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

Glaus, Peter

2012


http://hdl.handle.net/10138/37377
https://doi.org/10.1093/bioinformatics/bts260

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
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. Rice et al. 2009. As shown by other 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), understanding developmental changes (Kempkes et al. 2010) and microRNA target prediction (Xu et al. 2010). To perform an effective DE analysis, it is important to obtain accurate estimates of expression for each sample, but it is equally important to properly account for all sources of variation, technical and biological, to avoid spurious DE calls. Anderson and Huber 2011; Dabney et al. 2011; Robinson and Smyth 2007. In this contribution, we address both of these problems by developing integrated probabilistic models of the read generation process and the biological replication process in an RNA-seq experiment.

The length of the reads depends on the sequencing platform and currently typically ranges from 25 to 300 basepairs. Reads have to be either assembled into transcript sequences or aligned to a reference genome by an aligning tool, to determine the sequence they originate from. With proper sample preparation, the number of reads aligning to a certain gene is approximately proportional to the abundance of fragments of transcripts for that gene within the sample (Mortazavi et al. 2008). Counting researchers to study gene expression (Cloonan et al. 2008). However, during the process of transcription, most eukaryotic genes can be spliced into different transcripts which share parts of their sequence. As it is the transcripts of genes that are being sequenced during RNA-seq, it is possible to distinguish between individual transcripts of a gene. Several methods have been proposed to estimate transcript expression levels (Kang et al. 2011; Liao et al. 2011; Ngepants et al. 2011; Barro et al. 2011). Furthermore, Wang et al. 2010 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)
In Stage 2 of the analysis, the posterior distributions of transcript expression levels from multiple conditions and of within-condition expression are used to estimate the probability of positive log ratio (PPLR) between conditions, which is used to rank transcripts based on their DE status.

The BitSeq analysis pipeline consists of two main stages: transcript expression estimation and DE assessment. 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 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.

We use a generative model of the data, depicted in Figure 2, which models the RNA-seq data as independent observations of individual reads. 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 METHODS

The BitSeq analysis pipeline consists of two main stages: transcript expression estimation and DE assessment. For the transcript expression estimation, the 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 \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_M) \), 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 \( \hat{\theta}_{1m} = \theta_1 / \theta_m \sum_{i=1}^{m} \theta_i / \theta_i \), where \( \theta_m \) is the length of the \( m \)-th transcript. Alternatively, expression can be represented by reads per kilobase per million mapped reads (RPKM) \( \text{RPKM}_m = \frac{\theta_m}{10^6} \), introduced by Mortazavi et al. (2008).

We use a generative model of the data, depicted in Figure 2, which models the RNA-seq data as independent observations of individual reads \( r \in \mathcal{R} = (r_1, \ldots, r_N) \), depending on the relative abundance of transcripts’ fragments \( \theta \) and a noise parameter \( \sigma^2 \). The parameter \( \sigma^2 \) 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 \( \hat{\theta}_{1m} \), indicator variable \( Z_m^* = \text{Bern}(\hat{\theta}_{1m}) \) determines whether read \( r_1 \) is considered as noise or a valid sequence.
Fig. 2. Graphical representation of the RNA-seq data probabilistic model. We can consider the observation of reads $R = (r_1, ..., r_T)$ as conditionally independent events, with each observation of a read $r_i$ depending on the transcript (or isoform) it originated from $l_i$. The probability of sequencing a given transcript $l_i$ depends on the expression of fragments $\theta$ and the noise indicator $Z^{\text{noi}}_i$. The noise indicator variable $Z^{\text{noi}}_i$ depends on noise parameter $\theta^{\text{noi}}$ and indicates that the transcript being sequenced is regarded as noise, which enables observation of low-quality and unparseable reads.

For a valid sequence, the process of sequencing is being modelled. Under the assumption of reads being uniformly sequenced from the molecule fragments, each read is assigned to a transcript of origin by the indicator variable $l_i$, which is given by categorical distribution $l_i \sim \text{Cat}(\theta)$.

For a transcript, we can express the probability of an observed alignment as the probability of choosing a specific position $p$ and sequencing a sequence of given length with all its mismatches, $P(r_{l_i} | l_i = m) = P_T(p(m) | P_{r_{l_i}} | Z_{l_i})$. For paired-end reads, we compute the joint probability of the alignment of a whole pair, in which case, we also have to consider fragment length distribution $P(l_i)$. Details of alignment probability computation including optional position and sequence-specific bias correction methods are presented in Supplementary Material. For every aligned read, we also calculate the probability that the read is from neither of the aligned transcripts but is regarded as sequencing error or noise $P_{r_{l_i}}(\text{noise})$. This value is calculated by taking the probability of the least probable valid alignment corrupted with two extra base mismatches.

The joint probability distribution of the model can now be written as

$$P(R, \theta, Z^{\text{noi}}, \theta^{\text{noi}} = \theta_0) = P(\theta)P(\theta^{\text{noi}}) \prod_{i=1}^{T} P(r_{l_i} | l_i = m) \prod_{m=1}^{M} P(l_m | P_r | Z^m | P_r | \theta^m | \theta^{\text{noi}} | Z^{\text{noi}} | \theta^{\text{noi}})$$

where we use weak conjugate Dirichlet and Beta prior distributions for $\theta$ and $\theta^{\text{noi}}$, respectively. The posterior distribution of the model’s parameters given the data $R$ can be simplified by integrating over all possible values of $Z^{\text{noi}}$.

$$P(\theta, \theta^{\text{noi}} | R) \propto P(\theta)P(\theta^{\text{noi}}) \prod_{i=1}^{T} P(r_{l_i} | l_i = m) \prod_{m=1}^{M} P(l_m | P_r | Z^m | P_r | \theta^m | \theta^{\text{noi}} | Z^{\text{noi}} | \theta^{\text{noi}}) \prod_{l_i = 0} \left[ P_{r_{l_i}}(\text{noise}) (1 - \theta^{\text{noi}}) \right].$$

According to the model, any read can be a result of sequencing either strand

$$\begin{align*}
P_{r_{l_i}}( & \text{seq}) \mid Z_{l_i} = \text{seq} \sim \text{Bin}(n_{l_i}, \nu) \text{ and } P_{r_{l_i}}(\text{seq}) \mid Z_{l_i} = \text{noi} \sim \text{Bin}(n_{l_i}, \nu^{\text{noi}}) \end{align*}$$

Fig. 3. Graphical model of the biological variance in transcript expression experiment. For replicate $i$, condition $c$ and transcript $m$, the observed log-expression level $\mu (c, i, m)$ is normally distributed around the normalized condition mean expression $\mu (c, 0, m)$ with biological variance $1/\lambda (c, m)$. The condition mean expression $\mu (c, 0, m)$ for each condition is normally distributed with overall mean expression $\mu (0, 0, m)$ and scaled variance $1/(\lambda (0, m) \lambda (0, 0))$. The inverse variance, or precision $\lambda (c, m)$, for a given transcript $m$ follows a Gamma distribution with expression-dependent hyperparameters $\alpha_m, \beta_m$, which are constant for a group of transcripts $G$ with similar expression statistic is an estimate of a possible scale reduction of the marginal posterior variance and provides a measure of usefulness of producing more samples. We use the marginal posterior variance estimate and between chain variance statistic as the effective number of samples for each transcript as described by (Sellers et al. 2008), to determine the number of iterations needed for convergence.

Posterior samples of $\theta$ provide an assessment of the abundance of individual transcripts. We can also provide an accurate point estimate of the expression levels through the mean of the posterior, the probability distribution provides a measure of confidence for the results, which can be used in further analyses.

2.2 Stage 2: combining data from multiple replicates and estimating DE

To identify transcripts that are truly differentially expressed, it is necessary to account for biological variation by using replication for each experimental condition. Our method summarizes these replicates by estimating the biological variance and inferring permutation Mean expression levels for each transcript. During the DE analysis, we consider the logarithm of the expression levels $y_m = \log_2(\text{seq})$. The model for data originating from multiple replicates is illustrated in Figure 2B. We use a hierarchical log-normal model of within-condition expression. The prior over the biological variance is dependent on the mean expression level across conditions and the prior parameters (hyper-parameters) are learned from all of the data by fitting a nonparametric regression model. We fit a model for each gene using the expression estimates from Stage 1.

A novel aspect of our Stage 2 approach is that we fit models to posterior samples obtained from the MCMC simulation from Stage 1, which can be considered ‘pseudo-data’ representing expression corrupted by technical noise. A pseudo-data vector is constructed using a single MCMC sample for each replicate across all conditions. The posterior distribution over per-condition means is inferred for each pseudo-data vector using the model in Figure 2A described below. We then use Bayes’ rule to combine the evidence from each pseudo-data vector and determine the probability of DE. This approach allows us to account for the intrinsic technical variance in the data; it is also computationally tractable because the model for a single pseudo-data vector is conjugate and therefore inference can be performed exactly. This effectively regularizes our variance estimate in the case that the number of replicates is low. As a result, this approach provides improved control of error rates for weakly expressed transcripts where the technical variance is large.

For a condition $c$, we assume $R_c$ replicate datasets. The log expression from replicate $r$, $\mu_{(c, r, m)}$, is assumed to be distributed according to a normal distribution with condition mean expression $\mu_{(c, 0, m)}$ normalized by replication-specific constant $\tilde{\sigma}_{(c, r)}$ and precision $\lambda_{(c, r, m)} = \tilde{\sigma}_{(c, r)}^{-2} \sim \text{Gamma}(\alpha_{(c, r), m}, \beta_{(c, r), m})$. The expression mean for each group is estimated using the mean expression of the posterior, which is calculated as $\bar{\mu}_{(c, r, m)} = \frac{1}{R_c} \sum_{r=1}^{R_c} \mu_{(c, r, m)}$.
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 $n^{(m)}$, 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 (Robinson and Oshlack 2010) or any other suitable technique.

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

The hyperparameters $\alpha_c, \beta_c$ determine the distribution over pertranscript precision parameter $\lambda^{(c)}$, which varies with the expression level of a transcript (see Supplementary Figure S3). For this reason, we inferred these hyperparameters from the dataset for various levels of expression, prior to the estimation of precision $\alpha_c$ and mean expression $\mu_{(ac)}$. We used the same model as Figure 3 applied jointly to multiple transcripts with similar empirical mean expression levels $\mu_{(ac)}^{(0)}$. We set a uniform prior for the hyperparameters, marginalized out condition means and precision, and used an MCMC algorithm to sample $\alpha_c, \beta_c$. The samples of $\alpha_c, \beta_c$ were smoothed by Lowess regression (Cleveland 1981) against empirical mean expression levels $\mu_{(ac)}^{(0)}$. We used an MCMC algorithm to sample 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 to directly sample $\lambda^{(c)}$ and $\mu_{(ac)}$ given each pseudo-data vector $y_{ac}$ constructed from the Stage 1 MCMC samples:

$$P(\mu_{(ac)}, \lambda^{(c)}|y_{ac}) \propto \prod_{c} \text{Gamma}(\lambda^{(c)}|\alpha^{(c)}, \beta^{(c)}) \times \text{Norm}(\mu_{(ac)}|\mu_{(ac)}^{(0)}, \frac{1}{\lambda^{(0)}}).$$

The second dataset used in our evaluation was originally published by Xu et al. (2011) in a study focused on identification of microRNA targets and provides technical as well as biological replicates for both stated conditions. We use this data to illustrate the importance of biological replicates for DE analysis (Fig. 4 Supplementary Fig. S3) for biological variance) and the advantages of using a Bayesian approach for both expression inference and DE analysis (Fig. 5).

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 GRCh37/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.

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 (Roberts et al., 2011). We used RefSeq refGene transcriptome annotation, NCBI36/hg18 to keep results consistent with qRT-PCR data as well as previously published comparisons by Roberts et al. (2011).

The second dataset used in our evaluation was originally published by Xu et al. (2011) in a study focused on identification of microRNA targets and provides technical as well as biological replicates for both stated conditions. We use this data to illustrate the importance of biological replicates for DE analysis (Fig. 4 Supplementary Fig. S3) for biological variance) and the advantages of using a Bayesian approach for both expression inference and DE analysis (Fig. 5).

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 GRCh37/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 Fasq 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.
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 Shi et al. (2008) and was also used by Roberts et al. (2011) due to an error in an external package used for the bias correction. Instead, we provided the MMSEQ package with effective lengths computed by BitSeq bias correction algorithm to produce results for this comparison.

### 3.2 Expression-level inference

Figure 3 demonstrates the ambiguity that may be present in the process of expression estimation. In Figure 3A and 3B, we show the density of samples from the posterior distribution of expression levels for two pairs of transcripts. The expression levels of transcripts uc010bho.1 and uc010bho.1 (Fig 3B) are negatively correlated. On the other hand, transcripts uc010bho.1 and uc010bw.3 exhibit no visible correlation (Fig 3B) 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 surrounding sequence (Dohm et al., 2008; Roberts et al., 2011) which incurs bias. This bias is included in the effective length normalization. To compare the results with qRT-PCR, the relative expression coefficient ($R^2$) of the average expression estimate and the reads 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.

<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>0.960</td>
<td>0.823</td>
<td>0.996</td>
<td>0.998</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.

### 3.3 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 distribution which is included in the $P(n|m)$ term that is calculated as a preprocessing step. The current BitSeq implementation contains the options 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. 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 of differential expression. When biological variance is being considered by inferring condition mean expression, the significance of DE is decreased. The naive approach reports high confidence approaches are portrayed in Figure 5d with obvious difference in the two conditions. Resulting distributions of differences for 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 used 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 of differential expression. 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 5 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 a and b.

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 5c 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 the condition mean expression, the significance of DE is decreased to PPLR 0.836.

### 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 8 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 8), 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. 8a and 8b). The overall performance here is fairly low because of high level of biological variance. For highest expressed transcripts (Fig. 8c), 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 NCBI37/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 R statistic. The number of iterations necessary to produce the desired amount of effective samples seems to increase logarithmically with the number of reads.
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 in Section 3.4 took 97 min on the Intel Xeon 3.47 GHz CPU.

Table 3. Scalability and run-time complexity of BitSeq.

| Read pairs (M) | 4.9 | 9.1 | 19.8 | 39.6 | 79.2 |
| Preprocessing (m) | 8 | 15 | 29 | 57 | 115 |
| Total time (h) | 0.55 | 2.18 | 5.42 | 16.23 | 33.19 |
| Convergence it. | 5269 | 6900 | 8920 | 11970 | 15979 |

The table shows wall clock running times to preprocess the aligned reads, generate 1000 equally sized groups based on the mean generative read count [1,3), [3,19] and [19, ∞), respectively.

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/B000476/2 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


