DNA Research **15**, 63–72, (2008) doi:10.1093/dnares/dsm036

Computational Analysis of Full-length cDNAs Reveals Frequent Coupling Between Transcriptional and Splicing Programs

Tzu-Ming Chern, Nicodeme Paul, Erik van Nimwegen, and Mihaela Zavolan*

Division of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50-70, Basel CH-4056, Switzerland

(Received 31 August 2007; accepted on 25 December 2007; published online 14 February 2008)

Abstract

High-throughput sequencing studies revealed that the majority of human and mouse multi-exon genes have multiple splice forms. High-density oligonucleotide array-based measurements have further established that many exons are expressed in a tissue-specific manner. The mechanisms underlying the tissue-dependent expression of most alternative exons remain, however, to be understood. In this study, we focus on one possible mechanism, namely the coupling of (tissue specific) transcription regulation with alternative splicing. We analyzed the FANTOM3 and H-Invitational datasets of full-length mouse and human cDNAs, respectively, and found that in transcription units with multiple start sites, the inclusion of at least 15% and possibly up to 30% of the 'cassette' exons correlates with the use of specific transcription start sites (TSS). The vast majority of TSS-associated exons are conserved between human and mouse, yet the conservation is weaker when compared with TSS-independent exons. Additionally, the currently available data only support a weak correlation between the probabilities of TSS association of orthologous exons. Our analysis thus suggests frequent coupling of transcriptional and splicing programs, and provides a large dataset of exons on which the molecular basis of this coupling can be further studied.

Key words: alternative splicing; transcription initiation

1. Introduction

The most common form of splice variation is the inclusion of an exon in some, but not all, of the transcripts of a gene.[1,2] Numerous studies have been dedicated to specific instances of such exons, which are known by various names such as 'cassette', 'alternative', 'skipped', and 'cryptic' exons. The regulatory signals leading to the inclusion or exclusion of a cassette exon also form a vast topic of research. Computational studies are converging toward the view that cassette exons are generally less recognizable to the splicing machinery than constitutive exons due to their shorter length,[3] lower strength of splice sites,[4] and poor representation of general splice enhancers,[1,5] The tissue-specific inclusion of these exons appears to be dependent upon specific regulatory elements, at least some of which are located in the strongly conserved intronic regions that flank the cassette exons.[2,6–8]

One attractive hypothesis concerning the mechanism of tissue-specific inclusion of cassette exons involves the direct coupling between tissue-dependent transcription and splicing. It has been shown, for instance, that the promoter from which transcription is initiated can affect the inclusion of downstream exons through the recruitment of transcription factors and co-activators that modulate the elongation rate of RNA polymerase II[9,10] (kinetic model). In turn, a low polymerase elongation rate can promote the

Edited by Osamu Ohara

* To whom correspondence should be addressed. Tel. +41 61-267-1576. Fax. +41 61-267-1584. E-mail: mihaela.zavolan@unibas.ch

inclusion of some exons that are skipped when the elongation rate is high, as has been shown for the fibronectin EDI exon.[11] Alternatively, protein factors that are recruited at the stage of transcription initiation may interact with splicing factors (recruitment model). For instance, the inclusion of the fibronectin EDII exon, normally promoted by the SRp40 protein,[12] is inhibited when transcription is initiated at a promoter containing a binding site for the PPARγ transcription factor. This is because the PPARγ transcription factor recruits an SRp40 inhibitor, the PGC-1 co-activator.[13] The role of the promoter architecture on internal splicing has also been demonstrated for the cystic fibrosis transmembrane regulator[14] and for the steroid-sensitive genes.[15] Such studies have been limited, however, to very few genes, prompting us to evaluate the extent to which transcriptional and splicing programs appear to be coordinated at the level of the whole transcriptome.

2. Materials and methods

2.1. *Transcript mapping and splice analysis*

The sequences used in the study consist of the 102 797 Fantom3 mouse cDNAs,[16] 52 070 mouse mRNAs from Genbank, and 167 992 human cDNAs from version 3.0 (http://www.h-invitational.jp/) of the H-Invitational project.[17] We have mapped the mouse cDNAs to the mm7 assembly of the mouse genome and the human cDNAs to the hg18 assembly of the human genome, both available from the University of California at Santa Cruz.

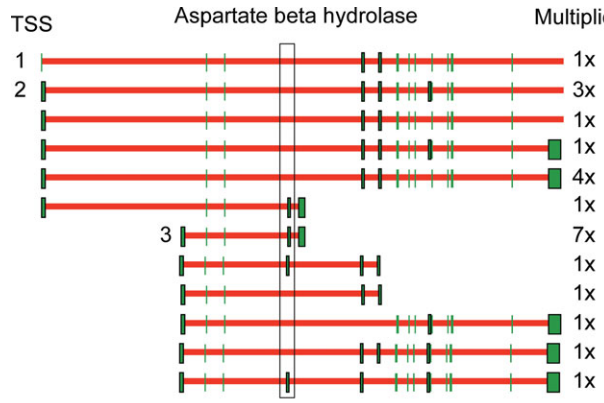
For the identification of splice variants, we used the automated splicing analysis pipeline that we have previously developed.[1,18] Briefly, we first mapped all cDNAs to their respective genome using our spliced alignment algorithm (SPA).[19] To avoid biases from transcripts that are badly mapped due to a high rate of sequencing errors or erroneous assembly, we select only those transcripts that have at least 75% of their nucleotides mapped to the genome, with at least 95% identity or less than ten mismatches in each exon. This procedure yielded 132 681 mouse and 110 978 human mapped transcripts, which we clustered such that the mapping of each transcript in a cluster shares at least one exonic nucleotide with at least one other transcript in the cluster.[1,18,20] We obtained 42 407 mouse and 22 116 human clusters (transcription units) that we analyzed for splice variation. We were interested only in cassette exons with no other form of splice variation. We identified these as internal exons that were completely contained in an intron implied by the mapping of another transcript in the cluster, having the same splice boundaries in all transcripts in which they were included. Our final mouse dataset consisted of 29 416 transcripts and 4 964 internal cassette exons, and the human dataset of 79 030 transcripts and 11 664 internal cassette exons.

2.2. *Quantifying the evidence for coupling between the choice of transcription start sites and the inclusion/exclusion of internal exons*

For each transcription unit, we first identified (1) the set of internal cassette exons and (2) the set of transcription start sites (TSSs). The cassette exon annotation was determined as outlined above. To identify different TSSs used within a transcription unit, we had to define precisely what we mean by a unique TSS. The analysis of mammalian TSSs by Carninci et al.[21] has shown that most TSSs show some amount of variability. Especially at TSSs located in CpG islands, one finds transcripts starting from many different nearby sites covering tens and sometimes hundreds of nucleotides. We, therefore, needed to group transcripts whose apparent start sites were 'near' each other and then identify different TSSs with the different clusters of apparent start sites. We decided to take a conservative approach to this clustering of apparent start sites in a transcription unit by considering all transcripts that started within the same exon to derive from the same TSS. That is, in our analysis different TSSs correspond to *different initial exons* in the transcripts.

In addition, we tested the validity of our results, on a separate dataset in which we use only transcripts whose initial exons were confirmed by CAGE tag data.[22] In the latter case, the initial exon was considered confirmed as a TSS if one or more CAGE tags were found within 100 bp of the start of the exon in the genome. Since the results did not change, and the requirement of CAGE validation of TSSs reduced the size of our data-set significantly, we did not use the CAGE validated TSSs further.

For each internal cassette exon, we collected all transcripts that could have included the exon as an internal exon, i.e. those transcripts that contained exons both upstream and downstream of the genomic location of the exon in question, and determined the TSS that was used for each of these transcripts. We thus obtained a list of TSSs that were used in the set of transcripts in which the cassette exon could have been included. For further analyses, we kept only cassette exons for which multiple TSSs were identified. We then counted, for each TSS in the list, how many transcripts starting from this TSS included the exon, and how many transcripts excluded the exon. For each internal exon, we thus obtained counts of the number of times each TSS was used in a transcript whose locus covered the



TSS	Exon inclusion	Exon skipping
1	0	1
2	1	9
3	9	3

Figure 1. Example of an internal cassette exon whose inclusion is strongly correlated with specific TSSs. Exons are indicated by (green) boxes and introns by red lines. Cassette exons are shown with a black frame. The internal cassette exon that has a high probability of TSS association is indicated by the thin black rectangle. The number of times a specific TSS-splicing pattern was observed in the data is indicated by the multiplicity on the right. For simplicity, we show only those transcripts whose genomic locus contains the cassette exon, and we truncate the first three transcripts past the TSS-associated cassette exon. Table 1 summarizes the data in the figure, indicating how many times the exon was included and how many times skipped when a particular TSS was used.

exon, and the number of times the exon was included and excluded with each of the TSSs.

To identify exons whose inclusion depends on which TSS was used, we used a Bayesian model selection procedure that compared the probabilities of the observed counts under a TSS-independent model and a TSS-dependent model. Considering a particular cassette exon, let t_i denote the total number of times TSS i was used, n_i the number of times the exon was included when TSS i was used, t the total number of transcripts, and n the total number of times that the exon was included. For the independent model, we assumed that the n inclusions are distributed at random among the t transcripts. Under this model, the probability of the observed counts $\{n_i\}$, given the counts $\{t_i\}$, and n is

$$P_{\text{indep}}(\{n_i\}|n,\{t_i\}) = \frac{n!(t-n)!}{t!}\prod_i \frac{t_i!}{n_i!(t_i-n_i)!}. \quad (1)$$

For the dependent model, we assumed that the rates of inclusion and exclusion for the different TSSs are set by some unknown mechanism. Given our general ignorance about the mechanism or mechanisms determining these rates, there is no reason to assume that any set of counts $\{n_i\}$ is more or less likely than any other set of counts. We, therefore, assumed that all possible counts $\{n_i\}$ that are consistent with the totals $\{t_i\}$ and the total number of inclusions n are all equally likely, meaning that

$$P_{\text{dep}}(\{n_i\}|n,\{t_i\}) = \frac{1}{C(n,\{t_i\})}, \quad (2)$$

where $C(n, \{t_i\})$ is the total number of different sets of counts $\{n_i\}$ that are possible, given the totals n and $\{t_i\}$. The total count numbers $C(n, \{t_i\})$ can be determined recursively. Let $C_i(r)$ be the number of different inclusion counts n_1 through n_i that can be assigned to TSSs 1 through i, such that r of the n total inclusions remain. We have the following recursion relation for $C_i(r)$:

$$C_i(r) = \sum_{n_i=0}^{t_i} C_{i-1}(r+n_i). \quad (3)$$

We initialize the recursion by setting

$$C_0(r) = \delta_{rn}, \quad (4)$$

that is, before we assign counts to any TSS there have to be precisely n inclusions left. Once we arrive at the last (pth) TSS, our count $C(n, \{t_i\})$ is given by $C_p(0)$, i.e. there should be no inclusions left.

To estimate the total fraction f of cassette internal exons whose inclusion is dependent on TSSs, we calculated the probability $P(D|f)$ of the data of all cassette exons assuming that a fraction f was dependent. Let $P_{\text{indep}}(k)$ and $P_{\text{dep}}(k)$ denote the probabilities of the counts for the kth cassette exon given the independent and dependent model, respectively. For each exon k, these quantities are computed according to Equations (1) and (2). We then have

$$P(D|f) = \prod_k \left[P_{\text{indep}}(k)(1-f) + P_{\text{dep}}(k)f\right]. \quad (5)$$

Using a uniform prior over f the posterior probability $P(f|D)$ for the fraction of dependent exons given the data simply becomes

$$P(f|D) = \frac{P(D|f)}{\int P(D|\tilde{f})\mathrm{d}\tilde{f}}. \quad (6)$$

This distribution is shown as the solid line in Fig. 2. We calculated the expectation value $\langle f\rangle = \int f\, P(f \mid D)$

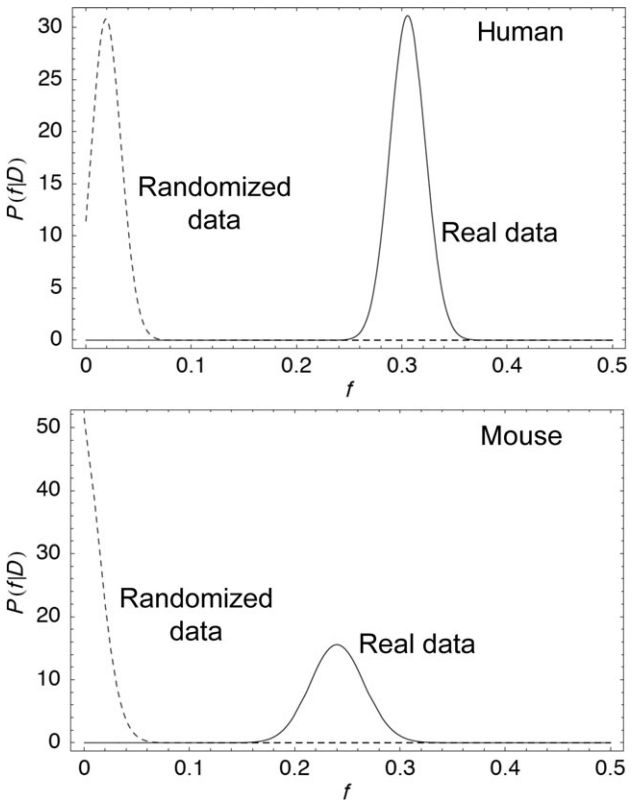


Figure 2. The posterior probability $P(f|D)$ that the inclusion of cassette exons is dependent on the TSS in a fraction f of all cassette exons (solid line). The dashed line shows the same distribution $P(f|D_{rand})$ for a randomized dataset containing the same marginal counts of inclusion and TSS usage for each exon. The upper panel shows the human data, and the lower panel the mouse data.

df and using $\langle f \rangle$ as a prior probability of the independent model we computed, for each individual exon, the posterior probability that the inclusion of the exon is TSS associated.

We also generated a randomized dataset D_{rand} by, for each exon, randomly distributing the n inclusions of the exon among the different TSSs, in such a way that the total number of transcripts t_i for each TSS i stays the same. That is, the data D_{rand} were generated in accordance with the independent model, keeping the total inclusion counts n, and total TSS counts $\{t_i\}$ of the real data. The distribution $P(f|D_{rand})$ is shown as the dotted line in Fig. 2.

2.3. *Extraction of constitutive exons*

We used our database to identify internal exons that were included in more than ten transcripts and did not have any splice variation. We obtained 5136 and 6377 such exons for mouse and human, respectively, and we used these as internal constitutive exons.

2.4. *Computation of the exon conservation statistics*

We used the whole-genome alignments provided by the University of California, at Santa Cruz (http://hgdownload.cse.ucsc.edu/goldenPath/mm7/vsHg18/axtNet/ and http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsmm7/axtNet/) to extract alignments of mouse exons in our dataset with the corresponding orthologous regions from the human genome, and of human exons with the corresponding orthologous regions from the mouse genome. We found orthologous human regions for 438 of the 496 TSS-associated, 469 of the 496 TSS-independent, and 5122 of the 5136 constitutive internal mouse exons. Similarly, we found orthologous mouse regions for 1003 of the 1166 TSS-associated, 1078 of the 1166 TSS-independent, and 6344 of the 6377 constitutive internal human exons. On the basis of the extracted alignments, we determined the fraction of all mouse exon nucleotides that are perfectly conserved in human, and the fraction of human exon nucleotides that are conserved in mouse.

To compute the correlation between posterior probabilities of TSS association in human and mouse, we used the following procedure. We started with cassette exons from transcription units with multiple TSSs from human. On the basis of the whole-enome alignments, we found the orthologous mouse exons as described above, and then we intersected the coordinates of these orthologs with the coordinates of mouse cassette exons that were part of transcription units with multiple TSSs. This procedure gave us the list of orthologous cassette exons that were part of transcription units with multiple TSSs in both human and mouse. We checked that we obtain the same list of exons if we start from the mouse cassette exons and compute their human orthologs. We did not set a threshold of minimal conservation between orthologous exons, as the dataset is already relatively small, and our previous results were not sensitive to the precise threshold of conservation that we used.

2.5. *Computation of the average distance to an inclusion-promoting TSS and to a skipping-promoting TSS*

We identified all transcripts in which a particular cassette exon was included, and for each of them, we determined the distance in the genome between the TSS and the start of the cassette exon. We then averaged these distances to obtain the average distance of the exon to an inclusion-promoting TSS. Similarly, we identified all transcripts in which a particular cassette exon was skipped, and we computed the average distance of the exon to a skipping-promoting TSS.

2.6. Analysis of 5′ EST data

To address the issue of biases in the coverage of gene structures that may have been introduced by the selection of clones for full-length cDNA sequencing, we have performed the same analysis using only 5′ EST sequences. On the basis of the October 2007 UniGene database, we extracted 3 738 929 human 5′ EST sequences. We removed from this set those ESTs with low sequence quality (over 3% ambiguous nucleotides), we then trimmed the polyA tails and discarded those ESTs whose length was <95 after polyA tail removal (using the trimpoly program of the SeqClean package from http://compbio.dfci.harvard.edu/tgi/software/). We thus obtained 3 700 028 ESTs, which we mapped to the hg18 assembly of the human genome using our SPA program.[19] After mapping, we retained for further analysis only ESTs, which contained at least two exons, were mapped with overall 95% identity, and whose every exon was mapped with at least 98% identity to the human genome. This selection left 1 276 669 good quality ESTs, which we clustered based on exon overlap and annotated for splice variation as described in Section 2.1.

3. Results and discussion

3.1. The inclusion/skipping of internal exons is correlated with the usage of specific TSSs

We used a Bayesian approach to estimate the fraction f of internal cassette exons whose inclusion depends on the choice of TSS. We thus considered two models: the first assumes that the probability of exon inclusion is independent of the TSS used to transcribe the pre-mRNA, i.e. the probability of exon inclusion is the same for all TSSs, and the second assumes that for each TSS there is an independent probability of exon inclusion, which can be different between different TSSs. For a given f, we can write the probability of the data as

$$P(D|f) = \prod_{k\in\text{exons}} \left[P_{\text{indep}}(k)(1-f) + P_{\text{dep}}(k)f\right], \qquad (7)$$

with $P_{\text{dep}}(k)$ and $P_{\text{indep}}(k)$ being the probabilities of the data for cassette exon k under the dependent and independent models, respectively. In order to obtain these probabilities we collected, for each internal exon, the set of transcripts in the dataset that could have contained the exon, i.e. those transcripts that contain exons both upstream and downstream of the genomic location of the cassette exon in question. We divided this set of transcripts into groups that use the same TSSs, and counted the number of transcripts in which the exon was included, and the number of transcripts in which the exon was excluded in each TSS group. A specific example of this computation is shown in Fig. 1.

As described in Section 2.2, we can use Equation (7) and Bayes' theorem to calculate the posterior probability $P(f|D)$ that a fraction f of all cassette exons is dependent on the TSS. The distributions $P(f|D)$ obtained for both the human and mouse data are shown as solid lines in Fig. 2. They suggest that the inclusion of about 24% (99% posterior probability interval 17.4−30.6%) of all cassette exons in our mouse dataset and 30% (99% posterior probability interval 26.2−34.9%) of all cassette exons in our human dataset is dependent on the TSSs of the corresponding transcripts. To additionally test the statistical significance of this result, we created randomized datasets D_{rand} by permuting, for each cassette exon, the inclusions and exclusions among the TSSs in such a way that the total number of times each TSS was used, and the total number of times the exon was included remained unchanged. The posterior distributions $P(f|D_{\text{rand}})$ obtained by applying the Bayesian procedure to the randomized data are shown as dashed lines in Fig. 2. These distributions show that the Bayesian procedure correctly infers that the inclusion of $<5\%$ (and likely none) of the cassette exons in D_{rand} depends on TSS. Moreover, Fig. 2 shows the striking difference between the real and randomized data, which is due to the enrichment of cassette exons with high posterior probability of TSS dependency in the real data.

Using the estimated fraction $\langle f \rangle$ (the mean of the posterior distribution $P(f|D)$) as a prior that the inclusion of a cassette exon depends on TSS, we computed the posterior probability that the inclusion of each exon in our dataset is TSS dependent (see Section 2.2). These data are given in the Supplementary table, and can also be further explored using the server that we established at http://www.spaed.unibas.ch/Promoter_data/TSS_cassette_exons_spaed_human.html and http://www.spaed.unibas.ch/Promoter_data/TSS_cassette_exons_spaed_mouse.html.

Fig. 1 shows the cassette exon with the highest posterior probability of TSS dependence in our mouse data. The exon belongs to the gene aspartate beta hydrolase has a posterior probability of TSS dependency of 0.89, and is indicated in the figure by the thin black rectangle. For clarity, we showed only the transcripts whose genomic loci contain the cassette exon, and we also truncated some of the transcripts after the exon in question. There are three different TSSs upstream of the exon, and a total of 23 transcripts that could have included this exon. Of the 11 transcripts originating in the two upstream TSSs,

only one includes the cassette exon. In contrast, nine of the 12 transcripts originating from the third TSS include the cassette exon.

One may wonder to what extent our results are affected by imperfect efficiency of full length cDNA capture. That is, if a significant fraction of the cDNAs is not full length, the apparent TSSs for these transcripts would be incorrect, and one may wonder how they would influence our results. We have addressed this question by performing the same analysis using only transcripts whose start site was confirmed by CAGE tag data,[22] and obtained essentially the same results. However, since requiring additional confirmation of TSSs substantially reduces the sizes of our datasets, we did not use this dataset further.

It is important to note that we do not need to find the precise locations of the TSSs which may in fact be much less precise than initially thought,[21] but we only need to separate our transcripts into sets that arose from the same transcription initiation regions, controlled by specific sets of regulatory signals. We decided to simply assume that transcripts with the same initial exon arose from the same TSS and that transcripts with different initial exons arose from different TSSs. Two types of errors may occur in this classification. First, transcripts that arose from two different, but nearby TSSs may be assigned to a single common TSS. Second, if a transcript is severely truncated due to cloning or sequencing errors, it may appear to start in an exon which is downstream from its real initial exon. Since the first type of error reduces our ability to distinguish different TSSs, and the second error per definition must be uncorrelated with splicing, the effect of both types of errors will be to *reduce* the correlations between TSS usage and splicing. Therefore, the clear correlations that we observe in spite of these potential errors should be considered to provide a lower bound on the correlations that do exist.

Another concern may be that the gene structures inferred from full-length cDNA are not representative, because the full-length cDNA sequencing projects generally included a prioritization step, that may have caused an apparent enrichment in rare splice variants. To address this issue, we have constructed a database of splice variants using solely human 5′-end ESTs, which we obtained based on the UniGene annotation. We analyzed these data using the same model as we used for full-length cDNAs. The 99% probability interval computed using this dataset was 0.74–0.78, compared with 0–0.009 obtained using the corresponding randomized dataset. This indicates that TSS-associated splice events are in fact even more frequent than initially estimated from full-length cDNA data. The likely reason why the estimate of the fraction of TSS-associated exons is larger when using 5′ EST data compared with full-length cDNA data is illustrated in Fig. 3. The exon indicated by the box belongs to the cAMP-dependent protein kinase catalytic beta subunit (PRKACB), is always skipped when the two upstream promoters are used (136 ESTs), and is generally included with the most downstream promoter (46 of 56 ESTs). These counts are very unlikely under a model in which the promoter usage and exon inclusion are uncoupled. Generally, many exons that in the cDNA data did not have sufficient coverage to allow us to detect their TSS association do have sufficient coverage in the 5′ EST data to allow this inference to be made, and consequently, the fraction of exons inferred to be TSS associated is larger.

3.2. *Evolutionary conservation of TSS association*

If the correlation that we inferred between transcriptional and splicing events is functionally relevant, one would expect that the TSS dependence tends to be conserved between orthologous cassette exons. This could, for instance, manifest itself in a correlation of the posterior probabilities of TSS dependence of orthologous cassette exons. To check this, we started with the human and mouse cassette exons that are part of transcription units with multiple start sites,

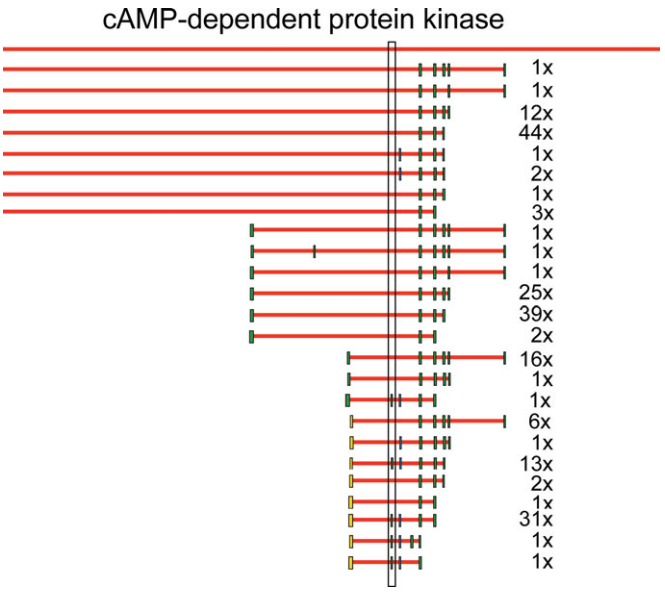


Figure 3. Example of an internal cassette exon whose inclusion is strongly correlated with specific TSSs as inferred from 5′ EST data. The conventions used in this representation are the same as for Fig. 1. Exons shown in green have invariant splice boundaries, those in yellow have variable splice donor sites, and those in cyan intron inclusion. The inclusion of the exon indicated by the black box occurs only when the most downstream TSSs are used.

and we used the UCSC human-to-mouse and mouse-to-human whole-genome alignments to identify orthologous exons (see Section 2.4). This procedure yielded 668 pairs of orthologs. We then computed the correlation coefficient between the posterior probabilities of orthologous exons. As shown in Fig. 4 we obtained a weak, but significant ($P = 0.002$) correlation between the probabilities of TSS association of orthologous exons, providing some evidence that TSS dependence of exon inclusion is evolutionarily conserved.

One of the best examples of an evolutionarily conserved relationship is shown in Fig. 5. The exon with a high probability of TSS association ($P = 1$ in human and $P = 0.73$ in mouse) is indicated by an arrow. It is included in the skeletal form of tropomyosin, which uses the most upstream TSS, and is excluded in other forms of tropomyosin, which also tend to use downstream TSSs.

The level of evolutionary conservation of cassette exons has previously been related to the rate of inclusion of the exons in mature mRNAs: the so-called 'major form' exons, which are predominantly included, are as conserved as constitutive exons, whereas 'minor form' exons, which are predominantly skipped, appear to be of a more recent evolutionary origin,[23] rarely having orthologs between human and rodents. To understand where TSS-associated exons fit in this evolutionary scenario, we analyzed the degree of human–mouse conservation of the following categories of exons: (1) TSS-associated exons—those with the top 10% values of posterior probability of TSS association, (2) TSS-independent exons—those with the bottom 10% values of posterior probability of TSS association, and (3) constitutive exons—exons included with no variation in more than ten transcripts. For each of these exons, we extracted the mouse—human and human–mouse alignments from the whole genome alignments provided by the UCSC (see Section 2.4). We then computed the fraction of exons that have orthologs in the other species and the fraction of nucleotides in each exon that are conserved. As shown in Table 1, we found that the large majority of TSS-associated exons are conserved between mouse and human ($438/496 = 88.3\%$ of mouse and $1003/1166 = 86.02\%$ of human TSS-associated exons). Particularly, TSS-associated exons are much more strongly conserved than 'minor form' exons, only 27–31% of which having been reported to be conserved between human and rodents,[23] However, TSS-associated exons are significantly less conserved than TSS-independent exons ($469/496 = 94.6\%$ in mouse and $1077/1166 = 92.4\%$ in human, P-value of the χ^2 test $= 6.7 \times 10^{-4}$ for mouse and 1.12×10^{-6} for human). These results are not sensitive to the precise threshold beyond which we consider an exon 'conserved'. Among those TSS-associated exons that do have orthologs, the proportion of conserved nucleotides is lower compared with TSS-independent exons (P-value: 2.9×10^{-12} for human and 2.2×10^{-3} for mouse), as well as compared with constitutive exons (P-value of the Wilcoxon test $< 2.2 \times 10^{-16}$ for human and 4.7×10^{-2} for mouse). Consistent with previous results relating the degree of evolutionary conservation to the inclusion rate of the exons,[23] we found that the overall inclusion rate of TSS-associated

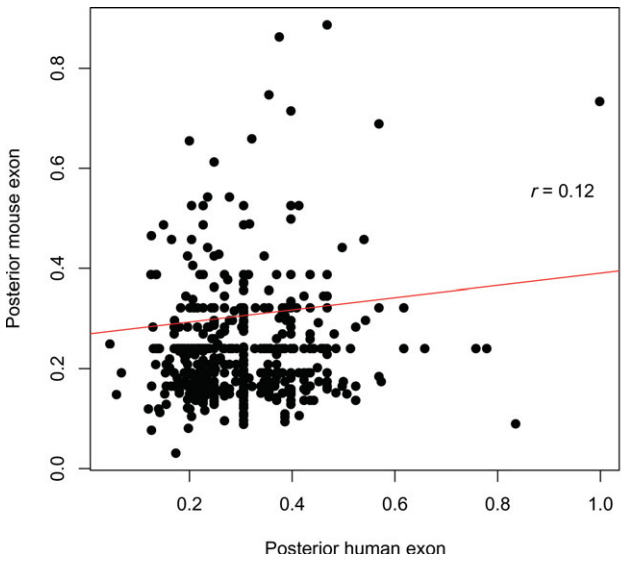


Figure 4. Correlation between the posterior probabilities of TSS association for orthologous human–mouse cassette exons. Each dot represents one exon, with the *x*-coordinate being the posterior probability of TSS association of the human exon and the *y*-coordinate being the the posterior probability of TSS association of the orthologous mouse exon. The correlation coefficient is $r = 0.12$, *P*-value = 0.002.

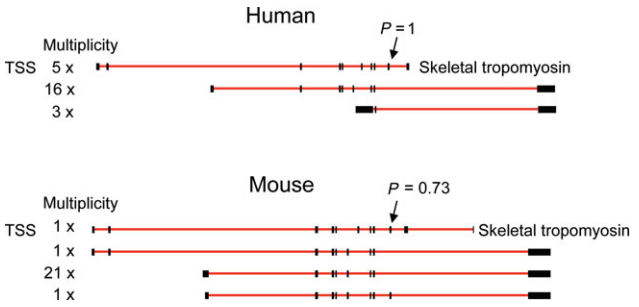


Figure 5. Conserved TSS association for a tropomyosin exon. The intron/exon structure of tropomyosin transcripts initiated from various TSSs is shown, with exons represented as black boxes and introns as red lines connecting the exons. The number of times each transcript form was observed in our dataset is indicated by the 'multiplicity' column on the right-hand side of the transcripts. The orthologous cassette exons are indicated by arrows, and their posterior probabilities of TSS association are indicated.

Table 1. Comparison of TSS associated, TSS independent, and constitutive exons

Data set	Number exons	With orthologs	Median proportion conserved nucleotides	Inclusion rate	Proportion symmetrical
Human exons					
TSS associated	1166	1003	0.84	0.75	0.41
TSS independent	1166	1077	0.87	0.86	0.45
Constitutive	6377	6343	0.88	1	0.38
Mouse exons					
TSS associated	496	438	0.86	0.75	0.42
TSS independent	496	469	0.88	0.88	0.48
Constitutive	5136	5117	0.87	1	0.4

exons is lower than the inclusion rate of TSS-independent exons (Table 1). Thus, the results suggest that some of the TSS-associated exons are of relatively recent evolutionary origin, and it will be interesting to establish whether the TSSs that promote their inclusion have also undergone recent evolutionary changes. On the other hand, the relatively weak correlation between the probabilities of TSS association of orthologous exons is likely to be in part due to the fact that the human and mouse sequencing projects did not cover sufficiently similar sets of tissues. This would be reflected in different relative usage of the alternative TSSs and exons between human and mouse and these, in turn, would be reflected in disparate probabilities of TSS association of orthologous exons in the two species.

Finally, we did not find a consistent trend in the proportion of 'symmetrical' exons, i.e. the proportion of exons whose length is a multiple of three, among our different categories of exons. We did not specifically select the exons for being part of coding regions, but rather we only considered internal exons in our analysis. As shown in Table 1, the proportion of symmetrical exons is significantly higher than expected by chance, i.e. 1/3, for all exon types (TSS-associated, TSS-independent, constitutive). In human, TSS-associated exons have a significantly higher tendency for symmetry compared with constitutive internal exons (*P*-value of the χ^2 test $= 0.02$), and lower, but not significantly, compared with TSS-independent exons (*P*-value of the χ^2 test $= 0.07$). The tendencies are similar in mouse, but the differences are not statistically significant (*P*-value of the χ^2 test $= 0.36$ in the comparison with constitutive exons and 0.097 in the comparison with TSS-independent exons).

3.3 *Insights into the mechanism of coupling between transcriptional and splicing events*

We can envision two mechanisms that could give rise to the correlation that we observe between transcriptional and splicing events. One is that tissue-specific transcriptional and splicing events are induced by independent, but tissue-specific transcription and splicing factors. The second mechanism involves a direct influence of tissue-specific transcription factors on internal splicing. We reasoned that if an exon is always included when one TSS is used and always skipped when another TSS is used, yet the two TSSs are both used in the same tissue, then the TSS association may be due to direct coupling of transcription with exon selection. To identify such cases, we used the library annotation of the transcripts in our datasets. We found that for 21% of the human and 14% of the mouse TSS-associated exons the inclusion- and skipping-promoting TSS have been both used in the same tissue.

One model that has been proposed for the coupling between transcriptional and splicing events is known as the 'kinetic model',[10] which postulates that cassette exons with weaker splice signals compared with constitutive exons tend to be skipped when the transcript is produced by a fast-elongating polymerase. In contrast, when the polymerase elongation rate is low, the spliceosome has sufficient time to assemble on these cassette exons, which are then spliced into the mature mRNA. We used a web server implementing the Shapiro and Senapathy model[24] (http://ast.bioinfo.tau.ac.il/SpliceSiteFrame.htm) to evaluate the strength of the splice sites of the cassette exons as well as of the exons flanking them. We found that the strength of the splice sites is comparable between the cassette exons and the exons flanking them (data not shown), suggesting that one of the pre-requisites of the kinetic model (weaker splice sites) is not met by our sets of TSS-associated exons.

Given that the elongation rate is not homogeneous along a gene, and that promoter-proximal pausing of RNA polymerase II is common,[25] we reasoned that the elongation rate of the polymerase may be generally lower at the start of the transcripts, allowing better recognition of cassette exons that are close to the TSS. Therefore, we asked whether the TSS-associated exons tend to be included when the TSS closest to

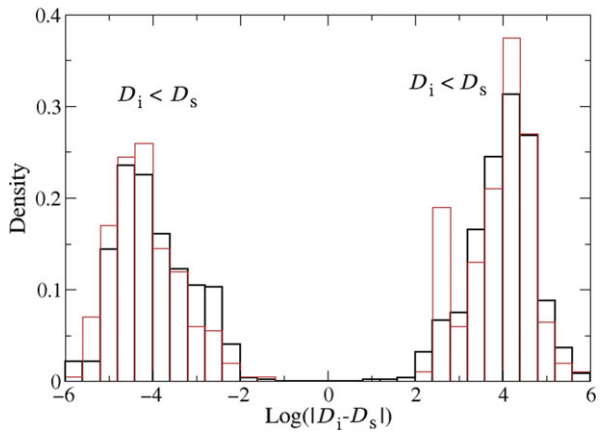


Figure 6. Histogram of the signed difference between the logarithm (base 10) of the average distance to TSS when the exon is included and the average distance to TSS when the exon is skipped. Human data are shown in black and mouse data are shown in red.

them is used, as this may allow sufficient time for spliceosome assembly. We found this not to be the case: for all TSS-associated exons, we computed the average distance D_i between the exon and the TSSs of transcripts that included the exon, and the average distance D_s between the exon and the TSSs of transcripts that excluded the exon. We then constructed a histogram of the difference $D_i - D_s$ across cassette exons (Fig. 6). To improve visibility of the histogram, we split it into a part where $D_i - D_s > 0$ and a part where $D_i - D_s < 0$, and plotted the absolute distance $|D_i - D_s|$ on a logarithmic scale. We found that approximately the same proportion of TSS-associated exons are preferentially included with the most upstream TSSs as are included with the most downstream TSSs (Fig. 6). This suggests that proximity to the TSS, which may be indicative of lower polymerase elongation rate, cannot explain the inclusion pattern of TSS-associated cassette exons.

To conclude, we estimated that exons whose inclusion in the mature mRNA is correlated with specific TSSs are rather common, i.e. they represent at least 15% of all internal cassette exons. The correlation may be due to direct coupling between transcription and splicing or indirect coupling, due, for instance, to a tissue-specific signaling pathway that activates both the transcription factors responsible for determining the TSS as well as the splicing factors responsible for the inclusion or exclusion of the cassette exon. Nonetheless, for at least 14–21% of the TSS-associated exons, TSSs that are associated with exclusion and TSSs that are associated with inclusion occur *both* in the same tissue, suggesting a more direct connection between the TSS and exon inclusion for at least these exons. The details of the molecular mechanism underlying this dependency remain to be uncovered, and may involve a gene-specific component, as suggested by our observation that some exons are predominantly included when the proximal TSSs are used, while other exons are predominantly skipped. One way of implementing a gene-(and tissue-) specific component could be through dynamic changes in the local chromatin structure that induce in turn local variations in the polymerase elongation rate.[26] This is a topic for future study and the cassette exons for which we estimated a high probability of TSS association provide good starting points for experimental investigations.

Supplementary Data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

This work has been supported in part by the European Network for Alternative Splicing (EURASNET). TM Chern is also supported by a fellowship from the South African National Research Foundation. We gratefully acknowledge the VitalIT team of the Swiss Institute of Bioinformatics, in particular Ioannis Xenarios, for computational support.

References

1. Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D., Hayashizaki, Y., et al., 2003, Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome, *Genome Res.*, **13**, 1290–1300.
2. Sugnet, C., Kent, W., Ares, M. and Haussler, D. 2004, Transcriptome and genome conservation of alternative splicing events in humans and mice, *Pac. Symp. Biocomput.*, 66–77.
3. Berget, S. 1995, Exon recognition in vertebrate splicing, *J. Biol. Chem.*, **270**, 2411–2414.
4. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M. 2000, An alternative-exon database and its statistical analysis, *DNA Cell Biol.*, **19**, 739–756.
5. Wang, J., Smith, P., Krainer, A. and Zhang, M. 2005, Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes, *Nucleic Acids Res.*, **33**, 5053–5062.
6. Sorek, R. and Ast, G. 2003, Intronic sequences flanking alternatively spliced exons are conserved between human and mouse, *Genome Res.*, **13**, 1631–1637.
7. Baek, D. and Green, P. 2005, Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing, *Proc. Natl Acad. Sci. USA*, **102**, 12813–12818.
8. Plass, M. and Eyras, E. 2006, Differentiated evolutionary rates in alternative exons and the implications for splicing regulation, *BMC Evol. Biol.*, **6**, 50.

9. Kornblihtt, A. 2005, Promoter usage and alternative splicing, *Curr. Opin. Cell. Biol.*, **17**, 262–268.
10. Kornblihtt, A. 2006, Chromatin, transcript elongation and alternative splicing, *Nat. Struct. Mol. Biol.*, **13**, 5–7.
11. Nogues, G., Kadener, S., Cramer, P., Bentley, D. and Kornblihtt, A. 2002, Transcriptional activators differ in their abilities to control alternative splicing, *J. Biol. Chem.*, **277**, 43110–43114.
12. Lim, L. and Sharp, P. 1998, Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats, *Mol. Cell. Biol.*, **18**, 3900–3906.
13. Monsalve, M., Wu, Z., Adelmant, G., Puigserver, P., Fan, M. and Spiegelman, B. 2000, Direct coupling of transcription and mRNA processing through the thermogenic coactivator PGC − 1, *Mol. Cell*, **6**, 307–316.
14. Pagani, F., Stuani, C., Zuccato, E., Kornblihtt, A. and Baralle, F. 2003, Promoter architecture modulates CFTR exon 9 skipping, *J. Biol. Chem.*, **278**, 1511–1517.
15. Auboeuf, D., Honig, A., Berget, S. and O'Malley, B. 2002, Coordinate regulation of transcription and splicing by steroid receptor coregulators, *Science*, **298**, 416–419.
16. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N. et al. 2005, The transcriptional landscape of the mammalian genome, *Science*, **309**, 1559–1563.
17. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. et al. 2004, Integrative annotation of 21,037 human genes validated by full-length cDNA clones, *PLoS Biol.*, **2**, 856–875.
18. Chern, T., van Nimwegen, E., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. et al. 2006, A simple physical model predicts small exon length variations, *PLoS Genet.*, **2**, e45.
19. van Nimwegen, E., Paul, N., Sheridan, R. and Zavolan, M. 2006, SPA: a probabilistic algorithm for spliced alignment, *PLoS Genet.*, **2**, e24.
20. Zavolan, M., van Nimwegen, E. and Gaasterland, T. 2002, Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome, *Genome Res.*, **12**, 1377–1385.
21. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J. et al. 2006, Genome-wide analysis of mammalian promoter architecture and evolution, *Nat. Genet.*, **38**, 626–635.
22. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M. et al. 2006, CAGE: cap analysis of gene expression, *Nat. Methods.*, **3**, 211–222.
23. Modrek, B. and Lee, C. 2003, Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss, *Nat. Genet.*, **34**, 177–180.
24. Shapiro, M. and Senapathy, P. 1987, RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression, *Nucletic. Acids Res.*, **15**, 7155–7174.
25. Krumm, A., Hickey, L. and Groudine, M. 1995, Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation, *Genes Dev.*, **9**, 559–572.
26. Batsche, E., Yavin, M. and Muchardt, C. 2006, The human SWI/SNF subunit Brm is a regulator of alternative splicing, *Nat. Struct. Mol. Biol.*, **13**, 22–29.