Biomarker detection and categorization in RNA-seq meta-analysis using Bayesian hierarchical model

Tianzhou Ma
Department of Biostatistics
University of Pittsburgh, Pittsburgh, PA 15261
email: tim28@pitt.edu

Faming Liang
Department of Biostatistics
University of Florida, Gainesville, FL 32611
email: faliang@ufl.edu

George C. Tseng
Department of Biostatistics
University of Pittsburgh, Pittsburgh, PA 15261
email: ctseng@pitt.edu

Author's Footnote:

Tianzhou Ma is Doctoral Candidate at Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261 (email: tim28@pitt.edu). Faming Liang is Professor at Department of Biostatistics, University of Florida, Gainesville, FL 32611 (email: faliang@ufl.edu). George C. Tseng is Professor at Department of Biostatistics (primary appointment), Department of Human Genetics, Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA, 15261 (email: ctseng@pitt.edu).

Abstract

Meta-analysis combining multiple transcriptomic studies increases statistical power and accuracy in detecting differentially expressed genes. As the next-generation sequencing experiments become mature and affordable, increasing number of ribonucleic acid sequencing (RNA-seq) datasets are available in the public domain. Count-data-based technology provides better experimental accuracy, reproducibility and ability to detect low expressed genes. A naive approach to combine multiple RNA-seq studies is to apply differential analysis tools such as edgeR and DESeq to each study and then to combine the summary statistics of *p*-values or effect sizes by conventional meta-analysis methods. Such a two-stage approach loses statistical power, especially for genes with short length or low expression abundance. We propose a full Bayesian hierarchical model (namely, BayesMetaSeq) for RNA-seq meta-analysis by modelling count data, integrating information across genes and across studies, and modelling potentially heterogeneous differential signals across studies via latent variables. A Dirichlet process mixture (DPM) prior is further applied on the latent variables to provide categorization of detected biomarkers according to their differential expression patterns across studies, facilitating improved interpretation and biological hypothesis generation. Simulations and a real application on multiple brain region HIV-1 transgenic rats demonstrate improved sensitivity, accuracy and biological findings of the method.

Keywords: Bayesian hierarchical model, differential expression (DE), meta-analysis, model-based clustering, Ribonucleic acid sequencing (RNA-seq)

1. INTRODUCTION

By using the next-generation sequencing technology to quantify transcriptome, ribonucleic acid sequencing (RNA-seq) has rapidly become a standard experimental technique in measuring RNA expression levels (Mortazavi et al., 2008; Wang et al., 2009). For RNA-seq, the abundance of transcript in each RNA sample is measured by counting the number of randomly sequenced fragments aligned to each gene. Compared to the popular microarray technology, RNA-seq has the advantage of detecting novel transcripts and quantifying a larger dynamic range of expression levels. It has been shown that it performs better than microarray techniques at detecting weakly expressed genes if sequencing is sufficiently deep (Wang et al., 2014). However, new statistical challenges emerge in the differential expression (DE) analysis of RNA-seq data. First, the sequencing data are discrete counts rather than continuous intensities, so a count model is more appropriate if a parametric approach is used. Secondly, since long transcripts usually have more mapped reads compared with short transcripts and the detection power of DE increases as the number of reads increases, short transcripts are always at a statistical disadvantage relative to long transcripts in the same dataset. Analysis of RNA-seq data needs to address such a read count bias considering that many important disease markers are of short length or low expression (Oshlack et al., 2009).

Many methods have been developed to identify differentially expressed genes between two or more conditions for RNA-seq count data. Two most popular tools edgeR and DESeq assume a negative binomial model that takes overdispersion into account and either a likelihood ratio test or an exact test is used to test for DE (Robinson et al., 2010; Anders and Huber, 2010). Other methods such as baySeq or EBSeq apply empirical Bayes approaches to detect patterns of DE (Hardcastle and Kelly, 2010; Leng et al., 2013). Recently, more methods have been developed using Bayesian hierarchical model and have used either approximation methods or Markov chain Monte Carlo (MCMC) sampling schemes to estimate the parameters (Van De Wiel et al., 2012; Chung et al., 2013). No single method has been shown to outperform the other methods under all circumstances in recent comparative studies (Rapaport et al., 2013; Soneson and Delorenzi, 2013). Bayesian approaches are advantageous in handling complex models and adopting more flexible modelling of effect size and variance, and thus may increase DE detection power for lowly expressed genes (Chung et al., 2013). However, all Bayesian hierarchical models are limited to single transcriptomic studies so far.

Meta-analysis in genomic research is a set of statistical tools for combining multiple "-omics" studies of a related hypothesis and can potentially increase the detection power of individual studies (Tseng et al., 2012). With the increasing availability of RNA expression data sets, many transcriptomic meta-analysis methods for microarray data have been developed in the past decade. These methods mainly fall into three categories. The first and the most popular is a two-stage method, where a single summary statistics is first computed for each study and then meta-analysis methods are used to combine the summary statistics. These methods include combining p-values (Fisher, 1925; Stouffer et al., 1949; Li et al., 2011), combining effect sizes (Choi et al., 2003) or combining rank statistics (Hong et al., 2006). The second category of methods merges the raw data from all microarray studies and normalizes simultaneously (also known as mega-analysis), then standard single-study analysis can be applied (Lee et al., 2008; Sims et al., 2008). These approaches have, however, been less favoured in the literature since they do not guarantee to remove cross-study discrepancy and may fail to retain study-specific biomarkers. Instead of using two-stage approaches (i.e. DE analysis in single studies plus meta-analysis summary statistics in the first category, and normalization and combined DE analysis in the second category), the third category integrates DE information from all studies by using a unified and joint stochastic model (Conlon et al., 2006; Scharpf et al., 2009). Since they are joint hierarchical models by nature, the more flexible Bayesian methods are usually applied. These approaches have the potential to offer additional efficiency over the two-stage methods and, at the same time, retain the study-specific features. This motivates us to develop a Bayesian hierarchical model for RNA-seq meta-analysis.

In the literature, almost no meta-analysis methods have been developed for RNA-seq so far. Two existing R packages claimed for RNA-seq meta-analysis – metaRNASeq (Rau et al., 2014) and metaSeq (Tsuyuzaki and Nikaido, 2013) – essentially apply naive two-stage methods by using DESeq or NOISeq methods in single studies and combining p-values by Fisher's or Stouffer's method. The two-stage approach leads to a loss of statistical power especially when the observed counts in a given gene are small. In this paper, we propose a Bayesian hierarchical model, BayesMetaSeq, under a unified meta-analytic framework, to analyze RNA-seq data from multiple studies jointly. Bayesian hierarchical modelling allows sharing of information across studies and genes to increase

DE detection power for genes with low read counts. In addition, a Dirichlet process mixture (DPM) prior is imposed on the DE latent variables to model the homogeneous and heterogeneous differential signals across studies. Model-based clustering embedded in the full Bayesian model provides categorization of detected biomarkers according to their DE patterns across studies. The result facilitates better biological interpretation and hypothesis generation.

Ramasamy et al. (2008) presented seven key issues when conducting microarray meta-analysis, including identifying and extracting experimental data, preprocessing and annotating each data set, matching genes across studies, statistical methods for meta-analysis and final presentation and interpretation. When combining RNA-seq studies for meta-analysis, most preliminary steps and data preparation issues will similarly apply. Identification and decision to include adequate transcriptomic studies into meta-analysis greatly impacts accuracy and reproducibility of biomarker detection (Kang et al., 2012). Many useful RNA-seq preprocessing tools such as fastQC, tophat and bedtools are instrumental for alignment and preparing expression counts for downstream analysis. Genes are matched across studies by using standard gene symbols or isoforms through a common reference genome (e.g. hg18 or hg19) (Oshlack et al., 2010). In the remainder of this paper, we assume that data collection and preprocessing have been carefully done and we focus only on downstream meta-analytic modelling and interpretation.

The paper is organized as follows. Section 2 describes the Bayesian hierarchical model and an MCMC algorithm for simulating posterior distributions of parameters. Section 3 explains how we perform DE analysis and cluster analysis based on Bayesian inference with multiple comparison addressed from a Bayesian perspective. In Section 4 and 5, we apply BayesMetaSeq to both simulation and a multiple brain region RNA-seq data set from HIV transgenic rat. Final conclusions and discussion are provided in Section 6.

2. BAYESIAN HIERARCHICAL MODEL

2.1 Notation and assumptions

In this paper, we denote by y_{gik} the observed count for gene g and sample i in study k, $T_{ik} = \sum\limits_{g=1}^{G} y_{gik}$ the library size (i.e. the total number of reads) for sample i in study k and $X_{ik} \in \{0, 1\}$ the phenotypic condition of sample i in study k. The observed data are:

$$D = \{(y_{gik}, T_{ik}, X_{ik}) : g = 1, \dots, G; i = 1, \dots, N_k; k = 1, \dots, K\},$$

where G is the total number of genes, N_k is the sample size of study k and K is the number of studies in the meta-analysis. The latent variable of interest $\delta_{gk} \in \{0, 1\}$ is the study-specific indicator of DE for gene g in study k, meaning that gene g is differentially expressed in study k if $\delta_{gk} = 1$ and non-differentially expressed if $\delta_{gk} = 0$.

Here we assume that the raw RNA-seq count values follow a negative binomial distribution under each condition. We also assume that genes are matched across studies, although the model could be readily extended to analyze multiple studies with similar but not completely overlapped gene sets. In the next three subsections, we shall introduce the generative model within each study (Section 2.2), describe integration of information on effect sizes across studies (Section 2.3) and model clusters of genes with different DE patterns across studies (Section 2.4). Figure 1 provides a graphical representation of the full Bayesian hierarchical model. Parameters within the rectangle form the main model and parameters outside the rectangle are hyperparameters. The grey-shaded parameters δ_{gk} (the latent variable of DE indicator) and λ_g (the DE effect size) are the parameters of interest in the model. The broken rectangle refers to a Dirichlet process mixture (DPM) model for DE gene categorization that will be described in Section 2.4.

2.2 Generative model within each study

Below, we describe the generative model for observed data within each study. We assume that the counts y_{gik}, conditioning on hyperparameters, are independent and follow a negative binomial distribution. Denote by $\mu_{gik} = E(y_{gik})$ the mean expression level and ϕ_{gk} the gene-specific dispersion parameter, we have:

$$y_{gik} \sim NB(\mu_{gik}, \phi_{gk}). \tag{1}$$

We then fit a log-linear regression model for the mean μ_{gik}, where α_{gk} denotes the baseline expression relative to the library size and β_{gk} denotes the effect size (i.e. the log-fold change of expression between the two conditions):

$$\log(\mu_{gik}) = \log(T_{ik}) + \alpha_{gk} + \beta_{gk} X_{ik}. \tag{2}$$

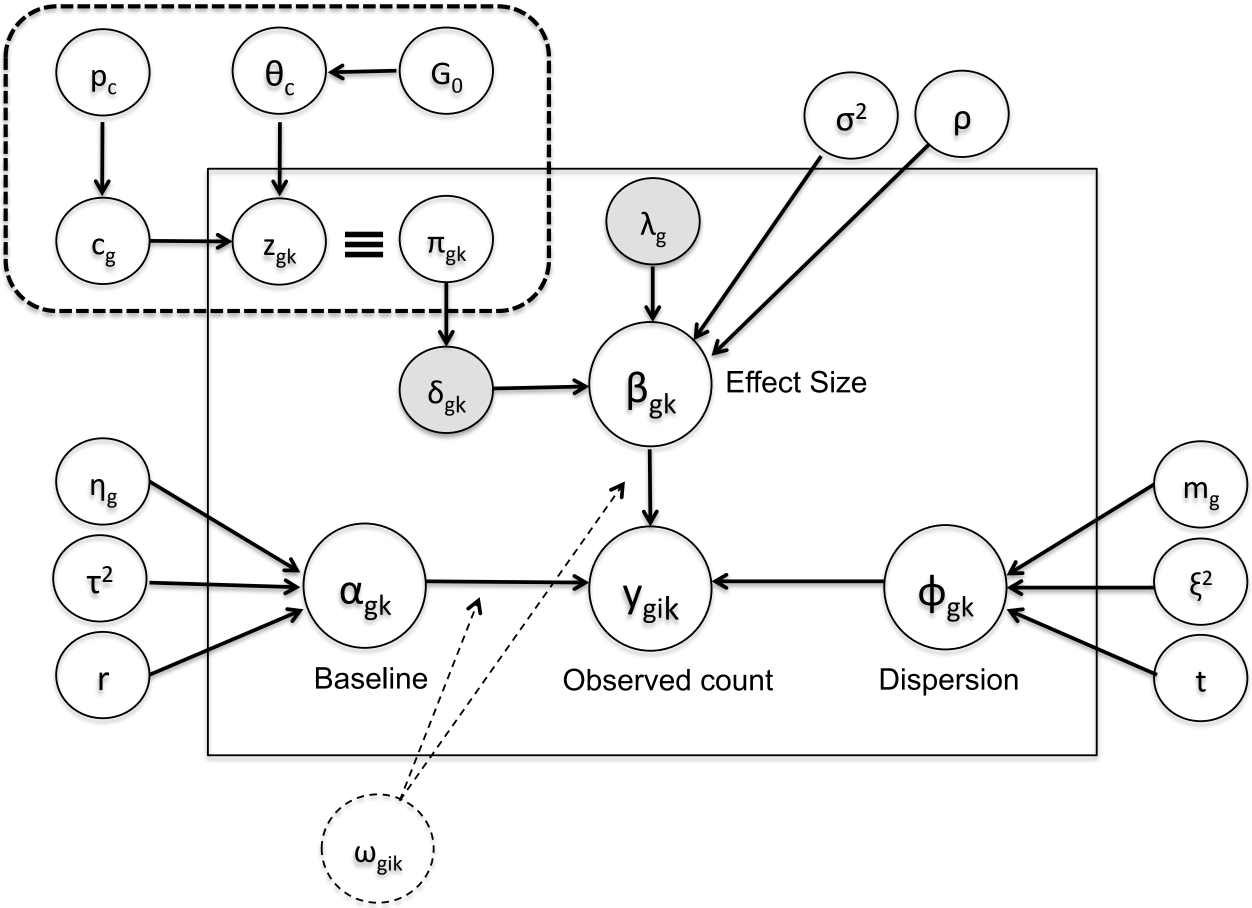

Figure 1: A graphical representation of the Bayesian hierarchical model

We set β_{gk} to depend on both g and k, allowing the existence of between-study heterogeneity for the same gene. If we reparametrize the negative binomial model in (1) in terms of proportion p ($\equiv \frac{\phi\mu}{1+\phi\mu}$) and dispersion ϕ, and let

$$\Psi = \text{logit}(p) = \log(\frac{\frac{\phi\mu}{1+\phi\mu}}{\frac{1}{1+\phi\mu}}) = \log(\phi\mu),$$

we can rewrite equation (2) as:

$$\Psi_{gik} = \log(T_{ik}) + \alpha_{gk} + \beta_{gk}X_{ik} + \log(\phi_{gk}). \tag{3}$$

This equation is useful when we later use Gibbs sampling to update the parameters α_{gk} and β_{gk}. Taking equation (1) and (2) together form our basic generalized linear model:

$$y_{gik}|\alpha_{gk}, \beta_{gk}, \phi_{gk} \sim NB(\log(T_{ik}) + \alpha_{gk} + \beta_{gk}X_{ik}, \phi_{gk}). \tag{4}$$

2.3 Information integration of effect size across studies among differentially expressed genes

Next, we select appropriate prior distributions for the model parameters in equation (4) to allow integration of information across studies. We first define the vectors:

$$\vec{\alpha}_g = (\alpha_{g1}, \ldots, \alpha_{gK})^T, \quad \vec{\beta}_g = (\beta_{g1}, \ldots, \beta_{gK})^T, \quad \log(\vec{\phi}_g) = (\log(\phi_{g1}), \ldots, \log(\phi_{gK}))^T,$$

which represent the baseline, effect size and dispersion vectors for gene g respectively. The three vectors are assumed to be *a priori* independent of each other. In addition, we define the vector for the DE indicators of gene g: $\vec{\delta}_g = (\delta_{g1}, \ldots, \delta_{gK})^T$. We assume both of the vectors $\vec{\alpha}_g, \log\vec{\phi}_g$ follow a multivariate Gaussian distribution:

$$\vec{\alpha}_g \sim N_K(\eta_g, \mathbf{\Lambda}), \quad \log\vec{\phi}_g \sim N_K(m_g, \mathbf{\Pi}), \tag{5}$$

where η_g and m_g are the gene-specific grand means for $\vec{\alpha}_g$ and $\log\vec{\phi}_g$, respectively. The covariance matrices $\mathbf{\Lambda}$ and $\mathbf{\Pi}$ are shared by all genes to be described below. For $\vec{\beta}_g$, we assume a multivariate Gaussian prior, with different means for differentially expressed and non-differentially expressed genes:

$$\vec{\beta}_g \sim N_K(\lambda_g\vec{\delta}_g, \mathbf{\Sigma}), \tag{6}$$

where λ_g is the gene-specific grand mean for differentially expressed genes (i.e. $\delta_{gk} \neq 0$ for some k). For non-differentially expressed genes ($\vec{\delta}_g = 0$), the grand mean is 0. We also allow a different covariance matrix of $\vec{\beta}_g$ for differentially expressed and non-differentially expressed genes, i.e. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_1}$ for differentially expressed genes and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_0}$ for non-differentially expressed genes.

Adopting the separation strategy on modelling covariance matrices by Barnard et al. (2000), we propose independent prior distributions on the diagonal variance components and the off-diagonal correlation matrix for all four covariance matrices that were mentioned above. Let $[\boldsymbol{\rho}_{(1)kk'}]_1^K$, $[\boldsymbol{\rho}_{(0)kk'}]_1^K$, $[\boldsymbol{r}_{kk'}]_1^K$ and $[\boldsymbol{t}_{kk'}]_1^K$ denote the correlation matrices corresponding to the covariance matrices $\boldsymbol{\Sigma_1}$, $\boldsymbol{\Sigma_0}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Pi}$ respectively, and let $[\boldsymbol{\sigma}^{\mathbf{2}}_{(1),k}]_1^K$, $[\boldsymbol{\sigma}^{\mathbf{2}}_{(0),k}]_1^K$, $[\boldsymbol{\tau}^{\mathbf{2}}_{k}]_1^K$ and $[\boldsymbol{\xi}^{\mathbf{2}}_{k}]_1^K$ denote the corresponding diagonal matrices with the variance terms on the diagonal. It is widely known that:

$$\boldsymbol{\Sigma_1} = ([\boldsymbol{\sigma}^{\mathbf{2}}_{(1),k}]_1^K)^{1/2}[\boldsymbol{\rho}_{(1)kk'}]_1^K([\boldsymbol{\sigma}^{\mathbf{2}}_{(1),k}]_1^K)^{1/2},$$

$$\boldsymbol{\Sigma_0} = ([\boldsymbol{\sigma}^{\mathbf{2}}_{(0),k}]_1^K)^{1/2}[\boldsymbol{\rho}_{(0)kk'}]_1^K([\boldsymbol{\sigma}^{\mathbf{2}}_{(0),k}]_1^K)^{1/2},$$

$$\boldsymbol{\Lambda} = ([\boldsymbol{\tau}^{\mathbf{2}}_{k}]_1^K)^{1/2}[\boldsymbol{r}_{kk'}]_1^K([\boldsymbol{\tau}^{\mathbf{2}}_{k}]_1^K)^{1/2},$$

$$\boldsymbol{\Pi} = ([\boldsymbol{\xi}^{\mathbf{2}}_{k}]_1^K)^{1/2}[\boldsymbol{t}_{kk'}]_1^K([\boldsymbol{\xi}^{\mathbf{2}}_{k}]_1^K)^{1/2}.$$

For each variance component, we propose a Jeffrey's prior, i.e.

$$\sigma^2_{(1),k} \propto \frac{1}{\sigma^2_{(1),k}}, \quad \sigma^2_{(0),k} \propto \frac{1}{\sigma^2_{(0),k}}, \quad \tau^2_k \propto \frac{1}{\tau^2_k}, \quad \xi^2_k \propto \frac{1}{\xi^2_k}.$$

For the correlation matrices, we propose an inverse Wishart prior distribution with identity matrix as its scale matrix and $v = K+1$ degrees of freedom, which is equivalent to putting a uniform prior on each element of the correlation matrices marginally (Gelman et al., 2014; Scharpf et al., 2009; Barnard et al., 2000), more specifically we have

$$[\boldsymbol{\rho}_{(1)kk'}]_1^K, [\boldsymbol{\rho}_{(0)kk'}]_1^K, [\boldsymbol{r}_{kk'}]_1^K, [\boldsymbol{t}_{kk'}]_1^K \sim W^{-1}(\boldsymbol{I}, v).$$

For gene-specific grand means λ_g, η_g and m_g, we assume that they follow normal priors, e.g. $\lambda_g \sim N(\mu_\lambda, \sigma^2_\lambda)$, $\eta_g \sim N(\mu_\eta, \sigma^2_\eta)$, $m_g \sim N(\mu_m, \sigma^2_m)$ with mean $\mu_\lambda = 0$, $\mu_\eta = 0$, $\mu_m = 0$, and variance $\sigma^2_\lambda = 10^2$, $\sigma^2_\eta = 10^2$, $\sigma^2_m = 10^2$. We performed sensitivity analysis on the hyperparameter values, since the variance σ^2_λ, σ^2_η and σ^2_m are fairly large; the results show little change when the means μ_λ, μ_η and μ_m change (see Appendix for the result of a sensitivity analysis on hyperparameter μ_η).

In addition to the informative parameters listed above, we introduce one supporting parameter ω_{gik} into the model to help to obtain closed form posterior distributions for β_{gk} and α_{gk} by exploiting conditional conjugacy (Polson et al., 2013; Zhou et al., 2012). The prior for ω_{gik} is specified as:

$$\omega_{gik} \sim PG(y_{gik} + \phi_{gk}^{-1}, 0),$$

where PG refers to the Polya-gamma distribution: details about this distribution and how the supporting parameter facilitates conditional conjugacy are provided in the Appendix. The closed-form posterior distribution for β_{gk} and α_{gk} by conditional conjugacy speeds up MCMC simulation.

2.4 Model-based clustering to categorize DE genes

We next utilize the DE indicators δ_{gk} to cluster the differentially expressed genes and to model the homogeneous and heterogeneous differential signals across studies. Since clustering based on the binary latent variable is unstable and does not take effect size into consideration, we first transform the binary vector into a standard normal vector and use a Dirichlet process Gaussian mixture model to cluster the differentially expressed genes, following Medvedovic et al. (2004). Suppose that $P(\delta_{gk} = 1) = \pi_{gk}$ is the prior probability that a gene g is differentially expressed in study k, the effect size is used to turn π_{gk} into a signed probability measure $\pi_{gk}^{\pm} = \pi_{gk} \times \text{sign}(\beta_{gk})$ where sign(.) is the sign function. We further rescale $\pi_{gk}^{*} = (\pi_{gk}^{\pm}+1)/2$, so the score falls in the range $[0, 1]$. Lastly, we transform π_{gk}^{*} to a Z-score $z_{gk} = \Phi^{-1}(\pi_{gk}^{*})$ where Φ is the standard normal cumulative distribution function. Following Ferguson (1983) and Neal (2000), we construct a DPM framework to cluster the differentially expressed genes:

$$\begin{aligned} \vec{z}_g | c_g, \boldsymbol{\theta} &\sim F(\vec{\theta}_{c_g}), \\ P(c_g = c) &= p_c, \\ \vec{\theta}_c &\sim G_0, \\ \vec{p} &\sim Dirichlet(a/C, \ldots, a/C). \end{aligned} \tag{7}$$

where $\vec{z}_g = (z_{g1}, \ldots, z_{gK})^T$ and c_g indicates the "latent cluster" for gene g, $F(.)$ is a mixture of K-dimensional multivariate Gaussian distributions with mean $\vec{\theta}_c$ and covariance matrix the identity matrix. C is the number of clusters, which is stochastic and allowed to go to ∞ under DPM. G_0 is the base distribution: in this case, $G_0 = N_K(\vec{0}, \boldsymbol{I})$ and $\vec{p} = (p_1, \ldots, p_C)$ is the mixing proportions for the clusters. a/C is the concentration parameter. In our model, we specify $a = C$ so the marginal prior distribution of each mixing proportion p_c would be Unif(0,1) under the constraint $\sum\limits_{c=1}^{C} p_c = 1$.

The above descriptions fully define the hierarchical Bayesian model that is proposed. The observed data are the raw counts, the library size and the phenotypic indicator $\{y_{gik}, T_{ik}, X_{ik}\}$, the parameters that we need to update through sampling include δ_{gk}, β_{gk}, α_{gk}, ϕ_{gk}, λ_g, η_g, m_g, σ_k^2, τ_k^2, ξ_k^2, $\rho_{kk'}$, $r_{kk'}$, $t_{kk'}$, ω_{gik}, c_g and C. The hyperparameters that we prespecify include $v = K + 1$, $\mu_\lambda = 0$, $\mu_\eta = 0$, $\mu_m = 0$, $\sigma_\lambda^2 = 10^2$, $\sigma_\eta^2 = 10^2$, $\sigma_m^2 = 10^2$ and $C_{\text{init}} = 10$.

2.5 Simulating posterior distribution via MCMC

We use the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) as well as the Gibbs sampling algorithm (Geman and Geman, 1984) to infer the posterior distribution of the parameters. Depending on the form of the distribution, five types of mechanisms are proposed to update the 16 groups of parameters.

1. The full conditionals for α_{gk} and β_{gk} are bivariate normal with known $\vec{\omega}_{gk}$. The full conditional for $\vec{\omega}_{gk}$ is Polya-gamma distribution with known α_{gk} and β_{gk} (Polson et al., 2013; Zhou et al., 2012). We use Gibbs sampling to update them sequentially for each gene g in study k.

2. The full conditionals for λ_g, η_g and m_g are multivariate Gaussian distribution for each gene g. The full conditional for each element in $[\boldsymbol{\sigma_k^2}]_1^K$, $[\boldsymbol{\tau_k^2}]_1^K$ and $[\boldsymbol{\xi_k^2}]_1^K$ is an inverse gamma distribution. The full conditionals for $[\boldsymbol{\rho}_{kk'}]_1^K$, $[\boldsymbol{r}_{kk'}]_1^K$ and $[\boldsymbol{t}_{kk'}]_1^K$ are inverse Wishart distributions. For all the above with closed form conditional distributions, we use Gibbs sampling to update them.

3. For $\vec{\phi}_g$, we propose a Metropolis-Hastings algorithm to update it for each gene g. In particular, we sample a new value of $\vec{\phi}_g$ from a multivariate log-normal jump distribution with mean equal to the old value and covariance matrix equal to $\mathbf{\Pi}$. The acceptance ratio r is defined as the ratio of two posterior density functions, and the new value is accepted with probability $\min[1, r]$.

4. For the pair $(\beta_{gk}, \delta_{gk})$, since the support for β_{gk} depends on δ_{gk}, we use a reversible jump MCMC to update them jointly (Green, 1995; Lewin et al., 2007). First, a potential new value of δ_{gk} is proposed by inverting the current value, i.e. $\tilde{\delta}_{gk} = 1 - \delta_{gk}$ and a new update $\tilde{\beta}_{gk}$ is then sampled from the associated full condition given $\tilde{\delta}_{gk}$. We define the ratio of the two

joint posterior distributions as r and jointly accept the new proposed values $(\tilde{\beta}_{gk}, \tilde{\delta}_{gk})$ with probability $\min[1, r]$.

5. Since our DPM model is in a conjugate context, to update the cluster assignment c_g, we follow algorithm 3 in Neal (2000) to draw a new value from $c_g | c_{-g}, \vec{z}_g$ for $g = 1, \ldots, G$ at each iteration, where c_{-g} is the cluster assignment of all genes other than g. The number of clusters C is updated at each iteration on the basis of c_g.

The detailed updating functions and algorithms for each group of parameters are described in the Appendix. For both simulation and real data, we ran 10,000 MCMC iterations. The selected traceplots (see Appendix) from simulation I below showed that all parameters reached convergence after relatively small number of iterations (roughly 3000). In light of this, the first 3000 iterations were dropped as burn-in period in all later analyses. The remaining 7000 of 10000 iterations are used for inference.

3. BAYESIAN INFERENCE AND CLUSTERING

3.1 Bayesian inference and control of false discovery rate

In the Bayesian literature, Newton et al. (2004) proposed a direct approach to control the false discovery rate (FDR) and defined a Bayesian FDR as:

$$\text{BFDR}(t) = \frac{\sum_{g=1}^{G} P_g(H_0|D) d_g(t)}{\sum_{g=1}^{G} d_g(t)},$$

where $P_g(H_0|D)$ is the posterior probability that gene g is non-differentially expressed (H_0) given data (D) and $d_g(t) = I\{P_g(H_0|D) < t\}$. The tuning parameter t can be tuned to control the BFDR at a certain α level. Throughout this paper, the Bayesian FDR BFDR will be used to address the multiplicity issue for the Bayesian method so that it is comparable with FDR control from the two-stage methods.

For fair comparison with the Fisher's method in meta-analysis, we adopt a union-intersection (UIT) hypothesis (also known as a conjunction hypothesis) setting following Li et al. (2011): $\mathrm{H}_0 : \bigcap\{\beta_k = 0\}$ vs $\mathrm{H}_a : \bigcup\{\beta_k \neq 0\}$, i.e. reject the null when the gene is differentially expressed in at least one study, where β_k is the effect size of study k, $1 \leq k \leq K$. Correspondingly, we define a null set $\Omega^0 = \{\vec{\beta}_g : \sum_{k=1}^{K} I(\beta_{gk} \neq 0) = 0\}$ and the respective DE set $\Omega^1 = \{\vec{\beta}_g : \sum_{k=1}^{K} I(\beta_{gk} \neq 0) > 0\}$. To control BFDR at the gene level, we introduce a Bayesian equivalent q-value. From the Bayesian posterior, we can calculate the probability that each gene falls in the null space: $\hat{P}_g(H_0|D) = \hat{P}(\vec{\beta}_g \in \Omega^0|D) = \frac{\sum_{t=1}^{T} I\{\vec{\delta}_g^{(t)} = \vec{0}\}}{T}$, where T is the total number of MCMC samples and $\vec{0}$ is a K-dimensional zero vector. We then define the Bayesian q-value of gene g as $q_g = \min_{t \geq \hat{P}_g(H_0|D)} \text{BFDR(t)}$. This q_g will be treated similarly as q-value in the frequentist approach by Fisher's method. Aside from detection of a differentially expressed gene list from meta-analysis, the posterior mean of δ_{gk}, $\text{E}(\delta_{gk}|D)$, can be used to infer DE for gene g in study k.

3.2 Summarization of clustering posterior to categorize differentially expressed genes

Addressing DE in multiple studies is more difficult than that in a single study because the gene may be concordantly or discordantly (up-regulated in some studies but not in the others) differentially expressed. The Bayesian method proposed is based on effect size; thus it would favor differentially expressed genes concordant across studies. Following Section 2.4, we use the posterior estimate of π_{gk} as an indicator of cross-study DE pattern to cluster the differentially expressed genes. To stabilize the estimation, we estimated π_{gk} by non-overlapping windows of every 20 MCMC simulations, i.e. $\hat{\pi}_{gk}^{(b)} = \sum_{t^{(b)}=1}^{20} \delta_{gk}^{t^{(b)}}/20$, for the bth simulation and then transformed into $\hat{z}_{gk}$ as in Section 2.4. After each chain of 20 simulations, the cluster assignment c_g is updated from the DPM model. At the end of all chains, to summarize the posterior estimates of c_g, we follow Medvedovic et al. (2004) and Rasmussen et al. (2009) and calculate the co-occurrence probability $p_{g,h}$ for any two genes g and h as the number of times that the two genes are assigned to the same cluster divided by the total number of assignments. Then we use $1 - p_{g,h}$ as a dissimilarity measure to further cluster the genes using consensus clustering (Monti et al., 2003). Consensus clustering is a stable clustering method by summarizing hierarchical clustering results with Ward linkage in repeated subsampling. The default consensus clustering method does not allow scattered genes (i.e. genes not belonging to any cluster) but one can apply other methods such as tight clustering for that (Tseng and Wong, 2005). As a result, genes with similar DE patterns over the chains are grouped together, while those with very different cross-study DE patterns will be separated.

3.3 Methods for comparison

Since other existing Bayesian methods in RNA-seq DE analysis such as “baySeq” and “EBSeq” have been developed for a single study (Hardcastle and Kelly, 2010; Leng et al., 2013), they cannot be immediately extensible to a meta-analysis framework and compare with our method. Thus, we shall compare our method with selected two-stage approaches that have been frequently adopted in the literature so far. For the first stage of single-study DE analysis of RNA-seq, we shall compare with the two most popular tools edgeR and DESeq (Robinson et al., 2010; Anders and Huber, 2010). For meta-analysis, since no other methods have been proposed specifically for RNA-seq, Fisher’s method will be applied to combine edgeR or DESeq p-values from multiple RNA-seq studies (Fisher, 1925). The meta-analysed p-values are then adjusted for multiple comparison by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). In this paper, we shall compare BayesMetaSeq with the two-stage edgeR-Fisher and DESeq-Fisher approaches.

4. SIMULATION

We performed three types of simulation to compare BayesMetaSeq, edgeR-Fisher and DESeq-Fisher. Details are described below.

4.1 Simulating homogeneous study effects to assess power and accuracy (simulation I)

In the first part of the simulation, we assessed the performance of BayesMetaSeq for genes with low, medium and high read counts when the effects were homogeneous across all studies. We simulated expression counts of $G = 1000$ genes for $K = 2, 5$ studies, $N_k = 10$ (five cases and five controls), $1 \leq k \leq K$. Library sizes for all samples were sampled from 0.3 to 0.5 millions so the average counts range roughly from 300 to 500. Baseline expressions were either high ($\alpha_{gk} \sim \text{Unif}\{-5.5, -4.5\}$; mean counts $\sim$ 1500-4500), medium ($\alpha_{gk} \sim \text{Unif}\{-8.5, -6.5\}$; mean counts $\sim$ 80-600) or low means ($\alpha_{gk} \sim \text{Unif}\{-11, -9\}$; mean counts $\sim$ 5-50). The log-scaled dispersions were generated accordingly ($\log(\phi_{gk}) \sim \text{Unif}\{-3.5, -2.5\}$ for high mean counts, $\log(\phi_{gk}) \sim \text{Unif}\{-2.5, -1.5\}$) for medium mean counts and $\log(\phi_{gk}) \sim \text{Unif}\{-1.5, -0.5\}$ for low mean counts), assuming genes with larger means had smaller dispersion (Anders et al., 2013). We let the first 20% genes ($N = 200$) be differentially expressed in all studies; among them, half were generated from high means and the other half from the low means. The rest of the genes ($N = 800$) were non-differentially expressed, a quarter of them were generated from high means, a quarter from the medium means, and the other half from the low means. For differentially expressed genes, the effect size β_{gk} was drawn from Unif$\{0.8, 2\}$ or Unif$\{-2, -0.8\}$ (a positive or negative log-fold change respectively). For non-differentially genes, β_{gk} was drawn from $N(0, 0.5^2)$. We repeated the above parameter sampling for all the K studies. Under the same homogeneous scenario, we also repeated the above simulations for weaker DE signals (Simulation IB), i.e. log-scaled effect size β_{gk} was drawn from Unif$\{0.7, 1.5\}$ or Unif$\{-1.5, -0.7\}$ for differentially expressed genes and from $N(0, 0.7^2)$ for non-differentially expressed genes.

4.2 Simulating heterogeneous study effects to assess power and accuracy (simulation II)

In the second part of simulation, we assessed the performance of BayesMetaSeq when the effects were heterogeneous in different studies. We simulated expression counts of $G = 1000$ genes for $K = 2, 5$ studies with $N_k = 10$, $1 \leq k \leq K$. The library size, baseline expression and the corresponding log-scaled dispersion were generated in the same way as in simulation I. We assumed that the first 30% of genes ($N = 300$) were differentially expressed. For $K = 2$, two-thirds of those genes are differentially expressed only in the first study or the second study, and a third are common DE; for $K = 5$, a third of those genes are differentially expressed only in one study, a third are differentially expressed only in two studies, and a third are differentially expressed in more than two studies. Similar to the previous simulation, half of the differentially expressed genes were from high means and half from the low means. The other 70% of genes ($N = 700$) were non-differentially expressed, a quarter of them were generated from high means, a quarter from the medium means, and half from the low means. For differentially expressed genes, the effect size β_{gk} was drawn from Unif$\{1, 2.5\}$ or Unif$\{-2.5, -1\}$, however, no discordance was allowed. For non-differentially genes, β_{gk} was drawn from Unif$\{-0.3, 0.3\}$.

4.3 Simulating cross-study differential patterns to evaluate differentially expressed gene clustering (simulation III)

In the fourth part of simulation, we assessed the clustering performance of BayesMetaSeq when the DE genes were generated from varying cross-study differential patterns. We simulated expression counts of $G = 1000$ genes for $K = 3$ with $N_k = 10$, $1 \leq k \leq K$. Library size was generated in the same way as in Simulation I. The baseline expression α_{gk} was drawn from Unif$\{-8.5, -6.5\}$ (mean counts $\sim$ 80-600) and the dispersion parameter ϕ_{gk} was drawn from Unif$\{-2.5, -1.5\}$. We assumed that the first 30% of genes ($N = 300$) were differentially expressed in at least 2 studies. Among them, a sixth were up-regulated in all studies ("+++"), a sixth were down-regulated in all studies ("- - -"), the other two-thirds were either up-regulated or down-regulated in two studies but non-differentially expressed in the third study (e.g. 50 genes with the pattern "++0", 50 genes with the pattern "- - 0", 50 genes with the pattern "+0+", 50 genes with the pattern "- 0 -"). For differentially expressed genes, the effect size β_{gk} was drawn from $N(2, 0.5^2)$ or $N(-2, 0.5^2)$ (up-regulated or down-regulated, respectively). For non-differentially expressed genes, β_{gk} was 0. For comparison with the other methods (edgeR-Fisher and DESeq-Fisher), we assessed both power and accuracy by plotting the number of true positives against the top number of declared differentially expressed genes, as well as the receiver operating characteristic (ROC) curves respectively for each method.

4.4 Simulation I and II

The posterior means and standard errors of selected parameters were summarized and compared with their true values from simulation IA as shown in the Appendix. The result demonstrated validity of BayesMetaSeq. In simulation IA of homogeneous study effects, we found that BayesMetaSeq was more powerful and accurate than the edgeR-Fisher and DESeq-Fisher methods in low mean counts whereas they performed almost equally well in high means counts (for simplicity, we combined both high mean and medium mean in this group), as shown in Figure 2(A). Comparing with the other two methods, only BayesMetaSeq had area under the curve (AUC) above 0.9 in the low mean region with both high sensitivity and high specificity. As the number of study K increased, we saw more noticeable advantage of the Bayesian method over the other methods in detecting differentially expressed genes with low means. Since the signals for high means were very strong, the three approaches performed almost perfectly even when $K = 2$. For simulation IB with weaker signals, the results were similar to simulation IA and the difference was more noticeable between BayesMetaSeq and the other two methods in the low mean region, whereas, for the high mean region, the performance for all three methods were alike (Figure 2(B)).

Similarly, in simulation II with heterogeneous study effects, though the overall signals became weaker, we found that BayesMetaSeq still performed better than the edgeR-Fisher and DESeq-Fisher methods in terms of both power and accuracy for low mean counts genes, whereas their performances were similar in high mean region, as shown in Figure 2(C). One thing to note here is that, even though the Bayesian method increased the power of detecting true DE signals in low mean regions, the detection power for low mean genes was still relatively weaker than high mean genes under the same scenario, owing to the inherent read count bias.

4.5 Simulation III

In simulation III, we found that BayesMetaSeq clearly identified the six clusters of differentially expressed genes with prespecified cross-study differential patterns (Figure 3 Left). Each of the six clusters corresponded to one particular cross-study differential pattern as reflected in the heatmap of signed $\mathrm{E}(\delta_{gk}|D)$ (Figure 3 Right), for example, cluster 1 included genes up-regulated in all studies and cluster 3 included genes down-regulated in all studies.

5. REAL DATA ANALYSIS

We applied BayesMetaSeq to a multiple brain region HIV-1 transgenic rat experiment (GSE47474) comparing the normal F344 strain and the HIV strain (Li et al., 2013). Samples from three brain tissues (hippocampus (HIP), striatum (STR), prefrontal cortex (PFC)) were sequenced and we regarded those as 3 studies to adopt our meta-analysis framework. There were 12 samples from each brain region in each strain ($N_1 = N_2 = N_3 = 24$; $K = 3$). The experiment was designed to determine differences in expression in brain regions of F344 and HIV-1 transgenic rats, in order to identify the mechanisms involved in HIV-1 neuropathology and develop efficient therapy for neuropsyhchiatric disorders associated with HIV-1 infection (Li et al., 2013). Following the guidance in edgeR (Robinson et al., 2010), we first filtered out genes with mean counts smaller than 1 in any study. After filtering, 10,280 genes remained for analysis. We applied BayesMetaSeq as well as edgeR-Fisher and DESeq-Fisher to the data. After we had obtained the DE genes from each approach, we performed pathway enrichment analysis by using Fisher's exact test based on the Gene Ontology (GO) database to annotate the identified genes (Khatri et al., 2012). In addition, we also analyzed the differentially expressed genes categories from BayesMetaSeq by using the Ingenuity Pathway Analysis (IPA) database for more biological insight. IPA is a commercial curated database

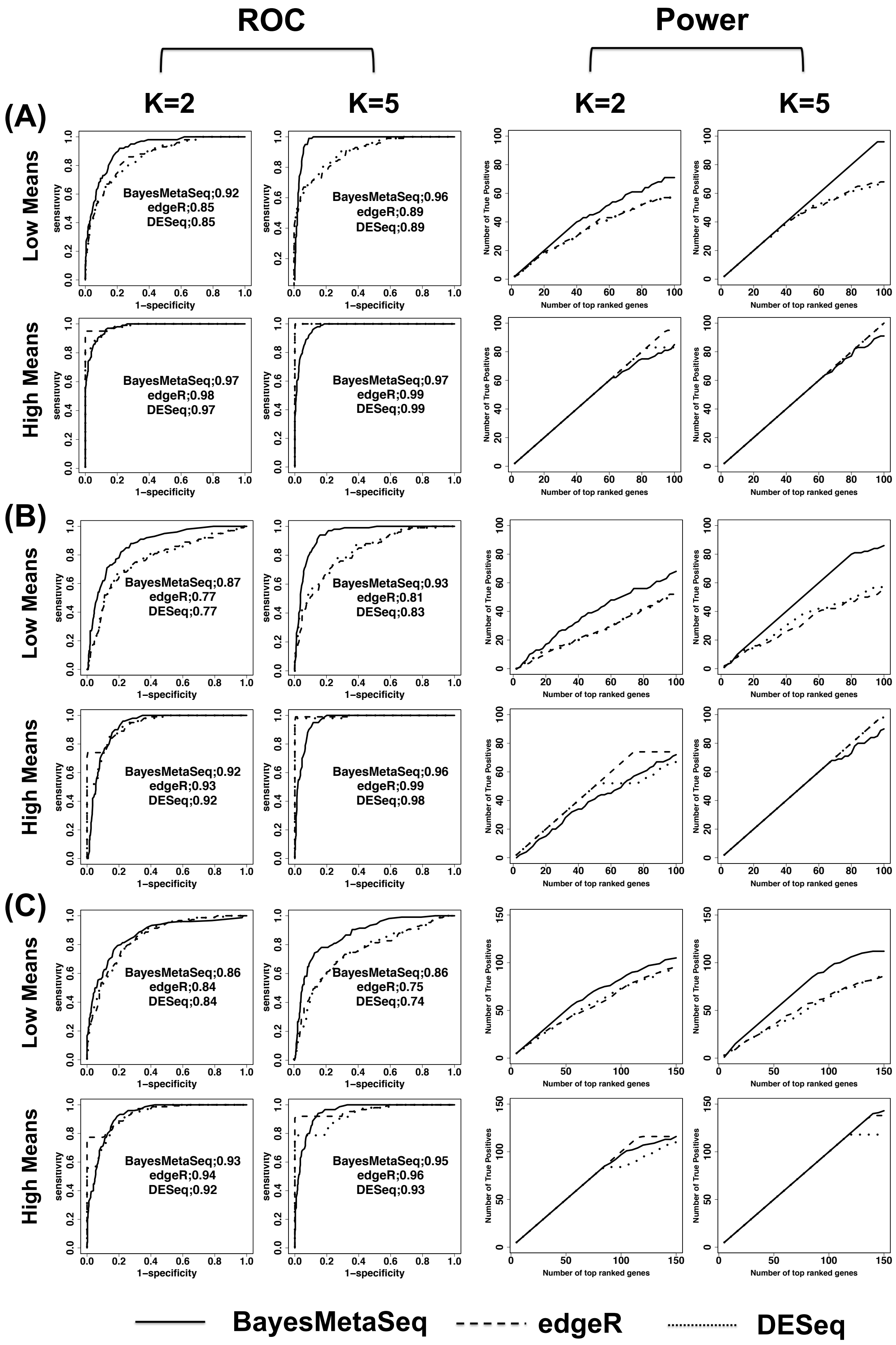

Figure 2: ROC Curve (left) and Power (right) comparison of BayesMetaSeq vs. edgeR-Fisher vs. DESeq-Fisher. (A) simulation IA; (B) simulation IB; (C) simulation II. The solid line is for BayesMetaSeq, the dashed line is for edgeR and the dotted line is for DESeq. The AUC values are attached to each ROC plot. For the power comparison, X axis refers to the top number of DE genes declared by each method, and Y axis refers to the number of true positives.

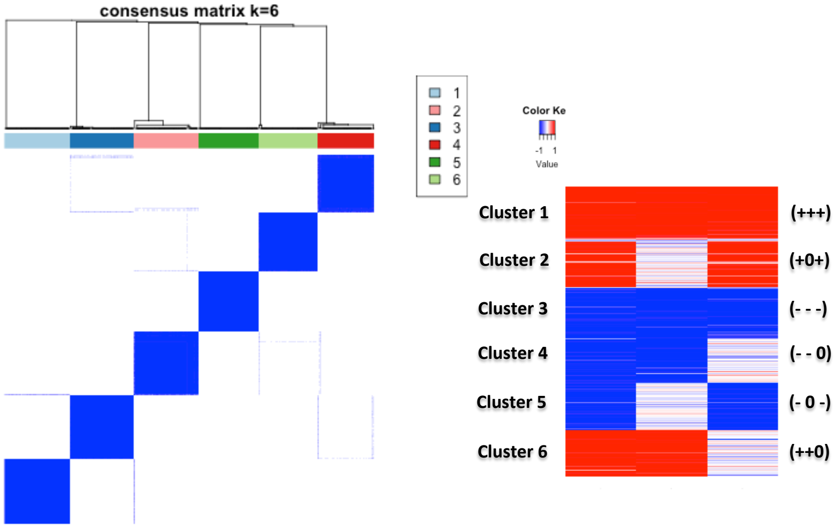

Figure 3: Simulation III. Left: Correlation heatmap of differentially expressed genes based on the co-occurrence probability $p_{g,h}$ with consensus clustering. Right: The heatmap of signed posterior mean of the DE latent indicator (i.e. $\mathrm{E}(\delta_{gk}|D)$ $\times\mathrm{sign}(\beta_{gk})$) in the six clusters.

that contains rich functional annotation, gene-gene interaction and regulatory information (IPA$^{®}$, QIAGEN Redwood City, www.qiagen.com/ingenuity).

5.1 Differential expression analysis

Controlling FDR at 0.1, edgeR-Fisher detected 51 DE genes and DESeq-Fisher 46 DE genes respectively, while BayesMetaSeq detected 245 DE genes (Table 1). A Venn Diagram showing the number of overlapping genes indicated good agreement among the three methods (see Appendix). As shown in Figure 4(A), the differentially expressed genes detected by BayesMetaSeq have wider detection range, especially for genes with smaller read counts, smaller RPKM (reads per kilobase per million) or shorter transcript length (Mortazavi et al., 2008). Table 2 lists three DE genes detected only by BayesMetaSeq but not by the other two methods. They typically have rare counts (a table and boxplots of normalized counts shown in the Appendix) due to short length of the transcripts (e.g. Mir212, Mir384) and/or small RPKM (e.g. Alb). microRNA-212 has been reported in previous studies to promote interleukin-17-producing T-helper cell differentiation (Nakahama et al., 2013). miRNA-384 has been found to regulate both amyloid precursor protein (APP) and β-site APP

Table 1: Comparison of 3 approaches in real rat data

Method	FDR at 0.05	FDR at 0.1
BayesMetaSeq	169	245
edgeR-Fisher	36	51
DESeq-Fisher	37	46

cleaving enzyme, which play an important role in the pathogenesis of Alzheimer's disease (Liu et al., 2014). Gene Alb encodes for albumin which is a primary carrier protein for steroids, fatty acids and steroid hormones in the blood, and has been used as markers of HIV disease progression in the highly active antiretroviral therapy (Shah et al., 2007).

5.2 Pathway enrichment analysis on detected differentially expressed genes

Detecting more differentially expressed genes does not necessarily indicate a better performance of our method. Since the underlying truth is not known in real data, we performed a pathway enrichment analysis on differentially expressed genes identified by each method. For fair comparison, we used the top 200 genes from each of the three methods and regarded them as differentially expressed genes in the pathway analysis. We tested on three pathway databases in MSigDB (`http://software.broadinstitute.org/gsea/msigdb`): GO, KEGG and Reactome, and only GO reported significant (q-value<0.05) pathways for all three methods. Controlling FDR at 0.05 by Benjamini-Hochberg correction, we found 50 GO pathways enriched with the DE genes from the BayesMetaSeq, while only 20 and 22 GO pathways were enriched for edgeR and DESeq, respectively. A cluster of enriched pathways was identified on the left of the Manhattan plot for BayesMetaSeq (circled), implying the enrichment in a major functional domain (Figure 4(B); pathways sorted by GO IDs). These pathways were mainly related to cell killing, leukocyte mediated cytotoxicity and T-cell mediated cytotoxicity (GO:0001906, GO:0001909, GO:0001910, GO:0001912, GO:0001913, GO:0001914, GO:0001916) and were enriched with BayesMetaSeq only (Table 3; p-values obtained from Fisher's exact test). The enrichment in these GO pathways might reflect changes in adaptive immune response against the HIV.

Table 2: Three example genes that show better detection power of BayesMetaSeq to detect low expressed or short length genes.

		edgeR		DESeq		BayesMetaSeq		Ave. normalized counts			
Gene	Study	p-value	Fisher's q-value	p-value	Fisher's q-value	Posterior means	Bayesian q-value	HIV strain	Normal strain	Ave. RPKM	Transcript length(bp)
Mir212	HIP	0.02	0.21	0.05	0.51	0.89	2e-3	2.08	3.92	20.27	23
	STR	0.02		0.03		0.99		2.20	4.93	21.31	
	PFC	0.07		0.09		0.83		2.99	5.08	22.56	
Mir384	HIP	0.06	0.39	0.10	0.86	0.88	8e-3	1.59	2.84	15.53	20
	STR	0.65		0.77		0.33		2.54	2.12	14.50	
	PFC	0.004		0.01		0.98		2.02	4.59	19.28	
Alb	HIP	0.002	0.10	0.06	0.88	0.95	6e-3	8.58	2.93	1.31	2676
	STR	0.61		0.60		0.41		9.17	7.65	1.67	
	PFC	0.006		0.03		0.99		20.09	10.90	2.89	

Table 3: Selected GO pathways enriched only with BayesMetaSeq from Figure 4(B). For fair comparison, the top 200 genes from each approach were regarded as differentially expressed genes and used for pathway analysis.

GO ID	GO Term	BayesMetaSeq p-value (logOR)	edgeR p-value (logOR)	DESeq p-value (logOR)
GO:0001906	cell killing	2.2e-4 (1.87)	0.033 (1.25)	0.12 (0.95)
GO:0001909	leukocyte mediated cytotoxicity	1e-3 (1.77)	0.105 (1.01)	0.102 (1.03)
GO:0001910	regulation of leukocyte mediated cytotoxicity	2.3e-4 (2.07)	0.056 (1.30)	0.055 (1.31)
GO:0001912	positive regulation of leukocyte mediated cytotoxicity	1.3e-4 (2.18)	0.04 (1.40)	0.043 (1.42)
GO:0001913	T cell mediated cytotoxicity	9.9e-5 (2.25)	0.039 (1.46)	0.038 (1.47)
GO:0001914	regulation of T cell mediated cytotoxicity	5e-5 (2.39)	0.029 (1.59)	0.028 (1.60)
GO:0001916	positive regulation of T cell mediated cytotoxicity	3.5E-5 (2.46)	0.024 (1.66)	0.24 (1.67)

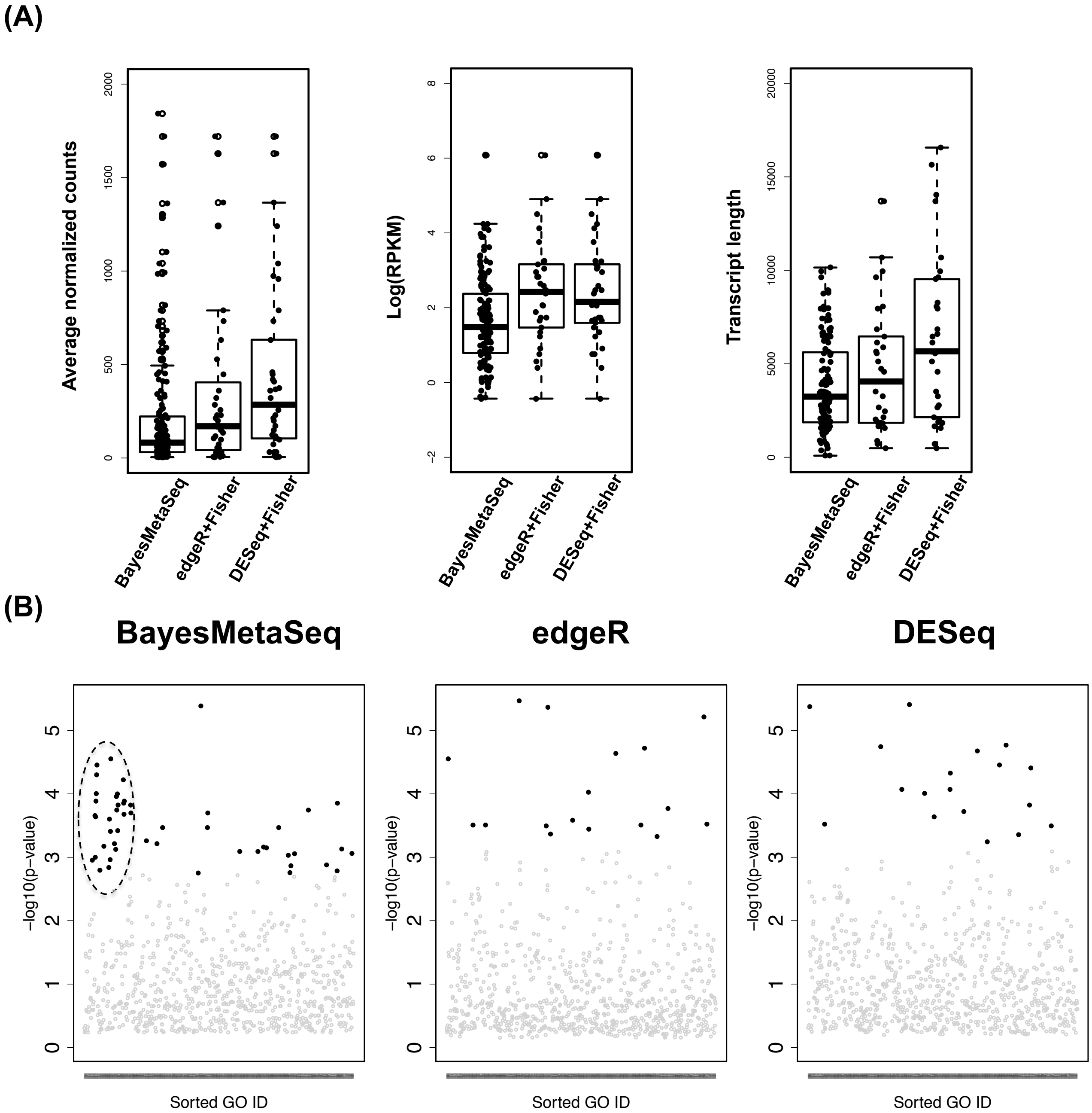

Figure 4: (A) Boxplot of average normalized counts, log(RPKM) and transcript lengths for the declared differentially expressed genes by each method. From left to right: BayesMetaSeq, edgeR-Fisher, DESeq-Fisher. (B) Manhattan plot of GO pathways enriched by the top 200 DE genes from each method. X axis refers to the GO pathways sorted by GO IDs, Y axis refers to the -log10(p-values) from the Fisher's exact test, the highlighted points are the GO pathways with FDR < 0.05.

5.3 Categorization of differentially expressed genes by study heterogeneity

We calculated the co-occurrence probability $p_{g,h}$ and used $1 - p_{g,h}$ as a dissimilarity measure to cluster the differentially expressed genes of BayesMetaSeq. As shown in Figure 5(A), we identified seven major clusters from the 245 differentially expressed genes. Each of the seven clusters corresponded to one particular cross-study differential patterns based on the signed $\mathrm{E}(\delta_{gk}|D)$ (Figure 5(B)). For example, genes in cluster 1 were up-regulated in all three studies and genes in cluster 5 were down-regulated only in STR, but not in HIP and PFC. Moreover, when we analyzed each cluster of genes separately through IPA pathway enrichment analysis, we noted that each cluster of genes represented different functional domains that were changed in the HIV strain as compared to the normal strain in different brain regions. For example, cluster 1 which included genes up-regulated in all three brain regions was mainly involved in antimicrobial response, whereas cluster 5 which included genes down-regulated in STR region only was mainly related to nervous system development (Figure 5(C)). Cluster 7 is not shown here since it included very few differentially expressed genes and only one enriched pathway was identified. A detailed list of significant pathways in each cluster with corresponding p-values and log-odds ratios can be found in the Appendix. In our analysis, we detected more region-specific DE markers (cluster 2-7) than common DE markers (cluster 1) which was consistent with the results reported from the original paper of this data (Li et al., 2013).

6. DISCUSSION AND CONCLUSION

In this paper, we proposed a Bayesian hierarchical model called BayesMetaSeq to conduct meta-analysis of RNA-seq data and biomarker categorization by study heterogeneity. Based on a negative binomial framework, the model assumed study-specific DE pattern for each gene and allowed shrinkage of multiple parameters. An MCMC algorithm was applied to update the posterior distribution of model parameters and the multiplicity issue was addressed by global FDR from a Bayesian perspective. A DPM model embedded in the Bayesian framework automatically clustered the detected biomarkers on the basis of cross-study differential patterns. Both the simulations and real rat data analysis showed that the Bayesian unified model was more powerful than two-stage methods (e.g. edgeR-Fisher and DESeq-Fisher), especially in lowly expressed genes without the loss of power

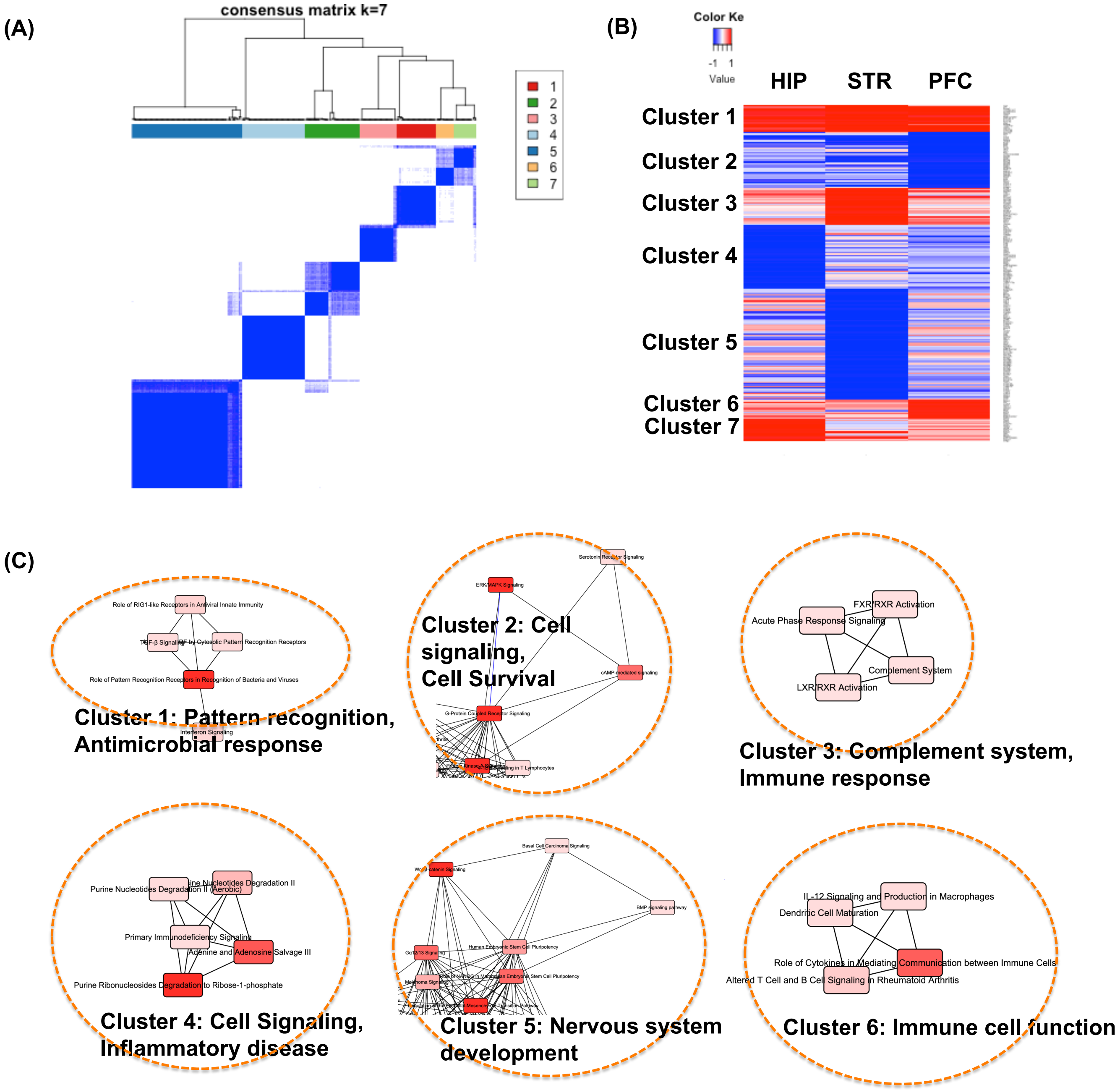

Figure 5: (A) Correlation heatmap of 245 Bayesian differentially expressed genes based on the co-occurrence probability $p_{g,h}$ with consensus clustering. (B) The heatmap of signed posterior mean of DE latent indicator (i.e.E$(\delta_{gk}|D)$ ×sign(β_{gk})) in the five major clusters. (C) A collection of overlapping IPA pathways enriched with each cluster of genes (deeper color refers to more significant pathways).

in highly expressed genes, and the FDR was well controlled. The differentially expressed genes identified by BayesMetaSeq between HIV strains and normal strains in the real data were further validated by pathway analysis and many differentially expressed genes were enriched in pathways related to immune response. Clustering analysis of the differentially expressed genes showed that genes with unique cross-study differential patterns were involved in specific functional domains such as antimicrobial response and inflammatory response.

Bayesian models have long been used in differential analysis of genomic studies such as microarray, RNA-seq and methylation (Hardcastle and Kelly, 2010; Leng et al., 2013; Van De Wiel et al., 2012; Chung et al., 2013; Park et al., 2014). Compared with other approaches, Bayesian methods can handle more complex generative mechanisms and allow the sharing of information across studies and across genes, both of which are essential for meta-analysis. Our unified Bayesian meta-analysis model increases the detection power for genes with low counts by accumulating small counts from multiple studies and encourages the sharing of information across different studies, which is not seen in the two-stage meta-analysis methods. In addition, the flexible and adaptable modelling of variance across samples in our approach also contributes to the improvement of detection power (Chung et al., 2013). A similar advantage of unified model over two-stage method has been seen in the categorical analysis literature where joint modelling of count data to combine multiple sparse contingency tables has been shown to be more powerful than traditional two-stage methods (Warn et al., 2002; Bradburn et al., 2007).

The current model relies on a fixed effects model, which assumes that differences of effect sizes are from sampling error alone. It can be readily extended to a random-effects scenario, where each effect size is assumed to be drawn from a study-specific distribution (Choi et al., 2003). Model checking can be performed to determine whether the fixed effect model or the random effect model is more adequate for a given dataset. Recent statistical research on RNA-seq proposed zero-inflated negative binomial model as an alternative to the regular negative binomial model and found that it fits better to real data since excessive zeros have always been observed in the next-generation sequencing (NGS) data (Van De Wiel et al., 2012). Our model can be easily extended to a zero-inflation framework, and its performance and computing feasibility for applications can be assessed through simulation or real data analysis. In our current approach, only a binary outcome is considered. The framework is applicable for a continuous outcome or multiclass outcome, where a dummy variable regression approach can be applied. Moreover, potential confounding covariates such as age, gender and other individual attributes can be included in the model.

Our real data application presents an example using the same RNA-seq platform across studies. In practice, it is possible that studies from different RNA-seq platforms are included and thus introduce significant bias. For example, the Sequencing Quality Control (SEQC) consortium performed extensive comparison on three RNA-seq platforms (Illumina HiSeq, Life Technology SOLiD and Roche 454) and determined pros and cons of different platforms (Consortium et al., 2014; Xu et al., 2013). As the end of July 2016, more than 95% of data in GEO used Illumina sequencing systems. As a result, unless different experimental protocols (e.g. mRNA preparation kits) are used in different studies, the platform bias in RNA-seq meta-analysis is not as severe as in microarray. We, however, acknowledge that platform bias may exist or may become more serious if new competing sequencing platforms become popular in the future. Practitioners should apply batch effect diagnostic or removal tools (Leek, 2014; Liu and Markatou, 2016), or extend with random effects in our model to account for cross-platform bias.

Currently, the Bayesian hierarchical model allows study-specific DE status, but favors concordant differential expression across studies. In some applications, discordant differentially expressed genes (e.g. a biomarker is up-regulated in one brain region but down-regulated in another brain region) may be expected and another hierarchical layer will be needed to accommodate. Another limitation of our method is the relatively high computational cost. To speed up the computation, we randomly partition the whole dataset into independent gene chunks and apply explicit parallelism using “snowfall” package in R (Knaus et al., 2009), while merging intermediate outputs for cluster analysis with all genes. It takes about 1 hour for 10000 MCMC iterations and 10280 genes with K=3 using 128 computing threads (8 CPUs each with Sixteen-core AMD 2.3GHz and 128GB RAM) in R code. Since the reduction of computing time is almost linear when more computing threads are used, we expect further computing time reduction when powerful computing clusters are used. Optimization of code in C++ and applying further parallel computing such as Consensus Monte Carlo Algorithm and Asynchronous Distributed Gibbs Sampling (Scott et al., 2013; Terenin et al., 2015) should further reduce computing time for general applications in the future. An R package, BayesMetaSeq, is publicly available to perform the analysis (`http://tsenglab.biostat.pitt.edu/software.htm`).

APPENDIX

Parameter estimation by Gibbs Sampling and the Metropolis-Hastings algorithm

In this section, we described the detailed updating conditional distributions or algorithms if there were no closed form conditional distributions for some parameters. The full conditional posterior is as follows:

$$\begin{aligned}P(-|Y_{gik}, T_{ik}, X_{ik}) \propto P(Y_{gik}|\alpha_{gk}, \beta_{gk}, \phi_{gk})\times \\ f(\alpha_{gk}|\eta_g, \tau_k^2, r) f(\beta_{gk}|\lambda_g, \delta_{gk}, \sigma_k^2, \rho) f(\phi_{gk}|m_g, \xi_k^2, t) \\ f(\eta_g|N(\mu_\eta, \sigma_\eta^2))(1/\tau_k^2) f(r|InvWishart(I, K+1)) \\ f(\lambda_g|N(\mu_\lambda, \sigma_\lambda^2))(1/\sigma_k^2) f(\rho|InvWishart(I, K+1)) \\ f(m_g|N(\mu_m, \sigma_m^2))(1/\xi_k^2) f(t|InvWishart(I, K+1)) \\ f(\delta_{gk}|\pi_{gk}) f(\pi_{gk}|\theta, c_g) f(\theta|G_0) f(c_g|p) f(p|a, C). \quad \text{(A.1)}\end{aligned}$$

To update each parameter, we simply integrate out the rest from the above.

Step 1

Gibbs sampling is used to update α_{gk}, β_{gk}. The two sets of parameters would be updated for each gene in each study, for simplicity, I will drop the suffix g and k here. The posterior distributions of these two parameters have closed form conditioning on the supporting parameter ω from the Polya-Gamma (PG) distribution. Following Polson et al. (2013), $\omega \sim PG(b, c)$ is an infinite convolution of gamma distributions defined as:

$$\omega \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{k - 1/2^2 + c^2/(4\pi^2)}.$$

where each $g_k \sim Gamma(b, 1)$ is an independent gamma random variable with $b > 0$, $c \in \Re$, and $\overset{D}{=}$ denotes equality in distribution.

The PG distribution has two important properties. Firstly, if $\omega \sim PG(b, 0)$, then by Laplace transform, we would have $E\{\exp(-\omega t)\} = cosh^{-b}(\sqrt{t/2})$, where $cosh(x) = \frac{e^x + e^{-x}}{2}$. Let $\omega \sim$

$PG(y+\phi^{-1},0)$, the negative binomial likelihood in terms of proportion p and dispersion ϕ can thus be expressed as:

$$L(p,\phi) \propto p^y(1-p)^{\phi^{-1}} = \frac{[\exp(\Psi)]^y}{[1+\exp(\Psi)]^{y+\phi^{-1}}} = \frac{2^{-(y+\phi^{-1})}\exp(\frac{(y-\phi^{-1})\Psi}{2})}{cosh^{y+\phi^{-1}}(\frac{\Psi}{2})}$$
$$\propto \exp(\frac{(y-\phi^{-1})\Psi}{2})E_\omega\{\exp(-\frac{\omega\Psi^2}{2})\}. \quad \text{(A.2)}$$

In other words, conditioning on ω, the above will end up with some negative quadratic form of Ψ (see equation (3) in the main text) within the exponential. Thus, the normal prior on Ψ would be a conjugate prior conditioning on ω. Let's go back to equation (3) in the main text, assume $\boldsymbol{B} = (\alpha,\beta)^T$ and $\boldsymbol{Z_i} = (1, X_i)^T$, then conditioning on known ω_i's, we know the likelihood of $\boldsymbol{B}$ is equal to:

$$L(\boldsymbol{B}) \propto \prod_{i=1}^{N} \exp\{-\frac{\omega_i}{2}(\boldsymbol{Z_i^T B} - (\frac{y_i-\phi^{-1}}{2\omega_i} - \log T_i))^2\}. \quad \text{(A.3)}$$

Let $\boldsymbol{\Omega} = diag(\omega_1,\ldots,\omega_n)$ and $u_i = \frac{y_i-\phi^{-1}}{2\omega_i} - \log T_i$, $\boldsymbol{U} = (u_1,\ldots,u_n)^T$, and we have the prior $\boldsymbol{B} \sim N(\boldsymbol{c},\boldsymbol{C})$, where $\boldsymbol{c} = (\eta,\lambda\delta)^T, \boldsymbol{C} = diag(\tau^2,\sigma^2)$, so the conditional posterior we used to update $\boldsymbol{B}$ would be:

$$(\boldsymbol{B}|-) \sim N(\boldsymbol{m},\boldsymbol{V}), \text{ where } \boldsymbol{V} = \boldsymbol{(Z\Omega Z^T + C^{-1})^{-1}}, \boldsymbol{m} = \boldsymbol{V(Z\Omega U + C^{-1}c)}. \quad \text{(A.4)}$$

Another important property of PG distribution is that any $PG(b,c)$ random variable ω has the following pdf (where the expectation in the denominator is taken w.r.t. $PG(b,0)$):

$$p(\omega|b,c) = \frac{\exp(-\frac{c^2}{2}\omega)p(\omega|b,0)}{E_\omega\{\exp(-\frac{c^2}{2}\omega)\}}$$

In other words, the posterior distribution of $\omega \sim PG(b,0)$ given c still belongs to the PG class (in our case, $b = y+\phi^{-1}, c = \Psi$):

$$P(\omega|\Psi) \propto \exp(-\frac{\omega\Psi^2}{2})PG(y+\phi^{-1},0) \propto PG(y+\phi^{-1},\Psi). \quad \text{(A.5)}$$

We can update each ω_i based on the above distribution using Gibbs sampling.

Step 2

For ϕ_g, since no closed form posterior distribution is available, we used Metropolis Hasting (MH) algorithm to update ϕ_g for all studies together. For each g, we proposed a new vector $\log(\vec{\phi}_{new}) =$

$(\log(\phi_1), \ldots, \log(\phi_K))^T$ from some jump distribution $N_K(\log(\vec{\phi}_{old}), \mathbf{\Pi})$. The proposal is accepted with probability $\min(1, r)$, where r is the acceptance ratio:

$$r = \frac{N_K(\log \vec{\phi}_{g_{new}}; m_g, \Pi) \prod_{k=1}^{K} \prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk} X_{ik}, \phi_{gk_{new}})}{N_K(\log \vec{\phi}_{g_{old}}; m_g, \Pi) \prod_{k=1}^{K} \prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk} X_{ik}, \phi_{gk_{old}})}. \tag{A.6}$$

If the proposal is accepted, we replace the old $\log(\vec{\phi})$ with the new one, otherwise, we keep the current value of $\log(\vec{\phi})$.

Step 3

We used Gibbs sampling to update λ_g, η_g, m_g based on their full conditional Gaussian distributions as follows:

$$\begin{aligned}
(\lambda_g|-) &\sim N_K(\lambda_\mu, \Sigma_\lambda),\ where\, \Sigma_\lambda = (diag(1/(\sigma^2_\lambda)) + K\Sigma^{-1})^{-1}, \lambda_\mu = \Sigma_\lambda(diag(1/(\sigma^2_\lambda))\vec{\mu}_\lambda + K\Sigma^{-1}\vec{\beta}_g) \\
(\eta_g|-) &\sim N_K(\eta_\mu, \Sigma_\eta),\ where\, \Sigma_\eta = (diag(1/(\sigma^2_\eta)) + K\Lambda^{-1})^{-1}, \eta_\mu = \Sigma_\eta(diag(1/(\sigma^2_\eta))\vec{\mu}_\eta + K\Lambda^{-1}\vec{\alpha}_g) \\
(m_g|-) &\sim N_K(m_\mu, \Sigma_m), \Sigma_m = (diag(1/(\sigma^2_m)) + K\Pi^{-1})^{-1}, m_\mu = \Sigma_m(diag(1/(\sigma^2_m))\vec{\mu}_m + K\Pi^{-1}\log\vec{\phi}_g)
\end{aligned} \tag{A.7}$$

To update λ_g, we will only use those β_{gk} for which $\delta_{gk} = 1$, if $\vec{\delta}_g = \vec{0}$, we would redraw from its prior $N(\mu_\lambda, \sigma^2_\lambda)$. Since we only need one value for each of the above parameters in every iteration, we took the average of each result.

Step 4

The full conditional for $[\boldsymbol{\sigma}^{\mathbf{2}}_{(1),k}]_1^K$, $[\boldsymbol{\sigma}^{\mathbf{2}}_{(0),k}]_1^K$, $[\boldsymbol{\tau}^{\mathbf{2}}_k]_1^K$ and $[\boldsymbol{\xi}^{\mathbf{2}}_k]_1^K$ have closed forms and are updated using Gibbs sampling for each k:

$$\begin{aligned}
\sigma^2_{(1),k} &\sim InvGamma(\frac{\sum_{g=1}^{G}\delta_{gk}}{2}, \frac{1}{2}\sum_{g=1}^{G}\delta_{gk}(\beta_{gk} - \lambda_g)^2) \\
\sigma^2_{(0),k} &\sim InvGamma(\frac{\sum_{g=1}^{G}(1-\delta_{gk})}{2}, \frac{1}{2}\sum_{g=1}^{G}(1-\delta_{gk})(\beta^2_{gk})) \\
\tau^2_k &\sim InvGamma(\frac{G}{2}, \frac{1}{2}\sum_{g=1}^{G}(\alpha_{gk} - \eta_g)^2) \\
\xi^2_k &\sim InvGamma(\frac{G}{2}, \frac{1}{2}\sum_{g=1}^{G}(\log\phi_{gk} - m_g)^2)
\end{aligned} \tag{A.8}$$

Step 5

The full conditional for $[\boldsymbol{\rho}_{(1)kk'}]_1^K$, $[\boldsymbol{\rho}_{(0)kk'}]_1^K$, $[\boldsymbol{r}_{kk'}]_1^K$, $[\boldsymbol{t}_{kk'}]_1^K$ have closed forms and are updated using Gibbs sampling:

$$\begin{aligned}\text{For } \vec{\delta}_g \neq 0 \ , \ [\boldsymbol{\rho}_{(1)kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^{K} (\bar{\beta}_k - \bar{\lambda})(\bar{\beta}_k - \bar{\lambda})^T, v = 2K+1) \\ \text{For } \vec{\delta}_g = 0 \ , \ [\boldsymbol{\rho}_{(0)kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^{K} (\bar{\beta}_k)(\bar{\beta}_k)^T, v = 2K+1) \\ [\boldsymbol{r}_{kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^{K} (\bar{\alpha}_k - \bar{\eta})(\bar{\alpha}_k - \bar{\eta})^T, v = 2K+1) \\ [\boldsymbol{t}_{kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^{K} (\log \bar{\phi}_k - \bar{m})(\log \bar{\phi}_k - \bar{m})^T, v = 2K+1)\end{aligned} \tag{A.9}$$

where the average is taken over all genes for β_k, λ, α_k, η, $\log \phi_k$ and m. After drawing a new covariance matrix from the above posterior, the actual correlation matrix can be obtained by integrating out the variance components.

Step 6

Since the support for β_{gk} depends on the choice of δ_{gk}, we use a reversible jump MCMC algorithm to update $(\delta_{gk}, \beta_{gk})$ together for each g and k. Specifically, a new value $\delta_{gk}^{new} = 1 - \delta_{gk}^{old}$ is proposed, and we then generate β_{gk}^{new} from the posterior in Step 1 based on δ_{gk}^{new}. The proposal is accepted with probability $\min(1, r)$, where r is the acceptance ratio:

$$r = \frac{N(\beta_{gk}^{new}; \delta_{gk}^{new} \lambda_g, \sigma^2) \prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk}^{new} X_{ik}, \phi_{gk})}{N(\beta_{gk}^{old}; \delta_{gk}^{old} \lambda_g, \sigma^2) \prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk}^{old} X_{ik}, \phi_{gk})} \tag{A.10}$$

We accept or reject the proposed values jointly from the above.

Step 7

Lastly, upon obtaining the updates of δ_{gk}, we can estimate π_{gk} for every 20 chains, and we transform it into z_{gk} through the steps described in Section 2.4. Based on the vector $\vec{z}_g$, we can update the cluster assignment c_g for each gene by Gibbs sampling using the following conditional probabilities:

$$\begin{aligned}\text{If } c = c_h \text{ for some } h \neq g : P(c_g = c | c_{-g}, \vec{z}_g) &= b \frac{n_c}{G - 1 + a} \int F(\vec{z}_g, \theta_c) dH_{-g,c}(\theta_c) \\ P(c_g \neq c_h \text{ for all } h \neq g | c_{-g}, \vec{z}_g) &= b \frac{a}{G - 1 + a} \int F(\vec{z}_g, \theta) dG_0(\theta)\end{aligned}$$

where $H_{-g,c}$ is the posterior distribution of θ_c based on the prior G_0 and all observations $\vec{z}_h$ for which $h \neq g$ and $c_h = c$, n_c is the cluster size of cluster c, b is the normalizing constant to make

Table S1: Comparison of posterior mean of the parameters estimated by BayesMetaSeq with their true values from Simulation IA, K=2

Parameters	True values	Posterior mean (SE)
β_0	0	-0.01 (0.42)
β_1^+	(0.8,2)	1.21 (0.50)
β_1^-	(-2,-0.8)	-1.25 (0.59)
α^{high}	(-8.5,-4.5)	-7.06 (1.32)
α^{low}	(-11,-9)	-11.17 (0.84)

Table S2: Sensitivity analysis on hyperparameter μ_η

Value of μ_η	α^{high} Posterior mean (SE)	α^{low} Posterior mean (SE)
0	-7.06 (1.32)	-11.17 (0.84)
-3	-7.05 (1.31)	-11.11 (0.79)
-5	-7.03 (1.31)	-11.10 (0.78)
-7	-7.04 (1.31)	-11.11 (0.78)

the probability sum to 1. More specifically, $\int F(\vec{z}_g, \theta_c) dH_{-g,c}(\theta_c) = f(\vec{z}_g; N_K(\mu_K = \frac{n_c}{n_c+1}\vec{z}_h, \Sigma = diag(\frac{n_c+2}{n_c+1}, K))$, $\int F(\boldsymbol{z}_g, \theta) dG_0(\theta) = f(\vec{z}_g; N_K(\mu_K = \mathbf{0}_K, \Sigma = diag(2, K))$.

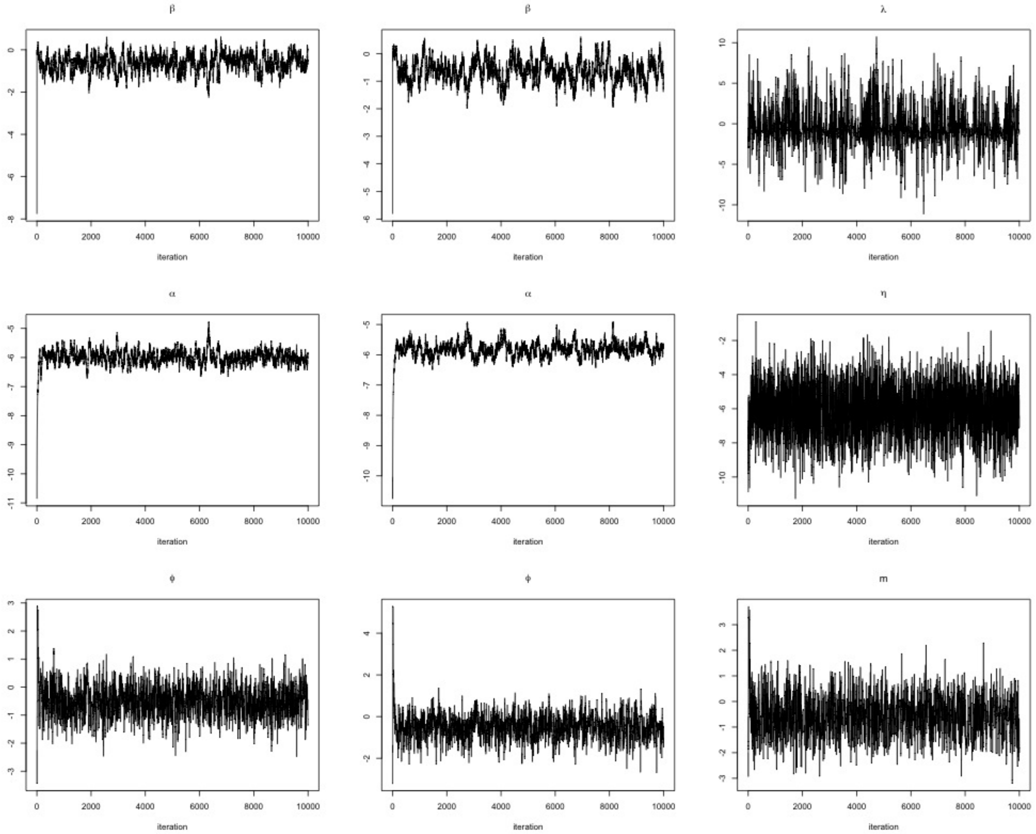

Figure S1: Traceplots of selected parameters from Simulation IA.

Table S3: Normalized counts (rounded) for the three genes shown in table 3.

Gene	Study	HIV strain	Normal strain
Mir212	HIP	(0,0,1,1,1,1,2,2,3,4,4,6)	(2,2,3,3,3,3,4,4,4,5,6,7)
	STR	(1,1,1,1,2,2,3,3,3,3,3,4)	(0,1,1,2,2,4,4,4,7,9,11,13)
	PFC	(0,1,1,1,2,2,3,3,4,5,6,7)	(0,1,2,4,5,6,6,6,7,7,8,9)
Mir384	HIP	(0,1,1,1,1,1,1,2,2,3,3,3)	(0,1,1,2,2,2,3,4,4,5,5,5)
	STR	(0,0,0,1,1,2,2,2,3,5,6,8)	(0,0,1,1,1,2,2,2,3,3,5,5)
	PFC	(1,1,1,1,1,1,2,2,3,3,4,4)	(2,2,3,3,4,4,5,5,7,7,7,7)
Alb	HIP	(3,4,4,4,5,5,5,6,6,8,13,41)	(1,2,2,2,2,2,3,3,3,3,4,8)
	STR	(0,3,3,4,4,4,8,11,14,18,20,21)	(1,1,2,3,3,4,7,8,11,12,13,26)
	PFC	(10,13,14,14,14,14,15,15,16,19,37,60)	(5,8,8,9,9,9,10,11,14,14,15,19)

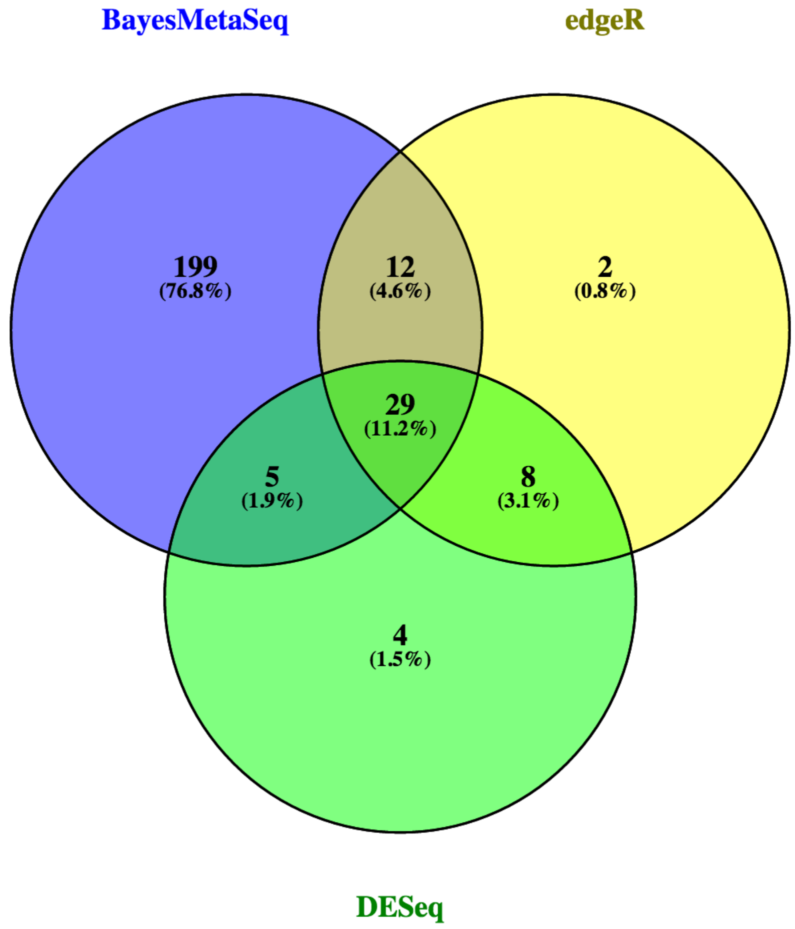

Figure S2: Venn Diagram of number of overlapping DE genes (FDR < 0.1) among the three methods applied in real data.

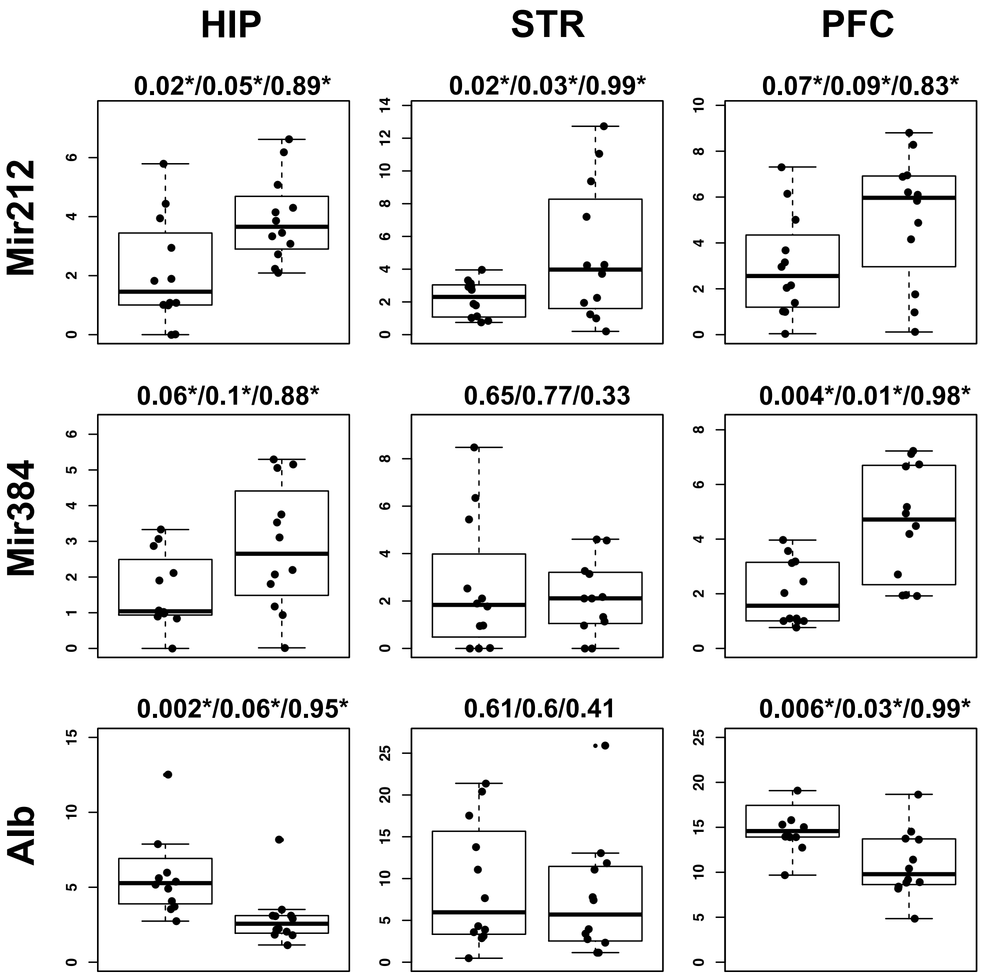

Figure S3: Distribution of normalized counts for the three genes shown in table 3 (left: HIV strain; right: Normal strain). The values above the boxplots correspond to the respective p-values or posterior means from edgeR/DESeq/BayesMetaSeq, with stars indicating the significance (e.g. p-value $\leq$0.1 or $E(\delta_{gk}|D) \geq 0.8$).

Table S4: List of significant IPA pathways (p-value < 0.05) from Cluster 1-4 in Figure 5.

Cluster	Pathway Name	p-value	logOR
Cluster 1	Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	0.003	3.79
	Role of JAK1, JAK2 and TYK2 in Interferon Signaling	0.017	4.67
	Allograft Rejection Signaling	0.023	4.33
	Autoimmune Thyroid Disease Signaling	0.023	4.33
	OX40 Signaling Pathway	0.030	4.08
	Role of RIG1-like Receptors in Antiviral Innate Immunity	0.031	4.03
	Interferon Signaling	0.033	3.98
	Activation of IRF by Cytosolic Pattern Recognition Receptors	0.046	3.60
Cluster 2	PI3K Signaling in B Lymphocytes	0.002	3.11
	G-Protein Coupled Receptor Signaling	0.002	2.55
	Protein Kinase A Signaling	0.003	2.45
	ERK/MAPK Signaling	0.005	2.72
	cAMP-mediated signaling	0.010	2.44
	Acute Phase Response Signaling	0.043	2.31
Cluster 3	Maturity Onset Diabetes of Young (MODY) Signaling	0.014	4.85
	Acyl-CoA Hydrolysis	0.016	4.67
	Stearate Biosynthesis I (Animals)	0.039	3.69
	Complement System	0.042	3.63
Cluster 4	Catecholamine Biosynthesis	0.008	5.99
	Adenine and Adenosine Salvage III	0.008	5.99
	Serotonin and Melatonin Biosynthesis	0.020	4.60
	Sphingomyelin Metabolism	0.020	4.60
	Adenine and Adenosine Salvage II	0.023	4.38
	Purine Nucleotides Degradation II (Aerobic)	0.035	3.91
	Tryptophan Degradation X (Mammalian, via Tryptamine)	0.046	3.59
	Primary Immunodeficiency Signaling	0.050	3.50

Table S5: List of significant IPA pathways (p-value < 0.05) from Cluster 5-7 in Figure 5.

Cluster	**Pathway Name**	**p-value**	**logOR**
Cluster 5	Regulation of the Epithelial-Mesenchymal Transition Pathway	7e-4	2.51
	Dendritic Cell Maturation	0.003	2.44
	Intrinsic Prothrombin Activation Pathway	0.003	3.77
	IL-4 Signaling	0.004	2.81
	Wnt/β-catenin Signaling	0.004	2.34
	Fc Epsilon RI Signaling	0.006	2.63
	Atherosclerosis Signaling	0.006	2.63
	Role of NANOG in Mammalian Embryonic Stem Cell Pluripotency	0.010	2.44
	Gα12/13 Signaling	0.010	2.44
	Human Embryonic Stem Cell Pluripotency	0.018	2.19
	Docosahexaenoic Acid (DHA) Signaling	0.018	2.78
	CTLA4 Signaling in Cytotoxic T Lymphocytes	0.020	2.74
	Melanoma Signaling	0.021	2.71
	Role of JAK1 and JAK3 in ?c Cytokine Signaling	0.024	2.64
	Virus Entry via Endocytic Pathways	0.026	2.58
	IL-15 Signaling	0.028	2.55
	Endometrial Cancer Signaling	0.029	2.52
Cluster 6	Role of Cytokines in Mediating Communication between Immune Cells	0.008	5.48
	Altered T Cell and B Cell Signaling in Rheumatoid Arthritis	0.029	4.16
Cluster 7	Serotonin Receptor Signaling	0.034	3.71

Acknowledgements

Research reported in this publication was supported by NCI of the National Institutes of Health under award number R01CA190766 to T.M. and G.C.T.

REFERENCES

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

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, 8(9):1765–1786.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312.

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)*, pages 289–300.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russell Localio, A. (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*, 26(1):53–77.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90.

Chung, L. M., Ferguson, J. P., Zheng, W., Qian, F., Bruno, V., Montgomery, R. R., and Zhao, H. (2013). Differential expression analysis for paired rna-seq data. *BMC bioinformatics*, 14(1):110.

Conlon, E. M., Song, J. J., and Liu, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC bioinformatics*, 7(1):247.

Consortium, S.-I. et al. (2014). A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24(1983):287–302.

Fisher, R. A. (1925). *Statistical methods for research workers.* Genesis Publishing Pvt Ltd.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

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.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. (2012). Metaqc: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic acids research*, 40(2):e15–e15.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375.

Knaus, J., Porzelius, C., Binder, H., and Schwarzer, G. (2009). Easier parallel computing in r with snowfall and sfcluster. *The R Journal*, 1(1):54–59.

Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., Farooqi, H. K., Liau, L. M., Horvath, S., Mischel, P. S., and Nelson, S. F. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC medical genomics*, 1(1):52.

Leek, J. T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research*, page gku864.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043.

Lewin, A., Bochkina, N., and Richardson, S. (2007). Fully bayesian mixture model for differential gene expression: simulations and model checks. *Statistical applications in genetics and molecular biology*, 6(1).

Li, J., Tseng, G. C., et al. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5(2A):994–1019.

Li, M. D., Cao, J., Wang, S., Wang, J., Sarkar, S., Vigorito, M., Ma, J. Z., and Chang, S. L. (2013). Transcriptome sequencing of gene expression in the brain of the hiv-1 transgenic rat. *PloS one*, 8(3):e59582.

Liu, C.-G., Wang, J.-L., Li, L., and Wang, P.-C. (2014). Microrna-384 regulates both amyloid precursor protein and β-secretase expression and is a potential biomarker for alzheimer's disease. *International journal of molecular medicine*, 34(1):160–166.

Liu, Q. and Markatou, M. (2016). Evaluation of methods in removing batch effects on rna-seq data. *Infectious Diseases and Translational Medicine*, 2(1):3–9.

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

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

Nakahama, T., Hanieh, H., Nguyen, N. T., Chinen, I., Ripley, B., Millrine, D., Lee, S., Nyati, K. K., Dubey, P. K., Chowdhury, K., et al. (2013). Aryl hydrocarbon receptor-mediated induction of the microrna-132/212 cluster promotes interleukin-17–producing t-helper cell differentiation. *Proceedings of the National Academy of Sciences*, 110(29):11964–11969.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

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

Oshlack, A., Wakefield, M. J., et al. (2009). Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, 4(1):14.

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). Methylsig: a whole genome dna methylation analysis pipeline. *Bioinformatics*, page btu339.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med*, 5(9):e184.

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

Rasmussen, C. E., De la Cruz, B. J., Ghahramani, Z., and Wild, D. L. (2009). Modeling and visualizing uncertainty in gene expression clusters using dirichlet process mixtures. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(4):615–628.

Rau, A., Marot, G., and Jaffrézic, F. (2014). Differential meta-analysis of rna-seq data from multiple studies. *BMC bioinformatics*, 15(1):91.

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):139–140.

Scharpf, R. B., Tjelmeland, H., Parmigiani, G., and Nobel, A. B. (2009). A bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, 104(488).

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E., and McCulloch, R. (2013). Bayes and big data: The consensus monte carlo algorithm. In *EFaBBayes 250 conference*, volume 16.

Shah, S., Smith, C., Lampe, F., Youle, M., Johnson, M., Phillips, A., and Sabin, C. (2007). Haemoglobin and albumin as markers of hiv disease progression in the highly active antiretrovial therapy era: relationships with gender*. *HIV medicine*, 8(1):38–45.

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., Miller, C. J., and Clarke, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1(1):42.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The american soldier: adjustment during army life.(studies in social psychology in world war ii, vol. 1.).

Terenin, A., Simpson, D., and Draper, D. (2015). Asynchronous distributed gibbs sampling. *arXiv preprint arXiv:1509.08999*.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, page gkr1265.

Tseng, G. C. and Wong, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61(1):10–16.

Tsuyuzaki, K. and Nikaido, I. (2013). metaseq: Meta-analysis of rna-seq count data.

Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2012). Bayesian analysis of rna sequencing data by estimating multiple shrinkage priors. *Biostatistics*, page kxs031.

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014). A comprehensive study design reveals treatment-and transcript abundance–dependent concordance between rna-seq and microarray data. *Nature biotechnology*, 32(9):926.

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

Warn, D., Thompson, S., and Spiegelhalter, D. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in medicine*, 21(11):1601–1623.

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

Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.