Published in final edited form as:

Stat Biosci. 2013 May 1; 5(1): 119–137. doi:10.1007/s12561-011-9052-3.

A Hierarchical Bayesian Model for Estimating and Inferring Differential Isoform Expression for Multi-Sample RNA-Seq Data

Saran Vardhanabhuti, Mingyao Li, and Hongzhe Li

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA, Tel.: 215-573-5038, Fax: 215-573-4865

Hongzhe Li: hongzhe@upenn.edu

Abstract

RNA-Seq has drastically changed our ways of studying transcriptomes in providing more precise estimates of gene expression, including isoform-specific expression. Most of the available methods for RNA-Seq data focus on one sample at a time. We present in this paper a Poisson-Gamma hierarchical model for multi-sample RNA-Seq data analysis in order to simultaneously estimate isoform-specific expression and to identify differentially expressed iso-forms. Our model has the advantage of borrowing information across all samples in estimating expression levels, which can improve the estimates drastically, particularly for low abundance isoforms. Furthermore, our hierarchical model has the ability to account for overdispersion in the data and also can incorporate sample-specific covariates in the underlying model, which facilitates the isoform-specific differential expression analysis. Simulation studies demonstrated that this Bayesian multi-sample approach can lead to more precise estimates of isoform-specific expression and higher power to detect differential expression by borrowing information across all samples than single sample analysis, especially for isoforms of low abundance. We further illustrated our methods using the RNA-Seq data of 10 Yoruban and 10 Caucasian individuals.

Keywords

Mixture of Poisson-Gamma model; Markov Chain Monte Carlo Sampling Next Generation Sequencing

1 Introduction

RNA-Seq data are millions of short cDNA reads with lengths between 25 to over 100 base pairs derived from fragmented mRNA sequence expressed in certain experimental conditions or tissues. These reads generated from the next generation sequencing (NGS) platforms allow one to interrogate transcriptome activities in the genome at higher resolution and wider dynamic range of expression compared to any existing methods such as microarray or EST-based sequencing (Nagalakshmi *and others*, 2008; Marioni *and others*, 2008; Mortazavi *and others*, 2008; Wang *and others*, 2008). Raw sequence reads are mapped to the genome and abundance of transcripts can be quantified using measure such as reads per kilobase per million mapped reads (RPKM) as described in Mortazavi *and others* (2008). It is relatively straightforward to obtain a gene level expression such as RPKM by summing read counts mapped to a gene normalized by the length of the gene and the total number of reads in a sample. Extension of RPKM to isoform-specific expression for a given gene is possible, however, uncertainties in estimating isoform expression arise when a gene has multiple isoforms, due to the fact that reads mapped to a certain exon in a gene maybe shared by two or more different isoforms. A few statistical methods have been developed to estimate the isoform-specific expression by using a likelihood-based approach and

importance sampling (Jiang and Wong, 2009), an EM-algorithm approach based on generative statistical model (Li *and others*, 2009) and recently a mixture-of-isoforms model based on Bayesian approach using Monte Carlo sampling (Katz *and others*, 2010). These methods, however, estimate isoform-specific expression for one sample at a time and do not provide inference on isoform-specific differential expression between groups of samples. Identification of isoforms that show differential expression between experimental groups is one of the most important problems in studying gene regulations.

In this paper, we present an integrated statistical approach for multiple RNA-Seq samples where isoform-specific variability and differential expression are modeled simultaneously. Our approach extends the Bayesian hierarchical Poisson regression model (Christiansen and Morris, 1997) that allows sharing of information across samples to obtain individual isoform-specific shrinkage estimates of expression and incorporation of sample-specific covariates in a regression framework. In contrast to a simple Poisson model of Jiang and Wong (2009) that only models RNA-Seq of one sample at a time, we model multi-sample RNA-Seq data by assuming a Poisson-Gamma hierarchical model which allows estimation of individual and isoform-specific expression and automatically accounts for overdispersion in the observed data. Compared to the Poisson hierarchical model of Christiansen and Morris (1997) where closed-form posterior distributions can be derived due to the use of conjugate priors, such conjugacy does not exist in our model due to the fact that observed reads can be derived from different isoforms. We estimate our model using the Markov chain Monte Carlo (MCMC) to obtain posterior distributions of parameters in the model. We exploit these posterior distributions to assess differential expression whereby the uncertainty in parameter estimates is incorporated in a natural manner. The proposed hierarchical model also facilitates sharing of information across different samples, which can lead to improved estimates of sample-specific gene expression of different isoforms.

The rest of the paper is organized as follows. We present our multi-sample Bayesian hierarchical model for inferring isoform-specific differential expression in Section 2. The model is justified with simulation results and model checks in Section 3. Applications to real RNA-Seq data analysis are given in Section 4. We conclude the paper with a brief discussion in Section 5.

2 A Hierarchical Bayesian Model for Multi-Sample RNA-Seq Data

2.1 Notations

Let $\mathbf{G}$ be the set of genes in the genome and $\mathbf{K}$ be the set of all possible isoforms of all the genes. In our modeling, we assume the set of genes and the set of all possible isoforms are the same across all individuals. In addition, we assume all the isoforms are known for the gene of interest. For a given gene $g \in \mathbf{G}$, there exists a set of isoforms $\mathbf{K}_g$ of size n_g . For isoform $k \in \mathbf{K}_g$ in the g^{th} gene, we let $m_{ik}^{(g)}$ be the number of isoform k transcripts in the i^{th} sample and $l_k^{(g)}$ be its length in kilobases. Based on this notation, for a given gene g , the probability of reads from isoform k in the i^{th} sample can be written as

$$p_{ik}^{(g)} = \frac{m_{ik}^{(g)} l_k^{(g)}}{\sum_{g \in \mathbf{G}} \sum_{k \in \mathbf{K}_g} m_{ik}^{(g)} l_k^{(g)}}.$$

Then the individual-specific isoform expression index, $\theta_{ik}^{(g)}$ can be expressed as,

$$\theta_{ik}^{(g)} = \frac{m_{ik}^{(g)}}{\sum_{g \in G} \sum_{k \in K_g} m_{ik}^{(g)} l_k^{(g)}} = \frac{p_{ik}^{(g)}}{l_k^{(g)}}.$$

Since our method analyzes each gene separately, to simplify the notation, we subsequently drop the index for gene g . The isoform expression indices for gene g in the i^{th} individual is then given by $\theta_{i1}, \theta_{i2}, \dots, \theta_{ik}, \dots, \theta_{iK}$ for $i = 1, 2, \dots, n$ where n is the number of samples.

2.2 Poisson-Gamma hierarchical model

For the observed data in a given gene, we have Y_{ij} as the number of reads for the i^{th} sample mapped to j^{th} region, where $j = 1, \dots, J$ represent exon or exon-exon junction regions. Let w_i be the total number of mapped reads for the i^{th} sample. We model the observed read counts data as Poisson distribution with parameter $\lambda_{ij} = w_i \sum_k p_{ik}$, thus we write $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$. We can then express λ_{ij} more precisely depending on whether reads fall within exons or exon-exon junctions. If reads fall within exon j , then $\lambda_{ij} = w_i \cdot l_j \cdot \sum_k c_{kj} \cdot \theta_{ik}$, where l_j is the length of the j^{th} exon and $c_{kj} = 1$ if isoform k contains exon j and 0 otherwise. On the other hand, if reads in region j belong to exon(p)-exon(q) junction, then $\lambda_{ij} = w_i \cdot l_j \cdot \sum_k c_{kp} \cdot c_{kq} \cdot \theta_{ik}$, where l_j is the length of exon p and q junction region, $c_{kp} = 1$ if isoform k contains exon p and 0 otherwise and $c_{kq} = 1$ if isoform k contains exon q and 0 otherwise. In general, λ_{ij} can be written as a linear combination of θ_{ik} as $\lambda_{ij} = \sum_k a_{ikj} \cdot \theta_{ik}$, where $a_{ikj} = w_i \cdot l_j \cdot c_{kj}$ and $c_{kj} = 1$ if isoform k contains region j . Based on the above notations, we can construct the structure of our hierarchical model which extended from Christiansen and Morris (1997) for mixture of Poisson data.

Level 1—The distribution of observed read counts Y_{ij} follows Poisson model (1) with θ_{ik} defined as a random variable and a_{ikj} as a known constant term,

$$Y_{ij} \sim \text{Poisson}(\sum_k a_{ikj} \cdot \theta_{ik}). \quad (1)$$

Level 2—The individual θ_{ik} parameter follows a gamma distribution,

$$\theta_{ik} \sim \text{Gamma}(\zeta_k, \frac{\mu_{ik}}{\zeta_k}). \quad (2)$$

The log link is assumed so that $\log(\mu_{ik}) = \mathbf{x}_i^T \boldsymbol{\beta}_k$, where $\mathbf{x}_i = (x_{i0}, \dots, x_{i,r-1})$ is a r -dimensional covariate vector associated with sample i and $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \dots, \beta_{k,r-1})$ are r unknown regression coefficients for specific isoform k . x_{i0} is taken to be one so β_{k0} is the intercept. Under this model, we have

$$\begin{aligned} E[\theta_{ik}] &= \zeta_k \cdot \frac{\mu_{ik}}{\zeta_k} = \mu_{ik}, \\ \text{Var}[\theta_{ik}] &= \zeta_k \cdot \left(\frac{\mu_{ik}}{\zeta_k} \right)^2 = \frac{\mu_{ik}^2}{\zeta_k}, \\ \text{CV}^2[\theta_{ik}] &= \frac{\mu_{ik}^2}{\zeta_k} \cdot \frac{1}{\mu_{ik}^2} = \frac{1}{\zeta_k}. \end{aligned}$$

It can also be shown that,

$$E[Y_{ij}] = \sum_k a_{ikj} \cdot \mu_{ik}$$

$$Var[Y_{ij}] = \sum_k a_{ikj} \cdot \mu_{ik} + \sum_k a_{ikj}^2 \cdot \frac{\mu_{ik}^2}{\zeta_k}$$

Note that in the above equation, an extra term for the variance of the observed data helps adjust for potential overdispersion that we may encounter in real data analysis.

Level 3—We assume β_k follows uninformative priors and $\zeta_k = \frac{z_0 \cdot B_0}{(1-B_0)}$ where

$B_0 = \frac{\zeta_k}{\zeta_k + z_0} \sim \text{uniform}(0, 1)$. The constant z_0 is median of ζ_k , so small values are less informative. For a given gene, β_k measures differential expression for the k^{th} isoform, i.e., the prior distribution of β_k is over different isoforms. Since this can vary from gene to gene, using an improper prior distribution provides more flexibility and a unified way of modeling genes with different isoform structures. The prior is also chosen to provide good repeated sampling properties to the resulting rules (Christiansen and Morris, 1997).

2.3 MCMC implementation

Based on a full Bayesian approach, the joint posterior density for parameters θ_{ik} and hyperparameters β_k and ζ_k given $\mathbf{Y} = \{Y_{ij}, i = 1, \dots, n, j = 1, \dots, J\}$ can be written as,

$$P(\theta_{ik}, \zeta_k, \beta_k | \mathbf{Y}) = \frac{\prod_{i=1}^n \prod_{j=1}^J [P(Y_{ij} | \theta_{ik}) \times P(\theta_{ik} | \zeta_k, \beta_k) \times P(\zeta_k, \beta_k)]}{P(\mathbf{Y})}$$

$$\propto \prod_{i=1}^n \left[\prod_{j=1}^J \left(e^{-\sum_{k=1}^K a_{ikj} \theta_{ik}} \cdot \left(\sum_{k=1}^K a_{ikj} \theta_{ik} \right)^{Y_{ij}} \right) \right] \times$$

$$\prod_{i=1}^n \left[\prod_{k=1}^K \left(\theta_{ik}^{\zeta_k - 1} \cdot \frac{e^{-\frac{\theta_{ik} \zeta_k}{\mu_{ik}}}}{\left(\frac{\mu_{ik}}{\zeta_k} \right)^{\zeta_k} \cdot \Gamma(\zeta_k)} \right) \right] \times \frac{z_0}{(\zeta_k + z_0)^2}.$$

However, it is difficult to sample from this joint distribution directly. This is in contrast to the Poisson hierarchical model of Christiansen and Morris (1997) where closed-form posterior distributions can be derived due to the use of conjugate priors. However, in our model, such conjugacy does not exist due to the fact that the observed reads data can be derived from different isoforms. To solve this problem, we propose to employ the MCMC technique to sample these parameters and to make inferences. In particular, the conditional distributions for unknown parameters and hyperparameters are given as follows,

$$P(\theta_{ik} | \zeta_k, \beta_k, \mathbf{Y}) \propto \prod_{j=1}^J \left(e^{-\sum_{k=1}^K a_{ikj} \theta_{ik}} \cdot \left(\sum_{k=1}^K a_{ikj} \theta_{ik} \right)^{Y_{ij}} \right) \cdot \theta_{ik}^{\zeta_k - 1} \cdot e^{-\frac{\theta_{ik} \zeta_k}{\mu_{ik}}}$$

$$P(\zeta_k | \theta_{ik}, \beta_k, \mathbf{Y}) \propto \prod_{i=1}^n \left(\theta_{ik}^{\zeta_k - 1} \cdot \frac{e^{-\frac{\theta_{ik} \zeta_k}{\mu_{ik}}}}{(\mu_{ik})^{\zeta_k} \Gamma(\zeta_k)} \right) \cdot \frac{z_0}{(\zeta_k + z_0)^2}$$

$$P(\beta_k | \zeta_k, \theta_{ik}, \mathbf{Y}) \propto \prod_{i=1}^n \left(e^{-\frac{\theta_{ik} \zeta_k}{e^{x_i^T \beta_k}}} \cdot e^{-(x_i^T \beta_k) \cdot \zeta_k} \right)$$

Since these conditional posterior distributions are not in standard distributional forms, we resort to use the MCMC approach to sample from these distributions. Specifically, we use Gibbs sampling to iteratively draw samples from approximate conditional distributions for

each set of the parameters. The samples are drawn sequentially, with the conditional distribution of the sampled draws depending on values of parameters at the previous iteration. We have shown that under some mild conditions, the conditional posterior distributions described are log-concave, which facilitates efficient sampling in WinBUGs (see Supplemental materials). In particular, if the conditional density functions above are log-concave, an adaptive rejection sampling (ARS) (Gilks, 1992) is used to obtain our sample draws in Gibbs sampling. We estimate our Bayesian hierarchical Poisson model and perform MCMC samplings using an R package called R2WinBUGs which provide a R interface to calling the WinBUGs software (Spiegelhalter *and others*, 1999).

2.4 Inference for isoform-specific expression and differential expression

In Bayesian inference framework, we summarize the posterior densities from the simulated draws to obtain the posterior means, standard deviations and 95% credible intervals. Estimation of individual-specific isoform expression index, θ_{ik} , is obtained by computing its posterior mean $\hat{\theta}_{ik}$, and uncertainty of this estimate is obtained based on 2.5 and 97.5 percentiles of its posterior distribution. Similarly, we make inference on isoform-specific differential expression by inspecting 95% credible interval of the regression coefficient for the group status, β_{k1} , assuming one covariate ($r=2$) for the k^{th} isoform. Nonzero coverage of β_{k1} 's credible interval suggests that isoform k is differentially expressed in one group compared to the other. This inference procedure is used in Sections 3 and 4 to assess differential expression in simulations and RNA-Seq data analysis.

Our proposed method can be applied to genome-wide RNA-Seq analysis, where each gene is analyzed separately by efficient parallel computing. Multiple testing then becomes an issue. There is no unified way of dealing with this issue in the framework of Bayesian inferences. One common approach is to calculate the Bayes factor (Kass and Raftery, 1995) for each isoform and to use the commonly used Jeffreys' scale to select the differentially expressed isoforms. Alternatively, we can calculate the posterior probability $Pr(|\beta_k| < \delta | \mathbf{Y})$ for some small threshold value δ and calculate the posterior probability-based false discovery rate as commonly used in differential expression analysis with microarray gene expression data (Muller *and others*, 2006).

3 Simulations Studies

In this section, we present simulation results to evaluate our proposed Bayesian hierarchical model for multi-sample analysis. We generated data sets based on the Poisson-Gamma hierarchical model satisfying (1) and (2) or a Poisson-normal model where a normal distribution is assume for the mean expression levels, with one covariate indicating the group status. We assume equal number of individual(s) in both case and control groups. All simulations are performed using the R and WinBUGs program with 21,000 MCMC runs and 1,000 burn-ins. No noticeable difference was observed when 100,000 runs and 50,000 burn-ins were used.

3.1 Simulation with one isoform

We first consider the simple case where a gene has only one isoform. We generated the read counts for each individual i based on

$$Y_{i1} \sim \text{Poisson}(\lambda_{i1}) = \text{Poisson}(w_i \cdot l \cdot \theta_{i1}),$$

$$\text{where } \theta_{i1} \sim \text{Gamma}(\zeta_1, \frac{\mu_{i1}}{\zeta_1}),$$

with $\log(\mu_{i1}) = \mathbf{x}_i^T \boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11})$. In this model, the observed read counts depend on w_i , l and θ_{i1} which are the total number of mapped reads for sample i in millions of reads (M), the length of a gene in kb and the isoform expression index respectively. In our simulations, we assume $w_i = w$ which implies constant number of mapped reads across samples. We create a quantity $m = w \cdot l$ to denote reads abundance that is a function of the total number of mapped reads and the length of a gene or an exon. We also note that when the fold-change in mean values of θ_{i1} between case and control groups is 2-fold or greater, it becomes increasingly easy to detect differential expression (data not shown). Therefore, we fixed the fold-change in isoform expression between case and control groups at 1.5 and studied the performance of our algorithm by varying m . In one-isoform simulation, we vary m from 2, 5, 10 and 50, which corresponds to $w = 2\text{M}$, 5M , 10M and 50M respectively for a gene of length of 1kb or equivalently to $l = 0.2\text{kb}$, 0.5kb , 1kb and 5kb respectively for 10M mapped reads in each sample. We set $\beta_{10} = -0.41$ and $\beta_{11} = 0.41$ where $\mathbf{x}_i = (1, 1)$ for a control individual and $\mathbf{x}_i = (1, 2)$ for a case individual. Based on this model, $E(\theta_{i1}) = 1$ if i is a control individual and $E(\theta_{i1}) = 1.5$ if i is a case individual, which corresponds to 1.5 fold-change in isoform expression between case and control groups. We consider different numbers of individuals in each group from 1, 5 and 10, i.e., 1 case vs. 1 control, 5 cases vs. 5 controls and 10 cases vs. 10 controls, respectively at various m values ranging from 2, 5, 10 and 50. To demonstrate that borrowing information across individuals results in more precise estimates for a given sample, we evaluated the average relative errors of isoform-specific expression and the average widths of 95% credible intervals for one particular individual across different scenarios over 100 runs of simulations.

The top panel of Figure 1 shows the relative errors and 95% credible intervals of the isoform-specific expression for one individual sample when analyzed in conjunction with different number of samples per group at varying degrees of reads abundance. We note decreases in relative errors and widths of the 95% credible intervals as the number of individuals increases in each group. Note that we did not average relative errors or widths of 95% credible intervals across individuals but only report one estimate of an individual. Marked improvements in estimates are seen for low reads abundance when $m = 2$ (corresponding to $w = 2\text{M}$ for 1kb gene or $l = 200\text{bp}$ for $w = 10\text{M}$), indicating the importance of borrowing information across samples in estimating the individual-specific expression. The middle panel of Figure 1 shows the boxplots of θ estimates for one case and one control sample in differential expression analysis between 10 cases and 10 controls at different reads abundance over 100 simulations. As seen in the plot, variability in θ estimates in one case or control individual among simulation runs decreases as number of mapped reads increases.

Similarly, we evaluated the relative errors and the width of the 95% credible intervals for $\hat{\beta}_{11}$ for different scenarios of number of individuals per group at varying degrees of reads abundance. We observed similar trends as seen in θ estimates (Figure 1). Also shown in this figure are the estimates of β_{10} and β_{11} over 100 simulation runs. Smaller variability and more accurate estimates are seen for high reads abundance ($m = 50$). We noted that at low reads abundance ($m = 2$), it is difficult to detect isoform differential expression between two groups as the whisker of $\hat{\beta}_{11}$'s boxplot extends to cover zero.

3.2 Simulation with multiple isoforms

We then evaluated the performance of our model based on hypothetical six-exon genes with two or three isoforms using 10 cases and 10 controls in each group. Even though the average number of exons per gene is around 8 in humans, many of these exons share common patterns across isoforms and they can be grouped into super-exons. In our 2- and 3-isoform case, we depict commonly seen patterns of splicing which is more relevant information compared to the number of exons in a given gene. See Figure 2 for an illustration of the gene

models used for simulations. For both 2-isoform and 3-isoform genes, we assign one isoform to be differentially expressed with 1.5 fold-increase for individuals in the case group compared to individuals in the control group, and the rest of the isoforms are not differentially expressed between the two groups. For the differentially expressed isoform k , we simulated read counts based on true underlying model with $\beta_{k0} = -0.41$ and $\beta_{k1} = 0.41$ where $\mathbf{x}_i = (1,1)$ for control individuals and $\mathbf{x}_i = (1,2)$ for case individuals. In this model, $E(\theta_{i,k}) = 1$ if i is a control and $E(\theta_{i,k}) = 1.5$ if i is a case, which corresponds to a 1.5 fold-change in expression between case and control groups. The read counts for each individual were generated as before but we fixed the number of individuals per group at 10 and varied m from 5, 10 and 50. In two- and three-isoform simulations, we simulated the data such that $m = 5, 10$ and 50 corresponds to $w = 5M, 10M$ and $50M$ with exon length of 500 bp for each of the exons in Figure 2, respectively. Table 1 presents the average estimates of the isoform-specific expression level θ , their average relative errors, the average widths of 95% credible intervals and results of differential expression analysis. Our results show that our estimates of isoform-specific expression, on average, are very close to the true values, the average relative errors and average of 95% credible intervals decrease as the reads abundance m increases, indicating the importance of high reads abundance. False differential expression (DE) rates are also observed to be closed to 5% and the power to detect differential expression increases as more reads are mapped to a sample or for longer exon length.

We then consider the scenarios that different isoforms have different expression levels and there exist major isoforms that are relatively highly expressed. In particular, Model 3 assumes the same isoforms as in Model 1 with a mean expression of isoform 1 of 5 for both case and control groups and a mean expression of isoform 2 of 1.5 for cases and 1 for controls, respectively. Model 4 assumes the same isoforms as in Model 2 with a mean expression of isoform 1 of 1 for both case and control groups, a mean expression of isoform 2 of 5 for both case and control groups, and a mean expression of isoform 3 of 1.5 and 1 for cases and controls, respectively. Table 2 presents the average estimates of the isoform-specific expression level θ , their average relative errors, the average widths of 95% credible intervals and results of differential expression analysis. Our results show that our estimates of isoform-specific expression, on average, are very close to the true values. False DE rates are also observed to be closed to 5%. As expected, we also observe that the power to detect differential expressions for lowly expressed isoforms decreases when there are abundant non-differentially expression isoforms.

We finally evaluate the sensitivity of our approach to misspecification of the isoform set. We define K to be set of three isoforms as depicted in Figure 2. We set the mean expression of isoform 1 to be 10 in both case and control groups, mean expression of isoform 2 to be 5 in both case and control groups. However, we set the isoform 3 as unexpressed in both case and control groups (i.e., has expression levels of 0). The number of individuals in each group is set to be 10 and $m = 50$. We can still estimate the expressions of isoforms 1 and 2 very well. For isoform 3, the means of the estimated expressions are 0.02 ($se = 0.004$) and 0.02 ($se = 0.003$) for the control group and case group, respectively. The results indicate that our method is not sensitive to including non-expressed isoforms in the model.

3.3 Comparison with single-sample analysis

In this section, we compare the relative errors, power to detect differential expression and false positive rates between multi-sample Bayesian hierarchical modeling and the single-sample procedure which involves estimating isoform-specific expression in the first step and performing differential expression analysis based on these estimates in the second step. However, such single-sample analysis does not account for the uncertainty of the estimates, which can lead to loss of power. To illustrate this, we simulated read counts in the same

manner as in the 1-isoform case with the expected fold-change of 1.5 between the two groups. In the single-sample approach, for 1-isoform gene, the maximum likelihood estimate (MLE) of gene expression can be obtained in closed-form (Jiang and Wong, 2009). Based on these MLEs, we use Wilcoxon two-sample test on the pooled sample to assess differential expression between the two groups. p -value of less than 0.05 is deemed significant and suggests differential expression between the groups.

Figure 3 compares the relative errors of the estimates of isoform expressions for one individual using the multi-sample and the single-sample approaches. In the multi-sample approach we present two curves ($n=10$ for 5 individuals per group and $n=20$ for 10 individuals per group). For a given sample, we observe that the multi-sample approach has much smaller relative errors in isoform-specific expression estimates compared to the single-sample approach. Increase in sample size from 5 to 10 individuals per group in the multi-sample analysis also reduces the relative errors in estimate of gene expression of a single sample. Since there is no borrowing of information in the single-sample analysis, the relative errors remain the same regardless of the sample size.

We then compared the proportion of correct differential expression calls using the multi-sample and the single-sample approaches in the settings with two different sample sizes (right plot of Figure 3). We clearly see that for a given sample size, the power to detect differential expression is greater using our multi-sample approach compared to the single-sample approach. When the sample size is small and isoform's abundance is low, the differences are quite pronounced. This illustrates the importance of pooling information across samples to provide more precise estimates of gene expression and increase power in detecting differential expression, particularly for low abundance isoforms. We also assessed the false positive rates when calling differential expression, where the expected fold-change in isoform expression is the same in both groups. The false positive rates for the multi-sample and single-sample approaches are very comparable with each ranges between 1–7% and each with mean of 5%. Average across simulations, false positive rates for both approaches are controlled at 5% level.

4 Application to RNA-Seq Data Analysis

4.1 Data description

To demonstrate the feasibility and the performance of our multi-sample approach on real data sets, we obtained published RNA-Seq data from lymphoblastoid cell lines (LCLs) for 10 Yoruban of Ibadan, Nigeria (YRI) samples based on 46 bp single-end reads (Pickrell *and others*, 2010) and 10 European (CEU) samples based on 37 bp paired-end reads (Montgomery *and others*, 2010). All the samples were sequenced on Illumina Genome Analyser platform. The average number of mapped reads for YRI and CEU samples are 5.9 and 10.6 millions reads, respectively (see Supplemental Table 1 for sample and read counts information). Raw sequence data were reformatted to FASTA file format and *Seqmap* (part of *rSeq* package) (Jiang and Wong, 2009) was used to map reads to the reference genome. We used output files generated from *Seqmap* to obtain the number of reads mapped to exons and exon-exon junctions in genes and used this as input to our program. We allowed number of mismatches to be 2 and by default the program only considers uniquely mapped reads. We studied six genes that were shown to be differentially expressed between the YRI and CEU samples based on microarray (Storey *and others*, 2007) and GeneChip Human Exon Array (Zhang *and others*, 2008). These includes cytokines and hemokine receptors genes, CCL22, CCL5, CCR7 and CXCR3 (Storey *and others*, 2007) and mRNA catabolism gene GSPT1 and cell-cell adhesion gene CD44 (Zhang *and others*, 2008). See Supplemental Tables 2–4 for detailed information on counts of the mapped reads on these genes. Storey et

al's analyses were based on eight YRI samples and eight CEU samples whereas Zhang et al's samples consisted of 30 YRIs and 30 CEUs.

4.2 Analysis results

For gene and isoform annotations, we used the UCSC hg19 refFlat format annotation file which is based on Feb 2009 assembly of the human genome. We estimated the gene and isoform-specific expression levels on these six genes, where CCL22, CCL5 and CCR7 have one isoform, CXCR3 has three exons with two isoforms, GSPT1 has 5 exons with three isoforms, and CD44 has 7 exons with 5 isoforms, using our multi-sample approach and compared the results to single-sample analysis using *rSeq* (Jiang and Wong, 2009). We let $\mathbf{x}_i = (1,1)$ for the YRI sample and $\mathbf{x}_i = (1,2)$ for the CEU sample in our model. Differential expression was assessed by $\hat{\beta}_{k1}$ and their credible intervals based on 100,000 MCMC runs with 50,000 burn-ins. In general, the estimates from these two approaches are very comparable when the number of read counts is large. However, for the isoforms with low read abundance, the multi-sample approach tends to shrink the estimates towards some common means. For example, the left panel of Figure 4 shows the estimated isoform-specific gene expression levels of gene CXCR3 for the 20 samples. The right panel of Figure 4 shows the widths of the 95% credible intervals of the estimated gene and isoform-specific expression levels for CXCR3 gene. We observe that for genes or isoforms with high expression levels, the results are very similar. However, for isoforms with low and moderate abundance, the widths of the 95% credible intervals from the multi-sample analysis are much smaller than those from the single-sample *rSeq* analysis, indicating smaller variances of our estimates of the isoform expressions.

Table 3 shows the results from differential expression analysis, indicating that CCL5, CCR7 and GSPT1 are differentially expressed at the gene level and GSPT1, isoform 1 and CD44, isoform 3 are differentially expressed at the isoform level. There was no evidence from these data that CCL22 or CXCR3 genes are differentially expressed between the YRI and CEU samples. As a comparison, we also report the results based on the Wilcoxon rank sum test or two-sample t-test for comparing gene and isoform-specific expression using the estimates obtained from single-sample *rSeq* analysis. Single-sample analysis was able to identify CCL5 and CCR5 as differentially expressed between the two populations but unable to identify any differentially expressed isoforms in GSPT1 and CD44 genes. Isoform 1 of GSPT1 is moderately expressed with $\hat{\theta}$ between 17–33, and isoform 3 of CD44 is of low abundance with $\hat{\theta}$ between 0.1–0.7. These results demonstrate that our multi-sample approach provides better estimates of gene and isoform expression with smaller variances and higher power to detect differential expression, particularly for genes and isoforms of low to moderate abundance.

Figure 4 shows that the sample #9 of the YRI data has much higher expression levels of the CXCR3 gene and isoforms than the other samples. To assess the robustness of our method, we redid the analysis leaving this sample out. Figure S1 in Supplemental Materials shows the estimates of the CXCR3 gene and its isoforms for the remaining samples. We observe that these estimates and their 95% credible intervals are very close to those in Figure 4, indicating that our MCMC procedure is robust to outlying observations. In addition, no differential expression of CXCR3 gene or its isoforms was observed.

5 Discussion

We have presented a Bayesian hierarchical model for multi-sample RNA-Seq data analysis in order to simultaneously estimate isoform-specific expressions and to identify differentially expressed isoforms. We illustrated through simulations and real RNA-Seq data analysis that our model has the advantage of borrowing information across all samples in

estimating expression levels, which can improve the estimates drastically for low abundance isoforms. We also demonstrated that the multi-sample approach for differential expression analysis led to higher power and more accurate estimates of isoform-specific expressions while controlling the number of false positives when compared to single-sample procedures. Large improvements in the power were seen when the sample size is small and isoform's abundance is low. Our Bayesian hierarchical regression framework also enables us to easily incorporate the sample-specific covariates such as age and gender on gene expression, which adjusts for possible confounders.

In our model, we assume that short reads are distributed uniformly across the genome and hence reads are sampled independently and uniformly from every possible nucleotides in the sample. However, this assumption may not hold true as recently demonstrated by Li *and others* (2010). Our model can be easily extended to accommodate non-uniformly in short-read rates following results from Li *and others* (2010) by modifying the coefficients a_{ikj} in our Poisson model (1). Specifically, we can replace the I_j with the sum of sequencing preferences of all nucleotides in region j , which can be estimated through a simple linear or nonlinear regression model (Li *and others*, 2010). Similar to *rSeq* of Jiang and Wong (2009), our model assumes that the gene model in terms of number of exons and isoform's structures are known *a priori*. Our simulation results also showed that when unexpressed isoforms are included in the model, the estimates of the expression levels of these isoforms are very close to zero and the estimates of the expressions of the expressed isoforms are not affected. As more and more RNA-Seq are being generated and analyzed, we expect to have more complete gene models, including new isoforms and splice junctions. One possible extension of our method is to include all possible isoforms that can be constructed from a set of known exons. However, for genes with many exons, the possible number of the isoforms can be very large. We will have to modify the prior distribution of the isoform-specific expression levels as a mixture of point mass zero and a gamma distribution and use the MCMC to obtain the posterior probability of a particular isoform being expressed in the samples.

We implemented the MCMC sampling using the WinBUGs software for ease of use. The R and WinBUGs codes are available upon request. Since the multi-sample analysis is performed one gene at a time, analysis of genome-wide RNA-Seq data and differential expression can be easily performed by parallel computing. In terms of computational time, our method is comparable to other existing methods for rigorous analysis of RNA-Seq data. For example, for low abundance isoforms, the one-sample analysis of *rSeq* has to apply the importance sampling technique to obtain the variance estimates for each isoform (Jiang and Wong, 2009). In addition, in order to account for the variability of the isoform estimates in the down-stream differential expression analysis, one has to apply some re-sampling methods, which also leads to increases in computing time. In contrast, our MCMC approach provides these estimates simultaneously. In the real data analysis we presented, for 100,000 MCMC runs with 50,000 burn-ins, on a standard desktop PC with 2G memory, it took 16, 54, 69 and 120 seconds for analysis of 1-isoform, 2-isoform 3-isoform and 5-isoform genes, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by NIH grants R01ES009911 and R01CA127334 (H. Li), R01HG004517 and R01HG005854 (M. Li) and T32CA093283 (S. Vardhanabhuti).

References

- Christiansen CL, Morris CN. Hierarchical Poisson Regression Modeling. *Journal of American Statistical Association*. 1997; 92:618–632.
- Gilks, WR. Bayesian Statistics. Vol. 4. Oxford University Press; 1992. Derivative-free adaptive rejection sampling for Gibbs sampling; p. 641-649.
- Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*. 2009; 25:1026–1032. [PubMed: 19244387]
- Kass R, Raftery AE. Bayes Factors. *Journal of the American Statistical Association*. 1995; 90:773–795.
- Katz Y, Wang ET, Airoidi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*. 2010; 7:1009–1015. [PubMed: 21057496]
- Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2009; 26:493–500. [PubMed: 20022975]
- Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*. 2010; 11:R50. [PubMed: 20459815]
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*. 2008; 18:1509–1517. [PubMed: 18550803]
- Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*. 2010; 464:773–777. [PubMed: 20220756]
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008; 5:621–628. [PubMed: 18516045]
- Muller, P.; Parmigiani, G.; Rice, K. FDR and Bayesian multiple comparisons rules. *Proc. Valencia/ISBA 8th World Meeting on Bayesian Statistics*; Benidorm (Alicante, Spain). 2006.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008; 320:1344–1349. [PubMed: 18451266]
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard J. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010; 464:768–772. [PubMed: 20220758]
- Spiegelhalter, DJ.; Thomas, A.; Best, NG. WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit; 1999.
- Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-Expression variation within and among human populations. *The American Journal of Human Genetics*. 2007; 80:502–509.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative iso-form regulation in human tissue transcriptomes. *Nature*. 2008; 456:470–476. [PubMed: 18978772]
- Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME. Evaluation of Genetic Variation Contributing to Differences in Gene Expression between Populations. *The American Journal of Human Genetics*. 2008; 82:631–640.

Appendix: Proof of log-concavity of the conditional distributions

Conditional distribution of θ_{ik}

$$\begin{aligned} \log P(\theta_{ik}|\zeta_k, \beta_k, Y_{ij}) &\propto \sum_{j=1}^J \left(-\sum_{k=1}^K a_{ikj} \theta_{ik} + Y_{ij} \log \left(\sum_{k=1}^K a_{ikj} \theta_{ik} \right) \right) + (\zeta_k - 1) \cdot \log \theta_{ik} - \frac{\theta_{ik} \zeta_k}{\mu_{ik}} \\ \frac{\delta \log P(\theta_{ik}|\zeta_k, \beta_k, Y_{ij})}{\delta \theta_{ik}} &= \sum_{j=1}^J \left(-a_{ikj} + \frac{Y_{ij} a_{ikj}}{\sum_{k=1}^K a_{ikj} \theta_{ik}} \right) + \frac{(\zeta_k - 1)}{\theta_{ik}} - \frac{\zeta_k}{\mu_{ik}} \\ \frac{\delta^2 \log P(\theta_{ik}|\zeta_k, \beta_k, Y_{ij})}{\delta \theta_{ik}^2} &= \sum_{j=1}^J \left(-\frac{Y_{ij} a_{ikj}^2}{(\sum_{k=1}^K a_{ikj} \theta_{ik})^2} \right) - \frac{(\zeta_k - 1)}{\theta_{ik}^2} \end{aligned}$$

Log-concavity is guaranteed if $\zeta_k \geq 1$.

Conditional distribution of ζ_k

$$\log P(\zeta_k|\theta_{ik}, \beta_k, Y_{ij}) \propto \sum_{i=1}^n \left[(\zeta_k - 1) \log \theta_{ik} - \frac{\theta_{ik} \zeta_k}{\mu_{ik}} + \zeta_k \log \zeta_k - \log \Gamma(\zeta_k) \right] - 2 \log(\zeta_k + z_0)$$

Since $\Gamma(\cdot)$ is log-convex, this implies $-\log \Gamma(\cdot)$ is concave, we denote $F(\cdot) = \log \Gamma(\cdot)$.

$$\begin{aligned} \frac{\delta \log P(\zeta_k|\theta_{ik}, \beta_k, Y_{ij})}{\delta \zeta_k} &= \sum_{i=1}^n \left[\log \theta_{ik} - \frac{\theta_{ik}}{\mu_{ik}} + (\log \zeta_k + 1) - F'(\zeta_k) \right] - \frac{2}{(\zeta_k + z_0)} \\ \frac{\delta^2 \log P(\zeta_k|\theta_{ik}, \beta_k, Y_{ij})}{\delta \zeta_k^2} &= \sum_{i=1}^n \left[\frac{1}{\zeta_k} - F''(\zeta_k) \right] + \frac{2}{(\zeta_k + z_0)^2} \end{aligned}$$

Log-concavity is guaranteed if $n F''(\zeta_k) \geq \frac{n}{\zeta_k} + \frac{2}{(\zeta_k + z_0)^2}$

$$\Rightarrow F''(\zeta_k) \geq \frac{1}{\zeta_k} + \frac{2}{n(\zeta_k + z_0)^2}$$

If we set $z_0 = \zeta_k$, log-concavity is guaranteed regardless of n .

If $z_0 = 0$, then we need $n \geq 4$ for guaranteed log-concavity.

Conditional distribution of β_k

$$\begin{aligned} \log P(\beta_k|\theta_{ik}, \zeta_k, Y_{ij}) &\propto \sum_{i=1}^n \left(-\theta_{ik} e^{-x_i^T \beta_k} - x_i^T \beta_k \right) \\ \frac{\delta \log P(\beta_k|\theta_{ik}, \zeta_k, Y_{ij})}{\delta \beta_k} &= \sum_{i=1}^n \left(\theta_{ik} e^{-x_i^T \beta_k} x_i^T - x_i^T \right) \\ \frac{\delta^2 \log P(\beta_k|\theta_{ik}, \zeta_k, Y_{ij})}{\delta \beta_k^2} &= \sum_{i=1}^n \left(-\theta_{ik} e^{-x_i^T \beta_k} x_i^T x_i \right) \end{aligned}$$

If we let $H = -dx_i^T x_i$ where $d = \theta_{ik} e^{-x_i^T \beta_k}$, we have $d \geq 0$ since $\theta_{ik} \geq 0$.

$$\begin{aligned} \Rightarrow yHy^T &= y(-dx_i^T x_i)y^T \\ &= -d(yx_i^T)(yx_i^T)^T \\ &= -d(yx_i^T)^2 \leq 0 \Rightarrow \text{concavity is guaranteed.} \end{aligned}$$

Fig. 1. Simulation for genes with one isoform. Top panel: relative errors (left) and 95% credible interval widths (right) for estimate of the isoform-specific expression level $\hat{\theta}$ for one individual when analyzed with different numbers of individuals. Middle panel: boxplots of the estimates of θ for one case and one control individual (left) and estimates of β_{10} and β_{11} (right) over 100 replications, where the horizontal lines indicate true values. Bottom panel: relative errors (left) and 95% credible interval widths (right) for the estimates of differential expression parameter β_{11} . In all plots, data with different read abundances are measured by different values of m .

Fig. 2. Gene models used for multiple-isoform simulations, where Model 1 assumes a gene with two isoforms and Model 2 assumes a gene with three isoforms.

Fig. 3. Simulation comparison between multi-sample and single-sample approaches. Left plot: relative error of estimate of the isoform-specific expression for one sample in single-sample and multi-sample approaches. Right plot: power to detect differential expression. Data with different read abundances are measured by different m values.

Fig. 4.

Analysis of real data set comparing the isoform expressions of CXCR3 gene between 10 YRI and 10 CEU samples. Left panels: estimated expressions of CXCR3 gene and its two isoforms for 20 individuals, where the first 10 individuals are YRI and the next 10 individuals are CEU. Right panels: widths of the 95% credible intervals for the estimated gene and isoform-specific expression levels for CXCR3 gene.

Simulation based on Poisson-Gamma model for 6-exon genes with two and three isoforms. Model 1: gene with 2 isoforms. Model 2: gene with 3 isoforms. In Model 1, $\mu_{control,1} = \mu_{control,2} = \mu_{case,1} = 1$ and $\mu_{case,2} = 1.5$. In Model 2, $\mu_{control,1} = \mu_{control,2} = \mu_{control,3} = \mu_{case,1} = 1$ and $\mu_{case,2} = 1.5$. Presented in table are the average estimates of isoform expressions for one individual, their average relative errors and average widths of the 95% credible intervals. For each model, data with different read abundances are measured by different values of m .

Table 1

	Model 1			Model 2		
	$m = 5$	$m = 10$	$m = 50$	$m = 5$	$m = 10$	$m = 50$
Avg. of estimate of $\alpha(\hat{\theta})$						
isoform1 (control)	1.01	1.02	1.00	1.09	1.05	1.01
isoform1 (case)	1.02	1.01	1.02	1.10	1.06	1.00
isoform2 (control)	1.00	0.97	1.00	0.94	0.95	0.99
isoform2 (case)	1.52	1.49	1.49	0.88	0.94	1.00
isoform3 (control)	-	-	-	0.94	0.99	1.00
isoform3 (case)	-	-	-	1.49	1.50	1.52
Avg. of relative error of $\hat{\theta}$						
isoform1 (control)	0.12	0.10	0.06	0.27	0.22	0.11
isoform1 (case)	0.13	0.09	0.06	0.34	0.24	0.10
isoform2 (control)	0.18	0.13	0.07	0.23	0.20	0.09
isoform2 (case)	0.13	0.10	0.06	0.28	0.20	0.09
isoform3 (control)	-	-	-	0.22	0.18	0.09
isoform3 (case)	-	-	-	0.16	0.14	0.07
Avg. of width of 95% CI of $\hat{\theta}$						
isoform1 (control)	0.69	0.53	0.29	1.33	1.03	0.52
isoform1 (case)	0.71	0.53	0.30	1.48	1.15	0.56
isoform2 (control)	0.89	0.67	0.37	1.21	0.93	0.48
isoform2 (case)	1.07	0.81	0.45	1.28	1.02	0.51
isoform3 (control)	-	-	-	1.13	0.90	0.47
isoform3 (case)	-	-	-	1.40	1.06	0.56
Differential expression (DE)						
% true DE	42	70	100	-	-	-
% false DE	4	6	5	-	-	-

	Model 1		Model 2			
	$m = 5$	$m = 10$	$m = 50$	$m = 5$	$m = 10$	$m = 50$
% true DE	-	-	-	31	42	94
% false DE	-	-	-	1	5	3
% false DE	-	-	-	5	2	1

Table 2

Simulation based on Poisson-Gamma model for 6-exon genes with two and three isoforms. Model 3: gene with 2 isoforms. Model 4: gene with 3 isoforms. In Model 3, $\mu_{control,1} = \mu_{case,1} = 5$, $\mu_{control,2} = 1$ and $\mu_{case,2} = 1.5$. In Model 4, $\mu_{control,1} = \mu_{case,1} = 1$, $\mu_{control,2} = \mu_{case,2} = 5$, $\mu_{case,3} = 1.5$ and $\mu_{control,3} = 1$. Presented in table are the average estimates of isoform expressions for one individual, their average relative errors and average widths of the 95% credible intervals. For each model, data with different read abundances are measured by different values of m .

	Model 3			Model 4		
	$m = 5$	$m = 10$	$m = 50$	$m = 5$	$m = 10$	$m = 50$
	Avg. of estimate of $\alpha(\hat{\theta})$					
isoform1 (control)	4.95	4.95	4.95	1.01	1.00	1.00
isoform1 (case)	5.11	5.02	5.03	1.04	1.05	1.01
isoform2 (control)	0.98	0.98	0.98	4.94	4.96	4.95
isoform2 (case)	1.30	1.36	1.47	4.96	4.96	4.95
isoform3 (control)	-	-	-	1.00	1.00	0.99
isoform3 (case)	-	-	-	1.44	1.48	1.48
	Avg. of relative error of $\hat{\theta}$					
isoform1 (control)	0.07	0.05	0.03	0.44	0.29	0.18
isoform1 (case)	0.07	0.06	0.04	0.50	0.37	0.16
isoform2 (control)	0.29	0.25	0.13	0.09	0.06	0.04
isoform2 (case)	0.27	0.20	0.09	0.08	0.07	0.03
isoform3 (control)	-	-	-	0.36	0.27	0.17
isoform3 (case)	-	-	-	0.31	0.20	0.11
	Avg. of width of 95% CI of $\hat{\theta}$					
isoform1 (control)	1.62	1.27	0.73	1.59	1.38	0.73
isoform1 (case)	1.73	1.34	0.77	1.95	1.57	0.77
isoform2 (control)	1.36	1.04	0.61	1.99	1.65	0.95
isoform2 (case)	1.68	1.31	0.73	2.14	1.73	0.98
isoform3 (control)	-	-	-	1.48	1.30	0.68
isoform3 (case)	-	-	-	1.94	1.56	0.75
	Differential expression (DE)					
% true DE	24	32	72	-	-	-
% false DE	3	5	6	-	-	-

	Model 3		Model 4			
	$m = 5$	$m = 10$	$m = 50$	$m = 5$	$m = 10$	$m = 50$
% true DE	-	-	-	12	22	62
% false DE	-	-	-	2	2	4
% false DE	-	-	-	6	4	4

Table 3

Differential expression analysis of six genes based on 10 Yoruban and 10 European RNA-Seq samples. Parameter estimate of β_{k1} and its 95% Bayesian credible interval from our proposed multi-sample analysis are given. p -values from the Wilcoxon or two-sample t -test based on single-sample analysis from *rSeq* are also provided.

Gene	$\hat{\beta}_{k1}$ (95% CI)	Wilcoxon	T-test
CCL22	0.48 (-0.14, 1.06)	0.089	0.096
CCL5	-0.73 (-1.14, -0.34)*	0.019*	0.008*
CCR7	0.69 (0.18, 1.17)*	0.015*	0.013*
CXCR3 - gene	-0.46 (-1.24, 0.35)	0.218	0.421
CXCR3 - iso1	-0.50 (-1.33, 0.34)	0.218	0.405
CXCR3 - iso2	-0.03 (-1.30, 1.20)	0.435	0.699
GSPT1 - gene	-0.15 (-0.32, -0.002)*	0.198	0.089
GSPT1 - iso1	-0.16 (-0.34, -0.001)*	0.166	0.103
GSPT1 - iso2	-0.24 (-1.53, 1.04)	1	0.967
GSPT1 - iso3	-0.20 (-0.95, 0.75)	0.655	0.910
CD44 - gene	0.02 (-0.34, 0.41)	0.739	0.616
CD44 - iso1	-0.84 (-2.39, 0.58)	0.970	0.761
CD44 - iso2	-0.43 (-1.05, 0.22)	0.084	0.070
CD44 - iso3	-1.37 (-2.39, -0.53)*	0.139	0.121
CD44 - iso4	-1.21 (-5.00, 1.52)	0.504	0.305
CD44 - iso5	0.05(-0.32, 0.41)	0.821	0.808