

Research

Studying alternative splicing regulatory networks through partial correlation analysis

Liang Chen* and Sika Zheng†

Addresses: *Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA. †Howard Hughes Medical Institute, University of California, Los Angeles, MRL 6-619, Los Angeles, California 90095, USA.

Correspondence: Liang Chen. Email: liang.chen@usc.edu

Published: 9 January 2009

Genome Biology 2009, **10**:R3 (doi:10.1186/gb-2009-10-1-r3)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2009/10/1/R3>

Received: 19 November 2008

Revised: 18 December 2008

Accepted: 9 January 2009

© 2009 Chen and Zheng; 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 pre-mRNA splicing is an important gene regulation mechanism for expanding proteomic diversity in higher eukaryotes. Each splicing regulator can potentially influence a large group of alternative exons. Meanwhile, each alternative exon is controlled by multiple splicing regulators. The rapid accumulation of high-throughput data provides us with a unique opportunity to study the complicated alternative splicing regulatory network.

Results: We propose the use of partial correlation analysis to identify association links between exons and their upstream regulators or their downstream target genes (exon-gene links) and links between co-spliced exons (exon-exon links). The partial correlation analysis avoids taking the ratio of two noisy random variables, exon expression level and gene expression level, so that it achieves a higher statistical power. We named this analysis procedure pCastNet (partial Correlation analysis of splicing transcriptome Network). Through studies of known alternative exons, conservation patterns, relative positions, functional annotations, and RT-PCR experiments, we concluded that pCastNet can effectively identify exon-gene or exon-exon links. We further found that gene pairs with exon-gene or exon-exon links tend to have similar functions or are present in the same pathways. More interestingly, gene pairs with exon-gene or exon-exon links tend to share *cis*-elements in promoter regions and microRNA binding elements in 3' untranslated regions, which suggests the coupling of co-alternative-splicing, co-transcription-factor-binding, and co-microRNA-binding.

Conclusions: Alternative splicing regulatory networks reconstructed by pCastNet can help us better understand the coordinate and combinatorial nature of alternative splicing regulation. The proposed tool can be readily applied to other high-throughput data such as transcriptome sequencing data.

Background

Alternative pre-mRNA splicing is an important gene regulation mechanism for expanding proteomic diversity in higher

eukaryotes. It has been estimated that 59-74% of human genes are alternatively spliced [1,2], and abnormal mRNA splicing contributes to many human diseases [3-5]. The alter-

native splicing of multiple pre-mRNAs is tightly regulated and coordinated, which is an essential component for many biological processes, including nervous system development and programmed cell death [6,7]. In the process of alternative splicing, splicing regulators bind to various pre-mRNAs and affect a large number of exons. Meanwhile the splicing pattern of a specific exon is determined by multiple pre-mRNA-binding proteins [8,9]. Therefore, it will be particularly interesting to study how the splicing of a group of exons is co-regulated and how the splicing of an exon is combinatorially controlled by multiple regulators.

With advancements in high-throughput technologies, such as Affymetrix exon arrays, various types of junction arrays, or high-throughput sequencing, it is feasible to study alternative splicing on a genomic scale. Current studies have centered on the differential analysis of alternative splicing. To identify exons with differential splicing, we must account for differential transcription of a gene itself. In Affymetrix exon arrays, both exon-level intensity and gene-level intensity are estimated. Gene-level-normalized exon intensity, which is defined as the ratio of the exon intensity to the gene intensity, has been widely used to remove the transcription effect when studying splicing. A significant difference in the normalized exon intensity (NI) indicates that this exon has different inclusion or exclusion rates between two conditions. For example, in ExACT, developed by Affymetrix [10,11], the NI is calculated as the ratio of the exon intensity to the gene intensity. Then, the 'splicing index' value is calculated by taking the log ratio of the NI in sample 1 to the NI in sample 2 to identify exons alternatively spliced between two samples.

Multiple groups have nicely surveyed the complexities of alternative splicing in various tissues and cell lines and observed tissue-specific alternative splicing events mainly through differential analysis [1,11,12]. These events are valuable for investigating the function of alternative splicing in phenotypic diversity. However, their regulatory interactions remain largely unknown; for example, one can hardly speculate on the relationship or the regulators of two exons co-enriched in a specific tissue. In combination with motif analysis, one can further study motif enrichment in a group of tissue-specific alternative exons [13,14]. However, such analysis is constrained by the limited knowledge of splicing regulators and their *cis*-regulatory motifs. The motifs of some splicing regulators have not yet been identified and some RNA binding proteins have almost identical binding motifs. Except for a few splicing factors (for example, FOX proteins), the degenerative nature of binding motifs of splicing regulators further confounds analysis. Several groups have used microarrays in conjunction with manipulation of splicing regulator expression or crosslinking immunoprecipitation (CLIP) of splicing regulators to identify their indirect or direct targets [15,16]. Such studies provide the most valuable data for dissecting alternative splicing regulation centered on one splicing regulator of interest.

Instead of performing differential analysis, we propose to study alternative splicing regulatory networks based on pairwise co-expression associations of exons and genes across multiple conditions. This can provide a direct association link between two exons or between one exon and one gene. Such association links can be used to infer regulatory or functional relationships between two nodes. In this study, we have used exon array data for human, mouse, and rat across 11 tissues to study alternative splicing regulatory networks. To study the co-splicing patterns of exons, we can intuitively calculate the NI for every exon across multiple conditions and then calculate the correlation between the NIs of two exons. However, the high-level of noise inherent to exon arrays will make the correlation unstable. Indeed, some studies using the NI approach have reported low validation rates (21-56%) for the identification of alternative splicing events [10,17,18]. The possible reason is that the distribution of the ratio of two random variables is often heavy-tailed if the noise level for the two random variables is high [19]. In other words, if the noise level is high, the ratio between the exon intensity and the gene intensity is not stable and it remains a special statistical challenge to derive appropriate test statistics. For example, we considered a constitutive exon and the gene it belongs to. Exon-level and gene-level intensities were simulated according to a bivariate normal distribution. The correlation between the exon-level intensity and the gene-level intensity was set as 0.9 to satisfy that the exon is a constitutive exon. A total of 1,000 expression levels were simulated. As shown in Additional data file 1, when the noise level is high, the NI can be as small as 0.5 or as high as 3 even if the exon is a constitutive exon.

Instead of using the ratio between the exon intensity and the gene intensity, we can perform correlation studies on the exon intensity directly. To remove the transcription effect in the exon intensity, we propose to apply partial correlation analysis. A partial correlation coefficient is the correlation between two variables, with the effects of other variables removed. For example, in order to exclude the possibility that a high exon-exon (EE) correlation is due to either the gene-level association or the association between one exon and the gene that the other exon belongs to, we calculate the partial correlation coefficients between the two exons conditioning on one or two genes. If the partial correlations are still high, we declare that there is an association between the two exons and this association represents a co-splicing relationship. In addition to EE co-splicing links, we also studied exon-gene (EG) links where the high correlation between an exon and a gene is not due to the gene-gene (GG) association. Partial correlation analysis has been applied to gene co-expression network studies [20-22].

In this study we have used exon array data for human, mouse, and rat across 11 tissues. The proposed methods can be readily applied to RNA-Seq data. We want to point out that the co-splicing relationship can be condition-specific. With the rapid

accumulation of high-throughput exon array or RNA-Seq data, we will be able to reconstruct dynamic regulatory networks under different conditions in the near future.

Results

Determining gene-gene, exon-gene and exon-exon links using pCastNet

Three types of associations were considered for a pair of gene: GG, EG, and EE associations. Using pCastNet (partial Correlation analysis of splicing transcriptome Network), the Pearson correlation coefficient for GG associations was calculated between gene 1 (g_1) and gene 2 (g_2) and denoted as $r_{g_1g_2}$. For EG associations, considering an exon (e_1) of gene 1 (g_1) and gene 2 (g_2), as well as the Pearson correlation coefficient $r_{e_1g_2}$, the partial correlation coefficient between e_1 and g_2 conditioning on g_1 was calculated as $r_{e_1g_2 \bullet g_1}$. The partial correlation can be interpreted as the association between e_1 and g_2 after removing the effect of g_1 . If the partial correlation is high, the association between e_1 and g_2 is not due to the correlation between g_1 and g_2 . Otherwise, e_1 can be a constitutive exon of g_1 and the association between e_1 and g_2 is due to the correlation between the two genes. For EE associations, the correlation between an exon (e_1) of gene 1 (g_1) and an exon (e_2) of gene 2 (g_2) was calculated as $r_{e_1e_2}$. We also calculated the partial correlations $r_{e_1e_2 \bullet g_1}$, $r_{e_1e_2 \bullet g_2}$ and $r_{e_1e_2 \bullet g_1g_2}$ to exclude the possibility that the EE correlation is due to the EG or GG correlation. In summary, if the p -value for $r_{g_1g_2}$ is significant, we declared a GG link between gene 1 and gene 2. If the p -values for both $r_{e_1g_2}$ and $r_{e_1g_2 \bullet g_1}$ are significant, we declared an EG link between e_1 and g_2 . This association is not due to GG association. If the p -values for $r_{e_1e_2}$, $r_{e_1e_2 \bullet g_1}$, $r_{e_1e_2 \bullet g_2}$, and $r_{e_1e_2 \bullet g_1g_2}$ are significant, we declared an EE link between the two exons e_1 and e_2 . The association is not due to GG or EG associations.

Simulation studies on the performance of pCastNet

We performed simulation studies to illustrate the relative performance of pCastNet. A total of five genes were considered. Each of them has five constitutive exons and one alternative exon. The five alternative exons have the same inclusion rate relative to their gene levels. Thus, the five alternative exons are co-spliced. Exon intensity data were simulated for ten tissues. Gene-level intensity was estimated as the average intensity of the five constitutive exons for each gene. In pCastNet, the correlation and partial correlations between

each pair of exons belonging to different genes were calculated as $r_{e_1e_2}$, $r_{e_1e_2 \bullet g_1}$, $r_{e_1e_2 \bullet g_2}$, and $r_{e_1e_2 \bullet g_1g_2}$. In the NI-based approach, the correlation between the NI values of each pair of exons across ten tissues was calculated as r . Different p -value thresholds were used to declare whether there is a co-splicing relationship between two exons. Figure 1 shows the ROC (receiver operating characteristic) curves of pCastNet (red) and the NI-based approach (black). Three scenarios were considered: the standard deviation of the exon intensity is 1 (circles), 2 (triangles), or 4 (crosses). pCastNet consistently performed better than the NI-based approach. When the variance of the exon intensity is large (2^2 or 4^2), the power (true positive rate) of pCastNet is almost 50-100% higher than that of the NI-based approach given the same false positive rate. The true positive rates and the false positive rates are the average values across 1,000 simulations for each scenario.

Choice of significance threshold

The choice of significance threshold remains a major challenge for co-expression network studies. Previous studies have typically relied on a data-independent constant correlation threshold. Zhang and Horvath [23] proposed a weighted gene co-expression network approach. They used soft thresholding instead of hard thresholding to better identify GG links. This method needs a scale-free topology criterion to estimate the involved parameters. Other topology-based approaches include clustering coefficient-based threshold selection developed by Elo *et al* [24]. Because there has been little study on the topology of alternative splicing regulatory networks, we avoided topology-based methods and instead propose a false discovery rate (FDR) approach. Specifically, we used the approach proposed by Efron [25] to control the expected FDR conditioning on a dependence effect parameter A . For GG, EG, and EE networks, hypotheses were performed to test the significance of pair-wise correlations. The dependence among hypotheses is largely ignored in traditional FDR control methods [26,27], despite the fact that correlations among hypotheses may be high for genomics studies [28]. In contrast, the conditional false discovery expectation takes the dependence of hypotheses into account and, therefore, achieves a more accurate estimate of FDR. For GG links, t -test statistics ($r_{g_1g_2} \sqrt{(n-2)/(1-r_{g_1g_2}^2)}$) were converted to z -values directly. For EG and EE links, t -test statistics ($r_{e_1g_2} \sqrt{(n-2)/(1-r_{e_1g_2}^2)}$, $r_{e_1g_2 \bullet g_1} \sqrt{(n-3)/(1-r_{e_1g_2 \bullet g_1}^2)}$, $r_{e_1e_2} \sqrt{(n-2)/(1-r_{e_1e_2}^2)}$, $r_{e_1e_2 \bullet g_1} \sqrt{(n-3)/(1-r_{e_1e_2 \bullet g_1}^2)}$, $r_{e_1e_2 \bullet g_2} \sqrt{(n-3)/(1-r_{e_1e_2 \bullet g_2}^2)}$, and

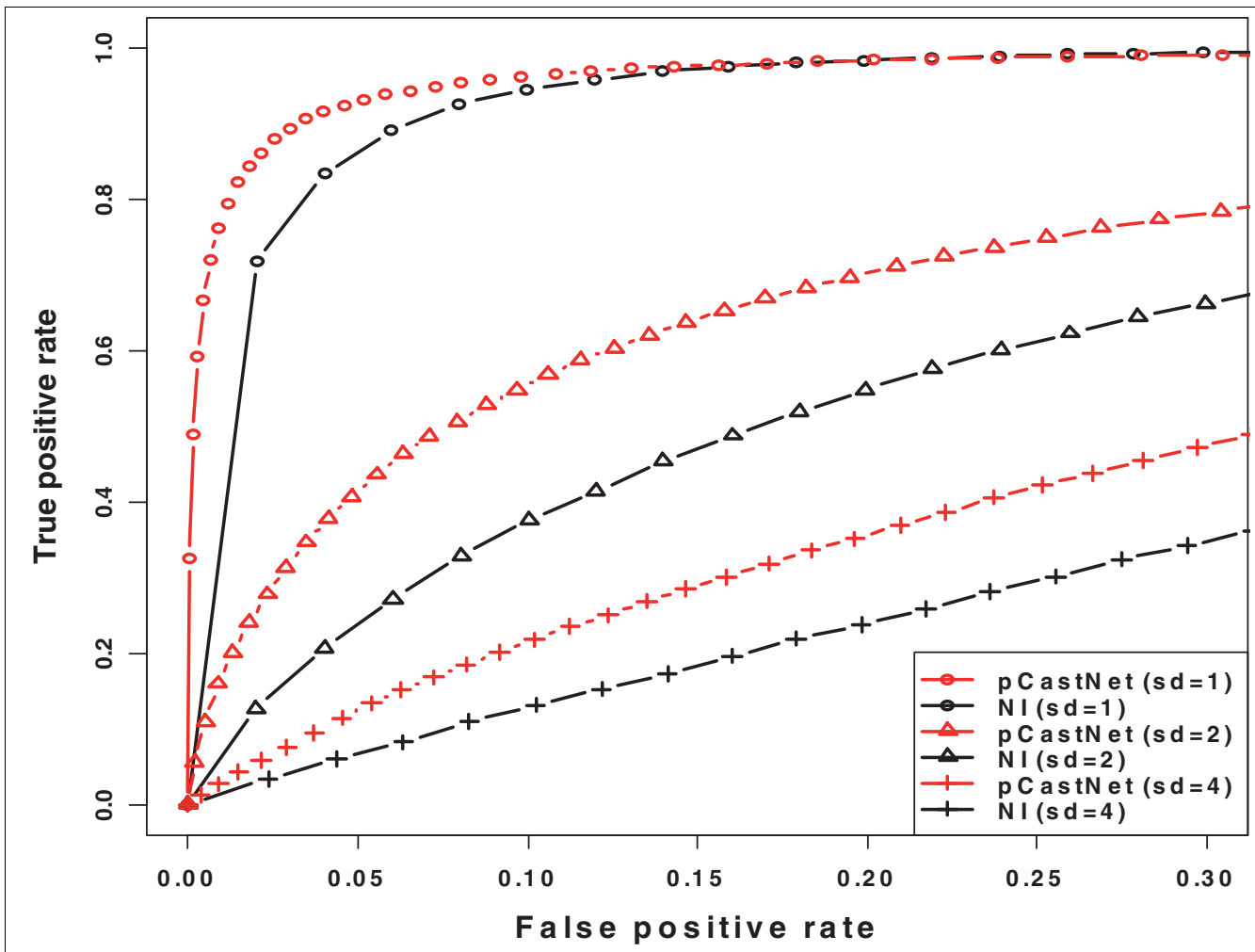


Figure 1
ROC (receiver operating characteristic) curves of pCastNet and the NI-based approach. The x-axis is the false positive rate and the y-axis is the true positive rate (power). Red lines are for pCastNet and black lines are for the NI-based approach. The standard deviation of expression level is 1 (circles), 2 (triangles), or 4 (crosses). Simulation procedures can be found in Materials and methods.

$r_{e_1e_2 \cdot g_1g_2} \sqrt{(n-4)/(1-r_{e_1e_2 \cdot g_1g_2}^2)}$ were first converted to z-values. The distribution of the minimum absolute z-value was estimated by a multivariate normal distribution; then, the minimum absolute z-value was further transformed to final z-values. Under the null hypotheses, the final z-values follow the standard normal distribution. By comparing the histogram of z-values and the standard normal distribution, we can estimate the dispersion parameter λ that reflects the dependence among hypotheses. Then we can calculate the conditional FDR. However, the number of declared links is very sensitive to the conditional FDR threshold (Table 1). Therefore, instead of applying a threshold on the conditional FDR directly, we estimated the sparseness of a network according to the conditional FDR and then chose a threshold on the sparseness. The sparseness of a network is defined as

the percentage of true links among all possible node pairs. The threshold selection has several advantages: first, the corresponding correlation thresholds are data dependent; second, we can derive an accurate estimate of the number of falsely declared links taking into consideration the dependence among hypotheses; and third, we can integrate prior information about the sparseness of networks if this information is available. Here we chose the sparseness threshold as 0.02%; this threshold corresponds to a reasonable conditional FDR and total number of declared GG, EG, and EE links. We also tried thresholds of 0.01% and 0.005%. The results discussed in the remaining of this paper are similar, although the number of links differs significantly (Table 1).

Table 1

Sparseness of networks and corresponding conditional FDR, z-value threshold, the number of GG, EG, EE links (n), and the range of correlations and partial correlations

Sparseness	Human			Mouse			Rat		
	0.02%	0.01%	0.005%	0.02%	0.01%	0.005%	0.02%	0.01%	0.005%
GG									
cFDR	0.227	0.206	0.185	0.05	0.031	0.02	0.045	0.025	0.015
z	4.42	4.61	4.79	4.77	5.02	5.25	4.8	5.07	5.31
n	13,552	6,523	3,202	12,014	5,878	2,893	2,672	1,307	653
$ r_{g1g2} $	≥ 0.947	≥ 0.957	≥ 0.965	≥ 0.964	≥ 0.972	≥ 0.979	≥ 0.965	≥ 0.974	≥ 0.981
EG									
cFDR	0.249	0.183	0.132	0.112	0.066	0.037	0.094	0.051	0.026
z	4.27	4.52	4.76	4.5	4.78	5.05	4.55	4.85	5.13
n	264,615	123,211	57,584	246,027	117,761	57,227	50,828	23,960	11,822
$ r_{e1g2} $	≥ 0.836	≥ 0.862	≥ 0.884	≥ 0.847	≥ 0.874	≥ 0.896	≥ 0.848	≥ 0.876	≥ 0.899
$ r_{e1g2g1} $									
EE									
cFDR	0.091	0.054	0.031	0.056	0.026	0.011	0.05	0.021	0.008
z	4.49	4.77	5.03	4.63	4.95	5.26	4.65	4.98	5.32
n	1,028,385	489,485	242,567	1,110,763	535,536	263,301	215,750	106,617	52,114
$ r_{e1e2} $	≥ 0.720	≥ 0.757	≥ 0.788	≥ 0.699	≥ 0.741	≥ 0.778	≥ 0.690	≥ 0.733	≥ 0.773
$ r_{e1e2g1} $									
$ r_{e1e2g2} $									
$ r_{e1e2g1g2} $									

The sparseness is the percentage of true links among all possible node pairs. Note that the number of declared links is very sensitive to the conditional FDR threshold. For example, for the GG network of human, when the conditional FDR (cFDR) changes from 4.42 to 4.61 (a 4% increase), the number of GG links changes from 13,552 to 6,523 (a 48% decrease). Meanwhile, the sparseness is from 0.02% to 0.01% (a 50% decrease).

Gene-gene, exon-gene and exon-exon links for human, mouse and rat

To study alternative splicing regulatory networks, we considered exon array data for human, mouse, and rat. For each organism, RNA samples from 11 tissues were profiled using Affymetrix exon arrays. The raw data were downloaded from the Affymetrix website [29] and the gene-level and the exon-level expressions were summarized using Affymetrix Power Tools.

GG association is the traditional GG co-expression association. EG association can be treated as the association between an alternatively spliced exon and its upstream regulators or its downstream target genes, which may not necessarily be direct regulators or direct target genes. Sophisticated models incorporating additional experiments (for example, CLIP experiments) are needed to infer the direct regulators or targets. EE association can be treated as the association between two alternatively spliced exons. The two exons could be regulated by the same direct or indirect splicing regulators. Another scenario could be that a specific transcript isoform of gene 1, which uniquely contains alternative exon 1 compared to other transcript isoforms of gene 1, regulates the exon of

gene 2. The latter case is a special exon-transcript association and 'transcript' here represents a particular transcript isoform instead of a family of gene splice variants. The above possible regulation relationships for EG and EE links are diagrammed in Additional data file 2. Additional data file 3 shows the Venn diagram of gene pairs with GG, EG, or EE associations. If GG links mainly reflect the transcriptional regulatory network whereas EG and EE links mainly reflect the alternative splicing regulatory network, it shows that these two networks are largely independent of each other.

Annotated alternative exons tend to have more exon-gene and exon-exon links

If an exon has association links with other exons or genes and such correlations are not due to the GG association, this exon is expected to be an alternatively spliced exon. Otherwise, if the exon is a constitutive exon that has a similar expression level to its gene, the EE or EG correlation is due to the GG correlation. We are interested to know whether EG or EE links can reflect the alternative splicing status of exons. Using the human data as an example, non-redundant transcript annotations were assembled from 14 sources (see details in Materials and methods). These transcripts may be experimentally

verified or just computationally predicted. Two groups of exons were then assembled from the large pool of transcript annotations: exons that are present in ≥ 14 transcript isoforms and are not spliced out in any transcript isoform; exons that are present in ≥ 7 transcript isoforms and are spliced out in another ≥ 7 transcript isoforms. The first exon group can be treated as constitutive exons and the second exon group can be treated as alternative exons. Figure 2 shows boxplots of the EG and EE links that the two groups of exons have; exons in group 2 clearly have more EG and EE links than exons in group 1. Specifically, for exons in group 1, 12% have ≥ 5 EG links and 11% have ≥ 50 EE links. For exons in group 2, the percentages increase to 23% for EG links and 21% for EE links. One-sided Wilcoxon tests show that exons in group 2 tend to have more EG and EE links with p -values $< 2.2 \times 10^{-16}$.

Conservation of exons with exon-gene or exon-exon links

It has been reported that the conservation level of alternative exons is lower than that of constitutive exons [30]. On the contrary, the intronic regions flanking alternative exons are more conserved than those flanking constitutive exons [30,31]. To assess whether exons with links to other exons (or genes) tend to be alternatively spliced, we plotted the conservation scores of exons and their flanking regions (Figure 3). Exons were divided into three groups: exons with node degree = 0 (black lines); exons with node degree > 0 and the node degree is in the top 10% of all non-zero node degrees (green lines); exons with node degree > 0 and the node degree is not in the top 10% list (red lines). Node degree is defined as the number of links that a node has to other nodes in the network. Here it represents the number of links that an exon has to

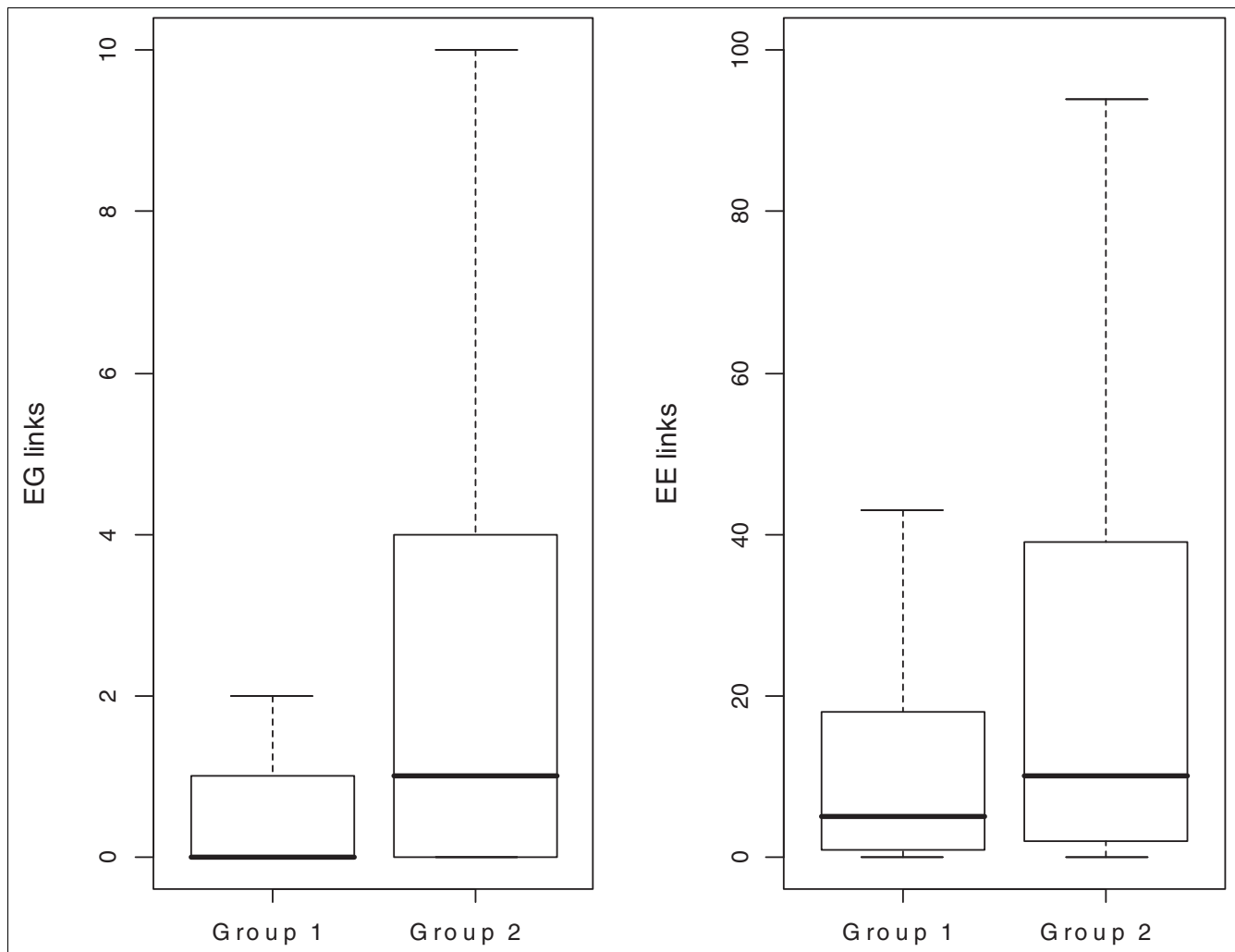


Figure 2
Boxplot of node degree of constitutive exons and alternative exons. Two groups of exons were assembled according to transcript annotations from 14 sources. Group 1 represents constitutive exons. Group 2 represents alternative exons. The boxplots of EG links and EE links are plotted (outliers are not drawn). Notice that alternative exons tend to have more EG and EE links than constitutive exons.

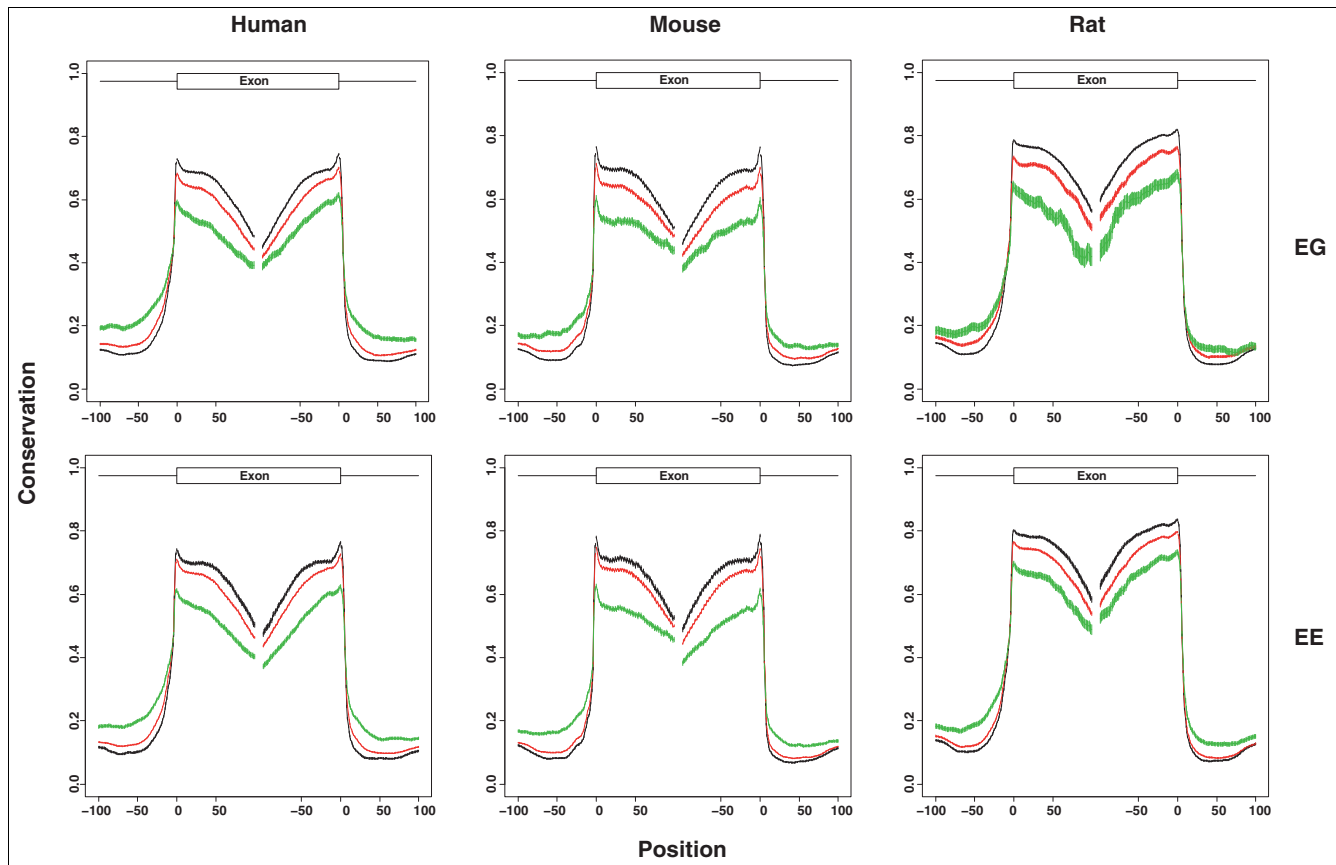


Figure 3
Conservation of exons with or without EG and EE links. For every site of an exon, x is defined as the position relative to the nearest splice site. It is positive for distances from the 5' edge and negative for distances from the 3' edge. The upstream intronic region is from -100 to 0 bp and the downstream intronic region is from 0 to 100 bp. Exons were divided into three groups: exons with node degree = 0 (black lines); exons with node degree > 0 and the node degree is in the top 10% of all non-zero node degrees (green lines); exons with node degree > 0 and the node degree is not in the top 10% list (red lines). The y-axis is the average conservation score for the three exon groups. The error bar indicates the standard error of the mean for each position.

other genes (EG) or exons (EE). The average PhastCons conservation score at each exon and flanking region position was calculated and plotted for the three exon groups. Exons with EG or EE links tend to be less conserved than exons without EG or EE links. The flanking intronic regions of exons with EG or EE links tend, however, to be more conserved than those of exons without EG or EE links, which is possibly related to the enriched *cis*-splicing regulatory elements in intronic regions. The more links an exon has, the less it is conserved and the more its flanking intronic regions are conserved. For Affymetrix exon arrays, an exon may represent a cluster of overlapping exons from transcript isoforms with different 5' or 3' splicing sites. The boundary of such an exon cluster may not be the real boundary of the exon in a cell. To eliminate this bias, we removed exons with more than one probe selection region (that is, exons with more than one pair of splicing sites). The results are similar (data not shown).

Relative position of exons with exon-gene or exon-exon links

The relative position from 5' to 3' was calculated for each exon, ranging from 0 to 1. The relative positions were partitioned into 10 windows. The proportion of exons with relative positions falling in each window was counted for exons with or without EG (EE) links and denoted as p_1 or p_2 , respectively. Figure 4 plots the ratio between p_1 and p_2 for each relative position window. It clearly shows that exons with EG or EE links tend to be enriched in the initial or terminal regions. Alternative promoters and alternative polyadenylation sites are two of the most prevalent mechanisms for generating transcript isoforms by including alternative first or last exons. Recent studies suggest that 30-50% of human and approximately 50% of mouse genes have multiple alternative promoters [32-36]. In addition, about 54% of human and 32% of mouse genes have alternative polyadenylation sites [37]. Exons with links to other genes or other exons are very likely to be alternatively spliced. Many of them, therefore, are close to the initial or the terminal regions of genes.

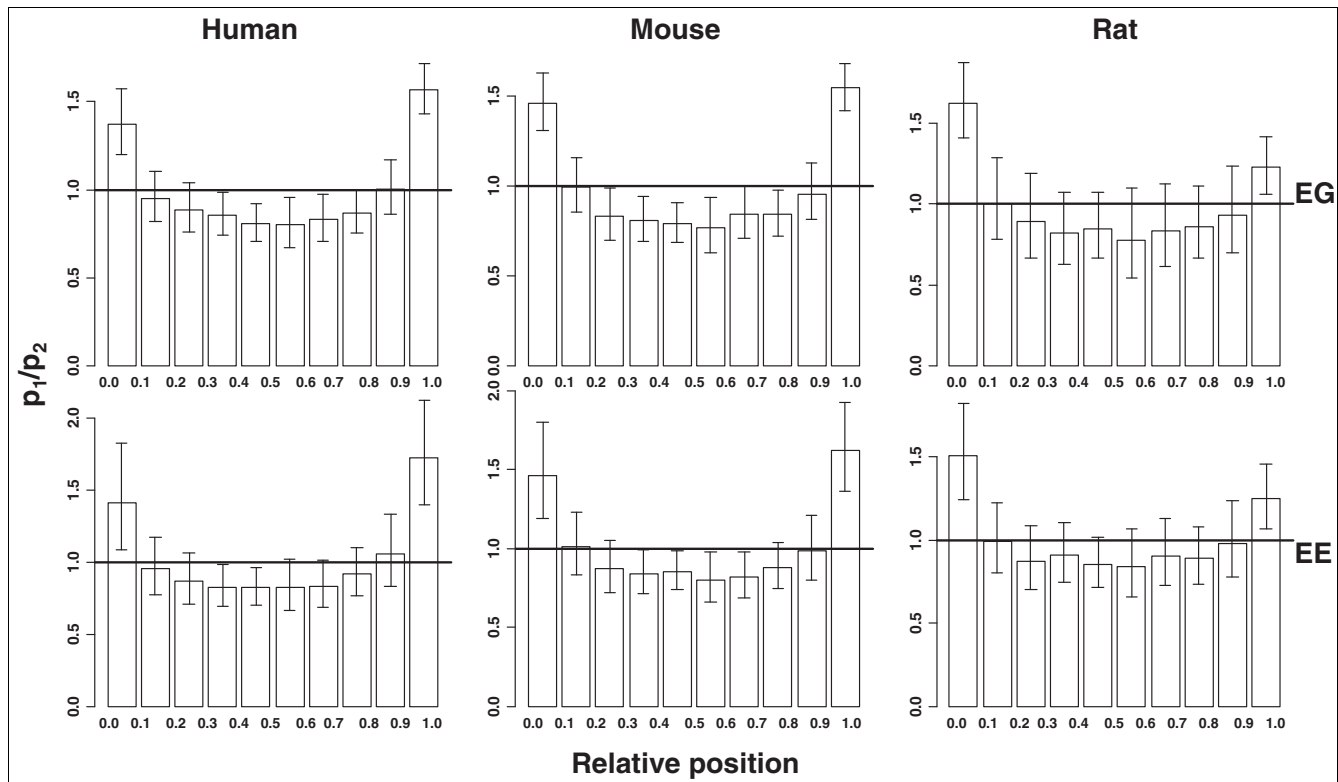


Figure 4
Enrichment of exons with EG or EE links at the termini of genes. For each gene, all of the core exons were sorted according to their genomic coordinates (from 5' to 3'). The relative position of the *i*-th exon is calculated as $(i - 1)/(n - 1)$, where *n* is the total number of exons. The relative positions were partitioned into ten windows. The proportion of exons with relative positions falling in each window was counted for exons with links and exons without links and denoted as p_1 or p_2 , respectively. The y-axis represents the p_1/p_2 ratio. Error bars represent the 95% confidence intervals of p_1/p_2 . Notice that p_1/p_2 is higher near the terminal regions.

Functional annotation analysis of hubs

We assembled exons with node degrees ranking in the top 1% in the EG network or the EE network. The DAVID functional annotation tool [38] was used with genes to which hub exons belong. The same was done for genes with node degrees ranking in the top 1% in the GG network. Table 2 lists the enriched annotation terms with at least five gene counts, with *p*-values after Bonferroni's correction ≤ 0.001 , and that appear at least twice in the nine groups (EG, EE, and GG for human, mouse, and rat). Bonferroni's correction is a very stringent multiple comparison correction. Here it restricts the probability of having one or more falsely declared significant annotation terms to ≤ 0.001 . The term 'alternative splicing' is a UniProt knowledgebase keyword meaning 'protein for which at least two isoforms exist due to distinct pre-mRNA splicing events'; it is enriched in genes with hub exons for all of the EG and EE networks. The Uniprot sequence feature 'splice variants' is also enriched in these hub exons. However, 'alternative splicing' and 'splice variants' are not enriched in the gene hubs of the GG networks.

Experimental validation

We experimentally examined the pCastNet results by RT-PCR across various tissues. In particular, the EE link is a relatively new correlation subject (in biology) and a very interesting phenomenon. We randomly chose EE links at the lower bound of the correlation cut-off (about 0.75-0.80) but favored cassette exons because of the ease of RT-PCR design. Due to the nature of our data, we also favored genes that are expressed in multiple tissues in order for PCR to amplify with the same number of cycles across the tissues. pCastNet found significant EE links among Kinesin-associated protein 3 (*Kifap3*) exon 20 (exon id 24930 in Affymetrix exon array), Suppression of tumorigenicity 7 (*St7*) exon 7 (exon id 685163), and Mitogen-activated protein kinase kinase kinase 7 (*Map3k7*) exon 12 (exon id 572746). For convenience, we refer to these exons as *Kifap3_20*, *St7_7*, and *Map3k7_12*. Specifically, pCastNet predicted that *Map3k7_12* has negative associations with both *Kifap3_20* and *St7_7* (correlations and partial correlations are about -0.75 for both), whereas *Kifap3_20* has a positive association with *St7_7* (correlation and partial correlations are about 0.80). NCBI EST database shows these exons are all alternative exons. Primers were designed in the flanking constitutive exons to

Table 2**Functional annotation analysis of exon hubs or gene hubs**

Category	Term	Corrected p -value for hubs of EG networks			Corrected p -value for hubs of EE networks			Corrected p -value for hubs of GG networks		
		Human	Mouse	Rat	Human	Mouse	Rat	Human	Mouse	Rat
SP	Alternative splicing	2.4×10^{-13}	6.3×10^{-25}	2.6×10^{-10}	1.0×10^{-21}	5.3×10^{-20}	3.0×10^{-19}	N	N	N
UP	Splice variant	1.7×10^{-06}	5.2×10^{-14}	5.3×10^{-05}	8.4×10^{-12}	4.9×10^{-08}	1.6×10^{-11}	N	N	N
MF	Binding	2.3×10^{-05}	8.0×10^{-05}	N	1.8×10^{-07}	3.1×10^{-06}	3.1×10^{-04}	N	N	N
SP	Phosphoprotein	1.4×10^{-17}	8.9×10^{-09}	N	1.7×10^{-16}	2.0×10^{-07}	5.3×10^{-07}	N	N	N
MF	Protein binding	2.2×10^{-08}	2.7×10^{-04}	N	8.0×10^{-08}	1.1×10^{-04}	1.7×10^{-07}	N	N	N
CC	Intracellular	2.6×10^{-07}	1.0×10^{-15}	N	1.2×10^{-20}	2.7×10^{-06}	N	N	N	N
CC	Intracellular part	8.6×10^{-09}	6.1×10^{-13}	N	9.4×10^{-20}	9.0×10^{-07}	N	N	N	N
SP	Cytoplasm	N	2.6×10^{-08}	N	N	6.4×10^{-08}	N	5.8×10^{-04}	N	N
CC	Cytoplasm	5.3×10^{-04}	5.5×10^{-11}	N	N	2.6×10^{-07}	N	N	N	N
CC	Intracellular organelle	2.2×10^{-05}	N	N	9.4×10^{-18}	N	N	N	6.5×10^{-04}	N
CC	Organelle	2.4×10^{-05}	N	N	1.0×10^{-17}	N	N	N	6.7×10^{-04}	N
SP	Coiled coil	N	1.6×10^{-05}	N	N	8.2×10^{-06}	N	N	N	N
BP	Cellular component organization and biogenesis	N	N	N	2.5×10^{-04}	N	3.8×10^{-04}	N	N	N
CC	Intracellular organelle part	N	N	N	4.6×10^{-05}	N	N	N	1.9×10^{-04}	N
BP	Macromolecule metabolic process	5.9×10^{-05}	N	N	2.5×10^{-08}	N	N	N	N	N
CC	Nucleus	N	N	N	1.2×10^{-14}	N	N	N	1.3×10^{-06}	N
SP	Nucleus	N	N	N	1.9×10^{-10}	N	N	N	8.1×10^{-05}	N
CC	Organelle part	N	N	N	5.9×10^{-05}	N	N	N	2.0×10^{-04}	N
CC	Synapse	N	N	N	N	3.0×10^{-04}	1.1×10^{-15}	N	N	N
BP	Transport	N	N	9.7×10^{-04}	N	N	7.1×10^{-14}	N	N	N

The DAVID functional annotation tool was applied to genes whose exons are the hubs of EG networks, genes whose exons are the hubs of EE networks, and genes that are the hubs of GG networks. The listed gene annotation terms have at least five gene counts, have p -values after Bonferroni's correction ≤ 0.001 , and appear at least twice in the nine groups (EG, EE, and GG for human, mouse, and rat). 'N' means the term is not significant for this group. The annotation terms considered here are from the default settings. SP, SP_PIR_KEYWORDS where PIR means protein information resource. UP, UP_SEQ_FEATURE, which means Uniprot sequence feature. BP, GOTERM_BP_ALL where BP means biological process. CC, GOTERM_CC_ALL where CC means cellular component. MF, GOTERM_MF_ALL where MF means molecular function.

amplify transcripts either containing or skipping these alternative exons. RT-PCR results and Pearson correlation analysis of exon inclusion levels (Figure 5b) show that *Map3k7_12* is negatively correlated with both *Kifap3_20* and *St7_7* while *Kifap3_20* is positively correlated with *St7_7* in these tissues. Besides the tissues surveyed in the exon array study, we also performed RT-PCR experiments in seven other tissues (Figure 5d). Based on the RT-PCR experiments, the correlation between *Kifap3_20* and *St7_7* is 0.60 and the correlation between *Map3k7_12* and *St7_7* is -0.82 whereas the correlation between *Map3k7_12* and *Kifap3_20* dropped to -0.29. Another example is a positive correlation between Solute carrier family 35, member B3 (*Slc35b3*) exon4 (exon id 226950, or *Slc35b3_4*) and Retinoic acid induced 14 (*Rai14*) exon 11

(exon id 300782, or *Rai14_11*). pCastNet predicts a positive association between these two exons (correlation and partial correlations are about 0.80). RT-PCR and Pearson correlation analysis (Figure 5c) show a positive correlation of 0.75 among the tested tissues used by the Affymetrix exon array. In the second set of tissues, RT-PCR experiments show that their correlation is about 0.87 (Figure 5e). Note that *Slc35b3* is not detectable in bladder, and thus has not been included in the correlation analysis.

Functional similarity of gene pairs with links

All of the above results indicate that pCastNet can effectively identify EG and EE links. We then further explored the possible functional relationship between two genes with an EG link

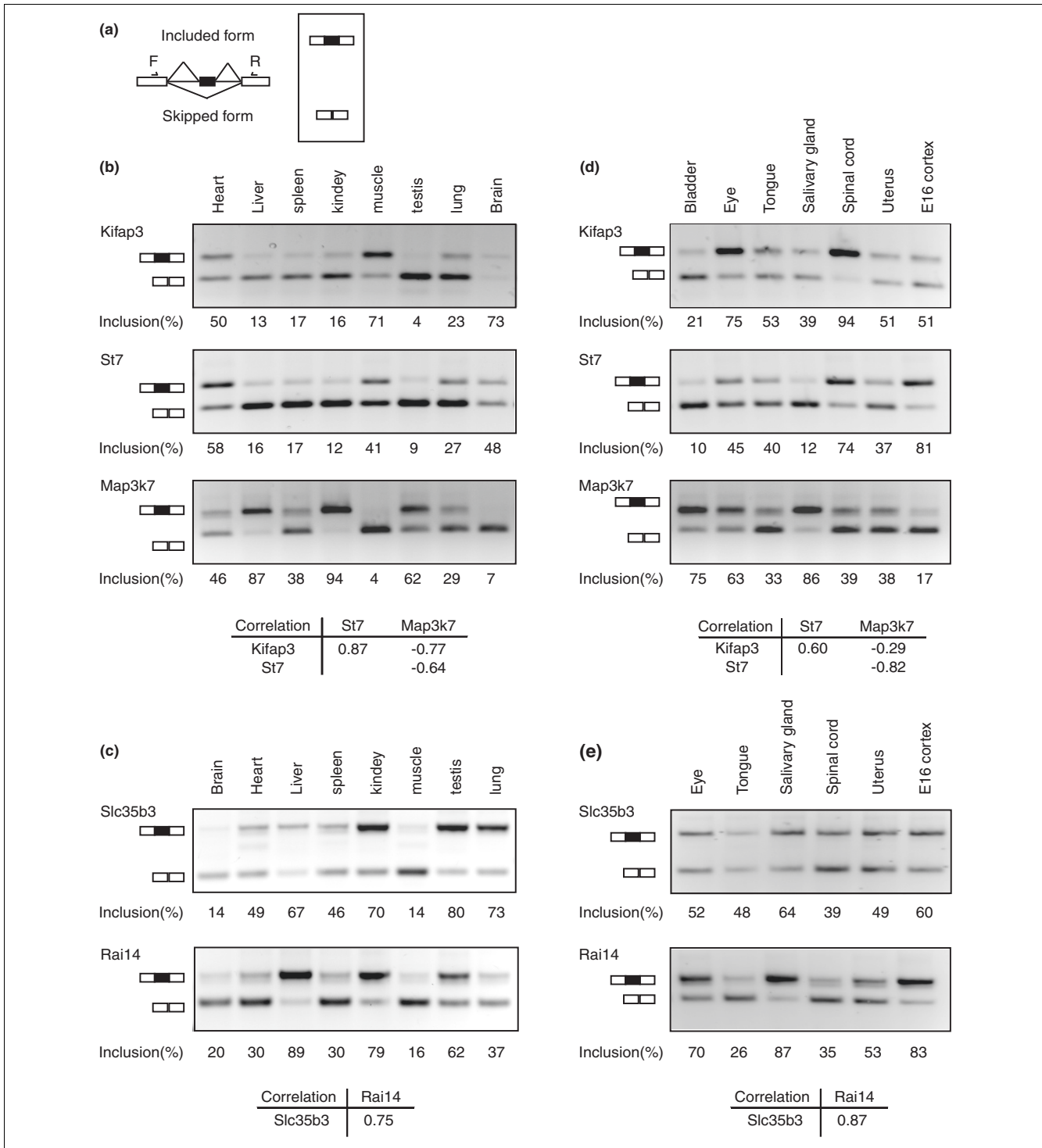


Figure 5
Examples of EE links illustrated by RT-PCR of tissue RNAs. (a) Scheme of RT-PCR design to examine splicing levels of alternative exons. Primers (arrows) are in the flanking constitutive exons. Inclusion levels of alternative exons (black box) are calculated as Included form/(Included form + Skipped form). (b). Alternative splicing of *Kifap3* exon20, *St7* exon 7, and *Map3k7* exon 12 in multiple mouse tissues. *Kifap3* exon 20 is positively correlated with *St7* exon 7 and negatively correlated with *Map3k7* exon 12. *St7* exon 7 is negatively correlated with *Map3k7* exon 12. Pair-wise Pearson correlations based on the RT-PCR experiments are shown. (c) *Slc35b3* exon 4 is positively correlated with *Rai14* exon 11. (d) Pair-wise correlations between *Kifap3* exon20, *St7* exon 7, and *Map3k7* exon 12 in a second set of tissues not surveyed by the Affymetrix exon array. (e) Pair-wise correlation between *Slc35b3* exon 4 and *Rai14* exon 11. Percentages of inclusion levels were averaged from three independent experiments.

or an EE link. Using the human data and the Molecular Signatures Database [39], genes were grouped into gene sets according to: their chromosome positions; curated information from pathway databases; shared conserved *cis*-regulatory motifs; and shared Gene Ontology (GO) terms. We tested whether genes with EG or EE links tend to be in the same gene sets using hypergeometric tests. The results are summarized in Table 3. Genes in the same chromosomal cytogenetic band ('c1') are more likely to have GG and EG links than EE links. Gene pairs with GG, EG, or EE links tend to be in the same pathways (these pathways are collected by the BioCarta, GenMAPP, and KEGG databases). More interestingly, gene pairs with EG or EE links tend to be in the same motif gene sets ('c3'). Specifically, genes in those sets share a motif in the promoter regions ('c3_promoter_known' and 'c3_promoter_unknown') or a microRNA (miRNA) binding site in the 3' untranslated regions ('c3_miRNA'). On the contrary, the *p*-values of GG links in the promoter motif sets are less significant than those of EG and EE links. And gene pairs with GG links are not enriched in the 'miRNA binding' gene sets. In addition, exons with EE links and sharing miRNA binding motifs tend to be enriched at the 3' terminals of the genes (Additional data file 4). Finally, genes with GG, EG or EE links all tend to share GO terms.

We also examined *p*-values for the enrichment of links in each individual gene set. We counted the number of GG, EG, or EE links between members of a gene set for each gene set. To test the significance of the enrichment of links, we simulated gene sets by randomly selecting the same number of genes. The simulated gene sets have no functional similarity. We then calculated the empirical *p*-values of the number of observed GG, EG or EE links as *Pr* (the number of links in the simulated gene set \geq the number of observed links) from 1,000 simulations. Figure 6 plots the histogram of the *p*-values of gene sets with at least one observed GG, EG, or EE link. For all gene set categories except category 1, there are more gene sets enriched with GG, EG or EE links compared with the random selections, where a uniform distribution of *p*-values is expected.

Examples

The motif (U)GCAUG has been reported as a binding motif for mammalian splicing factors FOX-1 (A2BP1) and FOX-2 (RBM9) [40-43]. We studied the enrichment of motif GCAUG in exons with EG links to FOX-1 and FOX-2. For each exon, we counted the occurrence of the pentamer GCAUG in the exonic region and the flanking 200 bp intronic regions. Table 4 shows the enrichment of this motif for exons correlated with

Table 3

Gene pairs sharing gene sets and having GG, EG, or EE links

Gene set category	No. of gene pairs sharing a gene set among a total of 53,721,795 gene pairs	Gene pairs having GG links (13,552)		Gene pairs having EG links (223,116)		Gene pairs having EE links (815,024)	
		No. of gene pairs also sharing gene set	<i>p</i> -value	No. of gene pairs also sharing gene set	<i>p</i> -value	No. of gene pairs also sharing gene set	<i>p</i> -value
c1	321,284	150	2.1×10^{-12}	1,584	1.1×10^{-11}	4,855	0.61
c2_BioCarta	30,023	40	2.0×10^{-17}	238	5.4×10^{-20}	601	2.3×10^{-11}
c2_GenMAPP	49,570	101	2.4×10^{-56}	421	4.4×10^{-40}	1,085	6.1×10^{-31}
c2_KEGG	182,069	348	1.1×10^{-179}	1,137	1.1×10^{-38}	3,048	3.1×10^{-08}
c3_miRNA	2,373,102	556	0.96	12,477	8.7×10^{-150}	49,006	$< 4.9 \times 10^{-324}$
c3_promoter_known	14,479,685	4,207	1.7×10^{-26}	67,435	7.5×10^{-261}	248,414	$< 4.9 \times 10^{-324}$
c3_promoter_unknown	3,186,951	915	3.5×10^{-05}	14,507	1.1×10^{-29}	55,308	1.1×10^{-227}
c4_bp	7,107,791	2,947	5.5×10^{-163}	35,289	6.9×10^{-272}	129,574	$< 4.9 \times 10^{-324}$
c4_cc	6,549,156	4,418	$< 4.9 \times 10^{-324}$	35,702	$< 4.9 \times 10^{-324}$	127,572	$< 4.9 \times 10^{-324}$
c4_mf	815,318	581	3.7×10^{-104}	4,294	4.3×10^{-52}	15,897	1.0×10^{-209}

Among the 10,366 human genes that were considered, there are 53,721,795 possible gene pairs. About 13,552 gene pairs were declared to have GG links; 223,116 gene pairs were declared to have at least one EG link; and 815,024 gene pairs were declared to have at least one EE link. The number of gene pairs that have GG, EG, or EE links and are in the same gene set is listed. The *p*-value of observing such a high or higher number of gene pairs that have GG, EG, or EE links and are in the same gene set was based on a hypergeometric test. Different gene set categories were considered. 'c1', genes sharing chromosomal cytogenetic bands; 'c2_BioCarta', 'c2_GenMAPP', and 'c2_KEGG', genes in the same pathways and the pathways were collected from the BioCarta, GenMAPP, or KEGG databases; 'c3_miRNA', genes sharing a microRNA binding site; 'c3_promoter_known' and 'c3_promoter_unknown', genes sharing a motif in the promoter regions and the motif matches a known transcription factor binding site or the motif does not match any known transcription factor binding site; 'c4_bp', genes sharing biological process ontology terms; 'c4_cc', genes sharing cellular component ontology terms; 'c4_mf', genes sharing molecular function ontology terms.

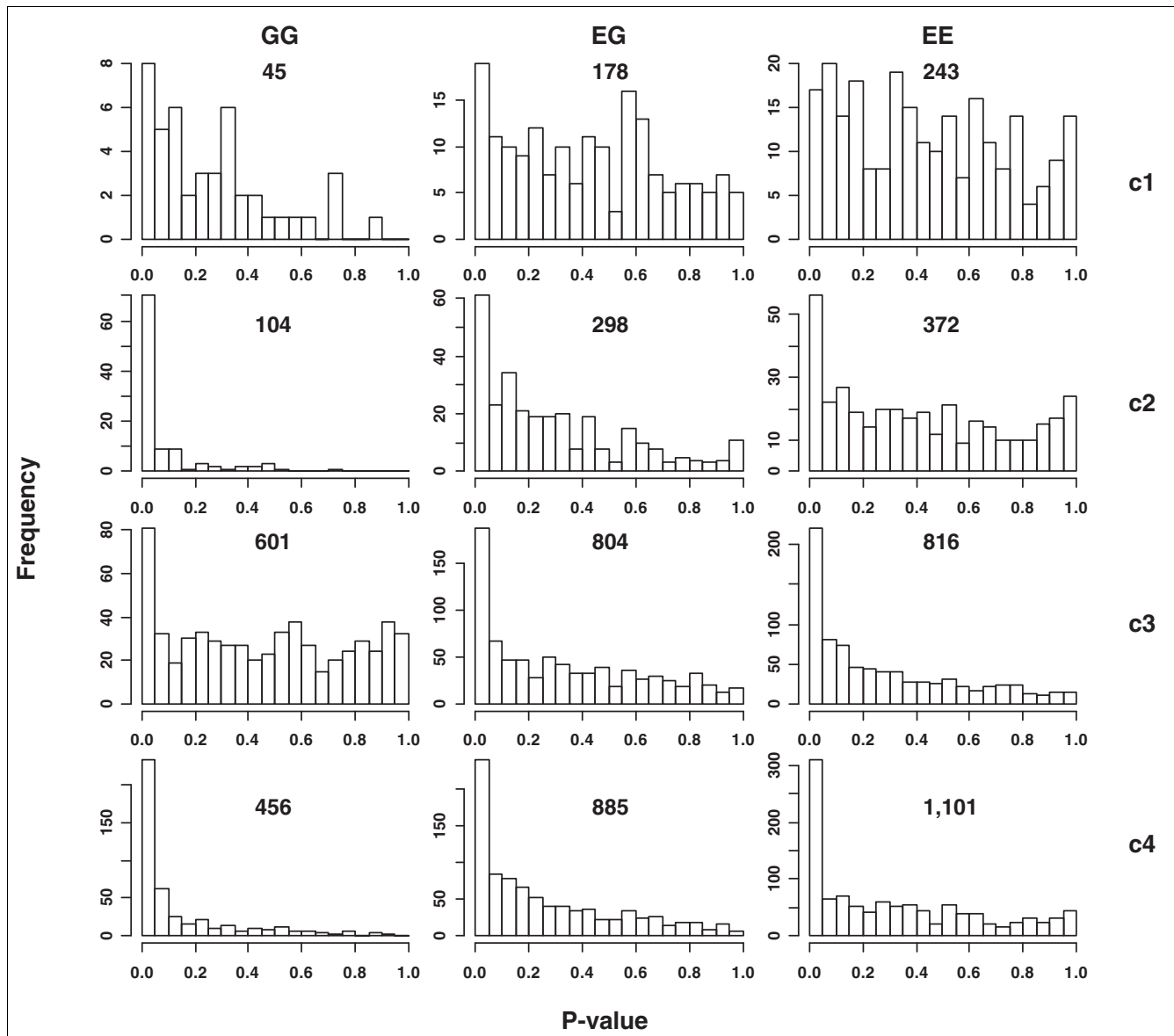


Figure 6
Enrichment of GG, EG, or EE links in functional gene sets. For each gene set with at least one GG, EG, or EE link, to test the significance of the enrichment of links, we simulated gene sets by randomly selecting the same number of genes as the tested gene set. The empirical *p*-value of the number of observed links was calculated as $\Pr(\text{the number of links in the simulated gene set} \geq \text{the number of observed links})$ from 1,000 simulations. Histograms of the *p*-values are plotted for those gene sets. The total number of test gene sets is listed on the histograms. C1: gene sets sharing a chromosomal cytogenetic band. C2: gene sets curated from pathway databases. C3: gene sets sharing a conserved *cis*-regulatory motif. C4: gene sets sharing a GO term.

FOX-1 or FOX-2. The empirical *p*-values of the enrichment were based on 1,000 simulated exon groups. Among the 64 exons correlated with FOX-1 in human, GCAUG occurs 96 times. And none of the 1,000 simulated exon groups has more than 96 occurrences of GCAUG. The expression level of FOX-1 and the inclusion rates of its associated exons are plotted in Additional data file 5; this clearly shows the co-expression patterns between FOX-1 and exons with EG links to FOX-1. Although the *p*-values for FOX-2 in human (0.031) and Fox-1 and Fox-2 in mouse (0.172, 0.060) are less significant, the

occurrences of GCAUG are about twice as many as the average occurrence among the 1,000 simulated groups. Note that after the filtering procedures for the raw data, FOX-1 and FOX-2 are not in the final gene list for rat.

The calcium signaling pathway has been shown to be intensively related to alternative splicing [44]. In our gene set analysis, the KEGG calcium signaling pathway is enriched with EG and EE links with empirical *p*-values of 0.01 and 0.001, respectively. However, GG links are not enriched in the path-

Table 4

Motif enrichment for genes with EG links to FOX-1 and FOX-2

	Human		Mouse	
	FOX-1(A2BP1)	FOX-2(RBM9)	Fox-1(A2bp1)	Fox-2(Rbm9)
Number of exons with EG links to FOX-1 or FOX-2	64	21	19	5
Number of GCAUG occurrences among exons with EG links to FOX-1 or FOX-2	96	28	21	9
Average number of GCAUG occurrences among 1,000 simulated exon groups	49.4	16.6	15.8	4.2
P-value of the motif enrichment	0.000	0.031	0.172	0.060

Exons with EG links to FOX-1 or FOX-2 were assembled. The occurrences of GCAUG were calculated for those exon groups. To test the significance of the motif enrichment, we simulated exon groups by randomly selecting exons. The p-value of the motif enrichment was calculated as Pr (the motif frequency in the simulated exon group ≥ the observed motif frequency) from 1,000 simulations.

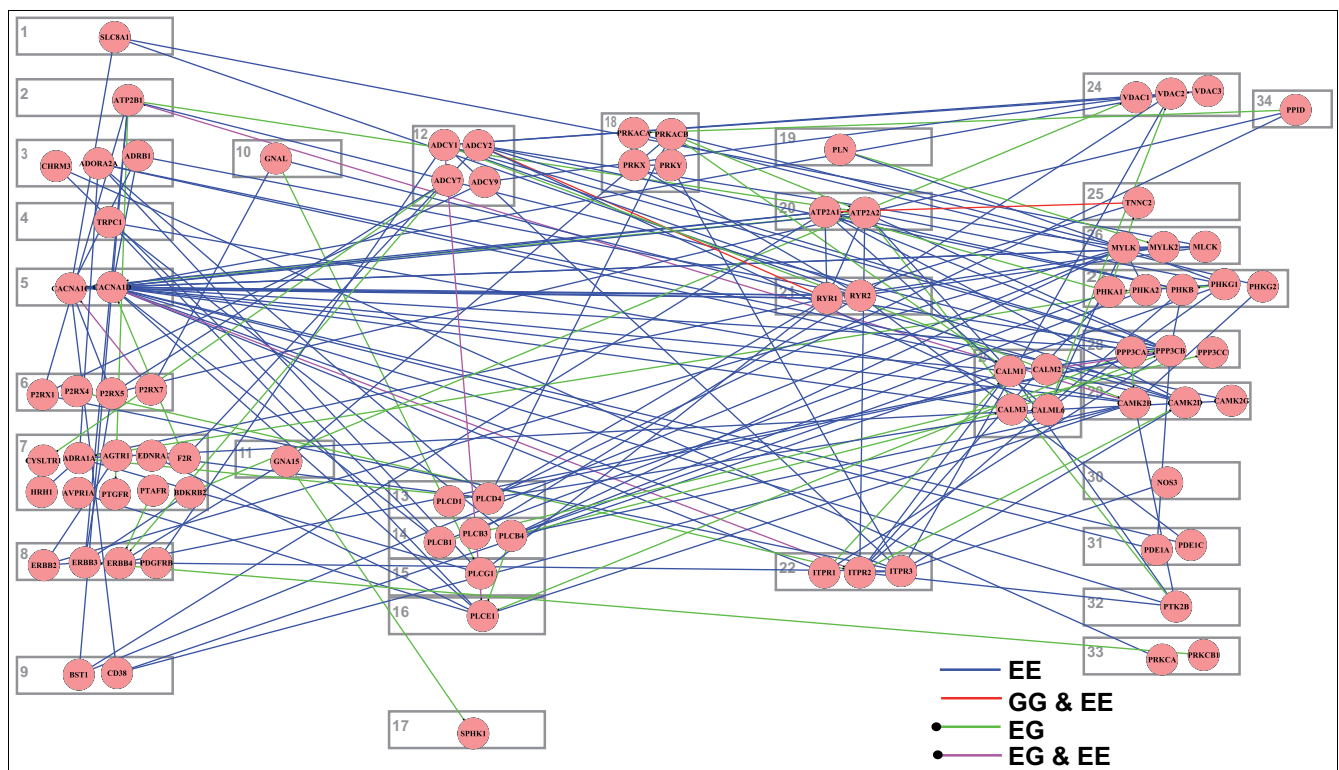


Figure 7

GG, EG, and EE links in the calcium signaling pathway. The gene layout is the same as the layout in the KEGG pathway database. A red line represents a GG link. The two gene pairs with GG links happen to also have EE links ('GG & EE'). A blue line represents an EE link and a green line represents an EG link. A dot at one end of a line is used to represent the exon in an EG link. The KEGG calcium signaling pathway is enriched with EG and EE links with p-values 0.01 and 0.001. However, GG links are not enriched in the pathway, with a p-value of 0.354. Each box represents the corresponding component in the KEGG database: 1, NCX; 2, PMCA; 3, GPCR; 4, SOC; 5, CaV1; 6, ROC; 7, GPCR; 8, PTK; 9, CD38; 10, Gs; 11, Gq; 12, ADCY; 13, PLCδ; 14, PLCβ; 15, PLCγ; 16, PLCE; 17, SPHK; 18, PKA; 19, PLN; 20, SERCA; 21, RYR; 22, IP3R; 23, CALM; 24, VDAC; 25, TnC; 26, MLCK; 27, PHK; 28, CaN; 29, CAMK; 30, NOS; 31, PDE1; 32, FAK2; 33, PKC; 34, PPIID. A circle in a box represents a gene in the corresponding component.

way, with a p -value of 0.354. Figure 7 plots the EG, EE, and GG links in the pathway. The gene layout is the same as the layout in the KEGG database. Most components of the calcium signaling pathway from the KEGG database have at least one link, as shown in Figure 7. Red lines represent GG links (the two gene pairs happen to have EE links also). Blue links represent EE links and green lines represent EG links. The above results indicate the important role of alternative splicing in signaling pathways and/or the important roles of calcium signaling pathways in alternative splicing regulation.

Discussion

In this paper, we propose the use of pCastNet to identify EE co-splicing links and EG co-expression links. pCastNet avoids taking the ratio between exon-level intensity and gene-level intensity and it achieves a higher statistical power compared to an NI-based approach (Figure 1). Such EG and EE links can provide information about alternative splicing. For example, alternative exons have significantly more EG or EE links than constitutive exons (Figure 2). Secondly, exons with EG or EE links tend to be less conserved in exonic regions than exons without EG or EE links. On the contrary, the flanking intronic regions of exons with EG or EE links tend to be more conserved than those of exons without EG or EE links (Figure 3). Such observations are consistent with the conservation patterns of alternative exons and constitutive exons [30]. In addition, exons with EG or EE links tend to be enriched in the 5' or 3' termini of genes where alternative splicing events are enriched (Figure 4). The functional annotation analysis also indicates that genes containing exon hubs of EG or EE networks tend to have multiple splicing isoforms (Table 2). All the results indicate that the EG or EE links can reflect the alternative splicing status of exons. Furthermore, they can provide information about the alternative splicing regulatory network.

The alternative splicing regulatory network reconstructed by pCastNet is composed of nodes (exon or gene) and their pairwise association links. It provides a different way to study alternative splicing from previous differential analysis. Typical differential analysis compares two tissues or conditions; for example, by studying differential alternative splicing between tissues, one can identify a cluster of tissue-specific alternative splicing events. By studying differential alternative splicing after the knockdown of splicing factors, one can identify a cluster of target candidates of splicing regulators. pCastNet considers multiple conditions at one time; by studying co-expression patterns of nodes across multiple conditions, we can identify pair-wise links between nodes. From these links regulatory or functional relationships can be inferred and they provide a comprehensive view of alternative splicing regulation. However, links identified by the current study are association links and not necessarily causal links. The possible regulatory relationships they reflect can be direct or indirect. CLIP based and knockdown experiments

are more powerful tools to identify direct causal links. Although pCastNet does not provide as strong evidence as CLIP studies for identifying downstream targets of splicing factors, it infers other spaces of regulation, for example, upstream regulatory genes besides the splicing factors of interest. If one is interested in a specific signaling pathway where multiple components can affect each other simultaneously, pCastNet can identify invaluable links to dissect the regulatory relationship. As we show in the calcium signaling pathway example, EE and EG links but not GG links are significantly enriched and the results provide clues to investigate the functional and regulatory relationships between nodes. In summary, pCastNet and differential study are complementary to each other and should be considered in combination to better understand the network of interest.

We validated pCastNet predictions using RT-PCR experiments. No studies have reported the co-splicing of exons *Kifap3_20*, *St7_7*, and *Map3k7_12* nor reported on the functional relationships between *Kifap3*, *St7*, and *Map3k7*. *Kifap3* is an auxiliary factor of the Kinesin family member 3 (*Kif3*) heterotrimer complex that links KIF3 with its cargos [45]. Alternative splicing of *Kifap3_20* generates two *Kifap3* isoforms that differ in the carboxyl terminus region. *St7* has been reported to be a tumor suppressor gene involved in multiple types of cancer [46]. In addition, Vincent *et al.* [47] reported that *St7* spans a translocation point in a patient with autism. They also observed the alternative splicing of exon 7. *Map3k7* is a mitogen-activated protein kinase kinase kinase that transduces intracellular signals from the interleukin-1 receptor [48] and tumor necrosis factor receptor [49]. Kondo and colleagues [50] reported strong biases of isoform (either including or excluding *Map3k7_12*) ratios in some lung cancer specimens. Why and how the splicing of *Kifap3_20*, *St7_7*, and *Map3k7_12* are coordinated in some tissues is an interesting question.

The selection of samples or conditions will impact the identification and biological interpretation of links. Although we had preferred to study a group of phenotypically relevant conditions (for example, different parts of the brain) to better infer the biological meanings of links, there were limited data available at the time of our study. As our data are from diverse tissues, the network more likely identifies links shared by the majority of the selected tissues. For example, pCastNet predicts a significant correlation between exon *Slc35b3_4* and exon *Rai14_11*. RT-PCR experiments have validated the correlation in tissues of brain, heart, liver, spleen, kidney, muscle, lung and testis (Figure 5c). Besides these eight tissues surveyed by the Affymetrix exon array, we also considered a different set of tissues not surveyed by the exon array study to see if this correlation is a general pattern extendable to other tissues. RT-PCR clearly shows the positive correlation among eye, tongue, salivary gland, spinal cord, ovary and E16 cortex (Figure 5e). This is consistent with the idea that due to the sources of our data, links identified by pCastNet in the cur-

rent study tend to be a general phenomenon shared by multiple tissues. Another example is the pair-wise correlations between the splicing of Kifap3_20, St7_7, and Map3k7_12 in brain, heart, liver, spleen, kidney, muscle, lung and testis (Figure 5b). For those seven additional tissues, Kifap3_20 and St7_7 are still positively associated, with a correlation of 0.60. Map3k7_12 and St7_7 are still negatively associated with a correlation of -0.82. However, the negative correlation between Map3k7_12 and Kifap3_20 decreases to -0.29 (Figure 5d). Therefore, although these exons are coordinately regulated in general, such a relationship remains context-dependent. If their correlation is directly caused by one or a few alternative splicing regulators, we could surmise that the tissue-specific expression of these splicing regulators and their differential *trans*-activity strength on the three exons confers the context-dependent correlation. Another explanation would be that the three exons have separate unique regulators besides the common regulators. The unique regulators counteract the effects of common regulators and are expressed in a tissue-specific manner. In the future, the power of pCastNet will be extended by combining it with transcriptome differential analysis and RNA binding protein motif analysis in order to elucidate the coordinate and combinatorial alternative splicing regulatory network.

We discovered the functional similarity of gene pairs with EG or EE links. Strikingly, gene pairs with EG or EE links tend to share a conserved sequence element in their promoter regions. However, the *p*-values for gene pairs with GG links are less significant. It has been reported and remains a puzzle that, in mammals, the direct correlation between regulatory *cis*-element similarity and expression similarity is not significant [51]. A second striking phenomenon is that gene pairs with EG or EE links tend to share miRNA targets (Table 3). However, the *p*-value for GG links is not significant, which is consistent with the general concept that miRNAs mainly affect protein translation but not transcript amounts in mammals. These results indicate the coupling of co-alternative splicing, co-transcription factor binding, and co-miRNA binding for a pair of genes. For example, genes sharing transcription factor binding sites may have co-regulated alternative promoters, which leads to the coupling of co-transcription-factor binding and co-alternative splicing. Besides alternative promoters, downstream alternative exons may also be involved in the coupling because alternative promoters have been reported to be correlated with downstream alternative splicing [52-56]. Thus, the transcription factor binding may be associated with the choice of promoters, while the choice of promoter again is associated with the inclusion or exclusion of a downstream alternative exon. Another explanation is that the conserved sequence elements could be RNA *cis*-elements for alternative splicing regulation instead of DNA *cis*-elements for transcription regulation because the considered promoter region is large (covering -2 kb to 2 kb around transcription start sites). Future work will need to explore the detailed mechanisms. The enrichment of

EG and EE links for genes in the same pathways or having the same GO terms also suggest that we can predict gene functions by considering neighboring genes in splicing regulatory networks.

Several groups published a few sets of transcriptome data from high-throughput sequencing techniques while this manuscript was under preparation. Such techniques improve the accuracy of expression level measurement and increase the efficiency of identifying novel alternative exons as exon junctions are sequenced. Our proposed methods can be directly applied on transcriptome RNA-Seq data to identify EE and EG links more accurately.

Conclusion

We propose a partial correlation analysis approach, pCastNet, to reconstruct EE and EG networks. EE and EG networks are part of alternative splicing regulatory networks. We confirmed that pCastNet can effectively identify EG and EE links through studying known alternative exons, conservation patterns, relative positions, and functional annotations, and by RT-PCR experiments. We also found that genes with EG or EE links with each other tend to have similar functions or are in the same pathways, and genes with EG or EE links tend to share *cis*-regulatory motifs in promoter regions and 3' untranslated regions. Through these networks, we can gain a better understanding of the role of alternative splicing in the gene regulatory network.

Materials and methods

Exon array pre-processing

Exon array (Human Exon 1.0 ST, Mouse Exon 1.0 ST, Rat Exon 1.0 ST) data for 11 tissues were downloaded from the Affymetrix website [29]. The profiled tissues for human include breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testis, and thyroid. The tissues for mouse and rat include brain, embryo, heart, kidney, liver, lung, muscle, ovary, spleen, testicle, and thymus. There are three replicates for each tissue. The probe intensities were quantile normalized and were adjusted based on the median intensity of probes with similar GC content. The PLIER method was used to summarize the probe-set-level intensity. The iter-PLIER method was used to summarize the gene-level intensity by iteratively calling PLIER with the core probe sets (that is, RefSeq supported) that correlate with signal estimates. In the design of Affymetrix exon arrays, gene annotations from databases were projected onto the genome to infer transcript clusters and exon clusters. A transcript cluster roughly corresponds to a gene. In many cases, an exon cluster represents a true biological exon and it acts as one probe selection region. In other cases, an exon cluster represents the union of multiple overlapping exons possibly due to alternative splice sites. Such exon clusters were further fragmented into multiple probe selection regions according to the hard

edges (for example, splice sites). In exon array, hard edge was defined as the end of the sequence that defines the boundary of a probe selection region and cannot be extended beyond the border by other annotation evidence. The probe set annotation, the transcript cluster annotation, and the exon annotation were downloaded from the Affymetrix website (version hg18 for human, mm9 for mouse, and rn4 for rat) [29]. To avoid knowledge bias, only core exons based on RefSeq transcripts or full-length mRNAs were considered. Note that 'core probe sets' in Affymetrix exon arrays only means that they are supported by RefSeq transcripts or full-length mRNAs; they do not contain any information about whether they are alternative exons or constitutive exons.

Presence or absence of probe sets was determined by a 'detection above background' *p*-value threshold of 0.05. Genes with more than 50% 'present' core probe sets were called 'present'. The following filtering procedures were performed for probe sets: remove probe sets that are not core probe sets; remove probe sets whose genes are present in < 11 arrays out of 33 arrays (11 tissues × 3 replicates) or whose genes are mapped to more than one Entrez gene ID; remove probe sets that are present in < 11 arrays out of 33 arrays; remove probe sets with (Maximum intensity)/(Minimum intensity) < 5. After the above procedures, a total of 97,293 probe sets corresponding to 76,038 exon clusters and 10,366 transcript clusters remain for human; a total of 102,729 probe sets corresponding to 82,145 exon clusters and 10,765 transcript clusters remain for mouse; and a total of 45,691 probe sets corresponding to 40,082 exon clusters and 5,077 transcript clusters remain for rat. The average intensity across the three replicates was log₂ transformed and used as the intensity level for each tissue.

Correlation and partial correlation calculation

The Pearson correlation coefficient is denoted as *r_{ab}* between variable *a* and variable *b*. The first-order partial correlation coefficient between *a* and *b* conditioning on *c* is [57]:

$$r_{ab \cdot c} = \frac{r_{ab} - r_{ac}r_{bc}}{\sqrt{(1-r_{ac}^2)(1-r_{bc}^2)}}$$

The second-order partial correlation coefficient between *a* and *b* conditioning on *c* and *d* is:

$$r_{ab \cdot cd} = \frac{r_{ab \cdot c} - r_{ad \cdot c}r_{bd \cdot c}}{\sqrt{(1-r_{ad \cdot c}^2)(1-r_{bd \cdot c}^2)}}$$

Note that *r_{ab \cdot cd}* = *r_{ab \cdot dc}* theoretically (see proof in Additional data file 6). In pCastNet, for GG associations, the Pearson correlation coefficient between gene 1 (*g₁*) and gene 2 (*g₂*) was calculated as *r_{g₁g₂}*. For EG associations, the partial correlation coefficient between an exon (*e₁*) of gene 1 and gene 2 (*g₂*) conditioning on gene1 (*g₁*) was calculated as *r_{e₁g₂ \cdot g₁}*. The

Pearson correlation was calculated as *r_{e₁g₂}*. For EE associations, the Pearson correlation *r_{e₁e₂}* and the partial correlations *r_{e₁e₂ \cdot g₁}*, *r_{e₁e₂ \cdot g₂}*, and *r_{e₁e₂ \cdot g₁g₂}* were calculated according to the above equations.

Simulation studies

In simulation studies, five genes were considered. Each of them has five constitutive exons and one alternative exon. The intensity data of the five constitutive exons of gene *g* (*g* = 1, ..., 5) in tissue *t* (*t* = 1, ..., 10) were simulated according to the normal distribution *N*(*μ_{gt}*, *σ*²), where *μ_{gt}* is a value sampled from the range 4 to 9 and it is different for different genes *g* and different tissues *t*. All of the five alternative exons have the same inclusion rate relative to their genes: (*τ₁*, *τ₂*, ..., *τ₁₀*) = (0.1, 0.2, ..., 1.0) for the 10 tissues. Thus, the intensity data of the alternative exon of gene *g* was simulated according to the normal distribution *N*(*μ_{gt}τ_t*, *σ*²). The gene-level intensity was estimated as the average value of the constitutive exons. pCastNet and the NI-based approach were conducted to calculate the correlations and partial correlations between exons belonging to different genes. Three scenarios were considered (*σ* = 1, 2, or 4). For each scenario, 1,000 simulations were performed and the average true positive rate and the average false positive rate were calculated according to different correlation thresholds.

Conditional false discovery rate control

For GG associations, the test statistic:

$$t = r_{g_1g_2} \sqrt{(n-2)/(1-r_{g_1g_2}^2)}$$

is converted to z-value:

$$z = \Phi^{-1}(G_o(t))$$

G_o(*t*) is the null cumulative distribution function for the *t*-values, *Φ*⁻¹ is the inverse function of the cumulative distribution function of standard normal, and *n* is the number of tissues being surveyed. Under the null hypothesis that there is no correlation between gene *g₁* and gene *g₂*, *t* follows a Student *t* distribution with degrees of freedom *n* - 2 and *z* follows the standard normal distribution.

For EG association, the *t* values are:

$$t_1 = r_{e_1g_2} \sqrt{(n-2)/(1-r_{e_1g_2}^2)}$$

and

$$t_2 = r_{e_1g_2 \cdot g_1} \sqrt{(n-3)/(1-r_{e_1g_2 \cdot g_1}^2)}$$

They are converted to z-values:

$$z_1 = \Phi^{-1}(G_{o1}(t_1))$$

and

$$z_2 = \Phi^{-1}(G_{02}(t_2))$$

Under the null hypothesis, t_1 follows a Student t distribution with degrees of freedom $n - 2$ and t_2 follows a Student t distribution with degrees of freedom $n - 3$. The test statistic is:

$$z_0 = \min(|z_1|, |z_2|)$$

$$p = \text{pr}(Z_0 \geq z_0) = \text{pr}(|Z_1| \geq z_0, |Z_2| \geq z_0)$$

Under the null, Z_1 and Z_2 follow a bivariate normal distribution with the correlation approximated as the sample correlation between Z_1 and Z_2 across different node pairs. The final z-value is:

$$z = \text{sign}(t_1)\Phi^{-1}(1 - p/2)$$

For EE associations, the t values are:

$$t_1 = r_{e_1, e_2} \sqrt{(n-2)/(1-r_{e_1, e_2}^2)}$$

$$t_2 = r_{e_1, e_2 \bullet g_1} \sqrt{(n-3)/(1-r_{e_1, e_2 \bullet g_1}^2)}$$

$$t_3 = r_{e_1, e_2 \bullet g_2} \sqrt{(n-3)/(1-r_{e_1, e_2 \bullet g_2}^2)}$$

$$t_4 = r_{e_1, e_2 \bullet g_2 g_1} \sqrt{(n-4)/(1-r_{e_1, e_2 \bullet g_2 g_1}^2)}$$

They are converted to z-values:

$$z_1 = \Phi^{-1}(G_{01}(t_1))$$

$$z_2 = \Phi^{-1}(G_{02}(t_2))$$

$$z_3 = \Phi^{-1}(G_{03}(t_3))$$

$$z_4 = \Phi^{-1}(G_{04}(t_4))$$

Under the null hypothesis, t_1 follows a Student t distribution with degrees of freedom $n - 2$; t_2 and t_3 follow a Student t distribution with degrees of freedom $n - 3$; t_4 follows a Student t distribution with degrees of freedom $n - 4$. The test statistic is:

$$z_0 = \min(|z_1|, |z_2|, |z_3|, |z_4|)$$

$$p = \text{pr}(Z_0 \geq z_0) = \text{pr}(|Z_1| \geq z_0, |Z_2| \geq z_0, |Z_3| \geq z_0, |Z_4| \geq z_0)$$

Under the null, Z_1, Z_2, Z_3, Z_4 follow a multivariate normal distribution with the correlations approximated as the sample correlations between Z_1, Z_2, Z_3, Z_4 across different node pairs. The final z-value is:

$$z = \text{sign}(t_1)\Phi^{-1}(1 - p/2)$$

For a threshold x , the conditional FDR is estimated as [25]:

$$\text{FDR}(x | A) = \frac{2N\Phi(-x)}{\#\{|z_i| \geq x\}} \left[1 + A \frac{x\phi(x)}{\sqrt{2}\Phi(-x)} \right]$$

where Φ is the standard normal cumulative distribution function, N is the total number of node pairs, ϕ is the standard normal density function, and $A = \frac{2\Phi(1)-1-\#\{z_i \in [-1,1]\}/N}{\sqrt{2}\phi(1)}$. For

a $\text{FDR}(x|A)$, the percentage of true links among all possible node pairs is estimated as $\#\{|z_i| \geq x\}(1 - \text{FDR}(x|A))/N$.

RNA preparation and RT-PCR

Various tissues from adult C57BL mice and embryonic cortex from E16 embryos were dissected and quickly submerged in Trizol (Invitrogen, Carlsbad, CA, USA) followed by immediate tissue homogenization. Total RNA samples were prepared according to manufacturer's protocol. RT-PCR was done as previously described [30]. Primer design was done with the Primer3 online software [58]. Primer sequences were. *Kifap3*, CCTCCAGAATGGAGATGTGG (forward), ACATGGGAGGGGTGATTTTA (reverse); *St7*, GCAGATGCAATAATGCAAAAAG (forward), GTAACAACCATCTCCAGCCTTC (reverse); *Map3k7*, TCTGAAATAGAAGCCAGGATCG (forward), CTTCTCTGAGGTTGGTCCTGAG (reverse); *Slc35b3*, AGCCTTACGGCTGGTACCTT (forward), AGTTTGGTGCAATTGTGCTG (reverse); *Rai14*, TCTCATGCTGGCTTGTGAAA (forward), GTTATTGATCGTGGGGAGGA (reverse). Identities of each RT-PCR product were confirmed by direct sequencing. PCR bands were quantified using ImageQuant TL software (GE Healthcare Bio-Sciences, Piscataway, NJ, USA). The correlation value of the each exon pair was calculated as the Pearson correlation coefficient between the inclusion levels of the two tested exons. Inclusion levels based on the PCR results were calculated as Inclusion form/(Inclusion form + Exclusion form).

Other datasets and analysis

Non-redundant human transcript annotations were assembled based on AceView gene [59], AUGUSTUS gene [60], CCDS gene [61,62], Ensembl gene [61], Geneid gene, Genescan gene [63], MGC gene, N-SCAN gene [64], ORFeome gene, RefSeq gene [62], SGP gene, SIB gene [65], UCSC genes [66], and ASTD gene [67]. The first 13 data sources were downloaded from the UCSC Genome Browser website (hg18) [68] and the last was downloaded from the ASTD database (release 1.0) [69]. A gene may have multiple transcript isoforms. For each exon cluster defined by the Affymetrix exon array, if one exon of a transcript isoform locates in the exon cluster region, the exon cluster is called 'present' in this transcript isoform. If the exon cluster locates in an intron region of a transcript isoform, the exon cluster is called 'spliced out' in this transcript isoform. Exon clusters (or exons for simplicity) were divided into two groups: exons that are present in ≥ 14 transcript isoforms and are not spliced out in any tran-

script isoform; exons that are present in ≥ 7 transcript isoforms and are spliced out in another ≥ 7 transcript isoforms.

The PhastCons conservation score [70] was downloaded from the UCSC Genome Browser (hg18) [68]. The score of each site is the posterior probability that the site is in the conserved state of the phylogenetic hidden Markov model for 17 vertebrates.

For each gene with multiple exons, all of the core exons were sorted according to their genomic coordinates (from 5' to 3'). The relative position of the i -th exon is calculated as $(i - 1)/(n - 1)$, where n is the total number of exons. The relative position ranges from 0 to 1.

Gene sets downloaded from the Molecular Signatures Database [39] belong to four categories: 'c1', positional gene sets for each chromosomal cytogenetic band; 'c2', curated gene sets from pathway databases; 'c3', motif gene sets sharing conserved *cis*-regulatory motifs [71]; 'c4', GO gene sets sharing the same GO term. We removed gene sets without any gene in our final gene list for exon arrays. Category 2 was further divided into 'c2_BioCarta', 'c2_GenMAPP', and 'c2_KEGG'; genes in the same pathways and the pathways were collected from the BioCarta database, the GenMAPP database, or the KEGG database. Category 3 was further divided into: 'c3_promoter_known' and 'c3_promoter_unknown' - genes sharing a motif in the promoter regions (covering -2 kb to 2 kb around transcription start sites) and the motif matches a known transcription factor binding site or the motif does not match any known transcription factor binding site; 'c3_miRNA' - genes sharing a miRNA binding site. Category 4 was divided into: 'c4_bp' - genes sharing biological process ontology terms; 'c4_cc' - genes sharing cellular component ontology terms; 'c4_mf' - genes sharing molecular function ontology terms.

Abbreviations

CLIP: crosslinking immunoprecipitation; EE: exon-exon; EG: exon-gene; FDR: false discovery rate; GG: gene-gene; GO: Gene Ontology; miRNA: microRNA; NI: normalized intensity; pCastNet: partial Correlation analysis of splicing transcriptome Network.

Authors' contributions

LC designed the study, developed the method, and performed the analysis. SZ designed and performed the validation experiments. LC and SZ wrote the paper.

Additional data files

The following additional data files are available with the online version of this paper. Additional data file 1 is a figure showing that the gene-level normalized exon intensity is not

stable when the noise level is high. Additional data file 2 is a diagram showing the possible regulation relationships for EG and EE links. Additional data file 3 is a figure showing Venn diagrams of gene pairs with GG, EG, or EE associations. Additional data file 4 is figure showing that exons with EE links and sharing miRNA binding motifs tend to be enriched at the 3' termini of the genes. Additional data file 5 is a figure showing the expression level of FOX-1 and exons with EG links to FOX-1. Additional data file 6 is a proof to show that $r_{ab-cd} = r_{ab-dc}$ theoretically.

Acknowledgements

We would like to specially thank Doug Black (HHMI, UCLA) for his scientific advice on the study and generous support for the validation experiments. We would also like to thank the reviewers for their insightful comments. This research was conducted while LC was an AFAR Research Grant recipient.

References

- Johnson JM, Castle J, Garrett-Engle 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**:2141-2144.
- Kan Z, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** *Genome Res* 2001, **11**:889-900.
- Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease.** *Genes Dev* 2003, **17**:419-437.
- Garcia-Blanco MA, Baraniak AP, Lasda EL: **Alternative splicing in disease and therapy.** *Nat Biotechnol* 2004, **22**:535-546.
- Blencowe BJ: **Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.** *Trends Biochem Sci* 2000, **25**:106-110.
- Li Q, Lee JA, Black DL: **Neuronal regulation of alternative pre-mRNA splicing.** *Nat Rev Neurosci* 2007, **8**:819-831.
- Jiang ZH, Wu JY: **Alternative splicing and programmed cell death.** *Proc Soc Exp Biol Med* 1999, **220**:64-72.
- Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** *Annu Rev Biochem* 2003, **72**:291-336.
- Matlin AJ, Clark F, Smith CW: **Understanding alternative splicing: towards a cellular code.** *Nat Rev Mol Cell Biol* 2005, **6**:386-398.
- Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y: **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7**:325.
- 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**:R64.
- Ip JY, Tong A, Pan Q, Topp JD, Blencowe BJ, Lynch KW: **Global analysis of alternative splicing during T-cell activation.** *Rna* 2007, **13**:563-572.
- Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM: **Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines.** *Nat Genet* 2008, **40**:1416-1425.
- Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S, Poliakov A, Dubchak I, Blume JE, Conboy JG: **A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing.** *Nucleic Acids Res* 2007, **35**:4845-4857.
- Xing Y, Stoilov P, Kapur K, Han A, Jiang H, Shen S, Black DL, Wong WH: **MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays.** *Rna* 2008, **14**:1470-1479.
- Hung LH, Heiner M, Hui J, Schreiner S, Benes V, Bindereif A: **Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis.** *Rna* 2008, **14**:284-296.

17. Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H, Blume JE, Hudson TJ, Sladek R, Majewski J: **Heritability of alternative splicing in the human genome.** *Genome Res* 2007, **17**:1210-1218.
18. Yeo GW, Xu X, Liang TY, Muotri AR, Carson CT, Coufal NG, Gage FH: **Alternative splicing events identified in human embryonic stem cells and neural progenitors.** *PLoS Comput Biol* 2007, **3**:1951-1967.
19. Brody JP, Williams BA, Wold BJ, Quake SR: **Significance and statistical errors in the analysis of DNA microarray data.** *Proc Natl Acad Sci USA* 2002, **99**:12975-12978.
20. Reverter A, Chan EK: **Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks.** *Bioinformatics* 2008, **24**:2491-2497.
21. de la Fuente A, Bing N, Hoeschele I, Mendes P: **Discovery of meaningful associations in genomic data using partial correlation coefficients.** *Bioinformatics* 2004, **20**:3565-3574.
22. Magwene PM, Kim J: **Estimating genomic coexpression networks using first-order conditional independence.** *Genome Biol* 2004, **5**:R100.
23. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
24. Elo LL, Jarvenpaa H, Oresic M, Laheasmaa R, Aittokallio T: **Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process.** *Bioinformatics* 2007, **23**:2096-2103.
25. Efron B: **Correlation and large-scale simultaneous significance testing.** *J Am Stat Assoc* 2007, **102**:93-103.
26. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100**:9440-9445.
27. Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *J Roy Stat Soc B Met* 1995, **57**:289-300.
28. Chen L, Tong T, Zhao H: **Considering dependence among genes and markers for false discovery control in eQTL mapping.** *Bioinformatics* 2008, **24**:2015-2022.
29. **Affymetrix** [http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx]
30. Chen L, Zheng S: **Identify alternative splicing events based on position-specific evolutionary conservation.** *PLoS ONE* 2008, **3**:e2806.
31. Sorek R, Ast G: **Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.** *Genome Res* 2003, **13**:1631-1637.
32. Baek D, Davis C, Ewing B, Gordon D, Green P: **Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters.** *Genome Res* 2007, **17**:145-155.
33. Sun H, Palaniswamy SK, Pohar TT, Jin VX, Huang TH, Davuluri RV: **MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data.** *Nucleic Acids Res* 2006, **34**:D98-103.
34. Takeda J, Suzuki Y, Nakao M, Kuroda T, Sugano S, Gojobori T, Imanishi T: **H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational.** *Nucleic Acids Res* 2007, **35**:D104-109.
35. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, Kondo H, Wagatsuma M, Murakawa K, Ishida S, Ishibashi T, Takahashi-Fujii A, Tanase T, Nagai K, Kikuchi H, Nakai K, et al.: **Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes.** *Genome Res* 2006, **16**:55-65.
36. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16**:1-10.
37. Tian B, Hu J, Zhang H, Lutz CS: **A large-scale analysis of mRNA polyadenylation of human and mouse genes.** *Nucleic Acids Res* 2005, **33**:2011-212.
38. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4**:P3.
39. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102**:15545-15550.
40. Ponthier JL, Schlupein C, Chen W, Lersch RA, Gee SL, Hou VC, Lo AJ, Short SA, Chasis JA, Winkelmann JC, Conboy JG: **Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16.** *J Biol Chem* 2006, **281**:12468-12474.
41. Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL: **Homologues of the *Caenorhabditis elegans* Fox-1 protein are neuronal splicing regulators in mammals.** *Mol Cell Biol* 2005, **25**:10005-10016.
42. Nakahata S, Kawamoto S: **Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities.** *Nucleic Acids Res* 2005, **33**:2078-2089.
43. Jin Y, Suzuki H, Maegawa S, Endo H, Sugano S, Hashimoto K, Yasuda K, Inoue K: **A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG.** *EMBO J* 2003, **22**:905-912.
44. Xie J, Black DL: **A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels.** *Nature* 2001, **410**:936-939.
45. Hirokawa N, Noda Y: **Intracellular transport and kinesin superfamily proteins, KIFs: structure, function, and dynamics.** *Physiol Rev* 2008, **88**:1089-1118.
46. Zenklusen JC, Conti CJ, Green ED: **Mutational and functional analyses reveal that ST7 is a highly conserved tumor-suppressor gene on human chromosome 7q31.** *Nat Genet* 2001, **27**:392-398.
47. Vincent JB, Herbrick JA, Gurling HM, Bolton PF, Roberts W, Scherer SW: **Identification of a novel gene on chromosome 7q31 that is interrupted by a translocation breakpoint in an autistic individual.** *Am J Hum Genet* 2000, **67**:510-514.
48. Li MG, Katsura K, Nomiya H, Komaki K, Ninomiya-Tsuji J, Matsumoto K, Kobayashi T, Tamura S: **Regulation of the interleukin-1-induced signaling pathways by a novel member of the protein phosphatase 2C family (PP2Cepsilon).** *J Biol Chem* 2003, **278**:12013-12021.
49. Yamaguchi K, Shirakabe K, Shibuya H, Irie K, Oishi I, Ueno N, Taniguchi T, Nishida E, Matsumoto K: **Identification of a member of the MAPKKK family as a potential mediator of TGF-beta signal transduction.** *Science* 1995, **270**:2008-2011.
50. Kondo M, Osada H, Uchida K, Yanagisawa K, Masuda A, Takagi K, Takahashi T: **Molecular cloning of human TAK1 and its mutational analysis in human lung cancer.** *Int J Cancer* 1998, **75**:559-563.
51. Kim RS, Ji H, Wong WH: **An improved distance measure between the expression profiles linking co-expression and co-regulation in mouse.** *BMC Bioinformatics* 2006, **7**:44.
52. Wang Y, Newton DC, Robb GB, Kau CL, Miller TL, Cheung AH, Hall AV, VanDamme S, Wilcox JN, Marsden PA: **RNA diversity has profound effects on the translation of neuronal nitric oxide synthase.** *Proc Natl Acad Sci USA* 1999, **96**:12150-12155.
53. Pecci A, Viegas LR, Baranao JL, Beato M: **Promoter choice influences alternative splicing and determines the balance of isoforms expressed from the mouse bcl-X gene.** *J Biol Chem* 2001, **276**:21062-21069.
54. Logette E, Wotawa A, Solier S, Desoche L, Solary E, Corcos L: **The human caspase-2 gene: alternative promoters, pre-mRNA splicing and AUG usage direct isoform-specific expression.** *Oncogene* 2003, **22**:935-946.
55. Landry JR, Mager DL, Wilhelm BT: **Complex controls: the role of alternative promoters in mammalian genomes.** *Trends Genet* 2003, **19**:640-648.
56. Parra MK, Tan JS, Mohandas N, Conboy JG: **Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene.** *EMBO J* 2008, **27**:122-131.
57. Crawley MJ: *Statistics: An Introduction Using R* John Wiley and Sons; 2005.
58. **Primer3** [<http://frodo.wi.mit.edu>]
59. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation.** *Genome Biol* 2006, **7 Suppl 1**:S12.1-S12.14.
60. Stanke M, Steinkamp R, Waack S, Morgenstern B: **AUGUSTUS: a web server for gene finding in eukaryotes.** *Nucleic Acids Res* 2004, **32**:3309-312.
61. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C,

- Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, et al.: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30**:38-41.
62. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**:D61-65.
 63. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.
 64. van Baren MJ, Brent MR: **Iterative gene prediction and pseudo-gene removal improves genome annotation.** *Genome Res* 2006, **16**:678-685.
 65. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update.** *Nucleic Acids Res* 2004, **32**:D23-26.
 66. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes.** *Bioinformatics* 2006, **22**:1036-1046.
 67. Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA: **ASD: a bioinformatics resource on alternative splicing.** *Nucleic Acids Res* 2006, **34**:D46-55.
 68. **UCSC Genome Browser** [<http://genome.ucsc.edu>]
 69. **Alternative Splicing Database Project** [<http://www.ebi.ac.uk/asd/>]
 70. Felsenstein J, Churchill GA: **A hidden Markov model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13**:93-104.
 71. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**:338-345.