Dual-coding Regions in Alternatively Spliced Human Genes

Han Liang, *Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA*

Laura F Landweber, *Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA*

By using different exon combinations, alternatively spliced genes may contain dual-coding regions, where more than one reading frame encodes amino acid sequences. These special coding regions generate functionally related but distinct protein products and evolve under unusual selective forces.

Advanced article

Article Contents

- Introduction
- Examples of Alternatively Spliced Genes with a Dual-coding Region
- Systematic Identification of Dual-coding Regions in the Human Genome
- Evolution of Dual-coding Regions in Humans
- Assessing the Functionality of Proteins Encoded by Dual-coding Regions
- Selective Advantages of Dual-coding Regions
- Concluding Remarks

doi: 10.1002/9780470015902.a0020780

Introduction

A stretch of deoxyribonucleic acid (DNA) sequence contains six possible reading frames (three on each strand) and thus may have the potential to encode multiple proteins. In most cases, a coding sequence only contains one valid reading frame, within which every nonoverlapped nucleotide triplet specifies a particular amino acid residue according to the genetic code. In contrast, a dual-coding region is defined as a stretch of DNA that encodes amino acids in overlapping reading frames. In these coding regions, codon positions in one reading frame are shifted relative to the other reading frame. Therefore, although the DNA sequence is exactly the same within the region of overlap, the two encoded peptide sequences are completely different. **See also**: Genetic Code: Introduction

Dual-coding regions are quite common in some bacteriophages and viruses, such as human immunodeficiency virus (HIV), influenza or hepatitis, where the limited genome size is a main constraint on genetic information storage (Normark *et al*., 1983). In these tiny genomes, all the genetic information is not encoded in a sequential manner, so that two genes often share a stretch of coding sequence but employ different reading frames (sometimes even on different strands). In complex eukaryotes, dual-coding regions mainly come from alternative splicing, a major mechanism that can increase genome complexity through generating multiple protein products from a single gene locus. **See also**: Alternative Splicing: Cell-type-specific and Developmental Control

In the human genome, $40 \sim 60\%$ of gene loci are estimated to produce alternatively spliced transcripts, but the vast majority of splicing events would not generate dual-coding regions because constitutively spliced exons (exons shared among different alternative transcripts) usually share the same reading frame. A dual-coding region arises when there is a reading frame switch in a constitutively spliced exon, thereby encoding amino acids in different reading frames on the same strand. **See also**: Messenger RNA in Eukaryotes

Long thought to be the realm of compact viral or streamlined microbial genomes, the description of several dual-coding regions in complex eukaryotes, including humans, has attracted substantial recent attention (Veeramachaneni *et al*., 2004; Liang and Landweber, 2006). The study of such regions is biologically important for decoding the complement of genes in the human genome and other eukaryotes for several reasons. First, dual-coding regions highlight the complexity of genetic information processing. Precise characterization of several well-studied dual-coding genes can improve gene prediction and annotation, unveiling more secondary reading frames when appropriate. Second, genes that contain a dual-coding region generate patches of distinct amino acid sequences in related protein products, providing crucial insights into the biological function of these proteins. Lastly, from an evolutionary point of view, dual-coding regions evolve under unusual selective forces and may display some special evolutionary patterns. **See also**: Gene Feature Identification

In this article, we will first present several well-characterized examples of dual-coding regions in human alternatively spliced genes, to provide an intuitive view of the concept. Then we will discuss how to identify dual-coding regions in the human genome and how these regions evolve at the molecular level. Finally, we will examine the functionality of potential dual-coding regions and the possible selective advantages of this special coding arrangement.

Examples of Alternatively Spliced Genes with a Dual-coding Region

Several human alternatively spliced genes with a dual-coding region have been reported in recent years. Here we discuss three diverse examples that influence unique aspects of biology.

This first example is the *GNAS1* locus, which encodes the stimulatory G-protein subunit α, a key signal transduction element that links receptor–ligand interactions with a variety of cellular responses (Kozasa *et al*., 1988). As shown in **Figure 1a**, the major transcript from this locus contains two completely overlapping reading frames on the same strand, with codon positions 1, 2 and 3 of the first reading frame overlapping codon positions 3, 1 and 2 of the second reading frame, respectively. The first reading frame encodes a 736-residue protein XLαs, while the second reading frame encodes a 322-residue protein ALEX. Binding essays showed that ALEX regulates the intracellular cAMP level through the specific interaction with XLαs, which is essential for normal human phenotypes. Disruption of the interaction between these two proteins can lead to mental retardation and growth deficiency. As an interesting consequence of reading-frame overlap and direct physical interaction, XLαs and ALEX evolve in an oscillatory fashion that constantly balances the rates of amino acid replacements in the two reading frames (Nekrutenko *et al*., 2005). **See also**: G Proteins

A second example is the *XBP1* locus, which regulates genes involved in a process called unfolded protein response (UPR), a complex mechanism that mitigates endoplasmic reticulum (ER) stress (Calfon *et al*., 2002). As shown in **Figure 1b**, the *XBP1* mRNA (messenger ribonucleic acid) contains two overlapping reading frames A and B. Under normal physiological conditions, only frame A is translated into $XBP1^{U}$, an 'unspliced' version of *XBP1*. Upon sensing misfolded proteins in the ER, a 26-bp spacer is removed by IRE1, a multifunctional kinase and endoribonuclease, producing $XBP1^{S}$, the spliced form of *XBP1*. $XBP1^{S}$ has been shown to have a key effect on UPR progression. Meanwhile, evolutionary analyses indicate the reading frame in the unspliced version, $XBP1^{U}$, is also under strong selective constraint, suggesting that it encodes a functional protein product (Nekrutenko and He, 2006).

Finally, the best-studied case of dual coding so far is the *INK4A*/*ARF* locus. This locus resides on the human chromosome 9q21, and is among the most frequent sites of genetic loss in human cancer. Deletion of this locus is associated with a variety of malignancies. As shown in **Figure 1c**, two tumour suppressor proteins $p16^{INK4A}$ and ARF are juxtaposed in this 30-kb locus: each has a unique first exon that splices to a common second exon, but in different reading frames. These two proteins have distinct biological functions: $p16^{INK4A}$ inhibits cdk4/6 activity, leading to RB hypophosphorylation, while ARF mainly inhibits MDM2-mediated degradation of P53. Due to this intriguing structure of overlapping reading frames, an initial supposition was that only $p16^{INK4A}$ was the true tumour suppressor at 9q21 and that loss of ARF was merely coincidental in cancer. However, recent human and mouse genetic data have indicated that both proteins possess significant but distinct *in vivo* tumour suppressor activity (Sharpless, 2005).

Systematic Identification of Dual-coding Regions in the Human Genome

Recently bioinformatic studies have examined human dual-coding regions in a systematic manner. One straightforward way to identify dual-coding regions is to analyse phase usage based on annotated transcript data. In this approach, all available transcript isoforms from the same genetic locus are first aligned against the genomic sequence. Then nucleotide positions in the coding region are annotated as first, second or third codon positions according to the reading frame of

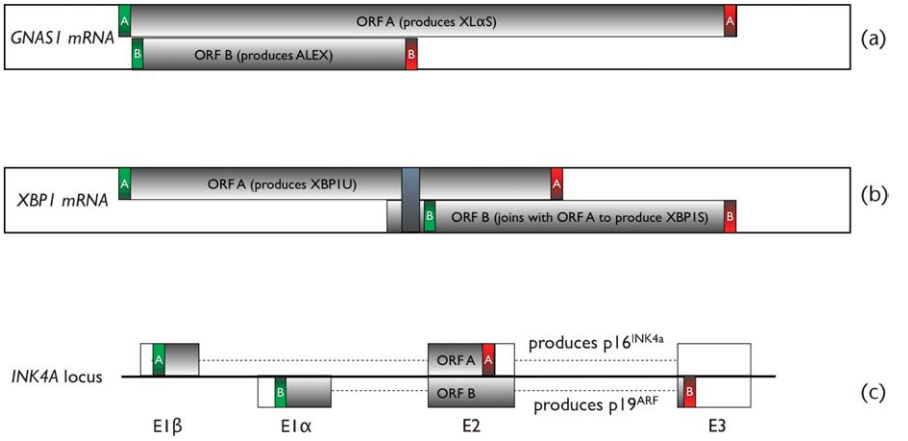


Figure 1 Three known examples of dual-coding genes in mammals. (a) A transcript of the Gnas1 gene contains two reading frames and produces two structurally unrelated proteins, XLαs and ALEX, using different translation start sites. (b) A newly transcribed XBP1 mRNA can only produce protein XBP1U from ORF A. Removal of a 26-bp spacer (dark gray rectangle) joins the beginning of ORF A with ORF B and translation produces a different product, XBP1S. (c) Ink4a generates two splice variants that use different reading frames within exon E2 to produce the proteins p16Ink4a and p19ARF. Reproduced with permission from Chung *et al.* (2007).

each annotated transcript. A dual-coding region can easily be identified as a DNA stretch in which nucleotide positions are annotated as more than one type of codon position. Not surprisingly, two key factors strongly influence the identification of dual-coding regions: the completeness of transcriptome data and the accuracy of reading frame annotation. Based on a set of high-quality and well-annotated transcripts, we originally estimated that approximately 7% of human alternatively spliced genes (179 out of 2585 alternatively spliced gene loci) contain a dual-coding region (Liang and Landweber, 2006). Furthermore, this number is probably an underestimate, because not all spliced versions of every alternatively spliced gene are available. More recently, Tress and colleagues examined alternatively spliced patterns in the ENCODE project comprising close examination of 1% of the human genome. The authors found that 23 out of 214 alternatively spliced loci contain a dual-coding region, suggesting an even higher proportion ($\sim$11%) (Tress *et al.*, 2007). **See also**: Transcriptomics and Proteomics: Integration?

Complementing the above approach, dual-coding regions may also be identified by comparative genomic analysis, which depends less on transcriptome data. Chung and colleagues developed a set of novel statistical tests to search for alternative reading frames that overlap known (or canonical) reading frames. Their predictions are based on two simple observations: first, given a coding sequence and a known reading frame, a second long reading frame (i.e. more than 500 bp) is unlikely to arise purely by chance; and second, functional dual-coding regions tend to be conserved in evolution. Using a set of conservative criteria, the authors identified 40 novel candidate genes with a frame-overlapping coding region. Interestingly, many of these regions show unusual substitution patterns that are consistent with the prediction of dual-coding regions (Chung *et al.*, 2007).

Evolution of Dual-coding Regions in Humans

Given the observation of dual-coding regions in humans, we may ask when the second reading frames arose in evolution. In general, this question cannot be effectively addressed by examining transcript data for orthologous sequences in mammals, because of the limited availability of experimental transcript data for other species and occasional biases. As an alternative, one can trace the evolutionary origin of a dual-coding region by examining the presence of in-frame stop codons in orthologous sequences, since the presence of in-frame stops usually indicates that the reading frame does not encode a protein. Using this approach, we found that most secondary reading frames (the reading frames that arose later in evolution) in human dual-coding regions evolved recently in mammals (Liang and Landweber, 2006). For example, about half of long dual-coding regions (i.e. $>$100 bp) in mice contain one or more stop codons in one of the two reading frames that overlap in humans. This method is intrinsically conservative for inferring nondual-coding regions, since the absence of in-frame stop codons does not guarantee that the reading frame is transcribed. **See also**: Homologous, Orthologous and Paralogous Genes

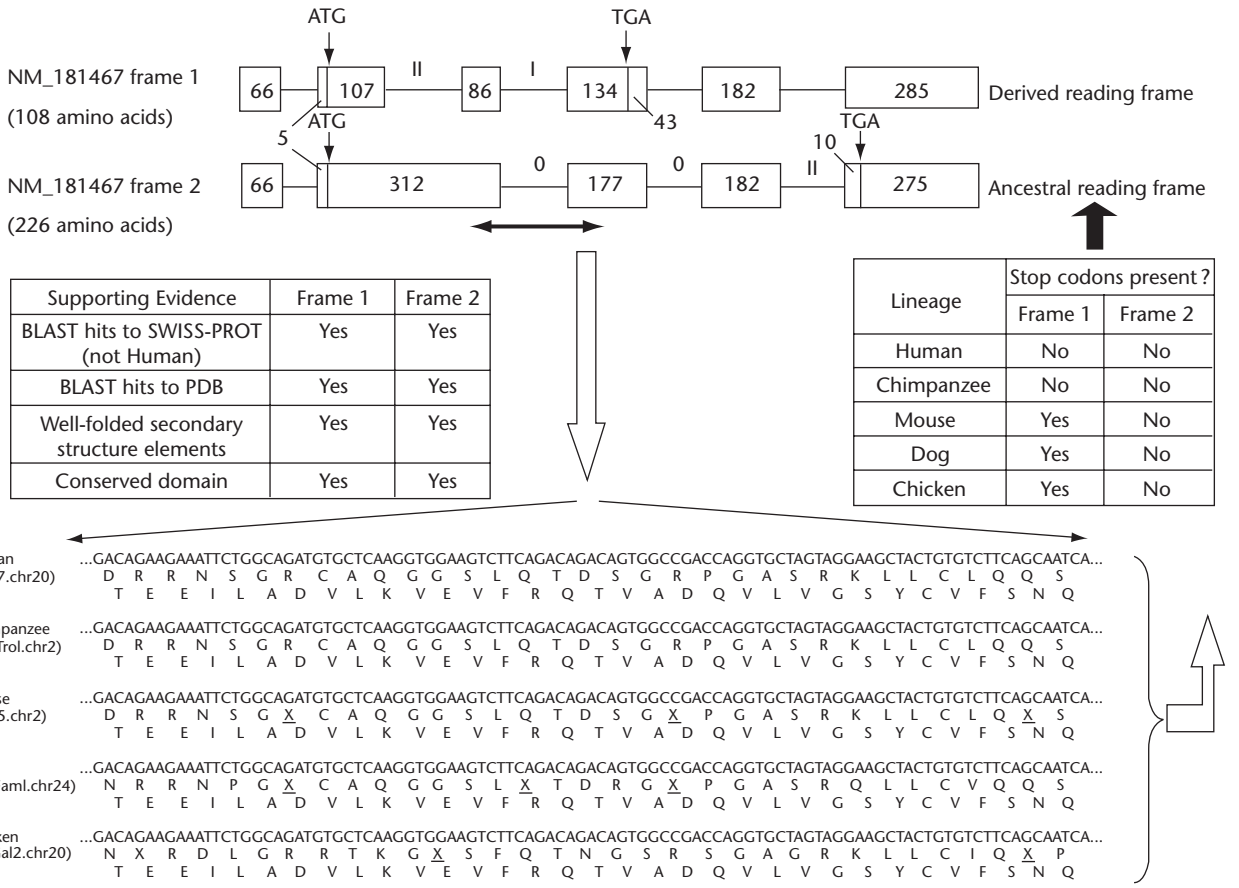
Dual-coding regions in alternatively spliced genes are certainly not unique to humans or mammals, and may also occur in *Drosophila* and *Caenorhabditis* (Liang and Landweber, 2006). Thus, it seems that dual-coding regions are a widespread phenomenon in multicellular eukaryotes, but only a small proportion of these regions are conserved over long periods of evolutionary time.

Based on the presence of in-frame stop codons in orthologous genes, one can further classify the overlapping reading frames in dual-coding regions as 'ancestral' or 'derived', since the former should be maintained (i.e. lack in-frame stop codons) for a longer period of evolution (**Figure 2**). With this classification, we showed that features of ancestral reading frames, such as GC codon-position bias and amino acid composition, are very similar to conventional coding regions in the genome; whereas the derived reading frames show greater variation in these properties (Liang and Landweber, 2006). **See also**: Base Composition Patterns

Assessing the Functionality of Proteins Encoded by Dual-coding Regions

While a relatively large number of candidate dual-coding genes have been identified based on mRNA data, a crucial question remains: what proportion of dual-coding regions encodes functional protein products in both reading frames? Currently, this topic has been under debate (Chung *et al.*, 2007; Tress *et al.*, 2007). One sceptical view is that some mRNAs containing an alternative reading frame might not encode functional proteins and may instead represent biological noise at the transcription level (i.e. incorrectly spliced forms or random products with a very low expression level). While experimental data for protein expression and genetic studies will ultimately answer this question, several types of computational studies may provide evidence for the functionality of a protein product. Here we summarize them as follows:

1. *Bioinformatic analysis of homologous functional features*. If a reading frame arises purely by chance, it is unlikely that the encoded amino acid sequence will share specific functional features with known proteins.
   (a) One can search for the presence of putatively homologous proteins in the database, using as a query the translated peptide sequences from the dual-coding region.
   (b) One can search for well-folded secondary structure elements (at the level of α helices and β sheets), since a random amino acid sequence is almost impossible to fold. **See also**: Protein Secondary Structures: Prediction



Supporting Evidence	Frame 1	Frame 2
BLAST hits to SWISS-PROT (not Human)	Yes	Yes
BLAST hits to PDB	Yes	Yes
Well-folded secondary structure elements	Yes	Yes
Conserved domain	Yes	Yes

Lineage	Stop codons present?	
	Frame 1	Frame 2
Human	No	No
Chimpanzee	No	No
Mouse	Yes	No
Dog	Yes	No
Chicken	Yes	No

Figure 2 Schematic representation of a dual-coding region in the human *ITGB4BP* gene. Exons are represented by boxes and introns by connecting lines. Numbers inside the boxes refer to base pairs. Roman numerals indicate intron phases. The dual-coding region is marked by a black horizontal arrow. Orthologous sequences for this region are shown in other species, and in-frame stop codons are marked by an underlined X. Bioinformatic supporting evidence for the use of both reading frames in humans is shown in the table on the *left*. The table on the *right* summarizes the presence of stop codons in orthologous sequences in two reading frames. White arrows indicate direction of data flow for bioinformatics analysis. NM_181466 and NM_181467 are RefSeq accession numbers. Reproduced from Liang and Landweber (2006) by permission of Cold Spring Harbor Laboratory Press.

(c) One can use sequence-3D-structure comparisons (e.g. threading) to infer distantly related protein homology. **See also**: Protein Tertiary Structures: Prediction from Amino Acid Sequences

(d) One can search for known functional motifs or conserved domains within a peptide sequence.

For example, **Figure 2** shows a bioinformatic analysis of the human *ITGB4BP* gene. The peptide sequences in both reading frames from the dual-coding region contain a well-defined domain and show sequence similarity to known protein sequences. This itself suggests that both reading frames encode functional products.

2. *Evolutionary and statistical analysis*. If both reading frames are functional, they should be maintained by selection.

(a) Sometimes the length of a dual-coding region can provide suggestive evidence, since most nonfunctional reading frames are frequently disrupted by stop codons. Indeed, given the pattern of human codon usage in one reading frame, simulations indicate that fewer than 0.1% of randomly generated second reading frames would extend more than 500 bp (Chung *et al*., 2007).

(b) If both reading frames are constrained by similar levels of purifying selection, then they would display averaged nonsynonymous substitution rates (K_A). Thus a similar K_A between the two reading frames would suggest 'dual functionality' of a dual-coding region (Chung *et al*., 2007). **See also**: Purifying Selection: Action on Silent Sites; Synonymous and Nonsynonymous Rates

3. *mRNA expression data analysis*. Additional expression data, such as expressed sequence tag (EST) abundance or tissue-specific annotation, may provide some confidence in the functionality of a protein product. **See also**: Expressed-sequence Tag (EST)

(a) If both transcript isoforms are expressed at a relatively high level, they are likely to encode functional protein products, since the level of

transcriptional noise should be much lower (Shao *et al*., 2006).

(b) If two transcript isoforms from a dual-coding gene locus are preferentially expressed in different tissues, this tissue specificity also suggests that both proteins may be functional.

Selective Advantages of Dual-coding Regions

Since two reading frames overlap in a dual-coding region, a synonymous substitution change in one reading frame almost always leads to an amino acid change in the other frame. This is a costly way to encode genetic information, because it severely limits the sequence space that either protein product can explore. Thus it is natural to ask, what are the potential selective advantages of a dual-coding region? There are several possible benefits. First, dual coding packs genetic information efficiently (i.e. double information density). However, unlike bacterial or viral genomes, information storage efficiency alone is unlikely to have a significant influence on mammalian genomes, because dual-coding regions only occupy a modest proportion of the human genome, which is replete with noncoding information. There are a few exceptional eukaryotes with many overlapping genes; however, these intriguing cases are compacted microbial genomes or relict nuclei of eukaryotic endosymbionts (Williams *et al*., 2005). Second, a recent theoretical study suggests that when dual-coding regions involve important residues, the number of vulnerable points in the DNA sequence will actually be reduced, making the encoded genetic information more robust against mutations (Peleg *et al*., 2004). Third, dual coding may help to achieve tight co-expression of the overlapped protein products. For example, at the *GNAS1* locus, two partner proteins have to bind to each other to perform their biological functions (Nekrutenko *et al*., 2005). By using nested reading frames, two proteins can be expressed at the same time and at the same place. Lastly, the products encoded in different reading frames may perform separate functions, accomplishing multitasking at the mRNA level.

Concluding Remarks

Dual-coding regions encode genetic information in an unusual manner and represent one of the most fascinating and unexpected aspects of eukaryotic genomes. Although much progress has been made in recent years, more effort is needed to characterize dual-coding regions at the functional level. The answers to the following questions will significantly advance our understanding of dual-coding regions. (1) Which dual-coding genes really produce functional protein products in both reading frames? (2) What is the functional and evolutionary connection between the two related but distinct protein products? Do they work in a coordinated or complementary manner? (3) What are the molecular mechanisms that regulate expression of dual-coding genes? In the next few years, we can hope to see in-depth examination of hundreds of cases of overlapping genes, helping us to understand this feature of our genomes.

References

Calfon M, Zeng H, Urano F *et al*. (2002) IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature* **415**: 92–96.

Chung WY, Wadhawan S, Szklarczyk R, Pond SK and Nekrutenko A (2007) A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Computational Biology* **3**: e91.

Kozasa T, Itoh H, Tsukamoto T and Kaziro Y (1988) Isolation and characterization of the human Gs alpha gene. *Proceedings of the National Academy of Sciences of the USA* **85**: 2081–2085.

Liang H and Landweber LF (2006) A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Research* **16**: 190–196.

Nekrutenko A and He J (2006) Functionality of unspliced XBP1 is required to explain evolution of overlapping reading frames. *Trends in Genetics* **22**: 645–648.

Nekrutenko A, Wadhawan S, Goetting-Minesky P and Makova KD (2005) Oscillating evolution of a mammalian locus with overlapping reading frames: an XLαs/ALEX relay. *PLoS Genetics* **1**: e18.

Normark S, Bergstrom S, Edlund T *et al*. (1983) Overlapping genes. *Annual Review of Genetics* **17**: 499–525.

Peleg O, Kirzhner V, Trifonov E and Bolshoy A (2004) Overlapping messages and survivability. *Journal of Molecular Evolution* **59**: 520–527.

Shao X, Shepelev V and Fedorov A (2006) Bioinformatic analysis of exon repetition, exon scrambling and trans-splicing in humans. *Bioinformatics (Oxford, England)* **22**: 692–698.

Sharpless NE (2005) INK4a/ARF: a multifunctional tumor suppressor locus. *Mutation Research* **576**: 22–38.

Tress ML, Martelli PL, Frankish A *et al*. (2007) The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the USA* **104**: 5495–5500.

Veeramachaneni V, Makalowski W, Galdzicki M, Sood R and Makalowska I (2004) Mammalian overlapping genes: the comparative perspective. *Genome Research* **14**: 280–286.

Williams BA, Slamovits CH, Patron NJ, Fast NM and Keeling PJ (2005) A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proceedings of the National Academy of Sciences of the USA* **102**: 10936–10941.

Further Reading

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

Freson K, Jaeken J, Van Helvoirt M *et al*. (2003) Functional polymorphisms in the paternally expressed XLαs and its cofactor ALEX decrease their mutual interaction and enhance

receptor-mediated cAMP formation. *Human Molecular Genetics* **12**: 1121–1130.

Huynen MA, Konings DA and Hogeweg P (1993) Multiple coding and the evolutionary properties of RNA secondary structure. *Journal of Theoretical Biology* **165**: 251–267.

Klemke M, Kehlenbach RH and Huttner WB (2001) Two overlapping reading frames in a single exon encode interacting proteins – a novel way of gene usage. *The EMBO Journal* **20**: 3849–3860.

Kozak M (2001) Extensively overlapping reading frames in a second mammalian gene. *EMBO Reports* **2**: 768–769.

Modrek B and Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics* **34**: 177–180.

Quelle DE, Zindy F, Ashmun RA and Sherr CJ (1995) Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83**: 993–1000.

Robertson U, Navik JA, Walden KKO and Honegger HW (2007) The bursicon gene in mosquitoes: An unusual example of mRNA trans-splicing. *Genetics* **176**: 1351–1353.

Scorilas A, Kyriakopoulou L, Yousef GM *et al.* (2001) Molecular cloning, physical mapping, and expression analysis of a novel gene, BCL2L12, encoding a proline-rich protein with a highly conserved BH2 domain of the Bcl-2 family. *Genomics* **72**: 217–221.

Szklarczyk R, Heringa J, Pond SK and Nekrutenko A (2007) Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. *Proceedings of the National Academy of Sciences of the USA* **104**: 12807–12812.

Zhao FQ, Zheng Y, Dong B and Oka T (2004) Cloning, genomic organization, expression, and effect on β-casein promoter activity of a novel isoform of the mouse Oct-1 transcription factor. *Gene* **326**: 175–187.