BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 2 2014, pages 151–156
doi:10.1093/bioinformatics/btt668

Genome analysis

Advance Access publication November 20, 2013

A novel splicing outcome reveals more than 2000 new mammalian protein isoforms

Laurence O. W. Wilson[1], Andrew Spriggs[2], Jennifer M. Taylor[2] and Aude M. Fahrer[1,*]

[1]Research School of Biology, Australian National University, Canberra, ACT 0200 and [2]CSIRO Plant Industry, Black Mountain Laboratories, Canberra, ACT 2601, Australia

Associate Editor: John Hancock

ABSTRACT

Motivation: We have recently characterized an instance of alternative splicing that differs from the canonical gene transcript by deletion of a length of sequence not divisible by three, but where translation can be rescued by an alternative start codon. This results in a predicted protein in which the amino terminus differs markedly in sequence from the known protein product(s), as it is translated from an alternative reading frame. Automated pipelines have annotated thousands of splice variants but have overlooked these protein isoforms, leading to them being underrepresented in current databases.

Results: Here we describe 1849 human and 733 mouse transcripts that can be transcribed from an alternate ATG. Of these, >80% have not been annotated previously. Those conserved between human and mouse genomes (and hence under likely evolutionary selection) are identified. We provide mass spectroscopy evidence for translation of selected transcripts. Of the described splice variants, only one has previously been studied in detail and converted the encoded protein from an activator of cell-function to a suppressor, demonstrating that these splice variants can result in profound functional change. We investigate the potential functional effects of this splicing using a variety of bioinformatic tools. The 2582 variants we describe are involved in a wide variety of biological processes, and therefore open many new avenues of research.

Contact: aude.fahrer@anu.edu.au

Supplementary Inforation: Supplementary data are available at *Bioinformatics* online.

Received on July 10, 2013; revised on November 8, 2013; accepted on November 12, 2013

1 INTRODUCTION

Alternative splicing is one of the most significant contributors to proteome diversity (Nilsen and Graveley, 2010). Through the production of multiple mRNA transcripts, a single gene can encode multiple and functionally distinct protein isoforms. These isoforms can differ subtly, varying in only a single amino acid, or demonstrate drastic changes in their peptide sequences accompanied by equally significant changes in their function. Automated pipelines have been developed which can identify and annotate alternative splice events and their encoded proteins, primarily through the alignment of Expressed Sequence Tags (ESTs) (Bonizzoni *et al*., 2009; Brett *et al*., 2000; Eyras *et al*., 2004; Hiller *et al*., 2004; Kan *et al*., 2002).

We have previously characterized an alternative splice form of the mouse non-SMC condensin II complex, subunit H2 (Ncaph2) gene that was incorrectly annotated by these pipelines (Gosling *et al*., 2007, 2008; Theodoratos *et al*., 2012). The gene can produce a transcript that lacks the last 17 bp of the first exon. While normally such a deletion would produce a frame-shift of the downstream sequence, translation at an alternate start codon rescues the reading frame of the sequence downstream from the splice event while shifting that of the upstream sequence. The result is a protein isoform that possesses a unique amino terminus, but in which the remainder of the protein is identical to the known isoform (Fig. 1).

An exhaustive literature search revealed only three other proteins displaying similar alternative splicing; mouse Innositol polyphosphate-4-phosphatase, type II (Inpp4b), human Leishmanolysin-like (metallopeptidase M8 family) (LMLN) and Otubain-1 (OTUB1) (Cobbe *et al*., 2009; Ferron and Vacher, 2006; Soares *et al*., 2004). All couple a frame-shifting splice event with an alternate start codon to change the reading-frame of the amino terminus, and all have been overlooked by the genome annotation pipelines, leading us to question if the few documented instances of this splicing are a true representation of their prevalence.

Here we report the genome-wide identification of new examples of this type of alternative splicing, using publically available EST data.

2 METHODS

2.1 Identification of novel protein isoforms

Reference gene and EST sequences were downloaded for human and mouse genomes from NCBI Unigene. The original search was conducted using Unigene Builds #231 for human and #192 for mouse, based on RefSeq Assembly 46. EST sequences were aligned to the gene reference sequence using BLAST (v2.2.12) (Altschul *et al*., 1990). In cases where a gene possesses more than one reference sequence, we chose the longest. Alignments were conducted using gap initiation and extension penalties of 5 and 2, respectively and an *E*-value threshold of 10^{-5}. Events were required to show an insertion or deletion of at least four nucleotides not divisible by three in an annotated region of translation. These events also needed to be flanked by a minimum of 20 unambiguous bases of sequence on either side and observed in at least two independent EST sequences. Events creating exons of <20 bases or insertion events contained entirely within annotated exons were discarded. Finally, transcripts arising from these events that could be: translated into a peptide sequence from an alternate start codon, with no premature stop codons, and thus possessing a conserved carboxy-terminus relative to the canonical protein were

[*]To whom correspondence should be addressed.

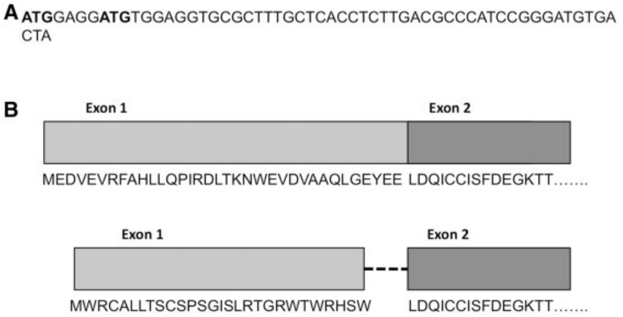
A
ATGGAGGATGTGGAGGTGCGCTTTGCTCACCTCTTGACGCCCATCCGGGATGTGA
CTA
B
Exon 1
Exon 2
MEDVEVRFAHLLQPIRDLTKNWEVDVAAQLGEYEE LDQICCISFDEGKTT.......
Exon 1
Exon 2
MWRCALLTSCSPSGISLRTGRWTWRHSW
LDQICCISFDEGKTT.......

Fig. 1. Schematic representation of the alternative splicing of the Ncaph2 gene. (A) Partial sequence of the Ncaph2 gene showing the two start codons (bold). (B) Schematic representation of the two splice variants of Ncaph2. In the canonical form, the first exon is translated from the first ATG. The alternate form begins translation at the second start codon and deletes the last 17 bases of the first exon, resulting in the first exon being translated in a different reading frame. The remaining 19 exons are the same between the two splice variants

selected. A summary of the full search strategy is shown in Supplementary Figure S1.

Prevalence of alternative splicing within gene ontology biological process functional categories was determined using DAVID gene ontology software (Huang da *et al.*, 2009).

2.2 Identification of conserved alternative splicing

Genes undergoing splicing in both mouse and human were identified and the resulting transcripts compared. If the transcripts were of a similar size (within 60 bp of each other) they were aligned using clustalw2 (v2.1) (Larkin *et al.*, 2007) and the results passed to PhyloCSF (Lin *et al.*, 2011). The protein sequences of alignments that received a positive score from PhyloCSF (which is indicative of conservation) were then checked manually.

2.3 Analyzing ribosome profiling data

Alternate start codons predicted in this study were considered supported by the results of Lee *et al.* (2012) if the Ribosome Profiling study identified a start codon in the same Reference Sequence at the same position. Identification of supported start sites was facilitated by the fact that both our studies used the same numbering method; the position of start codons was determined relative to their position within a Reference Sequence.

2.4 Secondary structure analysis of splice junctions

A 60-bp window of sequence centered on either the 5′- or 3′-splice junctions was identified by mapping the regions flanking the splice event (provided in Supplementary Tables S1 and S2) back to the genomic sequence (downloaded from ENSEMBL, Human Assembly GRCh37.p12; Mouse Assembly GRCm38.p1). The initial scan identified 97% of the predicted splice junctions. The remaining 3% were excluded from the analysis. An equal number of randomly chosen exons (downloaded from ENSEMBL) were used to generate a control group of splice junctions. The level of secondary structure was determined by predicting the Mean Free Energy (MFE) of the sequence using RNAfold (Denman, 1993), a component of the ViennaRNA package (v2.1.3) with default parameters. Differences between the groups of exons were assessed using an un-paired *t*-test.

2.5 Identification of supporting Mass Spectrometry evidence

We adopted a combined-database approach of peptide identification (Elias and Gygi, 2007). Our predicted alternate-proteins were appended to a list of all annotated human or mouse proteins downloaded from NCBI. We also included reversed versions of all the target proteins to act as decoys in order to calculate the False Detection Rate (FDR). Mass spectra were obtained from Peptide Atlas (Desiere *et al.*, 2005). We used data from two experiments for human (PAe000116, derived from the human erythroleukemia K562 cell line and PAe000053, derived from human blood) and mouse (PAe000373, PAe000370, PAe000359 and PAe000380 derived from the ctyoplasmic, microsomal, mitochondrial and nuclear fractions respectively, of mouse whole brain extracts and PAe000300, derived from mouse testes). Spectra were then analyzed using X!Tandem (v2010.12.01.1) (Craig and Beavis, 2004) with the following parameters: parent ion mass tolerance of ±100, monoisotopic mass, trypsin cleavages only and no more than one skipped cleavage in the peptide. Peptide assignments were statistically validated using PeptideProphet (from the Trans-Proteomic Pipeline Suite v4.4.1) (Keller *et al.*, 2002), discounting spectra assignments below 0.05 or to peptides <7-amino acids long. Proteins identifications and the protein FDR were made using MAYU (Reiter *et al.*, 2009), ignoring ambiguous peptides. We chose a peptide-spectrum match FDR of 0.009 as our cutoff, which corresponded to protein FDRs of 6.69% and 5.95% for human and mouse, respectively. The complete list of alternate protein sequences used in the MS analysis is provided in Supplementary Table S1 and S2.

3 RESULTS

3.1 Identification of >2000 novel protein isoforms in human and mouse

In order to identify potential frame-shifted protein isoforms, we designed a program that downloaded each gene's ESTs from Unigene and then aligned them back to the reference sequence using BLAST. Alignments were then parsed for splice events that produced frame shifts that could be rescued through use of an alternate start codon. An overview of the search program is given in Supplementary Figure S1. Our program predicted 1849 alternative splice events in the human genome and 733 instances in the mouse genome that result in proteins with a novel amino-terminus derived from an alternate reading frame. The full lists of these transcripts are provided in Supplementary Tables S1 and S2. Examples of genes displaying the splicing are provided in Table 1. The highest percentage of supporting ESTs recorded was for human LGALS14 (lectin, galactoside-binding, soluble 14), in which the 'alternative' transcript accounts for 58% of the total transcripts. Thus, this previously under-appreciated form of alternative splicing can affect a large percentage of the annotated transcripts for a given gene.

We compared our alternate protein sequences to a list of all annotated proteins present in the NCBI and ENSEMBL/GENCODE databases. In total 324 (18%) of the human and 44 (6%) of the mouse proteins have been previously annotated. Thus, >80% of the alternative translations we have identified are completely novel.

Examples of this unusual form of splicing were found in genes involved in a diverse range of biological processes. We used DAVID Gene Ontology software (Huang da *et al.*, 2009) to identify pathways enriched for this form of splicing. These

Table 1. Examples of genes identified as displaying the alternative splicing

	Number of supporting ESTs (%)	Number of codons read in two frames
Human Gene Name		
TBC1D15	211 (43.8)	32
LGALS14	26 (57.8)	5
MRPL10	184 (38.9)	18
FAM86B2	15 (53.6)	20
SLC2A14	111 (42.5)	6
NAE1	142 (38.2)	68
CABP2	11 (50)	25
CBWD1	180 (33.0)	18
MFAP4	3 (50)	77
SSX2	117 (38.1)	52
Mouse Gene Name		
1700067P10Rik	4 (28.6)	4
Asb18	2 (28.6)	9
Ccdc154	2 (28.6)	2
Agbl2	10 (27.0)	12
Rbm41	24 (25.3)	3
Ccdc159	8 (26.7)	4
Otub2	35 (24)	1
Zfp300	2 (22.2)	4
Slfn3	4 (21.1)	80
Eif2ak4	29 (18.4)	8

yielded enrichments in processes such as DNA replication and repair, programmed cell death, and RNA processing and translation pathways. Additionally we found multiple members of certain gene families displaying this splicing. Of the human genes identified, these included the Interleukin receptors (five members), the Eukaryotic Initiators of Translation (eight members) and the Solute Carrier families (41 members).

3.2 Conservation of splice events between human and mouse

Comparing the splicing events found in human and mouse identified 40 instances of conserved alternative splicing between the two species (Table 2). The conservation of these forms implies likely evolutionary selection. As well as the 40 genes displaying conserved splicing, we found an additional 102 genes that, while displaying this form of splicing in both human and mouse, did not produce conserved proteins. The relatively low species conservation is consistent with recent findings showing that alternative splicing is often species-specific (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012).

3.3 Secondary structure of splice sites

Analysis of the secondary structure surrounding the alternate splice junctions showed that in both human and mouse, the alternate 5′ junctions were predicted to possess a lower secondary structure and the 3′ junctions a higher level of structure when compared to a control group (Supplementary Figs S2 and S3). Given the influence of mRNA secondary structure on alternative splicing (Buratti and Baralle, 2004) these results may suggest a method of regulation, although the precise mechanism remains unclear.

Table 2. Genes identified as displaying conserved alternative splicing

Gene name	Number of codons affected (Human/Mouse)
ADSL	6/9
ALKBH3	12/12
AP1G2	6/2
ARL6IP1	12/5
ATF2	9/9
C1orf144	15/15
DENND1A	29/29
DNAJC1	10/10
EIF2A	39/39
ENO1	9/9
EYA1	9/9
FARSB	14/18
GAS7	34/34
GUSB	4/4
IQSEC2	35/35
ITGB6	5/5
LAPTM4A	14/14
MEIS2	2/2
NMT2	69/69
PBX1	29/29
PDK1	1/1
PHC2	2/2
PHF14	15/15
PHKB	26/26
POLR3E	6/6
PRLR	1/1
PUM1	2/2
RABAC1	25/25
RNF170	37/37
SCAMP5	27/27
SFSWAP	4/4
SLC6A9	25/25
SMC1A	4/4
STAU2	6/6
SYBU	9/9
UBE2B	5/5
UBE2G1	20/18
UFD1L	3/3
WDFY2	4/4
ZZZ3	8/8

3.4 Prevalence of Kozak sequences

Identification of these splice events using ESTs provides strong evidence for their transcription. Translation of these alternative splice forms however, relies on the translation machinery correctly identifying the alternate start codon, a process often mediated by the Kozak sequence (Kozak 1986, 1997). We therefore investigated the context in which the start codons were presented, paying attention to both the canonical and alternate start sites. We classified the surrounding sequence as either an optimal

Kozak sequence (defined as AnnATGn or GnnATGG), a sub-optimal Kozak sequence (defined as [C/T]nnATG[A/C/T]) or no Kozak sequence (Bazykin and Kochetov, 2010; Kochetov 2008; Volkova and Kochetov, 2010). Full results are given in Supplementary Tables S1 and S2. In human, 66% of the canonical start sites were found within an optimal Kozak sequence as opposed to 37% of the alternate sites (64% versus 37% for mouse). For both species, the alternate start sites were distributed evenly between optimal and sub-optimal Kozak sequences. Interestingly, the Kozak signals of the conserved alternate start sites appeared to be more likely to exist in an optimal Kozak sequence (47.5% of conserved sites versus 36.9% of non-conserved).

3.5 Proteomic evidence for the translation of the alternative splice forms

Ultimately, translation needs to be proven experimentally. We therefore adopted a proteomics approach, searching for evidence of the proteins in publicly available Mass Spectrometry (MS) data. As our proteins differ from the canonical forms only in the amino-terminus, such an approach is difficult because first, in order to be identified the alternate protein must produce unique peptides of appropriate length upon tryptic digestion, and second, such a peptide (matching either fully or in part to the altered amino terminus) must be identified in the MS experiment. By reanalyzing raw spectra from two human and mouse experiments, we found evidence supporting the translation of 19 novel human and seven novel mouse proteins. We used stringent search criteria corresponding to a protein false detection rate of 6.69% for human and 5.95% for mouse (i.e. no more than one match is likely to be a false positive).

The mouse MS data came from experiments on brain and testes. Four of the seven splice variants were represented by an EST derived from brain (or embryo upper head) and one by an EST derived from mouse testicles. Thus, the tissue derivation of the ESTs matches the provenance of the MS datasets, providing strong support for the validity of our search strategy. Such a correlation was not observed in human as the MS data were derived from blood and a leukemic cell line rather than specific tissues. The identification of examples of alternate proteins in these MS datasets therefore provides independent proof of concept, indicating that these alternate isoforms can indeed be translated into proteins.

3.6 Evidence for the selection of alternate start codons

Ribosome profiling is a technique that identifies mRNAs undergoing translation within a cell, by determining what transcripts are bound by ribosomes (Ingolia *et al.*, 2009). In 2012, Lee *et al.* modified the technique to identify translation initiation sites in both a human and mouse cell line (Lee *et al.*, 2012). By comparing their list of identified translation initiation sites to our list of predicted alternate start codons, we were able to find evidence supporting our annotations for 35 instances in human and three in mouse (full list provided in Supplementary Tables S1 and S2). This provides direct evidence that the translational machinery can utilize the alternate start codon predicted by our pipeline.

3.7 Prediction of functional effects of alternative splicing

The selective modification of the N-terminus may allow for the regulation or change of protein function. Given the preference for localization signals to be found within the amino-terminus, one possibility is that this method of splicing can regulate a protein's subcellular localization. To assess this, we compared the predicted localization of each of the canonical and alternate protein isoforms using the Wolf pSort program (Horton *et al.*, 2007). Of the mouse and human proteins >50% were predicted to change their localization due to these alternative splicing events (examples provided in Table 3, full results in Supplementary Tables S1 and S2).

Another key determinant in modification of protein function is the selective removal or addition of functional domains. In order to investigate this possibility, we used the Pfam software suite (v1.3) (Punta *et al.*, 2012) to identify differences in potential functional domains between the canonical and alternate isoforms (full data presented in Supplementary Tables S1 and S2; examples provided in Table 3).

One or more functional domains are predicted to be removed upon splicing in 33% and 38% of the human and mouse genes, respectively. In comparison, only a small number of the events were predicted to add a functional domain (six in mouse and 19 in human). Of the genes predicted to change their functional domains, 57% of the human proteins and 64% of the mouse proteins are also predicted to change their subcellular localizations. By pairing these two functional effects, this method of alternative splicing may produce protein isoforms that display unique functions and are located in separate cellular compartments.

4 CONCLUSION

The aim of this study was to investigate an unusual outcome of alternative splicing that results in the amino terminal exons of a gene being translated in two reading frames. Our pipeline

Table 3. Examples of human genes predicted to undergo some functional change upon alternative splicing

Gene name	Change in localization (Wolf pSort score)	Change in domains
C2	-	Loss of a Sushi domain
EIF3B	Nucleus (20.5) to cytoplasm (21.0)	Loss of an RNA recognition motif
CD40	Plasma membrane (24.0) to mitochondria (14.0)	-
NOTCH4	Extracellular (24.5) to cytoplasm (12)	Loss of EGF and hEGF domains
APOD	Extracellular (31) to cytoplasm (18.5)	Replacement of lipocalin-2 domain with lipocalin domain

predicted a total of 2582 instances of this splicing in the mouse and human genomes, the majority of which have not been previously annotated.

Previous large-scale studies have treated genes as possessing a single start codon, only considering alternate sites of translation initiation when a splice event removes the canonical start codon. A few studies have investigated the potential of alternate start codons, but these have focused on either downstream, in-frame, start codons that produce truncated versions of proteins (Bazykin and Kochetov, 2010), or those that result in a frame shift of the entire transcript (Chung *et al.*, 2007; Xu *et al.*, 2010). Our study is the first to investigate the prevalence of this unusual amino-terminus frame-shifting form of alternative splicing.

Our search strategy identified three of the four published instances of this alternative splicing. Upon further inspection, it was found that Inpp4b was not detected due to low EST coverage. The Unigene database contained only 58 ESTs associated with Inpp4b, none of which encode the experimentally validated splice variant. The 2582 splice forms we have identified are therefore likely to be an underestimate of the true frequency of this splicing. Low EST coverage of genes hides instances of this splicing (as seen in Inpp4b) in addition, variants transcribed at low levels or under restrictive conditions (such as in specific cell subsets or under particular metabolic conditions) may not be represented in databases.

Only one of the 2582 splice variants we have identified, otubain-1, has been previously functionally characterized and was found to have a dramatic effect on the function of the protein (Soares *et al.*, 2004). While the canonical form of the protein activated T-cells, the alternate form was found to promote T-cell anergy (turn off T-cells). The profound functional difference found in the only one of these genes studied, coupled with our predictions for alterations in domain composition and subcellular localization, highlights the potential importance of the novel proteins we have identified.

In this study, we have identified >2000 novel protein isoforms in human and mouse genomes. These result from an unusual outcome of alternative splicing, combining frame-shifting splice variants with alternative start codons. This largely overlooked form of alternative splicing therefore contributes significantly to transcriptome diversity. In each case, the novel proteins differ from their known isoforms by their amino terminus being translated from an alternate reading frame; the remainder of the protein is identical in both isoforms. This can result in a profound change to protein function. The proteins we have identified are involved in a wide range of biological processes. Research into the biological functions of these novel proteins will therefore open new avenues of research into almost every facet of cell biology.

ACKNOWLEDGEMENTS

We would like to thank Dr Iain W. Wilson, Dr Kristine Hardy, Prof. Kiaran Kirk and Mr Joshua Miazek for helpful discussions.

Funding: Australian Capital Territory Cancer Council; Australian Post Graduate Research Award (to L.O.W.W.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Barbosa-Morais,N.L. *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587–1593.

Bazykin,G.A. and Kochetov,A.V. (2010) Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic Acids Res.*, **39**, 567–577.

Bonizzoni,P. *et al.* (2009) Detecting alternative gene structures from spliced ESTs: a computational approach. *J. Comput. Biol.*, **16**, 43–66.

Brett,D. *et al.* (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.

Buratti,E. and Baralle,F.E. (2004) Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell Biol.*, **24**, 10505–10514.

Chung,W.Y. *et al.* (2007) A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput. Biol.*, **3**, e91.

Cobbe,N. *et al.* (2009) The conserved metalloprotease invadolysin localizes to the surface of lipid droplets. *J. Cell Sci.*, **122**, 3414–3423.

Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Denman,R.B. (1993) Using RNAFOLD to predict the activity of small catalytic RNAs. *Biotechniques*, **15**, 1090–1095.

Desiere,F. *et al.* (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.*, **6**, R9.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Eyras,E. *et al.* (2004) ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res.*, **14**, 976–987.

Ferron,M. and Vacher,J. (2006) Characterization of the murine Inpp4b gene and identification of a novel isoform. *Gene*, **376**, 152–161.

Gosling,K.M. *et al.* (2008) Defective T-cell function leading to reduced antibody production in a kleisin-beta mutant mouse. *Immunology*, **125**, 208–217.

Gosling,K.M. *et al.* (2007) A mutation in a chromosome condensin II subunit, kleisin beta, specifically disrupts T cell development. *Proc. Natl Acad. Sci. USA*, **104**, 12445–12450.

Hiller,M. *et al.* (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, **36**, 1255–1257.

Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.

Huang da,W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Ingolia,N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

Kan,Z. *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.

Keller,A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

Kochetov,A.V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays*, **30**, 683–691.

Kozak,M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283–292.

Kozak,M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J.*, **16**, 2482–2492.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lee,S. *et al.* (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl Acad. Sci. USA*, **109**, E2424–E2432.

Lin,M.F. *et al.* (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.

Merkin,J. *et al.* (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593–1599.

Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res*, **40**, D290–D301.

Reiter,L. *et al*. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell Proteomics*, **8**, 2405–2417.

Soares,L. *et al*. (2004) Two isoforms of otubain 1 regulate T cell anergy via GRAIL. *Nat. Immunol*., **5**, 45–54.

Theodoratos,A. *et al*. (2012) Splice variants of the condensin II gene Ncaph2 include alternative reading frame translations of exon 1. *FEBS J*., **279**, 1422–1432.

Volkova,O.A. and Kochetov,A.V. (2010) Interrelations between the nucleotide context of human start AUG codon, N-end amino acids of the encoded protein and initiation of translation. *J. Biomol. Struct. Dyn*., **27**, 611–618.

Xu,H. *et al*. (2010) Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res*., **20**, 445–457.