Panagiotis Papastamoulis* and Magnus Rattray

Bayesian estimation of differential transcript usage from RNA-seq data

https://doi.org/10.1515/sagmb-2017-0005

Abstract: Next generation sequencing allows the identification of genes consisting of differentially expressed transcripts, a term which usually refers to changes in the overall expression level. A specific type of differential expression is differential transcript usage (DTU) and targets changes in the relative within gene expression of a transcript. The contribution of this paper is to: (a) extend the use of cjBitSeq to the DTU context, a previously introduced Bayesian model which is originally designed for identifying changes in overall expression levels and (b) propose a Bayesian version of DRIMSeq, a frequentist model for inferring DTU. cjBitSeq is a read based model and performs fully Bayesian inference by MCMC sampling on the space of latent state of each transcript per gene. BayesDRIMSeq is a count based model and estimates the Bayes Factor of a DTU model against a null model using Laplace's approximation. The proposed models are benchmarked against the existing ones using a recent independent simulation study as well as a real RNA-seq dataset. Our results suggest that the Bayesian methods exhibit similar performance with DRIMSeq in terms of precision/recall but offer better calibration of False Discovery Rate.

Keywords: alternative splicing; false discovery rate; Laplace approximation; MCMC; within gene transcript expression.

1 Introduction

High throughput sequencing of cDNA (RNA-seq) (Mortazavi et al., 2008) is an important tool to quantify transcript expression levels and to identify differences between different biological conditions. RNA-seq experiments produce a large number (millions) of short reads (nucleotide sequences) which are typically mapped to the genome or transcriptome. Expression quantification requires estimating the number of reads originating from each transcript in a given sample. Quantifying the transcriptome between different samples allows the identification of differentially expressed (DE) transcripts between them. However, certain difficulties complicate the inference procedure. In higher eukaryotes, most genes are spliced into alternative transcripts which share specific parts of their sequence (exons). Hence, a given short read typically aligns to different positions of the transcriptome and statistical models are often used to infer the origin probabilistically (Trapnell et al., 2010, 2013; Li and Dewey, 2011; Nicolae et al., 2011; Glaus et al., 2012; Rossell et al., 2014; Hensman et al., 2015).

Differential transcript expression (DTE) refers to the event where the overall relative expression of a transcript changes between two conditions. In this case, θ_k refers to the relative expression of transcript k; $k = 1, \ldots, K$, with respect to the whole set of transcripts, with $\theta_k \geqslant 0$ and $\sum_{k=1}^{K} \theta_k = 1$. On the contrary, differential transcript usage (DTU) refers to the event that the relative within gene abundance of a transcript changes between conditions. Consider a gene $g = 1, \ldots, G$ with $K_g > 1$ transcripts. Then, the relative within gene transcript abundance is defined as $\theta_k^{(g)} = \frac{\theta_k}{\sum_{j \in g} \theta_j}$. Obviously, if a transcript belongs to a gene with $K_g = 1$ then it is always non-DTU. According to Gonzàlez-Porta et al. (2013) the dominant transcripts within a gene are likely to be the main contributors to the proteome and switching events between them is a common scenario of gene modification between conditions.

***Corresponding author: Panagiotis Papastamoulis,** Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PL, UK, e-mail: panagiotis.papastamoulis@manchester.ac.uk

Magnus Rattray: Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK

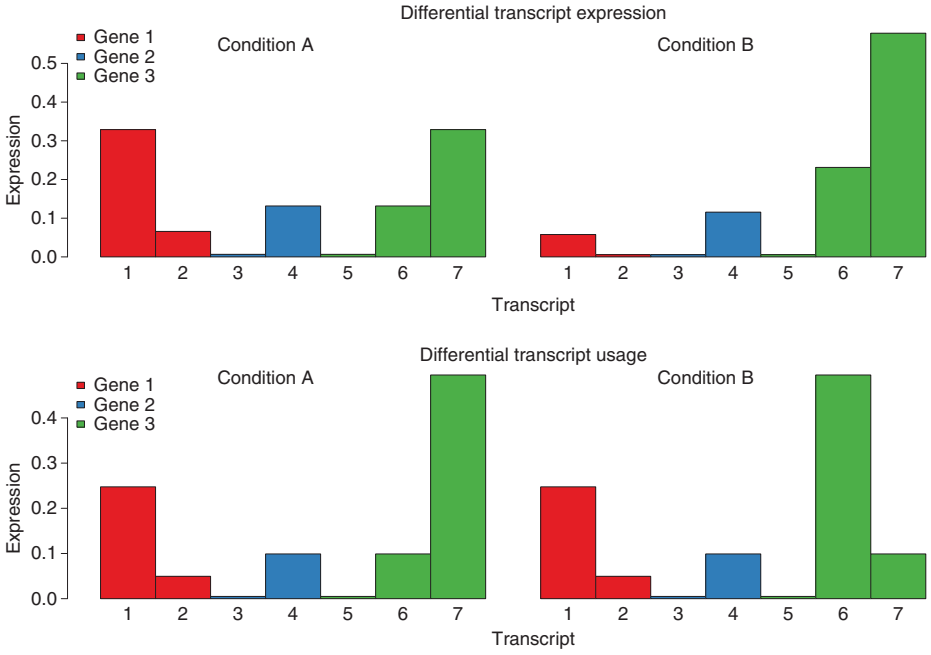

Figure 1: Differential transcript expression (up) and differential transcript usage (down).

Figure 1 illustrates the differences between DTE and DTU, considering a set of three genes (shown in red, blue and green) consisting of 2, 2 and 3 transcripts. In the case of DTE (upper panel) the overall expression of transcripts 1, 2, 6 and 7 change: in particular transcripts 1 and 2 are up-regulated in condition A while trancripts 6 and 7 are up-regulated in condition B. In the lower panel of Figure 1 note that only transcripts 6 and 7 are DTE. However, also note that now the relative expression of these transcripts conditionally on the set of the same-gene transcripts (green color) is not the same between conditions. In general, DTU implies DTE but the reverse is not necessarily true.

In this paper we extend the use of two available methods in order to perform Bayesian inference for the problem of DTU. cjBitSeq (Papastamoulis and Rattray, 2017) was originally introduced as a Bayesian read-based model for DTE inference and here we modify it for the DTU problem. We also propose a Bayesian version of DRIMSeq (Nowicka and Robinson, 2016), a count-based approach originally introduced as a frequentist model for DTU inference. Genome-scale studies incorporate a large number of multiple tests, typically at the order of tens of thousands. A crucial issue under a multiple comparisons framework is the control of the False Discovery Rate (FDR), that is, the expected proportions of errors among the rejected hypotheses (Benjamini and Hochberg, 1995). According to a recent benchmarking study (Soneson et al., 2015), the ability of frequentist count-based methods to control the FDR is drastically improved by pre-filtering low-expressed transcripts. This remains true for the Bayesian version of the count-based method presented here (DRIMSeq). However it is not possible to incorporate such a strategy for read-based methods (cjBitSeq) where transcript expression levels are not known a priori. Therefore, under our Bayesian framework, we also propose the use of transformations of the raw posterior probabilities and filtering the output based on the notion of trust regions which are motivated from realistic scenarios of gene regulation (Gonzàlez-Porta et al., 2013).

The rest of the paper is organized as follows. In Section 2 we briefly describe existing methods. The proposed Bayesian models are presented in Section 3. More specifically, Section 3.1 reviews the cjBitSeq framework and also introduces the necessary prior modifications for the problem of DTU. The likelihood of the DRIMSeq model is presented in Section 3.2 and a Bayesian version is introduced next, along with a detailed description of the inference. Section 4 deals with FDR control procedures. In Section 5 we report our findings on synthetic data using the carefully designed simulation study of Soneson et al. (2015). In Section 5.1 we compare cjBitSeq and BayesDRIMSeq with respect to the decision rules of Section 4 using power versus

achieved FDR plots. In Section 5.2 we benchmark these methods against existing ones and we also report more performance measures, such as ROC and precision/recall curves as well as comparisons in terms of run-time and memory requirements. A real RNA-seq dataset is analysed in Section 6. The manuscript concludes with a Discussion. A prior sensitivity analysis of BayesDRIMSeq as well as a comparison between alternative inputs of BayesDRIMSeq and DRIMSeq based on different quantification methods is provided in the Appendix.

2 Existing methods

cuffdiff

The cufflinks/cuffdiff (Trapnell et al., 2010, 2013) pipeline estimates the expression of a set of transcripts and then performs various differential expression tests both on the transcript and gene level. DTU at the gene level is based on comparing the similarity of two distributions using the square root of the Jensen-Shannon divergence (Osterreicher and Vajda, 2003; Endres and Schindelin, 2003). Following Soneson et al. (2015), we used the gene-wise FDR estimates from the `cds.diff` output file of cuffdiff (version 2.2.1).

DEXSeq

DEXSeq (Anders et al., 2012) is the most popular method for inferring DTU. The genome is divided into disjoint parts of exons (counting bins) and a matrix of read counts into the counting bins is used as input. The default method for counting reads for this purpose is HTSeq (Anders et al., 2015). Given the estimated reads from HTSeq, a negative binomial generalized linear model is fit and DTU is inferred by testing whether the interaction term between conditions is different from zero.

DRIMSeq

This recent package (Nowicka and Robinson, 2016) implements a dirichlet-multinomial model in order to describe the variability between replicates. A likelihood ratio test is performed in order to compare a full model with distinct parameters per condition and a null model which assumes that the parameters are shared. The input is a matrix of counts per transcript. We applied this method using the following filtering criteria:

- `min_gene_expr` $= 1$ (Minimal gene expression in cpm)
- `min_feature_prop` $= 0.01$ (Minimal proportion for feature expression)
- `min_samps_gene_expr` $= 3$ (Minimal number of samples where genes should be expressed)
- `min_samps_feature_prop` $= 3$ (Minimal number of samples where features should be expressed)

edgeR

The function `spliceVariants` from the edgeR (Robinson et al., 2010) package can be used to identify genes showing evidence of splice variation using negative binomial generalized linear models. For each gene (containing at least two transcripts) a likelihood ratio test compares a model with an interaction term between each condition against a null model with no interaction term. The input corresponds to a matrix of counts per transcript.

limma

The function `diffSplice` from the limma (Ritchie et al., 2015) package also tests for DTU by fitting negative binomial generalized linear models and performing a likelihood ratio test at the difference of log-fold changes. The input corresponds to a matrix of counts per transcript.

3 New Bayesian approaches

cjBitSeq was originally applied to problem of inferring transcripts with DTE and here this model is modified for the problem of DTU. DRIMSeq is a frequentist-based approach for the problem of DTU and this model is now extended under a Bayesian framework. cjBitSeq is a read-based model, that is, the observed data is a matrix of alignments of each read to the transcriptome. On the other hand, DRIMSeq is a count-based model, which uses as input a matrix of (estimated) counts corresponding to the number of reads originating from each transcript. Both methods report an estimate of the posterior probability of DTU per gene. cjBitSeq performs collapsed Gibbs sampling on the space of latent states of each transcript, that is, a binary vector with 0 corresponding to equally expressed (EE) transcripts and 1 otherwise. Bayesian DRIMSeq estimates the Bayes factor between a DTU and a null model. Therefore, cjBitSeq also reports a posterior probability of DTU for each transcript which may be of interest for transcript-level analysis. In this study we focus our attention at the gene-level summaries as done in Soneson et al. (2015).

Both models take advantage of distributions with richer covariance structures compared to standard sampling schemes: in particular, the generalized Dirichlet distribution is arising as a full conditional distribution at the cjBitSeq model, while DRIMSeq is based on the Dirichlet-Multinomial distribution. The Generalized Dirichlet distribution allows for positive correlations between proportions, something that it is not the case for a standard Dirichlet model, and the Dirichlet-Multinomial distribution exhibits extra variation compared to a multinomial model. Interestingly, we note that both distributions were introduced by the same author (Mosimann, 1962; Connor and Mosimann, 1969).

3.1 cjBitSeq

Let $\boldsymbol{x} = (x_1, \ldots, x_r)$, $x_i \in \mathscr{X}$, $i = 1, \ldots, r$, denote a sample of r short reads aligned to a given set of K transcripts. The sample space $\mathscr{X}$ consists of all sequences of letters A, C, G, T. Assuming that reads are independent, the joint probability density function of the data is written as

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \prod_{i=1}^{r} \sum_{k=1}^{K} \theta_k f_k(x_i). \tag{1}$$

The number of components (K) is equal to the number of transcripts and it is considered as known since the transcriptome is given. The parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \mathscr{P}_{K-1}$ denotes relative abundances, where

$$\mathscr{P}_{K-1} := \{p_k \geqslant 0, k = 1, \ldots, K-1 : \sum_{k=1}^{K-1} p_k \leqslant 1; p_K := 1 - \sum_{k=1}^{K-1} p_k\}.$$

The component specific density $f_k(\cdot)$ corresponds to the probability of a read aligning at some position of transcript k, $k = 1, \ldots, K$. Since we assume a known transcriptome, $\{f_k\}_{k=1}^{K}$ are known as well and they are computed according to the methodology described in Glaus et al. (2012), taking into account optional position and sequence-specific bias correction method.

Papastamoulis and Rattray (2017) proposed a Bayesian model selection approach for identifying differentially expressed transcripts from RNA-seq data. The methods builds upon the BitSeq model (Glaus et al., 2012; Papastamoulis et al., 2014; Hensman et al., 2015). Compared to other approaches, the main difference of cjBitSeq is that transcript expression and differential expression is jointly modelled. In contrast to other methods where the starting point of the DE analysis is a count matrix, the input of cjBitSeq is the matrix L containing alignment probabilities of each read to the transcriptome. According to Equation (1), the probability of read i aligning at transcript k is given by $L_{ik} = f_k(x_i)$ for $i = 1, \ldots, r$ and $k = 1, \ldots, K$.

Assume that we have at hand two samples $\boldsymbol{x} := (x_1, \ldots, x_r)$ and $\boldsymbol{y} := (y_1, \ldots, y_s)$, with r and s denoting the number of (mapped) reads for sample $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Now, let θ_k and w_k denote the unknown

relative abundance of transcript $k = 1, \ldots, K$ in sample $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Define the parameter vector of relative abundances as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K-1}; \theta_K) \in \mathscr{P}_{K-1}$ and $\boldsymbol{w} = (w_1, \ldots, w_{K-1}; w_K) \in \mathscr{P}_{K-1}$. Under the standard BitSeq model the prior on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$ would be a product of independent Dirichlet distributions. In this case the probability $\theta_k = w_k$ under the prior is zero and it is not straightforward to define non-DE transcripts. To model differential expression we would instead like to identify instances where transcript expression has not changed between samples. Therefore, we introduce a finite probability for the event $\theta_k = w_k$. This leads us to define a new model with a non-independent prior for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$.

Definition 1 (State vector). *Let $c := (c_1, \ldots, c_K) \in \mathscr{C}$, where $\mathscr{C}$ is the set defined by:*

1. $c_k \in \{0, 1\}$, $k = 1, \ldots, K$
2. $c_+ := \sum_{k=1}^{K} c_k \neq 1$.

Then, for $k = 1, \ldots, K$ let: $\begin{cases} \theta_k = w_k, & \text{if } c_k = 0 \\ \theta_k \neq w_k, & \text{if } c_k = 1. \end{cases}$ *We will refer to vector c as the state vector of the model.*

cjBitSeq was originally applied to the problem of DTE by introducing a cluster representation of aligned reads to transcripts. This clustering approach substantially reduces the dimensionality of the sampling space and makes the MCMC sampler converge to reasonable time. It is important to mention that clusters are defined under a data-driven algorithm, that is, by searching the alignments of each read and identifying groups of transcripts sharing reads.

Under the same approach, we would be able to infer clusters of transcripts with DTU. However, since in this work we focus on inference at the gene level, we impose the assumption that clusters are defined as the transcripts of each gene. Otherwise, in some instances it will not be straightforward to perform inference at the gene level, due to the possibility of clusters of transcripts merging multiple genes together. For example, we found that approximately 4.5% of mapped reads align to more than one gene in our simulation experiments of Section 5 using paired-end reads with length 101 base-pairs. In case that a read maps to more than one gene, we only keep the alignments corresponding to transcripts of the gene containing the best score for this specific read. Thus, the cjBitSeq algorithm is applied separately to each gene (consisting of at least two transcripts).

For the problem of DTU, cjBitSeq is applied under a modification in the prior distribution of DE per transcript. Under the Jeffreys' prior, which is used in the default cjBitSeq setting, the probability of a gene consisting of DE transcripts is an increasing function of the number of transcripts. This prior is reasonable at a transcript-level analysis and it has been shown that it outperforms other choices. However, this choice introduces a prior bias to the case of DTU since genes with larger number of transcripts are assigned larger prior probability of DTU than genes with small number of transcripts. Therefore, now it is a priori assumed that the probability of no differential expression within a gene is equally weighted with the event that at least two transcripts exhibit DTU, that is, $\mathbb{P}(c_+ = 0) = 0.5$. An equal prior probability is assigned to the rest possible configurations. Thus, the prior distribution on the state vector is defined as:

$$\mathrm{P}(c) = \mathrm{P}(c | c_+ \neq 1) = \begin{cases} 0.5, & c_+ = 0 \\ \frac{0.5}{2^K - K - 1}, & c_+ \geqslant 2. \end{cases} \tag{2}$$

This modification is necessary in order to ensure that no prior bias is enforced at the gene-level which is the aim of the analysis in the DTU setup.

A graphical model of the cjBitSeq prior assumptions is shown in Figure 2. The binary state vector $c = (c_1, \ldots, c_K)$ defines differentially or equally expressed transcripts within each gene. The prior distribution of $\boldsymbol{c}$ is given by Equation (2), although in the general implementation of Papastamoulis and Rattray (2017) an extra level of hierarchy is imposed by the hyper-parameter π, shown in Figure 2. The parameters $\boldsymbol{u}$ and $\boldsymbol{v}$ are a-priori independent Dirichlet random variables. The dimension of $\boldsymbol{u}$ is equal to K, i.e. the number of transcripts for a given gene. On the other hand, $\boldsymbol{v}$ is a random variable with varying

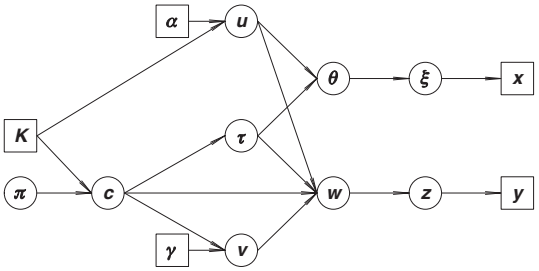

Figure 2: Directed Acyclic Graph representation for the cjBitSeq model. Squares and circles represent unknown and observed/fixed quantities, respectively.

dimension, which is defined by the number of differentially expressed transcripts, that is, $\sum_{k=1}^{K} c_k$. The parameters $\boldsymbol{u}$ and $\boldsymbol{v}$ along with an auxiliary parameter $\boldsymbol{\tau}$ define via a suitable one-to-one transformation the actual transcript expression parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$. According to Theorem 1 of Papastamoulis and Rattray (2017), $\boldsymbol{\theta}$ and $\boldsymbol{w}$ are marginally Dirichlet random variables, however they are not independent since the probability of the events $\{\theta_k = w_k; k = 1, \ldots, K\}$ is positive. At the next level of hierarchy, the latent allocation variables $\boldsymbol{\xi}$ and $\boldsymbol{z}$ define the transcript allocation of each read from sample $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, through the equations $P(\xi_i = k) = \theta_k$, independent for $i = 1, \ldots, r$, and $P(z_j = k) = w_k$, independent for $j = 1, \ldots, s$.

Papastamoulis and Rattray (2017) showed that the model is conjugate given c. But in order to update ($\boldsymbol{c}$, $\boldsymbol{v}$), a reversible-jump mechanism (Green, 1995; Richardson and Green, 1997; Papastamoulis and Iliopoulos, 2009) is required. However, this step can be avoided by analytical integration of ($\boldsymbol{u}$, $\boldsymbol{v}$). Thus, a collapsed Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990; Liu, 1994; Liu et al., 1995) updates the latent allocation variables ($\boldsymbol{\xi}$ and $\boldsymbol{z}$) of each read to its transcript of origin as well as the binary variables c_k of each transcript state (DE or EE). Let $\boldsymbol{x}_{-[i]}$ denote the vector arising from x after excluding its i-th entry. A pseudo-code description of the collapsed Gibbs MCMC sampler is:

1. Update allocation variables for sample $\boldsymbol{x}$: $\xi_i | \boldsymbol{\xi}_{[-i]}, \boldsymbol{z}, c, \boldsymbol{x}, \boldsymbol{y}, i = 1, \ldots, r$.
2. Update allocation variables for sample $\boldsymbol{y}$: $z_j | \boldsymbol{\xi}, \boldsymbol{z}_{[-j]}, c, \boldsymbol{x}, \boldsymbol{y}, j = 1, \ldots, s$.
3. Draw a random sample (without replacement) of indices (j_1, j_2) from $\{1, \ldots, K\}$ and update the block of state vector $c_{j_1, j_2} | \boldsymbol{c}_{-[j_1, j_2]}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}$.
4. Update $(\boldsymbol{\theta}, \boldsymbol{w}, \tau, \boldsymbol{u}, \boldsymbol{v}) | c, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}$ (optional).

Note that the update 4 is optional in the sense that it is not required by any of the previous steps, however one can include it in order to also obtain MCMC samples of the transcript expression parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$. For a detailed description of the conditional distributions involved in steps 1–4 (as well as the alternative RJMCMC sampler) see Papastamoulis and Rattray (2017).

According to our model, it is natural to call a gene as DE if at least two transcripts exhibit DTU. Hence, the posterior probability of DTU for a gene g is defined as

$$p_g = \mathbb{P}\{c_+ > 0 | \boldsymbol{x}, \boldsymbol{y}\}, \quad g = 1, \ldots, G, \tag{3}$$

and it is estimated by the corresponding ergodic average across the MCMC run (after burn-in).

3.2 BayesDRIMSeq

Let $n = n_g$ denotes the total number of reads aligning to a gene g with k transcripts, $g = 1, \ldots, G$. Assume that $\boldsymbol{X} = \boldsymbol{X}_g = (X_1, \ldots, X_k)$ is the vector of reads originating from each transcript, according to an underlying vector $\boldsymbol{\theta} = \boldsymbol{\theta}_g = (\theta_1, \ldots, \theta_k)$ of relative abundances which is unknown. A priori, a Dirichlet prior is imposed

on $\boldsymbol{\theta}$ and, given $\boldsymbol{\theta}$, the observed reads are generated according to a multinomial distribution, that is,

$$\boldsymbol{\theta} \sim \mathscr{D}(\delta_1, \ldots, \delta_k)$$

$$\boldsymbol{X}|\boldsymbol{\theta} \sim \text{Multinomial}(n, \boldsymbol{\theta})$$

Integrating out $\boldsymbol{\theta}$, this model leads to the Dirichlet-Multinomial (Mosimann, 1962) distribution:

$$\text{P}(\boldsymbol{X} = \boldsymbol{x}) = \binom{n}{\boldsymbol{x}} \frac{\Gamma(\delta_+)}{\Gamma(n+\delta_+)} \prod_{j=1}^{k} \frac{\Gamma(\delta_j + x_j)}{\Gamma(\delta_j)},$$

where the first term in the product denotes the multinomial coefficient and $\delta_+ = \sum_{k=1}^{K} \delta_k$. We will write: $\boldsymbol{X}|n, \boldsymbol{\delta} \sim \mathscr{DM}(n, \boldsymbol{\delta})$. It can be shown that

$$\mathbb{E}\boldsymbol{X} = n\boldsymbol{\pi}$$

and

$$\text{Var}\boldsymbol{X} = \left\{1 + \frac{n-1}{\delta_+ + 1}\right\} n\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\},$$

where $\boldsymbol{\pi} = \{\delta_j/\delta_+; j = 1, \ldots, k-1\}$ and $\text{diag}(\boldsymbol{\pi})$ denotes a diagonal matrix with diagonal entries equal to $\pi_1, \ldots, \pi_{k-1}$. Note that as $\delta_+ \to \infty$ the variance-covariance matrix of the Dirichlet-multinomial distribution reduces to $n\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\}$, that is, the variance-covariance matrix of the multinomial distribution. In any other case extra variation is introduced compared to standard multinomial sampling, a well known property of the Dirichlet-multinomial distribution [see e.g. Neerchal and Morel (1998)].

Consider now that a matrix of (estimated) read counts is available for two different conditions, consisting of n_1 and n_2 replicates. Given two hyper-parameter vectors $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2$, let

$$\boldsymbol{X}_i^{(g)}|n_{1i}, \boldsymbol{\delta}_1 \sim \mathscr{DM}(n_{1i}, \boldsymbol{\delta}_1), \quad \text{independent for } i = 1, \ldots, n_1$$

$$\boldsymbol{Y}_j^{(g)}|n_{2j}, \boldsymbol{\delta}_2 \sim \mathscr{DM}(n_{2j}, \boldsymbol{\delta}_2), \quad \text{independent for } j = 1, \ldots, n_2,$$

where $\boldsymbol{X}_i^{(g)}$, $\boldsymbol{Y}_j^{(g)}$ denote two independent vectors of (estimated) number of reads for the transcripts of gene $g = 1, \ldots, G$ for replicate $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$ for the first and second condition, respectively. Obviously, n_{1i} and n_{2j} denote the total number of reads generated from gene g for the first and second condition for replicates i and j.

In this context, DTU inference is based on comparing the hyper-parameters of the Dirichlet-Multinomial distribution. Note that $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ is proportional to the average expression level of the specific set of transcripts. Typically, there are large differences in the scale of these parameters, thus their direct comparison does not reveal any evidence for DTU. For this reason, it is essential to reparametrize the model as follows:

$$\boldsymbol{\delta}_1 = d_1\boldsymbol{g}_1 \tag{4}$$

$$\boldsymbol{\delta}_2 = d_2\boldsymbol{g}_2, \tag{5}$$

where $d_1 > 0$, $d_2 > 0$ and $\boldsymbol{g}_1 = (g_{11}, \ldots, g_{1k})$, $\boldsymbol{g}_2 = (g_{21}, \ldots, g_{2k})$, with $\sum_{i=1}^{k} g_{1i} = \sum_{i=1}^{k} g_{2i} = 1$ and $g_{1i}, g_{2i} > 0, i = 1, \ldots, k$.

In this case, DTU inference is based on comparing the null model:

$$\mathscr{M}_0 : \boldsymbol{g}_1 = \boldsymbol{g}_2$$

versus the full model where

$$\mathscr{M}_1 : \boldsymbol{g}_1 \neq \boldsymbol{g}_2.$$

A likelihood ratio test is implemented in the DRIMSeq package for testing the hypothesis of the null versus the full model. In this work, we propose to compare the two models by applying approximate Bayesian model selection techniques. In particular, a priori it is assumed that

$$d_i \sim \mathscr{E}(\lambda), \quad \text{independent for } i = 1, 2 \tag{6}$$

$$\boldsymbol{g}_i \sim \mathscr{D}(1, \ldots, 1) \quad \text{independent for } i = 1, 2, \tag{7}$$

and furthermore d_i and $\boldsymbol{g}_j$ are mutually independent.

In order to perform Bayesian model selection, the Bayes factor (Kass and Raftery, 1995) of the null against the full model is approximated using a two stage procedure. At first, the posterior distribution of each model is approximated using Laplace's approximation (Laplace, 1774, 1986), a well established practice for approximating posterior moments and posterior distributions (Tierney and Kadane, 1986; Tierney et al., 1989; Azevedo-Filho and Shachter, 1994; Raftery, 1996). Then, the logarithm of marginal likelihoods of $\mathscr{M}_0$ and $\mathscr{M}_1$ are estimated using independent samples from the posterior distribution via self-normalized sampling importance resampling (Gordon et al., 1993). Finally, the posterior probabilities $p(\mathscr{M}_0|\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)})$, and $p(\mathscr{M}_1|\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)})$ are estimated assuming equally weighted prior probabilities.

Denote by $\boldsymbol{g}_0$ the common value of $\boldsymbol{g}_1$, $\boldsymbol{g}_2$ in model $\mathscr{M}_0$. Let $\boldsymbol{u}_0 = (\boldsymbol{g}_0, d_1, d_2) \in \mathscr{U}_0$, $\boldsymbol{u}_1 = (\boldsymbol{g}_1, \boldsymbol{g}_2, d_1, d_2) \in \mathscr{U}_1$ denote the parameters associated with models $\mathscr{M}_0$ and $\mathscr{M}_1$, respectively. Obviously, the underlying parameter spaces are defined as $\mathscr{U}_0 = \mathscr{P}_{K_g-1} \times (0, +\infty)^2$ and $\mathscr{U}_1 = \mathscr{P}^2_{K_g-1} \times (0, +\infty)^2$. The marginal likelihood of data under model $\mathscr{M}_j$, is defined as

$$f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\mathscr{M}_j) = \int_{\mathscr{U}_j} f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\boldsymbol{u}_j) f(\boldsymbol{u}_j|\lambda) \mathrm{d}\boldsymbol{u}_j, \quad j = 0, 1.$$

According to the basic importance sampling identity, the marginal likelihood model can be evaluated using another density ϕ, which is absolutely continuous on $\mathscr{U}_j$, as follows

$$f(\boldsymbol{x}, \boldsymbol{y}|\mathscr{M}_j) = \int_{\mathscr{U}_j} \frac{f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\boldsymbol{u}_j) f(\boldsymbol{u}_j|\lambda)}{\phi(\boldsymbol{u}_j)} \phi(\boldsymbol{u}_j) \mathrm{d}\boldsymbol{u}_j.$$

The minimum requirement for ϕ is to satisfy $\phi(\boldsymbol{u}_j) > 0$ whenever $f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\boldsymbol{u}_j) f(\boldsymbol{u}_j|\lambda) > 0$. Assume that a sample $\{\boldsymbol{u}^{(i)}; i = 1, \ldots, n\}$ is drawn from $\phi(\cdot)$. Then, the importance sampling estimate of the marginal likelihood is

$$\widehat{f}(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\mathscr{M}_j) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\boldsymbol{u}_j^{(i)}) f(\boldsymbol{u}_j^{(i)}|\lambda)}{\phi(\boldsymbol{u}_j^{(i)})}, \quad j = 0, 1.$$

The candidate distribution ϕ is the approximation of the posterior distribution according to the Laplace's method. It is well known that basic importance sampling performs reasonably well in cases that the number of parameters is not too large. However, it can be drastically improved using sequential Monte Carlo methods, such as sampling importance resampling (Gordon et al., 1993; Liu and Chen, 1998). The `R` package `LaplacesDemon` (Statisticat and LLC., 2016) is used for this purpose.

Finally, the posterior probability of the DTU model is defined as

$$p_g = \mathbb{P}(\mathscr{M}_1|\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}) \propto f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\mathscr{M}_1) P(\mathscr{M}_1), \quad g = 1, \ldots, G, \tag{8}$$

by also assuming equally weighted prior probabilities, that is, $P(\mathscr{M}_1) = P(\mathscr{M}_0) = 0.5$. Note that the Bayes Factor of the null against the full model is then given by

$$B_{01}^{(g)} = \frac{\mathbb{P}(\mathscr{M}_0|\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)})}{\mathbb{P}(\mathscr{M}_1|\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)})} = \frac{f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\mathscr{M}_0)}{f(\boldsymbol{x}^{(g)}, \boldsymbol{y}^{(g)}|\mathscr{M}_1)}, \quad g = 1, \ldots, G$$

since the prior odds ratio is equal to one.

In case that low expressed transcripts are included in the computation, the Laplace approximation faces many convergence problems. We have found that this problem can be alleviated by pre-filtering low expressed transcripts, as also pointed out by Soneson et al. (2015).

4 Bayesian FDR control for the problem of DTU

In this section we consider various decision rules in order to control the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995; Storey, 2003; Müller et al., 2004, 2006). Decision rules (9) and (11) are taking into account the whole set of genes and make use of the raw and transformed posterior probabilities, respectively. Intuitively, the transformation of posterior probabilities prioritizes genes consisting of transcripts with large changes in their expression. Decision rules (10) and (12) are based on filtering the output of (9) and (11) according to a trust region.

A decision rule based on the raw gene-level posterior probabilities of DTU, as defined in Equations (3) and (8), is the following.

$$d_{1g} = \begin{cases} 1, & \widehat{p}_g \geqslant 1 - \alpha \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Note that for the problem of inferring DTE the decision rule (9) is the one used by Leng et al. (2013). However, the cjBitSeq model takes into account changes to any subset of transcripts within a gene, thus, (9) may identify a large number of genes consisting of relatively small changes in low expressed transcripts. A more conservative choice will focus our attention to the dominant transcripts, where more reads are available and potentially the results will be more robust.

Next we define a filtering of the output based on a "trust region." Let i and j denote the estimated dominant transcripts in condition A and B, respectively. The trust region corresponds to the subset of genes where the relative ordering of estimated expression levels of dominant transcript switches, that is,

$$G_0 = \{g = 1, \ldots, G : (\widehat{\theta}_i^{(g)} - \widehat{\theta}_j^{(g)})(\widehat{w}_i^{(g)} - \widehat{w}_j^{(g)}) < 0\}.$$

Switching events between dominant transcripts have been proposed as a major source of DTU in real RNA-seq data (Gonzàlez-Porta et al., 2013).

Note that in the previous expression we used the notation of transcript expression levels according to cjBitSeq. For BayesDRIMSeq $\boldsymbol{\theta}$ and $\boldsymbol{w}$ should be replaced by $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, respectively. The decision rule which corresponds to filtering (9) according to G_0 is the following:

$$d_{2g} = \begin{cases} 1, & \widehat{p}_g \geqslant 1 - \alpha \text{ and } g \in G_0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Note that decision rules d_1 and d_2 are solely based on the posterior probabilities of gene DTU and the trust region, respectively. However, it makes sense to also take into account additional information, such as the magnitude of the change of the within gene relative transcript expression, which is a by-product of our algorithm.

In order to clarify this, consider the following example. Assume that genes g_1 and g_2 both consist of two transcripts. For g_1, let $\theta_1^{(g_1)} = 0.1$, $\theta_2^{(g_1)} = 0.9$ and $w_1^{(g_1)} = 0.9$, $w_2^{(g_1)} = 0.1$. For g_2, let $\theta_1^{(g_2)} = 0.4$, $\theta_2^{(g_2)} = 0.6$ and $w_1^{(g_2)} = 0.6$, $w_2^{(g_2)} = 0.4$. Furthermore, assume that the posterior evidence of DE is the same for both genes, that is, $\widehat{p}_{g_1} = \widehat{p}_{g_2} = p$. In the case that the posterior probability p is sufficiently large, genes g_1 and g_2 will be given the same importance in our discovery list. Note however that for gene g_1 the absolute change in relative expression is 4 times larger than for gene g_2. Ideally, we would like our discovery list to rank higher gene g_1 than gene g_2. This is achieved using the following FDR control procedure.

Consider any (Bayesian) method that for each gene yields an estimate of the posterior probability of DTU per gene p_g, $g = 1, \ldots, G$.

- For a given permutation $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_G)$ of $\{1, 2, \ldots, G\}$ and let $q_g = p_{\tau_g}$, $g = 1, \ldots, G$.
- Define: $r_g = \frac{\sum_{j=1}^{g} 1 - q_j}{g}$, $g = 1, \ldots, G$.
- For $0 < \alpha < 1$, consider the decision rule:

$$d_{3g} = \begin{cases} 1, & 1 \leqslant g \leqslant g^* \\ 0, & g^* + 1 \leqslant g \leqslant G \end{cases} \tag{11}$$

  where $g^* := \max\{g = 1, \ldots, G : r_g \leqslant \alpha\}$.
- $\widehat{\mathbb{E}}(\text{FDR}|\text{data}) = \frac{\sum_{j=1}^{g^*} 1 - q_j}{g^*} \leqslant \alpha$

Here we mention that in the original implementation of cjBitSeq for the DTE problem, the permutation τ was defined as the one that orders the posterior probabilities of transcript DE in decreasing order.

The permutation that takes into account the previously described concept of magnitude change is defined as follows. Let $\rho_g = \max |\widehat{\theta}_k^{(g)} - \widehat{w}_k^{(g)}|$, $k = 1, \ldots, K_g$, where $\widehat{\theta}_k^{(g)}$ and $\widehat{w}_k^{(g)}$ denote the posterior mean estimates of within gene transcript expression for a given transcript k of gene g. Consider the permutation $\boldsymbol{\tau} = (\tau_1, \tau_2 \ldots, \tau_G)$ that orders the set $\{\rho_g; g = 1, \ldots, G\}$ in decreasing order, that is:

$$\rho_{\tau_1} \geqslant \rho_{\tau_2} \geqslant \ldots \geqslant \rho_{\tau_G}.$$

Finally, we combine decision rule d_3 with the trust region G_0 to obtain our final decision rule, that is,

$$d_{4g} = \begin{cases} 1, & 1 \leqslant g \leqslant g^* \text{ and } g \in G_0 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

5 Simulation study

In order to assess the performance of the proposed methods and decision rules as well as to compare against existing models, a set of simulation studies is used. Instead of setting up our own simulation scenarios, we followed the pipeline introduced in the recent study of Soneson et al. (2015), where a large number of count-based method is being benchmarked. Synthetic RNA-seq reads are generated from the Drosophila Melanogaster and Homo Sapiens transcriptomes using the RSEM-simulator (Li and Dewey, 2011). The model parameters for RSEM-simulator were estimated from real datasets using a Negative Binomial model described in Soneson and Delorenzi (2013). The transcriptomes of these two organisms exhibit strong differences as illustrated in Figure 3. The average number of transcripts per gene is considerably smaller for fruit fly, however the transcripts are longer than for human [see also Supplementary Table 1 of Soneson et al. (2015)].

Following Soneson et al. (2015), for each organism we simulated 3 replicates per condition. Each replicate consists of 25 million paired-end reads with length 101 base-pairs. Differential transcript usage was introduced for 1000 genes, by reversing the relative abundance of the two most abundant transcripts in one of

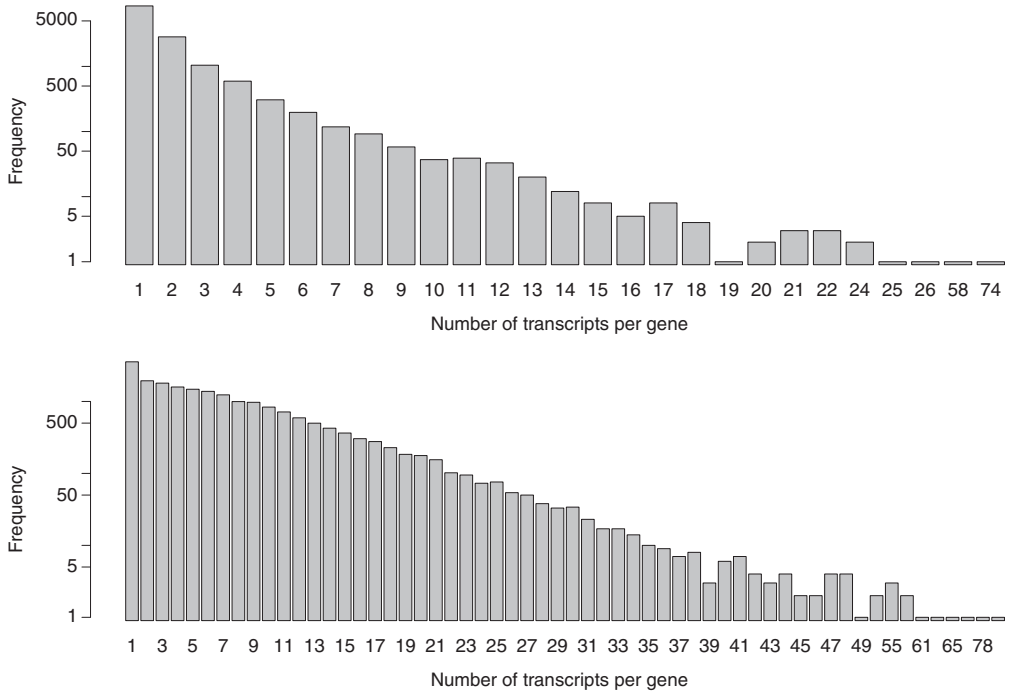

Figure 3: Frequencies (in log scale) of number of annotated transcripts per gene for drosophila (up) and human (down). The total number of genes and transcripts is 13937 and 26951 for drosophila and 20410 and 145342 for human, respectively.

the two conditions. The total number of reads for each transcript may or may not be equal across conditions. If the total number of reads generated from a gene is constant, no gene-level differential expression is evident. For the drosophila reads no gene-level differential expression was introduced. For human reads both cases are considered. Finally, the simulated reads are mapped to the genome or transcriptome with Tophat2 (Trapnell et al., 2009) and Bowtie2 (Langmead et al., 2009), respectively. Cufflinks and HTSeq used the alignment files produced by Tophat2, while BitSeqVB and cjBitSeq use the alignment produced from Bowtie2, allowing a maximum of 100 hits per read. The count matrix used as input to DEXSeq is estimated using the default HTSeq method, while BitSeqVB is used for input to edgeR, limma, DRIMSeq and BayesDRIMSeq.

5.1 Comparison of Bayesian decision rules

Figure 4 displays the power versus achieved FDR using the decision rules d_k; $k = 1, 2, 3, 4$ for the three simulated datasets. Each rule was evaluated at four typical values of expected FDR levels, $\alpha = 0.01, 0.025, 0.05, 0.1$, which are shown as dashed vertical lines. The plotted points correspond to the achieved FDR (x axis) and the proportion of true discoveries (y axis). The ability of each decision rule to control the FDR depends on the distance of each point from the corresponding vertical line: the closer, the better. On the other hand, a decision rule with higher y values is more powerful.

For cjBitSeq (upper panel) we conclude that the trust-region adjusted rules d_2 and d_4 achieve lower FDRs which are quite close to the expected values. However, note that d_4 yields better power compared to d_2, especially for the human datasets. BayesDRIMSeq is shown in middle and lower panel of 4. At the second panel of Figure 4 we have applied BayesDRIMSeq by filtering out transcripts with average number of reads less than 20. The results corresponding to the full set of transcripts (no pre-filtering) are shown at the lower panel of 4. We conclude that isoform pre-filtering is essential in order to achieve reasonable control of FDR in the case of human data. Note also that under isoform pre-filtering the trust region does not have a high impact on BayesDRIMSeq.

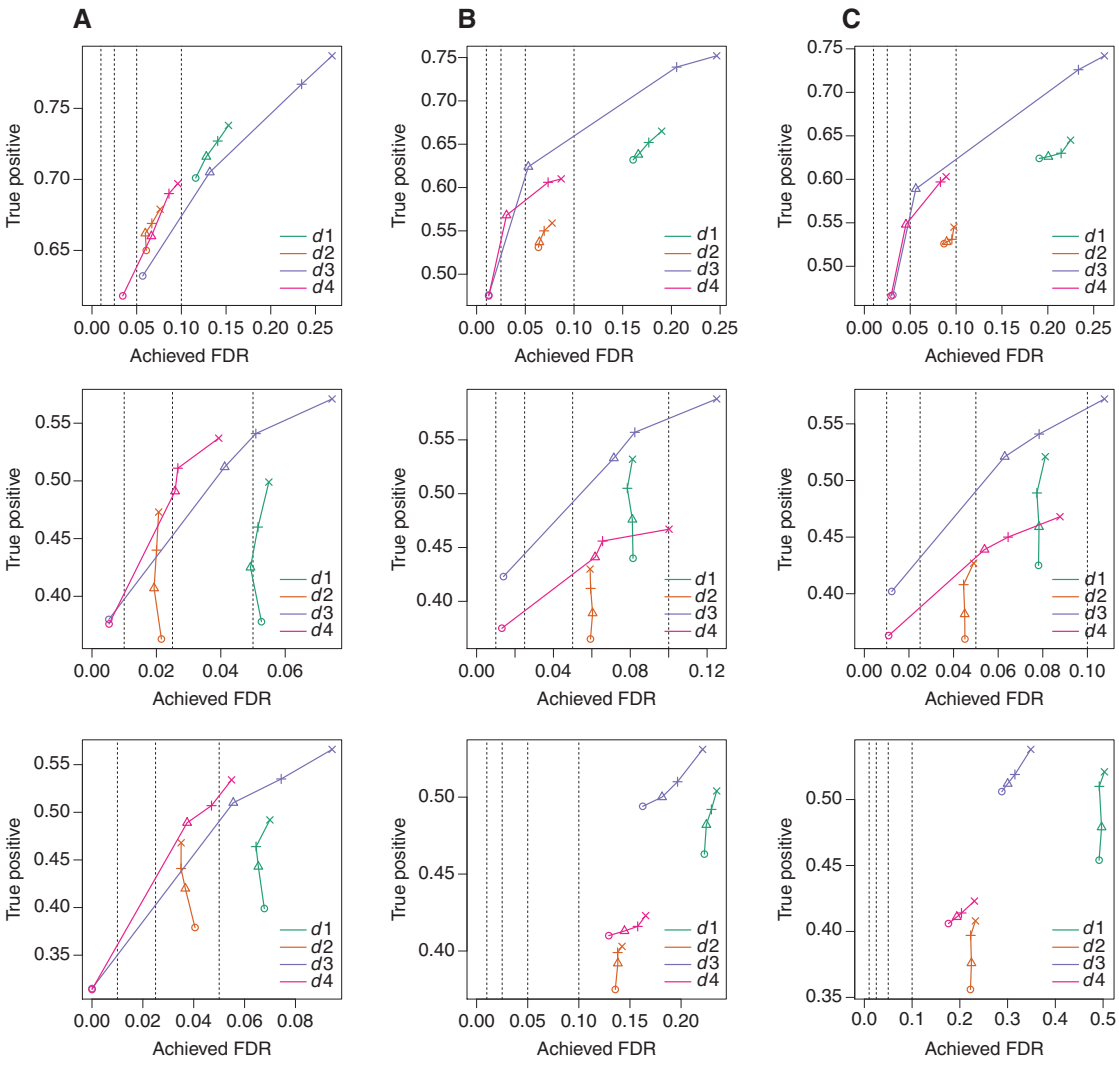

Figure 4: Power versus achieved FDR plot using the decision rules d_1, d_2, d_3 and d_4 for cjBitSeq (1st row) and BayesDRIMSeq with (second row) and without (third row) isoform pre-filtering on the simulated data. The vertical dashed lines show the expected FDR level (0.01, 0.025, 0.05, 0.1). (A) Drosophila, (B) human without DTE and (C) human with DTE.

5.2 Comparison against existing methods

For cuffdiff, DRIMSeq, edgeR and limma we use the gene-level p-values at the ROC and precision/recall plots and the adjusted q-values at the power versus achieved FDR plot. However, dexSeq reports only the adjusted q-values, hence this method is not shown at ROC and precision/recall curves. Note that for all these methods, the adjusted q-values correspond to the Benjamini and Hochberg (1995) FDR control procedure. For cjBitSeq we used the raw FDR rate (11) at the ROC and precision/recall curves and the adjusted FDR (12) at the power versus achieved FDR plots. For BayesDRIMSeq we used the raw FDR rate (11) at all plots, after pre-filtering isoforms with an average number of reads less than 20.

The performance measures of the evaluated methods are shown in Figure 5. Comparing results for the two organisms, it is clear that edgeR, limma, dexSeq and frequentist DRIMSeq exhibit large differences in their ability to control the FDR. In particular, these methods exhibit significantly larger False Discovery rates for the human datasets compared to drosophila. On the other hand, cjBitSeq and BayesDRIMSeq are able to produce consistent results in all cases, being able at the same time to control the FDR within the 0, 0.1 area.

More specifically, for the drosophila example observe that cjBitSeq exhibits smaller achieved FDR rate and larger True Positive Rate compared to DRIMSeq, dexSeq and edgeR. BayesDRIMSeq achieves even smaller

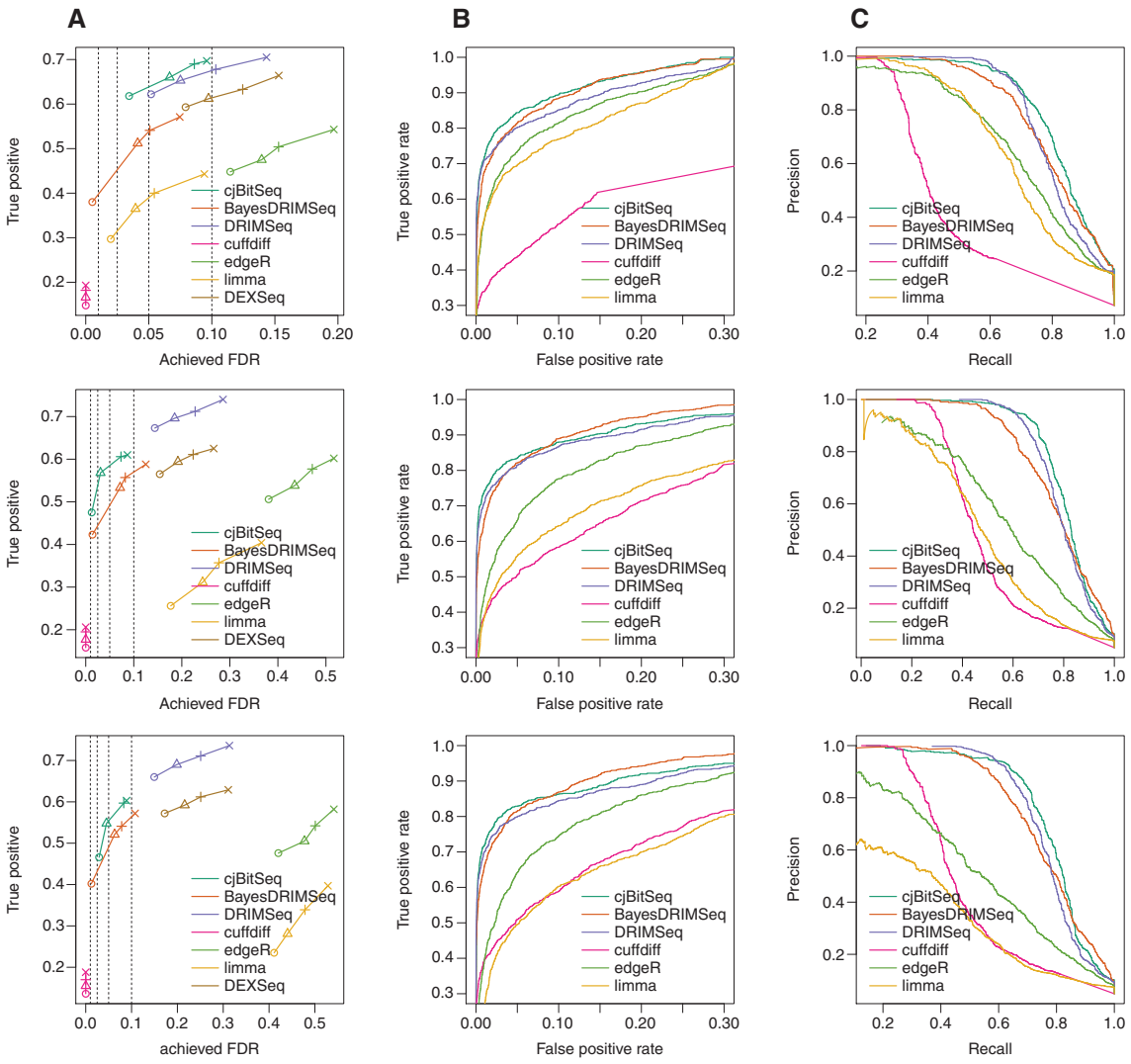

Figure 5: Performance measures for drosophila (1st row), human without DTE (2nd row) and human with DTE (3rd row). (A) Power versus achieved FDR plot. The vertical dashed lines show the expected FDR level (0.01, 0.025,0.05,0.1). (B) ROC curve. (C) Precision/recall curve.

FDR rates but the number of True Positives is reduced compared to cjBitSeq. For the human examples we conclude that cjBitSeq exhibit almost similar performance in terms of FDR control, however the former is able to discover a larger number of DTU genes in both cases. DRIMSeq and dexSeq achieve FDR rates between (0.12, 0.30) but DRIMSeq also achieves larger True Positive Rates compared to dexSeq. Cuffdiff exhibits an almost perfect control of the FDR, at the cost of substantially reduced power. The ROC and precision/recall curves, shown at Figure 5(B) and (C) respectively, suggest that cjBitSeq and DRIMSeq are consistently ranked higher than other methods. Overall, we conclude that cjBitSeq outperforms all other methods.

The run-time per method is illustrated in Figure 6, with respect to the maximum amount of virtual memory used by each process. For the counting-based methods, the main computational burden of the two-stage pipeline is due to the first stage (that is, either HTSeq or BitSeqVB). DRIMSeq, edgeR, limma which used BitSeqVB as input exhibit nearly identical computing performance so only DRIMSeq is shown. Compared to the counting-based methods, cjBitSeq requires longer computing times, which should be expected given that cjBitSeq performs MCMC sampling on the space of all possible configurations of each transcript using as input the read alignments. However, note that cjBitSeq is quite efficient with respect to the memory used and that both memory and computing time vigorously scale with the number of available cores. Therefore, it is

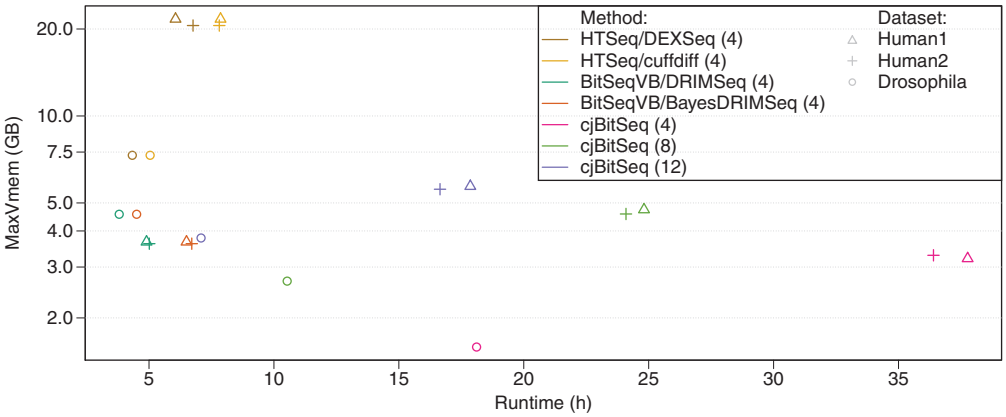

Figure 6: Wall clock runtime versus maximum value (in log-scale) of virtual memory used. The number of cores used by each process is shown in parenthesis. For each dataset the total number of reads is equal to 150 millions.

suggested to run cjBitSeq using at least 8 cores, since the memory requirements stay within reasonable levels. Finally, we mention that isoform pre-filtering is also essential for the computing time of BayesDRIMSeq. In case where no filtering takes place, the wallclock time is increased almost 2.5 times for drosophila and 4.3 times for the human datasets.

6 Adenocarcinoma dataset

In this section we benchmark the new Bayesian methods against DRIMSeq using real RNA-seq data from human lung normal and adenocarcinoma samples from six Korean female nonsmoking patients (Kim et al., 2013). The data corresponds to samples from GSM927308 to GSM927319 and was downloaded from NCBI's Gene Expression Omnibus (GEO) under the accession number GSE37764: SRR493937, SRR493939, SRR493941, SRR493943, SRR493945, SRR493947, SRR493949, SRR493951, SRR493953, SRR493955, SRR493957, SRR493959.

The data consist of paired-end reads with length equal to 78 base pairs which were mapped to the reference transcriptome using Bowtie2. The overall alignment rates and the total number of mapped reads range between (70%, 85%) and $(22 \times 10^6, 30 \times 10^6)$, respectively. Next, BitSeq was used in order to calculate the matrix of alignment probabilities (as input to cjBitSeq) as well as to obtain a matrix of estimated counts per transcript (as input to DRIMSeq and BayesDRIMSeq).

Following Nowicka and Robinson (2016), we benchmark our methods using two comparisons: (a) a two-group comparison of 6 normal versus 6 cancer samples and (b) "mock" comparisons where 3 versus 3 samples from the normal condition are compared. For the latter scenario the expectation is to detect no DTU since replicates of the same condition are compared, although the biological variation between the replicates of the normal condition is high (as noted by Nowicka and Robinson, 2016). The results are displayed in Figure 7, using different cutoff values for controlling the FDR. For the 6 normal versus 6 cancer samples comparison (Figure 7A), we conclude that all decision rules contain a large amount of genes which overlap with DRIMSeq (green colored regions), especially for the trust-region adjusted rules d_2 and d_4. For the "mock" comparison (Figure 7B), at first note that a smaller number of DTU genes is inferred. Second, observe that the decision rule d_4 is capable of substantially reducing the number of false discoveries compared to DRIMSeq and that this number is almost zero when using $\alpha = 0.01$.

7 Discussion

In this study we exemplified the use of Bayesian methods for inferring genes with differential transcript usage. For this purpose two previously introduced models were modified and extended: cjBitSeq and a Bayesian version of DRIMSeq. After defining proper decision rules we concluded that both methods exhibit

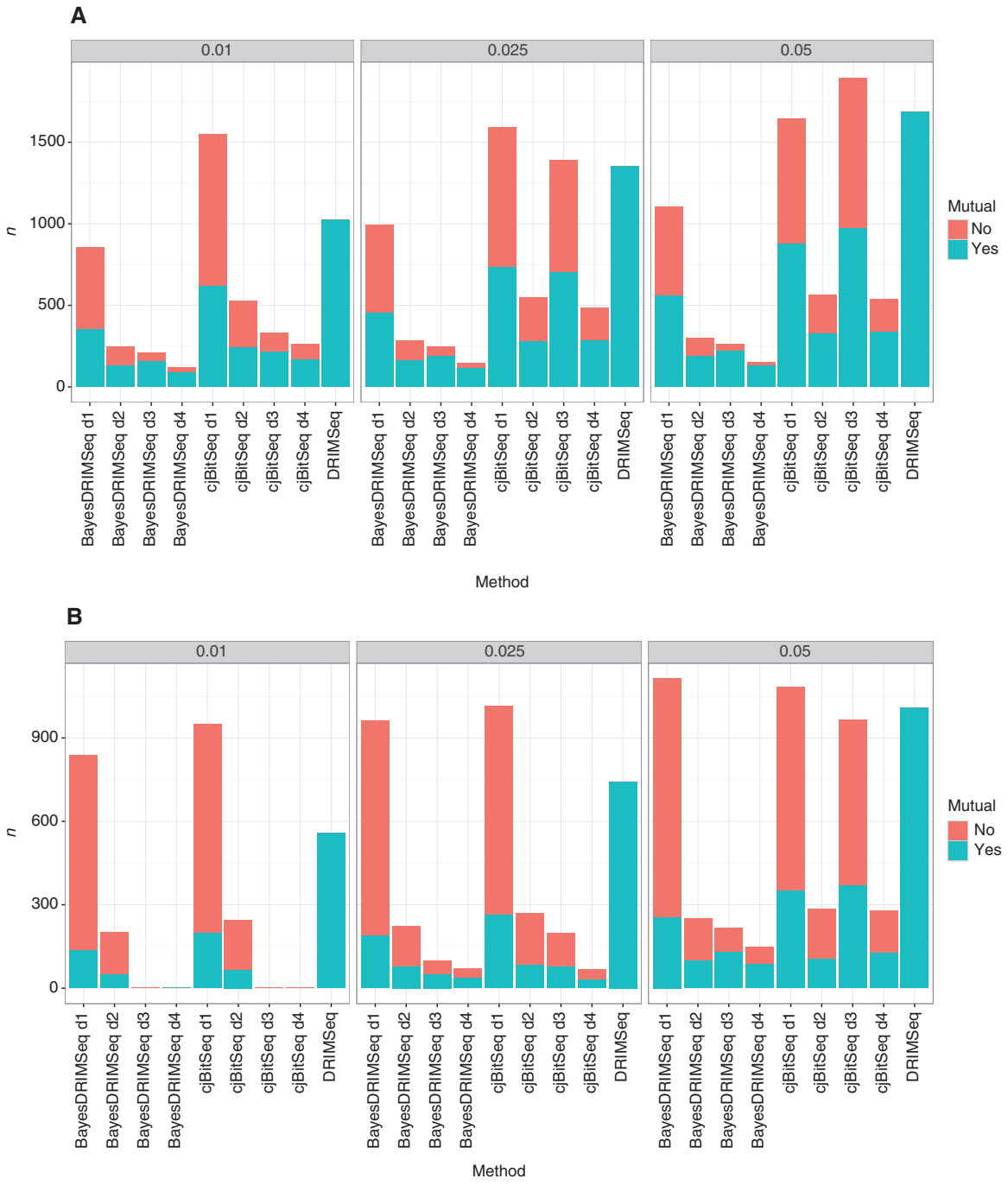

Figure 7: Inferred number of genes with DTU (n) at level $\alpha \in \{0.01, 0.025, 0.05\}$ for the comparison of 6 control and 6 tumor samples and null comparisons of 3 versus 3 control samples. For the null comparisons no differential splicing is expected. For the Bayesian methods cjBitSeq and BayesDRIMSeq all 4 decision rules are used. Green color corresponds to the number of DTU genes detected by each method that overlap with DRIMSeq and red corresponds to the opposite case. (A) 6 normal versus 6 cancer samples. (B) 3 normal versus 3 normal samples.

superior or comparable performance with other methods. This was achieved by using the decision rule defined in Equation (11), shown in the ROC and precision-recall curves. According to (11), the whole sequence of posterior probabilities is transformed with respect to the ordering of the magnitude change of relative expression between conditions. For the read-based method (cjBitSeq) FDR control is improved when the decision rule is combined with a trust region. For the count-based method (BayesDRIMSeq) FDR control

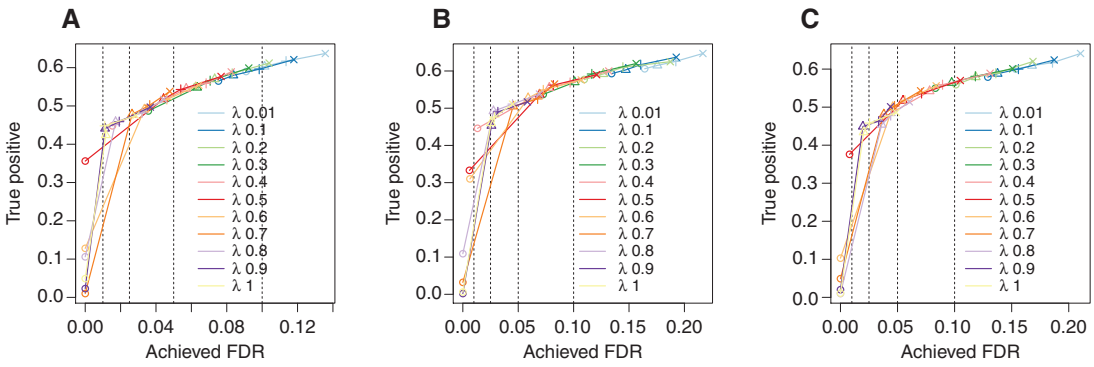

Figure 8: Prior sensitivity of BayesDRIMSeq with respect to λ.

is mainly affected by the filtering of low-expressed transcripts, as previously reported under a frequentist context by Soneson et al. (2015). BayesDRIMSeq exhibits slightly better FDR control than cjBitSeq for the drosophila dataset, however this effect is not so evident for the human datasets. In all cases cjBitSeq is more powerful than BayesDRIMSeq, but at the cost of increased computing time.

Regarding the analysis of real RNA-seq data, we compared our findings to DRIMSeq. We reported results based on a comparison of two different conditions, as well as "mock" comparisons of replicates within the same condition where no evidence of differential expression is expected. We concluded that our DTU lists contain a large number of genes also detected by DRIMSeq. Moreover, using conservative decision rules like d_4 we are able to substantially reduce the number of false discoveries when performing comparisons within the same condition.

The methods are available at `https://github.com/mqbssppe/cjBitSeq` (cjBitSeq) and `https://github.com/mqbssppe/BayesDRIMSeq` (BayesDRIMSeq). The source code for generating the simulated datasets of Soneson et al. (2015) is available from `https://github.com/markrobinsonuzh/diff_splice_paper`.

Acknowledgment: The research was supported by MRC award MR/M02010X/1, BBSRC award BB/J0 09415/1 and EU FP7 project RADIANT (grant 305626). The authors would like to acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester. Regarding BayesDRIMSeq and replication of simulations, helpful discussions with Mark Robinson, Malgorzata Nowicka and Charlotte Soneson (Institute of Molecular Life Sciences, University of Zurich) are gratefully acknowledged.

Appendix A: Prior sensitivity of BayesDRIMSeq

According to Equation (6), the prior assumptions of BayesDRIMSeq are depending on the fixed hyperparameter λ. Figure 8 displays the power versus achieved FDR curves based on the decision d_3 as a function of $\lambda \in \{0.01, 0.1, 0.2, \ldots, 1\}$ (after isoform pre-filtering). We conclude that the value $\lambda = 0.5$ offers, perhaps, the best trade-off between power and FDR control. In particular, we note that values smaller than 0.5 tend to have small power and, on the other hand, values larger than 0.5 have larger rates of False Discoveries. All results presented in the main paper correspond to $\lambda = 0.5$.

Appendix B: Using Kallisto counts

In the main text we used BitSeqVB count estimates as input to DRIMSeq and BayesDRIMSeq. According to the recent study of Hensman et al. (2015), BitSeqVB is ranked as one of the most accurate methods for estimating

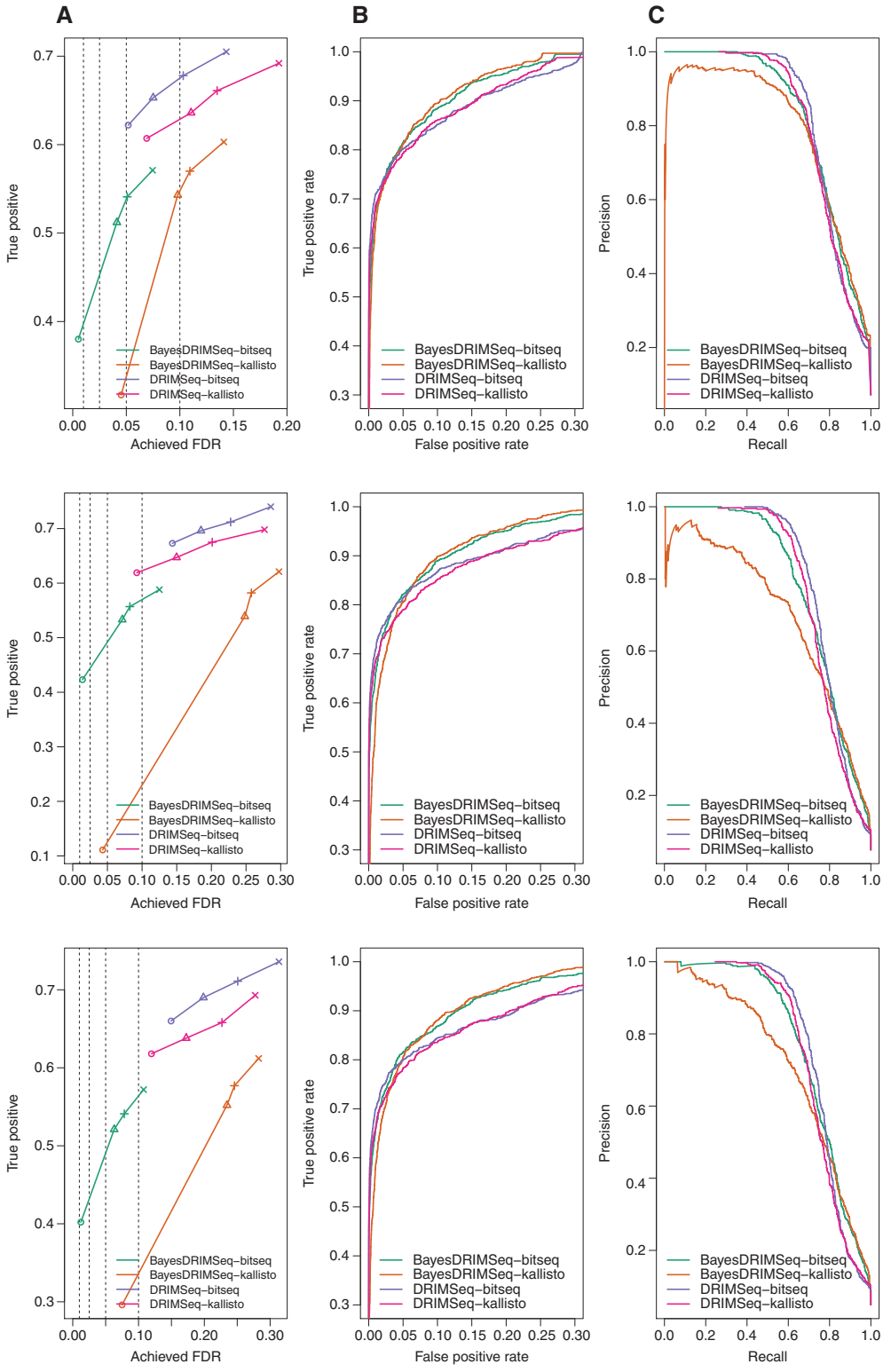

Figure 9: Comparison of DRIMSeq and BayesDRIMSeq using BitSeq and Kallisto counts for drosophila (first row) and human data (second and third row).

transcript expression levels. Since there is a variety of alternative methods for this purpose, we compare the performance when Kallisto (Bray et al., 2016) counts are being used as input. As shown in Figure 9, we conclude that in drosophila data both BayesDRIMSeq and DRIMSeq perform better when BitSeqVB counts are

used. However there is no clear ordering in the human datasets: in both cases BitSeqVB counts correspond to increased power but at the cost of slightly worse FDR calibration. Finally, ROC and precision-recall curves suggest that BitSeqVB leads to slightly increased performance for both methods.

References

Anders, S., P. T. Pyl, and W. Huber (2015): "HTSeq—a python framework to work with high-throughput sequencing data," Bioinformatics, 31, 166–169.

Anders, S., A. Reyes, and W. Huber (2012): "Detecting differential usage of exons from RNA-seq data," Genome Res., 22, 2008–2017.

Azevedo-Filho, A. and R. D. Shachter (1994): "Laplace's method approximations for probabilistic inference in belief networks with continuous variables," in *Proceedings of the tenth international conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., Burlington, MA, 28–36.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B Methodol., 57, 289–300.

Bray, N., H. Pimentel, P. Melsted and L. Pachter (2016): "Near-optimal RNA-Seq quantification," Nat. Biotechnol., 34, 525–527.

Connor, R. J. and J. E. Mosimann (1969): "Concepts of independence for proportions with a generalization of the Dirichlet distribution," J. Am. Stat. Assoc., 64, 194–206.

Endres, D. and J. Schindelin (2003): "A new metric for probability distributions," Inf. Theory IEEE Trans., 49, 1858–1860.

Gelfand, A. and A. Smith (1990): "Sampling-based approaches to calculating marginal densities," J. Am. Stat. Assoc., 85, 398–409.

Geman, S. and D. Geman (1984): "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6, 721–741.

Glaus, P., A. Honkela and M. Rattray (2012): "Identifying differentially expressed transcripts from RNA-Seq data with biological variation," Bioinformatics, 28, 1721–1728.

Gonzàlez-Porta, M., A. Frankish, J. Rung, J. Harrow and A. Brazma (2013): "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene," Genome Biol., 14, R70.

Gordon, N. J., D. J. Salmond and A. F. Smith (1993): "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE Proceedings* F (*Radar and Signal Processing*), volume 140, IET, 107–113.

Green, P. J. (1995): "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," Biometrika, 82, 711–732.

Hensman, J., P. Papastamoulis, P. Glaus, A. Honkela and M. Rattray (2015): "Fast and accurate approximate inference of transcript expression from RNA-seq data," Bioinformatics, 31, 3881–3889.

Kass, R. E. and A. E. Raftery (1995): "Bayes factors," J. Am. Stat. Assoc., 90, 773–795.

Kim, S. C., Y. Jung, J. Park, S. Cho, C. Seo, J. Kim, P. Kim, J. Park, J. Seo, J. Kim and S. Park (2013): "A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers," PLoS One, 8, e55596.

Langmead, B., C. Trapnell, M. Pop and S. Salzberg (2009): "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," Genome Biol., 10, R25.

Laplace, P. S. (1774): "Memoire sur la probabilite de causes par les evenemens," *Memoires de Mathematique et de Physique, Presentes a l'Academy Royale des Sciences, par divers Savans & lus dans ses Assemblees, Tome Sixieme*, 621–656.

Laplace, P. S. (1986): "Memoir on the probability of the causes of events (translated by S.M. Stigler, University of Chicago)," Stat. Sci., 1, 364–378.

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

Li, B. and C. N. Dewey (2011): "RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome," BMC Bioinf., 12, 323.

Liu, J. S. (1994): "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," J. Am. Stat. Assoc., 89, 958–966.

Liu, J. S. and R. Chen (1998): "Sequential Monte Carlo methods for dynamic systems," J. Am. Stat. Assoc., 93, 1032–1044.

Liu, J. S., W. H. Wong and A. Kong (1995): "Covariance structure and convergence rate of the Gibbs sampler with various scans," J. R. Stat. Soc. Ser. B Methodol., 57, 157–169.

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

Mosimann, J. E. (1962): "On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions," Biometrika, 49, 65–82.

Müller, P., G. Parmigiani and K. Rice (2006): "FDR and Bayesian multiple comparisons rules," *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics*.

Müller, P., G. Parmigiani, C. Robert, and J. Rousseau (2004): "Optimal sample size for multiple testing," *Journal of the American Statistical Association*, 99, 990–1001.

Neerchal, N. K. and J. G. Morel (1998): "Large cluster results for two parametric multinomial extra variation models," J. Am. Stat. Assoc., 93, 1078–1087.

Nicolae, M., S. Mangul, I. Mandoiu and A. Zelikovsky (2011): "Estimation of alternative splicing isoform frequencies from RNA-seq data," Algorithms Mol. Biol., 6, 9.

Nowicka, M. and M. Robinson (2016): "DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics," F1000Research, 5, 1356.

Osterreicher, F. and I. Vajda (2003): "A new class of metric divergences on probability spaces and its applicability in statistics," Ann. Inst. Stat. Math., 55, 639–653.

Papastamoulis, P., J. Hensman, P. Glaus and M. Rattray (2014): "Improved variational Bayes inference for transcript expression estimation," Stat. Appl. Genet. Mol. Biol., 13, 213–216.

Papastamoulis, P. and G. Iliopoulos (2009): "Reversible jump mcmc in mixtures of normal distributions with the same component means," Comput. Stat. Data Anal., 53, 900–911.

Papastamoulis, P. and M. Rattray (2017): "A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data," J. R. Stat. Soc. Ser. C Appl. Stat., doi:10.1111/rssc.12213.

Raftery, A. E. (1996): "Approximate Bayes factors and accounting for model uncertainty in generalised linear models," Biometrika, 83, 251–266.

Richardson, S. and P. J. Green (1997): "On Bayesian analysis of mixtures with an unknown number of components," J. R. Stat. Soc. Ser. B, 59, 731–758.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015): "limma powers differential expression analyses for RNA-sequencing and microarray studies," Nucleic Acids Res., 43, e47.

Robinson, M., D. McCarthy and G. Smyth (2010): "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, 26, 139–140.

Rossell, D., S.-O. C. Attolini, M. Kroiss and A. Stocker (2014): "Quantifying alternative splicing from paired-end RNA-sequencing data," Ann. Appl. Stat., 8, 309–330.

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

Soneson, C., K. L. Matthes, M. Nowicka, C. W. Law and M. D. Robinson (2015): "Differential transcript usage from RNA-seq data: isoform pre-filtering improves performance of count-based methods," Genome Biol., 17, 12.

Statisticat and LLC. (2016): LaplacesDemon: complete environment for Bayesian inference. URL https://CRAN.R-project.org/package=LaplacesDemon, R package version 16.0.1.

Storey, J. D. (2003): "The positive false discovery rate: A Bayesian interpretation and the q-value," Ann. Stat., 31, 2013–2035.

Tierney, L. and J. B. Kadane (1986): "Accurate approximations for posterior moments and marginal densities," J. Am. Stat. Assoc., 81, 82–86.

Tierney, L., R. E. Kass and J. B. Kadane (1989): "Fully exponential Laplace approximations to expectations and variances of nonpositive functions," J. Am. Stat. Assoc., 84, 710–716.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter (2013): "Differential analysis of gene regulation at transcript resolution with RNA-seq," Nat. Biotechnol., 31, 46–53.

Trapnell, C., L. Pachter and S. Salzberg (2009): "TopHat: discovering splice junctions with RNA-Seq," Bioinformatics, 25, 1105–1111.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010): "Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation," Nat. Biotechnol., 28, 511–515.