Cell Research (2010) 20:445-457.

www.nature.com/cr

ORIGINAL ARTICLE

# Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts

Heng Xu[1, *], Ping Wang[1, *], Yujie Fu[1, *], Yufang Zheng[3], Quan Tang[1], Lizhen Si[1], Jin You[1], Zhenguo Zhang[1], Yufei Zhu[1], Li Zhou[1], Zejun Wei[1], Bin Lin[1], Landian Hu[1, 2], Xiangyin Kong[1, 2]

*[1]The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine, 225 South Chong Qing Road, Shanghai 200025, China; [2]State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiaotong University, 197 Rui Jin Road II, Shanghai 200025, China; [3]Department of Physiology and Biophysics, School of Life Sciences, Fudan University, Shanghai 200433, China*

**A single mammalian transcript normally encodes one protein, but the transcript of *GNAS* (G-protein α-subunit) contains two reading frames and produces two structurally unrelated proteins, XLαs and ALEX. No other confirmed *GNAS*-like dual-coding transcripts have been reported to date, even though many such candidate genes have been predicted by bioinformatics analysis. In this study, we constructed a series of vectors to test how two protein products were translated from a single transcript *in vitro*. The length of the ORF (open reading frame), position of the first AUG and the Kozak motif were found to be important factors. These factors, as well as 55-bp NMD (nonsense-mediated mRNA decay) rule, were used in a bioinformatics search for candidate dual-coding transcripts. A total of 1307, 750 and 474 two-ORF-containing transcripts were found in human, mouse and rat, respectively, of which 170, 89 and 70, respectively, were found to be potential dual-coding transcripts. Most transcripts showed low conservation among species. Interestingly, dual-coding transcripts were significantly enriched for transcripts from the zinc-finger protein family, which are usually DNA-binding proteins involved in regulation of the transcription process.**

***Keywords*:** dual-coding transcripts, open reading frame, Kozak motif, first AUG, nonsense-mediated mRNA decay
*Cell Research* (2010) **20**:445-457. doi: 10.1038/cr.2010.25; published online 16 February 2010

## Introduction

According to the one-transcript-one-peptide rule, a single mammalian transcript normally encodes one protein. However, a transcript of the *GNAS* (G-protein α-subunit) gene contains two reading frames and produces two structurally unrelated proteins called XLαs and ALEX [1]. The open reading frame (ORF) of XLαs contains a second ORF for ALEX, which is frame shifted +1 compared to the start codon of XLαs. ALEX and XLαs are structurally distinct and can interact with each other [1]. Although ~7% genes contain dual-coding regions due to alternative splicing in different transcripts [2], no other *GNAS*-like (two proteins are produced with only one transcript) mammalian dual-coding transcript has been found to date.

Recently, a series of such candidate genes have been predicted using bioinformatics methods [3, 4]. Chung *et al.* [3] identified 149 two-ORF-containing transcripts conserved in human and mouse using 500 bp as the lower bound (that is, no ORF shorter than 500 bp), and suggested that dual coding is virtually impossible to happen by chance. When 150 bp was used as the threshold, the number of potential two-ORF transcripts dramatically increased: 9163 human-mouse-rat ortholog triplets were found in the report by Ribrioux *et al.* [4]. However, translation of these candidate transcripts has not been confirmed experimentally.

*These three authors contributed equally to this work.
Correspondence: Xiangyin Kong[a], Landian Hu[b]
[a]Tel: +86-21-63852639; Fax: +86-21-64678976
E-mail: xykong@sibs.ac.cn
[b]E-mail: ldhu@sibs.ac.cn
Received 21 July 2009; revised 29 September 2009; accepted 9 November 2009; published online 16 February 2010

Translation of cellular mRNA is thought to be impacted by two important factors: the position of first AUG codon and the length of ORF. In the most stringent test of the first-AUG rule, the first codon was shown to be the exclusive initiation site, even when the second AUG was positioned just a few bases downstream from, and in the same optimal context as, the first AUG [5]. The ORF length is also important for protein translation. The longest ORF is always considered to be the coding sequence (CDS) [6]. It seems that the longest-ORF rule could be more powerful than the first-AUG rule, as many upstream AUG-containing mRNAs have been identified [7].

The Kozak motif is located in the region around the initiator codon and has an optimal sequence of GCCRCCAUGG (R represents purine). The Kozak motif greatly impacts protein translation efficiency, and the two highly conserved nucleotides at positions −3 and +4 are especially important (the A of the AUG codon is designated as +1) [8]. If the AUG codon is flanked by optimum nucleotides at those two positions, the rest of the consensus sequence contributes only marginally. Therefore, an AUG codon can be classified as a strong or weak start site based on the nucleotides in positions −3 and +4. When either position contains the conserved nucleotide, the starting site is considered to be strong; sites with non-conserved nucleotides in these positions are considered weak starting sites [9, 10]. It has been demonstrated that nearly all ribosomes will initiate at the optimal AUGs [9]. Bioinformatics analysis has shown that 95% of main AUGs have an optimal context [11]. A weak first AUG start site could induce leaky scanning, which means 40S ribosomes can bypass this AUG. As a result, translation initiation occurs from both first and second AUGs, and two proteins in overlapping ORFs (including both frame-in and frame-shift patterns) from a single mRNA are produced [12]. It was reported in many cases that two in-frame proteins were translated from a single mRNA that has a weak Kozak motif around the first AUG [9] or a non-AUG start codon [13, 14]. However, transcripts of *GNAS* are the only experimentally confirmed example of overlapping frame-shifted leaky scanning.

Nonsense-mediated mRNA decay (NMD) is a conserved cellular surveillance mechanism that detects and eliminates aberrant mRNAs whose expression would result in truncated proteins that can be deleterious to the organism [15]. Most NMD targets follow the 55-bp rule, where nonsense codons that are more than 55 bp upstream of the last intron trigger NMD [16-18]. The translation of dual-coding mRNA could also be impacted by this pathway that screens out aberrant mRNAs.

In our work, three different factors (the position of the AUG, the ORF length and the type of Kozak motif) were used to construct eight subtypes of two-ORF-containing vectors. Through *in vitro* expression experiments, we found that two distinct proteins could be simultaneously translated in four out of the eight subtypes. To find other potential dual-coding transcripts, we screened human and mouse reference sequences that followed the dual-coding conditions we determined, as well as the NMD rule. The candidate dual-coding transcripts have been listed. However, most of them are not conserved between human and mouse, suggesting that there could be a high birth rate of new dual-coding mRNA, accompanied by a high death rate, similar to the case in microRNA (miRNA) [19]. Interestingly, transcripts from the zinc-finger family tend to be dual coding.

## Results

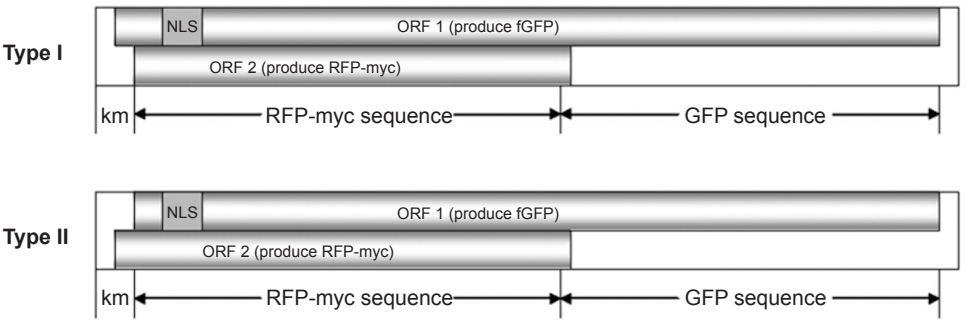
*Construction of artificial dual-coding transcript vectors*

We cloned the combinational DNA CDS containing RFP, the myc-tag and GFP into the plasmid pcDNA3.1 (+). The GFP ORF is frame-shifted −1 relative to the ORF of RFP-myc (Figure 1A and 1B). Fortunately, the −1 frame-shifted RFP ORF sequence contains neither translational stop codon nor start codon AUG (ATG in DNA sequence), so that the peptide product from this frame fused to GFP forms a fusion protein (fGFP). Additionally, this peptide includes a nuclear localization signal (NLS) with the amino-acid sequence RPRVRDRGRGR that can introduce fGFP into the nucleus based on protein subcellular localization prediction (http://www.rostlab.org/cgi/var/nair/resonline.pl and http://bioapps.rit.albany.edu/pTARGET/). Two ATGs with their Kozak motifs (shown as K1-ATG-K2-ATG in Figure 1A) were added to the beginning of the sequence, resulting in two different frames: (I) the frame of fGFP starts with the first AUG, while RFP-myc starts with the second; (II) the frame of RFP-myc starts with the first AUG, while fGFP starts with the second. Therefore, two ORFs could potentially be translated (an ~250 amino-acid product from RFP-myc and an ~500 amino-acid product from fGFP) (Figure 1B). Different types of Kozak motifs were added to the two AUGs. Alignments of all AUG start codon sequence contexts from human transcripts reveal that GCCACCAUGG is the most common set of nucleotides for each position, while UAUUUUAUGC is the least (Supplementary information, Figure S1). Findings by Kozak indicate that protein translation efficiency is affected by the Kozak motif. We used GCCACCAUGG as a strong Kozak motif and UAUUUUAUGC as a weak one. Using this convention, we named a type I frame with two strong Kozak motifs "ss", while sequences with a strong Kozak motif around the first AUG, followed by a weak

A

K1-ATG——K2-ATG GA ATT CTG GCC TCC TCC GAG AAC GTC ATC ACC GAG TTC ATG CGC
TTC AAG GTG CGC ATG GAG GGC ACC GTG AA**C GGC CAC GAG TTC GAG ATC GAG GGC GAG**
**GGC GA**G GGC CGC CCC TAC GAG GGC CAC AAC ACC GTG AAG CTG AAG GTG ACC AAG GGC
GGC CCC CTG CCC TTC GCC TGG GAC ATC CTG TCC CCC CAG TTC CAG TAC GGC TCC AAG
GTG TAC GTG AAG CAC CCC GCC GAC ATC CCC GAC TAC AAG AAG CTG TCC TTC CCC GAG
GGC TTC AAG TGG GAG CGC GTG ATG AAC TTC GAG GAC GGC GGC GTG GCG ACC GTG ACC
CAG GAC TCC TCC CTG CAG GAC GGC TGC TTC ATC TAC AAG GTG AAG TTC ATC GGC GTG
AAC TTC CCC TCC GAC GGC CCC GTG ATG CAG AAG AAG ACC ATG GGC TGG GAG GCC TCC
ACC GAG CGC CTG TAC CCC CGC GAC GGC GTG CTG AAG GGC GAG ACC CAC AAG GCC CTG
AAG CTG AAG GAC GGC GGC CAC TAC CTG GTG GAG TTC AAG TCC ATC TAC ATG GCC AAG
AAG CCC GTG CAG CTG CCC GGC TAC TAC TAC GTG GAC GCC AAG CTG GAC ATC ACC TCC
CAC AAC GAG GAC TAC ACC ATC GTG GAG CAG TAC GAG CGC ACC GAG GGC CGC CAC CAC
CTG TTC CTG TGC GGC CGC GAG GAG CAG AAG CTG ATC TCA GAG GAG GAC CTG CTC GAG
TGG **TGA** GCA AGG GCG AGG AGC TGT TCA CCG GGG TGG TGC CCA TCC TGG TCG AGC TGG
ACG GCG ACG TAA ACG GCC ACA AGT TCA GCG TGT CCG GCG AGG GCG AGG GCG ATG CCA
CCT ACG GCA AGC TGA CCC TGA AGT TCA TCT GCA CCA CCG GCA AGC TGC CCG TGC CCT
GGC CCA CCC TCG TGA CCA CCC TGA CCT ACG GCG TGC AGT GCT TCA GCC GCT ACC CCG
ACC ACA TGA AGC AGC ACG ACT TCT TCA AGT CCG CCA TGC CCG AAG GCT ACG TCC AGG
AGC GCA CCA TCT TCT TCA AGG ACG ACG GCA ACT ACA AGA CCC GCG CCG AGG TGA AGT
TCG AGG GCG ACA CCC TGG TGA ACC GCA TCG AGC TGA AGG GCA TCG ACT TCA AGG AGG
ACG GCA ACA TCC TGG GGC ACA AGC TGG AGT ACA ACT ACA ACA GCC ACA ACG TCT ATA
TCA TGG CCG ACA AGC AGA AGA ACG GCA TCA AGG TGA ACT TCA AGA TCC GCC ACA ACA
TCG AGG ACG GCA GCG TGC AGC TCG CCG ACC ACT ACC AGC AGA ACA CCC CCA TCG GCG
ACG GCC CCG TGC TGC TGC CCG ACA ACC ACT ACC TGA GCA CCC AGT CCG CCC TGA GCA
AAG ACC CCA ACG AGA AGC GCG ATC ACA TGG TCC TGC TGG AGT TCG TGA CCG CCG CCG
GGA TCA CTC TCG GCA TGG ACG AGC TGT ACA AG

B




C

| Subtype | **Sequence of k1-ATG-K2-ATG** |
|---|---|
| **ss** | gccaccATGggccaccATGg |
| **ww** | tattttATGctattttATGc |
| **ws** | tattttATGcgccaccATGg |
| **sw** | gccaccATGgtattttATGc |
| **ss2** | gccaccATGgcgccaccATGgc |
| **ww2** | tattttATGcgtattttATGcg |
| **ws2** | tattttATGcggccaccATGgc |
| **sw2** | gccaccATGgctattttATGcg |

**Figure 1** The artificial dual-coding transcript constructs. **(A)** Sequence of RFP-myc-GFP: two Kozak motifs with start codon ATGs are shown with border outlines (K1-ATG for the first ATG with its Kozak motif and K2-ATG for the second, sequence of K1-ATG-K2-ATG were shown in **C** to form different subtypes), followed by the RFP coding (uppercase) and myc-tag (gray background) sequences. The stop codon of RFP-myc is shown in bold with an outlined border. The GFP sequence (underlined) is added, making an ORF that is frame shift −1 related to the ORF of RFP-myc. The NLS signal encoded by the other frame of RFP is shown in bold; **(B)** sketch of two types of dual-coding models ("km" denotes a sequence with two ATGs and their Kozak motifs); **(C)** sequences of eight subtypes of Kozak motif combinations that replace the content with the outlined border in **A**.

Kozak motif around the second AUG, were labeled "sw". The type II frame with two strong Kozak motifs was labeled "ss2", etc. In total, eight subtypes of vectors were constructed ("ss", "ss2", "ww", "ww2", "sw", "sw2", "ws" and "ws2") (listed in Figure 1C).

*Dual-coding expression of different transcripts*

In order to test the protein expression of the two-ORF-containing transcripts, confocal fluorescence images were taken 30 h after transient transfection of HEK 293T cells with the vectors described above (recombinant protein expression of this cell line is rather high compared to other cell lines). Green and red fluorescence were used to detect the expression of fGFP and RFP-myc, respectively. DAPI was also used to locate the cellular nucleus. In all cells transfected with type I vectors (fGFP is in frame with the first ATG), green fluorescence that co-localized with DAPI could be detected, indicating that fGFP was translated and introduced into the nucleus. Furthermore, red fluorescence that was distributed diffusely throughout the cells was detected in types "ww" and "ws" (Figure 2A). These results demonstrate that RFP was translated from the second AUG through leaky scanning. Western blot results confirmed the expression of both fGFP (~56 kDa) and RFP-myc (~30 kDa) in cells transfected with both "ww" and "ws" vectors. fGFP yield was decreased relative to "ss" and "sw" subtypes (Figure 2B). Therefore, in type I vectors, protein translation was impacted mainly by the first-AUG rule and the Kozak motif.

Results were different in cells transfected with type II vectors, where RFP-myc is in frame with the first AUG. Red fluorescence was absent in the "ws2" vector-transfected cells, and these results were confirmed by western blot. Thus, translation initiation might bypass the first weak Kozak motif-containing AUG and start with the strong second one to read through the longer ORF. However, green fluorescence was absent in "sw2" vector-transfected cells, indicating that the longest ORF with a weak Kozak motif could not overcome the first AUG with a strong Kozak motif. Both green and red fluorescent proteins were simultaneously detected when the cells were transfected with "ss2" and "ww2" vectors (Figure 2C). Western blot results indicated that very low levels of fGFP were expressed in "ss2"-transfected cells compared with "ws2" and "ww2" (Figure 2D), while there was no significant difference among these subtypes in RNA level (data not shown). Therefore, four out of the eight subtypes of two-ORF-containing transcripts could be dual coding: "ww", "ws", "ss2" and "ww2".

*Screening for dual-coding transcript candidates in human and mouse with bioinformatics method*

Over 40 000 transcripts (including different splicing variants from a single gene) were found in both human and mouse species in the NCBI database, about half of which were annotated as "predicted". We first removed those "predicted" transcripts from the pool, leaving 26 009 transcripts for human and 21 792 transcripts for mouse in the database. We next screened this database for transcripts with two frame-shifted overlapping ORFs. Unlike previous work that aligned human, mouse and rat transcripts before searching for the conserved two-

**Table 1** Classification of two-ORF-containing transcripts in human and mouse

| Subtype | Total | | Non-NMD | | Conserved | | Conserved (non-NMD) | |
|---|---|---|---|---|---|---|---|---|
| | Human | Mouse | Human | Mouse | Human (with mouse) | Mouse (with human) | Human (with mouse) | Mouse (with human) |
| ss | 704 | 419 | 385 | 237 | 131 | 127 | 60 | 62 |
| ss2* | 50 | 23 | 18 | 9 | 3 | 4 | 0 | 0 |
| sw | 398 | 227 | 236 | 121 | 88 | 67 | 48 | 31 |
| sw2 | 21 | 3 | 10 | 3 | 0 | 0 | 0 | 0 |
| ws* | 77 | 39 | 40 | 23 | 10 | 9 | 6 | 3 |
| ws2 | 14 | 12 | 6 | 4 | 1 | 2 | 1 | 2 |
| ww* | 39 | 24 | 22 | 15 | 5 | 3 | 3 | 2 |
| ww2* | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sum | 1 307 | 750 | 717 | 412 | 238 | 212 | 118 | 100 |
| Dual coding | 170 | 89 | 80 | 47 | 18 | 16 | 9 | 5 |
| Dual coding/sum% | 13 | 11.9 | 11.2 | 11.4 | 7.56 | 7.55 | 7.63 | 5 |

Notes: Groups of total, non-NMD and conserved two-ORF-containing (ORF > 500 bp) transcripts were identified and classified into the eight subtypes by screening the referenced transcripts that were not annotated as "predicted" for human and mouse. Potential dual-coding subtypes are marked with "*".

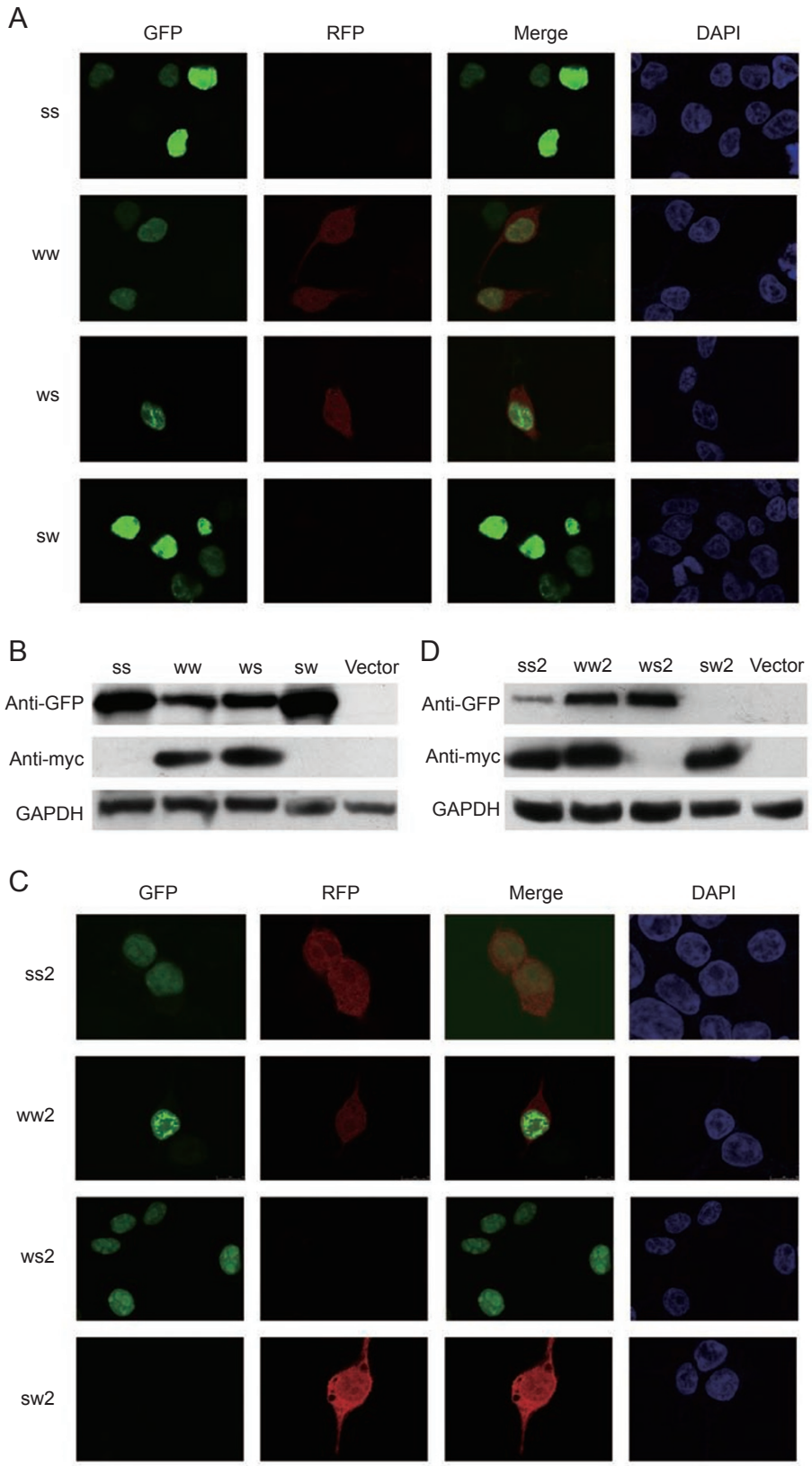



**Figure 2** Dual-coding expression of fGFP and RFP-myc **(A)** HEK 293T cells were transfected with type I plasmids. Thirty hours after transfection, cells were analyzed by confocal microscopy to detect green and red fluorescence. Cell nuclei were counterstained with DAPI (blue). **(B)** Cells transfected with the same plasmids were harvested and analyzed by western blot. **(C, D)** The same experiments were performed with type II plasmids.

ORF-containing transcripts [2, 3], we first searched for two-ORF-containing transcripts in each of these species individually. As in Chung's work, we searched for ORFs longer than 500 bp [2]. When there are multiple ORFs in one transcript, we considered only the two longest. Using these criteria, we found 1 307 (out of 26 009) human transcripts to be potential two-ORF-containing transcripts. Within these 1 307 two-ORF-containing transcripts, 238 were conserved between human and mouse (i.e., peptides encoded by both ORFs of these transcripts show conservation in different species). According to the Kozak rules [4, 7-9, 11], we classified Kozak motifs containing either R in the position of −3 or G in the position of +4 as strong and the rest as weak. Therefore, all the two-ORF-containing transcripts were classified into the eight subtypes described above ("ss", "ss2", "ww", "ww2", "sw", "sw2", "ws" and "ws2"). Only 13% (170 out of 1 307) belonged to the experimentally determined dual-coding subtypes ("ss2", "ww", "ww2", "ws"), and the percentage decreased to 7.6% (18 out of 238) in the conserved transcripts (Table 1). Similar results were obtained in mouse transcripts: 750 out of 21 792 transcripts contained two overlapping frame-shifted ORFs, while 212 were conserved with human (Table 1). The number is not the same as in the human case probably because the number of splicing variants from one gene could be different between the two species (e.g., four reported transcript variants from SULF1 in human compared with only one in mouse). Overall, 11.9% (89 out of 750) belonged to dual-coding subtypes in those two-ORF-containing mouse transcripts, and the percentage decreased to 7.5% (16 out of 212) in the conserved transcripts.

NMD is an RNA surveillance mechanism that rapidly degrades mRNAs harboring premature termination codons (PTCs). The 55-bp rule is a model that applies to PTC-containing mRNAs [15], where dual-coding transcripts that are potential NMD targets are not translated into protein products. To avoid false positives, transcripts with PTC-containing ORFs were excluded from our data. A total of 717 two-ORF-containing transcripts remained in the human data set (out of 1 307), of which 80 were potentially dual coding and 9 were conserved with mouse. In the mouse data set, 412 two-ORF-containing transcripts (out of 750) and 47 dual-coding transcripts remained after filtering by the NMD rule (Table 1); only 5 were conserved with human. *SMAD6* and *ZNF551* (*Zfp551* in mouse) were the only non-NMD target dual-coding transcripts conserved in both human and mouse (Table 2). We searched the SwissProt protein sequence database with web tool "blastp" in NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi), in order to find whether the proteins translated from the predicted ORFs (predicted proteins) of *SMAD6* and *ZNF551* could be conserved with reported proteins. We found that a fragment of the predicted protein of *ZNF551* was conserved with the referenced protein product of *ZNF781* (Figure 3A).

The conserved dual-coding transcripts account for a very small proportion of all two-ORF-containing

**Table 2** Conserved two-ORF-containing transcripts of human and mouse

| Dual coding | | | | Corresponding transcripts | | | |
|---|---|---|---|---|---|---|---|
| | RNA Gi no. | Gene symbol | Subtype | | RNA Gi no. | Gene symbol | Subtype |
| | 93141025 | *ADAM11** | ws | | 160707884 | *Adam11* | ss |
| | 118766338 | *NKX2-1* | ws | | 40254604 | *Nkx2-1* | ss |
| | 15042948 | *PRDM12** | ws | | 182765474 | *Prdm12* | ws2 |
| | 91206395 | *TAS1R3* | ws | | 118130760 | *Tas1r3* | ss |
| Human | 119372316 | *XIRP2* | ws | In mouse | 145046245 | *Xirp2* | ss |
| | 37537685 | *ZSCAN21* | ws | | 113374181 | *Zscan21* | ss |
| | 92859870 | **SMAD6** | ww | | 31560672 | **Smad6** | ww |
| | 190570171 | *ZNF180* | ww | | 113866021 | *Zfp180* | sw |
| | 190886459 | **ZNF551** | ww | | 162287257 | **Zfp551** | ws |
| | 34304110 | *Foxe1** | ws | | 21618324 | *FOXE1* | sw |
| | 169234959 | *Zscan20* | ws | | 148596976 | *ZSCAN20* | sw |
| Mouse | 118130551 | *Polr2a** | ww | In human | 169791021 | *POLR2A* | sw |
| | 31560672 | **Smad6** | ww | | 92859870 | **SMAD6** | ww |
| | 162287257 | **Zfp551** | ws | | 190886459 | **ZNF551** | ww |

Notes: Gi no., gene symbol and subtype of the conserved dual-coding potential transcripts are listed. Dual-coding transcripts in both human and mouse is shown in bold. CDS of the transcripts marked with "*" start from the first nucleotide, –3 position of these transcripts are the corresponding position in the DNA sequence after mapping the transcripts to genome.

A

```
Identities = 57/161 (35%), Positives = 77/161 (47%), Gaps = 0/161 (0%)

ZNF551  17  FTTRDVTLEECVMSVVNVGKPLATNLTSLNTREFTLEKGLMNVASAGNPLDKALAFFDTR  76
            F+ R+ TL    ++V NV K        LNTRE  L + LMNV +    L K     + +
ZNF781 161  FSIREFTLVRSPLNVRNVAKHSIIAQHLLNTRELILMRNLMNVRNVKRLLGKVHILLNIK 220

ZNF551  77  EFTLEKGLISAVSVGNPLDKSSISFDIEEFTLEKCLISAVIVGNLLAANRNSFNTREFTV 136
            EF L +  +S  +VG       I  DI EFTL K  ++ V    LL   +  FNTREFT+
ZNF781 221  EFILVRNHMSVSNVGRLSLVFLILIDIREFTLVKNPMNVKNVVELLTIVQLLFNTREFTL 280

ZNF551 137  EKDLMNAENVGNPLDNSLTSFDTAAFTLVIGLMSAVNVRNP 177
             + LMN  +VG  L      F+     L+  LM+  N R P
ZNF781 281  VRRLMNISSVGRFLSPVQHLFNIREHILMKNLMNVSNARRP 321
```

B

```
Identities = 83/180 (46%), Positives = 119/180 (66%), Gaps = 4/180 (2%)

DUSP13-p  141  PKIHGAVQASPYQPPTLASLQRLLWVRQAATLNHIDEVWPSLFLGDAYAARDKSKLIQLG 200
               P++ G  +A+P   P++  L+ LL   +++  + +DEVWP+LF+GDA  A ++ +L +LG
DUSP13-a    7  PELGGEDKATPC--PSILELEELLRAGKSSC-SRVDEVWPNLFIGDAATANNRFELWKLG  63

DUSP13-p  201  ITHVVNAAAGKFQVDTGAKFYRGMSLEYYGIEADDNPFFDLSVYFLPVARYIRAALSVPQ 260
               ITHV+NAA        G  FY G S+ Y G+ A D P FD+S YF   A +I  AL+ P
DUSP13-a   64  ITHVLNAAHKGLYCQGGPDFY-GSSVSYLGVPAHDLPDFDISAYFSSAADFIHRALNTPG 122

DUSP13-p  261  GRVLVHCAMGVSRSATLVLAFLMICENMTLVEAIQTVQAHRNICPNSGFLRQLQVLDNRL 320
                +VLVHC +GVSRSATLVLA+LM+ + ++L +A+ TV+ HR + PN GFL QL  LD +L
DUSP13-a  123  AKVLVHCVVGVSRSATLVLAYLMLHQRLSLRQAVITVRQHRWVFPNRGFLHQLCRLDQQL 182
```

C

```
Identities = 20/76 (26%), Positives = 29/76 (38%), Gaps = 11/76 (14%)

FGF19   45  RHSQPRSASRSPPSLPHPHRRSCAESPREV---PCGAGVWWSTYGSWPASGWPWPGAPSP  101
            +HSQP+     P        +SC   P +    PC +        + P    PWP    P
CARD6 953  KHSQPKPFHSVPSQPKSSQTKSCQSQPSQTKPSPCKS--------TQPKPSQPWPPQSKP 1004

FGF19  102  SRTRGPTCTTAGATPS 117
            S+ R P   ++   PS
CARD6 1005 SQPRPPQPKSSSTNPS 1020
```

**Figure 3** Conservation analysis of the predicted protein. The products translated from the frame-shifted ORFs of *ZNF551*, *DUSP13* and *FGF19* are partly conserved with *ZNF781* **(A)**, *DUSP13* **(B)** and *CARD6* **(C)**, respectively. DUSP13-p and DUSP13-a stand for predicted and annotated products of *DUSP13*, respectively.

transcripts. What accounts for the small number of conserved two-ORF-containing transcripts? One possibility is that dual-coding transcripts have both high birth and death rates. To test this hypothesis, we searched the rat mRNA database, which is more closely related to mouse, evolutionarily, than human (we did not test whether the transcripts are NMD targets, as there is no available exon information for rat mRNA). Of the 15 163 transcripts without "predicted" annotation, 474 transcripts are two-ORF-containing and 70 transcripts are potential dual-coding transcripts. More two-ORF-containing and potential dual-coding transcripts were found to be conserved between mouse and rat than those between human and rat (Table 3). Therefore, it is possible that dual-coding transcripts change rapidly during evolution. However, further investigations are needed.

**Table 3** Two-ORF-containing transcripts conserved with human and mouse in rat

| Subtype | Total | Conserved with human | Conserved with mouse |
|---|---|---|---|
| ss | 262 | 67 | 89 |
| ss2* | 31 | 4 | 2 |
| sw | 114 | 34 | 57 |
| sw2 | 22 | 2 | 0 |
| ws* | 23 | 5 | 7 |
| ws2 | 6 | 0 | 0 |
| ww* | 14 | 4 | 6 |
| ww2* | 2 | 0 | 0 |
| Sum | 474 | 116 | 161 |
| Dual coding | 70 | 13 | 15 |
| Dual coding/sum% | 14.8 | | |

Note: Potential dual-coding subtypes are marked with "*".

**Table 4** Two-ORF-containing transcripts with a longer unannotated ORF

| Subtype | Human unannotated > annotated | | Mouse unannotated > annotated | |
|---|---|---|---|---|
| | Total | Non-NMD | Total | Non-NMD |
| ss | 4 | 2 | 1 | 0 |
| ss2* | 12 | 9 | 6 | 4 |
| sw | 0 | 0 | 0 | 0 |
| sw2 | 10 | 9 | 1 | 1 |
| ws* | 2 | 1 | 0 | 0 |
| ws2 | 1 | 1 | 1 | 1 |
| ww* | 1 | 1 | 0 | 0 |
| ww2* | 0 | 0 | 0 | 0 |
| Sum | 30 | 23 | 9 | 6 |

Notes: Shown here is the number of the two-ORF-containing transcripts in human and mouse where the unannotated ORF is longer than the annotated one. Potential dual-coding subtypes are marked with "*".

We marked the annotated ORFs in NCBI (the annotated CDS region in GenBank) for each two-ORF-containing transcript, and defined the other ORFs as unannotated. Interestingly, although the longer ORF of the two was frequently the one that was annotated, there were some exceptions. The unannotated ORF is longer in 30 out of 1 307 two-ORF-containing transcripts in human and in 9 out of 750 in mouse (Table 4 and Supplementary information, Table S1). Furthermore, of those 30 human transcripts, 10 belonged to the "sw2" subtype, where only the short annotated ORF could be translated as shown in our experimental system. In addition, 12 transcripts belonged to the "ss2" subtype, which would have relatively low yield of the long ORF compared with the short ORF. In other words, 73.3% (22 out of 30) of these transcripts where the longer ORF was unannotated could still translate the annotated one. The percentage for mouse transcripts is 77.7% (18 out of 23). Also, we found the conservation of the predicted proteins of these genes with reported proteins (Figure 3B and 3C and data not shown). Interestingly, the predicted protein of *DUSP13* was conserved with the protein translated from its annotated ORF (Figure 3B). Both of these two products contain conserved DSPc (dual specificity phosphatases) domain. Additionally, the predicted protein likely contained a transmembrane domain (http://www.cbs.dtu.dk/services/TMHMM/). We speculate that these two proteins translated from one transcript could perform similar functions in different locations.

After classifying all the transcripts by function annotation analysis (http://david.abcc.ncifcrf.gov/), 26 009 transcripts in human were identified, with 16 504 identified as genes by the web tool. The decrease is mainly due to two effects: (1) some genes have two or more alterna-

tive splicing transcripts; (2) some genes could not be recognized by the web tool, especially zinc-finger proteins named as *ZFP* plus a number in mouse (e.g., *ZFP551*). As a result, the number of two-ORF-containing transcripts decreased from 1 307 to 956, and 138 out of 956 transcripts were potential dual-coding types. Only 67 transcripts were left after applying the non-NMD rule. We noticed that those two-ORF-containing and dual-coding transcripts were enhanced for genes associated with transcription (proteins involved in the transfer of genetic information from DNA to mRNA by DNA-directed RNA polymerase), transcriptional regulation (proteins involved in the regulation of the transcription process), DNA binding (factor which binds to DNA, typically to pack or modify the DNA, or to regulate gene expression) and zinc-finger proteins (proteins with at least one zinc finger). We found 10.2%, 10%, 9.7% and 8.7% of genes out of the 16 504 genes that belonged to these 4 groups, respectively (some genes belong to 2 or more of these 4 groups). In 956 two-ORF-containing transcripts, 138 had dual-coding potential and 67 were non-NMD dual-coding genes. These four groups of genes were significantly overrepresented in the two-ORF-containing transcripts ($P < 0.0001$) (Table 5). Similar results were also seen in mouse (Table 5). We noticed that zinc-finger proteins accounted for a large proportion of proteins in the transcription, transcription regulation and DNA-binding groups. This effect can be explained by the fact that zinc-finger proteins are thought to be involved in DNA binding, transcriptional processes, and protein-protein interactions [20, 21].

Many genes that encode zinc-finger proteins were named as *ZNF* or *ZFP* follow by a number, such as *ZNF551* or *Zfp551* (we simply call these genes *ZNFs*

**Table 5** Functions of annotated groups of genes in human and mouse

| Group | Total | | Two-ORF containing | | Dual coding | | Non-NMD dual coding | |
|---|---|---|---|---|---|---|---|---|
| | Human | Mouse | Human | Mouse | Human | Mouse | Human | Mouse |
| | (16 504) | (17 913) | (956) | (625) | (138) | (72) | (67) | (39) |
| Transcription | 1 690 | 1 224 | 224 | 71 | 34 | 11 | 26 | 9 |
| | (10.2%) | (6.8%) | (23.4%) | (11.4%) | (24.6%) | (15.3%) | (38.8%) | (23.1%) |
| | | | *P* = 6.52E-37 | *P* = 1.27E-5 | *P* = 3.25E-08 | *P* = 4.69E-3 | *P* = 1.88E-14 | *P* = 6.16E-5 |
| Transcription regulation | 1 656 | 1 201 | 222 | 70 | 33 | 10 | 26 | 8 |
| | (10%) | (6.7%) | (23.2%) | (11.2%) | (23.9%) | (13.9%) | (38.8%) | (20.5%) |
| | | | *P* = 1.74E-37 | *P* = 1.23E-5 | *P* = 7.58E-08 | *P* = 0.0152 | *P* = 7.10E-15 | *P* = 5.88E-4 |
| DNA binding | 1 604 | 1 230 | 219 | 73 | 31 | 11 | 25 | 8 |
| | (9.7%) | (6.9%) | (22.9%) | (11.7%) | (22.5%) | (15.3%) | (37.3%) | (20.5%) |
| | | | *P* = 1.92E-38 | *P* = 3.70E-06 | *P* = 5.47E-07 | *P* = 4.95E-3 | *P* = 3.70E-14 | *P* = 7.81E-4 |
| Zinc finger | 1 447 | 1 498 | 200 | 130 | 24 | 11 | 21 | 9 |
| | (8.7%) | (8.4%) | (20.9%) | (20.8%) | (17.4%) | (15.3%) | (31.3%) | (23.1%) |
| | | | *P* = 7.57E-36 | *P* = 3.48E-27 | *P* = 3.79E-4 | *P* = 0.0347 | *P* = 8.58E-11 | *P* = 9.33E-4 |

Notes: Two-ORF-containing and potential dual-coding genes are significantly enhanced for genes involved in transcription, transcription regulation, DNA binding and zinc fingers in both human and mouse.

**Table 6** Classification of ZNF/ZFP transcripts in human, mouse and rat

| (*ZNF* and *ZFP*)/total | Human | Mouse | Rat |
|---|---|---|---|
| Total | 703/26 009 | 326/21 792 | 152/15 163 |
| Two-ORF containing | 167/1 307 | 64/750 | 28/474 |
| | *P* = 4.44E-91 | *P* = 7.52E-48 | *P* = 6.35E-23 |
| Dual coding | 23/170 | 4/89 | 5/70 |
| | *P* = 1.05E-17 | *P* = 0.021 | *P* = 3.87E-7 |
| Non-NMD dual coding | 21/80 | 4/47 | |
| | *P* = 1.59E-37 | *P* = 8.22E-5 | |

Notes: ZNF/ZFP transcripts are significantly enhanced for groups of two-ORF-containing and dual-coding transcripts.

or *ZFPs*). We noticed that some *ZNFs* and *ZFPs*, especially examples from mouse, were not recognized by the functional annotation analysis web tool (http://david.abcc.ncifcrf.gov/). Therefore, we manually counted the *ZNF* and *ZFP* transcripts in all groups of transcripts. Our results showed that *ZNF/ZFP* was involved in 167 out of the 1 307 two-ORF-containing transcripts, in 160 of 717 non-NMD two-ORF-containing transcripts and in 21 of 80 non-NMD dual-coding potential transcripts. Significant zinc-finger overrepresentation was also seen in human transcripts, where 703 *ZNF/ZFP* were found out of 26 009 total transcripts ($P < 0.0001$). Similar results were seen in mouse and rat (Table 6). The *ZNF/ZFP* coding transcripts represent a large percentage of the dual-coding transcripts in all species we analyzed, indicating that dual-coding transcripts tend to fall into this family.

## Discussion

In this study, we showed that under certain conditions, two protein products can be translated from a single RNA. We constructed a series of artificial vectors to explore how three different factors impact translation: ORF length, the position of the first AUG and the Kozak motif. In these artificial vectors, the fGFP ORF is about two times longer than the RFP-myc ORF. Two classes of vectors were defined based on the position of the first AUG.

The ORF length was very important in determining the yield of protein products. In our study, fGFP, which was two times longer than RFP-myc, was expressed in seven out of eight subtypes of vectors. It is possible that protein translation will be regulated differently when the lengths of the two ORFs are nearly the same than when one ORF is several times longer than the other. The transcript of *PRDM12* is an interesting example of the first situation. The annotated CDS of human *PRDM12* covers 1 104 bp while the second ORF covers 723 bp. In contrast, in mouse, the annotated CDS is 1 098 bp, shorter than the 1 149 bp predicted second ORF. The transcripts of *PRDM12* (*Prdm12*) belong to the "ws" class in human and "ws2" in mouse based on our classification. Although conserved in these two species, the annotated CDS would not be translated in mouse according to our criteria. The explanation of this paradox might be that two ORFs with nearly the same length will be translated simultaneously. In the case of *LRP1B*, which belongs to the "ww" subtype, the annotated ORF covers 13 911 bp, which is nearly 25 times longer than the second longest ORF in another frame. It has been suspected that the short ORF could not be translated at all. Therefore, we infer that the relative length of the two ORFs has a great impact on dual-coding translation. However, how ORF lengths affect dual-coding translation requires further study.

According to the first-AUG rule, translation initiates at the AUG codon nearest the 5′ end of the mRNA [9, 22, 23]. With an increasing number of upstream AUGs being discovered [24, 25], two ancillary mechanisms (re-initiation and context-dependent leaky scanning) for escaping the first-AUG rule have been demonstrated. In our study, the translation process was greatly affected by the first AUG and in most ORFs the first AUG was translated (seven out of eight subtypes, Figure 2).

The number of two-ORF-containing and potential dual-coding transcripts obtained in our bioinformatics screening is quite low. One reason is that we used 500 bp as the lower bound, as in Chung *et al.* [3]. It has been speculated that more dual-coding transcripts would be found if a lower threshold was used. As described by Ribrioux *et al.* [4], 9 163 two-ORF-containing transcripts were found to be conserved in human, mouse and rat when 150 bp was used as the lower bound, while only 64 were left when 600 bp was used as the threshold.

In our artificial vectors, the two AUGs were designed to be very close to each other (only 7 or 8 bp apart). However, most of the two-ORF-containing transcripts we screened did not display this feature. In the case of *GNAS*, the AUG of ALEX is 32 nucleotides (187 bp through updated data [26]) downstream of the AUG of XLαs [1]. It is still unknown whether the distance between the AUGs affects the dual-coding translation.

Although we used the optimal sequence GCCRCCAUGG as the strong Kozak motif in our test, this motif is not very common as only ~85% of all transcripts contain R in the −3 position and ~50% contain G at +4 (Supplementary information, Figure S1). Thus, even fewer transcripts contain both conserved nucleotides. According to a previous *in vitro* study, a G in the +4 position with a U in the −3 position can lead to leaky scanning, while A in −3 with C in +4 did not [9], indicating that G in the +4 position has a relatively weaker initiation strength than R in −3. Thus, some two-ORF-containing transcripts that belong to the "ss" subtype could also be potential dual-coding "ws" subtype, such as *GTF2F1* (also known as *TFIIF*; its protein product is thought to be involved in both early and late transcriptional stages as a transcription factor [27]). The annotated CDS of the *GTF2F1* transcript contains a C in position −3 and G in +4, while the shorter predicted ORF contains As in both sites. Additionally, although it was thought that if an AUG codon is flanked by R in the position of −3 and G in +4, the rest of the consensus sequence contributes only marginally [9], some exceptions were reported. *CD40* is an example that contains a G in both −3 and +4 positions, neverthe-

less, the C/T SNP in the −1 site of *CD40* was reported to be associated with Grave's disease, probably because the protein expression levels with CT and TT genotypes were lower when compared with the levels in CC genotypes [28, 29]. Based on our criteria, the transcript of *GNAS* belongs to the "ss" subtype, and the short ORF encoding ALEX should not be translated. Kozak has indicated that the first AUG is in a good, but not perfect, context (it has an A in position −3, but lacks a G in position +4), and therefore a small fraction of ribosomes might reach the AUG of ALEX by leaky scanning [30]. We speculate that some "ss", "sw", "sw2" and "ws2" subtypes of two-ORF-containing transcripts could also yield two protein products in specific conditions.

All three features described above had a great impact on dual coding of two-ORF-containing transcripts. Our study found that changes in these features not only determined whether multiple proteins were translated from a single mRNA but also affected the relative yield of the total protein.

NMD is an important pathway that detects and eliminates aberrant mRNAs. The 3′-most intron within a valid transcript must be no more than 55 bp downstream of the true stop codon. When this 55-bp rule was first proposed and tested, just 2 out of 1 500 genes from many species were found to violate the rule [31]. Similar results were reported in the work by Scofield *et al.* [32]. Therefore, we thought that it was important to screen whether both ORFs of the two-ORF-containing transcripts were NMD targets. Fewer than half of the dual-coding potential and more than half of the two-ORF-containing transcripts remained after this filter was applied. Of these transcripts, only *SMAD6* and *ZNF551* were conserved between human and mouse. Currently, we are investigating whether the two ORFs of *SMAD6* could be translated simultaneously. However, the transcripts of *GNAS* were not in the final list as XLαs ORFs broke the 55-bp rule in human. Therefore, it is possible that in actuality, there may be more conserved dual-coding transcripts because some transcripts could disobey the 55-bp NMD rule.

miRNAs represent a dynamic class of genes because they have a high rate of birth and death during evolution [19]. We suspect that the evolution of dual-coding transcripts could undergo a similar process. Mutations generated in the evolution process could cause protein production from different frames of the subsistent transcripts. But, probably most of these products were neutral or redundant, or even harmful. As a result, they appear and disappear rapidly, and only functional ones were kept and evolved at a slower rate. For example, in *GNAS*, XLαs and ALEX evolve in an oscillating fashion, constantly balancing the rates of amino-acid replacement [33]. According to our study, the mRNAs of zinc-finger proteins could be the most active family of dual-coding transcripts in terms of birth and death rates. This effect may be associated with the complicated regulation pathways in different species since zinc-finger proteins are always involved in DNA binding and protein-protein interactions. However, this area requires further study.

## Materials and Methods

*Construction of expression plasmids for mammalian cells*

DNA sequences coding GFP and RFP were cloned from the plasmids of pEGFP-C2 and pDsRed, respectively, and linked to the myc tag sequence with *Not*I and *Xho*I to make RFP and myc in the same reading frame while the GFP ORF was −1 frame shifted. The full sequence of RFP-myc-GFP was cloned into the *Eco*RI and *Xba*I sites of pcDNA3.1 (+), which contains the CMV promoter and BGH poly (A). DNA sequences containing different AUG contexts were synthesized and cloned into the *Hin*dIII and *Eco*RI sites of pcDNA3.1 (+) to generate eight subtypes of vectors ("ss", "ww", "sw", "ws", "ss2", "ww2", "sw2" and "ws2") (Figure 1C).

*Cell culture, transfection, confocal microscopy and western analysis*

HEK 293T cells were cultured in DMEM supplemented with 10% fetal calf serum (Invitrogen). Cells were transfected with lipofectin 2000 (Invitrogen) following the protocol provided by Invitrogen. Thirty hours after transfection, cells were fixed with 4% paraformaldehyde in 1× PBS for 10 min then washed with 1× PBS three times. DAPI (Beyotime) with a final concentration of 1 μg/ml was used to stain the cellular nucleus. Cells were then imaged using confocal microscopy where blue, green and red fluorescence were excited at appropriate excitation wavelength. Transfected cells were also harvested and used in western blots with antibodies against GFP (Clonetech), the myc tag (Santa Cruz) and GAPDH (Beyotime).

*Searching for dual-coding transcripts by screening mRNA reference sequences in NCBI*

We screened human, mouse and rat RNA sequences from NCBI RefSeq release 31 and found 26 009, 21 792 and 15 163 protein coding transcripts, respectively, that were not annotated as "predicted" dual-coding transcripts. Of these mRNA sequences, we first found that all ORFs start with ATG (non-AUG start codon was not considered in our work) and stop with TAA, TAG and TGA. Then, we selected dual-coding candidates according to the following criteria: (1) The two longest ORFs were both longer than 500 bp and frame shifted in a single transcript; (2) These two ORFs were overlapped. To screen the transcripts with those two criteria, we screened all the start codon ATGs and their in-frame stop codons in each transcript and recorded their positions. When the A of ATG was in the same frame of the first nucleotide, the ORFs with this kind of ATG were labeled as frame 0, others as frames 1 and 2, respectively. The longest two ORFs with different frames were picked out and submitted to the 500-bp filter. By comparing the position of the start and stop codons of the two ORFs, we could select the transcripts with two overlapped ORFs. All transcripts that met these criteria are listed in Supplementary

information, Table S1. Next, we examined the nucleotides at positions −3 and +4 of each ATG Kozak motif and labeled the Kozak motifs with an R at −3 or G at +4 as "s" (strong) and the others as "w" (weak). The transcripts where the longest ORF correlates with the first AUG were labeled as type "I"; when the longest ORF correlates with the second AUG the sequence is labeled as type "II". In human and mouse, we downloaded the GenBank database from NCBI, and picked out the exon and CDS information of each transcript. Then, we calculated the length between stop codons of two ORFs and last exon-exon junction, respectively, and examined whether these ORFs met the 55-bp NMD rule. Using ortholog information from Mouse Genome Informatics (http://www.informatics.jax.org), we grouped mRNAs from the three species with the same gene symbol. We searched the conserved two-ORF-containing transcripts by aligning the two mRNAs in different species and checking whether both mRNAs in the different species were dual coding in the same frame and similar region. The annotated CDS was marked by comparing the information between NCBI and our ORF data. The program (dual-coding.pl) written with Perl language was provided as Supplementary information, Data S1.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (30530450 and 30871356), the National High Technology Research and Development Program of China (2006AA02Z330 and 2006AA02A301), the National Basic Research Program of China (2007CB512202 and 2004CB518603) and the Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX1-YW-R-74). We are grateful to Xuanfu Yi and Yang Shu (Institute of Health Sciences, Shanghai Institutes for Biological Sciences) for their technical assistance.

## References

1 Klemke M, Kehlenbach RH, Huttner WB. Two overlapping reading frames in a single exon encode interacting proteins–a novel way of gene usage. *EMBO J* 2001; **20**:3849-3860.
2 Liang H, Landweber LF. A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Res* 2006; **16**:190-196.
3 Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput Biol* 2007; **3**:e91.
4 Ribrioux S, Brungger A, Baumgarten B, Seuwen K, John MR. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics* 2008; **9**:122.
5 Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* 2002; **299**:1-34.
6 Furuno M, Kasukawa T, Saito R, *et al.* CDS annotation in full-length cDNA sequence. *Genome Res* 2003; **13**:1478-1487.
7 Vilela C, McCarthy JE. Regulation of fungal gene expression via short open reading frames in the mRNA 5′untranslated region. *Mol Microbiol* 2003; **49**:859-867.
8 Kozak M. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol Cell Biol* 1989; **9**:5073-5080.
9 Kozak M. Initiation of translation in prokaryotes and eukaryotes. *Gene* 1999; **234**:187-208.
10 Kozak M. 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* 1997; **16**:2482-2492.
11 Suzuki Y, Ishihara D, Sasaki M, *et al.* Statistical analysis of the 5′ untranslated region of human mRNA using "oligo-capped" cDNA libraries. *Genomics* 2000; **64**:286-297.
12 Kozak M. An analysis of vertebrate mRNA sequences: intimations of translational control. *J Cell Biol* 1991; **115**:887-903.
13 Packham G, Brimmell M, Cleveland JL. Mammalian cells express two differently localized Bag-1 isoforms generated by alternative translation initiation. *Biochem J* 1997; **328** (Pt 3):807-813.
14 Takayama S, Krajewski S, Krajewska M, *et al.* Expression and location of Hsp70/Hsc-binding anti-apoptotic protein BAG-1 and its variants in normal tissues and tumor cell lines. *Cancer Res* 1998; **58**:3116-3131.
15 Shyu AB, Wilkinson MF, van Hoof A. Messenger RNA regulation: to translate or to degrade. EMBO J 2008; 27:471-481.
16 Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* 2004; **5**:89-99.
17 Rehwinkel J, Raes J, Izaurralde E. Nonsense-mediated mRNA decay: target genes and functional diversification of effectors. *Trends Biochem Sci* 2006; **31**:639-646.
18 Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol* 1998; **18**:5272-5283.
19 Lu J, Shen Y, Wu Q, *et al.* The birth and death of microRNA genes in Drosophila. *Nat Genet* 2008; **40**:351-355.
20 Hartwig A. Zinc finger proteins as potential targets for toxic metal ions: differential effects on structure and function. *Antioxid Redox Signal* 2001; **3**:625-634.
21 Mackay JP, Crossley M. Zinc fingers are sticking together. *Trends Biochem Sci* 1998; **23**:1-4.
22 Kozak M. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* 1978; **15**:1109-1123.
23 Kozak M. The scanning model for translation: an update. *J Cell Biol* 1989; 108:229-241.
24 Iacono M, Mignone F, Pesole G. uAUG and uORFs in human and rodent 5′untranslated mRNAs. *Gene* 2005; **349**:97-105.
25 Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* 2000; **20**:8635-8642.
26 Abramowitz J, Grenet D, Birnbaumer M, Torres HN, Birnbaumer L. XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. *Proc Natl Acad Sci USA* 2004; **101**:8366-8371.
27 Cojocaru M, Jeronimo C, Forget D, *et al.* Genomic location of the human RNA polymerase II general machinery: evidence for a role of TFIIF and Rpb7 at both early and late stages of transcription. *Biochem J* 2008; **409**:139-147.
28 Tomer Y, Concepcion E, Greenberg DA. A C/T single-nucle-

otide polymorphism in the region of the CD40 gene is associated with Graves' disease. *Thyroid* 2002; **12**:1129-1135.
29 Jacobson EM, Concepcion E, Oashi T, Tomer Y. A Graves' disease-associated Kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: a case for translational pathophysiology. *Endocrinology* 2005; **146**:2684-2691.
30 Kozak M. Extensively overlapping reading frames in a second mammalian gene. *EMBO Rep* 2001; **2**:768-769.
31 Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* 1998; **23**:198-199.
32 Scofield DG, Hong X, Lynch M. Position of the final intron in full-length transcripts: determined by NMD? *Mol Biol Evol* 2007; **24**:896-899.
33 Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD. Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. *PLoS Genet* 2005; **1**:e18.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)