difference between the log-means of the null and non-null genes in the two processes (Δ; see (8)) and y-axis represents the true positive (TPP) and false discovery (FDP) proportions. Panels one through three represent TPPs, while panels four through six represent FDPs, for EdgeR, LPD-C, and SPLDA, respectively, when the number of genes (G) is 2000 (red) and 20,000 (blue). For both LPD-C and SPLDA, TPP increases and FDP decreases as Δ increases when $G = 2000$ and $20,000$; EdgeR also has a similar pattern except when $G = 2000$, where the TPP and FDP oscillates around 0.5 for all the Δ's. Although TPPs appear to increase with Δ for EdgeR, LPD-C, and SPLDA, the FDPs are much higher than their expected values. This pattern is expected because the number of non-null genes is same for $G = 2000$ and 20,000, and the power increase is accompanied by an increase in FDPs.

3.1.2 Results of EdgeR, LPD-C, and SPLDA LPD-C selects genes grouped in subsets, but EdgeR and SPLDA do not; therefore, we compare overall gene selection of LPD-C with that of EdgeR and SPLDA. Unlike LPD-C, both EdgeR and SPLDA select genes based on a response variable. We define a response variable (Y) that is 1 for the first ten samples and is 2 for samples 11 and 12, and EdgeR selects genes that are differentially expressed between samples with $Y = 1$ and $Y = 2$. Similarly, SPLDA finds a sparse list of genes that can classify samples as $Y = 1$ or $Y = 2$ based on their expression while minimizing the cross-validation (CV) error for classification. We use `edgeR` package (Robinson *et al.*, 2010) for EdgeR and `PoiClaClu` package (Witten, 2011) for SPLDA. We obtain the gene-wise p-values for differential expression using `edgeR`, correct for multiple comparisons using the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995), and choose the genes that corresponded to 200 smallest BH corrected p-values. For SPLDA, we choose the tuning parameters depending on G and Δ so that it select 200 genes.

Figure 1 shows the true positive and false discovery proportions for EdgeR, LPD-C, and SPLDA at different values of G and Δ. The proportion of true positives selected by LPD-C increases with

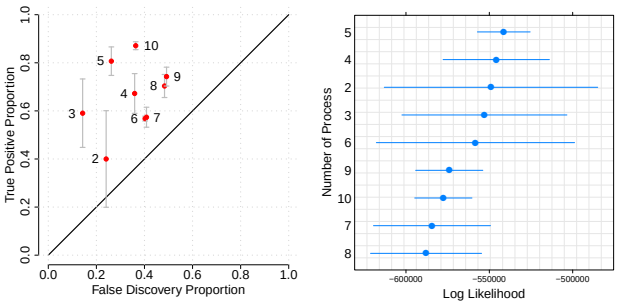
Δ when $G = 2000$ and 20,000; however, the proportion of false discoveries are much higher than their expected value when $G = 20,000$. This observation is expected because the number of non-null genes are 200 for both $G = 2000$ and 20,000, and the apparent increase in true positives when $G = 20,000$ comes at the cost of increased false discoveries. The true positive proportions for both EdgeR and SPLDA behave similar to that in LPD-C, but their values are lower than that of LPD-C. The false discovery proportions of SPLDA and LPD-C are much higher than that of EdgeR when $G = 20,000$. This observation for EdgeR is an artifact our gene selection procedure that only selects the first 200 genes based on the ascending order of p-values; however, the proportion of false discoveries of EdgeR when $G = 20,000$ is also much higher than those shown in Figure 1 when genes are selected based on the standard FDR cutoff 0.05.

3.2 Real data examples

We apply LPD-C to two publicly available NGS datasets. The cervical cancer data provide measurements of the digital expression for 714 small RNAs (miRNAs) in 29 tumor and 29 normal cervical tissue samples from humans (Witten *et al.*, 2010). The human data provide measurements of the digital expression for 22,925 genes in 14 samples from a single human male, which consists of seven technical replicates from liver and kidney, respectively (Marioni *et al.*, 2008). The cervical cancer data were collected for discovering miRNAs associated with human cervical cancer. The human data were collected for comparing microarrays and NGS technologies.

3.2.1 Selection of the number of processes K We suggest two practical approaches based on n-fold CV for selecting the number of processes K in real data. The problem of selecting K is similar to that of selecting the number of clusters, which is known to be a notoriously difficult problem; therefore, we suggest fitting LPD-C for a range of K's and selecting those K's which lead to results that agree closely with the biological knowledge. The first approach uses n-fold CV, varies K from 2 to a large integer, and calculates n held-out log likelihoods for each K. It chooses the K that maximizes the median of n held-out log likelihoods (Rogers *et al.*, 2005). This approach is well-suited for data with small sample sizes (e.g., the human data). The second approach chooses K using n-fold CV based on the true positive proportion (TPP) and false discovery proportion (FDP) determined from training and test data (Hastie *et al.*, 2009). Assuming that the genes selected by LPD-C in the training data represent the truth, this approach calculates TPP and FDP in the genes selected by LPD-C in the test data. This process is repeated n times to yield n TPPs and FDPs for each K. The values of K that have large TPPs and small FDPs represent good choices of K. This approach is more suitable for data with relatively large sample size (e.g., the cervical cancer data).

Figures 2a and 2b illustrate the determination of K for both the cervical cancer and the human data using 5-fold CV. We choose $K = 5$ for the human data because it has the maximum held-out log likelihood with a small median absolute deviation estimate compared to other values of K. For the cervical cancer data, both $K = 5$ and $K = 3$ are reasonable choices. As such, we selected genes in the cervical cancer data using LPD-C for both $K = 5$ and 3 and found that the results obtained using $K = 5$ agree closely with those from EdgeR and SPLDA; therefore, all our subsequent analyses for the cervical cancer data are based on $K = 5$.



(a) (b)

Fig. 2: Selecting the number of processes K in LPD-C for cervical cancer data (a) and human data (b) using 5-fold CV. (a) Median false discovery proportion (FDP; on x-axis) versus median true positive proportion (TPP; on y-axis) in cervical cancer data based on 5-fold CV when $K = 2, \ldots, 10$ (red points), respectively. The vertical lines (grey) show 1 median absolute deviation (MAD) intervals for the TPPs. Based on this plot, 3 and 5 are good candidates for K in cervical cancer data due to their relatively low FDPs and high TPPs. Further data analysis shows that $K = 5$ is a better candidate than $K = 3$. (b) The y-axis shows the number of processes and the x-axis showing the medians and 1 MAD intervals of the held-out log likelihoods in human data based on 5-fold CV when $K = 2, \ldots, 10$. Based on this plot, we choose $K = 5$ for human data.

3.2.2 Application of EdgeR, LPD-C, and SPLDA The first stage of LPD-C estimates $\boldsymbol{\alpha}$, Λ, Φ's, and $\boldsymbol{\gamma}$'s for both the cervical cancer and human data using $K = 5$ (see (6)). Similar to the simulation study, we tried various starting points for $\boldsymbol{\alpha}$ and Φ's until convergence to the posterior mode was stable; identified the process numbers based on the ascending order of α_k's; after estimating $\widehat{p}_{gk}$'s from Φ's, obtained $t_{gk} = \mathcal{F}_{t_3}^{-1}(\widehat{p}_{gk})$; and used columns of T to select genes using a locfdr cutoff of 0.2 for each subset. The two features of empirical Bayes hypothesis testing that are useful here are its mild distributional assumptions on, and no requirement for modeling the full error structure of t_{gk}'s (Efron, 2007, 2010).

We also apply EdgeR and SPLDA to the cervical cancer data using tumor status as the response variable. Similarly, EdgeR and SPLDA are applied to the human data using liver and kidney as values of the response variable. We obtain gene-wise p-values for differential expression using `edgeR`, correct for multiple comparisons using the BH procedure, and select genes using 0.05 as the cutoff for the BH corrected p-values. We used 5-fold CV for both the cancer and human data in SPLDA. While the choice of tuning parameter using CV is fairly stable for the human data, multiple tuning parameters lead to the same CV classification error for the cervical cancer data, which results in unstable gene-selection. For example, at the same value of CV error for classification, SPLDA selects 2 genes when the tuning parameter is 8.23 and selects 499 genes when the tuning parameter is 0.57. The reason for these unstable results is that a large range of tuning parameters yield the same classification error, 0.172, which corresponds to 10 out

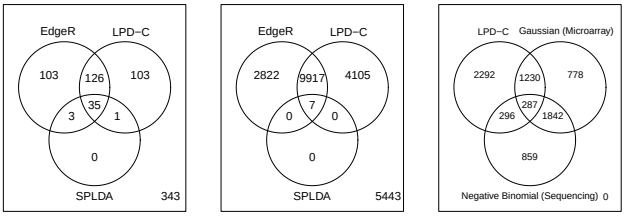
of 58 errors. Classification error is a very coarse measure (unlike, for example, the mean square error for regression), so many tuning parameter values are tied in terms of CV error for classification. We choose the tuning parameter as the mean of all the tuning parameter values that correspond to the minimum CV classification error (personal communication, D. Witten).

3.2.3 Results of EdgeR, LPD-C, and SPLDA Figures 3a and 3b summarize the number of genes selected in both the cervical cancer and the human data using EdgeR, LPD-C, and SPLDA, respectively. The total number of genes selected by EdgeR, LPD-C, and SPLDA are 267, 265, and 39 for the cervical cancer data, and 12746, 14029, and 7 for the human data. LPD-C selected 103 unique miRNAs in the cervical cancer data that are related to different types of cancers, including cervical cancer, and that are not selected by EdgeR or by SPLDA. Some of these 103 miRNAs are known to be in the *let-7*, *mir-7*, *mir-17*, *mir-24*,*mir-26*, *mir-27*, *mir-29*, *mir-124*, *mir-127*, *mir-192*, and *miR-744* families that have clinical applications in cancer diagnosis and therapy. Because the human data were collected for comparing microarray and NGS technology, we did not investigate the biological annotation of the genes selected by LPD-C.

LPD-C's results agree closely with those of EdgeR in that both of these methods select most of the genes chosen by SPLDA. When compared to EdgeR, about 61% and 70% of the genes selected by LPD-C in the cervical cancer data and the human data are also declared as differentially expressed. We also notice that the number of genes selected by SPLDA in both datasets is much smaller than the number of genes selected by either EdgeR or LPD-C. Further, unlike SPLDA and EdgeR, LPD-C groups the selected genes into subsets with desirable properties.

Marioni *et al.* (2008) employed both microarrays and NGS technologies to compare the differentially expressed genes. They used two sample t tests for the microarray data analysis using a Gaussian model and likelihood ratio tests for NGS data analysis using a negative binomial model. Figure 3c compares the 4105 genes selected only by LPD-C in the human data (Figure 3b) with the differentially expressed genes in microarray or NGS data reported by Marioni *et al.* (2008), excluding the 9924 genes that lie in the intersection of LPD-C and EdgeR (Figure 3b). We observe that almost half of the genes that are selected solely by LPD-C are also reported as differentially expressed in the microarray or NGS data results of Marioni *et al.* (2008).

We have demonstrated that LPD-C selects genes that compare favorably with existing approaches, such as EdgeR and SPLDA, even though SPLDA consistently selects a smaller number of genes than EdgeR and LPD-C. We note that this under-selection issue has been observed in applications of Lasso for variable selection to high-dimensional data that have dependence among the variables (Friedman *et al.*, 2010). Since most high-dimensional biological data, including NGS data, have dependence among their variables, this could be a potential reason for SPLDA selecting a smaller number of genes. Furthermore, since the tuning parameters is not identifiable in the cervical cancer data it leads to an unstable selection of genes which in turn makes SPLDA undesirable if used for gene selection. As an alternative suggestion for situations like these, we recommend using the glmnet algorithm for variable selection (Friedman *et al.*, 2010).



(a) (b) (c)

Fig. 3: (a) Comparison of the miRNAs selected in the cervical cancer data. (b) Comparison of the genes selected in the human data. (c) Comparison of the 4105 genes, selected solely by LPD-C in the human data, with the differentially expressed genes found in the microarray and NGS data analyses reported by Marioni *et al.* (2008). It excludes the 9924 genes that are in the intersection of EdgeR and LPD-C. Approximately 44% of the genes that are selected by LPD-C in the human data are differentially expressed in either the microarray or the NGS data.

Table 1. Number of genes, and fraction of differentially expressed genes, selected by LPD-C in the five gene-subsets for both the cervical cancer and the human data. The diagonal elements in columns 1-5 represent the total number of genes selected by LPD-C in the respective gene-subsets. The upper off-diagonal elements in columns 1-5 represent the number of genes that are in common between two gene-subsets, while columns 6-10 represent the fraction of differentially expressed genes among the processes as determined by `edgeR` (Robinson *et al.*, 2010).

	Number of Selected Genes					**Differentially Expressed Genes**				
CERVICAL CANCER DATA										
Process	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	5	1	2	0	0	0.60	1	0.5	0	0
2		44	13	4	3		0.59	0.62	0.5	0.33
3			61	8	6			0.59	0.75	0.67
4				103	25				0.67	0.80
5					113					0.62
HUMAN DATA										
Process	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	4128	806	935	1150	1773	0.72	0.74	0.72	0.71	0.79
2		3188	695	985	13		0.78	0.78	0.79	0.46
3			4107	469	1163			0.70	0.84	0.79
4				5476	1391				0.68	0.77
5					5090					0.77

The distinguishing feature of LPD-C is that it selects genes grouped as subsets having desirable properties. Table 1 summarizes the number of genes in each of the five gene-subsets as selected by LPD-C for both the cervical cancer data, and the human data. It also illustrates the number of genes that are in common when any two gene-subsets are compared, as well as the proportion of genes that are differentially expressed. For the human data, where SPLDA and EdgeR results agree, LPD-C results are similar. For the cervical cancer data, where SPLDA and EdgeR results do not agree, LPD-C leads to results that are close to those of EdgeR.

4 DISCUSSION

Due to the decreasing cost of using NGS technologies and the potential impact of large-scale genome-wide epidemiological projects, such as 1000 Genomes Project (Siva, 2008), genomic data

are becoming increasingly complex and large. Bayesian generative models offer an attractive approach to model the latent structure of genomic data. We have presented an application of LPD framework to NGS data, LPD-C, that addresses these challenges in two stages. LPD-C's first stage extends the generative Bayesian model of LPD for modeling microarray data to NGS data. Its second stage uses the parameter estimates from first stage to select genes, organized as gene-subsets, that are a small fraction of total number of genes and that belong their respective subsets with high probability. To achieve computationally tractable Bayesian inference, we have applied variational inference, and combined the results of variational inference with empirical Bayes hypothesis testing to select gene-subsets that control the number of false discoveries at a certain level. We have explored LPD-C's application in the context of simulated and real NGS data, and demonstrated that LPD-C's results agree with competing non-Bayesian approaches.

LPD-C can be easily modified to yield its supervised extension, SLPD-C. Representing $\mathbf{x}_s$ as the covariate information for sample s (e.g., disease status, survival time), SLPD-C extends (3) using a gene-specific analogue of mixed effects model

$$n_{sg}|\widetilde{\lambda}_{sg} \sim \text{Poisson}(\widetilde{\lambda}_{sg}) \text{ and } \log \widetilde{\lambda}_{sg} = \log \lambda_{gk} + \boldsymbol{\beta}_g^T \mathbf{x}_s + \epsilon_g,$$

where $\boldsymbol{\beta}_g$ is the mean covariate effect for gene g and ϵ_g is idiosyncratic noise. Another extension of LPD for modeling data from time course experiments follows immediately as an extension of dynamic linear models (Blei and Lafferty, 2006). The apriori choice of the number of processes, K, facilitates efficient parameter estimation. Sometimes, however, the apriori knowledge about K is unavailable or K is unidentifiable from approaches recommended in Section 3.2.1. In these scenarios it is desirable to adaptively select K based on the genomic data using Bayesian Nonparametrics (Hjort *et al.*, 2010). It is also desirable to develop MCMC algorithms tuned for LPD to estimate uncertainty in parameters of interest by sampling from their posterior densities. To this end, the collapsed Gibbs sampler of Griffiths and Steyvers (2004) can be easily extended for LPD.

The second stage of the LPD uses the parameter estimates from the Bayesian model and selects gene-subsets with desirable properties. Similar ideas about finding groups of differentially expressed gene-subsets have been explored starting with gene-set enrichment analysis (GSEA) (Subramanian *et al.*, 2005). We propose to investigate the relationship between the enriched gene-subsets obtained from GSEA and those obtained from LPD. Finally, we propose to incorporate sparsity in the second stage of the LPD using appropriate priors on Λ from Bayesian variable selection literature, such as horse-shoe prior and multiplicative gamma prior (Carvalho *et al.*, 2010; Bhattacharya and Dunson, 2011).

ACKNOWLEDGEMENT

This work is funded in part by a National Science Foundation (DBI-0733857) grant to RWD and her colleagues. SS benefited from discussions with Professors J.K. Ghosh and S. Kirshner.

REFERENCES

Airoldi, E., Blei, D., Xing, E., and Fienberg, S. (2005). A latent mixed membership model for relational data. In *Proceedings of the 3rd international workshop on Link discovery*, pages 82–89. ACM.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, **98**(2), 291–306.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer New York.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**, 993–1022.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, **103**(484), 1438–1456.

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dunson, D. and Herring, A. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, **6**(1), 11–25.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, **35**(4), 1351–1377.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge Univ Pr.

Efron, B., Turnbull, B., and Narasimhan, B. (2008). locfdr: Computes local false discovery rates. *R package*, page 195.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(Suppl 1), 5228.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, **11**(1), 422.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag.

Hjort, N., Holmes, C., Müller, P., and Walker, S., editors (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509.

Ormerod, J. and Wand, M. (2010). Explaining variational approximations. *The American Statistician*, **64**(2), 140–153.

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

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**(2), 321.

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

Rogers, S., Girolami, M., Campbell, C., and Breitling, R. (2005). The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 143–156.

Siva, N. (2008). 1000 genomes project. *Nature biotechnology*, **26**(3), 256–256.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. *Bayesian statistics*, **7**(2003), 723–732.

Witten, D. (2011). Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics*, **5**(4), 2493–2518.

Witten, D., Tibshirani, R., Gu, S., Fire, A., and Lui, W. (2010). Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC biology*, **8**(1), 58.

Young, M. *et al.* (2012). Differential expression for rna sequencing (rna-seq) data: Mapping, summarization, statistical analysis, and experimental design. *Bioinformatics for High Throughput Sequencing*, pages 169–190.