4740–4754 *Nucleic Acids Research, 2010, Vol. 38, No. 14* *Published online 12 April 2010*
doi:10.1093/nar/gkq197

Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome

I. G. Mollet[1],*, Claudia Ben-Dov[2], Daniel Felício-Silva[1], A. R. Grosso[1], Pedro Eleutério[1], Ruben Alves[1], Ray Staller[3], Tito Santos Silva[4] and Maria Carmo-Fonseca[1]

[1]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisbon, Portugal, [2]CRG-Centre de Regulació Genómica, Barcelona, Spain, [3]Independent Senior Consultant on Statistics & Mathematics, Amsterdam, the Netherlands and [4]Faculdade de Engenharia, Universidade Católica Portuguesa, Lisbon, Portugal

Received November 17, 2009; Revised February 22, 2010; Accepted March 9, 2010

ABSTRACT

Mining massive amounts of transcript data for alternative splicing information is paramount to help understand how the maturation of RNA regulates gene expression. We developed an algorithm to cluster transcript data to annotated genes to detect unannotated splice variants. A higher number of alternatively spliced genes and isoforms were found compared to other alternative splicing databases. Comparison of human and mouse data revealed a marked increase, in human, of splice variants incorporating novel exons and retained introns. Previously unannotated exons were validated by tiling array expression data and shown to correspond preferentially to novel first exons. Retained introns were validated by tiling array and deep sequencing data. The majority of retained introns were shorter than 500 nt and had weak polypyrimidine tracts. A subset of retained introns matching small RNAs and displaying a high GC content suggests a possible coordination between splicing regulation and production of noncoding RNAs. Conservation of unannotated exons and retained introns was higher in horse, dog and cow than in rodents, and 64% of exon sequences were only found in primates. This analysis highlights previously bypassed alternative splice variants, which may be crucial to deciphering more complex pathways of gene regulation in human.

INTRODUCTION

The overwhelming volume of currently available biological sequence data requires the creation of automated procedures for mining databases containing vast amounts of data. For the human genome alone, over 8 million transcribed sequences have been deposited in GenBank (1). An important piece of information is the identification of all alternatively spliced transcripts generated by each gene. Alternative splicing has evolved as a predominant mechanism in expanding the proteomic complexity encoded in a limited number of genes (2,3), a source of variation that is further exponentiated by the combinatorial possibilities inherent in the modular nature of protein complexes. Alternative splicing occurs in response to a wide range of stimuli and has been shown to be tissue-specific, developmental stage-specific and disease-specific (4). Early studies based on the large-scale analysis of cDNAs and expressed sequence tags (ESTs) indicate that more than 60% of human genes undergo alternative splicing (5), while microarray analyses have led to an increase in this estimate to between 73% (6) and 80% (7). More recently, high-throughput sequencing technologies are revealing that 92–94% of human genes undergo alternative splicing (8–10).

However, in recent years, it has become apparent that many of the alternatively spliced isoforms in a gene do not produce any protein. Several lines of evidence indicate that post-transcriptional regulation through alternative splicing involving non-protein-coding transcripts undoubtedly constitutes an important mechanism for the control of gene expression (11,12). Mechanisms of this nature include regulated unproductive splicing and

*To whom correspondence should be addressed. Tel: +46(0)40-391159; Fax: +46(0)40-391222; Email: ines.mollet@med.lu.se; inesmollet@yahoo.com
Present address:
I. G. Mollet, Lund University Diabetes Center, CRC 91-11, UMAS number 72, 205 02 Malmö, Sweden.

The authors wish it to be known that, in their opinion, the second and third authors should be regarded as joint Second Authors.

translation (RUST) (11) and nonsense-mediated decay (NMD) (13,14).

In the process of exploring alternatively spliced transcripts in several model organisms, we came to realize that there appeared to be greater amounts of apparently non-protein-coding alternative transcripts containing intron retentions or new exons in human genes than in orthologous genes from other organisms, which had not been annotated in existing alternative splicing databases (15). This prompted us to develop an automated method to mine EST and mRNA transcript alignments to genomic sequence with the goal of identifying all possible alternative splicing combinations, whether coding or non-coding, in a given gene locus. This work throws the spotlight on the potential functional importance of intron retention concomitant with the expression of short RNAs, and on unannotated exons in human with low sequence conservation.

MATERIALS AND METHODS

Data sources

Tables of BLAT alignments (16) of RefSeq (17), mRNAs and spliced ESTs to a genome assembly were obtained from the UCSC Genome Browser http://genome.ucsc.edu/ (18,19). BLAT alignments of spliced ESTs taken from the UCSC Genome Browser have a minimum of 96% base identity with the genomic sequence and only carry introns smaller than 750 000 bases; in addition, such spliced ESTs have at least one intron of minimum length 32 bases with a GT...AG consensus splice site. The chromosome sequences were downloaded from the UCSC Genome Browser for the human genome assembly, *Homo sapiens* (20) and mouse genome assembly, *Mus musculus* (21). Human ESTs used in this analysis originate from libraries generated from a variety of tissue sources (45% from normal tissue, 26% from tumor tissue or cell lines, 2% from disease tissue and 27% with no tissue annotation).

Data processing

The genomic region of a gene was determined as that to which all RefSeq BLAT alignments associated with a given Entrez GeneID (22) mapped, extended to any mRNA or spliced EST with at least one spliced junction in common with the RefSeq alignments. For unspliced RefSeq genes, all mRNA and spliced EST alignments falling within the boundaries of the gene region were considered. BLAT alignments of transcripts consist of a series of aligned blocks separated by gaps of genomic sequence. Several steps were taken to filter out data not considered sufficiently robust to permit extraction of information on alternative splicing. To establish the minimum intron size, we looked for the smallest RefSeq intron carrying the splice site consensus sequences GT...AG, GC...AG or AT...AC and found it to be 30 nt. Based on this, a minimum intron size of 30 nt was postulated and, as a result, any two adjacent blocks separated by gaps smaller than 30 nt were not considered reliable introns and were joined into a single block. These small gaps can be due to short repeats of variable length in the genomic sequence and the transcript or to short stretches of the transcript which the BLAT program did not succeed in aligning often due to single nucleotide polymorphisms; their occurrence was mostly seen in 5′ and 3′ untranslated regions. Blocks in the BLAT alignment of a transcript mRNA or EST are potential exons. The smallest exon, which we have detected in RefSeq annotated human genes, is 6 nt (in the *RELN* gene), but we found that it is not mapped correctly to the genome assembly in UCSC BLAT alignments. In total, eight blocks with 7 nt were found in RefSeq genes but of these only one mapped with consensus splice sites in UCSC BLAT alignments. In total, nine blocks with 8 nt were found in RefSeq genes and, of these, seven mapped with consensus splice sites. There are more than 50 blocks of 9 nt in RefSeq genes and, with few exceptions, these map with consensus splice sites in UCSC BLAT alignments. A 9-nt block was therefore chosen as a reliable lower threshold to impose on the size of blocks. The first two and last two nucleotides of each remaining gap in the alignment were then analyzed for splice site consensus sequences GT...AG, GC...AG, or AT...AC in the case of an mRNA source; and GT...AG or GC...AG in the case of an EST source. Where a splice site consensus did not exist, or a block was smaller than 9 nt, an alignment was cut into fragments at the non-consensus junction or on either side of the small block, respectively. Given that RefSeq sequences are constantly reviewed and are here used as reference sequences for a given gene, the program overrides this filtering procedure for RefSeq alignments and all of these are accepted regardless of the consensus sequence at the splice junction and the size of the aligned blocks. For the above set of filtered data, the start and end of each alignment were corrected to the nearest known spliced junction or, for the terminal regions, to the longest corresponding block, using a set of rules described in the Supplementary Data. The resulting blocks were considered to correspond to exons and the intervening gaps to introns and each sequence segment was numbered and assigned a type. Thus, only a single EST or mRNA is required to support a new exon, provided that the aligned sequence has 96% identity to the genome assembly, that it contains at least one intron with consensus splice sites and has at least one splice site in common with a RefSeq annotated exon. No special requirement was imposed on first or last exons other than that these must have at least 9 nt (the minimum block size required) and splice site consensus sequences. If a gap >30 nt between the first exon and the next exon does not have consensus splice sites the first exon is excluded, idem for terminal exons. The data on alternative splicing were generated in such a way as to allow visualization of all possible exons, within the context of a whole protein-coding gene. Alternative splicing patterns were determined as the longest pattern containing a distinct set of alternatively spliced exons. In this manner, information contained in a set of several thousand ESTs could be condensed by several orders of magnitude.

The collection of unannotated exons used in this study consists of the exons which are not annotated in RefSeq transcripts and excludes 3′ and 5′ extensions of exons. The complete set of data resulting from this treatment was integrated into a public interactive database called ExonMine freely available at: http://www.imm.fm.ul.pt/exonmine/.

Generation of data from random sets of ESTs

To assess the effect of the number of ESTs on estimates of levels of alternative splicing, the procedure discussed above was used with random sets of ESTs at intervals of 0.5 million ESTs, ranging from 0.5 to 2 million ESTs for mouse and 0.5 to 4 million ESTs for human. For each point, two random sets were generated and these displayed a variation of <0.1% in the percent of genes with more than one splicing pattern, <0.1 in the average number of exons per gene and <0.1 in the average number of splicing patterns per gene.

Analysis of poly-A signals in terminal exons and TSS in first exons

Since the data generated relies largely on fragments of transcripts, we assessed the likelihood of our terminal exons (exons with no 5′splice site) being true terminal exons by the presence of the poly-A signal AATAAA and single nucleotide variants of this signal (23). Similarly, the likelihood of first exons being true first exons (exons with no 3′ splice site) was estimated by the presence of transcription start sites (TSSs): to do this, the database of TSSs DBTSS, Version 7.0 15 Sep.2009, (24) was matched against first exons, including the region 200 nt upstream.

Tiling array data

Tiling array data covering the non-repetitive portion of the human genome (25) were obtained from NCBI Geo Dataset GSE7576. These data related to the 2004 human genome assembly sequence (NCBI Build 35, hg17); therefore, the genomic coordinates of the tiling array probe signals were lifted to the March 2006 human genome assembly sequence (NCBI Build 36.1, hg18) using BLAT. Data from the ExonMine August 2008 update on human genome assembly hg18 were matched to tiling array probe signal and to short and long transcribed fragments. Short transcribed fragments in the tiling array data range in size from 22 to 200 nt and were detected, in that study, in HeLa and HepG2 cell lines, both from the + and –strand, across the whole nonrepetitive fraction of the human genome.

Analysis of repetitive elements

The presence of repetitive elements in unannotated exons was determined by pairwise alignment against RepBase version 13.10 (26) using blastn (27). An exon was considered to contain a repetitive element if the alignment had a minimum of 80% identity over a minimum query aligned length of 50 nt.

P-value calculation for short RNAs associated with retained introns

To show that short RNAs are more strongly associated with retained introns than with all introns within the size range of retained introns (population 1) or within the population of all introns (population 2), we calculated the statistical significance of our result. As we are dealing with a yes/no issue we used a binomial distribution, which is fully defined by one parameter, *P*, in this instance, the probability of finding introns carrying short RNAs: P = (number of introns carrying short RNAs in a population)/(total number of introns in that population). To estimate this parameter, we assume the ‘largest group’ to be representative for the population. For population 1, $P = 31325/126292 = 0.2480$; and for population 2, $P = 64844/205975 = 0.3148$. To show that the set of retained introns confirmed by 50% coverage of tiling array transcribed fragments in the cytoplasm (50CytoTF set) is not a random sample taken from this large group we calculate the *P* value. The mean of a binomial distribution is $M = N \times P$ and the standard deviation is $s = \sqrt{}\,(M\,(1–P))$ where $N = 7381$ is the number of introns in the 50CytoTF set. For the 50CytoTF set this results in: $M = 7381 \times 0.2480 = 1830.76$ and $s = 37$ for population 1; and $M = 7381 \times 0.3148 = 2323.65$ and $s = 40$ for population 2. To calculate the *P*-value, we use the fact that a binomial distribution for $N = 7381$ can practically completely be represented by a normal distribution. This allows us to convert the 50CytoTF data to a standard normal distribution which has a mean of 0 and a standard deviation of 1; this standard normal distribution in its turn allows the use of standard tables, to derive the *P*-value. The formula for this conversion is $z = (x–M)/s$ where x is the observed number of introns carrying short RNAs in the 50CytoTF set. For population 1 $z = (2663–1830.76)/37 = 22.43$ and for population 2 $z = (2663–2323.65)/40 = 8.50$. Using the above-mentioned tables, we derived a $P\text{-value} < 0.0001$. According to conventional criteria this means that the number of short RNAs observed in the 50CytoTF set is extremely statistically significant.

GC content in introns

GC content was determined in three sets of introns of 2500 small introns, small introns being smaller than 1029 nt, as previously defined (28). Set Rs: retained introns that have at least 50% of their surface covered by tiling array transcribed fragments in the cytoplasm and also matching short transcribed fragments (25); set Rns: small retained introns that have at least 50% of their surface covered by tiling array transcribed fragments (25) in the cytoplasm but do not match short transcribed fragments; and set NR: small non-retained introns. We imposed the same number of introns and size distribution for sets Rns and NR as was observed for set Rs: this was done by random selection of the same number of introns in each quartile and outliers as that observed in set Rs. To test whether the difference in the mean GC content in two sets of introns is large enough that it is most likely that the means come from two different populations, the two-tailed *P*-value for a *t*-statistic was calculated for unequal variance.

Search for known small RNAs

To support the possibility of retained introns being involved in production of small RNAs, we searched UCSC table browser (19) data based on the miRBase (Release 13.0, March 2009) Sequence Database (29) for precursor of microRNAs; snoRNABase (version 3) for C/D and H/ACA box small nucleolar RNAs and small Cajal body-specific RNAs (scaRNAs) data from the Laboratoire de Biologie Moléculaire Eucaryote (30).

Sequence conservation analysis

Sequence conservation analysis of human unannotated exons and retained introns was performed using discontiguous megablast (27,31) across eight vertebrate species: chimp (*Pan troglodytes*, taxid:9598) and rhesus (*Macaca mulata*, taxid:9544) genomes were used to represent nonhuman primates; mouse (*Mus musculus*, taxid:10090) and rat (*Rattus norvegicus*, taxid:10116) genomes were used to represent non-primate supraprimates; dog (*Canis familiaris*, taxid:9615), horse (*Equus caballus*, taxid:9796) and cow (*Bos taurus*, taxid:9913) genomes were used to represent laurasiatheria, another large group of placental mammals; chicken (*Gallus gallus*, taxid:9031) was used to represent a nonmammalian vertebrate genome. Discontiguous megablast is a version of Mega BLAST designed specifically for comparison of diverged sequences, especially sequences from different organisms. The parameters, which we applied for this conservation analysis, are those recommended by the authors for cross-species analysis. The analysis was performed on chromosomes using the following parameters: a word size of 11; reward/penalty scoring parameters for matching and mismatching bases of 1/–1; associated gap existence penalty of 5; gap extension penalty of 2; a filter was applied to mask low complexity regions for the lookup table only; human specific repeats were filtered; and, finally, a discontiguous maximal template type with length 16 was used. A sequence was considered conserved if alignments displayed more than 70% identity over at least 80% query coverage with bitscores >50 and E (expect) values <0.05.

RESULTS

ExonMine detects a higher number of exons and alternatively spliced genes

Alternative splicing data generated using the method described above for the complete human genome resulted in the detection of 256 605 core exons (i.e. excluding exon extensions) covering 18 727 genes (haplotypes excluded) (Table 1). The position of these exons in transcripts is 16% first (40 418 exons), 73% internal (186 837 exons) and 11% terminal (28 292 exons). The identified exons are involved in a total of 199 045 distinct splicing patterns, giving an average of 10.6 splicing patterns per gene. All exons and splicing patterns can be tracked to the original set of mRNA and EST input data through fixed numbering of exons and introns in relation to a reference RefSeq sequence. The complete data set can be visualized online on the ExonMine database at: http://www.imm.fm.ul.pt/exonmine/.

Table 1 compares ExonMine data for human and mouse with data from two recently published alternative splicing databases (32,33). For both human and mouse, ExonMine detects a higher percentage of alternatively spliced genes, 88% compared to 78% in fastDB and 53% in ASAPII for human; and 79% compared to 54% in fastDB and 53% in ASAPII for mouse. Most striking is the fact that ExonMine data detects a 27% increase in the average number of exons detected per gene in human (14 exons per gene) compared to the highest detected by fastDB (11 exons per gene). This is most likely due to differences imposed on the selection of transcripts. For example, in fastDB transcripts are initially selected by blasting each exon, defined by EnsEMBL (34), against mRNA and EST alignments to the genome, whereas ExonMine relied on BLAT alignments that require a transcript to have only one splice site in common with RefSeq transcripts. Global percent identity of alignment of transcripts in fastDB is 98% versus 96% in ExonMine. In fastDB, at least 95% of the transcript has to be aligned, it must cover 10% of the genomic region, and the ratio of exon and intron lengths in a given transcript cannot exceed three times the average ratio of all defined exon and intron lengths; whereas ExonMine cuts alignments into fragments at non-consensus junctions and on either side of exons smaller than nine nucleotides and recovers any spliced fragments. As a result of the differences in data selection criteria, ExonMine succeeds in capturing more

Table 1. Comparison of our results with other databases

Database	Human				Mouse			
	No. genes	Genes with AS (%)	No. exons[a]	Average no. exons per gene	No. genes	Genes with AS (%)	No. exons[a]	Average no. exons per gene
ExonMine	18 727	88	256 605	14	19 110	79	215 343	II
fastDB	18 008	78	201 245	11	13 913	54	157 920	II
ASAP II	22 220	53	129 981	6	16 404	53	105 260	6

Alternative splicing data generated in this analysis (here referred to as ExonMine data) compared to FastDB (32), ASAP II (33).
[a] For our data this count includes only the core part of an exon, i.e. excluding exon extensions.

than three times the number of ESTs and mRNAs than fastDB. ExonMine also clustered the great majority of human spliced ESTs (97%), which suggests a near complete coverage of spliced genes by human RefSeq transcripts.

The complexity of alternative splicing is higher in human than in mouse

The evaluation of levels of alternative splicing based on ESTs in limited sets of EST data has been shown to be dependent on the number of ESTs available (35). However, there is an element of redundancy in EST coverage when large quantities of spliced ESTs are available, as in the case of human (3.9 million) and mouse (1.9 million) where the number of ESTs is an order of magnitude above that of the number of exons. In our analysis, for human and mouse, 78% and 79% of the total number of exons respectively are covered by two or more ESTs. This level of coverage allows us to begin to compare the levels of alternative splicing detected in these two organisms.

To assess how the number of ESTs can affect the determined level of alternative splicing we generated our results from random sets of ESTs at intervals of 0.5 million: ranging from 0.5 to 2 million in mouse and 0.5 to 4 million in human (Figure 1). For the percentage of genes with alternative splicing (Figure 1A), although we see an increase from 81.7% (estimate with 0.5 million ESTs) to 88.2% (estimate with 4 million ESTs) for human, at the latter point we appear to be reaching a threshold. The same can be said for the number of exons per gene (Figure 1B), which has reached 13.6 in human and 11.2 in mouse with the available data. However, when we estimate the number of splicing patterns per gene no threshold appears to have been reached yet (Figure 1C). This may either be due to spurious alternative splicing obtained from *in vitro* manipulations or, indeed, reflect a real increase in complexity of alternative splicing due to the increased combinatorial possibilities resulting from a greater number of exons per gene in human. On the whole, the results shown in Figure 1 suggest that the level of alternative splicing in mouse is consistently below that of human.

In relation to RefSeq, we detect seven times more splicing patterns in human and five times more in mouse. Mining for alternatively spliced exons in human mRNA and EST data resulted in the identification of 48 942 unannotated exons >25 nt, spliced to exons in annotated RefSeq transcripts, 60% of which occur in normal tissues and only 16% in cancerous tissues. Out of a total of 18 727 genes considered in this analysis, 13 497 (72% of genes) carry unannotated exons, 97.7% of which appear in completely spliced transcripts. Sixty percent of the unannotated exons are supported by at least one mRNA or two ESTs.

Both in human and mouse, the majority of unannotated exons are first exons, 45% in human (Figure 2B) and 52% in mouse, this contrasts with the distribution of first, internal and terminal exons found in the whole set of annotated exons where 16% of exons are first exons (Figure 2A). In our data, first exons are those for which no 3′ splice site was detected and terminal exons are those for which no 5′ splice site was detected. Since our data rely heavily on ESTs, which are fragments of transcripts, we estimated the likelihood of first exons being true first exons by matching these data with the latest version of the database of transcription start sites (DBTSS) (24). We found that 63% (Figure 2C) of new first exons match TSS in normal adult and fetal tissue (75% when cell lines are included). In known first exons we found matches for TSS

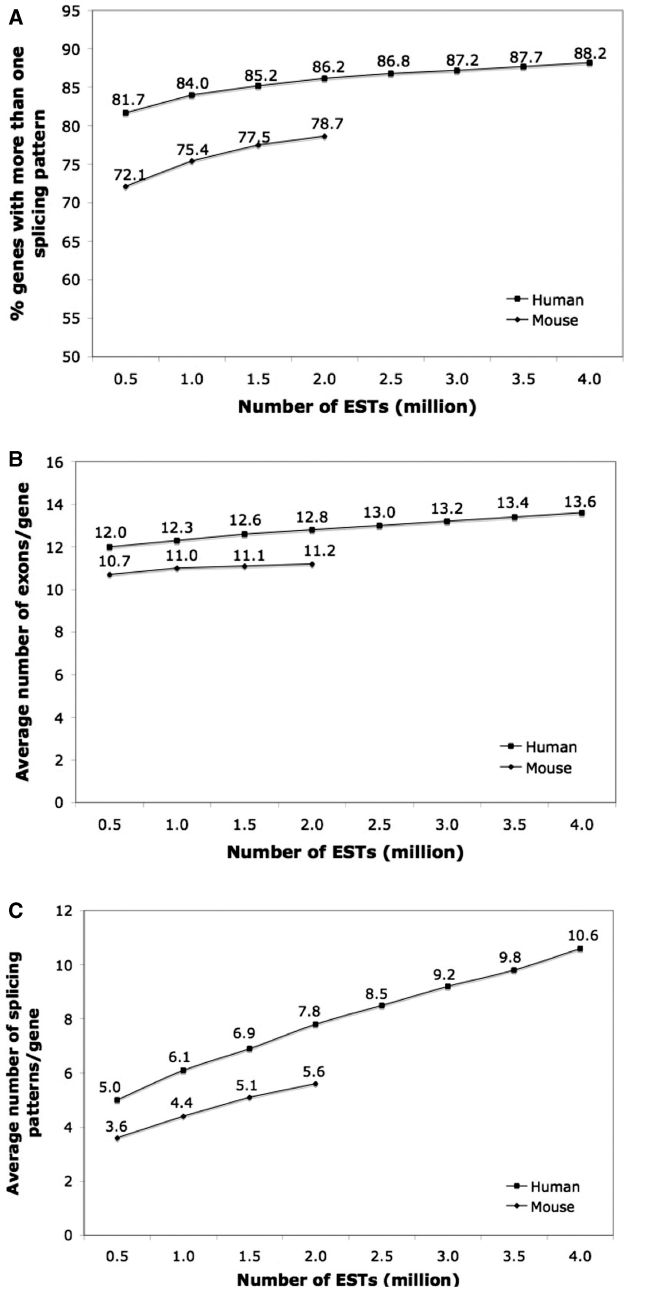
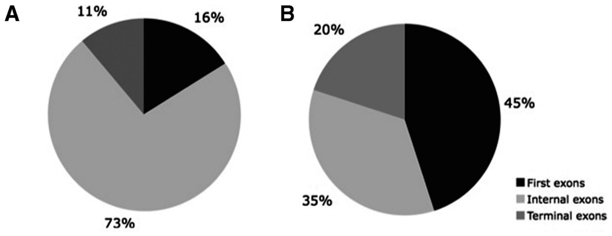


Figure 1. Effect of number of ESTs on estimates of levels of alternative splicing. Data generated from random sets of ESTs at intervals of 0.5 million ESTs, ranging from 0.5 to 2 million for mouse and 0.5 to 4 million for human. (**A**) Percent of genes with more than one splicing pattern. (**B**) Average number of exons per gene. (**C**) Average number of splicing patterns per gene.



C

Exon set	TSS in normal (adult and fetal) tissue	TSS in both normal (adult and fetal) tissue and cell lines
New fe	63%	75%
Known fe	86%	89%

D

Exon set	PolyA signal AATAAA	PolyA signal AATAAA and all variants
New te	30%	61%
Known te	78%	97%

Figure 2. Relative amounts of first, internal and terminal exons. (**A**) Relative amounts of first, internal and terminal exons in known (RefSeq annotated) exons. (**B**) Relative amounts of first, internal and terminal new (unannotated) exons. (**C**) Percent of first known and new exons containing transcription start sites (TSSs) within the exons or 200 nt upstream [DBTSS (24), Version: 7.0, 15 September 2009). (**D**) Percent of terminal exons containing the poly-A signal or 1-nt variants of the consensus AATAAA sequence (23). First exons are those for which no 3′ splice site was detected. Terminal exons are those for which no 5′ splice site was detected. Internal exons have both a 3′ splice site and 5′ splice site. This analysis includes only exons with minimum 25 nt and excludes chimeric transcript products.

in 86% of normal adult and fetal tissue (89% when cell lines were included). Although TSSs were also found in 21% of known internal exons, for normal adult and fetal tissue, this result suggests that only 25% of new first exons found are internal, derived from fragments of transcripts, rather than true first exons.

To assess the likelihood of new terminal exons being true terminal exons, we determined the number of terminal exons carrying the poly-A signal AATAAA and 1-nt variants of that signal (23) (Figure 2D). By true terminal exon we mean that it is the last exon to be spliced (i.e. no 5′ splice site was detected throughout the data) and it contains a polyadenylation signal. We find that only 30% of unannotated terminal exons contain the consensus poly-A signal (61% when one nucleotide variants of the poly-A signal are included) compared to 78% in known terminal exons (97% for variants of the poly-A signal). By contrast, in first exons and internal exons, the poly-A signal and its variants are only found in 4% of cases for the consensus AATAAA to a maximum of 12% for the AAGAAA variant. This result suggests that at least 39% of the terminal exons found might in fact be internal exons for which the complete transcript is still undetermined.

Overall, one-third of unannotated first exons and internal exons in human and mouse are in frame with downstream known exons and are thus expected to contribute putative coding sequence.

Novel exons are validated by tiling array expression data

Comparison of unannotated exons identified by our analysis with expression data from tiling arrays for eight cell lines (25) shows that 37% of the unannotated exons are not covered by sufficient tiling array probes to be detected as transcribed fragments, notwithstanding, 44% are represented in transcribed fragments in the tiling array data, with 35% expressed in the cytoplasm.

In light of recent reports suggesting that repetitive elements, such as the Alu element, play an important role in the evolution of alternative splicing in primates (36,37), we searched for the presence of vertebrate repeat elements, deposited in RepBase 13.10 (26), in unannotated exons using Blastn. This analysis revealed that 24% of unannotated exons align with repeats. By far the most abundant repetitive element found was the primate specific Alu element found in 12% of unannotated exons (~6000 exons). Since the tiling array data cover only the nonrepetitive fraction of the genome, the latter are expected to escape detection by array analysis.

To illustrate this analysis, an example of tiling array data for gene *ZRSR2* is presented in Figure 3, in which four unannotated exons were found. The unannotated exon 1A presents a signal comparable to that of known exons in all eight cell lines, whereas exons 2A, 3A and 3B are not covered by probes. The latter three exons are an example of exonization of tandem insertions of the primate specific Alu repeat element (38). On the other hand, tiling array data also indicate that exon 1A coincides with short transcribed RNA fragments ranging from 22 to 200 nt (25), suggesting a non-protein-coding function for this particular exon. However, we did not find a higher number of short transcribed fragments from the tiling arrays (putative short RNAs) associated with unannotated exons than with annotated exons (Table 2): whereas 64% of annotated first exons match short transcribed fragments from the tiling arrays, only 37% on unannotated exons do so. On the whole, it appears that first and terminal exons, but in particular first exons, are more often associated with short transcribed fragments than internal exons, suggesting a stronger regulatory role associated with short RNAs for first and terminal exons than for internal exons, the latter being naturally expected to perform mainly a protein coding function.

A large fraction of isoforms contain retained introns

Analysis of spliced transcripts carrying intron retentions revealed 16 288 retained introns in human, belonging to 7708 genes (41% of genes). Only 3% of these intron retentions are annotated in RefSeq transcripts. Among the RefSeq annotated retained introns, 37% do not alter the reading frame. In our set of 16 288 retained introns, 33%

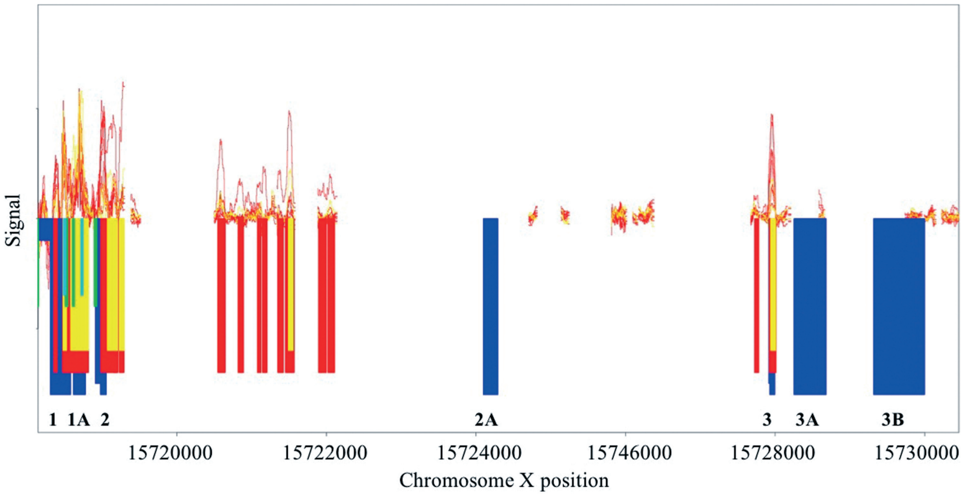


Figure 3. Tiling array data for fragment of gene *ZRSR2*. Coordinates for signal and transcribed fragments of tiling array data [Geo Accession GSE7576 (25)] of all eight cell lines used in that study for human genome assembly hg17 were lifted to assembly hg18 and matched to ExonMine data (August 2008 update). The graph represents data for the 5′-end of gene *ZRSR2*. Nuclear signal (yellow) and cytoplasmic signal (red) shown. Exon positions from our ExonMine data (blue) and transcribed fragments from tiling array data are represented superimposed on the negative axis: cytoplasmic (red), nuclear (yellow), short RNA top strand (green) and short RNA bottom strand (cyan). The figure shows that probe coverage on the tiling array is absent or too low for Alu containing unannotated exons 2A, 3A and 3B. For unannotated exon 1A, however, there is a clear nuclear and cytoplasmic signal as well as correspondence to short RNA transcribed fragments in that region. The figure also shows expression which is not detected in ExonMine, including: on the 5′-end of the intron downstream of exon 2; several transcribed fragments between exons 2 and 2A likely to correspond to a gene on the opposite strand for which there is only EST evidence (AA284226); and a transcribed fragment just upstream of exon 3 with a low signal.

Table 2. Exons matching short RNAs

ExonType	Known exons matching short transfrag (%)	New exons matching short transfrag (%)
fe	64	37
e	11	12
te	39	19

This data relates to exons confirmed by a minimum of 25 nts expressed in tiling arrays (25) both in the cytoplasm and in short transcribed fragments.

do not alter the reading frame. This corresponds to what would be expected in a random-sized set of introns. The human data contain approximately twice the amount of splicing patterns with intron retentions as the mouse data.

To investigate the biological significance of transcripts with retained introns in human, we analyzed the expression data obtained from tiling arrays covering the whole nonrepetitive fraction of the human genome for eight cell lines (25). Of the 16 288 introns detected by our analysis, 7.5% were covered by <11 tiling array probes and therefore could not be analyzed. In contrast, 70% of the retained introns were covered by at least 50 nt of transcribed fragments isolated from either the nucleus or the cytoplasm and 45% were detected as having at least 50% of their surface covered by transcribed fragments isolated from the cytoplasm. This contrasts with 14% expression in cytoplasmic transcribed fragments detected in tiling arrays for all other non-retained (spliced) introns within the same size range (109 817 introns). This result prompted us to further characterize these introns that apparently are unspliced in the nucleus and are exported to the cytoplasm (referred to in this text as the 50CytoTF retained intron set).

Retained introns are short and a subset, associated with expression of small RNAs, has a high GC content

Introns come in a very wide range of sizes, from as small as 30 nt, the smallest intron size detected in RefSeq transcripts, to as large as several hundred kilobases, with the majority of introns (71%) being longer than 500 nt. An analysis of the size range of the 50CytoTF set of retained introns shows that the vast majority (87%) are shorter than 500 nt. In most of these cases (78%), the transcripts contain a single retained intron, whereas in 23% of the cases there are transcripts with more than one intron retained. As many functional short RNAs are transcribed from intronic regions of protein coding genes, we analyzed the distribution of short RNAs associated with three sets of introns: all non-retained introns; retained introns that are detected in the cytoplasm (50CytoTF set); and retained introns not confirmed by tiling array data (Figure 4). According to the tiling array data, we found that short RNAs are generally associated with introns of all sizes (Figure 4A). However, within the size range of retained introns, we found that retained introns confirmed by tiling array data (50CytoTF set) have a higher correlation with short RNAs than non-retained introns (Figure 4B), P-value <0.0001.

Recently published deep sequencing data (8) were also analyzed for the presence of retained introns. In the published data, we found 105 retained introns ranging in size from 69 to 327 nt (Supplementary Data). With one exception, all of these intron retentions occur between RefSeq annotated exons; however, the intron retentions

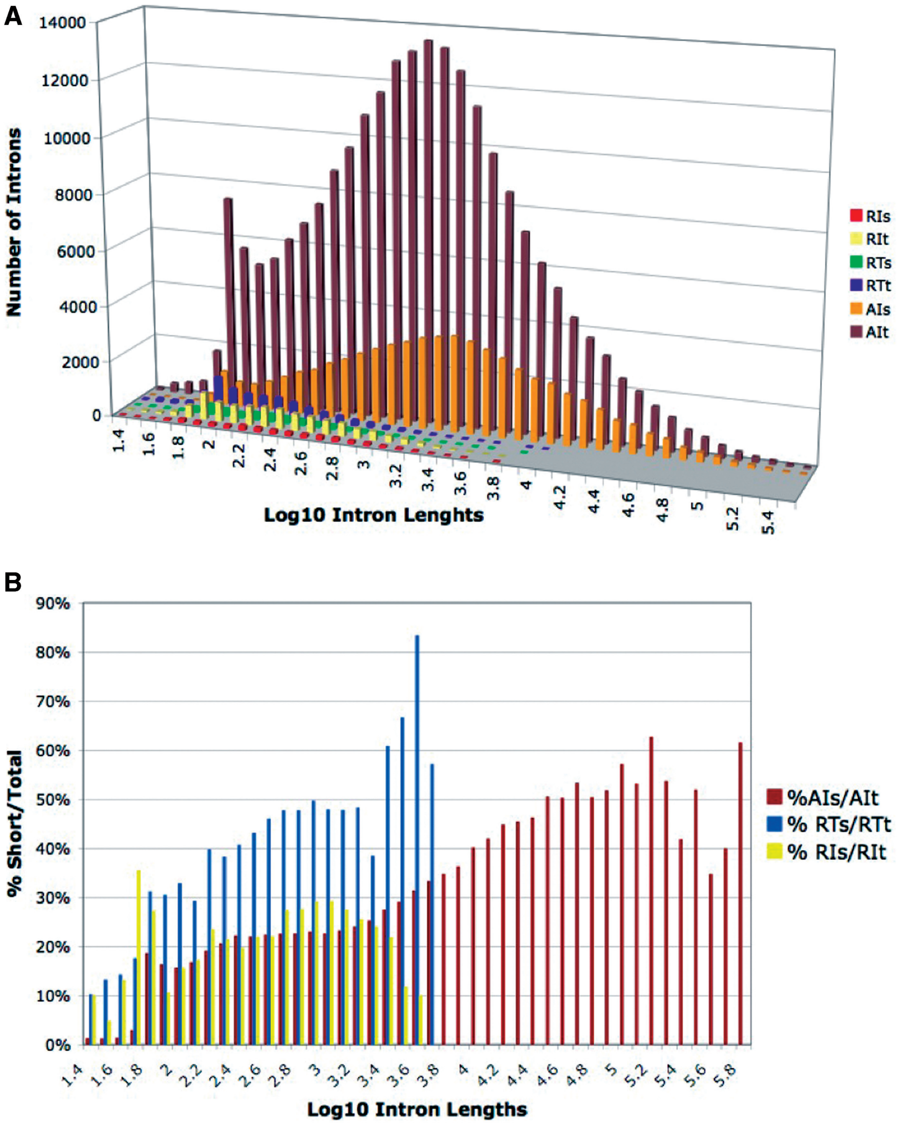


Figure 4. Distribution of intron size range and presence of short RNAs. (**A**) The plot represents the distribution of size range of three sets of introns. AIt: total non-retained introns (222721 introns); RTt: total retained introns with 50% of surface matching tiling array (25) transcribed fragments in the cytoplasm (50CytoTF set, 7381 introns); and RIt: total retained introns not confirmed by tiling array data (8907 introns); and their corresponding subsets matching short transcribed fragments detected in tiling arrays (AIs: 60 090 introns, RTs: 2663 introns and RIs: 1908 introns). Tiling array data for short RNAs (short transcribed fragments, 22–200 nt) was taken from (25). (**B**) Percentage of total introns in each set containing short RNAs plotted against Log10 of intron length. % AIs/AIt: non-retained introns; % RTs/RTt: 50CytoTF retained intron; % RIs/RIt: retained introns not detected in tiling array data. Within the size range of retained introns, this plot reveals that the 50CytoTF set of introns carries more short RNAs than non-retained introns. The calculated two tailed *P*-value for the difference observed, is <0.0001 (see 'Materials and Methods' section).

themselves are not annotated by RefSeq. Our analysis, on the other hand, detects 24 of these intron retentions. In 92% of the cases detected by deep sequencing, the retention of the intron does not disrupt the reading frame. Comparison of intron retention events detected by deep sequencing with the tiling array data shows that 72 of these introns (68%) have 50% coverage by cytoplasmic transcribed fragments and 34 (33%) are associated with short transcribed fragments (short RNAs).

The mean GC content of human introns in general has been reported as 43.51%, small introns (smaller than 1029 nt), having a higher GC content (28). We undertook an analysis of GC content in three sets of introns smaller than 1029 nt: small retained introns (from 50CytoTF set) matching short transcribed RNAs (set Rs); small retained introns (from 50CytoTF set) not matching short transcribed fragments (set Rns); and small introns not found to be retained (set NR). In addition to the size restriction, we required that the sets Rns and NR have the same size distribution as set Rs, by random selection of the same number of introns in each quartile and outliers as were observed in set Rs (Figure 5A). The GC content of each set was then measured and we found set Rs to have 63% GC content compared to 50% and 49% in sets Rns and NR, respectively (Figure 5B). We obtained a *P*-value = 1.9*e*-26 for comparing the mean GC content of sets

Rs and Rns, and a *P*-value = 0.14 for sets Rns and NR. A *P*-value that is so very small, in the former result, means that, with regard to GC content, it is close to certainty that Rs introns form a distinct set. As GC-rich regions can form more stable RNA secondary structures, this enrichment could reflect such a tendency and be implicated in involvement of splicing regulation in small RNA biogenesis.

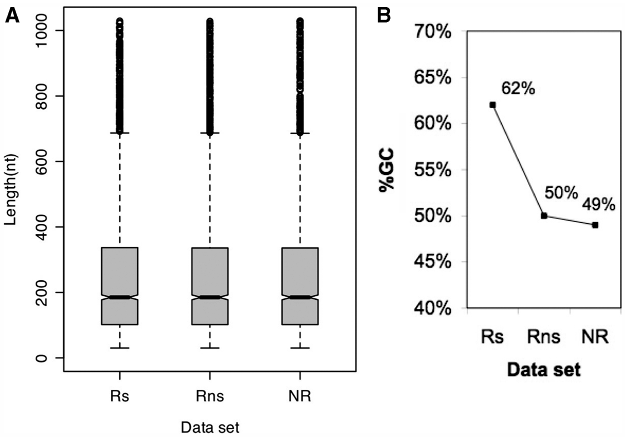


Figure 5. GC-content in retained and non-retained introns. GC content in three sets of 2500 introns each. Set Rs: small retained introns validated by tiling array data (25) in the cytoplasm and also matching short transcribed fragments; set Rns: small retained introns not matching short transcribed fragments; and set NR: small non-retained introns. Small introns are <1029 nt as defined in (28). (**A**) Boxplots of all three sets Rs, Rns and NR. Introns in sets Rns and NR are composed of a random selection of the same number of introns in each quartile and outliers as in set Rs: lower hinge = 30, extreme lower whisker = 102, median = 185, upper hinge = 337, extreme upper whisker = 687, lower extreme of notch = 177.6, upper extreme of notch = 192.4, 153 outliers. (**B**) Percent GC content in each of the three sets Rs, Rns and NR.

Next, we compared the frequency of occurrence of each nucleotide at the splice sites using WebLogo (39) (Figure 6), for the three above sets, Rs, Rns and NR and for all introns. At the 5′ splice site, the most prominent feature in small introns (Figure 6A, C and D) is a tendency for higher guanine content rather than a preference for uridines as is observed in all introns (Figure 6B); and in set Rs (Figure 6A) the higher GC content at this splice site is even more marked than for small non-retained introns (NR) and small retained introns not matching short transcribed fragments (Rns). The strength of the Py-tract at the 3′ splice site can be assessed in terms of its uridine content: at the 3′ splice site, set Rs appears to have a much weaker Py-tract (Figure 6A) than the other three sets. Set Rns of retained introns also appears to have a shorter and weaker Py-tract than non-retained small introns (set NR) with long introns showing the strongest Py-tract. The weakness of the Py-tract in retained introns (sets Rs and Rns) is likely to be implicated in regulation of alternative splicing requiring additional splicing factors (40).

Given the possibility of retained introns being involved in production of small RNAs we searched the miRBase Sequence Database for precursors of microRNAs, and for C/D and H/ACA box small nucleolar RNAs (snoRNAs) and small Cajal body-specific RNAs (scaRNAs) in snoRNABase data from the Laboratoire de Biologie Moléculaire Eucaryote. In the Rs set of retained introns we found 1 known microRNA, and 11 known C/D box and 7 H/ACA box snoRNAs (Supplementary Data). By contrast, in the Rns set of retained introns, no known small RNAs were found and only one known C/D box snoRNA was found in non-retained introns (set NR).

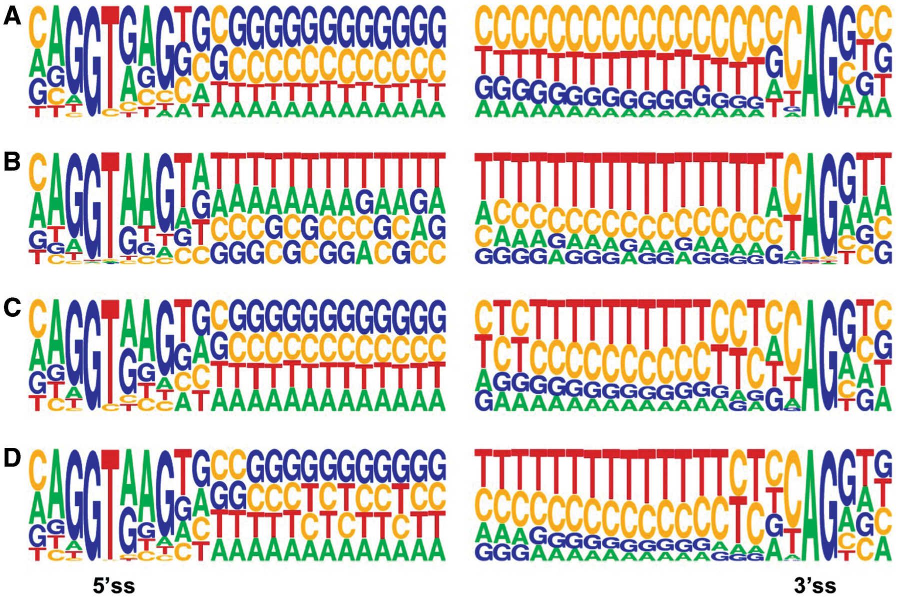


Figure 6. Frequency of nucleotide occurrence at splice sites. Logos representing the frequency of occurrence of nucleotides at each position at the 5′splice site (3 nt upstream and 20 nt downstream) and at the 3′splice site (30 nt upstream and 3 nt downstream) were produced using WebLogo (39). Uridines are represented by Ts. (**A**) Set Rs, 2500 small retained introns matching short RNAs, as described in Figure 5. (**B**) Random set of 2500 introns of all sizes. (**C**) Set Rns, 2500 small retained introns not matching short RNAs, as described in Figure 5. (**D**) Set NR, 2500 small non-retained introns, as described in Figure 5.

Conservation of retained introns and unannotated exons

Conservation analysis of unannotated exons and retained introns was performed using discontiguous megablast for eight species: nonhuman primates were represented by chimp and rhesus; nonprimate supraprimates were represented by the rodents mouse and rat; laurasiatheria, another large group of placental mammals, was represented by horse, dog and cow; and a nonmammalian vertebrate was represented by chicken. To evaluate conservation, we counted the number of exons or introns above the threshold of 70% identity over at least 80% of the sequence (Figure 7). The threshold of conservation imposed would be expected to detect sequences with protein coding potential. However, for chimp and rhesus, the high level of overall sequence similarity to the human genome, 98–99% in chimp (41) and 93% in rhesus (42), precludes the use of sequence conservation to postulate on the protein coding potential of a sequence.

Conservation in introns was analyzed for sets Rs and Rns of retained introns, and in the set NR of non-retained introns, all with the same size distribution as set Rs (Figure 7A). For retained introns not matching short transcribed fragments (Rns) we observed, in all placental organisms, a higher number of conserved introns than for retained introns matching short transcribed fragments (Rs). Non-retained introns within the same size range (set NR), on the other hand contain a lower (in the case of rodents, dog and cow) or equivalent (in the case of horse, rhesus and chimp) number of conserved introns as set Rs. This supports the non-coding hypothesis for set Rs setting it apart from set Rns. The number of introns conserved in chicken was negligible.

To evaluate conservation in exons, the analysis was performed on 9371 unannotated exons with expression validated by a minimum of 50% sequence coverage by tiling array (25) transcribed fragments detected in the cytoplasm (Figure 7B). Exons containing repetitive elements were excluded. This analysis was done alongside a random set of 2500 annotated exons. In all placental mammals, we found a significant number of exons conserved, 18% in rodents, 28% in dog and cow and 36% in horse, but only 4% in chicken.

One thing that stands out in the conservation analysis of both introns and exons is that a greater number of sequences appear to be conserved in horse, dog and cow than in rodents, which are generally accepted as being more closely related to primates.

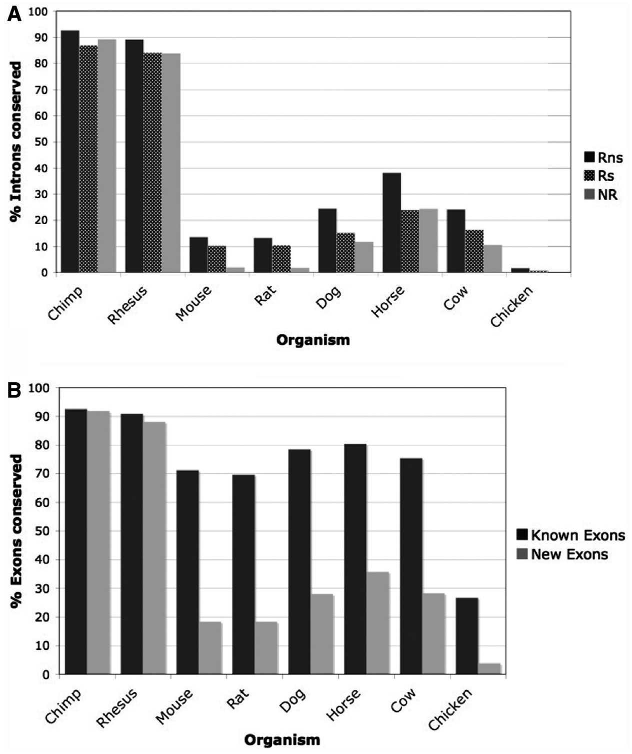


Figure 7. Conservation of unannotated exons and retained introns. Conservation estimated using discontiguous megablast (see 'Materials and Methods' section) against eight species: chimp, rhesus, mouse, rat, dog, horse, cow and chicken. The bars represent the percentage of exons or introns with a minimum of 70% sequence conservation over a minimum of 80% sequence coverage. (**A**) Conservation in intron sets Rs, Rns and NR as described in Figure 5, with the same size distribution. (**B**) Conservation of known (RefSeq annotated) and new (unannotated). Known exon set consists of a selection of 2500 random exons. Unannotated exons consist of the set of 9371 previously unannotated exons validated by more than 50% coverage of tiling array transcribed fragments in the cytoplasm. The error associated with the use of random sets of 2500 sequences was estimated at less than ± 1%.

DISCUSSION

Alternative splicing is often proposed to account for the differential complexity of organisms with similar number of genes and is considered an important driver of the evolution of phenotypic complexity in mammals (43). This assumption implies that more complex organisms are expected to have more alternative splicing. However, previous studies addressing this view have provided conflicting results (35,44–47). However, assessing the complete inventory of gene and mRNA isoform expression in different cell types of different organisms has been hampered by substantial technical challenges. Initial studies, using expressed sequence tags, have yielded relatively low estimates of alternative splicing, in part because EST data contain various sources of limitations—for example, EST coverage is typically biased toward the 3′- and 5′-ends of transcripts, EST quality and methods by which they were produced are very variable and, in general, there are insufficient numbers of sequenced transcripts covering different cell types in multiple species. More recently, microarray analyses achieved a more comprehensive coverage of the human transcriptome (6,7), but this approach is still constrained by relying on hybridization to specific probe sets. The most recently developed high-throughput sequencing technologies have the potential to circumvent these limitations and several recent studies have used mRNA-Seq data to survey alternative splicing in human and mouse transcriptomes (8–10,48–50). An important corollary of these studies was the identification of new exons and splice junctions not previously included in transcript databases.

Here, we show that many of the novel isoforms do not appear in current alternative splicing databases due to constraints imposed in the selection of transcripts. We therefore developed a less-constrained method and, having made the complete data set available, online at http://www.imm.fm.ul.pt/exonmine/, we proceeded to use this novel resource to investigate alternative splicing in the human and mouse transcriptomes.

Using our data mining method, we detected a higher rate of alternative splicing in the human compared to the mouse transcriptome. Consistent with our results, Kim *et al.* (35) reported that humans have more alternatively spliced genes and alternative exons than mice, and confirmed that this higher level of alternative splicing is not a result of ESTs derived from cancer cells that could increase the amount of aberrant splicing. However, given the limitations of EST data, future deep-sequencing analyses of multiple transcriptomes is likely to provide a clearer picture of the relationship between alternative splicing and organismal complexity.

In this study, extensive mining of mRNA and EST data resulted in the identification of previously unannotated exons in 72% of human genes. In the vast majority of cases (97.7%), the novel exons were detected in completely spliced transcripts. The results further reveal a high proportion of novel first exons, suggesting that alternative initial exons in the human transcriptome might be more frequent than previously thought. We also found overall a much greater number of first and terminal exons associated with short RNAs, which could reflect positional regulation of gene transcription at the start and end of a gene, leaving internal exons to perform mainly a protein coding function.

We found that 24% of unannotated exons aligned with repeat elements, suggesting that the majority probably evolved by a mechanism distinct from exonization of frequently transposed elements such as the Alu element, which was the most abundant element found in ~6000 (12%) unannotated exons. From a total of 48 942 unannotated exons, 44% were validated by tiling array expression data which covers the nonrepetitive part of the human genome, with 35% expressed in the cytoplasm, whereas 37% were not covered by probes in the array. This analysis of expression data from human tiling arrays provides a strong validation of unannotated exons identified by our data mining method.

A large fraction of unannotated exons appeared to be first exons. Previous strategies for collecting alternative splicing information have relied to a large extent on complete transcripts, whereas our approach relies extensively on ESTs, which are fragments of transcripts with a bias for the 3′- and 5′-end. First exons, in our data are simply exons for which no 3′ splice site was detected. We estimated the likelihood of first exons being true first exons by matching this data with the latest version of the database of TSSs DBTSS. This analysis confirmed that even though the data consist of fragments of transcripts and may be incomplete the majority of first exons, 63–75%, harbor TSS and are therefore likely to be true first exons.

Recent genome-wide studies of mammalian promoters reveal multiple TSSs on a given gene and although some may be cryptic, many are *bona fide* differentially regulated TSS generating alternative N-termini in protein coding genes (51). In our procedure for establishing first exons, the cut-Start of each first exon in the raw data, with a common 5′ss, is extended upstream to the farthest transcript start position in the data, so that our first exons are likely to include several TSSs. All original Genbank accessions for the data associated with each first exon (or indeed any exon) can be obtained, from our website (http://www.imm.fm.ul.pt/exonmine/), by inputting the exon number in 'Tissue distribution in Gene given the transcript, junction or exon' in the 'Downloads' section: if more than one accession is given for a first exon, then the first nucleotide of these GenBank transcripts will either start at the same position as an ExonMine first exon or further downstream within it. This means that our first exons, which match DBTSS, are more likely to correspond to a single promoter. However, if the RNA polymerase, originally docking at such a promoter, hypothetically, moved far enough along the gene to start transcription past the first 5′ss, in that case we would record, in our data, a new first exon or an extension of the first exon with an alternative 5′ss. On the other hand, mammalian promoter architecture is extremely complex and too diverse to allow accurate computational prediction, in fact it is thought that most genes may have multiple promoters, each one being associated with multiple TSSs, so that alternative promoter usage and regulation of TSS choice is likely to be responsible for generating considerable diversity and complexity in the mammalian transcriptome (52). Although it is clear that the biological determination of individual promoters, the initial docking sites of the polymerase, and the relation to the choice of a particular TSS, is a highly complex and unresolved problem, sometimes even involving more than one polymerase (53), this analysis could be approached using chromatin immunoprecipitation (ChIP) methods for polymerases coupled with validation by tissue specific microarray (ChIP–chip) or massively parallel sequencing methods (ChIP-Seq) (54,55).

The case for terminal exons, which have been extended downstream to the furthest known cut end in the GenBank data used, is similar. Our terminal exons (exons for which only a 3′ss has been detected, i.e. beyond which no more splicing is observed) can contain several alternative polyadenylation sites. We assessed the likelihood of terminal exons being true terminal exons through the presence of the poly-A signal and known variants of this signal. Here, we obtained a level of estimated true terminal exons of 30% for the AATAAA poly-A signal and 61% when known variants of the signal were considered. Terminal exons where no poly-A signals were found are likely to be internal exons from incomplete transcripts. An initial assessment of alternative polyadenylation site usage within our terminal exons can be done by looking at the GenBank data supporting a particular terminal exon.

Several forms of alternative splicing can generate, from a single gene with multiple introns, a variety of mRNAs containing different combinations of exons. The major forms of alternative splicing include: exon skipping, alternative 3′ splice site usage, alternative 5′ splice site usage, mutually exclusive exons, alternative initial exons, alternative terminal exons and intron retention (56). To date, there are only a few cases of intron retention events with known biological consequences in mammals. These include the transcription factor *Id3* (56), the splicing factor *9G8* (57), proinsulin (58), ion channels *CACNA1H* (59) and *KCNMA1* (60), and the Robo3 receptor involved in axon guidance (61). Previous bioinformatic studies of EST data found evidence for at least one intron retention event in 5–15% of human genes (5,62–64), and some of the new predicted variants containing retained introns have been experimentally validated (59). However, EST sequence information alone is generally considered insufficient to identify truly functional retained-intron isoforms due to possible artifacts derived from aberrant cDNA cloning (e.g. cDNAs produced from incompletely spliced mRNAs or genomic DNA contaminants).

Using our data mining method, we identified 16 288 intron retention events in 41% of human genes. Of the 16 288 retained introns, 7382 (45%) were detected in the cytoplasm of human cell lines by tiling array analysis (25). It is therefore unlikely that these transcripts are derived from unspliced or partially spliced nuclear pre-mRNAs. Intron retention involves excluding the normal recognition of a pair of splice sites and bypassing the surveillance mechanisms that typically prevent export of unspliced pre-mRNAs to the cytoplasm. Thus, intron retention is probably regulated by factors involved in both splicing and mRNA export. Retroviruses (including HIV) are the best studied systems involving regulated intron retention and export of unspliced mRNA to the cytoplasm (65). Unspliced viral RNA overcomes nuclear restriction by using a *cis*-acting RNA element known as the constitutive transport element (CTE), which interacts directly with the principal export receptor for cellular mRNA. Most probably, cellular mRNAs with retained introns use cellular CTE equivalents to be exported to the cytoplasm (66,67). Retained introns are therefore expected to contain some common sequence elements that might participate in the coordinate regulation of splicing and RNA export. Indeed, both intronic and exonic motifs that affect other types of alternative splicing have been shown to regulate intron retention when properly situated (68,69).

Here, we report several lines of evidence supporting the view that intron retention events in the human transcriptome are specifically regulated. First, retained introns have weak Py tracts contrasting with the majority of human introns which show a significant bias for strong Py tracts (70). Weak Py-tracts have previously been shown to be implicated in regulation of alternative splicing requiring additional splicing factors (40). Second, the vast majority (87%) of retained human introns are shorter than 500 nt contrasting with non-retained introns, which are typically longer than 500 nt. Third, both the weak Py tract (71) and size of retained human introns are reminiscent of plant introns [100–200 nt long (71)], and in plants intron retention is the most prevalent form of alternative splicing (72–74). Finally, a recent analysis of the human transcriptome using tiling arrays at 5-nt resolution revealed the presence of potentially functional short RNAs ranging in length from 22 to 200 nt, many of which were found in intronic regions (25). Mining these data sets, we found that transcripts containing retained introns that are likely precursors of short RNAs were detected in the cytoplasm, suggesting that intron retention events might correlate with the cellular production of short, probably non-coding RNAs. Consistent with this view, we show that the set of retained introns matching short RNAs was enriched in G and C nucleotides. G- and C-rich regions are likely to form more stable RNA secondary structures. In addition to this, we found a number of known small RNAs in retained introns matching short transcribed fragments (18 snoRNSs and 1 microRNA in total). This result further supports our hypothesis that a subset of retained introns contains small RNAs and calls for experimental validation of coordination between splicing regulation of intron retention and production of particular non-coding RNAs.

Conservation analysis is often used to indicate functionality. However, non-coding transcripts, which we have identified through consensus splice-sites alone, would be expected to have a low level of overall sequence conservation. If conservation exists, for example, at the level of RNA secondary structure, these sequences may have undergone divergence, which precludes their detection by primary sequence conservation alone. Our conservation analysis, of both tiling array validated unannotated exons and retained introns, shows a greater number of sequences conserved in horse, dog and cow than in mouse and rat. Phylogenetic studies performed in 2001 (75) suggested a common placental ancestor for human and rodents branching off from a distinct placental ancestor for horse, cow and dog. This would imply that rodents might have lost characteristics which could still be common to primates and laurasiatheria. However, more recent phylogenetic analyses from 2007 (76), performed using whole-genome data, instead resulted in overwhelming evidence for a human-carnivore clade with the exclusion of rodents. The latter lends support to our results, which indicates a greater number of human exons and Retained introns conserved in dog than in rodents.

Concerning conservation in primates alone, as the human sequences used in this conservation analysis were validated by tiling array data, it is rather the portion of exons and retained introns not conserved in the nonprimate mammalian species, that stand out from the overall background of conservation in the other model organism, which may function in a way unique to primate biology. However, given the *a priori* high level of sequence conservation between human, chimp and rhesus, and the lack of sequenced transcript information for nonhuman primate model organisms, it is not yet possible to determine precisely to what extent additional complexity in alternative splicing detected in human through novel exons and retained introns may be spurious or specific to primates. Notwithstanding the lack of concrete evidence for primate specific function, this study highlights a considerable number of sequences, coded within known gene regions and detected in the cytoplasm, which could be unique to primate biology.

Ongoing efforts by multiple research teams to determine levels of alternative splicing in all organisms will contribute to a comprehensive view of this complex problem. Each approach will bring in something that another approach may miss and likewise each approach will produce its own spurious data. Many efforts have focused on determining alternatively spliced protein-coding variants, and our approach expands this effort to non-protein-coding transcripts which consist of the majority of alternatively spliced variants. Recent evidence in the literature, indicating that non-protein-coding alternatively spliced transcripts may play an important regulatory function, strengthens this hypothesis. For example, in *Drosophila melanogaster* it has been shown that the conservation of short introns and of splicing can be used to identify spliced, capped and polyadenylated non-coding mRNA-like transcripts with low exon sequence conservation (77). Looking for conservation in introns flanking nonconserved unannotated exons would be a relevant approach. However, bioinformatics can only throw the spotlight onto potentially interesting avenues. The only way to show that the data are not spurious is to validate individual cases; and indeed, we are collaborating with groups who are using our data to do just that. For example, all unannotated exons detected in Factor VIII by our analysis have been recently validated in human pulmonary endothelium, all of these exons were first exons (78).

This bioinformatic analysis highlights potential avenues which merit attention and validation. Together with the data made available online to the scientific community, we hope that this study will contribute to a more comprehensive view of alternative splicing, in particular, to the role of intron retention and the role of novel exons in new species. Overall, these results reveal new levels of complexity in the regulation of alternative splicing in human.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted to Juan Valcárcel (CRG-Centre de Regulació Genómica, Barcelona, Spain) and members of his lab Britta Hartmann and Josefin Lundgren, to Samuel Aparicio (BC Cancer Agency, Vancouver, British Columbia, Canada) and to Célia Carvalho (Instituto de Medicina Molecular, Lisbon, Portugal) and Nuno Morais (Department of Oncology, University of Cambridge, UK) for very useful discussions on the data. We are also deeply indebted to the communities which provide the tools Perl, MySQL, PHP, R and Apache as open source software.

Author Contributions: I.G.M. designed the research, developed the ExonMine database, analysed the data and wrote the manuscript; C.B-D. revised the alternative splicing data; D.F-S. designed the ExonMine web interface; A.R.G. designed the ExonMine web interface tools; P.E. maintains of the ExonMine server; R.A. designed and implemented the ExonMine export features and maintains the ExonMine server and updates; R.S. supervised statistical analysis; T.S.S. supervised development of the ExonMine interface; M.C-F. supervised the research and revised the manuscript.

FUNDING

The Muscular Dystrophy Association (MDA3662); the European Commission (LSHG-CT-2005-518238, EURASNET); 'Fundação para a Ciência e Tecnologia, FCT', Portugal (PTDC/SAU-GMG/69739/2006). Funding for open access charge: 'Fundação para a Ciência e Tecnologia, Portugal (PTDC/SAU-GMG/69739/2006).

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
2. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell. Biol.*, **6**, 386–398.
3. Blencowe,B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37–47.
4. Wang,G.S. and Cooper,T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749–761.
5. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
6. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
7. Kampa,D., Cheng,J., Kapranov,P., Yamanaka,M., Brubaker,S., Cawley,S., Drenkow,J., Piccolboni,A., Bekiranov,S., Helt,G. *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, **14**, 331–342.
8. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
9. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
10. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
11. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
12. Hughes,T.A. (2006) Regulation of gene expression by alternative untranslated regions. *Trends Genet.*, **22**, 119–122.
13. Lareau,L.F., Brooks,A.N., Soergel,D.A., Meng,Q. and Brenner,S.E. (2007) The coupling of alternative splicing and nonsense-mediated mRNA decay. *Adv. Exp. Med. Biol.*, **623**, 190–211.

14. Grellscheid,S.N. and Smith,C.W. (2006) An apparent pseudo-exon acts both as an alternative exon that leads to nonsense-mediated decay and as a zero-length exon. *Mol. Cell. Biol.*, **26**, 2237–2246.
15. Mollet,I., Barbosa-Morais,N.L., Andrade,J. and Carmo-Fonseca,M. (2006) Diversity of human U2AF splicing factors. *FEBS J.*, **273**, 4807–4816.
16. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
17. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
18. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
19. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
20. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
21. Church,D.M., Goodstadt,L., Hillier,L.W., Zody,M.C., Goldstein,S., She,X., Bult,C.J., Agarwala,R., Cherry,J.L., DiCuccio,M. *et al.* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, **7**, e1000112.
22. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
23. Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J.M. and Gautheret,D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, **10**, 1001–1010.
24. Yamashita,R., Wakaguri,H., Sugano,S., Suzuki,Y. and Nakai,K. (2010) DBTSS provides a tissue specific dynamic view of Transcription Start Sites. *Nucleic Acids Res.*, **38**, D98–D104.
25. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
26. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
27. Zhang,Z., Scott Schwartz,S., Wagner,L. and Webb Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
28. Gazave,E., Marqués-Bonet,T., Fernando,O., Charlesworth,B. and Navarro,A. (2007) Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol.*, **8**, R21.
29. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **334**, D140–D144.
30. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
31. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
32. de la Grange,P., Dutertre,M., Correa,M. and Auboeuf,D. (2007) A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants. *BMC Bioinformatics*, **8**, 180.
33. Kim,N., Alekseyenko,A.V., Roy,M. and Lee,C. (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.*, **35**, D93–D98.
34. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
35. Kim,E., Magen,A. and Ast,G. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.*, **35**, 125–131.
36. Lev-Maor,G., Ram,O., Kim,E., Sela,N., Goren,A., Levanon,E.Y. and Ast,G. (2008) Intronic Alus influence alternative splicing. *PLoS Genet.*, **4**, e1000204.
37. Ram,O., Schwartz,S. and Ast,G. (2008) Multifactorial interplay controls the splicing profile of Alu-derived exons. *Mol. Cell. Biol.*, **28**, 3513–3525.
38. El-Sawy,M. and Deininger,P. (2005) Tandem insertions of alu elements. *Cytogenet. Genome Res.*, **108**, 58–62.
39. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
40. Pacheco,T.R., Coelho,M.B., Desterro,J.M., Mollet,I. and Carmo-Fonseca,M. (2006) In vivo requirement of the small subunit of U2AF for recognition of a weak 3′ splice site. *Mol. Cell. Biol.*, **26**, 8183–8190.
41. The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
42. Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.
43. Xing,Y. and Lee,C. (2006) Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, **7**, 499–509.
44. Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nat. Genet.*, **30**, 29–30.
45. Harrington,E.D., Boue,S., Valcarcel,J., Reich,J.G. and Bork,P. (2004) Estimating rates of alternative splicing in mammals and invertebrates. *Nat. Genet.*, **36**, 916–917, author reply.
46. Kim,H., Klein,R., Majewski,J. and Ott,J. (2004) Estimating rates of alternative splicing in mammals and invertebrates. *Nature Genet.*, **36**, 915–916.
47. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Seki,M., Irie,T., Takeuchi,N., Ueda,T., Nakao,M., Sugano,S. *et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**, 6386–6395.
48. Bainbridge,M.N., Warren,R.L., Hirst,M., Romanuik,T., Zeng,T., Go,A., Delaney,A., Griffith,M., Hickenbotham,M., Magrini,V. *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, **7**, 246.
49. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
50. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
51. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engström,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
52. Sandelin,A., Carninci,P., Lenhard,B., Ponjavic,J., Hayashizaki,Y. and Hume,D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.*, **8**, 424–436.
53. Listerman,I., Bledau,A.S., Grishina,I. and Neugebauer,K.M. (2007) Extragenic accumulation of RNA polymerase II enhances transcription by RNA polymerase III. *PLoS Genet.*, **3**, e212.
54. Smith,A.D., Sumazin,P., Das,D. and Zhang,M.Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics.*, **21(Suppl 1)**, i403–i412.
55. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.
56. Forrest,S.T., Barringhaus,K.G., Perlegas,D., Hammarskjold,M.L. and McNamara,C.A. (2004) Intron retention generates a novel Id3 isoform that inhibits vascular lesion formation. *J. Biol. Chem.*, **279**, 32897–32903.
57. Lejeune,F., Cavaloc,Y. and Stevenin,J. (2001) Alternative splicing of intron 3 of the serine/arginine-rich protein 9G8 gene. Identification of flanking exonic splicing enhancers and involvement of 9G8 as a trans-acting factor. *J. Biol. Chem.*, **276**, 7850–7858.

58. Mansilla,A., López-Sánchez,C., de la Rosa,E.J., García-Martínez,V., Martínez-Salas,E., de Pablo,F. and Hernández-Sánchez,C. (2005) Developmental regulation of a proinsulin messenger RNA generated by intron retention. *EMBO Rep.*, **6**, 1182–1187.
59. Zhong,X., Liu,J.R., Kyle,J.W., Hanck,D.A. and Agnew,W.S. (2006) A profile of alternative RNA splicing and transcript variation of CACNA1H, a human T-channel gene candidate for idiopathic generalized epilepsies. *Hum. Mol. Genet.*, **15**, 1497–1512.
60. Bell,T.J., Miyashiro,K.Y., Sul,J.Y., McCullough,R., Buckley,P.T., Jochems,J., Meaney,D.F., Haydon,P., Cantor,C., Parsons,T.D. *et al.* (2008) Cytoplasmic BK(Ca) channel intron-containing mRNAs contribute to the intrinsic excitability of hippocampal neurons. *Proc. Natl Acad. Sci. USA*, **105**, 1901–1906.
61. Chen,Z., Gore,B.B., Long,H., Ma,L. and Tessier-Lavigne,M. (2008) Alternative splicing of the Robo3 axon guidance receptor governs the midline switch from attraction to repulsion. *Neuron*, **58**, 325–332.
62. Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.
63. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
64. Galante,P.A., Sakabe,N.J., Kirschbaum-Slager,N. and de Souza,S.J. (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **10**, 757–765.
65. Cullen,B.R. (2003) Nuclear mRNA export: insights from virology. *Trends. Biochem. Sci.*, **28**, 419–424.
66. Li,Y., Bor,Y.C., Misawa,Y., Xue,Y., Rekosh,D. and Hammarskjöld,M.L. (2006) An intron with a constitutive transport element is retained in a Tap messenger RNA. *Nature*, **443**, 234–237.
67. Bor,Y.C., Swartz,J., Morrison,A., Rekosh,D., Ladomery,M. and Hammarskjöld,M.L. (2006) The Wilms' tumor 1 (WT1) gene (+KTS isoform) functions with a CTE to enhance translation from an unspliced RNA with a retained intron. *Genes Dev.*, **20**, 1597–1608.
68. Wang,Z., Xiao,X., Van Nostrand,E. and Burge,C.B. (2006) General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell.*, **23**, 61–70.
69. Marcucci,R., Baralle,F.E. and Romano,M. (2007) Complex splicing control of the human Thrombopoietin gene by intronic G runs. *Nucleic Acids Res.*, **35**, 132–142.
70. Schwartz,S.H., Silva,J., Burstein,D., Pupko,T., Eyras,E. and Ast,G. (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.*, **18**, 88–103.
71. Ner-Gaon,H., Leviatan,N., Rubin,E. and Fluhr,R. (2007) Comparative cross-species alternative splicing in plants. *Plant Physiol.*, **144**, 1632–1641.
72. Iida,K., Seki,M., Sakurai,T., Satou,M., Akiyama,K., Toyoda,T., Konagaya,A. and Shinozaki,K. (2004) Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucleic Acids Res.*, **32**, 5096–5103.
73. Ner-Gaon,H., Halachmi,R., Savaldi-Goldstein,S., Rubin,E., Ophir,R. and Fluhr,R. (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.*, **39**, 877–885.
74. Wang,B.B. and Brendel,V. (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl Acad. Sci. USA.*, **103**, 7175–7180.
75. Murphy,W.J., Eizirik,E., Johnson,W.E., Zhang,Y.P., Ryder,O.A. and O'Brien,S.J. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.
76. Cannarozzi,G., Schneider,A. and Gonnet,G. (2007) A phylogenomic study of human, dog, and mouse. *PLoS Comput. Biol.*, **3**, e2.
77. Hiller,M., Findeiss,S., Lein,S., Marz,M., Nickel,C., Rose,D., Schulz,C., Backofen,R., Prohaska,S.J., Reuter,G. *et al.* (2009) Conserved introns reveal novel transcripts in Drosophila melanogaster. *Genome Res.*, **19**, 1289–1300.
78. Shovlin,C.L., Angus,G., Manning,R.A., Okoli,G.N., Govani,F.S., Elderfield,K., Birdsey,G.M., Laffan,M.A., Mollet,I.G. and Mauri,F.A. (2010) Endothelial cell processing and alternatively spliced transcripts of Factor VIII. Potential implications for coagulation cascades and pulmonary hypertension. *PLoS ONE*, **5**, e9154.