

Methodology article

Open Access

## Prediction of alternatively skipped exons and splicing enhancers from exon junction arrays

Katerina Kechris\*<sup>†1</sup>, Yee Hwa Yang<sup>†2</sup> and Ru-Fang Yeh\*<sup>3</sup>

Address: <sup>1</sup>Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 4200 East 9th Avenue, B-119, Denver, CO 80262, USA, <sup>2</sup>School of Mathematics and Statistics, Sydney Bioinformatics, Carslaw Building F07, University of Sydney, NSW 2006, Australia and <sup>3</sup>Department of Epidemiology and Biostatistics, University of California San Francisco, Box 0560, San Francisco, CA 94143, USA

Email: Katerina Kechris\* - katerina.kechris@ucdenver.edu; Yee Hwa Yang - jeany@maths.usyd.edu.au; Ru-Fang Yeh\* - rufang@biostat.ucsf.edu

\* Corresponding authors †Equal contributors

Published: 20 November 2008

Received: 15 March 2008

BMC Genomics 2008, 9:551 doi:10.1186/1471-2164-9-551

Accepted: 20 November 2008

This article is available from: <http://www.biomedcentral.com/1471-2164/9/551>

© 2008 Kechris et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Alternative splicing of exons in a pre-mRNA transcript is an important mechanism which contributes to protein diversity in human. Arrays for detecting alternative splicing are available using several different probe designs, including those based on exon-junctions. In this work, we introduce a new method for predicting alternatively skipped exons from exon-junction arrays. Predictions based on our method are compared against controls and their sequences are analyzed to identify motifs important for regulating alternative splicing.

**Results:** Our comparison of several alternative methods shows that an exon-skipping score based on neighboring junctions best discriminates between positive and negative controls. Sequence analysis of our predicted exons confirms the presence of known splicing regulatory sequences. In addition, we also derive a set of development-related alternatively spliced genes based on fetal versus adult tissue comparisons and find that our predictions are consistent with their functional annotations. *Ab initio* motif finding algorithms are applied to identify several motifs that may be relevant for splicing during development.

**Conclusion:** This work describes a new method for analyzing exon-junction arrays, identifies sequence motifs that are specific for alternative and constitutive splicing and suggests a role for several known splicing factors and their motifs in developmental regulation.

### Background

Eukaryotic gene expression is controlled by a series of biological events involving various interactions among proteins, DNA and RNA that are subject to complex regulation. One of the essential processes is pre-mRNA splicing, in which the spliceosome complex recognizes splice sites (SS) in the precursors of mRNA following transcription, removes noncoding introns and joins neighboring exons to form mature messenger RNA that can be

further translated into protein sequence. Both 5' splice sites (marking exon-intron boundaries, or *donor* sites) and 3' splice sites (marking intron-exon boundaries, or *acceptor* sites) consist of short basal signals well conserved from yeast to human, but contain insufficient information to accurately identify exon-intron boundaries in vertebrates [1]. Consequently, there are numerous cryptic splices embedded throughout the pre-mRNA transcript. Variations in the splice site detection process can create multi-

ple transcript variants for a gene and is referred to as alternative splicing. Alternative splicing events may include the skipping or retention of entire exons, intron retention or alternative usage of 3' or 5' splice sites. These changes often lead to modifications in the encoded proteins and have been shown to play a critical role in development and disease [2-4]. Alternative splicing also provides an efficient means to expand protein diversity from the limited gene pool. It is estimated that 50–80% of the approximately 25,000 human protein-coding genes are subject to alternative splicing [5-7] underscoring the importance of splicing regulation.

The correct recognition of splice sites is facilitated by various protein-protein and protein-RNA interactions. In addition to exon-bridging or intron-bridging interactions of splice sites, there are many non-splice-site sequences in exons and adjacent introns, termed enhancers or silencers, that stimulate or repress the splicing of constitutive or alternative exons. Both enhancers and silencers have been identified through *in vitro* and *in vivo* methods [8-11]. Large-scale identification of these sequence elements are based on computational methods which analyze multiple sequences simultaneously. Some authors have focused on constitutively spliced exons [12-15], while others have focused on identifying the sequence signals that discriminate between constitutively and alternatively spliced exons [16-21]. More recent methods now take advantage of the availability of genomes from related species to make predictions based on evolutionarily conserved sequences [22-25]. Although motif search methods such as [26] identify sequences that are similar to previously determined enhancers and silencers, most computational approaches require as input a large collection of alternatively spliced exons.

The most widely used method to identify alternative splice variants is based on aligning EST and mRNA sequences to the genome [27-29]. Application of this method has provided estimates for the percentage of human genes with multiple splice variants between 40% and 60% [27-29]. However, EST-based methods are often biased towards the transcript ends, prone to sequencing errors, biased towards more highly expressed genes, limited by the availability of clone libraries for particular cell and tissue types and not uniformly annotated. The recent adaptations of microarray technology in the form of splice arrays are now providing new directions for detecting alternative splicing. DNA microarrays were originally developed to measure expression levels for a large number of genes simultaneously. Arrays for detecting alternative splicing (splice arrays) contain probes specifically designed to investigate splice events. They have been used to explore splicing in yeast [30,31] and in mammalian cells (*e.g.*, [6,32-34]). Several different designs reviewed in

[35] have been developed including tilling arrays, arrays with specific probes that distinguish between known splice forms [32,36], arrays based on exon probes [37,38], exon-junction probes [6,39] or both [32,36,40,41]. Most of the data analysis methods for splice arrays and levels of alternative splice inference depend on the array design (see review in [35]), and the design of these arrays typically rely heavily on known or predicted annotation of gene structures. In summary, splice arrays have been shown to complement the EST-based approaches by identifying a large number of novel alternative splices with no previous EST support, despite a moderate validation rate of ~50% by RT-PCR [6,42,43].

In this paper, we propose a new statistical approach for identifying alternatively spliced exons from exon-junction arrays, and predict a large set of alternatively skipped exons. This is achieved using the most comprehensive exon-junction array analysis to date [6], which monitored all exon junctions of over 10,000 RefSeq genes in 52 tissues samples. Our approach consists of a novel exon-skipping score that serves as a quantitative measure of evidence for alternatively skipped exons. By applying kmer-contrast and regression-based sequence analysis methods to the top ~8400 exons according to our score, we are able to recover several known splicing enhancers and identify additional novel candidate splicing regulatory motifs associated with skipped exons. Finally, we also identify a set of development-related alternative splices and their associated enhancers using a tissue-pair analysis followed by *de novo* motif finding algorithms. Enrichment analysis of gene ontology annotation supports the functional roles of the predicted development-related alternative splices and suggests a new scheme for identification of process-specific alternative splicing.

## Results and discussion

### Analysis of exon-junction array

The pre-processed array data described in [6] were obtained from the NCBI Gene Expression Omnibus (GEO) database with accession number GSE740. The dataset contains the average background subtracted intensity  $y_{i,j,k}$  from two dye-swapped arrays for each exon-junction probe  $j$  of gene  $k$  in tissue  $i$ , where  $i$  is one of the 52 tissues surveyed and  $j = 1, \dots, n_k$  for gene  $k$ , which has  $n_k$  probes bridging  $n_k+1$  exons. Note that intensity patterns of exon-junction (EJ) probes alone cannot distinguish between intron retention or alternative 5' or 3' splice site usage, and hence we only focus on the effective detection of exon skipping in the 8618 genes containing 5 exons or more ( $n_k \geq 4$  on the array). For each gene, we fit a linear model with terms for probe and tissue effects. Because of the large number of effects that need to be estimated, we must restrict the analysis to genes where there are enough data points. We found that a minimum of 5 exons was an

adequate cutoff that avoids over-fitting in our statistical analysis but does not filter too many genes. For each gene, 5 exons corresponds to 4 exon-junctions (4 data points per tissue) and consequently 3 adjacent exon-junction pairs.

If we treat exon inclusion as a simplified binary "on" or "off" event in each tissue, then conceptually, there are three main signal sources, other than noise, for the observed EJ probe intensity  $y_{i,j,k}$ : (i) probe-specific effect describing sequence binding affinity, (ii) tissue-specific effect resulting from differential expression, and (iii) alternative splice (AS) effect determined by exon inclusion or alternative splice site selection. A simple approach to recover the variance components of individual signal contribution is to use linear models on the properly transformed data (EJ probe intensities), essentially assuming that the transformed intensity is the sum of the probe, tissue, AS effect and white noise. A skipped exon will result in the two spanning EJ probes to be switched "off". This supports an exon-skipping score that measures the magnitude and concordance of AS effects in adjacent EJ probes spanning the same exon. Identifying exons that are alternatively skipped in some of the tissues surveyed is then equivalent to the detection of scores that deviate across tissues. In summary, our data analysis method can be described as a three-step procedure:

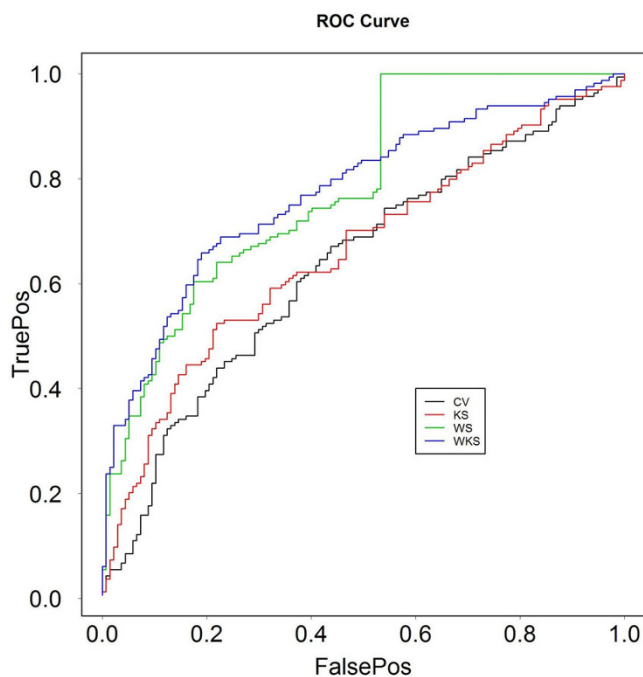
1. Estimate the variance stabilizing transformation [44] for the EJ probe intensities to satisfy the constant variance assumption for linear models.
2. Fit a linear model on the transformed data to estimate probe and tissue effects. The residuals of the fit  $r_{i,j,k}$  represent AS effects adjusted for probe- and tissue-specific effects.
3. Summarize exon-skipping events at the exon and gene level by defining an exon-skipping score based on the residuals  $r_{i,j,k}$  and test for large deviations of this score across tissues.

Note that [6], among others (*e.g.*, [45]), also used a linear model on log-transformed (step 1) intensities but estimated linear model parameters by medians (step 2), followed by *ad hoc* thresholding of the residuals  $r_{i,j,k}$  to define AS scores (step 3). We considered several alternatives to this approach. First, the original method of [6] only examined individual exon-junctions by thresholding the residuals into four categories from 0,1,2 and 3. The largest residuals, those given a score of 3, were predicted as splicing events. For each junction, the authors then tallied the number of tissues where the residuals had a score of 3. There appeared to be no systematic evaluation of neighboring pairs of exon-junctions in this scoring procedure.

Therefore, as an alternative, we calculated the novel statistic of residual products so that both flanking exon-junctions for an exon contribute to its score. Second, our preliminary analysis (data not shown) indicated that the logarithm function used by [6] to stabilize the variance may be insufficient. Therefore, we also tried alternative transformations. Third, we used statistical significance cutoffs to threshold the residuals (or residuals products).

In particular, at each stage of the three-step procedure, we considered several alternative methods for the same objective: applying the logarithm or arcsinh-based variance stabilizing transformation (VSN) function for stabilizing variances in step 1; using the standard least square fit (mean) or robust fit (median or median polish) for linear model parameter estimation in step 2; and measuring deviations from normality of residuals in step 3 using the non-parametric Kolmogorov-Smirnov (KS) test, a weighed non-parametric Kolmogorov-Smirnov test (WKS), parametric Wilk-Shapiro test (WS) [46] or simple thresholding of the coefficient of variation (CV, standard deviation normalized by the mean) on the residuals or residual products (see Methods). We viewed the selection of these options as an optimization problem, and determined the best procedure in terms of prediction accuracy using control data. We used two sets of alternatively and constitutively spliced exons curated from independent sources as controls (see Methods). The positive control (AS) consists of 164 genes supported by EST data. The negative control (CS) consists of 282 genes selected from the analysis of an Affymetrix exon array experiment that show constitutive splicing across a panel of 11 tissues. These two examples represent "approximate" controls because they rely on available ESTs or other data on selected tissues. However, we expect that they will be dominated by the desired events.

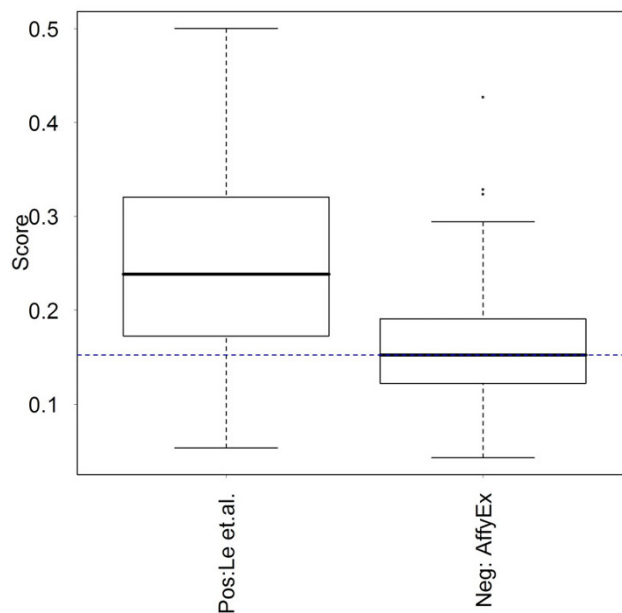
Using the control data sets and receiver operating characteristic (ROC) curve analysis, we found that the VSN transformation performs better than log-transformation as expected (data not shown), and that the fitting procedures in Step 2 were indistinguishable in terms of prediction errors (data not shown). Figure 1 shows the improvement of our procedure (VSN, least square fit, WKS) over our best reproduction of the method described in [6] using exon-level summaries (CV score, see Methods). Note that the AS calls in the previous work [6] were calculated from consistent AS scores from individual arrays which were not publicly available (only the average intensity of dye-swapped replicates were released). Therefore, we could not reproduce previous results in [6] but believe that the CV score is similar in spirit to that approach. Figure 2 shows the distribution of the WKS-statistic at the gene level for the two control sets. Overlaps in the distribution between the positives and negative may be due to the



**Figure 1**  
**ROC curve comparing step 3 options based on control data sets.** The x-axis is the percentage of false positives (FalsePos) and the y-axis is the percentage of true positives (TruePos). The step 3 options described in the text are CV (thresholding of the coefficient of variation), KS (Kolmogorov-Smirnov test), WS (Wilk-Shapiro test) and WKS (weighted Kolmogorov-Smirnov test).

caveats regarding the control sets discussed above. Nevertheless, the values of the WKS-statistic are generally higher for the positive controls versus the negative controls.

To further validate these results, we examined the WKS scores on another set of positive controls; the positive RT-PCR results in [6] and the 2656 RefSeq entries on the array with "cassetteExon" events from the UCSC "knownAlt" table hg18 annotation [47]. Although the RT-PCR set also contain genes with alternative splice site selections and intron insertions, the two new combined positive controls had WKS-statistics significantly greater than the negative control (Mann-Whitney test  $p$ -value =  $4.9 \times 10^{-4}$ ). Several of the highest scoring examples using the WKS method, such as UBA3 (ubiquitin-activating enzyme 3 isoform 1), MAPKAP1 (mitogen-activated protein kinase associated) and PPM1B (protein phosphatase 1B isoform 1) were identified as "cassetteExon" gene entries from UCSC but were low scoring using the CV method, which only uses data from a single exon-junction. This illustrates that the combined product score based on neighboring exon-junctions can predict known examples which are missed by single junction methods. In summary, given the original



**Figure 2**  
**Boxplot of WKS-statistic (y-axis) at the gene level for each of the control sets.** The positive control set is labeled "Pos: Le et al." and the negative control set is labeled "Neg: AffyEx".

linear model in [6], we derived a new exon-skipping score that 1) is based on an alternative transformation which improves the correction for non-constant variance and 2) draws on information from both splice junctions of an exon using the product residuals.

Using a score cutoff that balances the rate of true positive and false positives in our original two controls, we identified a set of alternatively spliced genes and corresponding exons (see Methods). We then determined the lowest ranking genes and exons according to our score to obtain an equivalently sized set of constitutive spliced exons. In total we predicted 8433 alternatively skipped exons and 8113 constitutive exons. An analysis of the Gene Ontology (GO) terms [48] for our predicted alternatively spliced genes (see Methods) using Gostat [49] showed enrichment in 114 GO terms. More than half of these GO terms are related to metabolism, cell death, regulation, transcription, splicing and protein targeting and localization (See Additional file 1, Table S1), which are categories consistent with prior studies [19,29,45].

**Sequences associated with alternatively splicing**

To identify sequence motifs associated with the regulation of alternatively skipped exons, we adapted the contrast-mer-based RESCUE-ESE algorithm for splicing [12] and the regression-based REDUCE program for transcription factor binding sites [50]. The RESCUE-ESE algorithm

accounts for the fact that splicing enhancers compensate for weak splice site signals, and finds candidate exonic splicing enhancers (ESE) for constitutive splicing from statistically over-represented hexamers in exons versus introns and associated with exons defined by weak splice sites versus strong splice sites. REDUCE enumerates all possible kmers in the promoters up to a specified length and uses a linear regression model to find kmers that show significant correlation with gene expression levels from a single microarray experiment. To identify putative splicing regulatory elements associated with alternative splicing, we used different contrast sets for RESCUE-ESE and employed our exon-level statistic for REDUCE on exons and their flanking intron sequences.

#### **Exonic splicing enhancers (ESE) for alternatively and constitutively spliced exons**

A direct contrast of alternatively skipped exons (AE) versus constitutively spliced exons (CE) in addition to differences in exons versus flanking introns gave 6 motifs associated with the 5' splice site of AE (204 hexamers in Additional file 1, Table S2) and 7 motifs for the 3' splice site of AE (192 hexamers in Additional file 1, Table S3), among which 5 motifs are in common. Several of the motifs overlap known ESE motifs from deletion experiments, functional SELEX or SR protein binding sequences (see survey in [12]), such as the purine-rich AAGA for SRp40, SRp55, SRp30a and SRp75, ACGA and TGAAG for 9G8, SC35 and ASF, and consensus GAAG for Tra2 $\beta$  [51] and motif variants for ASF/SF2. Three of the 8 nonredundant motifs (Figure 3) also match RESCUE-ESE predicted and experimentally verified motifs for constitutive splicing from [12], indicating that similar ESEs and splicing factors are involved in both alternative and constitutive splicing. For the kmers that define the predicted motifs, there are  $\sim 1.1$  to 2.2 times more kmer counts in alternative exons versus constitutive exons (Figure 3).

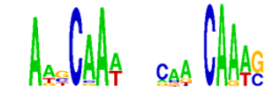
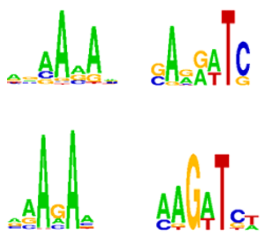


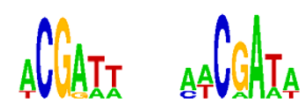

Interestingly, when the hexamers were subjected to the additional contrast of splice site strength association (weak splice site exons versus strong splice site exons) as in the RESCUE-ESE analysis for constitutive splicing, the number of significant kmers reduced drastically. The only significant AE motif associated with weak splice sites is the well known motif GAAGA from 13 and 6 non-redundant hexamers in the 5' and 3' analysis respectively (Figure 4 and Additional file 1, Table S4). Two other AE-associated ESE motifs, an A-rich and a TGGA motif, are linked to strong splice sites. These results suggest that the GAAGA motif and its associated splicing factors (9G8, ASF, SC35, SRp40, SRp55, R30A, SRp75 all have been implicated) may be critical for regulating alternative exons in the presence of very weak splices. This result may be due to the observation that splice sites involved in alternatively spliced exons tend to be weaker overall [25] and therefore, splice site strength dependency will only pick up the

strongest enhancers. Using position-specific scoring matrices to evaluate splice site strength, we did not observe a bias in weaker splice sites in the entire group of AE versus CE. However, correcting for overall gene effects, the splice site scores in the AE were relatively smaller than the scores in the CE (see Additional file 2).

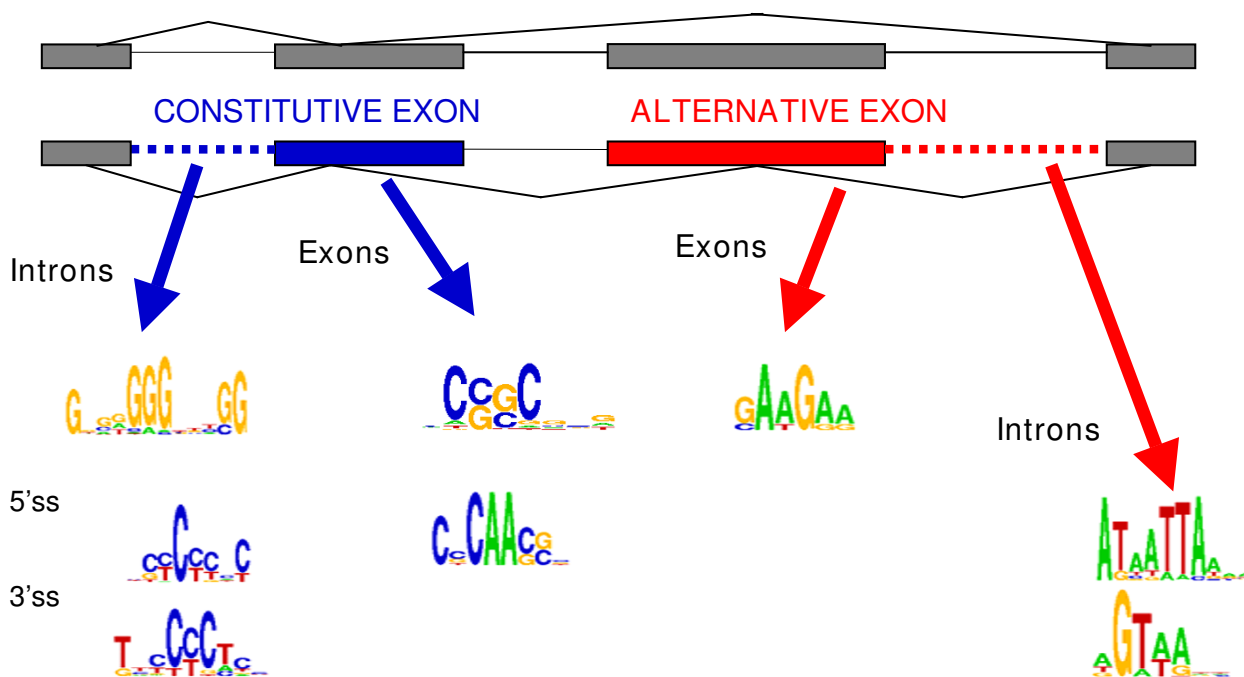
We also focused the contrast analysis on constitutively spliced exons to select for ESEs relevant for constitutive splicing using CE versus AE and exons versus introns comparisons and found 23 motifs for the 5' analysis and 12 motifs for the 3' analysis. The majority of motifs detected are CG-rich with diverse consensus patterns, most likely reflecting the coding bias of CE (Figure 4 and Additional file 1, Tables S5 and S6). Besides the CG-rich motifs, 6 of the 23 non-redundant motifs found also overlap known motifs or natural occurring enhancers, including CA-rich motifs (CAAC, CCAC) for splicing factor YB1, TGCCGTT for SC35 and functional SELEX motifs for Schaal-II-D (TCTCC) [12]. The resulting motifs were similar when we also restricted the comparison to weak splice sites. Finally, no motifs were found to be associated with strong splice sites in support of CE, consistent with the notion that splicing enhancers are required only when splice sites do not contain enough information in order to facilitate unambiguous exon recognition.

Alternatively, we applied the REDUCE software to correlate sequence features with the exon-skipping score (see Methods). Table 1 shows the most significant kmers ( $p$ -value  $< .01$ , Bonferroni corrected). Kmers positively correlated with the exon-level skipping score are associated with alternatively spliced exons, while negatively correlated kmers are associated with constitutively spliced exons. Note that this regression-based approach is a natural extension of contrast analysis of AE versus CE, and aims to evaluate the sequence association at the genome scale without the need to dichotomize scores. The REDUCE kmers were consistent with the above RESCUE-ESE analysis; the top negatively correlated kmers for CE are primarily G/C rich with a CACC-containing motif (tCACCg), and the AAGAA motif was found to be the top candidate for AE.

We applied similar contrast analysis to identify intronic motifs associated with AE by selecting for kmers over-represented in introns flanking AE versus CE-associated introns and in introns versus exons. We found A/T rich motifs associated in introns flanking AE and pyrimidine tract-like motifs and G-triplets in introns flanking CE (see Figure 4, Additional file 2). Randomization runs were performed on the exon data to assess the frequency of observing the predicted kmers in random data. The results indicate that we are observing more predicted kmers than expected by chance (see Additional file 2).

AE Motif ID	AE Motif Logo	Ratios	Overlapping Known Motifs
5'(3)/3'(2)		[1.28-2.04]/[1.18-1.74]	RESCUE-ESE 3C
5'(1,5)/3'(1,6)		[1.16-1.78]/[1.18-1.70] [1.15-1.92]/[1.17-1.63]	Known exonic splicing enhancer in natural context GAGGAAGAA ( <i>trans</i> -factors SRp40, SRp55, SRp30a, SRp75)
5'(2)/3'(7)		[1.18-1.72]/[1.19-1.58]	Known exonic splicing enhancer in natural context GATGAAGAG ( <i>trans</i> -factors 9G8/ASF,SC35)  RESCUE-ESE 5C/3D
3'(3)		[1.17-1.53]	RESCUE-ESE 5A/3G
5'(4)/3'(5)		[1.58-2.24]/[1.29-2.23]	Known exonic splicing enhancer in natural context GACGACGAG ( <i>trans</i> -factors 9G8/ASF,SC35) and high affinity SELEX binding motif for 9G8.
3'(4)		[1.40-2.04]	RESCUE-ESE 5B/3A

**Figure 3**  
**Predicted non-redundant 3' and 5'SS associated AE motifs and their overlapping known motifs.** Predicted AE motifs IDs and motifs are listed in columns 1 and 2. The motifs are displayed graphically as logos in the order of the IDs in column 1. The Motif ID includes a label for whether it was discovered relative to the 5' or 3' splice site and a numeric identifier for the motif in parenthesis. For all the kmers that define each motif in Column 2, Column 3 lists the range of ratios of the occurrence of these kmers in alternative exons versus constitutive exons. See Additional file 1, Tables S2 and S3 for the list of all motifs, their kmers and their numeric identifier. RESCUE-ESE and known enhancers (column 4) are taken from the survey in [12].



**Figure 4**  
**Summary of motifs found in alternative and constitutive exons and flanking introns.** Two alternative splice forms are indicated at the top. An alternatively spliced exon for the splice forms is labeled in red while a constitutively spliced exon is labeled in blue. The flanking intron for the alternative and constitutive exon is indicated by the dashed line. The arrows point to representative motifs associated with each of the sequences. Motifs found with respect to the 5' or 3'SS introns flanking constitutive exons are labeled separately.

**Development-related alternative splices derived from tissue-pair analysis**

To investigate whether *trans*-acting elements (*i.e.*, protein-RNA binding proteins) are relevant for development, we analyzed a comprehensive gene expression data set from the Gene Expression Atlas [52], which contains results

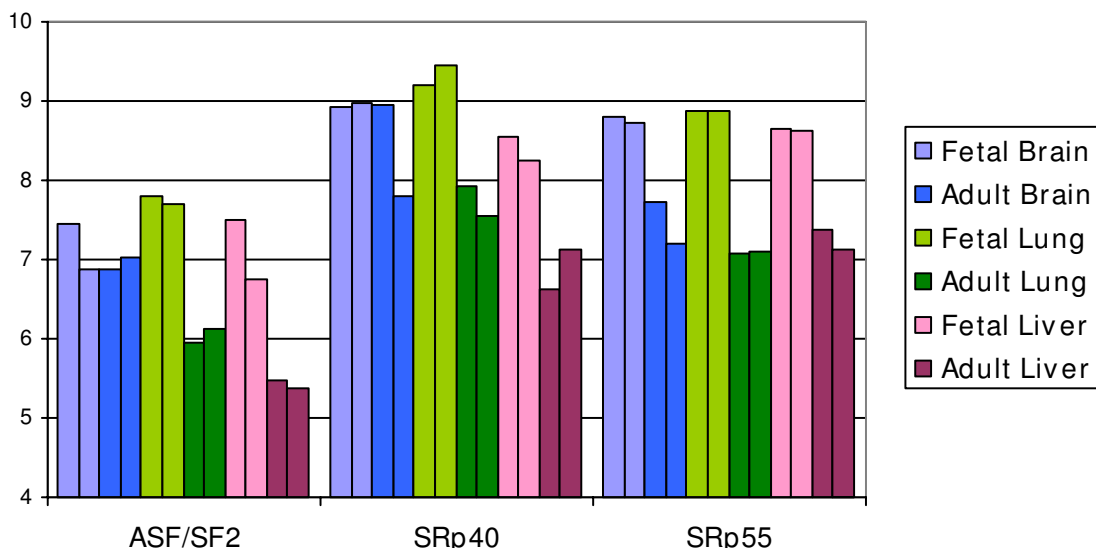
**Table 1: Significant kmers found by the REDUCE algorithm.**

Kmer	Correlation	P-value
cctgg	-0.0451	3.6e-36
aagaa	0.0310	1.8e-16
cgtgg	-0.0233	1.1e-08
gtttt	0.0195	1.0e-05
gggaga	-0.0199	2.1e-05
tatgg	0.0169	7.0e-04
tcaccg	-0.0178	7.0e-04
tcaac	-0.0161	2.2e-03
gaagat	0.0170	2.4e-03
tggagg	-0.0168	3.3e-03
gccgg	-0.0151	8.1e-03

The kmer, the correlation between the kmer counts and exon-skipping score for each exon, and the p-value for the correlation are listed in columns 1, 2 and 3 respectively.

from similar tissues lines as the exon-junction array. The CEL files of the Gene Expression Atlas data were provided by the authors and the data were preprocessed using robust multi-array analysis (RMA) proposed by [53]. We extracted genes encoding the family of serine/arginine (SR) proteins, which contain both RNA-binding and protein domains and are known to facilitate the assembly of the spliceosome by binding to ESEs [54]. Figure S1 in Additional file 3 shows the processed log-intensity values in adult and fetal lung, brain, and liver tissues for SR proteins listed in [20]. In particular, SRp55, SRp40 and ASF/SF2 (Figure 5) clearly show reduced expression in the adult tissues compared to the fetal tissues in at least two of the three examined tissue types.

These results imply that some SR proteins may be involved in a general developmental response. To explore this hypothesis, we further investigated the exon-junction array, which has extensive results from different tissues (*e.g.*, liver), developmental stages (*e.g.*, fetal) and disease cell lines (*e.g.*, lung carcinoma). Therefore, it is possible to identify AS involved in specific biological process by examining splice patterns between tissue pairs. For



**Figure 5**  
**Differences in log-intensities (y-axis) between fetal and adult tissues for three SR proteins.** There are two replicates for each fetal and adult tissue, which are labeled with the same color shade. Tissues (brain, lung and liver) are labeled in different colors.

instance, genes that show different splicing profiles in fetal versus adult brain tissues may be indicative of functional roles in the regulation of brain development through splicing control. By comparing fetal versus adult tissues for several different organs, we adapted our gene-level exon-skipping score to obtain alternatively spliced genes that are associated with development.

Within one gene, for each tissue  $i$  we have the product score  $r_{ij}$  across exons  $j$ . We are interested in genes where there are large differences in  $r_{ij}$  for pairs of tissues (e.g., fetal lung versus adult lung). Define  $d(i, i') = r_{ij} - r_{i'j}$  as the difference between two tissues,  $i$  and  $i'$ , in exon  $j$ . Extreme values (both positive and negative) for  $d$  indicate differences in exon-skipping levels between the tissues for that exon. We calculated a gene-level tissue pair difference score as the KS-statistic of  $d$  across all exons relative to a simulated null model assuming normally distributed residuals  $r_{ij}$  to gauge the differences in exon-skipping patterns between two tissues. Genes with Bonferroni-adjusted p-value  $< 0.05$  were selected as candidate genes for development-related alternative exon-skipping in brain, liver, and lung tissues (Table 2). Many of our predicted genes are known to have multiple splice variants which are regulated during development, including neurofibromin 1 (NF1) in the brain for the RAS signal transduction pathway [55,56], L1 cell adhesion molecule (L1CAM) for nervous system development [57] and ion

transport pump gene ATPase isoform 4 (PMCA4) [58] and fibroblast growth factor receptor 1 (FGFR1) for cell division in brain development [59,60].

To systematically assess the functional validity of our predictions, we analyzed the functional annotation of the predicted genes in the literature. Tissue-specific keywords were searched in the abstracts listed in the PubMed entries for each gene (see Methods). For example, for our lung-developmental genes (column 1 in Table 2), we searched for the keyword "lung" in each abstract, and a p-value was derived for tissue-specific keyword enrichment compared to all genes on the exon-junction arrays using Fisher's exact test (Table 2). We also specifically searched for a splicing-related keyword to check that the tissue annotations are not only due to transcription related information. In summary, all our predicted tissue-specific gene sets showed statistically significant ( $p < .05$ ) enrichment for the respective tissue-specific keyword, splicing-related keyword and for a combination of tissue-, splicing- and development-related keywords. A separate analysis of the GO terms in the 127 predicted brain-development associated genes with the "brain" keyword in PubMed showed functional enrichment in 26 GO terms using Gostat (Benjamini-corrected false discovery rate  $< 0.01$ ). The most common GO terms (11 out of 26) were those with functions related to signaling and ion transport (Additional file 1, Table S12), which are relevant for synaptic trans-



**Table 2: Development-related alternatively spliced genes.**

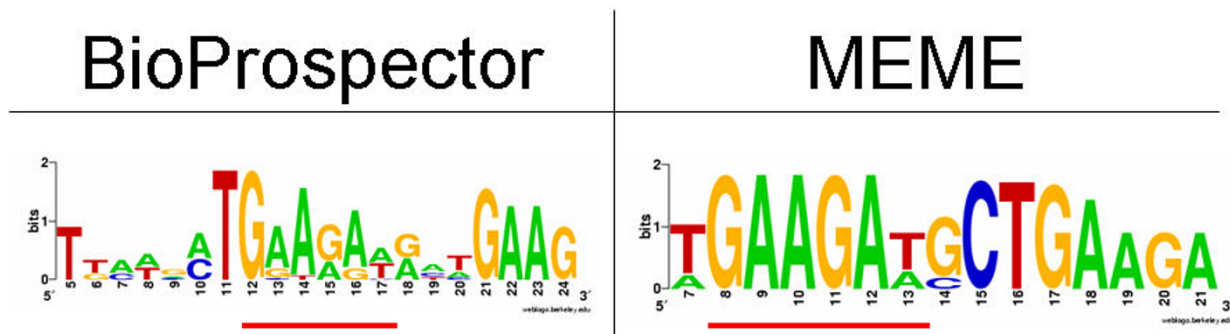
	Lung	Brain	Liver
Significant genes	395	631	482
Genes with PubMed tissue keyword	27 (7%)	127 (20%)	52 (11%)
% of all genes with keyword	5%	11%	7%
p-value	3.57e-02	1.04e-12	9.62e-04
Genes with "splicing" keywords	92 (23%)	135 (21%)	93 (19%)
% of all genes with keywords	12%	12%	12%
p-value	5.46e-11	1.26e-12	8.01e-07
Genes with PubMed tissue, "development" or "splicing" keywords	135 (34%)	258 (41%)	160 (33%)
% of all genes with keywords	23%	26%	25%
p-value	3.14e-07	4.57e-16	2.21e-05

A keyword hit is defined as a gene that has at least 2 abstracts with the keyword. The numbers in parentheses indicate the percentage of genes with a keyword hit within that gene set. Tissue keyword is the tissue name.

mission and neurogenesis. In summary, the two annotation analyses collectively support the validity of our predicted genes.

To identify potential cis-regulatory elements involved in splicing regulation for development, we applied *de novo* motif finding algorithms MEME [61] and BioProspector [62]. To keep the number of exons manageable for these algorithms, only genes predicted to undergo development-related AS for all three tissues and had a "splicing" or "development" keyword hit from the PubMed abstract search were considered. Applying the motif algorithms to alternatively spliced exons within these genes (see Methods and Additional file 1, Table S13), we repeatedly found a "GAAGAA" motif (Figure 6), which is very similar to the AE motif retrieved by our previous analyses and suggests that this motif and its binding factors are also involved in

the developmental regulation of splicing. Furthermore, since we previously found that this motif was more strongly associated with weak splice sites rather than strong splice sites, these results also suggest that the mode of regulation for splicing during development may be due to the combination of a weak splice site and the presence of this enhancer. However, the appearance of the "GAA-GAA" motif in exons predicted to undergo development-related AS may just be an artifact of this motif appearing in AS exons in general. To investigate this scenario, we checked whether the "GAAGAA" motif appeared in the development-related AS exons more frequently than all AS exons (AE from above). We counted the occurrence of the motif in both sets of exons and found that indeed the occurrence of the motif was significantly higher than expected in the development-related AS exons (p-value =



**Figure 6**  
**Motifs found in development-related exons.** Results are based on two motif-finding algorithms, BioProspector and MEME (see Methods). The common conserved sequence is underlined.

.02, binomial test) based on the frequency of the motif in general AS exons.

Experimentally verified binding sites in their natural context and from SELEX experiments are listed in Additional file 1, Table S14 for the three SR proteins that we identified based on the expression data as possibly involved in general developmental roles (SRp55, SRp40 and ASF/SF2). Most of these sequence elements contain the "GAA" sequence and are similar to the motifs discovered by our motif search in the development set of AS exons (Figure 5). The known binding sites, expression data, annotation results, and motif analysis provide consistent evidence supporting the hypothesis that these proteins may be involved in the regulation of alternative splicing during development.

### Conclusion

We have developed a novel approach for the analysis of exon-junction arrays that will specifically search for exon-skipping alternative splice events. Our proposed score evaluates the product of expression levels across exon-junctions to more accurately reflect exon-skipping. Our score also accounts for overall expression levels for a gene and uses an improved variance stabilizing correction. We have shown that the combination of these approaches improves the discrimination of positive and negative control sets determined from independent data sources. We could not, however, directly compare our method with existing methods in [6] because that analysis was based on individual array replicates, which we could not access. Nevertheless, utilizing another source of external validation, we demonstrate that the annotations for our alternatively spliced gene predictions were consistent with previous literature. There are several other exon-junction studies, but our method was not applicable because they were either disease specific or a detailed examination of a relatively smaller number of genes [39,63,64].

Following the array analysis, we examined our sequence predictions for dual purposes; to validate our analysis of the exon-junction array and to predict novel splicing enhancers and silencers. The results from the sequence analysis provide a further source of validation for the quality of our alternative and constitutive exon predictions. Using randomization trials, we found that our predictions were enriched in sequences that were not contained in random sets of sequences. We also identified several sequence signals that are consistent with the experimental literature (*e.g.*, GGG triplets and pyrimidine-rich motifs in introns) and identified several motifs that are highly specific. For example, the known exonic sequence enhancer CACC was discovered in the constitutive exons but not in the alternative exons. Another known enhancer, GAAGAA, appears to be more specific to alternative exons than constitutive exons. Furthermore, we

found that the occurrence of GAAGAA is biased towards exons with weak splice sites, while the other identified motifs associated with alternatively skipped exons do not have splice strength specificity. The sequence analysis also shows that, although originally developed for gene expression data [50], correlation-based methods utilizing whole genome data like REDUCE are applicable to splice arrays and corroborate the kmer enrichment analysis without relying on pre-determined cutoffs.

By our definition, the alternative exons show patterns of exon skipping among different tissues. The presence of known exonic enhancers in these sequences supports the hypothesis that depending on tissue-specific expression, the corresponding binding factors are enhancing the splicing of these exons, which would otherwise be skipped. An alternative hypothesis, not supported by these results, but also discussed in the literature [10], is that silencer sequences in the exon or flanking introns prevent proper splicing and are responsible for exon-skipping. These hypotheses are not mutually exclusive because both enhancer and silencer sequences may function cooperatively in splicing regulation. Furthermore, some sequence elements have been shown to act as both enhancers and silencers [65].

A useful feature of the Rosetta array is that both fetal and adult samples were included for three different tissues. We used this to develop a method for predicting genes with pairwise tissue differences in exon-skipping and applied our procedure to explore both the *cis* and *trans*-regulation of splicing during development. Our method for pairwise tissue analysis is not limited to the developmental comparison but can be extended to other tissues or cell lines (*e.g.*, cancer versus normal cell lines). We made gene predictions for three tissue types (lung, brain and liver) but only looked at the intersection to focus on general developmental alternative splicing, which was motivated by observed expression changes in specific SR proteins. Our final predictions for development-related alternative splicing were consistent with functional annotation and literature searches. In our sequence analysis of the development-related predictions, we found that a form of the GAAGAA motif may also have a role in alternative splicing during development. Furthermore, the changes in gene expression between fetal and adult tissues for several SR proteins that bind to this motif provide further evidence for their roles in developmental regulation.

Because of the array design, our work focuses on the detection of exon-skipping. However, these data can also be used to predict and analyze other types of alternative splice events. In particular, if alternative splice site selection or intron retention occurs, we would observe variation in a single exon-junction across tissues, but not necessarily in consecutive junctions, as observed with

exon-skipping. Therefore, a product summarization step would not be necessary, but we could adapt our procedure in other ways to predict alternative splice selection and intron retention. We will, however, not be able to discriminate between these other types of splice events because of the nature of the array design.

A recent direction taken by several groups is to use sequence conservation across multiple species to aid in the search for enhancers and silencers [22-25,66,67]. Comparative approaches have also been used to examine the conservation of alternative splice events by comparing genomic or EST data from multiple species [29,68-71]. As more splice array data become available from different organisms and tissues, there may also be opportunities to explore the conservation of splicing events from meta-analysis of splice arrays, without relying on ESTs from orthologous genes that are often limited across species.

**Methods**

**Exon-skipping scores**

Within each gene, the alternative splice (AS) effect for every probe *j* and tissue *i* is estimated by the residual  $r_{i,j,k}$  from a linear model fit. To identify whether exon-skipping has occurred in some tissues, we considered four exon-skipping scores that measure the deviations of the residuals across tissues. All calculations described in this and all other Methods sections are performed in the software R <http://www.r-project.org/> unless otherwise noted.

One score is based on the two-sample Kolmogorov-Smirnov (KS) test statistic of the residual products of adjacent exon-junction (EJ) probes spanning exons *j*,  $p_{i,j,k} = r_{i,j,k}r_{i,j+1,k}$ . For the exon-level score, residual products across tissues *i* for exon *j* are examined, and for the gene-level score, the residual products across tissues *i* and exons *j* for gene *k* are examined. To obtain p-values based on the two-sample KS test, we compared the observed residual products with randomly sampled residual products from the null hypothesis that they are a product of normally distributed random variables with mean 0,  $r_{i,j,k} \sim N(0, \sigma^2)$ , where  $\sigma$  is sampled from the empirical distribution for all genes.

Exploratory data analysis showed greater variability for genes with low overall expression signal and thus we considered a more appropriate exon-skipping score, the two-sample weighted Kolmogorov-Smirnov (WKS =  $w_kKS_k$ ) test statistic. The linear weight function is constructed to adjust for the overall expression signal strength and results in increasing the weight of more highly expressed genes. It is defined as  $w_k = a_k / (\max_k a_k - \min_k a_k)$  with

$a_k = \max_{tissue\ i} \min_{exon\ j} \gamma_{i,j,k}$ , where  $\gamma_{i,j,k}$  is the average background subtracted intensity. To obtain p-values, observed values

of WKS were compared with randomly sampled WKS statistics based on the product of randomly sampled weights, from a uniform distribution within the empirical range of calculated weights for all genes, and randomly sampled KS statistics, sampled from the theoretical distribution of the KS statistic.

Another score is based on the coefficient of variation (CV), a statistic that measures large variations and is defined as the standard deviation divided by the mean. For exon-level score, we calculated the CV of the residual products across tissues,  $CV_{tissue\ i}(p_{i,j,k})$  and for the gene-level score, we calculated the standard deviation across exons within the gene,  $sd_{exon\ j} [CV_{tissue\ i}(p_{i,j,k})]$ . This method is closest to the Rosetta approach [6] which relies on a statistic similar to the coefficient of variation.

Finally, we considered a score based on a direct test for normality using the Wilk-Shapiro test (WS) [46]. In this case we did not use the product summarization, but tested the following hypothesis using the WS test:  $r_{i,j,k} \sim X$  where *X* is distributed as  $N(0, \sigma^2)$ . As before, the tests are performed across exons or genes to obtain the exon- or gene-level scores respectively.

**Controls**

A collection of sample Affymetrix Human Exon 1.0 ST array data (from [72]) was used to generate an independent set of negative controls. These arrays represent a collection of 11 tissues with 3 replicates each. All of the 11 tissues were included in the Rosetta data set [6] except breast. Identifying gene sets from these arrays that demonstrate no changes across all 11 tissues provide an independent source of experimental evidence for constitutive exons. The data are background-corrected, quantile-normalized and summarized at the exon level using the Robust Multi-Array Average (RMA) method [73] to generate a logarithm base 2 RMA value,  $x_{i,j,k}$  for each exon probe set. To identify genes that are likely to be constitutively spliced, we rank genes according to their expression variability across the 11 tissues and across corresponding core exon probesets,  $affy_k = range_{exon\ j} range_{tissue\ i} x_{i,j,k}$ . The top 5% least variable RefSeq genes that are both present in the Affymetrix exon arrays and Rosetta exon-junction arrays (*n* = 282) were selected as our negative controls.

The positive control (AS) consists of 164 EST-supported alternatively spliced genes that were placed on the arrays in Le et al., 2004 [42].

### Constitutive and alternative exons

To determine a set of alternatively spliced genes, we used a cutoff for the gene-level WKS-statistic that produced a true and false positive rate of ~70% and ~25% respectively on the controls. This cutoff resulted in 3912 predicted alternatively spliced genes. Of the exons in these genes, we then extracted exons in the 75<sup>th</sup> percentile according to their exon-level KS-statistic, which resulted in 9178 predicted exons. The KS-statistic is used for the exon-level score because the WKS-statistic has a gene-level correction for expression variability that is the same for all exons in a gene. Using the UCSC Genome Browser collection of RefSeq sequences (July 2003), 8433 exons were identified along with their neighboring introns ( $n = 13813$ ) by sequence matching. Multiple occurrences of neighboring introns were removed. For the GO term analysis, we used a more stringent cutoff and only examined genes in the 95<sup>th</sup> percentile according to their WKS-statistic ( $n = 431$ ). GO term enrichment was determined using the Benjamini-Hochberg correction for controlling the false discovery rate at the .001 level.

To obtain a similar sized set of constitutive genes and exons so that the sequence comparisons would be balanced, we evaluated the lowest ~15<sup>th</sup> percentile of genes according to the gene-level WKS-statistic and extracted the exons within these genes that scored below the median exon-level KS-statistic. Using the UCSC Genome Browser collection of RefSeq sequences, 8113 of these exons were identified along with their neighboring introns ( $n = 10081$ ). Multiple occurrences of neighboring introns were removed.

### Motif analysis

We performed a search for over-represented kmers in our exonic and intronic sequences using the RESCUE-ESE method as described in [12]. Briefly, using the Perl programming language, counts for all kmers between two sets of sequences are tallied and evaluated for significant over-representation between the sets using a Z-test, and all kmers with Bonferroni-adjusted  $p$ -value  $< 0.01$  were clustered and aligned into groups of motifs. Distances between motifs were defined as the absolute deviation between the nucleotide frequencies of two aligned positions in the best local alignment. Motif pairs with distance less than 2 were deemed as a motif match. For kmer calculations with exons, the comparisons are 1) exons versus introns and 2) constitutive versus alternative exons. We also considered two comparisons for each of 5' and 3' SS based on splice site strength scored by the position-specific weight matrices and applied the 1st and 3rd quartiles of all known splice sites as the cutoffs for weak and strong splice sites: 3) weak versus strong splice site exons and 4) strong versus weak splice site exons. For kmer calculations with introns, the comparisons are 1) introns versus exons and 2) flanking introns of constitutive versus alternative

exons; 3) flanking introns of weak versus strong splice site exons and 4) flanking introns of strong versus weak splice site exons. All sequence analyses used intronic/exonic sequences up to 200 bp from the SS and excluded the splice site regions that cover the first 5 bp at each end of exons and the first 20 bp and 10 bp for the 3' and 5' ends of introns respectively.

In addition, we searched for novel sequence motifs utilizing the whole data set without the need to dichotomize exons into AE and CE by applying the correlation-based method REDUCE [50] to the splice site proximal sequences and our exon-level score for measuring exon skipping ( $n = 83789$ ). REDUCE enumerates all possible 5–6 base pair kmers and finds those that show significant correlation with expression values from a single microarray experiment. Multiple kmers are identified by iteratively removing the significant kmers from sequences and re-evaluating the correlation between the remaining kmer frequencies and residuals of the linear regression fit from the previous run (*i.e.*, subtracting the contribution from the previously selected kmer(s)). We masked the splice sites by removing 5 bp at the 5' and 3' end of each exon. The output lists all significant kmers with  $p$ -values adjusted using the Bonferroni correction.

To look for motifs in the development-related AS exon sets, we examined the genes that had appropriate keyword hits (see below). Within these genes, sequences were extracted for exons predicted to be AE based on our exon-skipping score. De novo motif finders MEME [61] and BioProspector [62] were applied to these selected exon sequences subtracting the first 5 bp at either end of each exon. BioProspector was run with the following options: only examine the forward stand (-d), assume every sequence has a motif (-a 1) and motif width range from 6 to 18 (-w 6,8,10,12,14,16). MEME was run with the following options: default value of only looking at the forward strand, DNA sequence (-dna), several choices for the expected number of motif occurrences (zero or one in each sequence -zoops, one in each sequence -oops, variable number -tcm) and the motif width (-w 6,8,10,12, and the range -minw 6 -maxw 20). For each of the 15 MEME runs and the top three motifs from the 6 BioProspector runs, we aligned the consensus sequences for the final predicted motifs with ClustalW [74]. All motifs and alignments are displayed in the Figures using WebLogo [75]

### PubMed keyword search

For all genes in the Genome Browser RefSeq transcript list, the software R <http://www.r-project.org/>, with the "annotate" and "XML" packages was used to search each abstract of the references listed in the PubMed report for this gene. A hit for a gene was defined if the keyword(s) occurred in at least two of the abstracts for the gene. For each keyword, we performed a Fisher's exact test to evaluate whether the

keyword appeared more frequently than expected by chance in our gene set compared to all genes. For the *splicing* keyword we searched for either "splicing" or "splice".

### Abbreviations

AS: alternative splice; AE: alternatively spliced exons; CE: constitutively spliced exons; CV: coefficient of variation; EJ: exon-junction; ESE: exonic splicing enhancers; KS: Kolmogorov-Smirnov; WS: Wilk-Shapiro; SS: splice site; ROC: receiver operating characteristic.

### Authors' contributions

KK and YHY performed the analysis, contributed to the design of the study and wrote the manuscript. RFY conceived the design of the study, coordinated the work and edited the manuscript. All authors read and approved the final manuscript.

### Additional material

#### Additional file 1

*This file contains supporting tables: Table S1: Significant GO terms for top scoring alternatively-spliced genes. Table S2: Motifs found in alternative exons within 200 bp of the 5'SS. Table S3: Motifs found in alternative exons within 200 bp of the 3'SS. Table S4: Motifs found in alternative exons within 200 bp of the 5'SS and 3'SS. (weak splice site strength comparison). Table S5: Motifs found in constitutive exons within 200 bp of the 5'SS. Table S6: Motifs found in constitutive exons within 200 bp of the 3'SS. Table S7: Motifs found within 200 bp of the 5'SS of 3' introns flanking alternative exons. Table S8: Motifs found within 200 bp of the 3'SS of 5' introns flanking alternative exons. Table S9: Motifs found within 200 bp of the 5'SS of 3' introns flanking constitutive exons. Table S10: Motifs found within 200 bp of the 3'SS of 5' introns flanking constitutive exons. Table S11: Number of significant kmers in predicted exons and flanking introns and randomly selected exons and flanking introns. Table S12: Significant GO terms for genes with development-associated alternative splicing in brain. Table S13: List of exons predicted to undergo development-related AS. Table S14: Binding sites for SR proteins.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-551-S1.xls>]

#### Additional file 2

*This file contains additional results and methods.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-551-S2.doc>]

#### Additional file 3

*This file contains one supporting figure: Figure S1: Differences in log-intensities (y-axis) between fetal and adult tissues for SR proteins listed in [20]. There are two replicates for each fetal and adult tissue, which are labeled with the same color. Tissues (brain, lung and liver) are labeled in different colors.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-551-S3.doc>]

### Acknowledgements

This research was supported by the Center for Bioinformatics and Molecular Biostatistics at University of California, San Francisco and a Sloan Foundation/DOE post-doctoral fellowship to KK.

### References

- Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci USA* 2001, **98(20)**:11193-11198.
- Lopez AJ: **Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation.** *Annu Rev Genet* 1998, **32**:279-305.
- Blencowe BJ: **Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.** *Trends Biochem Sci* 2000, **25(3)**:106-110.
- Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem* 2003, **72**:291-336.
- Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29(13)**:2850-2859.
- Johnson JM, Castle J, Garrett-Engel P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302(5653)**:2141-2144.
- Lander ES, Linton LM, Birren B, Nussbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822)**:860-921.
- Coulter LR, Landree MA, Cooper TA: **Identification of a new class of exonic splicing enhancers by in vivo selection.** *Mol Cell Biol* 1997, **17(4)**:2143-2150.
- McCullough AJ, Berget SM: **G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.** *Mol Cell Biol* 1997, **17(8)**:4562-4571.
- Fairbrother WG, Chasin LA: **Human genomic sequences that inhibit splicing.** *Mol Cell Biol* 2000, **20(18)**:6816-6825.
- Schaal TD, Maniatis T: **Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences.** *Mol Cell Biol* 1999, **19(3)**:1705-1719.
- Fairbrother WG, Yeh RF, Sharp PA, Burge CB: **Predictive identification of exonic splicing enhancers in human genes.** *Science* 2002, **297(5583)**:1007-1013.
- Sironi M, Menozzi G, Riva L, Cagliari R, Comi GP, Bresolin N, Giorda R, Pozzoli U: **Silencer elements as possible inhibitors of pseudoexon splicing.** *Nucleic Acids Res* 2004, **32(5)**:1783-1791.
- Zhang XH, Chasin LA: **Computational definition of sequence motifs governing constitutive exon splicing.** *Genes Dev* 2004, **18(11)**:1241-1250.
- Zhang XH, Heller KA, Hefter I, Leslie CS, Chasin LA: **Sequence information for the splicing of human pre-mRNA identified by support vector machine classification.** *Genome Res* 2003, **13(12)**:2637-2650.
- Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, Conboy JG: **Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing.** *Nucleic Acids Res* 2001, **29(11)**:2338-2348.
- Clark F, Thanaraj TA: **Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.** *Hum Mol Genet* 2002, **11(4)**:451-464.
- Miriami E, Margalit H, Sperling R: **Conserved sequence elements associated with exon skipping.** *Nucleic Acids Res* 2003, **31(7)**:1974-1983.
- Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T: **Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome.** *Genome Res* 2003, **13(6B)**:1290-1300.
- Yeo G, Holte D, Kreiman G, Burge CB: **Variation in alternative splicing across human tissues.** *Genome Biol* 2004, **5(10)**:R74.
- Lian Y, Garner HR: **Evidence for the regulation of alternative splicing via complementary DNA sequence repeats.** *Bioinformatics* 2005, **21(8)**:1358-1364.

22. Yeo GW, Nostrand EL, Liang TY: **Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements.** *PLoS Genet* 2007, **3(5)**:e85.
23. Sorek R, Ast G: **Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.** *Genome Res* 2003, **13(7)**:1631-1637.
24. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G: **Comparative analysis identifies exonic splicing regulatory sequences – The complex definition of enhancers and silencers.** *Mol Cell* 2006, **22(6)**:769-781.
25. Itoh H, Washio T, Tomita M: **Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes.** *Rna* 2004, **10(7)**:1005-1018.
26. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB: **Inference of splicing regulatory activities by sequence neighborhood analysis.** *PLoS Genet* 2006, **2(11)**:e191.
27. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9(12)**:1288-1293.
28. Kan Z, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** *Genome Res* 2001, **11(5)**:889-900.
29. Modrek B, Lee CJ: **Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss.** *Nat Genet* 2003, **34(2)**:177-180.
30. Clark TA, Sugnet CW, Ares M Jr: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays.** *Science* 2002, **296(5569)**:907-910.
31. Srinivasan K, Shiue L, Hayes JD, Centers R, Fitzwater S, Loewen R, Edmondson LR, Bryant J, Smith M, Rommelfanger C, et al.: **Detection and measurement of alternative splicing using splicing-sensitive microarrays.** *Methods* 2005, **37(4)**:345-359.
32. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, et al.: **Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform.** *Mol Cell* 2004, **16(6)**:929-941.
33. Shai O, Morris QD, Blencowe BJ, Frey BJ: **Inferring global levels of alternative splicing isoforms using a generative model of microarray data.** *Bioinformatics* 2006, **22(5)**:606-613.
34. Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, Wang H, Williams A, Kulp D, Blume JE, Haussler D, et al.: **Unusual intron conservation near tissue-regulated exons found by splicing microarrays.** *PLoS Comput Biol* 2006, **2(1)**:e4.
35. Cuperlovic-Culf M, Belacel N, Culf AS, Ouellette RJ: **Microarray analysis of alternative splicing.** *Omnics* 2006, **10(3)**:344-357.
36. Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ: **Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression.** *Genes Dev* 2006, **20(2)**:153-158.
37. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE: **Discovery of tissue-specific exons using comprehensive human exon microarrays.** *Genome Biol* 2007, **8(4)**:R64.
38. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, et al.: **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7**:325.
39. Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, et al.: **Nova regulates brain-specific splicing to shape the synapse.** *Nat Genet* 2005, **37(8)**:844-852.
40. Fehlbaum P, Guihal C, Bracco L, Cochet O: **A microarray configuration to quantify expression levels and relative abundance of splice variants.** *Nucleic Acids Res* 2005, **33(5)**:e47.
41. Heinzen EL, Yoon W, Weale ME, Sen A, Wood NW, Burke JR, Welsh-Bohmer KA, Hulette CM, Sisodiya SM, Goldstein DB: **Alternative ion channel splicing in mesial temporal lobe epilepsy and Alzheimer's disease.** *Genome Biol* 2007, **8(3)**:R32.
42. Le K, Mitsouras K, Roy M, Wang Q, Xu Q, Nelson SF, Lee C: **Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data.** *Nucleic Acids Res* 2004, **32(22)**:e180.
43. Hu GK, Madore SJ, Moldover B, Jatkoa T, Balaban D, Thomas J, Wang Y: **Predicting splice variant from DNA chip expression data.** *Genome Res* 2001, **11(7)**:1237-1245.
44. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl 1)**:S96-104.
45. Cline MS, Blume J, Cawley S, Clark TA, Hu JS, Lu G, Salomonis N, Wang H, Williams A: **ANOSVA: a statistical method for detecting splice variation from expression data.** *Bioinformatics* 2005, **21(Suppl 1)**:i107-115.
46. Shapiro SS, Wilk MB: **An analysis of variance test for normality (complete samples).** *Biometrika* 1965, **52**:591-611.
47. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res* 2004, **32(suppl\_1)**:D493-496.
48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1)**:25-29.
49. **GOstat** [<http://gostat.wehi.edu.au/>]
50. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27(2)**:167-171.
51. Hofmann Y, Lorson CL, Stamm S, Androphy EJ, Wirth B: **Htra2-beta I stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2).** *Proc Natl Acad Sci USA* 2000, **97(17)**:9618-9623.
52. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al.: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101(16)**:6062-6067.
53. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31(4)**:e15.
54. Hastings ML, Krainer AR: **Pre-mRNA splicing in the new millennium.** *Curr Opin Cell Biol* 2001, **13(3)**:302-309.
55. Vandenbroucke I, Vandesompele J, De Paepe A, Messiaen L: **Quantification of NFI transcripts reveals novel highly expressed splice variants.** *FEBS Lett* 2002, **522(1-3)**:71-76.
56. Danglot G, Regnier V, Fauvet D, Vassal G, Kujas M, Bernheim A: **Neurofibromatosis 1 (NFI) mRNAs expressed in the central nervous system are differentially spliced in the 5' part of the gene.** *Hum Mol Genet* 1995, **4(5)**:915-920.
57. Jouet M, Rosenthal A, Kenwrick S: **Exon 2 of the gene for neural cell adhesion molecule 1I is alternatively spliced in B cells.** *Brain Res Mol Brain Res* 1995, **30(2)**:378-380.
58. Strehler EE, James P, Fischer R, Heim R, Vorherr T, Filoteo AG, Peniston JT, Carafoli E: **Peptide sequence analysis and molecular cloning reveal two calcium pump isoforms in the human erythrocyte membrane.** *J Biol Chem* 1990, **265(5)**:2835-2842.
59. Fu L, Abu-Khalil A, Morrison RS, Geschwind DH, Kornblum HI: **Expression patterns of epidermal growth factor receptor and fibroblast growth factor receptor 1 mRNA in fetal human brain.** *J Comp Neurol* 2003, **462(2)**:265-273.
60. Jiao J, Greendorfer JS, Zhang P, Zinn KR, Diglio CA, Thompson JA: **Alternatively spliced FGFR-1 isoform signaling differentially modulates endothelial cell responses to peroxynitrite.** *Arch Biochem Biophys* 2003, **410(2)**:187-200.
61. Bailey T, Elkan C: **Learning of Multiple Motifs in Biopolymers using EM.** *Machine Learning* 1995, **21(1-2)**:51-80.
62. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001:127-138.
63. Zhang C, Li HR, Fan JB, Wang-Rodriguez J, Downs T, Fu XD, Zhang MQ: **Profiling alternatively spliced mRNA isoforms for prostate cancer classification.** *BMC Bioinformatics* 2006, **7**:202.
64. Nagao K, Togawa N, Fujii K, Uchikawa H, Kohno Y, Yamada M, Miyashita T: **Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays.** *Hum Mol Genet* 2005, **14(22)**:3379-3388.
65. Ladd AN, Cooper TA: **Finding signals that regulate alternative splicing in the post-genomic era.** *Genome Biol* 2002, **3(11)**:reviews0008.
66. Minovitsky S, Gee SL, Schokrpur S, Dubchak I, Conboy JG: **The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons.** *Nucleic Acids Res* 2005, **33(2)**:714-724.

67. Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S, Poliakov A, Dubchak I, Blume JE, et al.: **A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing.** *Nucleic Acids Res* 2007, **35(14)**:4845-4857.
68. Thanaraj TA, Clark F, Muilu J: **Conservation of human alternative splice events in mouse.** *Nucleic Acids Res* 2003, **31(10)**:2544-2552.
69. Sorek R, Shamir R, Ast G: **How prevalent is functional alternative splicing in the human genome?** *Trends Genet* 2004, **20(2)**:68-71.
70. Yeo GW, van Nostrand E, Holste D, Poggio T, Burge CB: **Identification and analysis of alternative splicing events conserved in human and mouse.** *Proc Natl Acad Sci USA* 2005, **102(8)**:2850-2855.
71. Ohler U, Shomron N, Burge CB: **Recognition of unknown conserved alternatively spliced exons.** *PLoS Comput Biol* 2005, **1(2)**:113-122.
72. **Exon 1.0 ST Array Sample Data** [[http://www.affymetrix.com/support/technical/sample\\_data/exon\\_array\\_data.affx](http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx)]
73. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2)**:249-264.
74. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22)**:4673-4680.
75. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14(6)**:1188-1190.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

