Bioinformatics, 32, 2016, i147–i155
doi: 10.1093/bioinformatics/btw283
ISMB 2016

OXFORD

Analysis of differential splicing suggests different modes of short-term splicing regulation

Hande Topa[1,2,*] and Antti Honkela[2,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo 00076, Finland and [2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki 00014, Finland

*To whom correspondence should be addressed.

Abstract

Motivation: Alternative splicing is an important mechanism in which the regions of pre-mRNAs are differentially joined in order to form different transcript isoforms. Alternative splicing is involved in the regulation of normal physiological functions but also linked to the development of diseases such as cancer. We analyse differential expression and splicing using RNA-sequencing time series in three different settings: overall gene expression levels, absolute transcript expression levels and relative transcript expression levels.

Results: Using estrogen receptor α signaling response as a model system, our Gaussian process-based test identifies genes with differential splicing and/or differentially expressed transcripts. We discover genes with consistent changes in alternative splicing independent of changes in absolute expression and genes where some transcripts change whereas others stay constant in absolute level. The results suggest classes of genes with different modes of alternative splicing regulation during the experiment.

Availability and Implementation: R and Matlab codes implementing the method are available at https://github.com/PROBIC/diffsplicing. An interactive browser for viewing all model fits is available at http://users.ics.aalto.fi/hande/splicingGP/

Contact: hande.topa@helsinki.fi or antti.honkela@helsinki.fi

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

1 Introduction

Alternative splicing is an important mechanism for increasing proteome complexity in eukaryotes. A great majority of human genes have been found to exhibit alternative splicing with a growing number of annotated spliceforms (Djebali *et al.*, 2012; Sultan *et al.*, 2008; Wang *et al.*, 2008). Changes in splicing are important for cell differentiation (Trapnell *et al.*, 2010). Abnormal splicing has been associated with many diseases, including cancer (Barrett *et al.*, 2015; David and Manley, 2010) as well as neurodegenerative diseases (Cooper-Knock *et al.*, 2012).

Our ability to study and understand alternative splicing is limited by the technology to measure it. The most widely used method is RNA-sequencing (RNA-seq). There are emerging sequencing techniques that enable sequencing of full-length mRNAs (Tilgner *et al.*, 2014), but they do not match the sequencing depth and economy of short-read sequencing technologies which are needed at least to complement the long read sequencing for more reliable quantification of low-abundance genes and transcripts. Analysis of short-read RNA-seq data raises a difficult problem to identify and infer the expression levels of transcript isoforms from reads that are too short to uniquely map to a single isoform. Several methods have been developed to solve this problem (e.g. Glaus *et al.*, 2012; Jiang and Wong, 2009; Li *et al.*, 2010; Trapnell *et al.*, 2010), whereas others have focused on inference of individual alternative splicing events instead of full transcript quantification (Katz *et al.*, 2010). A recent evaluation found that especially the transcript assembly problem is currently too difficult to solve reliably from short-read data (Jänes *et al.*, 2015), and recommended quantification based on known annotated transcripts. Even for this problem there is significant variation between alternative methods (Kanitz *et al.*, 2015; SEQC/MAQC-III Consortium, 2014).

Our study is motivated by the desire to understand the principles of the regulation of splicing. On a large scale, DNA/RNA sequence motifs (Barash *et al.*, 2010; Xiong *et al.*, 2015) and epigenetics (Luco *et al.*, 2010) are important factors in regulation of splicing

(Luco and Misteli, 2011), especially between individuals as well as between tissues. In this article, we study short-term changes in splicing during signaling response within a single tissue or cell line, happening on a time scale of minutes to a few hours. We use estrogen receptor α signaling response on MCF7 breast cancer cell line as our model system here using data from Honkela *et al.* (2015). The first studies performing genome-wide RNA-seq analyses on similar time scale (Äijö *et al.*, 2014; Trapnell *et al.*, 2010) have investigated cell differentiation, while ours is the first to study signaling in this detail.

Methodologically, our work resembles that of Äijö *et al.* (2014), except they only focus on analysis of gene expression from RNA-seq and do not study splicing. A similar dynamical model and test for generic gene expression analysis that does not take the properties of RNA-seq data into account was proposed by Kalaitzis and Lawrence (2011).

2 Materials and methods

2.1 Methods overview

We present a method for ranking the genes and transcripts according to the temporal change they show in their expression levels. In order to identify differential splicing and its underlying dynamics, we model the expression levels in three different settings: overall gene expression level, absolute transcript expression level and relative transcript expression level expressed as a proportion of all transcripts for the same gene.

An outline of our method is shown in Figure 1. Having the RNA-seq time series data, we first start by aligning the RNA-seq reads to the reference transcriptome by Bowtie (Langmead *et al.*, 2009) and then estimate the transcript expression levels by BitSeq (Glaus *et al.*, 2012) separately at each time point. We use BitSeq because it was found to deliver state-of-the-art performance in recent evaluations (Kanitz *et al.*, 2015; SEQC/MAQC-III Consortium, 2014). The same procedure could be applied to other methods that provide reliable uncertainties on quantification results, such as RSEM with posterior sampling (Li and Dewey, 2011). Finally, we model the time series of log-expression or relative expression by two alternative Gaussian process (GP) models, namely *time-dependent* and *time-independent* GPs. In time-dependent GPs, we combine a squared exponential covariance matrix to model the temporal dependency and a diagonal covariance matrix to model the noise whereas in the time-independent GP, we use only the diagonal noise covariance matrix. Finally, we rank the time series by Bayes factors which are computed by the ratio of the marginal likelihoods under alternative GP models.

Our GP-based ranking method utilizes the expression posterior variances from BitSeq in the noise covariance matrices of our GP models, which allows us to set different lower bounds on the noise levels at different time points. A similar approach for modeling the variance from count data has recently been shown to yield higher precision than the naive application of GP models in detecting SNPs (single-nucleotide polymorphisms) selected under natural selection in an experimental evolution study (Topa *et al.*, 2015).

We further introduce a method for improving the variance estimation in situations where the replicates are available only at a small number of time points. More specifically, we perform a simulation with an L-shaped experiment design which consists of three replicates only at the first time point and only one observation at each of the subsequent time points. We then develop a mean-expression-dependent variance model in order to identify the relation between the mean and the variance of the expression levels by using the replicated data available at the first time point and extrapolate this relation to the other time points in order to determine the variance estimates depending on the mean expression level estimates.

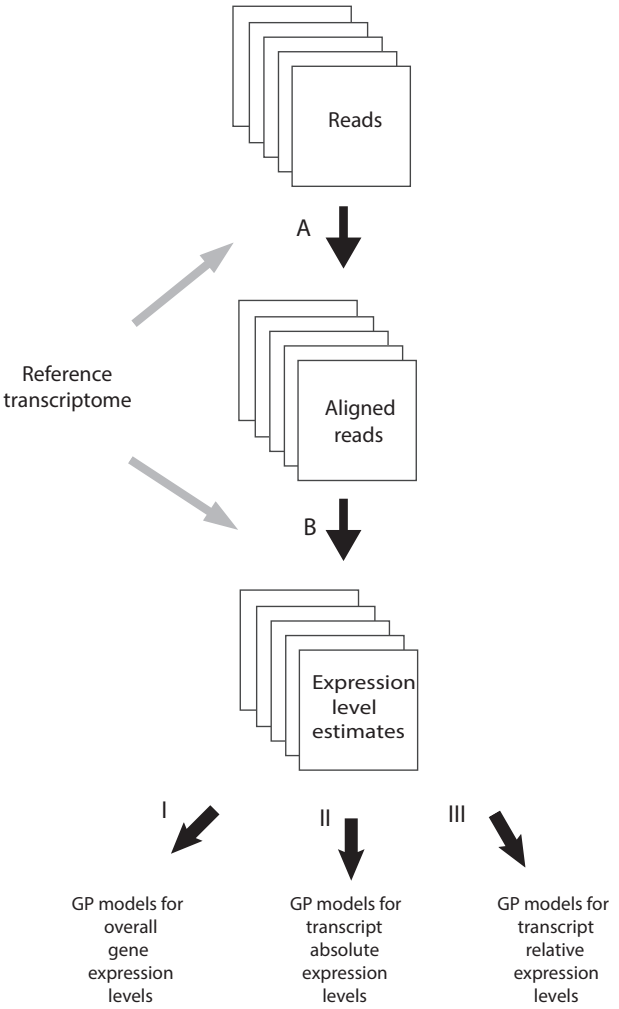


Fig. 1. Methods pipeline: (**A**) The reads are aligned to the reference transcriptome at each time point. (**B**) Expression levels are estimated for each transcript at the given time points. After appropriate normalization and filtering, time series are ranked by the Bayes factors which are computed by dividing the marginal likelihoods under time-dependent and time-independent GP models in three settings: (I) overall gene expression; (II) absolute transcript expression and (III) relative transcript expression.

With a small-scale simulation study, we evaluate the performances of our GP-based ranking method under different scenarios in which the variance information is obtained or used in different ways. We then apply the best-performing variance method in genome-wide real data set and present interesting short-term splicing modes observed in the absolute and relative transcript expression levels. In the following subsections, we will elaborate the intermediate steps in the methods pipeline which have been summarized in Figure 1.

2.2 Gene and transcript expression estimation

As the data were based on an rRNA depletion protocol, we constructed the reference transcriptome by combining cDNA sequences of the protein-coding transcripts, long non-coding RNA and pre-mRNA sequences from gencode.v19 human transcriptome files,

which we downloaded from ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/. Then we ran Bowtie (Langmead *et al.*, 2009) to align the RNA-seq reads to our reference transcriptome according to instructions of the BitSeq package.

Having obtained the aligned reads, we estimated the transcript absolute expression levels by BitSeq (v.0.7.0). BitSeq is a Bayesian method for inferring transcript expression levels from RNA-seq experiments (Glaus *et al.*, 2012) and it returns a posterior distribution over expression levels represented as Markov chain Monte Carlo (MCMC) samples from the distribution.

After obtaining the BitSeq MCMC samples of the expression level estimates for each transcript, we focused to mature mRNAs by removing the pre-mRNAs and renormalizing the reads per kilobase per million reads (RPKM) values of the remaining transcripts with respect to the new number of total mapped reads after exclusion of the reads mapped to the pre-mRNAs. This was necessary to standardize the samples against possible changes in mRNA/pre-mRNA ratio. In addition, we normalized the gene expression levels across time points using the method of Anders and Huber (2010).

2.3 GP modeling of expression time series

A GP is defined as a collection of random variables, any finite subset of which have a joint Gaussian distribution (Rasmussen and Williams, 2006). A GP is specified by its mean function $m(t)$ and covariance function $\Sigma(t,t')$:

$$f(t) \sim GP(m(t), \Sigma(t,t')). \tag{1}$$

Let us assume that we have noisy observations y_t measured at time points t for $t = 1,\ldots,n$ and the noise at time t is denoted by ϵ_t. Then,

$$y_t = f(t) + \epsilon_t. \tag{2}$$

To make the computation simpler, let us subtract the mean from the observations and continue with a zero-mean GP. From now on, y_t will denote the mean-subtracted observations and hence $f(t) \sim GP(0, \Sigma(t,t'))$. Let us combine all the observations in the vector $\mathbf{y}$ such that $\mathbf{y} = [y_1, y_2, \ldots, y_n]$. Assuming that the noise ϵ_t is also distributed with a Gaussian distribution with zero mean and covariance Σ_ϵ, and combining the sampled time points in vector $T = [1,\ldots,n]$ and the test time points in vector T_*, the joint distribution of the training values $\mathbf{y}$ and the test values $f_* = f(T_*)$ can be written as:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma(T,T) + \Sigma_\epsilon(T,T) & \Sigma(T,T_*) \\ \Sigma(T_*,T) & \Sigma(T_*,T_*) \end{bmatrix}\right). \tag{3}$$

Applying the Bayes' theorem, we obtain

$$p(f_*|\mathbf{y}) = \frac{p(\mathbf{y}, f_*)}{p(\mathbf{y})}, \tag{4}$$

where

$$\mathbf{y} \sim N(0, \Sigma(T,T) + \Sigma_\epsilon(T,T)). \tag{5}$$

The computation of Equation 4 leads to:

$$f_*|\mathbf{y} \sim N(\mathbf{m}_*, \Sigma_*), \tag{6}$$

where

$$\mathbf{m}_* = E[f_*|\mathbf{y}] = \Sigma(T_*,T)[\Sigma(T,T) + \Sigma_\epsilon(T,T)]^{-1}\mathbf{y} \tag{7}$$

and

$$\Sigma_* = \Sigma(T_*,T_*) - \Sigma(T_*,T)[\Sigma(T,T) + \Sigma_\epsilon(T,T)]^{-1}\Sigma(T,T_*). \tag{8}$$

The covariance function $\Sigma(t,t')$ of the GP determines the shape of the model, and for estimation purposes it can be constructed based on the assumptions of the underlying model. Squared exponential covariance (Σ_{SE}) is one of the commonly used covariance functions which is suitable for modeling smooth temporal changes with its two parameters: the length scale, ℓ, and the variance, σ_f^2. Each element of the matrix Σ_{SE} can be computed as

$$\Sigma_{SE}(t,t') = \sigma_f^2 e^{-\frac{(t-t')^2}{2\ell^2}}. \tag{9}$$

As demonstrated in Topa *et al.* (2015), the performance of the GP-based ranking methods can be improved by incorporating the available variance information into the GP models. For this reason, we modify the noise covariance matrix such that the variances given in the diagonal have lower bounds which are determined by the variances estimated at each time point separately:

$$\Sigma_\epsilon = \mathrm{diag}(\sigma_N^2 + s_1^2, \ldots, \sigma_N^2 + s_n^2). \tag{10}$$

Σ_ϵ resembles the white noise covariance $\sigma_N^2 I$ except for the fact that the variances are not identical at each time point, being restricted by a lower bound. Note that the only parameter of Σ_ϵ is σ_N^2 since the variances s_t^2 are considered fixed for $t = 1,\ldots,n$.

The log marginal likelihood of the GP model can be written as:

$$\ln p(\mathbf{y}|T) = -\frac{1}{2}\mathbf{y}^T{\Sigma_{\mathbf{obs}}}^{-1}\mathbf{y} - \frac{1}{2}\ln|\Sigma_{\mathbf{obs}}| - \frac{n}{2}\ln 2\pi, \tag{11}$$

where $\Sigma_{\mathbf{obs}} = \Sigma(T,T) + \Sigma_\epsilon(T,T)$. We estimate the parameters of the covariance matrices by maximizing the log marginal likelihoods by using the *gptk* R package which applies scaled conjugate gradient method (Kalaitzis and Lawrence, 2011). In order to prevent the algorithm from getting stuck in a local maximum, we try out different initialization points on the likelihood surface.

2.4 Ranking by Bayes factors

For ranking the genes and transcripts according to their temporal activity levels, we model the expression time series with two GP models, one time-dependent and the other time-independent. While time-independent model has only one noise covariance matrix Σ_ϵ, time-dependent model additionally involves Σ_{SE} in order to capture the smooth temporal behavior. Then, the log marginal likelihoods of the models can be compared with Bayes factors, which are computed by their ratios under alternative models where the log marginal likelihoods can be approximated by setting the parameters to their maximum likelihood estimates instead of integrating them out, which would be intractable in our case. Therefore, we calculate the Bayes factor (K) as follows:

$$K = \frac{P(\mathbf{y}|\widehat{\theta}_1, \text{`time - dependent model'})}{P(\mathbf{y}|\widehat{\theta}_0, \text{`time - independent model'})}, \tag{12}$$

where $\widehat{\theta}_0$ and $\widehat{\theta}_1$ contain the maximum likelihood estimates of the parameters in the corresponding models. According to Jeffrey's scale, log Bayes factor of at least 3 is interpreted as strong evidence in favor of our 'time-dependent' model (Jeffreys, 1961).

2.5 Application of the methods in three different settings

Assuming we have M transcripts whose expression levels have been estimated at n time points, let us denote the kth MCMC sample from the expression level estimates (measured in RPKM) of transcript m at time t by θ_{mt}^k, for $t = 1,\ldots,n$, $m = 1,\ldots,M$ and $k = 1,\ldots,500$. Here we will explain how we determine the

observation vector $\mathbf{y}$ and the fixed variances $(s_1^2, \ldots, s_n^2)$ which we incorporated into the noise covariance matrix Σ_ϵ in our GP models in three different settings:

2.5.1 Gene-level

We compute the overall gene expression levels by summing up the expression levels of the transcripts originated from the same gene, and we calculate their means and variances as following:

$$y_{jt,\text{gen}} = \mathrm{E}_k\left(\log\left(\sum_{m\in I_j}\theta_{mt}^k\right)\right), \tag{13}$$

where I_j is the set of the indices of the transcripts which belong to gene j.

$$s_{jt,\text{gen}}^2 = \max\left(s_{jt,\text{gen}}^{2^{\text{bitseq}}}, s_{jt,\text{gen}}^{2^{\text{modeled}}}\right), \tag{14}$$

where

$$s_{jt,\text{gen}}^{2^{\text{bitseq}}} = \mathrm{Var}_k\left(\log\left(\sum_{m\in I_j}\theta_{mt}^k\right)\right) \tag{15}$$

and modeled variances $(s_{jt,\text{gen}}^{2^{\text{modeled}}})$ are obtained by a mean-dependent variance model which will be explained in Section 2.6.

2.5.2 Absolute-transcript-level

Note that in order to remove the noise that could arise from lowly expressed transcripts, we filtered out the transcripts which do not have at least 1 RPKM expression level at two consecutive time points. Subsequent transcript-level analyses, both in absolute and relative level, were performed by keeping these transcripts out. Then we computed the means and the variances for the absolute transcript expression levels as:

$$y_{mt,\text{abs}} = \mathrm{E}_k(\log(\theta_{mt}^k)), \tag{16}$$

$$s_{mt,\text{abs}}^2 = \max\left(s_{mt,\text{abs}}^{2^{\text{bitseq}}}, s_{mt,\text{abs}}^{2^{\text{modeled}}}\right), \tag{17}$$

where

$$s_{mt,\text{abs}}^{2^{\text{bitseq}}} = \mathrm{Var}_k(\log(\theta_{mt}^k)) \tag{18}$$

and modeled variances $(s_{mt,\text{abs}}^{2^{\text{modeled}}})$ are obtained by a mean-dependent variance model which will be explained in Section 2.6.

2.5.3 Relative-transcript-level

We computed the relative expression levels of the transcripts by dividing their absolute expressions to the overall gene expression levels:

$$y_{mt,\text{rel}} = \mathrm{E}_k\left(\frac{\theta_{mt}^k}{\sum_{m\in I_j}\theta_{mt}^k}\right), \tag{19}$$

and

$$s_{mt,\text{rel}}^2 = \max(s_{mt,\text{rel}}^{2^{\text{bitseq}}}, s_{mt,\text{rel}}^{2^{\text{modeled}}}), \tag{20}$$

where

$$s_{mt,\text{rel}}^{2^{\text{bitseq}}} = \mathrm{Var}_k\left(\frac{\theta_{mt}^k}{\sum_{m\in I_j}\theta_{mt}^k}\right) \tag{21}$$

and modeled variances for transcript relative expression levels $(s_{mt,\text{rel}}^{2^{\text{modeled}}})$ are obtained by Taylor approximation using the modeled variances of logged gene and logged absolute transcript expression levels:

$$s_{mt,\text{rel}}^{2^{\text{modeled}}} = (s_{mt,\text{abs}}^2 + s_{jt,\text{gen}}^2)(y_{mt,\text{rel}})^2. \tag{22}$$

2.6 Modeling the mean-dependent variance

In this section, we will explain how we model the mean-dependent variances by utilizing the MCMC samples generated by BitSeq for each of the replicates available at one time point. Our variance model resembles that of BitSeq Stage 2 (Glaus *et al.*, 2012) except for the fact that we have only one condition and we assume the mean expression levels are fixed. A similar approach is also used by DESeq (Anders and Huber, 2010). Let us assume that at a time point we have R replicates, each of which can be estimated by the mean of the MCMC samples generated by BitSeq. We start by dividing the genes into groups of ≈500 such that each group contains the genes with similar mean expression levels. Let us denote the expression level (log RPKM) of the rth replicate of the jth gene in the gth group by $y_{g,j}^{(r)}$, and the mean expression level by $\mu_{g,j}$, which is calculated as

$$\mu_{g,j} = \mathrm{E}_r(y_{g,j}^{(r)}). \tag{23}$$

Let us also assume that $y_{g,j}^{(r)}$ follows a normal distribution with mean $\mu_{g,j}$ and variance $\frac{1}{\lambda_{g,j}}$:

$$y_{g,j}^{(r)} \sim \mathrm{Norm}\left(\mu_{g,j}, \frac{1}{\lambda_{g,j}}\right), \tag{24}$$

where

$$\lambda_{g,j} \sim \mathrm{Gamma}(\alpha_g, \beta_g) \tag{25}$$

and

$$P(\alpha_g, \beta_g) \sim \mathrm{Uni}(0, \infty). \tag{26}$$

Setting $\mu_{g,j}$ fixed to the mean of the MCMC samples over replicates, we apply a Metropolis-Hastings algorithm to estimate the hyperparameters α_g and β_g for each gene group g. Then we estimate the modeled variance $s_{j^*}^{2^{\text{modeled}}}$ for any given expression level y_{j^*} by Lowess regression which is fitted by smoothing the estimated group variances $(\widehat{\frac{1}{\lambda_g}})$ $(=\frac{\widehat{\beta_g}}{\alpha_g})$ across group means.

The details about the estimation of the hyperparameters with Metropolis-Hastings algorithm can be found in 'Supplementary text'.

2.7 Evaluation of the variance estimation and feature transformation methods with synthetic data

Although high-throughput sequencing technologies have become less costly during the last decade, the trade-off between the cost and the number of replicates still remains as an important factor which needs to be handled with caution. Especially in time series experiments, having replicated measurements at each and every time point could still be very costly.

Here, we evaluate our method under different experiment designs with different numbers of replicates by developing appropriate variance estimation methods for each design.

For this aim, we simulated small-scale RNA-seq time series data and compared the performances of different variance estimation methods in GP models when replicates are available only at some time points or are not available at all. We simulated RNA-seq reads

at 10 time points ($t \in \{1, \ldots, 10\}$) for 15 530 transcripts originating from 3811 genes in chromosome 1 in the transcriptome Homo_sapiens.GRCh37.73. Expression levels of 384 ($\approx 10\%$) genes are changing in time while the rest are constant except for noise. Similarly, 2868 ($\approx 18\%$) and 1530 ($\approx 10\%$) of the transcripts have been generated from a time-dependent model in absolute and relative expression levels, respectively. As RNA-seq data is generally known to follow a negative binomial distribution (Robinson *et al.*, 2010), we generated three replicates at each time point from a negative binomial distribution in which the variance (σ^2) depends on the mean (μ) and the overdispersion parameter (ϕ) with the function $\sigma^2 = \mu + \phi^2\mu^2$. We simulated three sets of experiments with overdispersion parameter (ϕ) set to 0.05, 0.1 and 0.2.

We compare average precision (AP) values of the methods in which the variances which are incorporated into the noise covariance matrix of the GP models are estimated in different ways. We can list the variance estimation methods as following:

- *unrep_naive*: Standard GP regression which does not incorporate the variance information into the noise covariance matrix. In other words, the noise covariance matrix in Equation 10 does not include any fixed variances s_t^2.
- *n-rep_naive*: Standard GP regression which does not incorporate the variance information into the noise covariance matrix. However, there are n replicates available at all time points.
- *unrep_bitseq*: Only one observation is available at each time point. The means and the variances of the expression level estimates are computed by using the BitSeq MCMC samples.
- *n-rep_bitseq*: The ideal case in which n replicates are available at all time points. BitSeq variances are computed separately for each replicate and are included in the noise covariance matrix.
- *unrep_modeled*: There are three replicates only at the first time point and only one observation at the other time points. At the first time point, genes are divided into groups with similar mean expression levels and mean-dependent variances are estimated

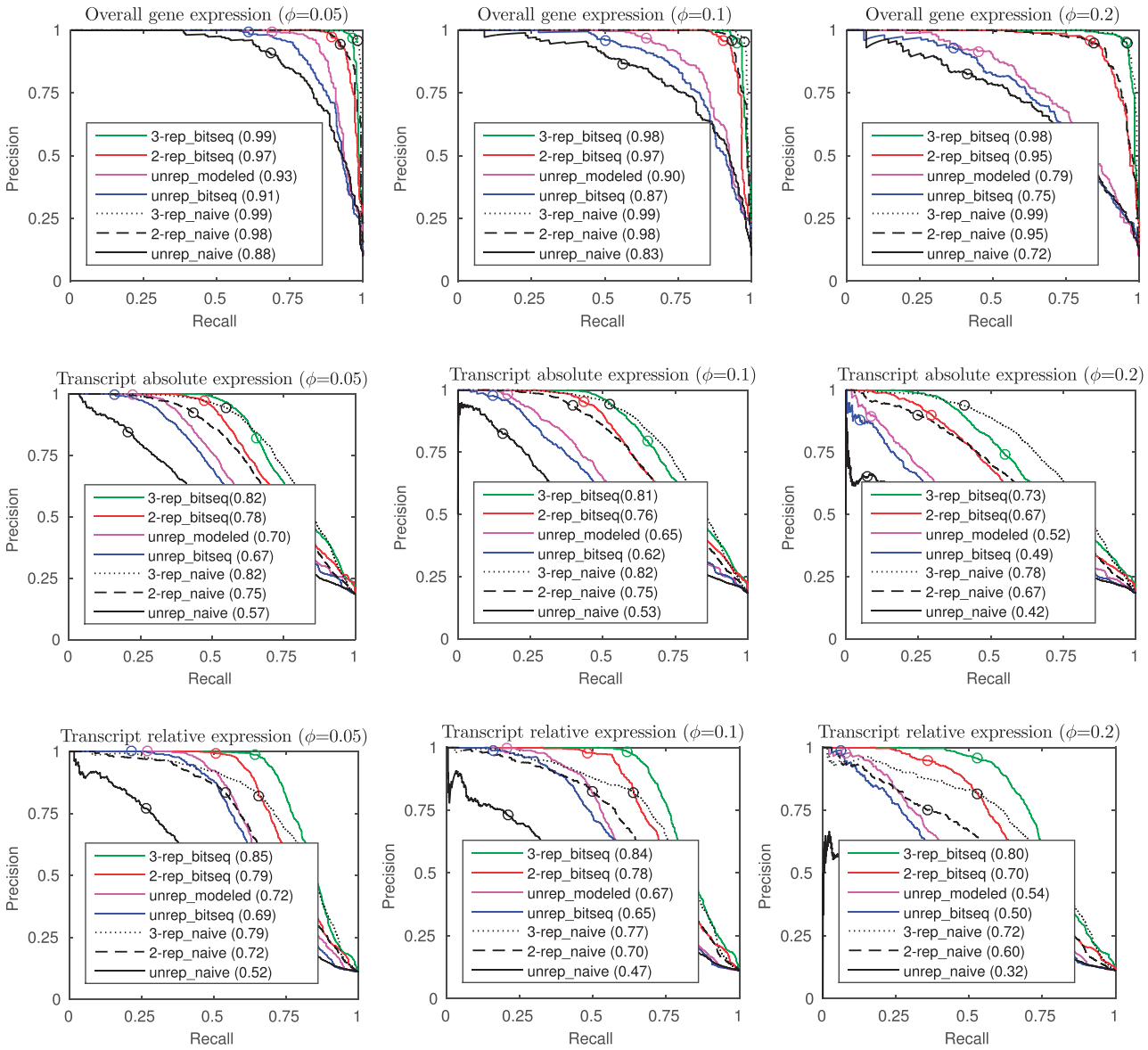


Fig. 2. Precision–recall curves for the GPs with different variance estimation methods and overdispersion parameters (ϕ). The numbers in the legend denote APs of the methods (equivalent to area under the curve). The circles indicate the cut-off $\log(BF) > 3$. The low precision values obscured by the legend correspond to high false discovery rate (FDR) that would not be used in practice.

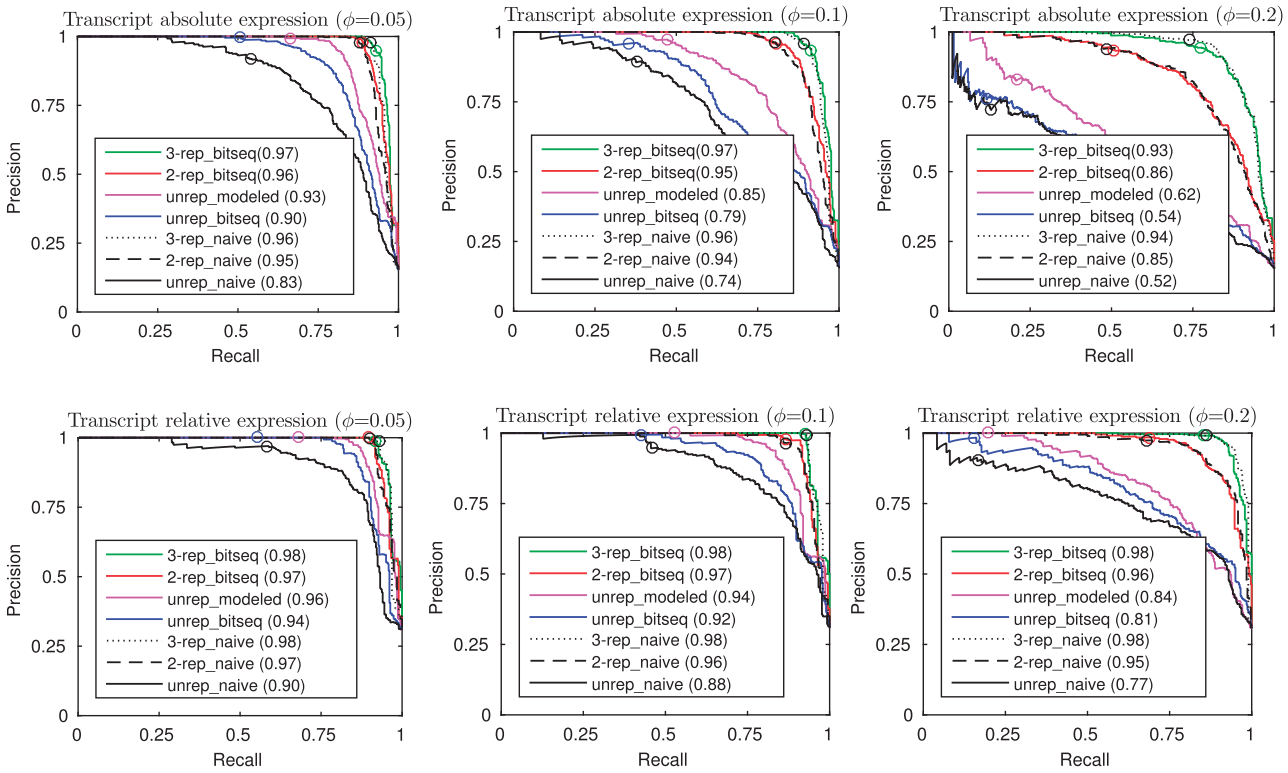


Fig. 3. Precision–recall curves for the GPs with different variance estimation methods and overdispersion parameters (ϕ) for the highly expressed (mean log-RPKM $\geq$ 4) transcripts. The numbers in the legend denote APs of the methods (equivalent to area under the curve). The circles indicate the cut-off $\log(BF) > 3$.

for each group. Then, the variances for the gene and transcript expression levels at the unreplicated time points are modeled by smoothing the group variances as described in Section 2.6. We use the modeled variances at the unreplicated time points if they are larger than the BitSeq variances, and we use the BitSeq variances for each replicate at the first time point.

Additionally, we compute the BitSeq variances for the relative transcript expression levels after applying the following transformations:

- *Isometric log ratio transformation (ILRT)*: It is a popular transformation which is used for transforming compositional data into linearly independent components (Aitchison and Egozcue, 2005; Egozcue *et al.*, 2003). ILRT for a set of m proportions $\{p_1, p_2, \ldots, p_m\}$ is applied by taking component wise logarithms and subtracting the constant $\frac{1}{m}\sum_k \log(p_k)$ from each log-proportion component. This results in the values $q_i = \log(p_i) - \frac{1}{m}\sum_{k=1}^{m} \log(p_k)$ where $\sum_k \log(q_k) = 0$.
- *Isometric ratio transformation(IRT)*: Similar to the above transformation, but without taking the logarithm, that is, $q_i = \frac{p_i}{(\prod_{k=1}^{m} p_k)^{\frac{1}{m}}}$.

Table 1 Numbers of nonDE and DE genes which have at least one transcript belonging to the corresponding absolute-relative (abs-rel) transcript groups

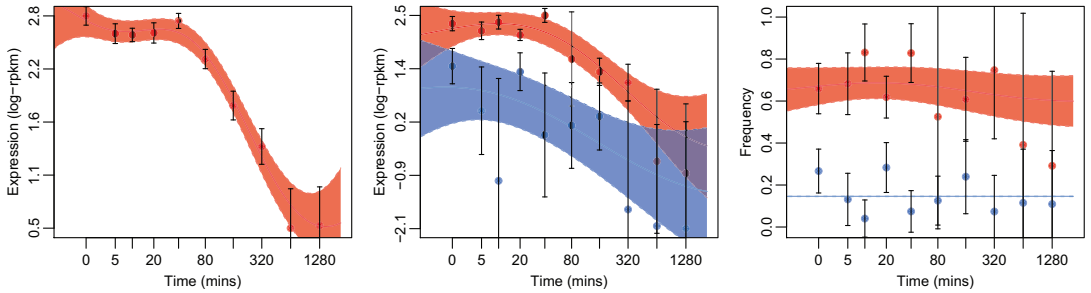
		Gene		
		NonDE	DE	Sum
Transcript Abs-rel	**DE-DE**	336	88	424
	NonDE-DE	152	12	164
	DE-nonDE	1014	700	1714
	NonDE-nonDE	16 511	449	16 960
	Sum	18 013	1249	19 262

The values in the table have been calculated by excluding the single-transcript genes, and only expressed transcripts have been taken into account, i.e. transcripts which had at least 1 RPKM expression level at two consecutive time points.

3 Results and Discussion

3.1 Comparison of variance estimation methods with simulated data

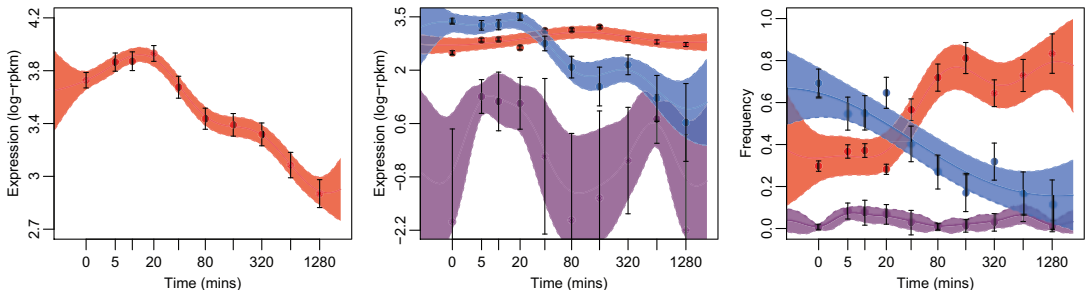
Having simulated the RNA-seq data, we estimated the mean expression levels and variances from the samples generated by BitSeq separately for each replicate at each time point. We evaluated our GP-based ranking method with different variance estimation methods under the scenario where the replicates are not available at all time points. As can be seen in Figure 2, using BitSeq variances in the GP models in unreplicated scenario yields a higher AP than the naive application of GP models without BitSeq variances. An L-shaped design with three replicates at the first time point and the mean-dependent variance model increase the precision of the methods further. In this model, we use the BitSeq samples of these replicates for modeling the mean-dependent variances and we propagate the variances to the rest of the time series, and use these modeled variances if they are larger than the BitSeq variances of the unreplicated measurements. Comparison of the precision recall curves in Figure 2 indicates that this approach leads to a higher AP for all settings. We also observed that the modeled variances become more helpful for highly expressed transcripts when overdispersion increases as can be seen in Figure 3, in which the precision and recall were computed by considering only the transcripts with mean log expression of at least 4 log-RPKM. The figures also show the conventional $\log(BF) > 3$ cut-off. This highlights the fact that the naive model can be very anti-conservative, leading to a large number of false positives.



(a) Gene expression levels of gene *GRHL3*. log-BF= 9.86

(b) Absolute transcript expression levels of gene *GRHL3*. log-BFs:
GRHL3-008(blue): 1.26
GRHL3-201(red): 6.05

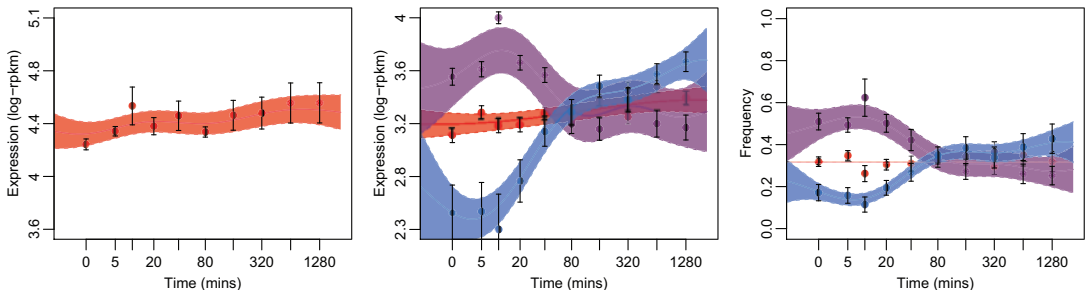
(c) Relative transcript expression levels of gene *GRHL3*. log-BFs:
GRHL3-008(blue): 0
GRHL3-201(red): 0.50



(d) Gene expression levels of gene *RHOQ*. log-BF= 8.16

(e) Absolute transcript expression levels of gene *RHOQ*. log-BFs:
RHOQ-001(red): 0.56
RHOQ-006(purple): 1.07
RHOQ-007(blue): 4.80

(f) Relative transcript expression levels of gene *RHOQ*. log-BFs:
RHOQ-001(red): 3.66
RHOQ-006(purple): 0.22
RHOQ-007(blue): 4.10



(g) Gene expression levels of gene *MTCH2*. log-BF= 0.27

(h) Absolute transcript expression levels of gene *MTCH2*. log-BFs:
MTCH2-001(red): 0.98
MTCH2-002(blue): 7.83
MTCH2-201(purple): 1.64

(i) Relative transcript expression levels of gene *MTCH2*. log-BFs:
MTCH2-001(red): 0
MTCH2-002(blue): 6.56
MTCH2-201(purple): 3.64

Fig. 4. GP profiles of three example genes and their transcripts. Error bars indicate ±2 fixed-standard-deviation (square root of the fixed variances) intervals and the colored regions indicate the ±2 standard-deviation confidence regions for the predicted GP models. The transcripts are shown in the same color in absolute (b,e,h) and relative (c,f,i) transcript-expression-level plots. Prior to GP modeling, time points were transformed by $\log(t+5)$ transformation.

Figure 2 also shows results for fully two-way and three-way replicated time series. Introducing the second replicate at each time point improves the performance very significantly while the marginal benefit from the third replicate is much smaller. Introducing the BitSeq variances increases the accuracy significantly for transcript-level analyses, especially for transcript relative expression.

3.2 Comparison of feature transformation methods on relative transcript expression levels with synthetic data

Transcript relative expression levels represent a special type of data called compositional data because they always sum to 1 for each gene. This property generates an artificial negative correlation between the transcripts which can make analysis more challenging. Several

transformation techniques have been recommended in the literature for this task. ILRT is one of the most commonly used transformations for breaking the linear dependency between the proportions.

We applied ILRT as well as its unlogged version (IRT) to the relative transcript expression levels. Calculating the BitSeq variances for the transformed values, we compared the performance of our method with the performance when no transformation is applied. As can be seen in Supplementary Figure 1, we observed that the feature transformations were not useful for increasing the performance of our method. Therefore, we did not apply any transformation to the relative expression levels in real data analysis. The reason for their poor performance may be that the new transformation was poorly compatible with our GP model and variance models.

3.3 Differential splicing in ER-α signaling response

Encouraged by the good performance of the modeled variances and especially their good control of false positives, we apply that method for real data using the estrogen receptor-α (ER-α) signaling as a model system using RNA-seq time series data from Honkela *et al.* (2015) (accession GSE62789 in GEO).

The data set contains RNA-seq data obtained from MCF7 breast cancer cell lines treated with estradiol at 10 different time points (0, 5, 10, 20, 40, 80, 160, 320, 640 and 1280 min). We treat the first three time points as if they were the replicates measured at the same time point to fit the variance model. This approach is reasonable because the system starts from a quiescent steady state and only very little new transcription is expected to occur during the first 10 min.

We build our reference transcriptome from gencode.v19 by combining the protein-coding cDNA sequences, long non-coding RNA sequences and pre-mRNA sequences. The reference transcriptome contains 34 608 genes and their 119 207 transcripts. We exclude 15 346 single-transcript genes from our transcript-level analyses.

The numbers of non-differentially expressed (nonDE) and DE genes which have at least one transcript belonging to the corresponding abs-rel (absolute-relative) transcript groups (DE-DE, nonDE-DE, DE-nonDE, nonDE-nonDE) are given in Table 1. We assumed that a transcript is expressed only if it has at least 1 RPKM expression level at two consecutive time points, and we ignored the unexpressed transcripts which do not satisfy this criterion in order to avoid the noise originated from lowly expressed transcripts. We call genes and transcripts DE in absolute expression levels if the GP-smoothed fold change (the ratio of the maximum GP mean expression to the minimum GP mean expression) is at least 1.5, and the log-Bayes factor is larger than 3. We set the same thresholds for the relative transcript expression levels except for the fold change which we replaced with the condition that the difference between the GP-smoothed maximum and minimum proportions be larger than 0.1.

According to the table, $\sim$11% of genes undergo either differential splicing or have DE transcripts. There is a significant number of genes which are not called DE or differentially spliced, but have at least one DE transcript. The model fits for these genes can be viewed in the online model browser, which shows that many of these examples are probably due to lower sensitivity of relative expression change detection. There are also many cases where the absolute expression signal of a single transcript appears very clean, but the other transcripts mess up the gene and relative expression signals making them appear more like noise.

3.4 Evidence for different modes of splicing regulation

The results in Table 1 suggest that different genes employ different strategies for the regulation of splicing. This is confirmed by visual observation of the model fits, available in the online model browser. Illustrative examples of genes from the different classes are shown in Figure 4.

The gene *GRHL3* in the top row shows an example of a gene where the relative proportions of the different transcripts remain constant throughout the experiment even though the expression of the gene changes. This appears to be a relatively common case. Even using stringent criteria for no change in relative expression ($\log{-}\mathrm{BF} < 1$) almost 450 genes follow this pattern.

The *RHOQ* and *MTCH2* genes in the middle and bottom rows show two slightly different interesting examples where the absolute expression level of one of the transcripts remains constant while the others change, suggesting highly sophisticated regulation of the individual transcript expression levels. These are both examples of the class with both differential relative and absolute expression which covers more than 400 genes. The behavior of these genes is extremely diverse and hard to categorize further, but by visual inspection one can find many more examples where the gene and some of its transcripts are changing whereas some expressed transcripts remain constant, such as *ARL2BP*, *RB1CC1*, *HNRNPD*, *TBCEL*, *OSMR*, *ESR1*, *ADCY1*, *PMPCB*, *AP006222.2*, *EPS8*, *RAVER2* and *P4HA2*.

4 Conclusion

In this article, we have presented a method for detecting temporal changes in gene expression and splicing as well as transcript expression patterns that successfully incorporates uncertainty arising from RNA-seq quantification in the analysis.

We evaluated the performance of our method under different experiment designs in a simulation study. Our results again confirm the importance of replication in genomic analyses. In our clean synthetic data adding a second replicate gives a dramatic boost but improvements from having more than two replicates of the entire time course are modest. Things may of course not be as simple for real data where a third replicate could at least be very useful for detecting corrupted and otherwise significantly diverging measurements that could otherwise decrease the power.

We compared approaches based on noise variances inferred only from the data and using posterior variance from BitSeq as a lower bound on the noise for the GP. The BitSeq variances were found to be very useful in unreplicated case as well as for transcript-level analyses.

We also experimented with a computational method for modeling variances to fill in missing replicates with information propagated from a single replicated time point. The results indicate that this method can increase the accuracy of the analyses. However, in the case of transcript relative expression there are still unsolved technical challenges that may have a role in the performance. As the variance of the relative transcript expression levels depends on the variances of the overall gene expression levels and the absolute transcript expression levels as well as the covariance between them, which we did not take into account here, it is not straightforward to model the variance for the relative transcript expression levels and it would require more powerful methods which would be suitable for compositional data.

Application of our method to the analysis of splicing patterns during estrogen receptor signaling response in a human breast cancer cell line lead to the discovery of classes of genes with different kinds of splicing and expression changes. We found several genes for which the relative expression levels of different transcripts remain approximately constant whereas the total gene expression level changes and for which the relative expression levels change apparently independently of the total expression level, consistent with a

model of independent regulation of total expression level and relative splicing levels. There appears however to also be a potentially more interesting set of genes where the absolute expression of some transcripts remains constant whereas the expression level of others changes. These examples suggest a link between regulation of gene expression and splicing, but further research with careful controls is needed to assess how common this phenomenon is. The finding nevertheless suggests that alternative splicing analyses need to combine both absolute and relative transcript expression analyses.

Acknowledgements

We thank Peter Glaus for providing his Python code for creating the FASTA files in the simulation of RNA-seq reads. We also acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

H.T. was supported by Alfred Kordelin Foundation, and A.H. was supported by the Academy of Finland [259440, 251170].

Conflict of Interest: none declared.

References

Äijö,T. *et al*. (2014) Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics*, **30**, i113–i120.

Aitchison,J.J. and Egozcue,J. (2005) Compositional data analysis: where are we and where should we be heading? *Math. Geol.*, **37**, 829–850.

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

Barash,Y. *et al*. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.

Barrett,C.L. *et al*. (2015) Systematic transcriptome analysis reveals tumor-specific isoforms for ovarian cancer diagnosis and therapy. *Proc. Natl. Acad. Sci. USA*, **112**, E3050–E3057.

Cooper-Knock,J. *et al*. (2012) Gene expression profiling in human neurodegenerative disease. *Nat. Rev. Neurol.*, **8**, 518–530.

David,C.J. and Manley,J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.*, **24**, 2343–2364.

Djebali,S. *et al*. (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Egozcue,J. *et al*. (2003) Isometric logratio transformations for compositional data analysis. *Math. Geol.*, **35**, 279–300.

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

Honkela,A. *et al*. (2015) Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proc. Natl. Acad. Sci. USA*, **112**, 13115–13120.

Jänes,J. *et al*. (2015) A comparative study of RNA-seq analysis strategies. *Brief Bioinform.*, **16**, 932–940.

Jeffreys,H. (1961). *Theory of Probability*, 3rd edn. *Oxford Classic Texts in the Physical Sciences*. Oxford University Press, Oxford.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Kalaitzis,A.A. and Lawrence,N.D. (2011) A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, **12**, 180.

Kanitz,A. *et al*. (2015) Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.*, **16**, 150.

Katz,Y. *et al*. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

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.

Li,B. *et al*. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Luco,R.F. and Misteli,T. (2011) More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr. Opin. Genet. Dev.*, **21**, 366–372.

Luco,R.F. *et al*. (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.

Rasmussen,C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy. *Nat. Biotechnol.*, **32**, 903–914. reproducibility and information content by the Sequencing Quality Control Consortium.

Sultan,M. *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Tilgner,H. *et al*. (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA*, **111**, 9869–9874.

Topa,H. *et al*. (2015) Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, **31**, 1762–1770.

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.

Wang,E.T. *et al*. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Xiong,H.Y. *et al*. (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science*, **347**, 1254806.