BIOINFORMATICS

Vol. 26 ISMB 2010, pages i325–i333
doi:10.1093/bioinformatics/btq200

Model-based detection of alternative splicing signals

Yoseph Barash[1,2,3,*], Benjamin J. Blencowe[1,3,4] and Brendan J. Frey[1,2,3,*]

[1]Banting and Best Department of Medical Research, [2]Department of Electrical and Computer Engineering, [3]Donnelly Centre for Cellular and Biomolecular Research and [4]Department of Molecular Genetics, University of Toronto, ON, Canada

ABSTRACT

Motivation: Transcripts from ∼95% of human multi-exon genes are subject to alternative splicing (AS). The growing interest in AS is propelled by its prominent contribution to transcriptome and proteome complexity and the role of aberrant AS in numerous diseases. Recent technological advances enable thousands of exons to be simultaneously profiled across diverse cell types and cellular conditions, but require accurate identification of condition-specific splicing changes. It is necessary to accurately identify such splicing changes to elucidate the underlying regulatory programs or link the splicing changes to specific diseases.

Results: We present a probabilistic model tailored for high-throughput AS data, where observed isoform levels are explained as combinations of condition-specific AS signals. According to our formulation, given an AS dataset our tasks are to detect common signals in the data and identify the exons relevant to each signal. Our model can incorporate prior knowledge about underlying AS signals, measurement quality and gene expression level effects. Using a large-scale multi-tissue AS dataset, we demonstrate the advantage of our method over standard alternative approaches. In addition, we describe newly found tissue-specific AS signals which were verified experimentally, and discuss associated regulatory features.

Contact: yoseph@psi.utoronto.ca; frey@psi.utoronto.ca

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The proper function of a living cell depends on constant regulation of its biomolecular content, both spatially and temporally. One important regulated process is *splicing*, where exonic segments of the transcribed pre-mRNA are spliced together while intronic segments are removed. In some cases, different exons of the pre-mRNA may be retained, leading to different transcripts called isoforms, a process known as *alternative splicing* (AS). There are several known types of AS, with the most common one being cassette AS (Wang *et al*., 2008), illustrated in Figure 1A. AS plays a critical role in shaping how genetic information is expressed in cells of metazoan species. Moreover, 15–50% of human disease mutations affect splice site selection (Wang and Cooper, 2007). Recent high-throughput sequencing studies estimate that transcripts from ∼95% of multiexon human genes undergo AS, with the majority of AS exons displaying differential expression across different tissues (Pan *et al*., 2008; Wang *et al*., 2008). Such high-throughput studies can be probed to identify condition-specific[1] splicing changes that subsequently can be linked to genetic disease (Scheper *et al*., 2004). Alternatively, groups of exons identified by these studies as exhibiting concerted changes across functionally related conditions (e.g. muscle tissues) can be used to look for a common regulatory program. For example, several works (Castle *et al*., 2008; Fagnani *et al*., 2007; Wang *et al*., 2008) correlated splicing changes with potential binding motifs. Recently, Barash *et al*. (2010) developed a computationally derived regulatory code that includes combinations of motifs and non-motif features such as transcript structure characteristics to directly predict splicing changes from genomic sequence. The focus of the work presented here is to facilitate such downstream analysis by accurately identifying biologically informative condition-specific splicing changes, or AS signals, underlying high-throughput measurements.

High-throughput AS measurements are typically acquired using microarrays or high-throughput sequencing technologies. Accurate quantification of every isoform still poses computational and experimental challenges. Consequently, much of the research involving AS and derived datasets focuses on the simpler task of quantifying splicing changes at the single exon level. Such data can be quantified as the fraction of gene isoforms including an exon, with additional information conveying the overall gene expression level in each condition (Pan *et al*., 2004; Shai *et al*., 2006). This representation is not only easier to derive, but also corresponds well to the hypotheses accounting for regulated AS. The regulatory mechanisms involved include various structural features and *cis* elements in the proximity of the alternative exon, with splicing and transcription forming two distinct networks of regulation (Hartmann and Valcarcel, 2009; Wang and Burge, 2008).

Formally, the measurements from high-throughput AS studies can be represented as a matrix whose entries convey the fraction of inclusion and exclusion isoforms for each of the exons and each of the conditions monitored in the study (Fig. 1B). Similarly, the overall expression levels of the genes containing each exon can be represented by a corresponding matrix. A widely used approach for identifying common patterns in such data is clustering (Eisen *et al*., 1998; Segal *et al*., 2004). Clustering the columns of the matrix derived from Fagnani *et al*. (2007), containing ∼3700 cassette exons profiled across 27 mouse tissues, easily identifies a cluster of all central nervous system (CNS) tissues, evident on the left side of the clustergram in Figure 1C. Similarly, clustering rows of the matrix, which represent *AS profiles* of exons, identifies groups of exons that share a similar AS profile across conditions.

While readily available and easy to apply to AS datasets, standard clustering methods such as hierarchical agglomerative clustering and K-means clustering, suffer from several drawbacks. First, many splicing changes occur in functionally related conditions, such as CNS or muscle tissues. Standard clustering is not easily modified so as to incorporate such prior knowledge. Second, a splicing change

*To whom correspondence should be addressed.

[1]We use the term condition not only for specific cellular conditions but also to denote tissue or cell types, etc.

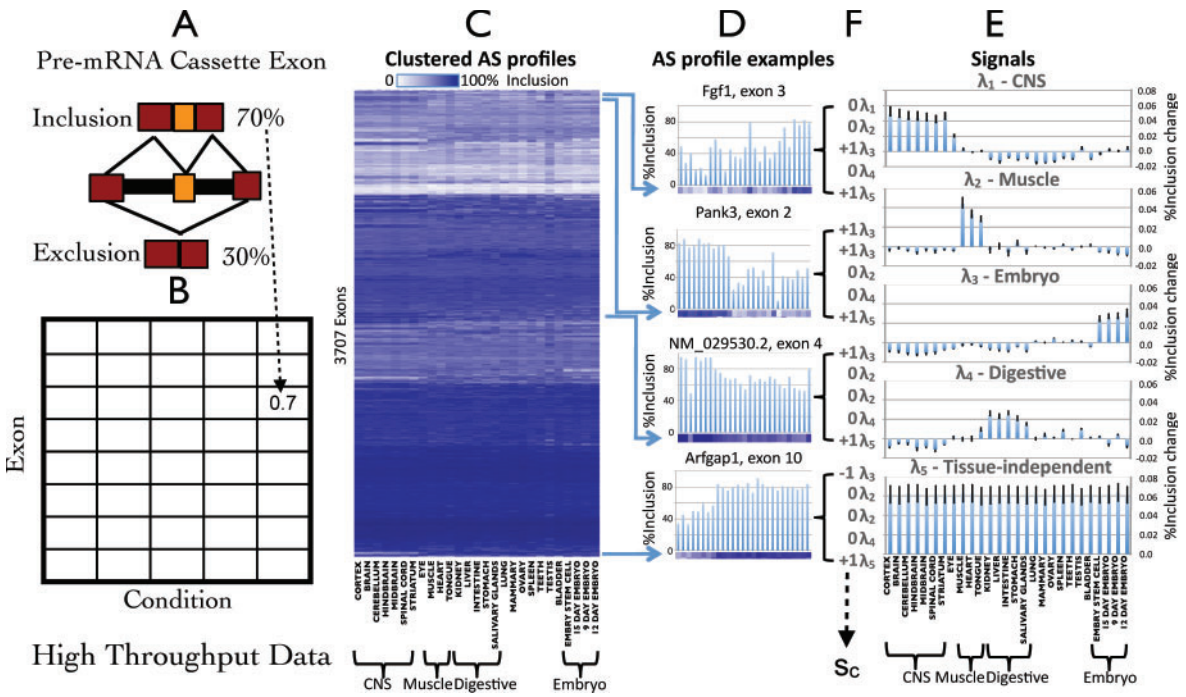

Fig. 1. High-throughput AS data representation and analysis: (**A**) isoforms including and excluding a cassette exon can be quantified using a single number, representing the percent of isoforms that include the exon. A separate number gives the overall gene expression level (not shown here). (**B**) High-throughput AS data as a matrix of percent inclusion values for different exons (rows) under different conditions (columns). (**C**) The same matrix for a real dataset (Fagnani *et al*., 2007), after agglomerative clustering. Inclusion levels are displayed as a heat-map, with the subset of CNS tissues visible on the left. (**D**) Four examples of exons exhibiting condition-specific splicing changes in CNS, muscle and embryo tissues. Arrows show the position of each exon in the clustergram. These relative positions do not convey well the tissue groups in which each of these exons exhibit splicing changes. (**E**) The five underlying AS signals identified by our model and (**F**) how each of these signals is associated with the four exon examples.

in a subset of conditions may involve two types of effects, one being a relative increase of exon inclusion levels and the second being a relative decrease. Common distance measures used for clustering either distinguish between these two effect types or ignore the difference between the two (e.g. correlation and absolute correlation coefficients). In practice, however, it is desirable to be able to distinguish between these two types of effects, but still associate them since they may share common regulatory elements. For instance, splice factors such as Nova, Fox and (n)PTB have been linked to both effects in the same tissues (Ashiya and Grabowski, 1997; Boutz *et al*., 2007; Minovitsky *et al*., 2005; Ule *et al*., 2006). Third, another characteristic of AS datasets not easily incorporated into standard clustering methods, is that exons may exhibit a combination of splicing changes in several functionally related tissue groups, such as muscle and CNS (Zhang *et al*., 2008), but occurrence of regulated splicing changes across cellular conditions is expected to generally be sparse. This expected sparsity is related to the experimental setup, where thousands of exons are monitored but most of these do not exhibit condition-specific profiles. Fourth, previous work has shown that while some exons exhibit a type of on/off splicing profile, others exhibit continuous splicing changes across tissues, and may have different tissue-independent baseline inclusion levels. A closer look at the clustering results (Fig. 1D), illustrates the problems that arise from the inability to easily incorporate these domain characteristics into standard clustering. Exon 2 of Pank3 and exon 10 of Arfgap1 are an example of two exons positioned on opposite ends of the clustergram, despite both exhibiting a profile containing a splicing change in CNS tissues. Exon 2 of Pank3 and exon 3 of NM_029530.2 are positioned far apart since their baseline levels of inclusion are distinctly different, but both exhibit a similar pattern of increased inclusion in CNS. The Pank3 exon also exhibits increased inclusion in muscle tissues, yet it is positioned adjacent to exon 3 of Fgf1, which exhibits an unrelated splicing profile. Switching to different similarity measures [e.g. from the default L_1 norm used here to Pearson's correlation, or mean squared error (MSE)] or between clustering algorithms, may help address some of these problems, but does not offer an overall solution to the issues raised.

Matrix factorization algorithms, such as principal component analysis (PCA), factor analysis (FA) and singular value decomposition (SVD), offer an alternative approach for analyzing high-throughput AS measurements. Following the terminology of SVD previously used for this task (Wang *et al*., 2008), these algorithms are able to identify a set of $C \leq T$ underlying 'eigen-exons' (termed 'components' in PCA and 'factors' in FA), and assign to each exon in the dataset a matching set of values that represent how much each of these eigen-exons contributes to a given exon AS profile. This approach is thus more naturally suited for modeling AS measurements as continuous combination of components, where each component can have either a positive (increased inclusion) or negative (increased exclusion) effect, and with different magnitude. However, these algorithms still lack

some basic elements needed to properly model AS. For example, it is not obvious how to incorporate prior knowledge about the domain (e.g. groups of related experiments) or possible noise in the measurements. Specifically, some measurements are more likely to be noisy because a gene is insignificantly transcribed in a certain tissue, or suffers from low read coverage. In addition, since sparsity is not enforced in many of these algorithms, they can account for an AS profile using small amounts of many eigen-exons, and such contributions are usually meaningless in terms of underlying condition-specific regulation.

We propose a probabilistic model for high-throughput AS measurements that stems from signal processing and FA (Rubin and Thayer, 1982). In this framework, given a dataset, our objective is to identify what are the underlying AS signals that together explain the observed data, and what combination of those make up each of the observed exon AS profiles. Our generative model treats the observed AS profile of an exon as a vector of random variables which is the result of a combination of underlying (hidden) condition-specific AS signals. Each AS signal, just as the eigen-exons in SVD, is a vector across the experimental conditions. However, unlike in standard matrix factorization, the multiplicative factor modulating the contribution of each AS signal to each exon is modeled by an assignment to a random variable that can come from three different distributions: the first distribution corresponds to the signal being 'off' (i.e. contributes nothing to the AS profile), while the other two distributions represent the signal being 'on' or 'reversed', corresponding to the signal contributing to differential inclusion or exclusion, respectively. We include a sparse prior favoring the 'off' state to reflect the fact that most exons monitored in high-throughput experiments are not expected to exhibit condition-specific splicing changes. In addition to the AS signals, our generative model also includes a competing background model. Whether an observed measurement was generated by the signal or the background model is determined by the assignment of a matching binary noise variable, which is generally unknown (i.e. hidden). However, when additional information is available, such as the overall expression level of the exon's gene in that condition, we incorporate it into the model to increase or decrease the posterior belief that specific measurements came from the background model.

We derive an algorithm to efficiently learn the proposed model given a high-throughput AS dataset. Performing inference in the learned model, one can identify which combination of signals make up the AS profile of each exon monitored, or test exons not included in the original study. The result of such analysis is illustrated in Figure 1E and F. Three of the identified signals in Figure 1E correspond to previously known splicing changes in CNS, muscle and embryo tissues, while the last signal corresponds to a tissue-independent signal that captures the (hidden) baseline inclusion level. One tissue-specific signal is a novel signal representing splicing changes in digestive tissues identified by the algorithm. As we discuss later, this signal was not discovered by alternative methods but was supported by a predictive model derived from putative regulatory features followed by experimental testing of the model's predictions (Barash *et al*., 2010). The combinations of these five AS signals used to account for the AS profiles of the four exon examples of Figure 1D are shown in Figure 1F. We see that exon 3 of Fgf1 was correctly identified as including the signal for a change in embryo tissues set to 'on', the three other exons correctly identified to include the CNS tissues signal set to either 'on' or 'reverse', and exon 2 of Pank3 was also found to include the signal for muscle.

In the following section, we provide more indepth details of the model and the algorithm used to learn it. In the results section, we discuss the AS signals identified and experimental verifications of exons exhibiting those signals. Then, we demonstrate the advantages of our model, both quantitatively and qualitatively, over the related SVD method and over manually defined AS signals, two approaches previously used for this task (Castle *et al*., 2008; Fagnani *et al*., 2007; Wang *et al*., 2008). We finish the results section relating the AS signals identified to a compendium of known regulatory features we compiled, briefly reviewing some of the mechanistic insights suggested by our analysis before concluding with a discussion of related work and future directions.

2 METHODS

The input data includes two matrices, one for AS measurements $\{x_t^e\} \in [0,1]$, giving for each exon $e \in \{1,\ldots,E\}$ and condition $t \in \{1,\ldots,T\}$ the estimated fraction of isoforms that include the cassette exon, and the second matrix, $\{v_t^e\} \in \mathcal{R}$ representing the estimated log-abundance of each exon's corresponding gene in each condition. We term the vector $\mathbf{x}^e = x_1^e,\ldots,x_T^e$ the *AS profile* of exon e. In our probabilistic framework, the AS profile is a vector of random variables $\mathbf{x} = x_1,\ldots,x_T$, and an observed AS profile of an exon ($\mathbf{x}^e$) is an instantiation of it. According to our generative model, an observed AS profile is a result of an unknown combination of a set of unknown components, or AS signals $\{\boldsymbol{\lambda}_c\}_{c=1}^C$ where $\boldsymbol{\lambda}_c$ is a vector over the measured conditions $\lambda_{c,1},\ldots,\lambda_{c,T}$, where $\lambda_{c,t} \in \mathcal{R}$. We know from previous studies (Castle *et al*., 2008; Wang *et al*., 2008) that splicing changes across conditions may occur at different levels of magnitudes. Accordingly, the contribution of each signal $\boldsymbol{\lambda}_c$ depends on a multiplicative factor, modeled by a matching random variable $m_c \in \mathcal{R}$. To reflect the fact that each AS signal (e.g. inclusion change in CNS tissues) may contribute to an observed AS profile either positively (increase inclusion) or negatively (increase exclusion), and that most exons surveyed in such high-throughput datasets are not expected to exhibit a condition-specific splicing profile, we use a mixture distribution for the magnitude modulation variable m_c. The class of this mixture distribution is represented by a ternary random variable s_c, and corresponds to three components: $s_c = 0$ which means the signal is absent ($m_c \doteq 0$) , $s_c = +1$ and $s_c = -1$ which imply positive ($m_c > 0$) or negative ($m_c < 0$) contributions of the matching signal $\boldsymbol{\lambda}_c$. To summarize, we have:

$$P(x_t^e, \mathbf{s}^e, \mathbf{m}^e) = P(\mathbf{s}^e) P(\mathbf{m}^e | \mathbf{s}^e) \mathcal{N}(x_t^e; \sum_{c=1}^{c} \lambda_{c,t} m_c^e, \psi_t), \quad (1)$$

where $\mathcal{N}$ denotes a Gaussian with variance ψ_t. To reflect the fact that most exons are not expected to exhibit condition-specific AS profiles, we use a sparse prior where $\forall c\ P(s_c = 0) \gg P(s_c = \pm 1)$. When an AS signal is absent ($s_c = 0$) we have m_c set to zero. For cases where an AS signal is present ($s_c = \pm 1$) we use $P(m_c | s_c) = \mathcal{N}(m_c; v_c, \gamma_c)$ and initialize it with $v_c = \pm 4, \gamma_c = 1$,[2] to avoid situations where $s_c^e \neq 0$ (indicating a change in splicing), but the magnitude of the change is close to zero. We note that while the Gaussian assumption carries the benefits of mathematical tractability for the derivations that follow, it is not ideal for modeling a distribution over a random variable confined to a finite range. However, we stress that the marginal distribution $P(x_t^e)$ as defined above is actually a mixture distribution, with the assignment s_c determining the mixture component.

[2]In FA, changing either v_c or γ_c of m_c prior distribution have the same effect on the model's likelihood as the magnitude of the signal multiplied by m_c is absorbed by the values of the matching $\boldsymbol{\lambda}_c$.

The assignments to those mixture variables in turn determine which exon exhibits each of the inferred AS signal.

Exons surveyed in high-throughput experiments tend to have varying baseline inclusion levels. Indeed, when performing SVD analysis of AS data the first and most prominent signal identified is a tissue-independent signal (Section 3). To model this we incorporate into the model a signal $\boldsymbol{\lambda}_B$ that is forced to be used in the positive sense in all cases (i.e. $s_B^e = +1, \forall e$ with $m_B \sim \mathcal{N}(\nu_B, \gamma_B)$, initialized using $\nu_B = 4, \gamma_B = 1$). This is the fifth AS signal depicted at the bottom of Figure 1E. While the exact value of $\boldsymbol{\lambda}_B$ is updated during learning, the m_B variable guarantees varying magnitudes of this baseline signal to appear in the AS profile of each exon.

Next, we incorporate into the model gene expression levels and possibly other knowledge about the experiments. Intuitively, if a gene is not expressed in a specific cellular condition t then a corresponding entry x_t^e for one of its exons should be ignored. In practice, either biological factors (e.g. low gene expression) or technical ones (e.g. non-specific hybridization or fluctuations in read coverage) usually lead us to ascertain higher or lower confidence to measurements. We assume additional measurement information such as gene expression levels is given as a matrix $\{v_t^e\}$ and generally $v_t^e \in \mathcal{R}$. To utilize this information, our generative framework contains alongside the signal model a competing background or outlier model. A hidden binary variable n_t indicates whether each observation x_t^e was generated from the signal model ($n_t^e = 0$) or from the background model ($n_t^e = 1$). Formally:

$$\begin{aligned} P(\mathbf{x}^e, \mathbf{m}^e, \mathbf{s}^e, \mathbf{v}^e) &= \prod_c P(s_c^e) P(m_c^e | s_c^e) \\ &\cdot \prod_t \Big[P(n_t^e = 1) P(v_t^e \mid n_t^e = 1) \mathcal{N}(x_t^e; \mu_t, \phi_t) \\ &+ P(n_t^e = 0) P(v_t^e \mid n_t^e = 0) \mathcal{N}(x_t^e; \sum_{c=1}^{C} \lambda_{c,t} m_c^e, \psi_t) \Big]. \end{aligned} \tag{2}$$

If prior knowledge about the quality of the AS measurements, given by $P(v_t^e \mid n_t^e)$, is not available, we simply set $P(n_t = 1)$ to a low value, indicating low outlier probability. More details on how setting $P(v_t^e \mid n_t^e)$ using gene expression information are given in Supplementary Material.

Gene expression levels have an additional role, besides guiding the model to assign higher or lower confidence in the AS measurements. Specifically, high gene transcript levels correlate with skipping of alternative exons, possibly because of 'kinetic coupling' (Kornblihtt, 2007) where changes in the rate of transcriptional elongation in turn affect the timing in which splice sites are presented to the splicing machinery. To account for this phenomena we include an additional signal in our model, based on the expression measurements. Unlike other signals in the model (or in standard matrix factorization), this signal is not learned but determined directly from the expression values of each exon's corresponding gene. We therefore denote it $\boldsymbol{\lambda}_V^e$, where $\lambda_{V,t}^e = v_t^e$, $t \in \{1, \ldots, T\}$. Since in this case the modulation levels may be positive or negative (corresponding to a positive or negative correlation with gene expression) and some may be relatively small, we set the distribution over it to be $P(m_V) = \mathcal{N}(m_v; \nu_V = 0, \gamma_V)$ and learn the variance γ_V.

To summarize, the addition of a baseline and expression signals via the m_b and m_V modulation variables imply that the condition-specific AS signals identified by the model and their assignments to the various exons are inferred after tissue-independent inclusion level and gene expression effects are accounted for. The resulting model can be represented as a Bayesian network (Fig. 2), with the model elements that are shared with standard FA denoted by a dashed line. Compared with FA, the model includes four major changes outlined above: the addition of a separate baseline and expression signal; the introduction of a mixture distribution over the factor modulation variable m_c via a matching s_c variable to enforce sparse signal activation with either positive or negative effects; and the addition of a competing outlier model with gene expression or other experimental information serving as noisy sensors for it, via the $\{n_t\}$ variable set.

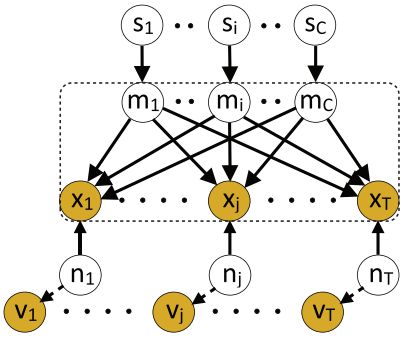

Fig. 2. A Bayesian network representation of the model. Observed variables are colored and dependencies are denoted with directed edges. The dashed frame denotes elements shared with standard FA.

2.1 Learning the model

Given input data $\{x_t^e\}$ and $\{v_t^e\}$, the objective of our probabilistic generative framework is to learn the model parameters $\Theta = \{\boldsymbol{\lambda}_c, \psi_t, \gamma_c, \nu_B, \gamma_V, \mu_t, \phi_t\}$ that maximize the likelihood of the data. We develop an efficient learning algorithm, based on generalized expectation maximization (EM), to optimize a bound on the likelihood termed free energy (Neal and Hinton, 1998). Similar to standard FA and independent factor analysis (IFA) (Attias, 1999) (and unlike SVD/PCA that are optimized analytically), convergence to the global optimum is not guaranteed. We therefore follow the learning algorithm description with a review of how it can be effectively initialized and directed to find good solutions. Finally, we review how the free parameters that control the number of AS signals C and the sparsity of those $P(s_c)$ can be set.

The structure of the graphical model in Figure 2 illustrates the computational complexity of running standard parameter learning and inference methods given the data, since the dependency between $\mathbf{s} = s_1, \ldots, s_C$ and $\mathbf{n} = n_1, \ldots, n_T$ makes such inference intractable for moderate values of T and C. We therefore use a variational approximation (Neal and Hinton, 1998) for the joint posterior distribution $Q(\mathbf{s}^e, \mathbf{m}^e, \mathbf{n}^e)$ given the observed inclusion levels $\mathbf{x}^e$ and additional data $\mathbf{v}^e$:

$$\begin{aligned} P(\mathbf{s}^e, \mathbf{m}^e, \mathbf{n}^e \mid \mathbf{x}^e, \mathbf{v}^e) &\approx Q(\mathbf{s}^e, \mathbf{m}^e, \mathbf{n}^e) = \\ &= \prod_{c=1}^{C} \Big(Q(s_c^e) \mathcal{N}(m_c^e; \eta_{s_c}^e, \sigma_{s_c}^e) \Big) \prod_{t=1}^{T} Q(n_t^e), \end{aligned} \tag{3}$$

where $\eta_{s_c}^e, \sigma_{s_c}^e$ are the variational parameters governing the posterior distribution over the modulation assignments. Given this variational approximation, learning a maximum likelihood model is done using an EM-like iterative procedure. We defer additional technical details to the Supplementary Material, and only note that for the results presented in the following sections, the learning procedure converges to a set of parameters that define the AS signals (given by $\{\boldsymbol{\lambda}_c\}$) and that the posterior belief that a given exon's AS profile $\mathbf{x}^e$ includes a specific signal as either 'off', 'on' or 'reversed', is given by $Q(s_c^e)$ with $s_c \in \{0, +1, -1\}$ in the above equation.

As is usually the case with EM-based learning algorithms, it is imperative to initialize the model parameters properly. When there is no prior knowledge, we create a random initialization point by setting $\boldsymbol{\lambda}_c$ to the average AS profile plus independent Gaussian noise with variance equal to the AS profile variance. Similarly, μ is initialized to the average tissue profile, while Ψ and Φ are initialized to the variance of the AS profiles. In addition, to avoid poor local minima for a given training set, we repeat this procedure N times and select the model with the highest likelihood ($N = 50$ in the experiments described below).

Finally, we describe how the model's free parameters can be set. We employed a standard approach of cross-validation to test how well different model settings do on train and test data. The most crucial parameter to set

is C, the number of underlying AS signal assumed to comprise a given dataset. We evaluate different values for C in the following section. For the signal sparsity prior $P(s_c)$, we wanted the model to avoid assigning AS signal with low values and therefore used the following preprocessing. Initial SVD analysis identified the well-known AS signals for CNS, muscle and embryo tissues (Section 3), with the cumulative distribution over the singular values associated with these signals typically shaped like a sigmoid. Consequently, we set the prior $P(s_c = \pm 1)$ at 0.08 to reflect the probability mass of both edges of this sigmoid. The AS signals identified and their assignments to exons were not sensitive to changes (± 0.05) to these settings as long as sparsity was maintained (data not shown).

2.2 Setting signals using biological knowledge

An additional benefit of the model-based approach described here is that specific initialization points and model constraints can be easily incorporated. These initialization points or model constraints can be used to reflect prior biological knowledge about the underlying AS signals in the data. While the uninformed initialization scheme described above works generally well (Section 3), several reasons may lead researchers to prefer biologically directed solutions. First, learning such solutions may be computationally efficient, leading to quick convergence and avoiding excessive numbers of random restarts. Perhaps more importantly, our generative model ultimately serves as an approximation for the physical process that yielded the observed measurements. As such, there may be several solutions the model can converge to, with some that are biologically plausible yet quite different. For these reasons, as we demonstrate in Section 3, it is beneficial to have the ability to use different biologically directed and undirected settings, exploring the space of possible solutions.

To direct the learning algorithm toward a certain solution for the AS signals, we use the following procedure: for any subgroup of conditions $T' \subset \{1,\ldots,T\}$ corresponding to a known signal we initialize a matching signal λ_c so that $\text{sign}(\lambda_{c,t}) = \text{sign}(\lambda_{c,t'})\ \forall t,t' \in T'$, while $\lambda_{c,t} = 0\ \forall t \notin T'$. If this is a good solution in terms of the likelihood surface, the learning algorithm can quickly converge to a similar solution in the neighborhood of this starting point. As we later show, we can also initialize a subset of the signals in such a way, and learn the rest of the signals using random initializations. Alternatively, we can constrain the model so that $\lambda_{c,t} = 0\ \forall t \notin T'$ and learn only the subset of the parameters $\{\lambda_{c,t}\}, \forall t \in T'$. If we use only such constrained signals, this is equivalent to inferring the contribution of AS signals from predefined condition groups, along with the baseline inclusion level. We note, however, that unlike many clustering and bi-clustering algorithms, even in such a constrained scenario, each condition t may be included in more than a single AS signal, and for a specific signal c, the contribution of the conditions T' that define it need not be the same. The advantage of this modeling ability is nicely illustrated in the AS signals depicted in Figure 1E, derived using unconstrained learning with a biologically derived initialization point. While the eye tissue is obviously rich with nerve cells, it is not exclusively part of the CNS. Consequently, it appears in the inferred CNS signal ($\boldsymbol{\lambda}_1$), but has a lower value associated with it.

3 RESULTS

We used the dataset of Fagnani *et al*. (2007), comprising 3707 cassette exons measured across 27 mouse tissues to evaluate our computational method. The condition-specific AS signals, corresponding to splicing changes in CNS, muscle, embryo and digestive tissues, are shown in Figure 1. The error bars for the signals were derived by randomly sampling subsets containing 80% of the original data. The four tissue-specific AS signals identified were also supported by a recent work where splicing changes corresponding to these four signals where predicted directly from genomic sequence and verified experimentally using RT-PCR experiments (Barash *et al*., 2010).

3.1 Comparison to alternative approaches

The comparison to alternative computational approaches for signal extraction includes two main parts: the AS signals identified, and their assignment to exons. Computational alternatives can be broadly divided into two subgroups: supervised and unsupervised. The supervised approach is based on prior knowledge of condition groups (e.g. CNS tissues) and is executed by computing a statistic for each exon such as the difference between its mean inclusion level in a predefined group of conditions compared with all the others. The set of exons for which the inclusion levels in these groups deviates the most are subsequently assigned the AS signal matching these groups (Castle *et al*., 2008; Fagnani *et al*., 2007). The second, unsupervised, computational alternative includes clustering algorithms discussed in Section 1, and variants of matrix factorization.

We start by comparing the identified AS signals. For the supervised approach, comparing the three major AS signals corresponding to splicing changes in CNS, muscle and embryo tissues is not particularly informative as the signals are both known and highly robust (see below). However, it is important to note that the supervised approach can only be applied to signals that are already known. Unless additional exploratory analysis is performed, this approach cannot detect unknown signals such as those corresponding to splicing changes in digestive tissues or in two subgroups of CNS tissues described below.

Matrix factorization algorithms, such as SVD, PCA and FA, are implemented in various software packages and represent the second alternative approach for AS signal extraction. We term this approach 'unsupervised' since, unlike the first approach described, the underlying signals are not set in advance. For comparison, we focused on SVD, which was recently applied to this task (Wang *et al*., 2008). In SVD, the input matrix of observed inclusion levels X is decomposed so that $X = USV^T$. The diagonal matrix S contains the singular values, ordered by magnitude, while the rows V_k represent 'eigen-exons'. SVD guarantees that for any $C \leq \text{rank}[X]$ using only the first C components of the matrices U, S, V^T produces a matrix X^C that is the closest, by MSE, to X from all rank C matrices. This can be given a probabilistic interpretation when assuming a fixed Gaussian noise model for the observations.

Figure 3 shows the results of SVD analysis. Based on the magnitude of the singular values (Fig. 3A), we included in subsequent analysis the first six eigen-exons. The first and by far most dominant singular value (119.2), corresponds to a tissue-independent eigen-exon, while the following five eigen-exons, denoted $E_1,\ldots,E_5$, are shown in Figure 3B. SVD clearly retrieves CNS, muscle and embryo AS signals represented by the first three eigen-exons. To test how robust the signals identified by SVD are, we performed the following procedure: we created ten subsets containing 80% of the data, then computed the pairwise correlation between the signals identified for each of the 10 data subsets. The pairwise correlations between the five eigen-exons derived for each of these data subsets are plotted as a heat-map in Figure 3C. The first three eigen-exons, matching splicing changes in CNS, muscle and embryo tissues, were highly robust but the fourth and fifth signals were less clear (Fig. 3B). Over the 10 data subsets, the fourth and fifth eigen-exons contained various combinations of tissues with at

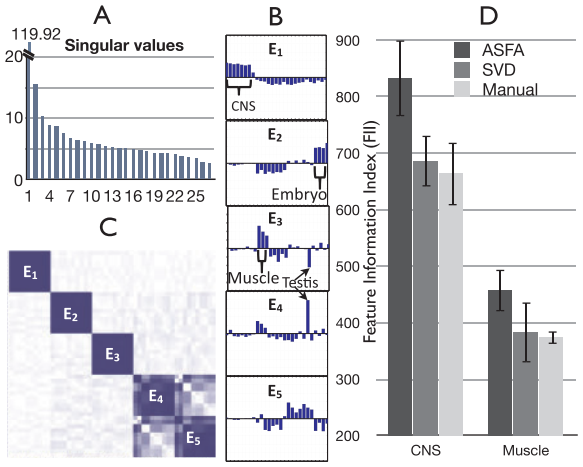

Fig. 3. Comparison to alternative approaches. (**A**–**C**) SVD analysis, including the singular values (**A**), examples of the first five condition specific eigen-exons (**B**), and a heat map (**C**) of the pair-wise correlation between the first five eigen-exons identified from ten random subsets of the data. (**D**) Comparison of the FII, which measure enrichment of previously reported regulatory features in groups of exons assigned the CNS (left) and muscle (right) AS signals. Signal assignment was performed using our model (denoted ASFA), SVD analysis, and by computing for each exon the difference between the mean inclusion level in the pre-defined tissue group and the other tissues (denoted Manual).

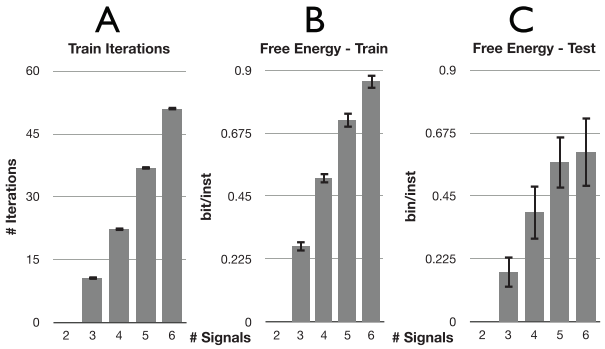

Fig. 4. The effect of varying the number of condition-specific AS signals between two and six: (**A**) the number of iterations until convergence. (**B**) Free energy (average bits per instance) for the train set. (**C**) Free energy for the test set. In all plots the baseline is a model with only two signals, given on the far left, and therefore all values for it are by definition zero.

least one always having a strong peak in testis. The testis tissue, which is a clear outlier in this dataset, also appeared in other eigen-exons, such as the muscle one (E_3) depicted in Figure 3B. This result is probably due to the fact that SVD analysis, based on a uniform Gaussian noise model and no additional modulation, is more sensitive to outliers.

Next, we evaluated the quality of the signal assignments to exons. When analyzing real-life high-throughput data, the correct assignment of signals to exons is generally not known. We therefore defined an independent measure for the quality of the signal assignment, termed the feature information index (FII). In short, the FII measures the enrichment of previously reported regulatory elements in groups of exons assigned a specific AS signal (e.g. splicing changes in CNS tissues). See Supplementary Material for more details. The rational behind the FII measure is that a better definition of a set of exons as those that exhibit splicing changes in certain conditions should consequently lead to finding higher enrichment levels within this exon set of elements known to regulate splicing in these conditions. We note that the algorithms only had access to AS and expression measurements; thus, the FII can serve as an independent quality measure. Moreover, the FII may also be indicative of the quality of further downstream analysis of high-throughput AS datasets, as many of the works producing these datasets subsequently try to identify the regulatory elements and underlying mechanisms that govern condition-specific AS.

Our model allows the assignment of signals to exons by setting a threshold that represents high confidence according to the signal posterior $Q(s_c^e)$ defined in Section 2. In the case of SVD, each column vector U^c defines an ordering over the amount each eigen-exon c contributes to each exon. Similarly, for the supervised approach the magnitude of the difference between the mean inclusion level in a predefined condition group (e.g. CNS tissues) and the rest also defines an ordering over the exons. However, both of these computational alternatives have no built-in method to set a threshold for signal assignments. In order to avoid biasing the FII measure due to differences in group sizes, we therefore used the orderings of these methods defined over exons for each AS signal to create groups of the same size as our model defined. Specifically, in the experiments described we used a confidence threshold of $Q(s_c^e) \geq 0.9$ to define both the exons that had a signal ($s_c = \pm 1$) or did not have it ($s_c = 0$). Changing the threshold to 0.99 yielded similar results (data not shown).

The results of the FII evaluation are summarized in Figure 3D. Only CNS and muscle tissue groups are shown as these are the only tissues for which a substantial knowledge of condition-specific *cis* regulatory elements is currently available (see Supplementary Material for details). SVD and the supervised method gave similar results, while our method scored significantly higher for both signals. The similar performance of SVD to the manually constructed tissue groups is to be expected in this case as both matching eigen-exons are dominated by these tissue groups, and SVD uses a fixed MSE model to assign those to exons. In contrast, our modeling approach identified the same two AS signals but is able to reject noisy measurements and assign signals to exons with varying degrees of splicing changes due to the modulation factor.

3.2 Evaluation of model settings

We start the evaluation of the model settings by addressing the most prominent question of how many AS signals are we able to identify in the data. We employed a train and test procedure, randomly choosing 80% of the exons for training, and keeping the other 20% for testing. Figure 4 shows the effect of increasing the number of underlying AS signals C in our model, where the baseline is a simple model containing only two tissue-specific AS signals, a tissue-independent signal (λ_B) and the expression-dependent signal (λ_V). Error bars were derived by repeating the above procedure 10 times. As expected, when the number of signals increases, so does the time it takes for the algorithm to explore the search space and converge (Fig. 4A). While the free energy keeps dropping for

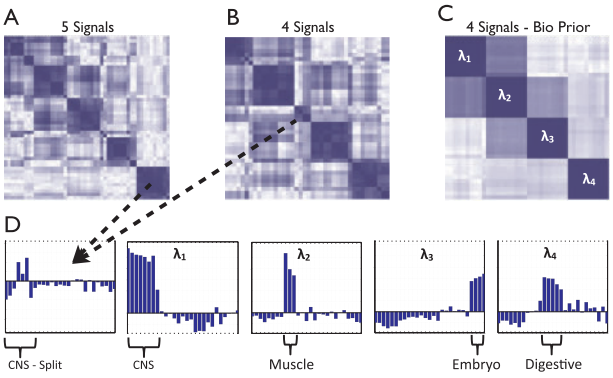

Fig. 5. The effect of different model settings on the identified AS signals. (**A**–**C**) Heat maps of the pairwise correlation between all AS signals identified in 10 random subsets of the data when (**A**) learning five AS signals (**B**) learning four signals, and (**C**) learning four signals with three signal initialized to CNS, muscle and embryo tissues. (**D**) Examples of the AS signals identified. The four on the right match the ones shown in Figure 1. The leftmost signal corresponds to a possible split of CNS tissues into two subgroups.

the training set, for test data we see a clear saturation effect after reaching five signals.

Since the algorithm converges during learning to a specific set of signals that represent a local optimum in the search space, a highly related question is how robust are these signals. In one extreme, convergence to very different solutions under slight data perturbation that make no biological sense would indicate a problem in the modeling approach or with the ability to overcome local minima, while, on the other hand, consistent convergence to a single solution would suggest a distinct global optimum under the modeling assumptions. To test the robustness of the identified signals, we repeated the same procedure that we used for evaluating SVD, using the same 10 randomly chosen subsets containing 80% of the original data. For each subset we learned the AS signals, then computed the pairwise correlation between the signals. Figure 5 shows a heat-map for the pairwise correlations between all AS signals learned under different settings of the algorithm. The tested settings include either four or five tissue-specific signals, using random initialization and initializing the model to include three known AS signals corresponding to splicing changes in CNS, muscle and embryo tissues. In general, we found that randomly initialized runs always converged to solutions that included CNS, muscle and embryo signals, sometimes slightly combined. The best scoring solutions had distinct CNS, muscle and embryo signals, with a split between the CNS tissues sometimes occurring when learning either four or five signals (Fig. 5D, left panel). This split may represent a novel distinction, in terms of AS signals, between two subgroups of CNS tissues: one that is dominated by spinal cord and hindbrain and another that is dominated by striatum and cortex.

When the model was initialized with three AS signals corresponding to known splicing changes in CNS, muscle and embryo tissues, the algorithm convergence was highly robust for all 10 data subsets (Fig. 5C). We note that in this setting the first three signals were only initialized to these tissue groups but not held fixed. Moreover, the initialization for the fourth signal was kept random as we had no prior knowledge for additional signals, yet the algorithm consistently converged to it given the other settings. The four tissue-specific signals the algorithm converged to are shown in Figure 5D. These signals match with those shown in Figure 1, where we also included the variance of the signals. We noticed that adding additional random restarts to several data subsets that originally converged to a different set of signals, eventually yielded this set of AS signals, implying this may be a slightly better solution and possibly a global optimum for these data subsets too. Taken together with the other results, we conclude that while the search space of the algorithm contains many local optima, all of these include a combination of CNS, muscle and embryo signals. The ability to switch between directed and undirected exploration of the search space was beneficial for the analysis of the data, with the undirected search identifying one local optimum that includes a distinction between two groups of CNS tissues and the directed search leading to a more stable solution that includes a novel splice pattern in digestive tissues verified by direct experiments.

Next, we tested the effect of modeling each measurement as generated from either the AS signals or a competing background model, with a posterior belief derived from the observed gene expression levels. Compared with an uninformative sparse noise prior, the derived AS signals remained similar but convergence time was reduced by 30%. This moderate effect was probably due to the fact that only ~11% of the measurements in the data of Fagnani *et al.* (2007) were suspected as noise based on expression values; thus, using this prior allowed the algorithm to converge more quickly but had little effect on the actual result. We suspect the noise prior may play a more critical role in cases where a larger portion of the data is expected to be noise (see Supplementary Material for an example). We also tried varying the sparsity of the signal prior ($P(s_c = \pm 1)$), within a ±5% range, with no substantial effect in terms of the signals identified or their assignment to exons. Standard FA, where sparsity of the signals is not enforced, gave similar results to the SVD analysis described before (data not shown).

3.3 Regulatory features associated with AS signals

While not being the main focus of this work, we conclude this section with a review of the correspondence between the AS signals we identified and known regulatory elements included in the FII. In general, we found excellent agreement between our results and previously reported ones. Nova YCAY motifs were found to be enriched mostly in the downstream introns of exons associated with increased inclusion in CNS and mostly downstream of exons downregulated in those tissues. Some of the more distant regions enriched in Nova motifs were also identified (Ule *et al.*, 2006). Fox motif variants ([U]GCAUG) were associated with inclusion in muscle and to lesser extent brain tissues when appearing in the downstream intron, and mostly with exclusion in those tissues when in the upstream intron. However, this motif was correlated with a general change in exon inclusion in those tissues, indicating a reversed effect too, a result in accordance with a recent study (Zhang *et al.*, 2008). CU-rich motifs known to bind (n)PTB were found to be highly enriched both up-and downstream of exons exhibiting splicing changes in several tissue groups, most prominently CNS. This result is inline with (n)PTB known role as a derepressor in CNS tissues and its ability to loop across relatively short exons, as in the well-studied case of src N1 (Chan and Black, 1997). An interesting result involved the Quaking-like motif ACUAAY, which was previously reported enriched downstream of exons exhibiting

increased inclusion in muscle (Das *et al*., 2007; Wang *et al*., 2008). We not only identified this enrichment, but also an enrichment upstream of exons exhibiting splicing changes in CNS, which is in line with known roles of this class of splicing factors in neuronal disease mutations.

4 DISCUSSION

In this work we presented a model-based approach to identifying condition-specific AS signals from high-throughput data. Unlike other approaches, our generative model is specifically tailored for this task. It incorporates prior knowledge about AS, including known correlation to expression levels, modeling of a tissue- independent signal, the expected sparsity of the AS signals and the fact that AS signals may have either a positive or negative effect. The model is also able to incorporate specific knowledge about a given dataset, including information about the quality of the measurements, related gene expression levels and knowledge of specific AS signals in the data. We compared our approach with commonly used alternatives and showed that on real data it was able to produce superior results in terms of the signals identified, their robustness and their assignments to biologically important groups of exons. For the latter, we defined an independent measure of quality, the FII, performed a literature search for tissue-specific splicing regulatory *cis* elements, and showed that the assignment of AS signals by our method correlated significantly better with those elements. Our method was able to detect a novel split of the signal for splicing changes in CNS tissues into two separate subgroups of tissues, and a previously unreported AS signal associated with digestive tissues. The four main tissue-specific AS signals identified by our model are supported by a predictive model derived from putative regulaotry features, including experimental testing of the model's predictions (Barash *et al*., 2010).

Previous work developing related models include IFA and another form of a mixture of factor analyzers (Attias, 1999; Ghahramani and Hinton, 1996). Both of these models, developed for different applications, differ from ours in the mixture model used and do not include domain-specific elements, such as the background model, the gene expression levels and the AS baseline signals. More recent work involving general matrix factorization algorithms include the work by (Dueck *et al*., 2005), which was applied to gene expression data and Robust PCA (Candes *et al*., 2009), which involves sparse signal assignments and noisy data. In general, the computational alternatives currently available can be broadly divided into two: supervised and unsupervised methods. The first mostly consists of computing statistics such as the mean inclusion level for a predefined group of conditions, while the other includes clustering methods such as hierarchical agglomerative clustering, as well as SVD, PCA and other variants of matrix factorization algorithms. Compared with those, our modeling approach can range between supervised and unsupervised signal identification, depending on the amount of additional information incorporated into the model. We were able to demonstrate the usefulness of this trade-off between search space exploration and prior knowledge exploitation for the identification of AS signals in the data.

There are several direct computational extensions to the model presented here. One extension involves defining the number of components in the model as a random variable and marginalizing over it. A recent work by (Paisley and Carin, 2009) implemented such a model for factor mixtures using a beta process prior. We note though that in our context we are not simply interested in marginalizing out the component number, since the identity of the AS signals and the exons associated with them are biologically meaningful. Another possible direction for future work is to replace the Gaussian mixture components used in our model with alternative distributions that would fit the data better. This change would be of practical use if the better fit to the data would also lead to more accurate signal assignments to exons.

We briefly reviewed some of the regulatory elements we identified as enriched in groups of exons associated with the AS signal reported. Many of these are in excellent agreement with previously published results, including the role and positional bias of *cis* element known to bind Nova, PTB and Fox. Some identified features, such as the Quaking motif upstream of exons highly included in CNS, offer possible insights to additional regulatory mechanisms.

The correlation between signals identified by our model and regulatory elements points to possible extensions and applicative potential of the modeling approach we propose. Our probabilistic framework can naturally be extended to a unified framework where combinations of regulatory features are used to explain identified AS signals and the assignment of these signals to exons is subsequently used to refine the regulatory programs learned. Such a unified approach for modeling regulatory programs has been applied successfully in other domains (Bar-Joseph *et al*., 2003; Beer and Tavazoie, 2004; Segal *et al*., 2002, 2003). In our recent work (Barash *et al*., 2010), the model described here was utilized to construct a regulatory code that predicts tissue-specific splicing changes directly from genomic sequence. Our framework can be extended to incorporate additional information such as secondary structure elements, nucleosome positions and splice factor binding measurements (Licatalosi *et al*., 2008), to gain further insights into the underlying regulatory mechanisms associated with each AS signal.

Based on the analysis of AS signals we report and the comparison of our method to standard computational alternatives, we believe this work will facilitate the study of future high-throughput AS datasets, extending our understanding of the transcriptome complexity.

ACKNOWLEDGEMENTS

We thank Inmar Givoni for helpful comments on the manuscript. B. F. holds a Canada Research Chair and is an NSERC EWR Steacie Fellow and a Fellow of the Canadian Institute for Advanced Research.

Funding: Canadian Institute of Health Research grants (to B.J.F. and B.J.B.); a Canada Foundation for Innovation and the Ontario Innovation Trust grant (to B.J.F.); an National Cancer Institute of Canada grant (to B.J.B.); grant from Genome Canada through the Ontario Genomics Institute (to B.J.F., B.J.B and others).

Conflict of Interest: none declared.

REFERENCES

Ashiya,M., and Grabowski,PJ. (1997) A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the

polypyrimidine tract binding protein and a brain-specific PTB counterpart, *RNA*, **3** 996–1015.

Attias,H. (1999) Independent factor analysis. *Neural Comput.*, **11,** 803–851.

Barash,Y. *et al*. (2010) Deciphering the splicing code. *Nature*, **464**, 7294.

Bar-Joeph,Z. *et al*. (2003) Computational discovery of gene modules and regulatory networks, *Nat. Biotechnol.*, **21,** 1337–1342.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence, *Cell*, **117,** 185–198.

Boutz,P. *et al*. (2007) MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development, *Gen. Dev.*, **21,** 71–84.

Candes,E. *et al*. (2009) Robust principal component analysis? *Stanford Technical Report.* Available at http://www-stat.stanford.edu/∼candes/publications.html.

Castle,J.C. *et al*. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines, *Nat. Gen.*, **40**, 1416–1425.

Chan,R. and Black,D. (1997) The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream, *Mol. Cell Biol.*, **17** 4667–4676.

Das,D. *et al*. (2007) A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing, *Nucleic Acids Res.*, **35,** 4845–4857.

Dueck,D.*et al*. (2005) Probabilistic sparse matrix factorization with an application to discovering gene functions in mouse mRNA expression data. *Bioinformatics*, **21** (Suppl. 1), i144–i151.

Eisen,B.M. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl Acad. Sci.USA*, **95**, 14863–14868.

Fagnani,M. *et al*. (2007) Functional coordination of alternative splicing in the mammalian central nervous system, *Genome. Biol.*, **8,** R108.

Ghahramani,Z. and Hinton,G.E. (1996) The EM algorithm for mixtures of factor analyzers, *University of Toronto Technical Report*, CRG-TR-96-1. Available at http://www.cs.toronto.edu/∼hinton/absps/tr-96-1.pdf.

Hartmann,B. and Valcarcel, J. (2009) Decrypting the genome's alternative messages, *Curr. Opin. Cell Biol.*, **21,** 377–386.

Kornblihtt,A.R. (2007) Coupling transcription and alternative splicing, *Adv. Exp. Med. Biol.*, **623,** 175–189.

Licatalosi,D. *et al*. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing, *Nature*, **456,** 464–469.

Minovitsky,S. *et al*. (2005) The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons, *Nucleic Acids Res.*, **33,** 714–724.

Neal,R. and Hinton,G. (1998) A view of the Em algorithm that justifies incremental, sparse, and other variants. In Jordan,M.I. (ed.), *Learning in Graphical Models*. Kluwer Academic Publishers, Norwell, MA, USA, pp. 355–368.

Paisley,J. and Carin,L. (2009) Nonparametric factor analysis with beta process priors, In *International Conference on Machine Learning (ICML) 2009*. Vol. 382, Montreal, Canada, pp. 777–784.

Pan,Q. *et al*. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform, *Mol. Cell*, **16,** 929–941.

Pan,Q. *et al*. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, *Nat. Genet.*, **40**, 1413–1415.

Rubin,D. and Thayer,D. (1982) EM algorithms for ML factor analysis. *Psychometrika*, **47,** 69–76.

Scheper,W. *et al*. (2004) Alternative splicing in the N-terminus of Alzheimer's presenilin 1. *Neurogenetics*, **5**, 223–227.

Segal,E. *et al*. (2002) From promoter sequence to expression:a probabilistic framework. In *RECOMB*. ACM Press, Washington, DC, USA, pp. 263–272.

Segal,E. *et al*. (2003) Modules networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34,** 166–176.

Segal,E. *et al*. (2004) GeneXPress: a visualization and statistical analysis tool for gene expression and sequence data. In *Proceedings of the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

Shai,O. *et al*. (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data, *Bioinformatics*, **22,** 606–613.

Ule,J. *et al*. (2006) An RNA map predicting Nova-dependent splicing regulation, *Nature*, **444,** 580–586.

Wang,G.S. and Cooper,T. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749–761.

Wang,Z. and Burge,CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code, *RNA*, **14,** 802–813.

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

Zhang,C. *et al*. (2008) Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Gen. Dev.*, **22,** 2550–2563.