BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 17 2013, pages 2153–2161
doi:10.1093/bioinformatics/btt363

Gene expression

Advance Access publication June 21, 2013

Differential gene expression analysis using coexpression and RNA-Seq data

Ei-Wen Yang[1,*], Thomas Girke[2,3] and Tao Jiang[1,3,*]

[1]Department of Computer Science and Engineering, [2]Department of Botany and Plant Sciences and [3]Institute of Integative Genome Biology, University of California, Riverside, CA 92521, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: RNA-Seq is increasingly being used for differential gene expression analysis, which was dominated by the microarray technology in the past decade. However, inferring differential gene expression based on the observed difference of RNA-Seq read counts has unique challenges that were not present in microarray-based analysis. The differential expression estimation may be biased against low read count values such that the differential expression of genes with high read counts is more easily detected. The estimation bias may further propagate in downstream analyses at the systems biology level if it is not corrected.

Results: To obtain a better inference of differential gene expression, we propose a new efficient algorithm based on a *Markov random field* (MRF) model, called MRFSeq, that uses additional gene coexpression data to enhance the prediction power. Our main technical contribution is the careful selection of the clique potential functions in the MRF so its *maximum a posteriori* estimation can be reduced to the well-known maximum flow problem and thus solved in polynomial time. Our extensive experiments on simulated and real RNA-Seq datasets demonstrate that MRFSeq is more accurate and less biased against genes with low read counts than the existing methods based on RNA-Seq data alone. For example, on the well-studied MAQC dataset, MRFSeq improved the sensitivity from 11.6 to 38.8% for genes with low read counts.

Availability: MRFSeq is implemented in C and available at http://www.cs.ucr.edu/~yyang027/mrfseq.htm

Contact: yyang027@ucr.edu or jiang@cs.ucr.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 27, 2013; revised on June 6, 2013; accepted on June 10, 2013

1 INTRODUCTION

Next-generation sequencing technologies have been widely used in genomics research. RNA-Seq, one of the most exciting applications of next-generation sequencing technologies, is used to reveal the complexity of tanscriptomes in biological systems (Wang *et al.*, 2009). Many unprecedented discoveries are being made by RNA-Seq, such as the inference of novel isoforms, characterization of the modes of antisense regulation and study of intergenic expression patterns (Carninci *et al.*, 2005; Graveley *et al.*, 2011; Nagalakshmi *et al.*, 2008; Trapnell *et al.*, 2010). In recent years, RNA-Seq has taken a major role in the quantitative analysis of gene expression and transcript variant discovery. In the past decade, most of these applications were dominated by microarray-based technologies. In these quantitative assays, RNA populations are partially sequenced and the obtained read sequences are aligned back to the reference genome. The aligned reads are then assigned to genes based on the common regions that they share in the alignment. The number of reads assigned to a gene is called the *read count* of the gene, which has been shown to be nearly linearly correlated with the expression level of a gene (Marioni *et al.*, 2008).

Differential gene expression analysis is to identify if genes express differently between biological conditions of interest. Given RNA-Seq read count data, detecting differentially expressed (DE) or equally expressed (EE) genes can be done by checking if the observed difference of the read counts is significant or not, i.e. greater than some natural random variation. To test the significance of the difference between RNA-Seq read counts, the distribution of read counts was first assumed to be Poisson in (Marioni *et al.*, 2008; Srivastava and Chen, 2010; Wang *et al.*, 2010). However, the Poisson distribution may underestimate the variance of read counts and cause unexpected false positives in differential gene expression analysis (Nagalakshmi *et al.*, 2008; Robinson and Smyth, 2007). To solve the problem, negative binomial distributions were applied to RNA-Seq read data (Anders and Huber, 2010; Robinson and Smyth, 2007, 2008; Robinson *et al.*, 2010) and have become the state-of-the-art statistical model. Other than the methods based on the Poisson or negative binomial distributions, two data-driven probabilistic methods, baySeq (Hardcastle and Kelly, 2010) and NOISeq (Tarazona *et al.*, 2011), have also been proposed. Moreover, given annotated or inferred mRNA transcripts (or isoforms) of genes, some statistical methods for detecting differential expression at the transcript level have been published recently (Griffith *et al.*, 2010; Li and Dewey, 2011; Trapnell *et al.*, 2013; Zheng and Chen, 2009). Because the expression level of a gene with known (or inferred) isoforms can be calculated by simply summing up the expression levels of its isoforms, these transcript-level methods can be used as alternative methods for detecting differential expression of isoforms (Trapnell *et al.*, 2013), although the accuracy of these methods clearly depends on the quality of the provided isoforms.

Although the statistical properties of RNA-Seq data have been well studied and taken into account in the above statistical methods, these methods suffer from the following issues. First, it has been observed that statistical power increases with read count

*To whom correspondence should be addressed.

values (Oshlack and Wakefield, 2009; Oshlack *et al*., 2010; Young *et al*., 2010). Note that the read count of a gene is proportional to the gene expression level multiplied by the gene length. As a result, long or highly expressed genes are more likely to be detected as DE genes compared with their short and/or lowly expressed counterparts. This bias in DE gene detection is unavoidable even when normalization or rescaling is applied to read count data (Oshlack and Wakefield, 2009; Young *et al*., 2010). It is known that the selection bias on DE genes, if uncorrected, may lead to biased downstream analyses (Oshlack and Wakefield, 2009; Oshlack *et al*., 2010; Young *et al*., 2010). Second, the dependency among the expression of genes is not used in these methods. In gene expression analysis based on microarray data, the prior knowledge of gene coexpression patterns has been used to improve the performance of algorithms for detecting phenotype-related pathways (Rahnenfuhrer *et al*., 2004), searching for significant pathway regulators (Sivachenko *et al*., 2005), identifying differential gene expression patterns (Jacob *et al*., 2012) and the classification of microarray data (Rapaport *et al*., 2007). In particular, to obtain more accurate inference of DE genes, Wei and Li (Wei and Li, 2007) proposed a *Markov random field* (MRF) model that integrates the gamma-gamma model based on microarray data (Kendziorski *et al*., 2003; Newton *et al*., 2001) and gene coexpression networks extracted from KEGG pathways (Kanehisa and Goto, 2000) such that DE genes can be determined by the *maximum a posteriori* (MAP) estimation of the MRF model. Their experimental results demonstrate that the additional gene coexpression information can help detect more subtle changes of gene expression (e.g. local disturbances within known pathways) and significantly improve the overall prediction accuracy of DE genes (Wei and Li, 2007). However, due to the difference between continuous microarray intensity values and discrete RNA-Seq read counts, The MRF model in (Wei and Li, 2007) cannot be applied to RNA-Seq data immediately. Moreover, because the MAP estimation problem for an MRF model is generally NP-Hard (Boykov, 2001), the MRF model in (Wei and Li, 2007) was solved by a heuristic method, *iterated conditional modes*, which provides an approximately optimal prediction with no confidence scores.

In this work, we propose a novel MRF model, MRFSeq, combining RNA-Seq read counts with the prior knowledge of gene coexpression networks to infer DE genes. Different from the MRF model in (Wei and Li, 2007), we choose the clique potential functions of the MRF model carefully so that the MAP estimation of DE genes can be reduced to the well-known maximum flow problem on flow networks based on the work of Kolmogorov and Zabih (Kolmogorov and Zabih, 2004). Because the maximum flow problem is polynomial-time solvable, our MRF model can be solved exactly in polynomial time. Moreover, we introduce a *loopy belief propagation* method (Mooij, 2007; Weiss, 2000) to calculate the confidence of each inferred DE or EE gene. Our extensive experiments on simulated and real RNA-Seq data demonstrate that MRFSeq achieves a much improved overall estimation performance by gaining considerable sensitivity without losing precision. A detailed analysis of the prediction results indicates that the DE genes predicted by MRFSeq are distributed more evenly across different values of read counts than those recovered by the existing methods using RNA-Seq data alone. Hence, MRFSeq can help alleviate the selection bias of DE genes against genes with low read counts. Our analysis further shows that most of the DE or EE genes that can be correctly predicted from RNA-Seq data alone are also correctly predicted by MRFSeq, implying that the use of the prior knowledge of gene coexpression does not introduce new biases in the differential analysis result. Moreover, we compare MRFSeq with a recently published transcript-level method, Cuffdiff 2 (Trapnell *et al*., 2013), on the real RNA-Seq data using the annotated transcriptome from UCSC hg19 (Meyer *et al*., 2013). The comparison shows that MRFSeq is much more sensitive than Cuffdiff 2.

The rest of this article is organized as follows. Section 2.1 defines the terms and notations used in our algorithms, while Section 2.2 provides the formulation of the MRF model and the design of its clique potential functions. The reduction to the maximum flow problem is shown in Section 2.3. The experimental results are described in Section 3, which also contains a comparison between MRFSeq and existing differential expression analysis methods including edgeR (Robinson *et al*., 2010), DESeq (Anders and Huber, 2010), baySeq (Hardcastle and Kelly, 2010), NOISeq (Tarazona *et al*., 2011) and Cuffdiff 2 (Trapnell *et al*., 2013). In particular, Section 3.4 compares the performance of the methods on genes with low read counts and shows that MRFSeq achieves not only an overall significantly higher accuracy but also provides a less biased prediction. A few concluding remarks are given in Section 4. The loopy belief propagation method for calculating the confidence level of each prediction as well as some figures and tables are omitted in the main text owing to page limit and are provided in the Supplementary Materials.

2 MATERIALS AND METHODS

2.1 Terminology and notations

Let $G = \{g_1, g_2, ..., g_n\}$ be the genes to be tested for differential expression and $X = \{x_1, x_2, ..., x_n\}$ the binary random variables such that each $x_i \in \{0, 1\}$ indicates the DE state of gene g_i. The random variable $x_i = 1$ if the gene g_i is a DE gene and $x_i = 0$ indicates that the gene is an EE gene. Two random variables x_i and x_j are assumed to be correlated if the two genes g_i and g_j form a pair of coexpressed genes. A configuration x is a 0–1 assignment to the random variables X. Assume that there are p and q replicates in the two conditions, A and B, of interest, respectively. Let the read counts a_j^i and b_j^i be the number of the reads aligned to gene g_i in the j-th replicate of the conditions A and B, respectively. For each gene g_i, two sets of the read counts $R_{A,i} = \{a_1^i, a_2^i, ..., a_p^i\}$ and $R_{B,i} = \{b_1^i, b_2^i, ..., b_q^i\}$ are summarized from all the replicates of the two conditions A and B after mapping all the reads to the reference genome. Popular statistical measurements for the observed difference of read counts are the false discovery rates [FDRs, i.e. the p-value corrected for multiple testing (Benjamini and Hochberg, 1995)] and prior probability. The current statistical methods infer DE genes by checking independently for each gene if the difference measurement of its read count exceeds a certain threshold (Oshlack *et al*., 2010). In our method, DE genes are determined by the configuration that maximize a likelihood function of both observed difference of read counts and gene coexpression while no prior knowledge of the thresholds is required. MRFSeq uses, but is not limited to, the FDR q_i from DESeq (Anders and Huber, 2010) as the difference measurement of the read counts $R_{A,i}$ and $R_{B,i}$, where $q_i \in [0, 1]$. To improve the computational efficiency of our algorithm, the FDR q_i is further discretized by binning the interval [0,1] into 20

intervals of the same length 0.05. Let $y_i \in \{1, 2, \ldots, 20\}$ denote the interval where the observed difference q_i belongs to, and $Y = \{y_1, y_2, ..., y_n\}$ be, the collection of all the discretized FDRs. The joint probability of the hidden variables X given its observed values Y is then formulated by an MRF model, a graphical model capable of capturing the statistical dependency of random variables (Kindermann and Snell, 1980), described in the next subsection. Given the joint probability of X conditional to Y, estimating the DE states of the genes actually involves two inference problems. The first is the MAP estimation problem, i.e. searching for a configuration x^* such that $Pr(x^*|Y)$ is maximized. The algorithm for the MAP estimation problem will be discussed later in the section. The second is the *marginal probability* problem, i.e. computing the probability $Pr(x_i|Y)$ as a confidence level of the configuration on each gene g_i. The loopy belief propagation method for the *marginal probability* problem is given in the Supplementary Materials.

2.2 MRF model

Let $H = (V_X, E)$ be an undirected graph representing the coexpression network for G such that every node $v_{x_i} \in V_X$ is associated with the random variables $x_i \in X$ and every edge (i,j) shared by the nodes v_{x_i} and v_{x_j} encodes the dependency of the two correlated random variables x_i and x_j. Two variables x_i and x_j are assumed to be correlated if the two genes g_i and g_j are coexpressed. To determine which pair of the genes are the coexpressed genes, the correlation coefficient $c_{i,j}$ defined in COXPREdb (Obayashi and Kinoshita, 2011) is used as the measurement of gene coexpression between the two genes g_i and g_j. Two genes are considered as a pair of coexpressed genes if $c_{i,j}$ is greater than a threshold ρ. We use $\rho = 0.5$ throughout this work because it is widely used in the literature (Patil *et al.*, 2011; Watson, 2006).

In our model, we think that the DE state of each gene should depend on its observed difference in read counts and the DE states of its coexpressed genes. In other words, we can assume that every random variable is conditionally independent to the variables indexed by non-adjacent vertices in H. Hence, the following property is satisfied:

$$Pr(x_i|X) = Pr(x_i|x_j, v_{x_j} \in N(v_{x_j})), \tag{1}$$

where $N(v_{x_j})$ represents the neighbors of v_{x_i} in H. By the Hammersley–Clifford theorem (Besag, 1974), a joint distribution of the random variables X given Y can be factorized as a form of clique potential functions $T_C(C)$, the positive functions for configurations over cliques in the given graph H such that $Pr(X|Y) = \prod_{C \in H} T_C(C)$.

To model the pairwise dependency between coexpressed genes, we may use an MRF model consisting of only potential functions for cliques of sizes at most 2. This type of MRF is called the *pairwise* MRF (Besag, 1986) and will be used in our work. There are two types of potential functions adopted in our MRF model. One is the unary functions $\phi_i(x_i)$ that score how compatible the random variable x_i is with its observed evidence y_i. The other is the pairwise potential functions $\psi_{(i,j)}(x_i, x_j)$ that measure the statistical dependency between every pair of correlated variables x_i and x_j. By the definition of the potential functions, the joint distribution of X given Y can be written as

$$Pr(X|Y) = \frac{1}{Z} \prod_{(i,j) \in E} \psi_{(i,j)}(x_i, x_j) \prod_{i=1}^{n} \phi_i(x_i), \tag{2}$$

where Z is the normalized term to assure that the joint probability $Pr(X|Y)$ sums up to 1. Let $P_{(1,i)} = Pr(x_i = 1|y_i)$ and $P_{(0,i)} = Pr(x_i = 0|y_i)$. The unary function $\phi_i(x_i)$ is defined as follows:

$$\phi_i(x_i) = \begin{cases} P_{(1,i)}/P_{(0,i)}, & \text{if } P_{(1,i)} > P_{(0,i)}, x_i = 1 \\ P_{(0,i)}/P_{(1,i)}, & \text{if } P_{(0,i)} > P_{(1,i)}, x_i = 0 \\ 1, & \text{otherwise,} \end{cases} \tag{3}$$

To calculate the unary functions, the ratio between the two prior probabilities $Pr(x_i = 1|y_i)$ and $Pr(x_i = 0|y_i)$ should be given as a known parameter in our MRF model. To estimate the parameter, the read counts of four replicates (two per condition) for 10 000 DE genes and 10 000 EE genes are first synthesized. Our simulation of the read counts of the DE and EE genes follows the same steps as used in the simulation study of DESeq (Anders and Huber, 2010). For the DE genes, the $\log_2$ fold change rate of the observed read counts between two conditions is randomly drawn from the normal distribution with mean 0 and variance 0.7. For the EE genes, the mean is set to be 0 and the variance 0.2. After the simulation of read counts, the discretized FDRs introduced previously are calculated as the observed difference in the synthesized read counts. Assume that there are m_{y_i} DE genes and n_{y_i} EE genes whose discretized FDR is y_i in this simulation. We further assume the equality of the two background probabilities of x_i holds, i.e. $Pr(x_i = 1) = Pr(x_i = 0)$. By Baye's rule, the ratio of the prior probabilities is obtained as follows:

$$\frac{Pr(x_i = 0|y_i)}{Pr(x_i = 1|y_i)} = \frac{Pr(y_i|x_i = 0)Pr(x_i = 0)}{Pr(y_i|x_i = 1)Pr(x_i = 1)} = \frac{n_{y_i}}{m_{y_i}}, \tag{4}$$

Symmetrically, we have $Pr(x_i = 1|y_i)/Pr(x_i = 0|y_i) = m_{y_i}/n_{y_i}$.

For the pairwise function $\psi_{(i,j)}(x_i, x_j)$ of every pair of coexpressed genes g_i and g_j, the correlation coefficient $c_{i,j}$ defined in COXPREdb (Obayashi and Kinoshita, 2011) is used as the measure of the statistical dependency between x_i and x_j. The pairwise potential functions are thus defined as follows:

$$\psi_{(i,j)}(x_i, x_j) = \begin{cases} e^{c_{i,j}}, & \text{if } x_i = x_j, \\ 1, & \text{otherwise,} \end{cases} \tag{5}$$

This completes the specification of the joint distribution of X. To facilitate the presentation of our algorithms, the joint distribution of X can be rewritten by taking negative logarithm on both sides of Equation (2) as below:

$$E(X|Y) = -\gamma - \sum_{i=1}^{n} \alpha_i(x_i) - \sum_{(i,j) \in E} \beta_{(i,j)}(x_i, x_j), \tag{6}$$

where γ is a constant, $\alpha_i(x_i) = ln\phi_i(x_i)$ and $\beta_{(i,j)}(x_i, x_j) = ln\psi_{(i,j)}(x_i, x_j)$. $E(X|Y)$ is called the pseudo-energy function when each α_i is a unary term and each $\beta_{(i,j)}$ is a pairwise term of the energy. A configuration maximizing the joint probability $Pr(X|Y)$ is actually the configuration minimizing the pseudo-energy function $E(X|Y)$ (Besag, 1986).

2.3 MAP Estimation

Different from the heuristic method, *iterated conditional modes*, used to approximate the MAP of the MRF model of Wei and Li (2007), we show in this subsection that, by designing the potential functions in MRFSeq carefully, the MAP estimation problem for MRFSeq is no longer an NP-Hard problem because it can be reduced to the maximum flow problem on flow networks and solved optimally in polynomial time.

A random variable x_i is said to be *inverted* by a configuration x if the state assignment to x_i violates its prior probability, i.e. $x_i = 1$ if $Pr(x_i = 0|y_i) > Pr(x_i = 1|y_i)$ or $x_i = 0$ if $Pr(x_i = 1|y_i) > Pr(x_i = 0|y_i)$. For an inverted random variable x_i, $\alpha_i(x_i) = 0$ instead of $|ln\phi_i(1) - ln\phi_i(0)|$. We define $|ln\phi_i(1) - ln\phi_i(0)|$ as the cost of the inversion. Two correlated variables x_i and x_j are said to be *separated* by a configuration x if the assigned states of x_i and x_j are different, *i.e.* $x_i \neq x_j$. For a pair of separated variables x_i and x_j, $\beta_{(i,j)}(x_i, x_j) = 0$ instead of $c_{i,j}$. The cost of the separation is $c_{i,j}$. Kolmogorov and Zabih (2004) proved that when the pairwise term $\beta_{(i,j)}(x_i, x_j)$ of the pseudo-energy function $E(X|Y)$ is *submodular*, that is, the following property is satisfied:

$$\beta_{(i,j)}(0,0) + \beta_{(i,j)}(1,1) \geq \beta_{(i,j)}(0,1) + \beta_{(i,j)}(1,0), \tag{7}$$

searching for a configuration that minimizes the pseudo-energy function can be done by looking for a configuration minimizing the total cost of inversion and separation. That is, the MAP estimation problem on an

MRF model can be reduced to the maximum flow (or minimum cut) problem over a flow newtwork H' such that a minimum cut of H' corresponds to a MAP estimation of the MRF model and the saturated capacity of the cut is exactly the total cost of the inversion and separation.

It is easy to verify that our paiwise term is submodular. $\beta_{(i,j)}(0,0)+\beta_{(i,j)}(1,1)$ sums up to $2c_{i,j}$, where $c_{i,j}\geq 0.5$, while $\beta_{(i,j)}(0,1)+\beta_{(i,j)}(1,0)$ is 0. The reduction from our MRF model whose graph representation is $H=(V_x,E)$ to the flow network $H'=(V_x\cup\{s,t\},E')$ can be done as follows. The nodes of H' are the union of the nodes of H and two additional nodes, the source s and sink t. Every undirected edge (i,j) of H is transformed into two directed edges (i,j) and (i,j) with capacity $c_{i,j}$. For every node x_i, two directed edges (s,i) and (i,t) are added to E'. The capacity of the edge (s,i) is $|ln\phi_i(1)-ln\phi_i(0)|$ if $Pr(x_i=1|y_i)>Pr(x_i=0|y_i)$. Otherwise, the capacity of the edge (s,i) is 0. Symmetrically, the capacity of the edge (i,t) is $|ln\phi_i(1)-ln\phi_i(0)|$ if $Pr(x_i=0|y_i)>Pr(x_i=1|y_i)$. Otherwise, the capacity of the edge (i,t) is 0. After running a standard maximum flow algorithm, e.g. the Edmond and Karp algorithm (Edmonds and Karp, 1972), on the flow network H', a minimum cut $Q=\{V_s\cup s, V_t\cup\{t\}\}$ is obtained, where V_s are the nodes adjacent to s and V_t the nodes adjacent to t. It represents a 0–1 assignment such that all the random variables corresponding to the nodes of V_s are assigned 1 and all the random variables corresponding to the nodes of V_t are assign 0. Then, a gene g_i is inferred as a DE gene if x_i is 1, or an EE gene otherwise.

2.4 RNA-Seq Datasets

Two publicly available human RNA-Seq datasets, the MAQC dataset (Bullard *et al.*, 2010; Shi *et al.*, 2006) and Griffth's dataset (Griffith *et al.*, 2010), will be used as the benchmark datasets to assess the performance of our selected differential gene expression analysis methods. Each of the dataset is associated with an additional qRT-PCR dataset to validate the DE states of genes. The MAQC dataset consists of two samples, Ambion's human brain reference RNA (brain) and Stratagene's human universal reference RNA. Each sample provides seven replicates and a total of 45 million single-end RNA-Seq reads of length 35 bp. The read counts for the MAQC dataset is obtained from 71 million uniquely mapped reads calibrated by ReCounts (Frazee *et al.*, 2011). Griffith's dataset was made from the qRT-PCR validation for the DE or alternatively expressed genes highlighted by ALEXA-Seq (Griffith *et al.*, 2010). It contains 96 and 198 million pair-end reads across two human colorectal cancer cell lines that only differ in fluorouracil resistance phenotypes. To equilibrate sequencing depth in both samples, as done in (Tarazona *et al.*, 2011), the read library size is set to be about 100 million reads per condition. Raw RNA-Seq reads of the MAQC dataset were downloaded from the SRA database (Leinonen *et al.*, 2011), while the RNA-Seq reads of Griffith's dataset were downloaded from the FTP site of the ALEXA-Seq Web site. The gene association across platforms was performed with BioMart (Zhang *et al.*, 2011). Unmatched genes were discarded in downstream analysis steps. To obtain the read counts for Griffith's dataset, the raw RNA-Seq reads were aligned against the high-coverage assembly of the human genome UCSC hg19 (Meyer *et al.*, 2013) using Tophat (Trapnell *et al.*, 2009) where two mismatches were allowed and reads mapped to multiple locations were removed. Finally, the read counts for each gene in Griffith's dataset were summarized by using the R packages GenomicFeatures and RSamtools from Bioconductor along with the genome annotation information from Ensembl (version 60) (Flicek *et al.*, 2011) and only exonic reads. For a fair comparison, a pseudo read count, 1, was applied to all genes with zero read counts to avoid the divided-by-zero problem in some statistical calculations.

2.5 Evaluation Metrics

Following the assessment method of Bullard *et al.* (2010), all our experimental results are evaluated in terms of precision (PRE), $PRE=TP/(TP+FP)\times 100\%$, and sensitivity (SEN), $SEN=TP/(TP+FN)\times 100\%$, where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. To combine the two evaluation measures, the F-score (FS) (van Rijsbergen, 1979), defined as $FS=[2\times(PRE\times SEN)/(PRE+SEN)]\times 100\%$, is used as a measure of the overall performance of a prediction method in our tests.

3 EXPERIMENTAL RESULTS

3.1 Selection of differential gene expression analysis methods

To compare our method with the existing gene differential analysis methods, the same selection criteria proposed by Tarazona *et al.* (2011) was followed. However, Fisher's exact test (Fisher, 1922), which was compared in (Tarazona *et al.*, 2011), was excluded here because its performance was shown to be far lower than those of the other methods. At the end, four methods including edgeR (Robinson *et al.*, 2010), DESeq (Anders and Huber, 2010), baySeq (Hardcastle and Kelly, 2010) and NOISeq (Tarazona *et al.*, 2011) were selected to be compared in our tests. Note that NOISeq has two versions, NOISeq_real and NOISeq_sim, and the version NOISeq_real is used in our experiments because numbers of replicates in our simulated and real datasets are always greater than one. Some reasonable cutoff values are required in these methods (except MRFSeq) to decide the significance of a statistical difference measurement. To obtain comparable performance analysis scenarios, the cutoff values adopted in the literature are applied in our experiments. More specifically, the FDR 0.1 chosen in DESeq is used for DESeq and edgeR. We choose the probability 0.8 and 0.999, as done in the work of Tarazona *et al.* (2011), for NOISeq and baySeq, respectively. Experiments at two levels of difficulty are conducted to compare our method MRFSeq with the other selected methods. At the first level, all read counts of the benchmark datasets are generated from the same distribution as assumed in the simulation studies of DESeq. At the second level, all read counts of the genes are accumulated from the two real datasets, the MAQC and Griffith's datasets, and may contain low read counts. In addition to the comparisons with the gene-level methods, MRFSeq is also compared with the recently published transcript-level method Cuffdiff 2 on the two RNA-Seq datasets.

3.2 Simulation studies

3.2.1 Simulation experiments Our simulation experiments follow the framework in (Wei and Li, 2007). All gene sets associated with the 186 KEGG pathways in MSigDB (Subramanian *et al.*, 2005) were downloaded. The coexpression networks of the 186 gene sets were then defined using COXPREdb (Obayashi and Kinoshita, 2011) and they formed 186 undirected graphs. A gene set was discarded if the number of the edges in its coexpression network is less than the number of the nodes. After the filtration, 37 gene sets consisting of 2194 different genes were kept. The 37 coexpression networks (see Supplementary Table S1) were merged as a global network consisting of 2194 nodes and 8512 edges. All the methods are tested at five different abundance levels of true DE genes. The performance assessment is categorized into five classes, where each class represents an

abundance level interval of 10% such that the five classes cover abundance levels of DE genes ranging from 0 to 50% as done in (Wei and Li, 2007). At each of the five levels, we randomly choose 10 combinations of the pathways to form the sets of true DE genes, while keeping the rest of the genes as true EE genes, such that the percentage of the true DE genes is within the range of the level. The 10 different combinations form 10 benchmark datasets and read counts are randomly obtained by following the same steps for simulating read counts used in DESeq. The simulated read counts range from 25 to 401. All the methods are applied to the 50 benchmark datasets. Owing to the page limit, the performance assessment on the first interval [0,10) is summarized in Table 1 and the complete assessment on all five intervals is presented in Supplementary Table S2. For the convenience of the reader, the precision-sensitivity curves are also provided in Supplementary Figure S1.

3.2.2 Comparisons of the methods on simulated data MRFSeq has clearly the best F-scores (i.e. the overall performance) and significantly improved sensitivity over the other methods. Its F-score is 14.2, 6.8, 6.8, 3.7 and 4.8% greater than the second best in the five interval, while its improvement on sensitivity is 13.6, 10.6, 9.1, 1.7 and 2%, respectively. Although baySeq provides close sensitivity scores in the intervals [30,40) and [40,50), it fails to obtain comparable precision scores and hence has an inferior overall performance. While achieving a considerable improvement on sensitivity in the interval [0,10), MRFSeq improves the precision by at least 6.9%. In the other four intervals, MRFSeq's precision is slightly lower than those of the other methods. The difference between the precision of MRFSeq and the best precision in these intervals is 2.3, 1.4, 2.2 and 1%, respectively, which are actually smaller than the standard deviations. The standard deviations of the sensitivity and F-score of MRFSeq are greater than the standard deviations of the other methods. This is because the performance of MRFSeq is somewhat correlated to the topological distributions of the true DE genes on the coexpression network. The amount of improvement achieved by MRFSeq may vary depending on the topological distribution. Nevertheless, the simulation results demonstrate that coexpression data could help improve differential gene expression analysis by increasing the coverage of true DE genes significantly.

Table 1. Comparison of different methods on simulated datasets

Levels[a]	Avg (%)[b]	Methods[c]	PRE (%)	SEN (%)	FS (%)
[0,10)	5.7	MRFSeq	75.55 (11.0)	71.99 (12.8)	73.36(10.3)
		baySeq	66.23 (10.2)	53.49 (4.3)	59.02 (6.4)
		DESeq	68.57 (10.3)	47.78 (4.7)	55.87 (4.3)
		edgeR	63.07 (12.8)	57.07 (2.9)	59.11 (4.8)
		NOISeq	50.04 (17.3)	58.32 (3.0)	52.29 (9.3)

[a]The range of the abundance levels of DE genes.
[b]The average percentage of DE genes among the 10 test datasets at the level.
[c]The names of the methods.

3.3 Performance on real RNA-Seq data

3.3.1 Experiments on the MAQC dataset In addition to the previous simulation study, the performance for inferring DE genes is assessed on the MAQC dataset. Previously, Tarazona *et al.* (2011) tested the selected methods on different numbers of replicates (or lanes), from 2 replicates to 7 replicates per condition, in the MAQC dataset to see how sequencing depth would affect the performance of the methods. The results indicated that increasing the sequencing depth would decrease the precision of all selected methods except NOISeq. To compare the performance and understand how the precision of MRFSeq would change as the sequencing depth increases, the experiments designed by Tarazona *et al.* are used in our work. Different numbers of replicates are considered such that the read library size varies from 14 to 45 million reads in each of the two samples. The expression levels of the genes in the MAQC dataset were measured by the normalized threshold cycle values (CT) of qRT-PCR. To validate the true DE genes of the MAQC dataset, a gene is defined as a true DE gene if the $\log_2$ fold change ratio (LR) of its CT values is greater than a certain threshold b, e.g. 0.5 or 2, while a gene is a true EE gene if its LR is smaller than threshold a, e.g. 0.2 (Bullard *et al.*, 2010). Any gene whose LR is between the two thresholds a and b is considered as a borderline gene. In the previous studies, all borderline genes were discarded (Bullard *et al.*, 2010; Tarazona *et al.*, 2011). Owing to the detection limitation of qRT-PCR, lowly expressed genes may be absent in some of the qRT-PCR assays. A gene that was detected in at least one of the qRT-PCR assays would also be removed if it failed to appear in at least three-fourth of the qRT-PCR assays (Bullard *et al.*, 2010). Different from the previous studies, we do not throw away those borderline genes. To further test the inference power on genes with low read counts, lowly expressed genes are also kept in our experiments. This gives us a total of 836 genes. We define a gene as a true DE gene if its LR is larger than the threshold b. Otherwise, the gene is a true EE gene. There are 669 true DE genes when the threshold b is set to be 0.5 and 373 true DE genes when b is 2.0. The coexpression network of the 836 genes forms a graph of 836 nodes and 2426 edges. All the methods are tested at these two different abundance levels (or LR values) of DE genes. The prediction results are again assessed in terms of precision, sensitivity and F-score as summarized in Figure 1.

3.3.2 Comparison of the performance on the MAQC dataset Similar to the results in the simulation study, MRFSeq achieves significantly improved sensitivity scores and F-scores at both abundance levels of true DE genes. The improvement on sensitivity is at least 9.2 and 8.8% for all sequencing depths considered when $b = 0.5$ and 2, respectively. While achieving the best sensitivity scores, the precision scores of MRFSeq are also comparable with the precision of the others except NOISeq who exhibits extremely high precision. Note that although NOISeq has the best precision among all methods, its sensitivity is much lower than the scores of the others and its overall performance (as measured by F-score) suffers from this. As the sequence depth increases, the precision of NOISeq remains stable while all other methods lose some precision. The precision of DESeq drops 4.0 and 4.7%, respectively, for the two

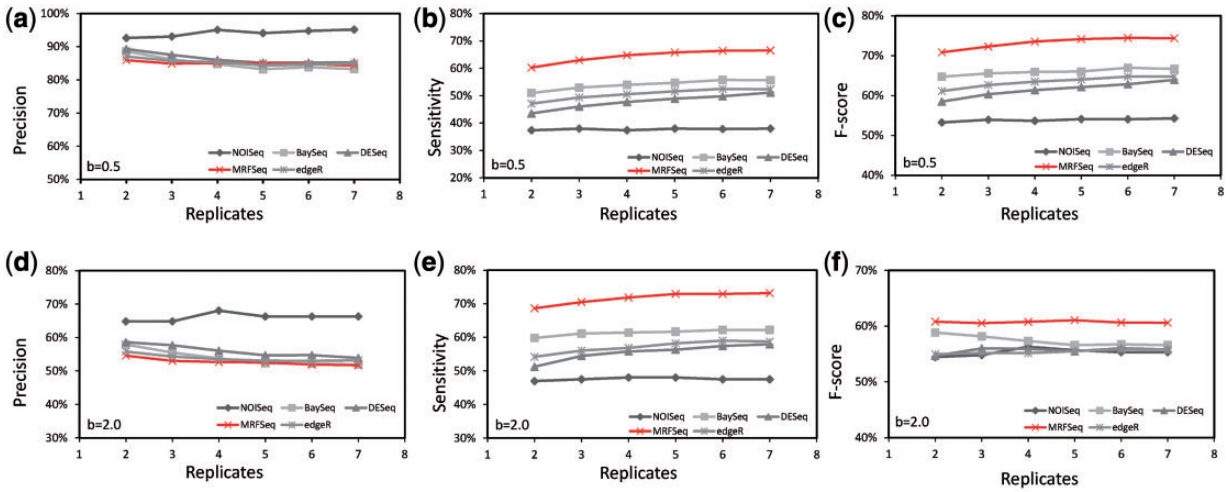

Fig. 1. Performance assessment at various sequencing depths. The *X*-axis shows the number of used lanes and the *Y*-axis indicates various assessment measures. In the upper row of plots, the LR threshold *b* is set as 0.5 and in the lower row $b = 2.0$. Plots (**a**) and (**d**) compare the precision scores at different sequence depths. Plots (**b**) and (**e**) depict the sensitivity scores, while plots (**c**) and (**f**) illustrate the F-scores

values of b, when the number of replicates increases from two to seven. The decrease in precision is 5.4 and 6% for baySeq, while edgeR loses 2.0 and 2.6%. At the same time, the precision of MRFSeq only decreases 1.6 and 2.8%. The relative small loss of the precision for MRFSeq can be explained by the fact that many false positives, if not predicted at a strong confidence level, could be eliminated by MRFSeq using the coexpression information. Hence, these results on the MAQC dataset show that coexpression information not only helps gaining more coverage of the true DE genes but also keeps precision relatively stable against the increase of sequencing depth. Moreover, it could help reduce our reliance on deeply covered RNA-Seq data in differential gene expression analysis.

3.3.3 Taking confidence scores into consideration Like the FDRs of DESeq and edgeR or the prior probability of baySeq and NOISeq, MRFSeq estimates the confidence (i.e. marginal probability) for each predicted DE gene and a confidence threshold can be applied to select DE genes for the output (instead of following the MAP estimation algorithm). We are interested in the performance of MRFSeq on the MAQC dataset when different thresholds are applied to the confidence. To calculate the confidence scores, the loopy belief propagation algorithm is run on all seven replicates in the MAQC dataset. To compare the performance of MRFSeq with the other methods, a precision-sensitivity curve where each point represents the precision and sensitivity under a certain threshold, is depicted for each of the selected methods, as done in (Tarazona *et al.*, 2011). To depict the precision-sensitivity curves for DESeq and edgeR the range of the FDR threshold from 10^{-6} to 1 is selected. Note that this range for FDR cutoffs covers all the threshold values used in practice and these FDR thresholds yield sensitivity values between 45 and 100%. For the other methods that do not use FDRs, equivalent thresholds that lead to sensitivity within the same range, i.e. 45–100%, are applied to draw the precision-sensitivity curves. The precision-sensitivity curves in Supplementary Figure S2 show that, in general, MRFSeq provides more accurate confidence scores than the other methods. Note that unlike the MAP estimation algorithm, using the marginal probabilities obtained by the loopy belief propagation algorithm to infer DE genes requires additional knowledge to choose an approproate confidence (marginal probability) threshold. Besides, the loopy belief propagation algorithm is a heuristic and thus does not guarantee correct marginal probabilities. Hence, MRFSeq uses the MAP estimation to select DE genes and the loopy belief propagation algorithm only to estimate the confidence score of each selected DE gene.

3.3.4 Comparisons of the performance on Griffith's dataset The qRT-PCR data of Griffith's dataset consists of 193 exon assays on 94 protein coding genes. Different from the LR of the MAQC dataset, a two-tailed *t*-test was applied to identify the true DE genes from the qRT-PCR data of Griffith's dataset. A *P*-value of the *t*-test was considered significant if it is smaller than 0.05 (Griffith *et al.*, 2010). Under this criterion, 83 true DE genes and 11 true EE genes are identified and used in testing the selected methods. The coexpression network of the 94 genes extracted from COXPREdb forms a graph of 94 nodes and 25 edges. The performance of the methods on Griffith's dataset is shown in Supplementary Table S3. MRFSeq still has the best overall performance, although its improvement over the other methods is not as significant as on the MAQC data. Please see the Supplementary Materials for a detailed discussion.

3.4 Performance on genes with low read counts

3.4.1 Genes with low read counts To understand how the methods perform on genes with different read count levels, the prediction on the real datasets is further analyzed. The genes in the datasets are separated into two classes, genes with low read counts and genes with decent read counts. In (Bullard *et al.*, 2010), a gene is said to have a low read count if it has fewer than 10 reads in every replicate of the two conditions. Otherwise, the gene is said to have a decent read count. Because Griffith's

dataset contains only genes with decent read counts, we consider the MAQC dataset only below. Among the 836 genes in the MAQC dataset, there are 453 genes with low read counts and 383 genes with decent read counts. The methods baySeq and NOISeq provide an additional option for normalizing gene lengths. These two methods with normalized gene lengths are denoted as baySeq_{len} and NOISeq_{len}, respectively. To further study the effect of the normalization on genes with low read counts, baySeq_{len} and NOISeq_{len} are also applied to the MAQC dataset. By choosing a threshold of $b = 0.5$ for the LR values, the prediction results on genes with low read counts by different methods are compared in Table 2. The detailed numbers of true and predicted DE genes with low or decent read counts are listed in Supplementary Table S4.

3.4.2 Significant improvement on genes with low read counts On the genes with low read counts, the sensitivity of MRFSeq is 38.8%, while the second best sensitivity is only 11.6%. Similarly, MRFSeq achieves an F-score of 53.3%, while the second best F-score is only 20.4%. In addition to these significant improvements, the prediction of MRFSeq shows a more balanced pattern between genes with low read counts and genes with decent read counts. The $\text{RTP}_{l/h}$ of MRFSeq is 43.1% while its $\text{RPP}_{l/h}$ is 42.7%. The second best $\text{RTP}_{l/h}$ and $\text{RPP}_{l/h}$ are only 13.0 and 13.3% (obtained by edgeR). This result shows that all the other methods are biased against genes with low read counts. Most of their predicted DE genes are from the genes with decent read counts. After applying the normalization of gene lengths on genes with low read counts, the performance of baySeq and NOISeq is slightly improved. However, the length normalization does not really improve the overall performance on genes with low read counts much or correct the selection bias.

3.5 Comparison with Cuffdiff 2

Different from gene-level methods that use raw read counts, Cuffdiff 2 requires the mapping of reads to the given transcripts of genes as input to call differential gene expression (Trapnell *et al.*, 2013). To assess the performance of Cuffdiff 2 on the MAQC and Griffith's datasets, the RNA-Seq reads of the two real datasets are mapped to the annotated transcriptome UCSC hg19 using Tophat as done in (Trapnell *et al.*, 2013). The same threshold 0.1 for the FDR values is used to call DE genes for Cufflink 2. The prediction accuracies of MRFSeq and Cuffdiff 2 on the two datasets are summarized in Supplementary Table S5, with the LR threshold $b = 2$ and the cutoff P-value 0.05 for the MAQC and Griffith's datasets, respectively. The precision-sensitivity curves also are illustrated in Supplementary Figure S3. The table shows that MRFSeq has a significantly better F-score (and thus overall performance) by achieving a higher sensitivity, while Cuffdiff 2 achieves a better precision. The precision-sensitivity curve also suggests that MRFSeq has a better overall performance than Cuffdiff 2 when we consider the full spectrum of FDR or restricting the FDR value to at most 0.1. A detailed analysis shows that Cuffdiff 2 predicts fewer true DE genes with relatively small LR values than MRFSeq. In the MAQC dataset, there are 290 true DE genes with the LR values from 0.5 to 2. The prediction of MRFSeq covers 171 of the 290 genes, while Cuffdiff 2 can only detect 140 of the true DE genes. In Griffith's dataset, 9 true DE genes are associated with P-values, which measure the significance of the difference between the LR values, from 0.005 to 0.001. All of the 9 true DE genes are predicted by MRFSeq, but Cuffdiff 2 could only identify 4 of the DE genes. This result is consistent with the discussion in (Trapnell *et al.*, 2013). In general, Cuffdiff 2 may report fewer DE genes with relatively low LR rates because of its control of variance in expression owing to fragment count uncertainty.

Table 2. Comparison of the prediction results on genes with low read counts

Methods	$\text{RTP}_{l/h}$ (%)[a]	$\text{RPP}_{l/h}$ (%)[b]	PRE (%)	SEN (%)	FS (%)
MRFSeq	43.1	42.7	84.8	38.8	53.3
baySeq	6.8	6.2	100.0	5.5	10.4
baySeq_{len}	7.4	6.8	100.0	6.1	11.5
DESeq	12.5	13.0	82.6	11.0	19.4
edger	13.0	13.3	83.3	11.6	20.4
NOISeq	0.0	0.0	—	0.0	—
NOISeq_{len}	4.5	5.0	84.6	3.2	6.1

[a]The ratio of true positives with low read counts over the true positives with high read counts.
[b]The ratio of predicted positives with low read counts over the predicted positives with high read counts.

3.6 Consistency of predictions by DESeq and MRFSeq

A gene is defined to be *incorrectly inverted* if its DE state is correctly predicted by using RNA-Seq data alone but incorrectly predicted by MRFSeq. Although our above results demonstrate that using the prior knowledge of gene coexpression significantly improves the overall accuracy of differential gene expression analysis and helps to alleviate the bias against genes with low read counts, it raises the question if the prior knowledge might introduce some new prediction biases. In this subsection, we estimate the number of incorrectly inverted genes in the prediction of MRFSeq compared with prediction by a popular RNA-Seq based method DESeq and analyze the types of genes in coexpression networks that are more likely to be incorrectly inverted. The detailed prediction results of DESeq and MRFSeq on our above simulation and real datasets are compared. In the 40 simulation benchmark datasets, only 3092 of the 73 619 (4.2%) correctly predicted genes by DESeq are incorrectly inverted by MRFSeq. In the MAQC and Griffith's datasets, only 16 (3.5%) and 0 (0%) genes correctly predicted by DESeq are incorrectly inverted by MRFSeq, respectively. Generally, most of the correctly predicted genes by DESeq remain correct in the MRFSeq prediction. Moreover, we observe that the incorrectly inverted genes tend to have higher edge degrees in gene coexpression networks than the other genes. The comparison of the average edge degree of all genes and that of the incorrectly inverted genes in gene coexpression networks is shown in Supplementary Figure S4. The significance of the difference between the edge degrees is confirmed by using one-tailed t-test (Goulden, 1956). The P-value of the t-tests on the simulation and MAQC datasets are 5.1×10^{-14} and 5.1×10^{-4}, respectively. However, as gene

coexpression networks usually possess the well-known scale-free property, only a small number of genes have high edge degrees (Carlson *et al*., 2006; Stuart *et al*., 2003). This property should limit the number of incorrectly inverted genes, and thus most of the DE or EE genes correctly predicted based on RNA-seq data alone (by, e.g. DESeq) are well preserved in the result of MRFSeq.

4 CONCLUSION AND DISCUSSION

In this work, we have proposed a new statistical method, MRFSeq, that combines both RNA-Seq data and coexpression information and obtains a MAP estimation of the DE/EE genes efficiently. The discussion of the improvement benefits from our graphical model is provided in Section 1.3 of the Supplemmentary Materials. Using extensive experiments on both simulated and real data, we have shown that MRFSeq is able to take advantage of coexpression information and this additional piece of information can help provide a more accurate and less biased differential gene expression analysis. Clearly, our improved performance (especially on genes with low read counts) critically depends on the existence of a high-quality gene coexpression network (See the discussion in Section 1.4 of Supplementary Materials.) Finally, MRFSeq uses the DE analysis results of DESeq. It would be interesting to study how MRFSeq performs when the DE analysis results of other tools are used instead. Section 1.5 of Supplementary Materials compares the performance of MRFSeq with the variant where NOISeq is incorporated. We plan to make MRFSeq flexible so it can be combined with any DE analysis tool in the near future.

Our experiments used COXPREdb (Obayashi and Kinoshita, 2011), which consists of coexpression data for seven model organisms. One could also consider using other sources of coexpression data such as ACT (Manfield *et al*., 2006), ATTED-II (Obayashi *et al*., 2007), CSB.DB (Steinhauser *et al*., 2004), CoP(Ogata *et al*., 2010), etc. or constructing custom gene coexpression networks from publicly available expression data such as GEO (http://www.ncbi.nlm.nih.gov/geo/), ENCODE (http://encodeproject.org/ENCODE/), modENCODE (http://www.modencode.org/), etc., especially for organisms (or tissues) not covered by COXPREdb. Moreover, the threshold ρ used for extracting pairs of coexpressed genes from a given gene coexpression network may also have an impact on the performance of our algorithm. We set $\rho = 0.5$ empirically in our experiments based on the literature (Patil *et al*., 2011; Watson, 2006) and some preliminary tests on the MAQC data. Clearly, a higher ρ may decrease the sensitivity of MRFSeq, while a lower ρ may decrease the precision of MRFSeq. We plan to explore the impact of different coexpression networks (including the choice of ρ) on the performance of MRFSeq and study automatic methods for choosing an optimal ρ in future work.

ACKNOWLEDGEMENT

The authors are grateful to the anonymous referees for their constructive comments. The research was partially supported by a University of California, Riverside, Presidential Chair Professorship fund, the National Science Foundation grant DBI-1262107, and the National Natural Science Foundation of China grant 61175002.

Conflict of Interest: none declared.

REFERENCES

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

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

Besag,J. (1974) Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc.*, **B 36**, 192–236.

Besag,J. (1986) On the statistical analysis of dirty pictures. *J. R. Stat. Soc.*, **B 48**, 259–302.

Boykov,Y. (2001) Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 1222–1239.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in MRNA-seq experiments. *BMC Bioinformatics*, **11**, 94.

Carlson,M.R. *et al.* (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, **7**, 40.

Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–63.

Edmonds,J. and Karp,R.M. (1972) Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, **19**, 248–264.

Fisher,R.A. (1922) On the interpretation of χ^2 from contingency tables, and the calculation of p. *J. R. Stat. Soc.*, **85**, 87–94.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Frazee,A.C. *et al.* (2011) Recount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.

Goulden,C.H. (1956) *Methods of Statistical Analysis*. 2nd edn. Wiley, New York.

Graveley,B.R. *et al.* (2011) The developmental transcriptome of drosophila melanogaster. *Nature*, **471**, 473–479.

Griffith,M. *et al.* (2010) Alternative expression analysis by RNA sequencing. *Nat. Methods*, **7**, 843–847.

Hardcastle,T.J. and Kelly,K.A. (2010) bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Jacob,L. *et al.* (2012) More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.*, **6**, 561–600.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kendziorski,C.M. *et al.* (2003) On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.*, **22**, 3899–3914.

Kindermann,R. and Snell,J.L. (1980) *Markov Random Fields and Their Applications*. American Mathematical Society, Providence, RI.

Kolmogorov,V. and Zabih,R. (2004) What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 147–159.

Leinonen,R. *et al.* (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.

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

Manfield,I.W. *et al.* (2006) Arabidopsis co-expression tool (act): web server tools for microarray-based gene expression analysis. *Nucleic Acids Res.*, **34**, W504–W509.

Marioni,J.C. *et al.* (2008) RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Meyer,L.R. *et al.* (2013) The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.

Mooij,J.M. and Kappen,H.J. (2007) Sufficient conditions for convergence of the sum-product algorithm. *IEEE Trans. Inf. Theory*, **53**, 12.

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, **320**, 1344–1349.

Newton,M.A. *et al.* (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol.*, **8**, 37–52.

Obayashi,T. *et al.* (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in arabidopsis. *Nucleic Acids Res.*, **35**, D863–D869.

Obayashi,T. and Kinoshita,K. (2011) Coxpresdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res*., **39**, D1016–D1022.
Ogata,Y. *et al*. (2010) COP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics*, **26**, 1267–1268.
Oshlack,A. *et al*. (2010) From RNA-Seq reads to differential expression results. *Genome Biol*., **11**, 220.
Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-Seq data confounds systems biology. *Biol. Direct*, **4**, 14.
Patil,A. *et al*. (2011) Assessing the utility of gene co-expression stability in combination with correlation in the analysis of protein-protein interaction networks. *BMC Genomics*, **12** (**Suppl. 3**), S19.
Rahnenfuhrer,J. *et al*. (2004) Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol*., **3**, 16.
Rapaport,F. *et al*. (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.
Robinson,M.D. *et al*. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, **9**, 321–332.
Shi,L. *et al*. (2006) The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*., **24**, 1151–1161.
Sivachenko,A. *et al*. (2005) Identifying local gene expression patterns in biomolecular networks. *IEEE Comput. Syst. Bioinform. Conf*., 180–184.
Srivastava,S. and Chen,L. (2010) A two-parameter generalized poisson model to improve the analysis of RNA-Seq data. *Nucleic Acids Res*., **38**, e170.
Steinhauser,D. *et al*. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, **20**, 3647–3651.
Stuart,J.M. *et al*. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
Tarazona,S. *et al*. (2011) Differential expression in RNA-Seq: a matter of depth. *Genome Res*., **21**, 2213–2223.
Trapnell,C. *et al*. (2009) Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–515.
Trapnell,C. *et al*. (2013) Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nat. Biotechnol*., **31**, 4653.
van Rijsbergen,C.J. (1979) *Information Retrieval*. 2nd edn. Butterworth, London.
Wang,L. *et al*. (2010) DEGSeq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.
Wang,Z. *et al*. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*., **10**, 57–63.
Watson,M. (2006) CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, **7**, 509.
Wei,Z. and Li,H. (2007) A markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**, 1537–1544.
Weiss,Y. (2000) Correctness of local probability propagation in graphical models with loops. *Neural Comput*., **12**, 1–41.
Young,M.D. *et al*. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*., **11**, R14.
Zhang,J. *et al*. (2011) BioMart: a data federation framework for large collaborative projects. *Database (Oxford)*, **2011**, bar038.
Zheng,S. and Chen,L. (2009) A hierarchical bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res*., **37**, e75.