Gene expression

Identifying differentially expressed transcripts from RNA-seq data with biological variation

Peter Glaus1,* , Antti Honkela2,*,† and Magnus Rattray3,*,†

1School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK, 2Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, FI-00014 University of Helsinki, Finland and 3Department of Computer Science and Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield S10 2HQ, UK

ABSTRACT

Motivation: High-throughput sequencing enables expression analysis at the level of individual transcripts. The analysis of transcriptome expression levels and differential expression (DE) estimation requires a probabilistic approach to properly account for ambiguity caused by shared exons and finite read sampling as well as the intrinsic biological variance of transcript expression.

Results: We present Bayesian inference of transcripts from sequencing data (BitSeq), a Bayesian approach for estimation of transcript expression level from RNA-seq experiments. Inferred relative expression is represented by Markov chain Monte Carlo samples from the posterior probability distribution of a generative model of the read data. We propose a novel method for DE analysis across replicates which propagates uncertainty from the sample-level model while modelling biological variance using an expression-level-dependent prior. We demonstrate the advantages of our method using simulated data as well as an RNA-seq dataset with technical and biological replication for both studied conditions.

Availability: The implementation of the transcriptome expression estimation and differential expression analysis, BitSeq, has been written in C++ and Python. The software is available online from http://code.google.com/p/bitseq/, version 0.4 was used for generating results presented in this article.

Contact: glaus@cs.man.ac.uk, antti.honkela@hiit.fi or m.rattray@sheffield.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on September 6, 2011; revised on April 21, 2012; accepted on April 27, 2012

1 INTRODUCTION

High-throughput sequencing is an effective approach for transcriptome analysis. This methodology, also called RNA-seq, has been used to analyze unknown transcript sequences, estimate gene expression levels and study single nucleotide polymorphisms. Wang et al. (2009) have shown that others researchers. Mortazavi et al. (2008). RNA-seq provides many advantages over microarray technology, although effective analysis of RNA-seq data remains a challenge.

A fundamental task in the analysis of RNA-seq data is the identification of a set of differentially expressed genes or transcripts. Results from a differential expression (DE) analysis of individual transcripts are essential in a diverse range of problems such as identifying differences between tissues. Mortazavi et al. (2008) showed that estimating gene expression as a sum of transcript expression levels yields more precise results than inferring the gene expression by summing reads over all exons. As the transcript of origin is uncertain for reads aligning to shared subsequence, estimation of transcript expression levels has to be completed in a probabilistic manner. Initial studies of transcript expression used the expectation-maximization (EM)
The normalized expression samples are further used to infer expression-dependent variance hyperparameters in Step 5. Using these results, replica tes are used to infer the posterior distribution in Equation (3). In Stage 2 of the analysis, the posterior distributions of transcript expression levels from multiple conditions and within-condition expression are used to estimate the probability of positive log ratio (PPLR) between conditions, which is used to rank transcripts based on DE belief.

In many gene expression studies, expression levels are used to select genes with differences in expression in two conditions, a posterior distribution of model parameters. More advantages for transcript-level DE analyses. Furthermore, this method accounts for decreased technical reproducibility of RNA- seq for low-expressed transcripts recently reported by Katz et al. (2010) and can decrease the number of transcripts falsely identified as differentially expressed.

2 METHODS

The BitSeq analysis pipeline consists of two main stages: transcript expression estimation and DE assessment (Fig 1). For the transcript expression estimation, the input data are single-end or paired-end reads from a single sequencing run. The method produces samples from the inferred probability distribution over transcripts’ expression levels. This distribution can be summarized by the sample mean in the case that only expression level estimates are required.

The DE analysis uses posterior samples of expression levels from two or more conditions and all available replicates. The conditions are summarized by inferring the posterior distribution of condition mean expression. Samples from the posterior distributions are compared with score the transcripts based on the belief in change of expression level between conditions.

2.1 Stage 1: transcript expression estimation

The initial interest when dealing with RNA-seq data is estimation of expression levels within a sample. In this work, we focus on the transcript expression levels, mainly represented by $\theta = (\theta_1, \ldots, \theta_N)$, the relative abundance of transcripts’ fragments within the studied sample, where $M$ is the total number of transcripts. This can be further transformed into relative expression of transcripts $\theta_{rn}^m = \theta_m / (\sum_{l=1}^{n} \theta_l / n)$, where $n$ is the length of the m-th transcript. Alternatively, expression can be represented by reads per kilobase per million mapped reads, RPKMm = $\theta_m / 10^3 / n$, introduced by Mortazavi et al. (2008).

We use a generative model of the data, depicted in Figure 1 which models the RNA-seq data as independent observations of individual reads $r_i \in R = \{r_1, \ldots, r_N\}$, depending on the relative abundance of transcripts’ fragments $\theta$ and a noise parameter $\omega^m$. The parameter $\omega^m$ determines the number of reads regarded as noise and enables the model to account for unmapped reads as well as for low-quality reads within a sample.

Based on the parameter $\omega^m$, indicator variable $Z_{rn}^m = \mathbb{B}ern(\omega^m)$ determines whether read $r_n$ is considered as noise or a valid sequence.
We then precompute fragments, each read is assigned to a transcript of origin by the indicator
the assumption of reads being uniformly sequenced from the molecule
for a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads

For a valid sequence, the process of sequencing is being modelled. Under
as noise, which enables observation of low-quality and unmappable reads
a given transcript
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can
consider the observation of reads \( R = (r_1, \ldots, r_T) \) as \( N \) conditionally independent events, with each observation of a read \( r_j \) depending on the transcript (or isoform) it originated from \( \lambda_j \). The probability of sequencing a given transcript \( \lambda_j \) depends on the relative expression of fragments \( \theta \) and
the noise indicator \( Z_{m,n} \). The noise indicator variable \( Z_{m,n} \) depends on noise parameter \( \epsilon_{\text{m,n}} \) and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unmappable reads
As our parameters represent the relative expression levels in the sample, BitSeq implicitly incorporates normalization by the total number of reads or the RPKM measure, as was done when generating the results in this publication. Further more, normalization can be implemented using the normalization constant $n_{i}^{-1}$, which is constant for all transcripts of a given replicate and can be estimated prior to probabilistic modeling using, for example, a quantile-based method. 

The condition mean expression is normally distributed $\mu_{i}^{c} \sim \text{Normal}(\mu_{i}^{0},1)$ with mean $\mu_{i}^{0}$, which is empirically calculated from multiple samples and scaled precision $\lambda_{i}^{c}$. The prior distribution over pertranscript, condition-specific precision $\lambda_{i}^{c}$ is a Gamma distribution with hyperparameters $a_{c}, b_{c}$, which are fixed for a group of transcripts with similar expression level, $G$.

The hyperparameters $a_{c}, b_{c}$ determine the distribution over pertranscript precision parameter $\lambda_{i}$, which varies with the expression level of a transcript (see Supplementary Figure 3). For this reason, we inferred these hyperparameters from the dataset for various levels of expression, prior to the estimation of condition $\lambda_{i}$ and mean expression $\mu_{i}$. We used the same model as Figure 3 applied jointly to multiple transcripts with similar empirical mean expression levels $\mu_{i}^{(0)}$. We set a uniform prior for the hyperparameters, marginalized out condition means and precision, and used an MCMC algorithm to sample $a_{c}, b_{c}$. The samples of $a_{c}, b_{c}$ were smoothed by Lowess regression $G$-levelled [9] against empirical mean expression to produce a single pair of hyperparameters for each group of transcripts with similar expression level.

This model is conjugate and thus leads to a closed-form posterior distribution. This allows us to directly sample $\lambda_{i}$ and $\mu_{i}$ given each pseudo-data vector $\mathbf{y}_{m}$ constructed from the Stage 1 MCMC samples:

$$P(\theta | \mathbf{y}_{m}) \propto \prod_{i=1}^{N} \text{Gamma}(\lambda^{(i)}_{c} | a_{c}, b_{c}),$$

$$\text{Normal}(\mu_{i}^{(i)} | \mu_{i}^{(0)}, 1) \lambda_{i}^{(i)} = \frac{1}{\lambda_{i}^{(i)}},$$

where $\mu_{i}^{(i)} = \mu_{i}^{(i)} + \mathbf{y}_{m}^{(i)}$. The samples of $\mu_{i}^{(i)}$ and $\mu_{i}^{(i)}$ are used to compute the probability of expression level of transcript $m$ in condition $c$ being greater than the expression level in condition $c_2$. This is done by counting the fraction of samples in which the mean expression from the first condition is greater, that is $P(\mu_{i}^{(i)} > \mu_{i}^{(i)})=1/N \sum_{m=1}^{N} P(\mu_{i}^{(i)} > \mu_{i}^{(i)})$, which we refer to as the PPLR. Here, $n=1...N$ represents one sample from the above posterior distribution for each of $N$ independent pseudo-data vectors. Subsequently, ordering transcripts based on PPLR produces a ranking of most probable upregulated and downregulated transcripts. This kind of one-sided Bayesian test has previously been used for the analysis of microarray data [25, 26].

3 RESULTS AND DISCUSSION

3.1 Datasets

We performed experiments evaluating both gene expression estimation accuracy as well as DE analysis precision. For the evaluation of bias correction effects as well as comparison with other methods (Table 1), we used paired-end RNA-seq data from the microarray quality control (MAQC) project [27, 28] (Short Read Archive accession number SRA012427), because it contains 907 transcripts which were also analyzed by TaqMan qRT-PCR, out of which 893 matched our reference annotation. The results from qRT-PCR probes are generally regarded as ground truth expression estimates for comparison of RNA-seq analysis methods [29]. We used RefSeq refGene transcriptome annotation, assembly NCBI36/hg18 to keep results consistent with qRT-PCR data as well as previously published comparisons by Roberts et al. [29].

The second dataset used in our evaluation was originally published by Xu et al. [30] in a study focused on identification of microRNA targets and provides technical as well as biological replicates for both studied conditions. We use this data to illustrate the importance of biological replicates for DE analysis (Fig. E Supplementary Fig. 3 for biological variance) and the advantages of using a Bayesian approach for both expression inference and DE analysis (Fig. B).

For the purpose of evaluating and comparing BitSeq to existing DE analysis methods, we created artificial RNA-seq datasets with known expression levels and differentially expressed transcripts. We selected all transcripts of chromosome 1 from human genome assembly NCBI37/hg19 and simulated two biological replicates for each of the two conditions. We initially sample the expression for all replicates using the same mean relative expression and variation between replicates as were observed in the Xu et al. data estimates. Afterwards, we randomly choose one-third of the transcripts and shift one of the conditions up or down by a known fold change. Given the adjusted expression levels, we generated 300k single-end reads uniformly distributed along the transcripts. The reads were reported in Fastq format with Phred scores randomly generated according to empirical distribution learned from the SRA012427 dataset. With the error probability given by a Phred score, we generated base mismatches along the reads.
other transcript expression estimation methods: Cufflinks v0.9.3 for read distribution was analyzed using the SRA012427 dataset and can account for positional and sequence bias. The effect of correcting model or using the model proposed by Roberts implementation contains the option of using a uniform read density distribution bias correction and provide improvement over the other methods. In spite of providing slightly better results for this conversion leads to the higher correlation with qRT-PCR for within-gene expression levels, BitSeq is more accurate with BitSeq and RSEM showing almost equally good results. For the relative within-gene expression levels, BitSeq is more accurate than the other methods. In spite of providing slightly better results in Table 2. MMSEQ provides the best absolute expression accuracy for the selected transcripts.

The dataset contains three technical replicates. These were analyzed separately and the resulting estimates for each method were averaged together. Subsequently, we calculated the squared Pearson correlation coefficient ($R^2$) of the average expression estimate and the results of qRT-PCR analysis. All four methods used with the default uniform read distribution model provide similar level of accuracy with BitSeq performing slightly better than the other three methods.

Both BitSeq and Cufflinks use the same method for read distribution bias correction and provide improvement over the uniform model similar to improvements previously reported by Roberts et al. (2011). We used version 0.9.3 of Cufflinks (as used by Roberts et al.) since we found that the most recent stable version of Cufflinks (version 1.3.0) leads to much worse performance for both uniform and bias-corrected models (see Supplementary results).

Section 2.5. The RSEM package uses its own method for bias correction based on the relative position of fragments, which in this case did not improve the expression estimation accuracy for the selected transcripts.

In the case of BitSeq, the major improvement of accuracy originates from using the effective length normalization. To compare the results with qRT-PCR, the relative expression of fragments $\theta$ has to be converted into either relative expression of transcripts ($\theta^*$) or RPKM units. Using the bias-corrected effective length for this conversion leads to the higher correlation with qRT-PCR

To the Supplementary Table 1. This means that using an expression measure adjusted by the effective length, such as RPKM, is more suitable than normalized read counts for DE analysis.

We also evaluated the accuracy of the four methods using three different expression measures on simulated data. First, we compared with transcripts’ RPKM as an absolute expression measure. Second, we used relative within-gene expression in which transcript expression is the relative proportion within transcripts of the same gene. Finally, we used gene expression RPKM, the sum of transcript expression levels for each gene. The results are presented in Table 2. MMSEQ provides the best absolute expression accuracy with BitSeq and RSEM showing almost equally good results. For the relative within-gene expression levels, BitSeq is more accurate than the other methods. In spite of providing slightly better results in absolute measure, RSEM and MMSEQ show worse correlation in the relative within-gene measure as they tend to assign zero expression to some transcripts within one gene. This is most likely due to the use of maximum-likelihood parameter estimates as the starting point for the Gibbs sampling algorithm.

### Table 1. Comparison of expression estimation accuracy against TaqMan qRT/PCR data

<table>
<thead>
<tr>
<th>Read model</th>
<th>BitSeq</th>
<th>Cuff. 0.9.3</th>
<th>RSEM</th>
<th>MMSEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>0.7677</td>
<td>0.7503</td>
<td>0.7632</td>
<td>0.7614</td>
</tr>
<tr>
<td>Non-uniform</td>
<td>0.8011</td>
<td>0.8056</td>
<td>0.7633</td>
<td>0.7990*</td>
</tr>
</tbody>
</table>

The table shows the effect of non-uniform read distribution models using correlation coefficient $R^2$ of average expression from three technical replicates with the 893 matching transcripts analyzed by qRT/PCR. Highest correlation is highlighted in bold. The sequencing data (SRA012427) are part of the MAQC project and was originally published by Roberts et al. (2011). We were not able to use the default bias correction provided by MMSEQ. Instead, we provided the MMSEQ package with effective lengths computed by BitSeq bias correction algorithm to produce results for this comparison.

### Table 2. The $R^2$ correlation coefficient of estimated expression levels and ground truth

<table>
<thead>
<tr>
<th>Expression</th>
<th>Cutoff</th>
<th>BitSeq</th>
<th>Cuff. 0.9.3</th>
<th>RSEM</th>
<th>MMSEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript</td>
<td>1</td>
<td>0.994</td>
<td>0.764</td>
<td>0.905</td>
<td>0.997</td>
</tr>
<tr>
<td>Relative</td>
<td>10</td>
<td>0.945</td>
<td>0.724</td>
<td>0.876</td>
<td>0.886</td>
</tr>
<tr>
<td>Relative</td>
<td>100</td>
<td>0.963</td>
<td>0.773</td>
<td>0.946</td>
<td>0.948</td>
</tr>
<tr>
<td>Gene</td>
<td>1</td>
<td>0.994</td>
<td>0.823</td>
<td>0.936</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Three different expression measures were used: absolute transcript expression, relative within-gene transcript expression and gene expression. Comparison includes sites with at least 1 read per transcript for transcript expression, either 10 or 100 reads per gene for within-gene transcript expression and at least 1 read per gene for gene expression. The highest correlation is in bold.

3.2 Expression-level inference

Figure 4 demonstrates the ambiguity that may be present in the process of expression estimation. In Figure 4A and 4B, we show the density of samples from the posterior distribution of expression levels for two pairs of transcripts. The expression levels of transcripts uc010oho.1 and uc010ohp.1 (Fig 4A) are negatively correlated. On the other hand, transcripts uc010oho.1 and uc001bwm.3 exhibit no visible correlation (Fig 4B) in their expression-level estimates. Even though this kind of correlation does not have to imply biological significance, it does point to technical difficulties in the estimation process. These transcripts share a significant amount of sequence and significance, it does point to technical difficulties in the estimation process. These transcripts share a significant amount of sequence and

3.3 Expression estimation accuracy and read distribution bias correction

Initially, it was assumed that high-throughput sequencing produces reads uniformly distributed along transcripts. However, more recent studies show biases in the read distribution depending on the position and surrounding sequence. Roberts et al. (2011). Our generative model for transcript expression inference (Fig 3) includes a model of the underlying read distribution which is included in the $P(n_i|\theta_i=m)$ term that is calculated as a preprocessing step. The current BitSeq implementation contains the option of using a uniform read density model or using the model proposed by Roberts et al. (2011) which can account for positional and sequence bias. The effect of correcting for read distribution was analyzed using the SRA012427 dataset and results are presented in Table 1. We also compare BitSeq with three other transcript expression estimation methods: Cufflinks v0.9.3 Roberts et al. (2011), MMSEQ v0.9.18 Turro et al. (2011) and RSEM v1.1.4 Li and Dewey (2011).
We use the Xu et al. method to infer condition mean expression levels using all replicates. For comparison, we used a naive way of combining two replicates by combining the posterior distributions of expression into a single distribution. The resulting posterior distributions for both conditions were joined into one dataset for each condition. (c) Inferred posterior distribution of mean expression level for each condition using the probabilistic model in Figure 3. (d) Distribution of differences between conditions from both approaches show that the naive approach leads to overconfident conclusion.

For more details and results comparing the transcript expression estimation accuracy, please refer to Supplementary Material Section 2.3.

### 3.4 DE analysis

We use the Xu et al. dataset to demonstrate the DE analysis process of BitSeq. This dataset contains technical and biological replication for both studied conditions. We observed significant difference between biological and technical variance of expression estimates. Furthermore, the prominence of biological variance increases with transcript expression level. We illustrate how BitSeq handles biological replicates to account for this variance in Figure 5 by showing the modelling process for one example transcript given only two biological replicates for each of two conditions.

Figure 5a shows histograms of expression-level samples produced in the first stage of our pipeline. BitSeq probabilistically infers condition mean expression levels using all replicates. For comparison, we used a naive way of combining two replicates by combining the posterior distributions of expression into a single distribution. The resulting posterior distributions for both approaches are depicted in Figures 5b and 5c.

The probability of DE for each transcript is assessed by computing the difference in posterior expression distributions of the two conditions. Resulting distributions of differences for both approaches are portrayed in Figure 5d with obvious difference in the level of confidence. The naive approach reports high confidence of upregulation in the second condition, with the PPLR being 0.995. When biological variance is being considered by inferring condition mean expression levels using all replicates, the posterior distributions from replicates are joined into one dataset for each condition. (e) Inferred posterior distribution of mean expression level for each condition using the probabilistic model in Figure 3. (d) Distribution of differences between conditions from both approaches show that the naive approach leads to overconfident conclusion.

### 3.5 Assessing DE performance with simulated data

Using artificially simulated data with a predefined set of differentially expressed transcripts, we evaluated our approach and compared it with four other methods commonly used for DE analysis: DESeq v1.6.1 [Anders and Huber, 2010], edgeR v2.4.3 [Robinson et al., 2010] and baySeq v1.8.1 [Hardcastle and Kelly, 2010] were designed to operate on the gene level and Cuffdiff v1.3.0 [Trapnell et al., 2010] on the transcript level. Despite not being designed for this purpose, we consider the first three in this comparison as the use case is very similar and there are no other well-known alternatives besides Cuffdiff that would use replicates for transcript level DE analysis. All other methods besides Cuffdiff use BitSeq. Stage 1 transcript expression estimates converted to counts. Details regarding use of these methods are provided in the Supplementary material Section 2.5. Figure 6 shows the overall results as well as split into three parts based on the expression of the transcripts. The receiver-operating characterization curves were generated by averaging over five runs with different transcripts being differentially expressed and the figures are focused on the most significant DE calls with false-positive rate below 0.2.

Overall (Figure 6b), BitSeq is the most accurate method, followed first by baySeq, then edgeR and DESeq with Cuffdiff further behind. This trend is especially clear for lower expression levels (Fig. 6b and 6c). The overall performance here is fairly low because of high level of biological variance. For highest expressed transcripts (Fig. 6c), DESeq and edgeR show slightly higher true positive rate than BitSeq and baySeq, especially at larger false-positive rates. Furthermore details and more results from the DE analysis comparison can be found in Supplementary material Section 2.5.

### 3.6 Scalability and performance

As BitSeq models individual read assignments, the running time complexity of the first stage of BitSeq increases with the number of aligned reads. Preprocessing the alignments and sampling a constant number of samples scales linearly with the number of reads. However, with more reads, the data become more complex and the Gibbs sampling algorithm needs more iterations to capture the whole posterior distribution.

In Table 3, we present the running time for Stage 1, using simulated data generated from the UCSC NCB37/hg19 knownGene reference. We ran the preprocessing of the reads with a uniform read distribution model on a single CPU and sampling with four parallel chains on four Intel Xeon 3.47 GHz CPUs. We set the sampler to run until it generates 1000 effective samples for at least 95% of transcripts. At the end, almost all transcripts converged according to the $\hat{R}$ statistic. The number of iterations necessary to produce the desired amount of effective samples seems to increase logarithmically with the number of reads.
Alignments (M) 16 32 64 129 258

Table 3. Scalability and run-time complexity of BitSeq on different-sized datasets using simulated data with 9.9 up to 158.5 million paired-end reads.

<table>
<thead>
<tr>
<th>Read pairs (M)</th>
<th>Alignments (M)</th>
<th>Preprocessing (m)</th>
<th>Total time (h)</th>
<th>Convergence it.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9</td>
<td>16</td>
<td>8</td>
<td>0.55</td>
<td>5269</td>
</tr>
<tr>
<td>9.1</td>
<td>32</td>
<td>15</td>
<td>2.18</td>
<td>6900</td>
</tr>
<tr>
<td>19.8</td>
<td>64</td>
<td>29</td>
<td>5.42</td>
<td>8920</td>
</tr>
<tr>
<td>39.6</td>
<td>129</td>
<td>57</td>
<td>16.23</td>
<td>11970</td>
</tr>
<tr>
<td>79.2</td>
<td>258</td>
<td>115</td>
<td>33.19</td>
<td>15979</td>
</tr>
</tbody>
</table>

The table shows wall clock running times to preprocess the aligned reads, generate 1000 samples and full time for the sampling algorithm on four CPUs. The last row contains the estimated number of iterations needed to reach convergence for at least 95% of transcripts.

Running time of the DE analysis in Stage 2 does only depend on the number of reference transcripts, replicates and samples generated in Stage 1 for the analysis. Producing the result presented in Section 4 took 97 min on the Intel Xeon 3.47 GHz CPU.

4 CONCLUSION

We have presented methods for transcript expression level analysis and DE analysis that aim to model the uncertainty present in RNA-seq datasets. We used a Bayesian approach to provide a probabilistic model of transcriptome sequencing and to sample from the posterior distribution of the transcript expression levels. The model incorporates read and alignment quality, adjusts for non-uniform read distributions and accounts for an experiment-specific fragment length distribution in case of paired-end reads. The accuracy of inferred expression is comparable and in some cases, outperforms other competing methods. However, the major benefit of using BitSeq for transcript expression inference is the availability of the full posterior distribution which is useful for further analysis.

The inferred distributions of transcript expression levels can be further analyzed by the second stage of BitSeq for DE analysis. Given biological replicates, BitSeq accounts for both intrinsic technical noise and biological variation to compute the posterior distribution of expression differences between conditions. It produces more reliable estimates of expression levels within each condition and associates these expression levels with a degree of credibility, thus providing fewer false DE calls. We want to highlight that to make accurate DE assessment, experimental designs must include biological replication and BitSeq is a method capable of combining information from biological replicates when comparing multiple conditions using RNA-Seq data.

In our current work, we aim to reduce the computational complexity of BitSeq by replacing MCMC with a faster deterministic approximate inference algorithm and we are generalizing the model to include more complex experimental designs in the DE analysis stage.

**Funding:** European ERASysBio+ initiative project SYNERGY by the Biotechnology and Biological Sciences Research Council [BB/B004769/2 to M.R.] and Academy of Finland [135311 to A.H.]; Academy of Finland [121179 to A.H.]; and IST Programme of the European Community, under the PASCAL2 Network of Excellence [IST-2007-216888]. This publication only reflects the authors’ views.

**Conflict of Interest:** none declared.

**REFERENCES**


