BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 2 2014, pages 180–188
doi:10.1093/bioinformatics/btt624

Gene expression
Advance Access publication November 26, 2013

Flexible analysis of RNA-seq data using mixed effects models

Ernest Turro[1,2,*], William J. Astle[3] and Simon Tavaré[1]

[1]Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK, [2]Department of Haematology, University of Cambridge, NHS Blood and Transplant, Long Road, Cambridge CB2 0PT, UK and [3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal QC H3A 1A2, Canada

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Most methods for estimating differential expression from RNA-seq are based on statistics that compare normalized read counts between treatment classes. Unfortunately, reads are in general too short to be mapped unambiguously to features of interest, such as genes, isoforms or haplotype-specific isoforms. There are methods for estimating expression levels that account for this source of ambiguity. However, the uncertainty is not generally accounted for in downstream analysis of gene expression experiments. Moreover, at the individual transcript level, it can sometimes be too large to allow useful comparisons between treatment groups.

Results: In this article we make two proposals that improve the power, specificity and versatility of expression analysis using RNA-seq data. First, we present a Bayesian method for model selection that accounts for read mapping ambiguities using random effects. This polytomous model selection approach can be used to identify many interesting patterns of gene expression and is not confined to detecting differential expression between two groups. For illustration, we use our method to detect imprinting, different types of regulatory divergence in *cis* and in *trans* and differential isoform usage, but many other applications are possible. Second, we present a novel collapsing algorithm for grouping transcripts into inferential units that exploits the posterior correlation between transcript expression levels. The aggregate expression levels of these units can be estimated with useful levels of uncertainty. Our algorithm can improve the precision of expression estimates when uncertainty is large with only a small reduction in biological resolution.

Availability and implementation: We have implemented our software in the `mmdiff` and `mmcollapse` multithreaded C++ programs as part of the open-source MMSEQ package, available on https://github.com/eturro/mmseq.

Contact: et341@cam.ac.uk

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

Received on May 9, 2013; revised on September 2, 2013; accepted on October 28, 2013

1 INTRODUCTION

High-throughput sequencing of RNA (RNA-seq) is superseding microarrays as the standard technology for genome-wide profiling of RNA samples and for differential expression analysis. RNA-seq is a technique for generating millions of short subsequences called reads from a population of RNA transcripts in a biological sample. Roughly speaking, the number of reads generated by a transcript is proportional to its length and to the number of copies of the transcript in the sample. Consequently, given an accurate mapping of reads back to transcripts (e.g. from sequence alignments), it is possible to quantify transcript expression levels from read counts.

Sequencing offers greater dynamic range than microarrays and, by providing direct observations of complementary DNA sequences, it allows improved discrimination among isoforms and haplotypes. Although the methodological literature for gene expression profiling with microarray data is mature, there remain aspects of RNA-seq data analysis that require further development. In particular, the principal approaches currently used for differential expression analysis with RNA-seq (Anders and Huber, 2010; Hardcastle and Kelly, 2010; Robinson and Oshlack, 2010; Tarazona *et al*., 2011) make comparisons between treatment classes using statistics derived from sequence alignment counts. Such approaches are useful for assessing differential gene expression because, for most genes, an exact read count can be obtained that is proportional to its expression level. However, methods based on read counts are not robust to variability in the relative expression of isoforms because the length of each isoform—not only its expression level—influences the expected number of gene-level read counts. It can be more accurate to estimate expression at the gene level by adding up estimates made at the transcript level than by counting raw alignments as though each gene produced a single transcript of canonical length (Wang *et al*., 2010).

A distinct advantage of RNA-seq is that sequencing along splice junctions and heterozygous loci facilitates the estimation of expression levels for gene isoforms and even haplotype-specific isoforms in polyploid organisms. At the isoform level, raw counts are not available because the sharing of exons between isoforms of the same gene means that reads may align to multiple transcripts. Moreover, sequence homology between genes can result in reads that map to transcripts belonging to different genes. For haplotype-specific analysis, sequence-sharing is even more extreme as only reads mapping to heterozygous loci allow transcripts from different haplotypes to be distinguished. Deconvolution algorithms, which account for the read mapping ambiguity, can provide estimates of the expression levels of each transcript and estimates of the associated uncertainties (Glaus *et al*., 2012; Turro *et al*., 2011). Given the large amounts of

*To whom correspondence should be addressed.

sequence-sharing between transcripts in mammalian transcriptomes, these uncertainties can be considerable and highly variable. Consequently, it is important to account for differential uncertainty in any statistical analysis based on expression estimates, e.g. when comparing expression levels between treatment groups. This idea has been pursued previously in the context of microarray analysis (e.g. Hein and Richardson, 2006; Liu *et al.*, 2006) but has received less attention in the context of RNA-seq analysis.

The Cuffdiff 2 software (Trapnell *et al.*, 2012) implements frequentist hypothesis tests to detect differential expression and regulation. The methods rely on a number of distributional approximations, are limited to two-class comparisons and cannot be used for haplotype-specific inference. Recently, a Bayesian model of Markov chain Monte Carlo (MCMC) traces for expression parameters has been proposed (Glaus *et al.*, 2012) that is attractive because it accounts for the shape of the posterior uncertainties in the parameters. However, it too is limited to a two-condition study design and relies on a computationally intensive read-level algorithm to generate the MCMC traces.

2 APPROACH

In this article, we propose a Bayesian mixed model approach to the analysis of multisample RNA-seq expression data that accounts for the posterior uncertainty in expression estimates, including the uncertainty due to read mapping ambiguity. Figure 1 is a motivating illustration showing how a method accounting for the uncertainty in the expression parameters can have more power to detect differential expression between two treatment groups than a method that assumes that all expression parameters are known with equal accuracy. Up-weighting observations with lower posterior uncertainty helps to recover the signal from the noise. In real datasets, the variability in posterior uncertainty can be considerable (Supplementary Fig. S1).

Our method identifies transcripts with scientifically interesting patterns of expression by making a statistical comparison of regression models. Consequently, it is flexible and can be applied to a wide range of experimental designs. The uncertainty in expression estimates is incorporated via random effects. In general, this means that the more precise an estimate the more information it will contribute to the comparison of models. We present an MCMC algorithm for posterior inference that uses the method of Carlin and Chib (1995) for making the model comparisons. The method we propose can be used to identify more intricate variations in patterns of expression than the usual gene-level differentiation between treatment groups. For example, we can model estimates of haplotype-specific expression in first-generation crosses of inbred strains to detect imprinted genes or transcripts. We can also detect differential isoform usage by applying our method to the probit transform of the expression level of an isoform expressed as a proportion of the overall expression of its gene.

Finally, in cases where the posterior uncertainty in expression estimates is extreme, it can be reduced by collapsing transcripts into identifiable aggregates. We propose an algorithm to generate such a collapsing using the output of the MMSEQ (Turro *et al.*, 2011) method for estimating transcript-specific expression levels. Briefly, MMSEQ infers transcript expression levels using an MCMC algorithm for a Bayesian model of read counts. The model accounts for Poisson noise and various sources of technical bias. In addition, it accounts for any uncertainty about which transcript generated which read by integrating over all possibilities. We demonstrate that aggregation of transcripts based on the MMSEQ output does not preclude biologically meaningful inference, as it tends to apply to sets of transcripts that share considerable stretches of sequence and important biological attributes (such as skipped or retained exons). This approach helps counteract some of the estimation difficulties posed by the existence of highly complex gene structures.

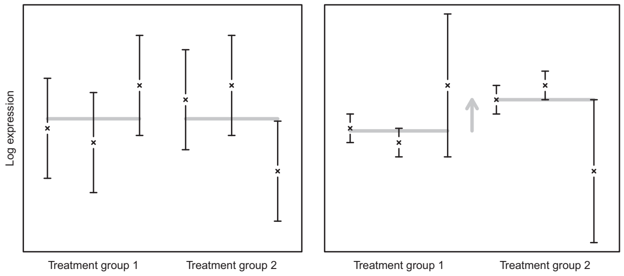

Fig. 1. Accounting for the posterior uncertainty in log-expression can improve the power to detect differential expression. The left panel illustrates a hypothetical analysis that ignores heterogeneity in posterior uncertainty, so that the total error (measurement plus experimental) has the same variance (bars) across the observations (crosses). The right panel illustrates a hypothetical analysis, which accounts for the error in the point measurements of log-expression (posterior means) using random effects, thereby exposing a difference between the two treatment groups (grey lines)

3 METHODS

3.1 Linear mixed model

Consider a single feature (i.e. gene, isoform or haplotype-specific isoform) with expression μ_i in sample i. If μ_i is known with complete precision for each i, regressions of the following type may be used to model the biological variation across samples:

$$\log \mu_i = \mathbf{P}_i^{(m)} \eta^{(m)} + \epsilon_i^{(m)} \tag{1}$$

Here, $\mathbf{P}_i^{(m)}$ is the ith row of design matrix $\mathbf{P}^{(m)}$, which defines the statistical model m, $\eta^{(m)}$ is a corresponding vector of regression coefficients and $\epsilon_i^{(m)}$ is a Gaussian error with mean zero. Biological inferences may be drawn through a statistical comparison of competing models, where the true model is denoted by $\gamma \in \{0, 1\}$. Although we consider only two choices for γ in any given model comparison, an arbitrary number of models can be considered through multiple pairwise comparisons. The choice of $\mathbf{P}^{(m)}$ should depend on the structure of the experiment and the scientific hypothesis. For example, in a differential expression experiment comparing two treatment groups of size three, the following $\mathbf{P}^{(m)}$s would be appropriate:

$$\mathbf{P}^{(0)} = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)^T \tag{2}$$

$$\mathbf{P}^{(1)} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix}^T \tag{3}$$

In practice, the $\log \mu_i$ are not known precisely but are estimated with varying degrees of statistical uncertainty. The MMSEQ method (Turro *et al.*, 2011) summarizes this statistical uncertainty in the Monte Carlo samples of $\log \mu_i$ generated by an MCMC algorithm. The posterior distribution of each $\log \mu_i$ can be approximated using a Gaussian distribution with mean and variance equal to the empirical mean y_i and variance v_i^2 of the corresponding Monte Carlo sample. The posterior mean y_i can be treated as an estimator of $\log \mu_i$. Because the MMSEQ prior on $\log \mu_i$ is vague, y_i and $1/v_i^2$ should closely approximate the maximum likelihood estimator and Fisher information of $\log \mu_i$, respectively. Consequently, approximately:

$$y_i|\mu_i \sim N(\log \mu_i, \upsilon_i^2) \tag{4}$$

Substitution of (4) in (1) leads to the following mixed model with a random effect, which accounts for the uncertainty in the $\log \mu_i$:

$$y = \mathbf{P}^{(m)}\eta^{(m)} + \nu^{(m)} + \epsilon^{(m)}$$
$$\nu_i^{(m)} \sim N(0, \upsilon_i^2).$$

We have yet to specify the variance structure of the $\epsilon^{(m)}$. One option is to assume homoscedasticity across all samples, but grouped data are common in gene expression experiments and thus it may be more appropriate to specify a different variance parameter for each grouping. For convenience, we adopt the latter parameterization in this article:

$$\epsilon_i^{(m)} \sim N\left(0, \sigma^{2^{(m)}}_{c^{(m)}(i)}\right)$$

where $i \mapsto c^{(m)}(i)$ maps observation i to grouping category $c^{(m)}(i)$ under model m. Note that the υ_i^2 are fixed quantities, and therefore the $\sigma_c^{2(m)}$ are identifiable.

Having specified a probability model for the data, we now consider inference. The parameters of primary scientific interest are the model indicator γ and the regression coefficients of the true model $\eta^{(\gamma)}$. We prefer the Bayesian over the frequentist approach on foundational grounds, but there are also practical advantages. It is difficult to apply standard frequentist procedures, such as the likelihood ratio test, to this problem because there is a non-zero (frequentist) probability that the maximum likelihood estimate of each of the $\sigma_c^2(m)$ will lie on the boundary of the parameter space (i.e. at 0). The boundary event occurs when the estimates of the posterior variance, given by the v_i^2, are sufficiently large to explain the empirical variability of the y. When there is a non-zero probability of a maximum likelihood estimate lying on a boundary, standard asymptotic results for the distribution of frequentist test statistics (e.g. Wilks's theorem) may fail to apply (Ferguson, 1996). Even when frequentist theory does hold (e.g. if all $v_i^2 = 0$), standard frequentist approaches may nonetheless suffer from overfitting when sample sizes are low (in the order of ten to a hundred), as is common in gene expression experiments, leading to unreliable inference (Gelman, 2004, p. 371). The Bayesian approach allows us to specify priors for hyperparameters, which restrict overfitting. Specifically, our prior for η penalizes large coefficients, reflecting a prior belief that big log fold changes are rare. However, because prior belief about the location of the data is vague, we prefer to separate the intercept term α from η and avoid including a constant column in $\mathbf{P}$ [cf. (2) and (3)]:

$$y = \alpha^{(m)} + \mathbf{P}^{(m)}\eta^{(m)} + \nu^{(m)} + \epsilon^{(m)}$$

It is not always possible to design an experiment controlling for all extraneous variables, such as gender, thought to affect gene expression levels. However, such variables can be accounted for statistically by including covariates in the regression model. We generalize the probability model for the data previously given, by specifying a model-independent covariate matrix $\mathbf{M}$ with corresponding vector of regression coefficients β:

$$y = \alpha^{(m)} + \mathbf{M}\beta^{(m)} + \mathbf{P}^{(m)}\eta^{(m)} + \nu^{(m)} + \epsilon^{(m)}$$

We now describe the choice of priors for the parameters. We specify independent Gaussian priors for the $\alpha^{(m)}$ and $\beta^{(m)}$ and Student's t priors for the $\eta^{(m)}$. For the error term variance parameters, we specify $\sigma_c^{2(m)} \sim \text{Inverse} - \text{Gamma}(\text{shape} = k/2, \text{scale} = k\rho^{(m)}/2)$. We shrink the $\sigma_c^{2(m)}$ towards a common mean by specifying a common Gamma prior on $\rho^{(m)}$, reflecting our prior belief that the error term variance will be similar across categories (see Supplementary Material for details). Finally, we place a Bernoulli prior on γ with fixed hyperparameter p, reflecting our degree of prior belief that model 1 rather than model 0 is true.

We have implemented an MCMC algorithm to generate samples from the joint posterior distribution of these models. For model comparison, we use the pseudo-prior method of Carlin and Chib (1995), which can be used to generate a Bayes factor and therefore an estimate of the posterior probability for $\gamma = 1$. Details of the MCMC algorithm can be found in the Supplementary Material.

3.2 Assessment of random effects model

To assess our method, we simulated expression values and standard deviations for 80 000 features across 10 samples. For the first 20 000 features, there is no $\mathbf{P}$ (homogeneous expression model) and for the remaining features $\mathbf{P}$ has a single column such that $\mathbf{P}_{i1}^{(D)} = \frac{1}{2}$ if $i \in \{1, \ldots, 5\}$ and $\mathbf{P}_{i1}^{(D)} = -\frac{1}{2}$ if $i \in \{6, \ldots, 10\}$ (two-condition differential expression model). For the first 20 000 features, $\alpha = 0$, $\beta = \eta = 0$ and $\sigma_c = 0.3$.

To mimic the heteroscedasticity of the posterior variances and their correlation with the expression values, we sampled the v_i from posterior standard deviations obtained using a real dataset (Brooks *et al.*, 2011). Specifically, we sampled the standard deviation from values in the real dataset where the estimate of $\log \mu$ was within 0.5 log fold change of the simulated expression value. In this way, our simulation accounts for the fact that more highly expressed features tend to be estimated more precisely (on the log scale) (Supplementary Fig. S1).

The data for the remaining 60 000 features are generated in the same way except using the two-condition model with log fold changes of 1, 2 and 4 and α values of 0.5, 1 and 2, respectively, for successive sets of 20 000 features (thus keeping the baseline expression fixed at around zero). For the inference, we specify a model without a $\mathbf{P}^{(0)}$ and with a $\mathbf{P}^{(1)} = \mathbf{P}^{(D)}$, and we use a prior probability that $\gamma = 1$ of 0.1. For the first 20 000 features, therefore, as we do not simulate any differential expression, model 0 is the true model whilst for the remaining features, as we simulate a non-zero log fold change, model 1 is the true model.

Each box plot in the left panel of Figure 2 summarizes the distribution (of the MCMC point estimates) of the posterior probabilities for $\gamma = 1$ in the group of features corresponding to a particular simulated log fold change. Evidently, increases in the simulated fold-change lead to model 1 being selected preferentially over model 0.

The main motivation for developing the random effects method was to account for posterior uncertainty about the values of expression parameters (including uncertainty due to multi-mapping of reads to transcripts) in the statistical model linking expression levels to treatment groups. To assess the impact of accounting for heterogeneous errors on the accuracy of inference, we reanalyzed the simulated data using the method already described but setting the $v_i = 0$. This inference model does not account for heteroscedasticity in uncertainties about expression parameters.

We found that when the log fold change is 0 and therefore the true model is model 0, the posterior probability that $\gamma = 1$ tends to be lower in the analysis with the non-zero v_i, which corresponds to a lower type I error rate. Furthermore, when the log fold change is positive, the posterior probability that $\gamma = 1$ tends to be larger when the v_i are non-zero, which corresponds to greater power to detect differential expression. The advantage in sensitivity and specificity held by the method taking account of the posterior uncertainty in expression estimates is clearly visible in the receiver operator characteristic (ROC) curves comparing the two approaches (Fig. 2, right panel).

As ROC curves evaluate a classifier assuming that the population proportion of each of the two values of the binary variable are equal, we also show positive predictive value (PPV) and negative predictive

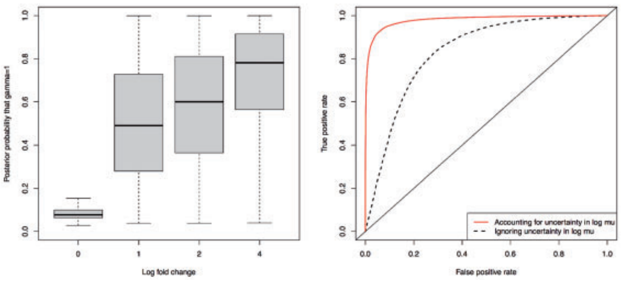

Fig. 2. Assessment of model selection algorithm. The box plots (left panel) summarize distributions (over simulated features) of MCMC estimates of the posterior probability in favour of a two treatment group differential expression model ($\gamma = 1$), when comparing that model to a null model assuming no differential expression ($\gamma = 0$). Each plot corresponds to analyses of simulated data with a different log fold change in expression between treatment groups. Therefore, the leftmost plot corresponds to features for which $\gamma = 0$, whereas the other plots correspond to features for which $\gamma = 1$. The ROC curves (right panel) illustrate the superiority of a classifier that accounts for differential uncertainty in the $\log \mu_i$ over one that does not. Both classifiers threshold on the MCMC estimate of the posterior probability that $\gamma = 1$. The posterior probability estimates for the green/dashed curve were generated by the model accounting for uncertainty in $\log \mu_i$ through random effects, whereas those for the black/solid curve were generated by the same model without random effects ($\upsilon_i^2 = 0$)

value (NPV) plots for varying proportions of differentially expressed (DE) versus non-differentially expressed features (Supplementary Fig. S2). The inclusion of the posterior uncertainty in read counts improves both PPV and NPV. The PPV improvement is most noticeable when the proportion of DE features is small while the NPV improvement is most noticeable when the proportion of DE features is large. When we simulate 20 000 non-DE and 6000 DE features (2000 features per fold-change class), we observe a drastic improvement: for a posterior probability threshold of 0.45, incorporating read count uncertainty results in almost perfect PPV, whereas assuming no uncertainty results in a PPV of around 0.5.

3.3 Collapsing sets of transcripts

We have shown how it is possible to account for posterior uncertainty in $\log \mu_i$ using random effects. When the variance of the random effects is large, there will be low power to distinguish between the models using the analysis described earlier in the text. Unfortunately, in all RNA-seq experiments there is a large class of transcripts for which the posterior variance is large—specifically, transcripts that are not mapped to uniquely by any read (e.g. black points in Supplementary Fig. S1).

Often, sets of transcripts with poorly estimated expression parameters are anti-correlated because reads can be mapped to the combined set of transcripts with confidence, but no reads can be mapped to specific transcripts within the set. In these circumstances, it may be more informative to treat the set of transcripts as the unit of inference, rather than the individual transcripts.

We now propose an algorithm for collapsing transcripts with anti-correlated expression parameters into informative sets. Given a set S in a partition of the transcripts, we define an aggregate expression parameter $\phi_{iS} \equiv \sum_{t \in S} \mu_{it}$, where μ_{it} is the expression parameter for transcript t in sample i. At each iteration of the algorithm, we coarsen the current partition of the transcripts by combining the two sets S and S' for which the mean posterior correlation between ϕ_{iS} and $\phi_{iS'}$ is the least, where the mean is taken over i (i.e. we combine the pair of sets that have on average the most anti-correlated expression parameters). The algorithm terminates when the minimum (over transcripts) mean (over samples) posterior correlation exceeds a stopping threshold. We use the right tail of the empirical distribution of maximum mean correlations as a control and set the stopping threshold to minus the 97.5th percentile by default. We have found that this threshold leads to reasonably symmetric distributions between the minimum and maximum mean correlations while being robust to small numbers of spurious highly correlated transcripts (Supplementary Fig. S4). To initialize the algorithm, we need to choose the set of transcripts to collapse and a partition on that set. We now discuss concrete choices.

Empirically, the log expression parameters of transcripts with u uniquely mapping reads almost always have higher posterior standard deviations than the log expression parameters of transcripts with $u+1$ uniquely mapping reads (see the stratified pattern in Supplementary Fig. S1). In practice, the expression of transcripts for which $u \geq 1$ can usually be estimated with a reasonable degree of accuracy. In order not to collapse transcripts for which we have reasonably precise estimates in at least some of the samples, we apply our collapsing algorithm only to transcripts in the $u=0$ stratum and for which the posterior standard deviation is greater than the maximum posterior standard deviation among transcripts with $u \geq 1$. We call transcripts in this set 'low-information transcripts'. We initially collapse transcripts with identical sequences as, by necessity, their expression parameters are unidentifiable. We then collapse the remaining low-information transcripts. However, low-information transcripts that are not observed (i.e. have zero alignments across all samples) are excluded from collapsing.

3.4 Assessment of collapsing algorithm

We use the well-known Pasilla dataset (Brooks *et al.*, 2011), which includes RNA-seq reads from seven *Drosophila melanogaster* cell cultures in two conditions [4 control and 3 ps(RNAi) samples], to assess our collapsing algorithm. The top-right panel of Figure 3 shows a trace of the minimum mean correlation as the algorithm progresses. There is a gradual increase in the minimum mean correlation interrupted by sudden downward fluctuations. Supposing transcript sets S and S' are such that $\text{mean}_i\text{cor}(\phi_{iS}, \phi_{iS'})$ is lowest at iteration q, a sudden downward fluctuation in the minimum mean correlation occurs at iteration $q+1$ if there exists a set S'' such that $\text{mean}_i\text{cor}(\phi_{iS} + \phi_{iS'}, \phi_{iS''}) \ll \text{mean}_i\text{cor}(\phi_{iS}, \phi_{iS'})$. For example, the expression parameters for transcripts RC and RB of the five-transcript gene CG42671 have a mean correlation of -0.61. However, the expression of the collapsed pair has a much lower correlation of -0.95 with that of a third transcript, RE. The collapsed triplet shares an extremely similar structure, including a $3'$ exon that is unique to the three transcripts (Fig. 4). As another example, consider Supplementary Figure S3, which shows how nine transcripts from the up gene sharing the same two exons near the $5'$ end cluster together while a 10th transcript with a different $5'$ exon, up-RD, is estimated separately.

Overall, the algorithm collapsed 1368 low-information transcripts from the *Drosophila* dataset into 570 higher-precision sets. Of these 570 transcript sets, 546 solely contained isoforms belonging to the same gene, even though the algorithm uses only posterior correlations and is thus blind to each isoform's gene membership, sequence and coordinates. Only 22 and two sets contained isoforms from two and three different genes, respectively. The sets of three consisted of paralogues of the U2 snRNAs (14B, 38ABa and 38ABb) and the TEKTIN protein family (CG32819, CG32820, CG17450). The sets of two consisted mostly of duplicated neighbouring genes (e.g. CG31809 and CG31810) or highly overlapping genes (e.g. SP555 and CG14042). These examples illustrate how our algorithm can improve precision by collapsing groups of resemblant transcripts without a significant loss in biological resolution.

We next examined the consistency of our algorithm through cross-validation, comparing collapsed sets obtained using different bipartitions

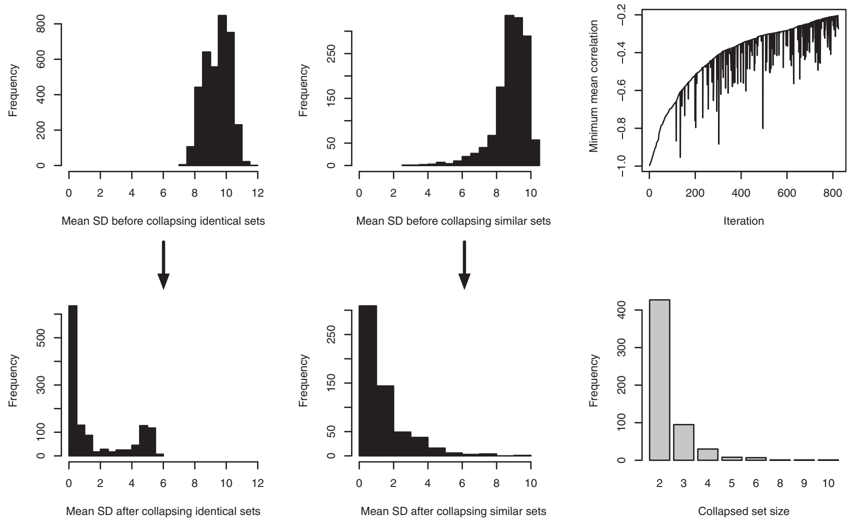

Fig. 3. Left panels: histograms of mean posterior standard deviations for sets of identical transcripts before (top) and after (bottom) collapsing. Central panels: histograms of mean posterior standard deviations for sets of non-identical low-information transcripts before (top) and after (bottom) collapsing. Right panels: trace of the minimum mean correlation as the collapsing algorithm progresses. Bottom-right panel: histogram of the set sizes after collapsing sets of non-identical low-information transcripts

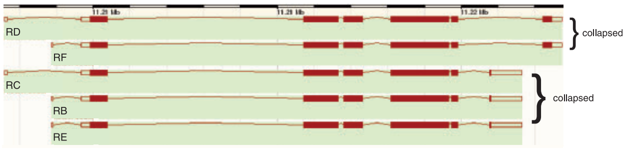

Fig. 4. The algorithm collapses the five transcripts of gene CG42671 into two groups. One group contains transcripts RB, RC and RE, and the other group contains transcripts RD and RF. Transcripts within the same group share the same 3′ exon

of six of the seven samples in the *Drosophila* dataset. We assessed the consistency between the two groups formed by each of the 10 (i.e. $\binom{6}{3}/2$) possible bipartitions of equal size. A collapsed set S in group 1 is said to be consistent with the collapsed sets in group 2 if either of the following conditions is met:

- S is a subset of a set in group 2,
- all sets in group 2 containing an element in S are subsets of S.

Thus, $\{A, B, C\}$ in group 1 would be consistent with $\{\{A\}, \{B, C\}\}$ or $\{A, B, C, D\}$ in group 2 but not with $\{A, D\}$. Also, note that $\{\{A, B\}, \{C, D\}\}$ is not consistent with $\{\{A, C\}, \{B, D\}\}$. For each bipartition, the consistency of each set in group 1 was assessed with respect to all sets in group 1 and *vice versa*. Despite the dataset containing mixed single and paired-end libraries and two different conditions, our consistency rate averaged 96.7%. The lowest consistency, of 95.84%, was obtained in the bipartition separating the treatment and the control group exactly. Thus, even in the presence of structured biological variability, our algorithm merges transcripts in a highly consistent manner.

4 APPLICATIONS

4.1 Finding imprinted genes in mice

Mammalian cells can express the (usually) two copies of autosomal genes differentially, leading to allele-specific imbalance in their RNA products. When the direction of the imbalance between the two copies depends on the sex of the parent from whom each copy was inherited, imprinting is said to occur. It is thought that imprinting is mostly determined by differential methylation during gametogenesis (Li and Sasaki, 2011). We analyzed previously published RNA-seq data obtained from the livers of six initial and six reciprocal crosses of inbred mice (Goncalves *et al.*, 2012) to assess whether our model selection algorithm can be used to detect imprinting. The initial crosses inherited the C57BL/6J (BL6) strain genome from the father and the CAST/EiJ (CAST) strain genome from the mother. The reciprocal crosses inherited the CAST genome from the father and the BL6 genome from the mother. We obtained haplotype- and gene-specific estimates for each strain within each cross using MMSEQ. Thus we obtained two sets of posterior means and standard deviations for each initial and reciprocal mouse.

We applied our model selection algorithm comparing a null versus an imprinting model. Under the null model, it is assumed that any difference between the two strains in the initials is the same in the reciprocals. Under the imprinting model, it is assumed that the difference between the two strains has the same magnitude in the initials as the reciprocals, but has opposite sign. The collapsed design matrices (i.e. with repeat rows removed) are as follows:

$$\mathbf{P}^{(\text{null})} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix} \quad \mathbf{P}^{(\text{imp})} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & -1 \\ -1 & 1 \end{pmatrix} \begin{array}{l} \ldots\text{F1i BL6} \\ \ldots\text{F1i CAST} \\ \ldots\text{F1r BL6} \\ \ldots\text{F1r CAST} \end{array}$$

Each row of the two matrices corresponds to a cross-classification of the mice by haplotype (BL6/CAST) and cross type (F1 initial/F1 reciprocal, abbreviated F1i/F1r). There are six observations of each kind so the uncollapsed matrices used in the analysis have 24 rows, a six-fold replication of those above. The regression coefficient $\eta_1^{(\text{null})}$ (which corresponds to the first column of the design matrix $\mathbf{P}^{(\text{null})}$) represents the log fold change between the initial and the reciprocal crosses while the coefficient $\eta_2^{(\text{null})}$ represents the log fold change between the two haplotypes. The coefficient $\eta_1^{(\text{imp})}$ has the same interpretation as $\eta_1^{(\text{null})}$, but the $\eta_2^{(\text{imp})}$ represents the log fold change between the maternally and the paternally inherited haplotype.

We restricted our analysis to polymorphic genes, as there is no power to detect imprinting unless there is a difference in sequence between the BL6 and CAST haplotypes. We used a prior probability of imprinting of 10% to calculate the posterior probabilities. For each gene, we checked whether it was listed in the following four online imprinting databases: WAMIDEX (Schulz *et al.*, 2008), geneimprint.com, mousebook.org and the Catalogue of Imprinted Genes and Parent-of-Origin Effects in Humans and Animals (Morison *et al.*, 2001).

The 10 genes with the highest posterior probability of being imprinted are well known to be imprinted in mouse, as they are listed in all four of the aforementioned databases. These genes are H13, Igf2r, Meg3, Slc22a3, Rian, Snrpn, Sgce, Impact, Zrsr1 and Peg3, all of which have a posterior probability of imprinting >99%. If we relax the threshold on the posterior probability to 90%, we find an additional five genes, three of which have supporting evidence in the literature or in the databases of being imprinted in mouse: Mas1, Mirg and Mcts2 (Fig. 5, left panel). Our results are in general agreement with the findings of Goncalves *et al.* (2012) (Fig. 5, right panel), with a few additional identifications of known genes such as Rian [posterior probability (pp) = 1.00] Peg3 (pp = 1.00), Mirg (pp = 0.99), H19 (pp = 0.70), Rtl1 (pp = 0.58), Mkrn3 (pp = 0.54), Peg10 (pp = 0.43) and Igf2 (pp = 0.31).

4.2 Classifying mouse genes by type of regulatory divergence

Thus far, we have focused on pairwise comparisons between pairs of models, i.e. $\gamma \in \{0, 1\}$. However, it is straightforward to perform polytomous classification through pairwise comparisons of Bayes factors between each model and an arbitrary baseline model (e.g. model 0):

$$\mathbf{P}(\gamma = m|y) = \frac{\text{Bayes factor}(0, m) \times \text{P}(\gamma = m)}{\sum_{m'} \text{Bayes factor}(0, m') \times \text{P}(\gamma = m')} \tag{5}$$

This type of polytomous analysis lends itself to problems where there are several plausible models for the expression summaries, which correspond to different biological mechanisms. Returning to the mouse liver dataset, it is possible to discern different types of regulatory divergence for each gene by comparing expression summaries obtained in the pure-strain mice (F0s) with those obtained in first-generation crosses of the two strains (F1s). Briefly, genes for which there is no difference between the strains in the F0s and no difference between the haplotypes in the F1s are considered to have conserved regulation. Genes for which the difference between the strains is the same in the F0s as in the F1s are considered to have diverged through *cis*-acting regulatory mutations. If a difference in the F0s is completely lost in the F1s, then the gene is said to have diverged through *trans*-acting mutations. Finally, if there are different fold changes in the F0s compared with the F1s, then the gene is said to have diverged through a combination of *cis* and *trans*-acting mutations (Goncalves *et al.*, 2012).

Here, we assess whether our general model selection framework can be used to discern the four patterns of gene expression divergence. We define the following four collapsed design matrices for the different models of regulatory divergence where, for each row of each design matrix, we show the corresponding class of observation on the right:

$$\mathbf{P}^{(\text{conserved})} = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{array}{l} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{F0 (both strains)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{F1 (both strains)} \end{array}$$

$$\mathbf{P}^{(\text{cis})} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix} \begin{array}{l} \ldots\ldots\ldots\ldots\ldots \text{F0 BL6} \\ \ldots\ldots\ldots\ldots\ldots \text{F0 CAST} \\ \ldots\ldots\ldots\ldots\ldots \text{F1 BL6} \\ \ldots\ldots\ldots\ldots\ldots \text{F1 CAST} \end{array}$$

$$\mathbf{P}^{(\text{trans})} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 0 \end{pmatrix} \begin{array}{l} \ldots\ldots\ldots\ldots\ldots \text{F0 BL6} \\ \ldots\ldots\ldots\ldots\ldots \text{F0 CAST} \\ \ldots\ldots\ldots\ldots\ldots \text{F1 (both strains)} \end{array}$$

$$\mathbf{P}^{(\text{cis+trans})} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \\ -1 & 0 & -1 \end{pmatrix} \begin{array}{l} \ldots\ldots\ldots \text{F0 BL6} \\ \ldots\ldots\ldots \text{F0 CAST} \\ \ldots\ldots\ldots \text{F1 BL6} \\ \ldots\ldots\ldots \text{F1 CAST} \end{array}$$

The 5804 genes that are not polymorphic between the two strains were excluded from analysis, as there is no power to detect differences in the hybrid strains using sequencing. Imprinting is a confounding factor in this study design because it results in strain-specific imbalance in the F1s that is not driven by regulatory divergence. Therefore, we excluded from analysis all 542 genes with a moderate posterior probability (>0.25) of being imprinted. We attempted to classify the remaining 28 630 genes into one of the four categories. We ran our model selection algorithm comparing the conserved model with each of the three other models. Then, assuming a flat prior probability of 0.25 that any of the four models is true, we calculated the joint posterior probability of the models using Equation (5).

For many of the genes, the data did not favour any model strongly and the joint posterior probabilities resembled the prior model probabilities. In total, 11 135 genes had a posterior probability greater than 0.5 of belonging to one of the categories. We classified these genes into the conserved, *cis*, *trans* and *cis+trans* categories. Figure 6 contrasts the log fold change between the strains within the F0s and the log fold change between the strains within the F1s in each of the classes. Our algorithm produces a clear clustering of genes into four different patterns. The conserved genes cluster around (0,0) (allowing for some error along $y = 0$ due to lower sequencing depth per mouse and strain in the F1s), *cis* genes cluster around $y = x$, *trans* genes cluster around $x = 0$ and *cis+trans* genes are scattered away from the rest. Notice how the *cis+trans* panel shows distinctive gaps at $y = x$ and $x = 0$, where the *cis* and *trans* genes cluster, respectively. The plots show that our model comparison algorithm is able to distinguish multiple patterns of gene expression effectively, while accounting for the major sources of statistical uncertainty.

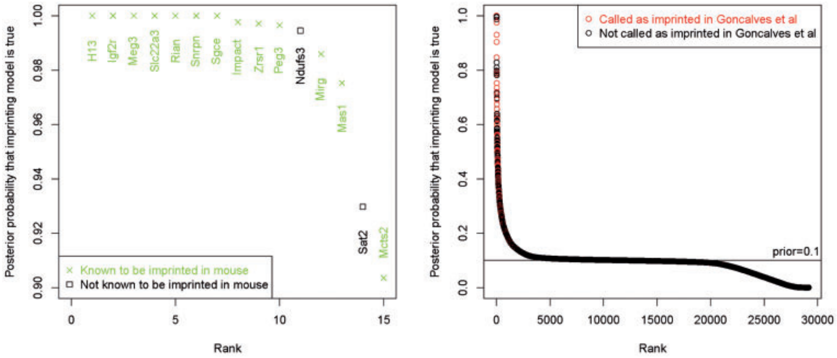

Fig. 5. The left plot shows the posterior probability that the imprinting model for the top-ranking genes with a posterior probability >90%. Genes known to be imprinted in mice are shown as crosses and coloured in green, whereas the two genes not known to be imprinted in mice are shown as squares and coloured in black. The right plot shows the posterior probabilities for all genes, coloured in red if they were identified as imprinted by the custom analysis done by Goncalves *et al.* (2012). The prior probability that the imprinting model is true was set to 0.1 (horizontal line)

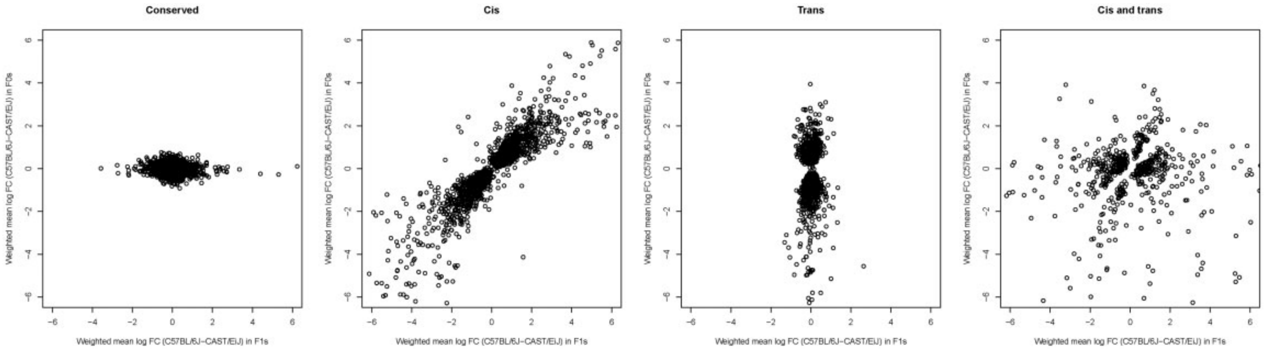

Fig. 6. Each scatterplot shows crude estimated log fold change between the two strains in the F0s (*y*-axes) and the F1s (*x*-axes) for genes confidently classified into one of the four categories. The crude estimates are the mean differences of the posterior means weighted by the inverse of the posterior variances

4.3 Detecting differential isoform usage

Differential regulation of splicing is an important biological phenomenon that can be associated with disease (Garcia-Blanco *et al*., 2004). Methods exist (e.g. Anders *et al*., 2012) for detecting differential usage of exons, which serves as an indicator that differential isoform usage is taking place but does not necessarily pinpoint specifically which isoform is being differentially regulated. It may sometimes be more interesting to quantify changes in the proportions of alternative isoforms of a gene directly rather than through changes at the exon level, as isoforms are the ultimate determinants of protein products. Furthermore, in studies of inbred crosses, as in the aforementioned example, standard exon-level analyses cannot be applied because the detection of differential exon or isoform usage must be haplotype-specific. In contrast, our method can straightforwardly accommodate analyses for arbitrary units of inference (see Section 5).

Here we investigate whether our model selection algorithm can be used to detect differential isoform usage. Given the high level of ambiguity in the assignment of reads to isoforms when the genomic coordinates of isoforms from the same gene overlap, accounting for uncertainty in the observations is potentially more important in this scenario than in gene-level or haplotype-specific gene-level analyses.

We return to the Pasilla dataset, for which 16 RT-PCR-validated cases of differential splicing between the control samples and the ps(RNAi) samples are available from the authors. We first adapted the MMSEQ estimation program to output the posterior distributions of the probit-transformed isoform proportions for each isoform in each gene. At each iteration of the MMSEQ MCMC, the expression level of each isoform is expressed as a proportion of the total expression of its gene, which is probit-transformed and recorded. The empirical means and standard deviations of these traces are then downsampled and used as input to our model selection program (see Fig. 7, left panel for an example).

As in the imprinting example mentioned earlier in the text, we found a clear pattern for the distribution of posterior probabilities consisting of a large population near or below the prior and a small population of posterior probabilities, which rise well above the prior. The maximum posterior probability of isoform usage per gene was below the isoform-wise prior (0.1) for 84.6% of genes (Fig. 7, right panel), showing that in the vast majority of cases there is no evidence of differential isoform usage. Only 225 genes showed moderate evidence of differential isoform usage (maximum posterior probability >0.2). This is somewhat fewer than the number of genes (323) found by Brooks *et al.* (2011)

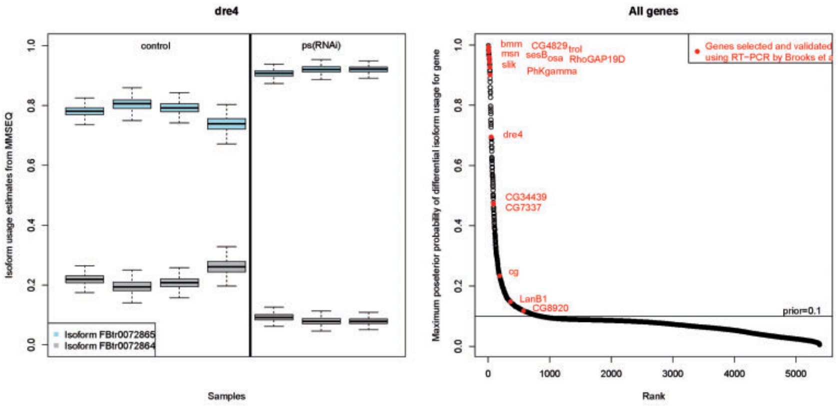

Fig. 7. The left panel shows boxplots of the MMSEQ estimates of the posterior distributions of isoform usage for the two isoforms of the dre4 gene in four control samples and three ps(RNAi) samples. As dre4 has only two isoforms, there is symmetry between the two sets of boxplots. The posterior probability of differential usage for dre4 was calculated to be 0.69 based on a prior probability of 0.1. The right panel shows the maximum posterior probabilities across all isoforms of each gene for all genes. Genes are coloured in red if they were selected and identified as differentially spliced by Brooks *et al.* (2011). The prior probability that the differential isoform usage model is true was set to 0.1 (horizontal line)

using various heuristics and Fisher's exact tests of genome alignment counts. We cannot assess the false discovery rate exactly because the truth is known only for a small subset of genes that were validated by RT-PCR (all of which tested positive). We can, however, determine the posterior probabilities for the set of genes that were validated. The right panel of Figure 7 shows that the genes cluster on the vertical high-probability section of the plot. Genes bmm, CG4829, trol, msn, sesB, osa, RhoGAP19D, slik and PhKgamma all had maximum posterior probabilities greater than 0.9.

Having observed a marked enrichment of validated genes within the small group for which we found compelling evidence of differential isoform usage, we conclude that our method has reasonable power and specificity for detecting differential regulation of isoform usage by modelling posterior summaries of probit-transformed isoform usage proportions. By way of comparison, a standard Cuffdiff 2 splicing analysis of the same dataset declared only three of the validated genes as 'significant': trol, CG4829 and bmm. The minimum (over transcription start sites) unadjusted *P*-values ranged from 0.00005 to 0.53085 for the validated genes while LamB1 and CG8920 were given a 'NOTEST' result by Cuffdiff 2. Supplementary Figure S5 shows that our approach is much more sensitive at recovering the validated genes than Cuffdiff 2.

5 DISCUSSION

The work we have presented can be applied to a wide range of problems in the study of gene expression thanks to two important features. First, due to the linear mixed regression framework, we are able to model many different patterns of expression. We have shown, for example, how the framework can be used to detect imprinting in mice. Second, by using posterior summaries of expression as the outcomes in our regression models rather than alignment counts, we are able to apply our methods in cases where counts are not readily available, such as haplotype-specific and isoform-specific analyses. We have described earlier in the text how transformation of expression proportion variables to the probit scale allows us to detect differential isoform usage. Crucially, having demonstrated that posterior uncertainty in expression estimates may be large and heteroscedastic, we have shown that accounting for this uncertainty can improve precision and sensitivity.

Ordinarily, a different method would be applied to comparisons of expression intensities at different levels of granularity. For example, standard gene expression analyses might be based on comparisons of genome-alignment counts, allele-specific expression analyses might compare counts on a heterozygote-by-heterozygote basis and splicing analyses might compare counts on a splice junction by splice junction basis. It is difficult to integrate the results of such varied kinds of analyses to make inference. Returning to the mouse dataset mentioned earlier in the text, we have shown how our method can be used to perform a differential expression analysis between the two strains in the (pure-strain) parents together with a differential expression analysis between the two strains in the first-generation crosses (i.e. between haplotypes). Our model can straightforwardly accommodate both types of analysis, producing comparable posterior probabilities. More generally, our model selection method can be applied universally to any arbitrary collection of features through aggregation of joint posterior distributions. Analyses could be performed at the gene, haplo-gene, isoform, haplo-isoform, exon and 5′-untranslated region (UTR) level, to name a few, applying a single method of inference across comparison types to an RNA-seq dataset. In the latter example, it would be possible to detect changes in 5′ UTR usage between treatment groups simply by collapsing isoforms sharing the same 5′ UTR start sites. Finally, we have shown how polytomous classification can be performed effectively when more than two models are postulated for the data, which is often the case in studies of gene expression.

ACKNOWLEDGEMENTS

The authors thank Susan Shen (Washington University School of Medicine in St Louis) for providing the list of imprinted genes in mice and Edward Morrissey (University of Cambridge) for helpful discussion on model selection.

Funding: E.T. and S.T. were funded by Cancer Research UK grant C14303/A10825 and E.T. was further funded by the Cambridge Biomedical Research Centre. W.J.A. was funded by UK BBSRC grant BB/E020372/1 and a Team Grant from the Fonds de recherche du Québec—Nature et technologies.

Conflict of Interest: none declared

REFERENCES

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

Anders,S. *et al*. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res*., **22**, 2008–2017.

Brooks,A.N. *et al*. (2011) Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res*., **21**, 193–202.

Carlin,B.P. and Chib,S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Stat. Soc. B Methodol*., **57**, 473–484.

Ferguson,T. (1996) *A Course in Large Sample Theory*. Chapman & Hall/CRC, London, UK.

Garcia-Blanco,M.A. *et al*. (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol*., **22**, 535–546.

Gelman,A. (2004) *Bayesian Data Analysis*. 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

Glaus,P. *et al*. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, **28**, 1721–1728.

Goncalves,A. *et al*. (2012) Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res*., **22**, 2376–2384.

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

Hein,A.-M.K. and Richardson,S. (2006) A powerful method for detecting differentially expressed genes from GeneChip arrays that does not require replicates. *BMC Bioinformatics*, **7**, 353.

Li,Y. and Sasaki,H. (2011) Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. *Cell Res*., **21**, 466–473.

Liu,X. *et al*. (2006) Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, **22**, 2107–2113.

Morison,I.M. *et al*. (2001) The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res*., **29**, 275–276.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*., **11**, R25.

Schulz,R. *et al*. (2008) Wamidex: a web atlas of murine genomic imprinting and differential expression. *Epigenetics*, **3**, 89–96.

Tarazona,S. *et al*. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res*., **21**, 2213–2223.

Trapnell,C. *et al*. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol*., **31**, 46–53.

Turro,E. *et al*. (2011) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*., **12**, R13.

Wang,X. *et al*. (2010) Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *J. Bioinform. Comput. Biol*., **8** (**Suppl. 1**), 177–192.