4220–4234 *Nucleic Acids Research, 2011, Vol. 39, No. 10* *Published online 25 January 2011*
doi:10.1093/nar/gkr007

Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences

Ivaylo P. Ivanov[1,2,*], Andrew E. Firth[1,3], Audrey M. Michel[1,4], John F. Atkins[1,2] and Pavel V. Baranov[4,*]

[1]BioSciences Institute, University College Cork, Cork, Ireland, [2]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112-5330, USA, [3]Department of Pathology, University of Cambridge, Cambridge, UK and [4]Biochemistry Department, University College Cork, Cork, Ireland

Received October 6, 2010; Revised December 16, 2010; Accepted January 3, 2011

ABSTRACT

In eukaryotes, it is generally assumed that translation initiation occurs at the AUG codon closest to the messenger RNA 5′ cap. However, in certain cases, initiation can occur at codons differing from AUG by a single nucleotide, especially the codons CUG, UUG, GUG, ACG, AUA and AUU. While non-AUG initiation has been experimentally verified for a handful of human genes, the full extent to which this phenomenon is utilized—both for increased coding capacity and potentially also for novel regulatory mechanisms—remains unclear. To address this issue, and hence to improve the quality of existing coding sequence annotations, we developed a methodology based on phylogenetic analysis of predicted 5′ untranslated regions from orthologous genes. We use evolutionary signatures of protein-coding sequences as an indicator of translation initiation upstream of annotated coding sequences. Our search identified novel conserved potential non-AUG-initiated N-terminal extensions in 42 human genes including VANGL2, FGFR1, KCNN4, TRPV6, HDGF, CITED2, EIF4G3 and NTF3, and also affirmed the conservation of known non-AUG-initiated extensions in 17 other genes. In several instances, we have been able to obtain independent experimental evidence of the expression of non-AUG-initiated products from the previously published literature and ribosome profiling data.

INTRODUCTION

Translation initiation is the only step during protein biosynthesis where an incoming aminoacyl tRNA is bound directly in the ribosomal P-site (1,2). There is a greater flexibility for mismatches in the codon:anticodon duplex in the P-site (3,4) in comparison with the A-site where, in contrast, proper geometry of messenger RNA (mRNA):tRNA interactions is strictly monitored by the decoding center for the first two positions of a codon (5). As a result, initiator Met-tRNA$_i$ can be incorporated into the ribosome at a wider range of codons than in the case of elongator Met-tRNA whose incorporation is strictly limited to AUG codons. Such a flexible usage of initiation codons by Met-tRNA$_i$ requires additional control mechanisms for discrimination between potential start sites.

These control mechanisms differ between eukaryotes and bacteria. In fact, the initiation step represents the biggest mechanistic difference in the process of protein biosynthesis between organisms from different domains of life. In bacteria, with a single notable exception of psyllid endosymbiont *Carsonella* (6), the small subunit of the ribosome binds the mRNA a short distance upstream of the initiation codon through interaction between the 3′-end of the 16S rRNA and a complementary sequence, termed Shine–Dalgarno (7). In contrast, in eukaryotes, the scanning model for the initiation of translation postulates that the small ribosomal subunit, in complex with initiation factors and Met-tRNA$_i$, binds first to the 5′ cap. Then it scans 5′–3′ until it finds a suitable initiator codon (8). Exceptions include internal ribosome entry sites, shunting and reinitiation mechanisms, see references (9) and (10) for invigorating discussion of contrasting viewpoints. During scanning, the efficiency of initiation is dependent on the nucleotide context surrounding the initiator codon with the optimal context being known as the Kozak motif (11).

The initiation context—comprising the 6 nt before and the 1 nt immediately following a potential initiation codon—has significant influence on the recognition of that codon as an initiation site, through as yet

*To whom correspondence should be addressed. Tel: +353 21 4901313; Fax: +353 21 4904259; Email: iivanov@genetics.utah.edu
Correspondence may also be addressed to Pavel V. Baranov. Tel: +353 21 4904212; Fax: +353 21 4904259; Email: p.baranov@ucc.ie

poorly understood mechanisms requiring the activity of eIF1 (12,13). The optimal context in mammals is GCCRCC***AUG***G (14) with the identity of the underlined nucleotides in the −3 and +4 positions (relative to the 'A' of the AUG) being the most important. An 'A' in the −3 position is preferred over a 'G' and a purine in that position is more important than a 'G' in the +4 position (15). The identity of the nucleotides in the other positions plays an important role only when there is no purine in position −3 and no 'G' in position +4 (15).

Although relatively rare, initiation of translation of eukaryotic mRNAs can occur at non-AUG codons that differ from AUG at a single nucleotide position. Initiation on the nine possible such codons occurs with a varying degree of efficiency that is partially taxon dependent. For example, in mammals and plants CUG appears to be the most efficient non-AUG initiation codon, while AAG and AGG are the least efficient (16). For initiation at non-AUG codons, the presence of a good Kozak context is crucial (17–19). The rules for the most favorable context are believed to be identical to those for an AUG codon. In addition, a strong RNA secondary structure starting ∼15 nt downstream of the initiation site significantly increases the efficiency of initiation at non-AUG codons (20).

In mammals, at least 38 mRNAs from 23 genes have been reported to have or else have annotated non-AUG initiation codons in GenBank (21). In most cases, the non-AUG initiation provides an alternative longer isoform in addition to an isoform resulting from initiation at a standard AUG codon downstream *via* a process termed 'leaky scanning' (22). Where alternative isoforms are produced as a result of leaky scanning, the longer isoform frequently contains a signal for subcellular localization that is absent in the shorter form (23–26). In some cases of non-AUG-initiated genes, the principles of leaky scanning ensure that the non-AUG initiation codon is the sole initiation site of the main CDS even when downstream in-frame AUG codons exist. This happens when the region between the non-AUG initiation codon and the first available downstream in-frame AUG codon is populated by one or even many out-of-frame AUG codons, especially if some of them occur in a good initiation context as, for example, in the case of EIF4G2 (27). Conversely, another peculiar feature of eukaryotic initiation can be exploited to suppress initiation on intervening out-of-frame AUG codon(s). Translation of open reading frames (ORFs) shorter than approximately 35 codons results in only partial dissociation of terminating ribosomes from the mRNA, with the 40S subunit able to resume scanning downstream (11,28). Scanning following translation of such an ORF, however, precludes efficient reinitiation until initiation factors have been reloaded onto the 40S subunit. As a result AUG codons, even in perfect (Kozak) context, are inefficiently recognized if they are present ∼100–200 nt downstream of an efficiently translated short upstream ORF (29). Translation of an ORF longer than approximately 35 codons usually precludes resumption of scanning downstream. An additional layer of complexity comes from the fact that the efficiency of 'reloading' the scanning 40S subunit following translation of a short upstream ORF (uORF) is under regulatory control through the phosphorylation status of Ser-51 of eIF2α (29). Leaky scanning, combined with the peculiar features of reinitiation following translation of short uORFs, can result in diverse and complex scenarios for predicting which codons, and under what conditions, are recognized as initiation sites for translation on eukaryotic mRNAs (22).

Earlier *in vitro* work with rabbit reticulocyte lysates showed that the concentration of Mg^{2+} negatively regulates the fidelity of translation initiation (18). A recent study found that non-AUG-initiated uORFs regulate expression of ornithine decarboxylase homologs in many eukaryotic organisms (30). In one of the mammalian homologs, and perhaps in all other homologs that have this particular regulatory feature, polyamines induce initiation on an AUU codon. It is noteworthy that polyamines are essentially organic polycations and, as such, their biophysical role overlaps with that of Mg^{2+} ions which, at least *in vitro*, show an identical effect on the fidelity of initiation. A recently developed technique for profiling the ribosomal density on mRNAs revealed the presence of translating ribosomes on >200 non-AUG-initiated uORFs in yeast. Moreover, initiation on these uORFs was shown to be upregulated under amino acid starvation conditions (31). Another recent study showed that variations in the intracellular levels of eIF1 lead to significant differences in the stringency of start codon selection (32). These results suggest that the fidelity of initiation is variable under physiological conditions and strongly supports the idea that non-AUG initiation might perform regulatory roles in eukaryotic cells. Therefore, the utilization of non-AUG initiation in mammalian mRNAs merits further exploration, as the regulated alternative usage of start codons provides an opportunity for global alterations of a proteome in a coordinated manner. An earlier finding that sequences in the 5′-untranslated regions (UTRs) are highly conserved and that the level of conservation globally increases toward the UTR/CDS boundaries (33) is suggestive that this conservation could be due, in part, to the 3′-ends of a portion of 5′-UTRs encoding N-terminal extensions to the annotated AUG-initiated proteins. The protein coding potential of 5′-UTRs and modes of its utilization have been discussed in detail elsewhere (34).

Here we performed a systematic analysis of the 5′-UTRs of human GenBank RefSeq mRNAs to investigate the extent of non-AUG initiation in humans. Our methodology is based on the analysis of codon substitution rates in pairwise alignments of human and mice orthologous sequences that allows us to detect 5′-UTR fragments evolving under the constraints of protein coding evolution. If a segment of a 5′-UTR evolves under such constraints, it is likely that the corresponding protein sequence is produced and is positively functional. Candidates selected by this criterion were subjected to further computational analysis using multiple alignments of sequences from other vertebrate species followed by rigorous manual evaluation.

MATERIALS AND METHODS

A pipeline for identification of candidates for non-AUG translation initiation

A schematic outline of the pipeline for potential 5 extension of CDS (P5EC) detection is illustrated in Figure 2. A total of 46 500 human mRNA sequences were downloaded from the NCBI RefSeq database (35) on 12 February 2009 (release 33) using the Refseq ftp site: ftp://ftp.ncbi.nih.gov/refseq/release/. We chose RefSeq annotations to minimize the effect from erroneous and inconsistent misannotations as well as to avoid redundancy in our dataset. All sequences that have in-frame stop codons within 50 codons upstream of annotated AUG initiator codons were discarded. For the remaining mRNA sequences, the region surrounding the annotated AUG codon was translated from the nearest upstream in-frame stop codon to the 100th codon downstream of the AUG. The inclusion of the 5′-terminal 100 codons of the annotated CDS was a necessary element to ensure detection of correct mouse orthologs in the following steps. This procedure resulted in a set of 3437 peptide sequences. This set of peptide sequences was then queried against the mouse Refseq mRNA database using tblastn as implemented in NCBI BLAST client netblast-2.2.19. The best hits for each of the peptides were considered as candidate orthologs. Those sequences that produced alignments with a span of at least 50 codons upstream of the annotated CDS were extracted for further processing; 2194 human–mouse orthologous pairs were obtained. Sequences were conceptually translated, re-aligned using muscle (36) and back-translated using t_coffee utilite (37) to yield nucleotide alignments. A custom perl script was designed to trim P5EC alignments prior to K_a/K_s calculation in the following way. Columns containing gaps in either the human or mouse sequence were removed. Because the alignment of sequences was carried out at the amino acid level, all gaps at the nucleotide level occur in multiples of 3 and therefore their removal does not affect reading frame. For alignments containing in-frame stop codons in the mouse sequence, columns corresponding to the stop codons and all columns upstream were removed. This was done based on the assumption that, in general, functionally important P5ECs and alternative translation initiation starts (ATISs) should be conserved between mouse and human sequences and that the presence of a stop codon in a mouse sequence indicates that the sequence is not translated in mouse. If, after trimming, the resulting alignment was shorter than 25 codons, then the corresponding mouse P5EC could not be longer than 25 codons. Also it would not be practical to include such short alignments into further analysis. Therefore, only alignments with a length of at least 25 codons upstream of the annotated AUG (1193 alignments) were subjected to K_a/K_s analysis as implemented in the codonml program of the PAML package (38). Then the alignments were scored based on K_a/K_s values, length of the pairwise alignments and identity at the protein level.

Identification of candidates with known alternative transcript variants

An additional step, as discussed in the 'Results' section, was the identification of those P5EC-containing mRNAs whose genes are known to have alternative transcripts containing 5′-extensions to the annotated CDS. For this purpose, annotations of mRNA sequences from the pool above (1193) were searched for the following text strings: 'transcript variant' or 'shorter' preceding 'N-terminal' within the same annotation field of GenBank format RefSeq entries. Due to the lack of uniformity in annotation descriptions, this procedure does not guarantee identification of all mRNAs with alternatively spliced transcripts where the CDS is 5′-terminally extended. However, this step allowed us to achieve two goals. First, we eliminated a large portion (204 instances) of genes where P5ECs exist because they correspond to translated regions in alternative transcripts and hence most likely do not utilize non-AUG initiation. Second, we created a set of sequences that could be utilized as a positive control, because these genes represent cases where sequences upstream of annotated start codons are known to be translated (see 'Results' section).

As described in the Results section, a relatively high K_a/K_s ratio for the entire P5EC does not necessarily exclude the possibility that at least the 3′ part of the P5EC might be evolving under purifying selection. Therefore, we used a relatively relaxed threshold for choosing the candidates for further detailed semi-manual analysis, i.e. $K_a/K_s < 1$, and the existence of a significant human–mouse pairwise alignment that spans at least 25 codons in length. Identity at the nucleotide or protein level was not taken into account at this point. This resulted in 742 candidates for further analysis.

Analysis of the approximately 700 candidates using deep phylogenetic analysis and manual evaluation

As it has been mentioned above, our filter for RefSeq mRNAs containing P5ECs due to alternative splicing was not exhaustive because of idiosyncrasies in RefSeq annotation comments. Therefore, all 742 candidates were subjected to further examination for the presence of alternative splice forms using the Entrez Gene database (39). All candidates containing alternative splice or transcription initiation forms resulting in noticeably variable 5′-ends were excluded from further analysis. The sequences of those candidates that passed this step were queried against human ESTs from dbEST (40) using blastn. Retrieved ESTs were used to search for the possible existence of non-annotated splice variants in the vicinity of the 5′-UTR-CDS junctions in the candidates under examination. Sequences that showed multiple 5′ variants, either resulting from alternative transcription initiation or alternative splicing, not included in RefSeq, but deemed likely to be real were, with a few exceptions, removed from further consideration. The exceptions were made for alternative transcripts that nevertheless lacked any upstream in-frame AUG codons. Among the set of genes that passed the filters above, we found numerous cases where the 5′-end extensions contained

in-frame AUG codons not present in the annotated RefSeq sequence. Such examples were also excluded from further analysis. For the candidates that passed this step, the human extension up to the upstream in-frame stop codon was conceptually translated and queried against the dbEST, wgs and RefSeq_rna databases to retrieve sequences from non-human species that share significant similarity with the query. If the coding sequence was interrupted either by an in-frame stop codon or a frameshift in sequences from multiple organisms (i.e. >20% of the total available), especially if the conservation at the amino acid level was weak, such candidates were deemed questionable and were excluded from further analysis. The final criterion used to select for positives in the screen was the presence of a well-conserved 'near cognate' initiation codon for the putative extension. In total, 36 genes passed all these tests.

Multiple sequence alignments were constructed for the final list of candidates, plus the previously identified cases of upstream non-AUG-initiated extensions that passed the same selection criteria, and analyzed using PAML codonml (38,41) and MLOGD (42). Codonml was used to recalculate K_a/K_s, now using the multiple species alignments instead of just human–mouse and restricting to the region between the annotated AUG and the predicted upstream non-AUG initiation codon. MLOGD calculates coding potential by using the pattern of substitutions observed within an input sequence alignment to compare a coding model with a non-coding model via a likelihood ratio test. Although K_a/K_s could not be calculated with MLOGD (it is fixed via the BLOSUM62 matrix used within the coding model), MLOGD could be used to produce graphs of coding potential versus sequence position and could therefore be used to help identify the most probable non-AUG initiation sites.

These alignments were also used to calculate the probability, *P* (no stop codons), that the ORF of the extension would be preserved by chance if the sequence were non-coding. For each alignment, nucleotide columns were randomized 10 000 times and the proportion of randomizations in which the ORF was preserved (i.e. no stop codons in the zero-frame in any sequence) was taken as an estimate for *P* (no stop codons). The procedure controls for the phylogenetic non-independence of aligned nucleotides and any gene-specific nucleotide biases in the extension region (e.g. GC-rich regions tend to have fewer stop codons and therefore tend, by chance, to have longer ORFs). It also controls for gene-specific conservation biases (e.g. in a 5′-UTR that happens to be highly conserved due to some non-coding functional element, a chance ORF in one sequence is more likely to be preserved throughout the alignment). One caveat with interpretation of these *P*-values is that, for a small number of alignments, stop-codon-containing sequences from a small number of species had been previously discarded from the alignment either as presumed sequencing errors or on the assumption that the extension has been lost in some species. The *P* (no stop codons) value was not used as part of the selection criteria. In fact, most of the non-AUG-initiated extensions identified here are not sufficiently long to have a statistically significant *P* (no stop codons) value. However, it is useful for a small number of cases where we identified a very long extension that, nonetheless, was not subject to strong purifying selection and hence had weak K_a/K_s and MLOGD scores.

A similar randomization procedure was used to determine *P*-values for the MLOGD and K_a/K_s scores. Because MLOGD and K_a/K_s are codon-based statistics, alignment columns containing gaps were removed prior to randomization (otherwise, after randomization of nucleotide columns, alignment gaps would no longer occur in groups of three so reading frames would not be conserved between species with and species without alignment gaps). Due to CPU constraints, only 1000 randomizations were used for each alignment, which restricts the resolution of raw *P*-values to ~0.001. In many cases, however, the observed MLOGD and $\log(K_a/K_s)$ values (recalculated for the same degapped alignments) were many standard deviations away from the corresponding mean values for the randomizations. To obtain higher resolution *P*-values (shown in Table 1), a normal approximation was assumed for the distribution of the randomization statistics for any particular alignment (after log transform for K_a/K_s; MLOGD is already on a log scale) and the probabilities of obtaining the observed $\log(K_a/K_s)$ and MLOGD statistics in randomized alignments were calculated with respect to this distribution.

Analysis of ribosomal profiling data

Short reads generated during ribosomal profiling experiments in HeLa cells described in Guo *et al.* (43) were obtained from the Gene Expression Omnibus (accession GSE22004). Short reads from all available experiments (SRR057511, SRR057512, SRR057516, SRR057517, SRR057521, SRR057522, SRR057526, SRR057529, SRR057532) were aggregated and aligned to the 59 mRNA sequences using Bowtie short read aligner (44), allowing zero mismatches in the first 25 nt 'seed' region. According to Guo *et al.* (43), on average, mRNA nucleotide positions corresponding to the 5′-end of the short reads in the alignments are located 15-nt upstream of the P-site tRNA in the ribosome. Therefore, coordinates of ribosome positions were calculated accordingly, i.e. shifted by 15-nt downstream from the beginning of the alignment between a short read and mRNA sequence. During this analysis, we did not check whether particular reads could be potentially aligned to other positions within the human genome and therefore the absolute numbers of footprints may be skewed. The density of ribosomes in CDSs and non-AUG extensions was calculated as the absolute number of footprints corresponding to a particular region divided by its length.

RESULTS

Analysis of P5ECs

Most protein-coding sequences evolve under purifying selection and this feature can be used for detection of protein coding regions in nucleotide sequences (45). If, in a particular gene, translation initiates upstream of the annotated start codon, then the sequence located between

Table 1. Ranking of newly identified and known non-AUG-initiated N-terminal extensions

Rank	RefSeq ID	Gene name	Start	Extension length, codons	MLOGD	*P*(MLOGD)	K_a/K_s	$P(K_a/K_s)$	Blast score, bits	*P* (no stop codons)	Ribosomal footprints (abs)/density	
											Extension	Annotated CDS
A) Newly identified non-AUG-initiated N-terminal extensions in this study												
1	NM_001042589	TMEM8B	CUG	295	71.69	2.4×10^{-28}	0.036	1.5×10^{-46}	566	<0.001	ND	ND
2	NM_001037335	PRIC285	GUG	247	53.32	9.2×10^{-14}	0.185	2.5×10^{-17}	301	<0.001	(4)/0.01	(43)/0.01
3	NM_001010858	RNF187	CUG	109	45.37	4.4×10^{-17}	0.105	8.8×10^{-17}	168	0.094	(88)/0.27	(865)/2.26
4	NM_001136108	R3HCC1	CUG	187	51.31	1.2×10^{-14}	0.199	2.1×10^{-18}	264	<0.001	(158)/0.28	(0)/0.0
4	NM_003760	EIF4G3	AUC	187	26.61	5.4×10^{-09}	0.134	3.4×10^{-12}	217	<0.001	(43)/0.08	(616)/0.13
6	NM_006375	ENOX2	UUG	60	34.12	2.2×10^{-07}	0.139	2.3×10^{-09}	113	<0.001	(3)/0.02	(46)/0.03
7	NM_176677	FLJ36208	CUG	156	33.61	9.0×10^{-10}	0.211	1.1×10^{-10}	235	0.001	ND	ND
8	NM_153756	FNDC5	AUA	75	18.47	9.0×10^{-09}	0.061	4.5×10^{-10}	132	0.04	ND	ND
9	NM_006688	C1QL1	AUU	30	17.5	7.5×10^{-06}	0.023	5.8×10^{-11}	92.3	0.418	ND	ND
10	NM_145008	YPEL4	ACG	35	27.12	3.1×10^{-07}	0.026	3.7×10^{-12}	70.5	0.318	ND	ND
11	NM_182528	C1QL2	AUU	34	19.52	1.3×10^{-06}	0.022	5.3×10^{-11}	67.4	0.362	ND	ND
12	NM_000314	PTEN	CUG	173	11.26	1.5×10^{-05}	0.176	1.7×10^{-05}	190	<0.001	(33)/0.06	(292)/0.24
12	NM_139239	NFKBID	CUG	142	25.37	1.7×10^{-09}	0.27	5.1×10^{-08}	182	<0.001	(16)/0.04	(1)/0.0
14	NM_001015072	UFSP1	CUG	76	18.84	9.2×10^{-08}	0.191	1.6×10^{-07}	122	0.056	ND	ND
15	NM_020153	C11orf60	AUA	54	31.79	2.0×10^{-07}	0.197	5.2×10^{-06}	90.9	0.001	(3)/0.02	(110)/0.1
16	NM_001005404	YPEL2	ACG	33	25.32	1.2×10^{-08}	0.041	1.6×10^{-12}	63.9	0.232	ND	ND
17	NM_020335	VANGL2	AUA	48	9.04	8.8×10^{-04}	0.069	1.7×10^{-05}	92	0.125	ND	ND
18	NM_017457	CYTH2	CUG	45	16.03	3.7×10^{-04}	0.172	4.3×10^{-04}	78.2	0.292	(11)/0.08	(355)/0.29
18	NM_001010908	C1QL3	AUU	32	11.21	9.0×10^{-05}	0	1.6×10^{-33}	61.6	0.309	ND	ND
20	NM_001008223	C1QL4	AUU	31	23.04	3.1×10^{-06}	0.071	4.6×10^{-08}	58.5	0.202	ND	ND
21	NM_001002914	KCTD11	AUU	39	14.33	3.3×10^{-06}	0.035	1.9×10^{-10}	55.5	0.241	(1)/0.01	(19)/0.03
22	NM_025160	WDR26	ACG	100	14.32	1.6×10^{-05}	0.255	1.0×10^{-04}	61.2	0.004	(42)/0.14	(785)/0.4
23	NM_005078	TLE3	CUG	52	3.68	2.3×10^{-03}	0.399	9.8×10^{-03}	87.4	0.248	(10)/0.06	(346)/0.15
24	NM_002250	KCNN4	GUG	25	2.74	3.4×10^{-02}	0.367	6.2×10^{-02}	74.9	0.255	(44)/0.58	(346)/0.27
25	NM_004494	HDGF	GUG	50	9.53	2.5×10^{-04}	0.274	1.2×10^{-03}	61.2	0.405	(119)/0.79	(3373)/4.66
26	NM_013313	YPEL1	ACG	35	9.86	1.2×10^{-04}	0.194	2.3×10^{-05}	39.7	0.009	ND	ND
27	NM_022106	C20orf177	AUA	58	−1.56	1.2×10^{-01}	0.388	4.4×10^{-02}	76.6	<0.001	(43)/0.25	(14)/0.01
28	NM_006079	CITED2	CUG	21	1.81	4.6×10^{-02}	0.246	2.5×10^{-02}	59.2	0.28	(76)/1.19	(165)/0.2
29	NM_182603	ANKRD42	CUG	300	−94.17	3.7×10^{-04}	0.42	1.5×10^{-14}	157	<0.001	(21)/0.05	(27)/0.02
30	NM_014310	RASD2	CUG	107	−9.31	2.5×10^{-03}	0.424	6.2×10^{-05}	111	0.006	ND	ND
31	NM_002506	NGF	UUG	44	0.11	9.4×10^{-03}	0.556	2.2×10^{-01}	72	0.002	(47)/0.35	(97)/0.19
32	NM_152283	ZFP62	GUG	60	8.05	3.2×10^{-03}	0.345	2.7×10^{-03}	41.2	<0.001	(32)/0.18	(47)/0.02
33	NM_001102654	NTF3	UUG	23	−3.4	9.3×10^{-02}	0.426	6.6×10^{-02}	70.6	0.029	ND	ND
34	NM_003252	TIAL1	GUG	90	−12.92	6.5×10^{-02}	0.676	1.6×10^{-01}	103	0.09	(34)/0.13	(1258)/1.11
35	NM_024794	EPHX3	ACG	37	8.88	4.2×10^{-04}	0.391	1.9×10^{-03}	27.3	0.044	ND	ND
36	NM_018646	TRPV6	ACG	40	1.56	7.3×10^{-04}	0.438	2.1×10^{-03}	52.4	0.016	ND	ND
36	NM_033315	RASL10B	UUG	33	−4.78	1.3×10^{-01}	0.418	3.7×10^{-02}	55.1	0.111	ND	ND
38	NM_001080510	C17orf95	CUG	63	−0.36	8.2×10^{-03}	0.432	5.1×10^{-03}	49.3	0.01	(8)/0.04	(205)/0.35
39	NM_023110	FGFR1	ACG	43	−9.22	5.1×10^{-01}	0.684	4.2×10^{-01}	62.8	0.022	(21)/0.16	(141)/0.06
40	NM_153369	KIAA1919	CUG	54	−2.32	1.1×10^{-02}	0.559	4.2×10^{-02}	52.4	0.024	ND	ND
41	NM_001144886	CITED1	CUG	17	−2.86	6.6×10^{-02}	1.039	7.5×10^{-01}	33.7	0.082	ND	ND
42	NM_006645	STARD10	GUG	34	−10.8	4.7×10^{-01}	0.739	4.0×10^{-01}	47.4	0.015	(2)/0.02	(12)/0.01
B) Previously reported non-AUG-initiated N-terminal extensions												
1	NM_002097	GTF3A	CUG	235	126.73	2.8×10^{-34}	0.104	2.0×10^{-33}	404	<0.001	(144)/0.2	(78)/0.2
2	NM_001418	EIF4G2	GUG	206	16.21	9.1×10^{-06}	0	4.7×10^{-46}	421	<0.001	(2832)/4.58	(3638)/1.83
3	NM_001017371	SP3	AUA	217	55.29	7.9×10^{-16}	0.129	3.5×10^{-15}	346	<0.001	(96)/0.15	(570)/0.38
3	NM_175886	PRPS1L1	ACG	67	46.73	5.2×10^{-14}	0.005	1.5×10^{-36}	132	0.002	(494)/2.45	(244)/0.32
5	NM_003213	TEAD4	UUG	73	40.13	1.8×10^{-09}	0.054	2.2×10^{-12}	135	0.014	(257)/1.17	(247)/0.23
6	NM_003214	TEAD3	AUA	65	34.03	7.2×10^{-11}	0.045	1.9×10^{-12}	110	0.045	(52)/0.27	(63)/0.06
6	NM_031895	CACNG8	CUG	34	49.14	6.0×10^{-17}	0.02	2.0×10^{-27}	69.3	0.007	ND	ND
8	NM_016178	OAZ3	CUG	48	47.48	2.8×10^{-08}	0.249	8.4×10^{-05}	88.6	<0.001	ND	ND
9	NM_021961	TEAD1	UUG	22	6.75	5.5×10^{-03}	0.112	4.5×10^{-03}	73.6	0.335	(3)/0.04	(451)/0.36
9	NM_001098504	DDX17	ACG	57	21.96	3.5×10^{-06}	0.192	8.2×10^{-06}	68.9	0.044	(81)/0.34	(4333)/2.21
11	NM_001025366	VEGFA	CUG	180	−4.09	9.8×10^{-04}	0.539	6.6×10^{-03}	190	<0.001	(162)/0.3	(89)/0.13
11	NM_022002	NR1I2	CUG	55	19.39	1.5×10^{-05}	0.289	3.8×10^{-04}	68.2	<0.001	ND	ND
11	NM_001172131	HCK	CUG	21	13.67	3.9×10^{-05}	0.157	2.8×10^{-04}	62.1	0.396	ND	ND
11	NM_000378	WT1	CUG	73	8.27	2.9×10^{-05}	0.339	2.8×10^{-06}	73.2	0.005	ND	ND
15	NM_001172415	BAG1	CUG	71	−4.11	2.8×10^{-03}	0.49	3.4×10^{-03}	82	0.096	(222)/1.04	(1136)/1.38
16	NM_001099456	NPW	CUG	52	0.71	4.6×10^{-03}	0.315	4.1×10^{-04}	48.5	0.109	ND	ND
17	NM_002467	MYC	CUG	15	−8.4	8.2×10^{-01}	1.046	6.8×10^{-01}	0	0.118	(279)/6.07	(4864)/3.68

Column 1, combined ranking; column 2, GenBank accession number; column 3, gene name; column 4, predicted non-AUG initiation codon in human (see Supplementary Dataset 1 for full details); column 5, length in codons of predicted N-terminal extension in human; column 6, MLOGD score (negative values are shown in red and indicate candidates subject to weak or no purifying selection); column 7, *P*-value for MLOGD score based on randomizations; column 8, K_a/K_s ratio (with gap-containing columns removed from the alignment); column 9, *P*-value for K_a/K_s based on randomizations; column 10, BLAST bits score measured on the alignment of the human and mouse extensions; column 11, probability of the ORF of the extension being preserved by chance if non-coding. Note that the MLOGD score scales with alignment length and divergence and the BLAST bits score scales with length. The final ranking is based on the average of rankings by the individual scores (MLOGD, K_a/K_s and BLAST bits). Columns 12 and 13 give information on the number of mRNA fragments protected by ribosomes for extension (column 10) and annotated CDSs (column 11). The absolute number of footprints and density of footprints are separated by a backslash. Density is calculated as the absolute number divided by the length of the mRNA fragment (extension or CDS). 'ND', not detected. A) Ranking of the 42 newly identified non-AUG extensions. GenBank accession numbers highlighted in green represent extensions that are conserved beyond mammals and for which all available sequences from vertebrates appear to utilize non-AUG initiation codon(s); accession numbers highlighted in light blue represent extensions that are conserved beyond mammals and for which some or all non-mammalian sequences appear to utilize AUG instead of non-AUG initiation; accession numbers highlighted in magenta represent extensions that are initiated by AUG codons in at least some mammals; accession numbers highlighted in yellow represent extensions that are conserved only in mammals (in some cases only eutherian mammals) and which are never initiated by AUG codons. B) Ranking of previously reported non-AUG extensions.

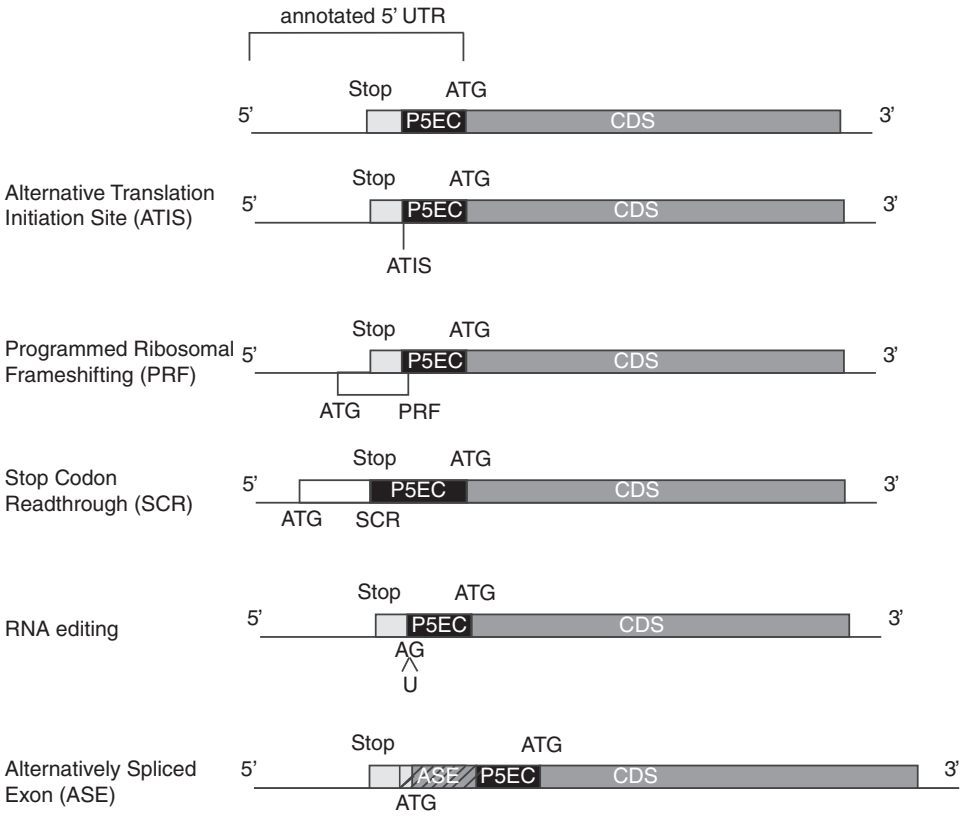

Figure 1. Five known molecular mechanisms responsible for the initiation of translation upstream of the first 5′ in-frame AUG codon. mRNAs are shown as horizontal lines. Dark grey boxes represent annotated CDS regions. Light grey boxes represent extensions of CDSs upstream of annotated AUG codons up to the closest in-frame stop codon. Black boxes denoted as P5EC represent upstream regions where codons in-frame with annotated CDSs evolve under purifying selection. Diagonal stripes are used to denote alternatively spliced exons.

the annotated start codon and the actual (upstream) start codon should evolve under constraints of purifying selection. For brevity, we will refer to such regions as P5ECs. We used the existence of P5ECs as an initial indicator of utilization of alternative initiator codons. In principle, the presence of a P5EC does not guarantee that an alternative in-frame start codon is used for initiation. Figure 1 illustrates five possibilities (examples for which are known) where the sequence upstream of, and in-frame with, an annotated start codon would evolve under purifying selection: (i) Initiation at an upstream in-frame non-AUG codon (ATIS), the subject of this study. (ii) Programmed Ribosomal Frameshifting (PRF), where initiation occurs at the start codon of a uORF and then ribosomes enter the main protein-coding ORF (pORF) by shifting reading frames at a specific location. The most prominent human examples are the three antizyme paralogs (46). (iii) Stop Codon Readthrough (SCR). Similar to the above but where the uORF and the pORF occur in the same translational phase and are separated by a single Stop Codon. No chromosomal genes are known to utilize this phenomenon in humans but, in flies, three examples have been experimentally identified (47) and >100 additional candidates have been identified by comparative sequence analysis (48). (iv) RNA editing. A start codon is generated post-transcriptionally by the insertion of a U between an A and a G, as has been suggested for the linker histon H1F0 and HMGN1 protein genes (49). (v) Alternative Splicing. An exon containing a start codon in one transcript variant could be skipped in another transcript variant. In this case the latter transcript would use an initiator codon located downstream of the start codon of the former transcript. However the region between the 3′-end of the Alternatively Spliced Exon (ASE) and the start codon used in the second transcript would still evolve under the constraints of protein coding sequence, because this region is translated in the first transcript variant. This fifth class of P5EC-containing mRNAs is the largest class. Therefore, during our analysis we paid particular attention to discriminate P5ECs occurring as a result of alternative splicing from those resulting from non-annotated translation events. The pipeline for the initial computational analysis of human Refseq mRNAs is outlined in Figure 2 and is described in detail in the 'Materials and Methods' section.

Comparison of P5ECs due to alternative splicing with the other P5ECs

After obtaining K_a/K_s values for P5EC regions, we generated a set of sequences where the presence of a P5EC is due to the existence of alternative splice variants (see 'Materials and Methods' section). We compared the distribution of K_a/K_s values for these P5ECs with the K_a/K_s values for the remaining P5ECs.

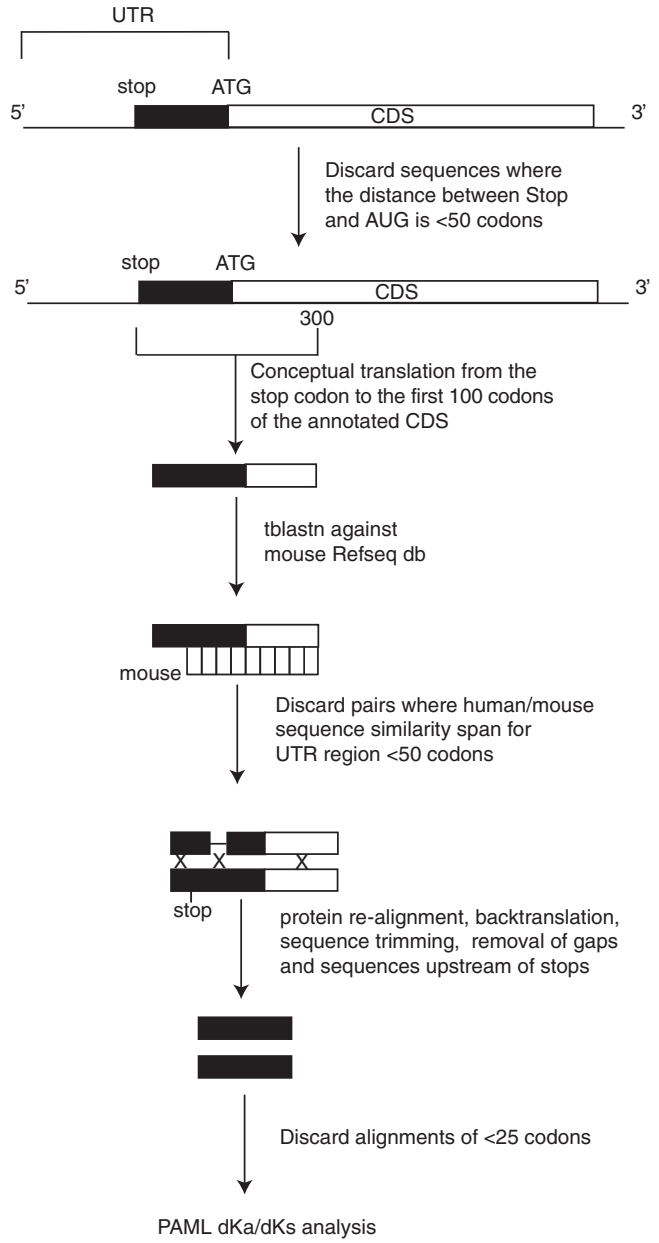

Figure 2. Pipeline of RefSeq mRNA analysis for the identification of conserved 5′ CDS extensions (P5ECs). White boxes indicate annotated CDSs. Black boxes correspond to 5′ in-frame codon extensions up to the closest in-frame stop codon. Xs correspond to the deleted regions of human–mouse alignments prior to K_a/K_s analysis.

The distribution is shown in Figure 3 (see Supplementary Data for the actual values for each sequence). The distribution of K_a/K_s values for mRNAs with known alternative transcript variants (containing CDS 5′ extensions) is significantly sharper than the distribution of K_a/K_s values for the rest of the P5ECs, with the great majority of them falling under 0.2. Therefore, Figure 3 clearly illustrates that K_a/K_s can be used as a predictor of *bona fide* CDS 5′ extensions. Figure 4 shows scatter plot distributions of K_a/K_s ratios for sequences from both data sets in relation to the length of P5EC (upper panels) and the level of identity between mouse and human orthologs at the protein level. While P5ECs from both datasets have highly variable length, it is clear that those resulting from alternative transcript variants have, on average, higher identity at the protein level as well as lower K_a/K_s values. While low K_a/K_s ratio and high protein identity are good indicators of translated P5ECs, high K_a/K_s ratio and low protein identity does not necessarily means that a P5EC is not translated. This is because, at this stage of the analysis, the statistics were calculated for the entire region between the annotated AUG codon and the nearest in-frame stop codon in the 5′-UTR. However, if alternative initiation occurs closer to the 3′-end of a P5EC, then the region of the P5EC upstream of the ATIS would not evolve under the constraints of protein coding evolution and the cumulative values of K_a/K_s and of protein identity for the entire region would be intermediate between values typical of coding and non-coding sequences. Therefore, we used a relatively relaxed K_a/K_s ratio threshold for selecting the candidates for further detailed analysis (see 'Materials and Methods' section).

Identification and analysis of candidates

After a series of automated, semi-automated and manual steps that are described in detail in 'Materials and Methods', 36 candidates satisfied the criteria of our search. One of them, Wilms tumor 1 homolog, has a reported non-AUG initiation in mouse which is homologous to the extension in human and, therefore, the gene was grouped with the known examples of non-AUG initiation discussed elsewhere in this study. In almost all cases, the putative 5′ extensions in the candidates are bracketed upstream by in-frame stop codons in at least one and often many species, thus precluding the existence of an upstream AUG codon for initiation of translation for the extension. In some of the 35 examples, it appears that the occurrence of non-AUG initiation is conserved throughout vertebrates. This is also the case for the known non-AUG extensions in TEAD1, TEAD3, TEAD4 (50) and EIF4G2 (51). Vertebrate genes have numerous paralogs that originated as a result of large-scale DNA duplications in an early chordate (52). For example, TEAD1 has three human paralogs in two of which a non-AUG-initiated N-terminal extension is present and conserved throughout vertebrates. When utilization of non-AUG initiation is ancient for a particular gene, as evident by conservation up to the root of the vertebrate tree, it can be expected that its paralogs may also employ this mechanism. We investigated whether non-AUG extensions are present in paralogs of the 35 provisional positives in our search, especially deeply conserved ones. This led to the identification of seven additional candidates that had escaped initial detection due to limitations in our rigorous selection criteria at various stages of the pipeline.

In the next step, the nature and depth of conservation of the extension in the 42 candidates were probed in non-human/mouse species. In 10 cases (C1QL1, C1QL2, C1QL3, C1QL4, YPEL1, YPEL2, YPEL4, WDR26, FLJ36208 and VANGL2), 7 of them belonging to two

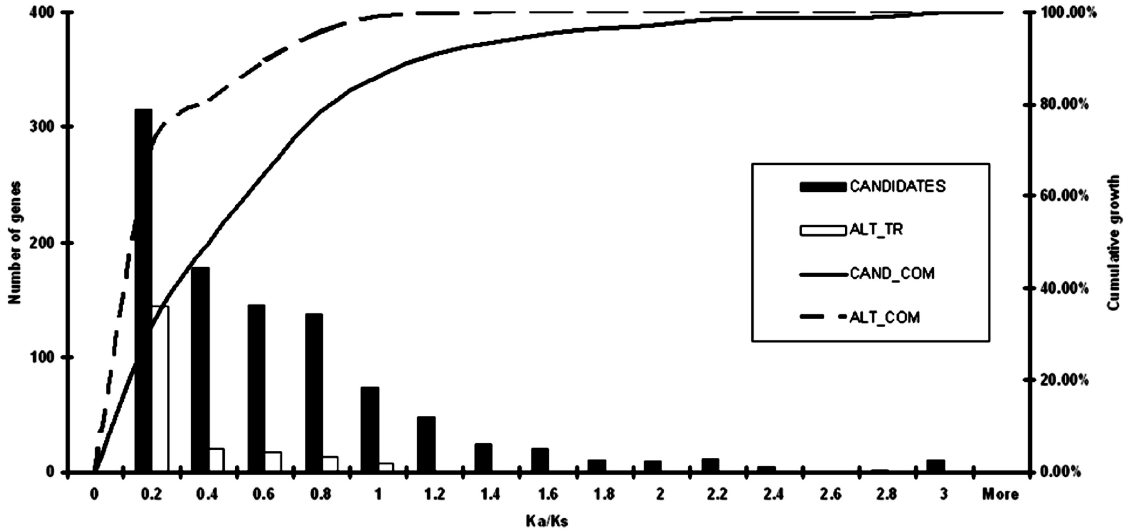

Figure 3. Histogram of K_a/K_s values for mRNA sequences with known 5′ extensions. White bars represent mRNAs for which alternative transcripts with extended CDSs are known and therefore corresponding extensions are known to be translated in alternative transcripts. Sequences of these extensions are expected to evolve as protein coding sequences and were used as an internal control in this study. Black bars represent the remaining mRNAs for which it is not known whether alternative mRNA isoforms exist. Curves indicate the number of genes (y-axis) with K_a/K_s below a particular value (x-axis).

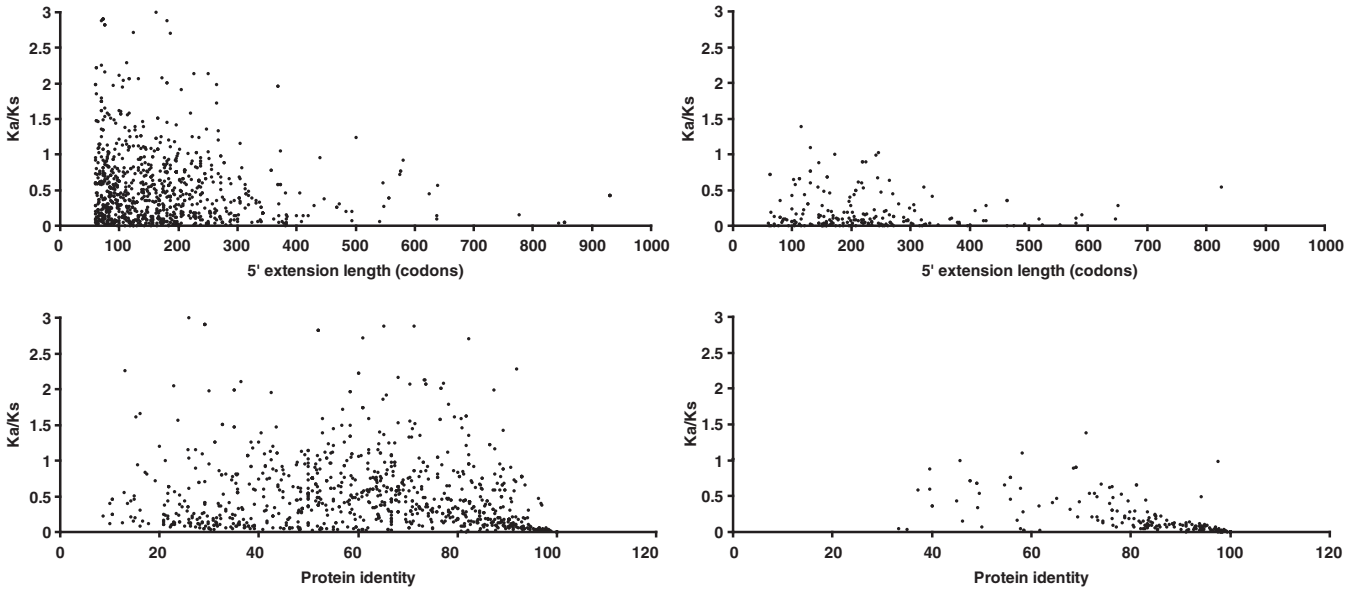

Figure 4. Scatter plots of K_a/K_s ratios for the alignments of the sequences corresponding to P5ECs from different mRNAs (y-axis) in relation to the level of protein identity (bottom panels), and the lengths of P5ECs (top panels). The right-hand panels correspond to mRNAs for which transcript variants with 5′-extended CDSs are known. The left panels correspond to the remaining mRNAs.

paralogous clusters, both the extension and its non-AUG initiation appear to be conserved in all vertebrates for which sequence is available (examples highlighted in green in Table 1). In nine other cases, the extension also appears to be conserved in most or all vertebrates but the non-AUG initiation is conserved only in mammals (highlighted in blue in Table 1). In a further eight cases, the extension appears to be conserved only in mammals and in some of them (usually a minority) it is initiated by AUG either at the position of the putative human non-AUG initiation codon or in its vicinity (highlighted in magenta in Table 1). In two of them the non-AUG initiation evolved recently as all non-primate species have AUG initiation in place of the human non-AUG initiation codon. In one of these, FNDC5, the putative non-AUG initiation is specific to humans while in chimp and gorilla an AUG codon appears to be utilized instead. In the final 15, the extension is conserved only in mammals and is always initiated by a non-AUG codon (highlighted in yellow in Table 1).

In half (i.e. 21) of the newly identified non-AUG-initiated extensions, one to six out-of-frame AUG codons exist between the putative non-AUG start site and the next available in-frame AUG codon downstream. In

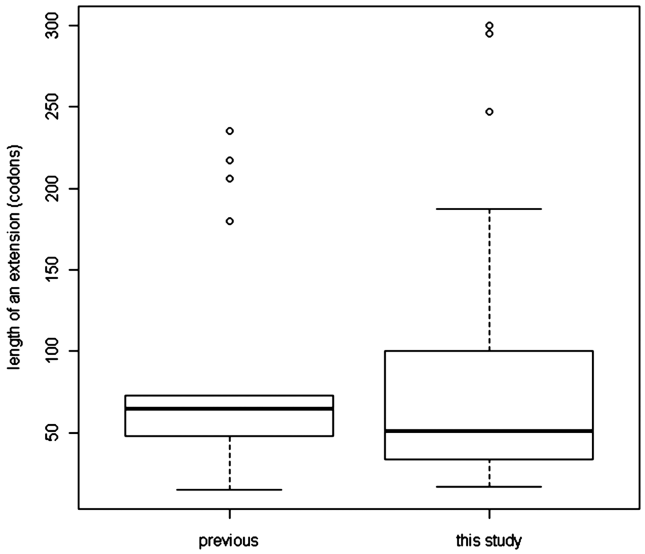

Figure 5. Boxplots of non-AUG CDS extension length distributions for previously known cases and those identified in this study.

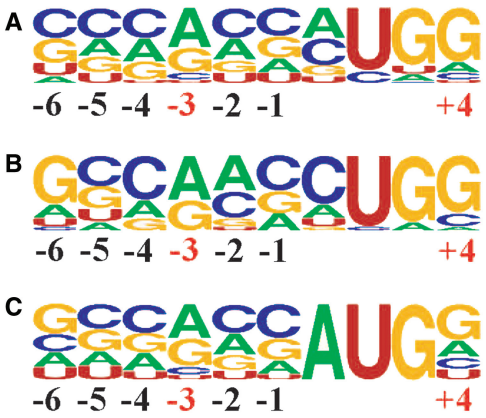

Figure 6. Weblogo representation of the region surrounding the known and putative conserved non-AUG initiation sites in humans. Numbering is relative to the first nucleotide of the start codon. (**A**) Representation for the 42 sequences with newly identified extensions. (**B**) Representation for the 17 sequences with previously identified and conserved extensions. (**C**) Representation of all AUG start sites of humans [the frequencies for nucleotide occurrence at each position for the human mRNAs were obtained from the Transterm database (73)].

the example with six out-of-frame AUG codons, R3HCC1 and perhaps in some others, the downstream in-frame AUG would not be available for initiation, and translation of the main CDS would occur solely via the upstream non-AUG initiation site. Conversely, in the 21 extensions lacking any out-of-frame AUG codons, a significant proportion of ribosomes would be predicted to reach the in-frame AUG codon via leaky scanning, resulting in the translation of multiple C-terminally coincident isoforms.

In addition to the 23 human non-AUG utilizing genes compiled previously (21), we found six others either in the published literature or with annotated non-AUG initiation in RefSeq. To help with the evaluation of the candidates reported here for the first time, these 29 known human cases of non-AUG initiation were subjected to the same qualitative analysis as described above for the new candidates. Seventeen known extensions exhibit a conservation pattern similar to the 42 new candidates, while the remaining 12 do not pass the rigorous qualitative tests for conservation used to identify the 42 new candidates. This analysis also showed that in TEAD1, a UUG codon further upstream, in addition to the previously identified AUU codon (53), is perfectly conserved and is likely to also serve as an initiation codon. In DDX17, the previously proposed CUG initiator (54) is not well conserved; however, tandem ACG-GUG codons downstream are perfectly conserved and are the more likely initiation site(s).

The total size of all 42 newly identified extensions is 3374 codons; the average size is 80.3 codons and the median is 51 codons. In the 17 known cases that passed the qualitative test, the average extension is 87.7 codons and the median is 65 codons. Among the newly identified non-AUG-initiated extensions, the shortest is 17 codons and the longest is 300 codons. Among the 17 known and conserved non-AUG-initiated extensions, the shortest is 15 codons and the longest is 235 codons, see Figure 5 for distribution of extension lengths. The annotated sequences of mRNAs described in Table 1 are available in the Supplementary Dataset 1.

Of the 9 possible codons that differ from AUG in a single position, 5 are used to initiate the extensions of the 17 known cases that passed our qualitative analysis for conservation of the extension. By far the most commonly used is CUG, 10 times (59%), followed by ACG, AUA and UUG with two occurrences of each, and GUG with one occurrence. The distribution of the putative initiation codons of the 42 new candidates in humans is not radically different (notwithstanding the potential for observer bias in locating the precise initiation codon in a small number of cases). Seven of the nine possible codons appear to be utilized, with only the very inefficiently recognized AGG and AAG not used. The distribution of the 7 used is: 15 CUGs (36%), 7 ACGs, 6 GUGs, 5 AUUs, 4 each of UUG and AUA and, finally, 1 AUC. This order correlates with the efficiency of initiation for each non-AUG codon, with CUG the most efficient of them all (16). In addition to the identity of the initiation codon, the initiation context of the 17 previously known examples that passed the screening process and the 42 newly identified candidates were examined (Figure 6). Once again the pattern is similar in both cases though the previously identified cases are closer to the optimal context for mammalian genes.

The 42 candidates and the 17 known examples were subjected to a more precise quantitative analysis, including calculation of K_a/K_s (38,41) and MLOGD (42) scores for multiple sequence alignments of up to 32 vertebrate species, and blast bits scores for human–mouse alignments, as described in 'Materials and Methods'. Candidates were ranked separately by the K_a/K_s ratio, MLOGD score and the BLAST bits score and then the rankings were averaged. The results of these analyses are shown in Table 1. As a general rule, extensions that rank

near the top and that show conservation beyond mammals are far more likely to be physiologically significant. Conversely, extensions that rank near the bottom are conserved only in mammals, are never initiated by AUG codons in any species and are less likely to be physiologically significant even if they are in fact translated.

Independent experimental evidence for the non-AUG extensions from Western blot data

For 6 of the 42 candidates identified in our analysis, published western blot data provide independent experimental support for translation of the non-AUG-initiated extension *in vivo* (see Supplementary Dataset 2). These include:

Ankyrin repeat domain 42 (ANKRD42), also known as several ankyrin repeat protein (SARP), is known to interact with protein phosphatase 1. Browne *et al*. (55) described the original cloning of the human gene and reported the presence of two protein isoforms *in vivo*, one with an apparent molecular weight of 92–95 kDa and another with a molecular weight of 65 kDa. Without supporting evidence, they assumed that the two products result from translation on two alternatively spliced mRNA variants represented by two cloned complementary DNAs (cDNAs): DQ508934 (for the longer isoform) and DQ508815 (for the shorter isoform). This, however, is contradicted by a northern blot presented in the same report [Figure 3F in (55)] which shows a single mRNA species of approximate size 3.6 kb. The authors also noted that their longest clone, DQ508934, does not have an in-frame AUG that could initiate a 92- to 95-kDa protein. They then assumed that the AUG must be present in the genomic region just upstream of the 5′-end of the longer cDNA clone. Such an AUG is indeed present in the genomic DNA of humans, 16 nt upstream of the 5′-end of DQ508934, but there is no evidence that this region is transcribed. In fact, although approximately 50 human ESTs exist for the 5′-end of ANKRD42, not a single one extends beyond the 5′-end of DQ508934. In the orthologs in mouse, dog, orangutan, cattle and other mammals, delimiting upstream in-frame stop codons exist that unambiguously occur 3′ of any in-frame AUG codons that might come from the upstream genomic region. The simplest explanation for the protein species observed *in vivo* is that they correspond to alternative translation initiation from a conserved CUG codon in a good initiation context (producing the extension ranked 29th in Table 1), and the next in-frame AUG codon located ~900 nt downstream. These products have expected molecular weights of 89.6 and 57.3 kDa, respectively, close to the observed 92–95 and 65 kDa. It is significant that, in the 900 nt space between the conserved CUG and first in-frame AUG, there is a single out-of-frame AUG in humans that is immediately followed by a stop codon and is therefore unlikely to significantly affect scanning ribosomes that failed to initiate on the CUG. In mice and rats, however, six out-of-frame AUG codons are present between the CUG and the first in-frame AUG. Thus, in these species, one would predict synthesis of only a single protein isoform, i.e. the longer CUG-initiated form (unless alternative transcripts are produced).

Surprisingly, considering the direct evidence for its *in vivo* expression, the data shown in Table 1 suggest that in this case the 5′ extension does not actually evolve under strong purifying selection ($K_a/K_s \sim 0.4$). Yet the existence of such a long extension (lacking any in-frame stop codons) in most mammalian orthologs strongly argues that the extension is indeed translated.

Hepatoma-derived growth factor (HDGF), also known as high-mobility group protein 1-like, is a developmentally regulated pulmonary endothelial cell-expressed angiogenic factor. HDGF northern blots identified a single 2.3-kb mRNA in many rat tissues (56). Western blot analysis on human microvascular endothelial cells revealed the presence of at least two protein species—one major with apparent size of 43 kDa and one minor with apparent size of 48 kDa (57). Examination of available ESTs aligning with human HDGF reveals evidence for alternative mRNA variants resulting from different transcription initiation sites and a different exon 1, but none of the ESTs contains an additional in-frame AUG upstream of the previously identified initiation codon. Our analysis predicts that the human HDGF mRNA encodes a 50-amino acid extension initiated by a GUG codon in good context, with both the GUG codon and the context conserved in mammals (ranked 25th in Table 1). In several species, e.g. pig and dog, an upstream delimiting stop codon exists. Initiation on the GUG and on the downstream AUG codon is expected to produce protein products with estimated molecular weights of 31.6 and 26.8 kDa, respectively. Neither of these coincides with the observed products on SDS western blots (48 and 43 kDa), but the difference between the two observed products (~5 kDa) is nearly identical to the difference between the two predicted products. No out-of-frame AUG codons exist between the GUG and AUG codons. Thus, it appears that standard leaky scanning is sufficient to account for the observed protein products.

Mammalian eukaryotic translation initiation factor 4 γ 3 (EIF4G3, more commonly known as eIF4GII) is a part of the cap-binding protein complex of eukaryotic translation initiation factor 4F. EIF4GII has two known paralogs, eIF4G1 (also known as eIF4GI) and eIF4G2. The latter is known to be initiated at a non-AUG codon (27,51). The study reporting the cloning and initial characterization of eIF4GII identified a single mRNA species by northern blot in humans with a size of 6.0 kb (58). The measured size of the endogenous protein was ~220 kDa. Another band is present on western blots of the native eIF4GII which is ~20 kDa above the main band. This band is unremarked in the main text, but it is clearly visible in Figure 3B lane 2 of Gradi *et al*. (58). The size difference between the two native bands is ~20 kDa. This is close to the difference between the predicted sizes of the AUC- and AUG-initiated products (extension ranked fourth in Table 1). The proposed AUC initiation codon and its context are conserved in all mammals for which sequence information is available. Curiously, the 4 nt surrounding the putative AUC codon, AAA***AUC***C (−3 and +4 positions underlined), which include the initiation

context, are identical to the equivalent positions of AUU-initiated uORFs in mammalian antizyme inhibitor and the 5′-CUG-initiated N-terminal extension of mammalian antizyme 3 (30).

The non-AUG initiation in EIF4G2 and EIF4G3 is very intriguing because it implies possible autoregulation and hence a role for these two proteins in modulating the stringency of start codon selection. If such autoregulation exists, it would be analogous to the recently discovered autoregulation of eIF1 (see 'Discussion' section). So far, however, neither protein has a known role in the stringency of start codon selection.

RASD family, member 2 (RASD2, alternatively known as Rhes) is a Ras homolog and is also related to other Ras GTP-binding proteins. It is predominantly expressed in the striatum of mammalian brains. Mouse experiments with RASD2 reveal behavioral abnormalities when the gene is inactivated (59). Our analysis identified a CUG-initiated extension encoded by the RASD2 mRNA that is conserved in mammals and ranked 30th in Table 1. The knockout analysis is accompanied by western blots on striatum extracts from wild-type and *Rasd2*$^{-/-}$ animals (59). In mice the CUG-initiated peptide has an expected size of 41.2 kDa while downstream in-frame AUG initiation would result in a protein with a predicted molecular weight of 30.2 kDa. The western blots show three protein products present in the wild-type animals but not in the knockout. One, with an apparent molecular weight of 31 kDa, is nearly identical to a band that appears in HeLa cells transfected with a *Rasd2* construct lacking the extension, thus suggesting that this is the product of initiation from the AUG codon. The two other *Rasd2*-specific products on the western blot have apparent sizes of 39 and 48 kDa. It seems likely that one (or both) of these corresponds to the CUG-initiated protein product. Available human and mouse ESTs do not provide support for the existence of alternative transcripts that could be used to explain the presence of the larger product.

Neurotrophin 3 (NTF3) is a target-derived neurotrophic factor. NTF3 and its paralogs, nerve growth factor, brain-derived neurotrophic factor, neurotrophin 4/5 and neurotrophin 6, bind to both low- and high-affinity receptors on target cells to elicit a cascade of intracellular responses to produce their biological effects. Disruption of NTF3 leads to severe sensory and sympathetic deficits that are incompatible with postnatal life in mice (60). Our analysis indicates the presence of a UUG-initiated extension in human NTF3 (ranked 33rd in Table 1). An antibody for NTF3 is commercially available from Abcam. The product description page shows a western blot with NTF3 antibody detecting two bands—one with an apparent size of 30 kDa and one with a size of 32 kDa, closely matching the predicted sizes of NTF3 initiated at the UUG (33.3 kDa) and at the downstream in-frame AUG (30.8 kDa).

WD repeat domain 26 (WDR26) is a WD40 repeat-containing protein that might be involved in signal transduction and might affect transcriptional regulation (61). The protein is highly conserved from humans to yeast. Our analysis identified a well-conserved ACG/CUG-initiated extension in vertebrates that is present from human to fish, and ranked 22nd in Table 1. In humans the AUG- and ACG-initiated proteins have expected molecular weights of 72.1 and 81.3 kDa, respectively. An antibody for WDR26 is commercially available from Abcam. The product description page shows a western blot that displays two different bands corresponding to proteins with apparent sizes of 80 kDa, for the main product, and 90 kDa for the minor product. The ~10 kDa difference between the two corresponds closely to the difference expected for initiation on the conserved ACG and on the first in-frame AUG codon.

Curiously, the published annotations of this gene (and the previous version of the Refseq mRNA NM_025160.5) not only omit the non-AUG initiated extension but also the well-conserved first in-frame AUG, and instead assume that initiation occurs on the second in-frame AUG codon (61,62). (Note, however, that the latest version of the Refseq entry NM_025160.6 [updated during preparation of this article on 21 July 2010] has the CDS starting from the first in-frame AUG.)

Orthologs of WDR26 in non-vertebrate metazoans were identified and investigated for the presence of the non-AUG-initiated extension. The gene in *Drosophila* has an AUU-initiated extension that is conserved in other flies though not in mosquitoes. It appears that the extension arose independently in vertebrates and in flies.

Evidence of non-AUG initiation from ribosomal profiling data

The second set of experimental data supporting our predictions was obtained from recently published ribosomal profiling experiments carried out in human cells (43). In this work, mRNA fragments protected by ribosomes were converted into a library of oligonucleotides and subjected to massively parallel sequencing. The alignment of corresponding sequences onto sequences of mRNAs allows detection of ribosomal locations and the density of the ribosomes on mRNA. We have aligned raw sequences obtained by Guo *et al*. (43) using the bowtie program (44) and quantified locations of the ribosomes on the mRNA sequences where non-AUG initiation was predicted (see 'Materials and Methods' section). We have been able to extract reads for 32 of 59 genes. Examples of ribosomal profiles for NM_004494 and NM_001010858 are shown in Figure 7. For the other sequences, the number of footprints corresponding to CDS and non-AUG extensions, as well as ribosomal density in these regions, are given in Table 1. The presence of ribosomal profiles in the extensions upstream of annotated CDSs indicates that initiation takes place upstream of annotated AUG initiation codons and supports our predictions regarding non-AUG initiation.

Other noteworthy candidates for non-AUG initiation

Ecto-NOX disulfide-thiol exchanger 2 (ENOX2, also known as tNOX) is a hydroquinone and NADH oxidase with protein disulfide-thiol interchange activity (63). It is associated with the outer leaflet of the plasma membrane at the surface of cancer cells and in sera of cancer patients

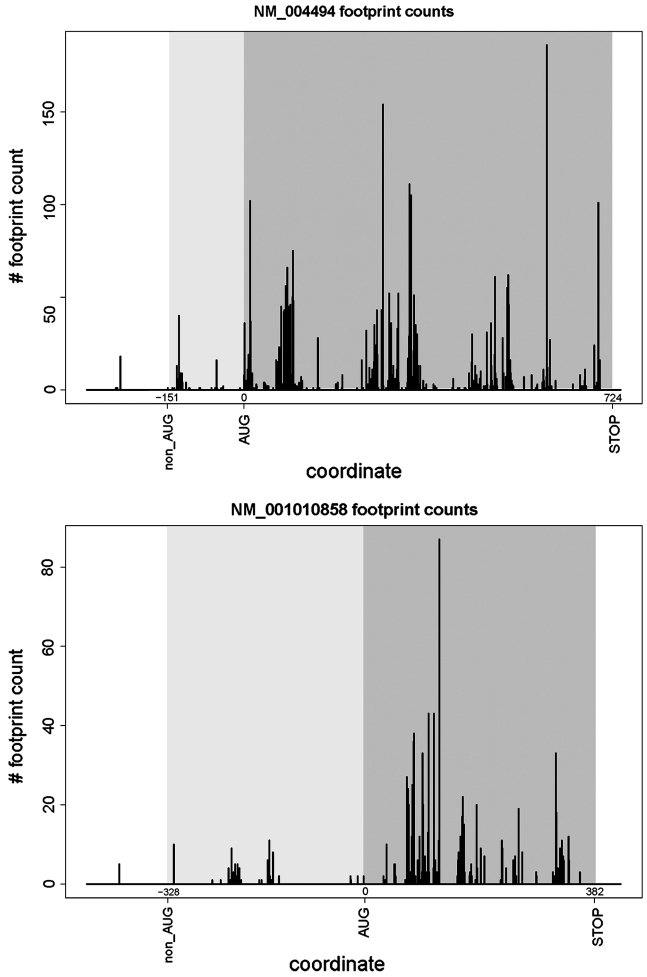

Figure 7. Plots showing density of mRNA fragments protected by ribosomes for NM_004494 and NM_001010858. The position of the annotated AUG codon was taken as zero; relative coordinates of stop codons and predicted non-AUG initiators are indicated. Regions corresponding to annotated CDSs are highlighted in dark grey; regions corresponding to non-AUG-initiated extensions are highlighted in light grey. The presence of ribosomal footprints in the region of an extension indicates that the initiation of translation takes place upstream of the annotated CDS.

but is absent from the surface of non-cancer cells and from sera of healthy individuals (64). Full-length tNOX mRNA is present in both normal and cancer cells, but its translation is apparently inhibited in normal cells. Our analysis identified an N-terminal extension of 60 amino acids, likely initiated at an upstream UUG codon. The non-AUG initiation is conserved only in mammals, but the extension itself appears conserved in most non-mammalian vertebrates where it is initiated by an AUG codon. The gene ranks sixth in the analysis shown in Table 1 and is strongly supported by evolutionary evidence derived from codon substitutions in multiple alignments.

The mRNA of tNOX has a feature that perhaps explains the low level of expression of this protein in normal cells. The region between the UUG and the downstream in-frame AUG contains an out-of-frame AUG codon in very strong initiation context which starts a 22-codon ORF. The length of this ORF is close to that considered to prevent reinitiation following its translation. Therefore, the role of this short ORF is likely to inhibit AUG initiation of the main CDS thus ensuring that translation of the main CDS is almost solely dependent on initiation at the upstream inefficient UUG.

Western blots for tNOX are available but are not informative about the utilization of the extension *in vivo*. The detected product is 34 kDa which is only a fraction of the predicted size, 66.6 kDa, of even the AUG-initiated polypeptide, thus indicating extensive post-translational processing.

Not much is known about the function of the gene R3H domain and coiled coil containing 1 (R3HCC1). The protein is well conserved in mammals. Two features of R3HCC1 make it particularly striking with regard to its proposed non-AUG initiation. The first is the existence of six out-of-frame AUG codons between the putative CUG initiation codon in humans, present in perfect Kozak context, and the next available in-frame AUG codon, as mentioned above. This configuration is also present in the other six available mammalian sequences which, on average, have 5.2 out-of-frame AUG codons in this region. The second is the fact that the position of this first available in-frame AUG codon is variable and, in cattle (*Bos taurus*), it is just eight codons away from the stop codon. In cattle, as indeed in all other mammals, an upstream in-frame delimiting stop codon exists, suggesting that translation of the main CDS would occur solely via the non-AUG initiation site. Considering the large number of out-of-frame AUGs following the conserved CUG (in near perfect Kozak context) and the next in-frame AUG, and the variable position of the AUG, it seems very likely that the CUG codon is the sole position at which translation of this CDS is initiated in most or all mammals.

Ufm1-specific peptidase 1 (UFSP1) is responsible for release of the ubiquitin-like protein Ufm1 from Ufm1-conjugated cellular proteins as well as for activating the Ufm1 precursor (65). UFSP1 is a cysteine protease in which Cys53 (numbering relative to the mouse sequence) is part of the catalytic center and is believed to carry out the nucleophilic attack on the carbonyl carbon of the substrate (66). UFSP1 is a single-exon gene in mammals. Unlike the gene in *Mus musculus* and three other species, the other 16 available mammalian sequences do not appear to have in-frame AUG codons capable of initiating translation of the full-length protein. The first available in-frame AUG is located at a position equivalent to Met85 of the mouse sequence—i.e. more than 30 residues downstream of the critical Cys53. This means that, even if UFSP1 initiated with Met85 has some biological activity, it would be predicted to be non-functional as a cysteine protease. In fact because of this the human ortholog is currently designated as 'non-functional' although our analysis suggests that initiation at a CUG codon (in a good Kozak context and in a position nearly identical to Met1 of mouse Ufsp1) would produce a fully functional enzyme.

DISCUSSION

In this study, we identified 42 novel candidates for conserved non-AUG-initiated N-terminal extensions in RefSeq human mRNAs. The purpose of this study was to find the most reliable candidates for non-AUG initiation and the approach used was highly conservative. Therefore, a large number of filters were used to increase selectivity at the expense of decreased sensitivity. First, the initial pool of data were limited only to those human mRNAs that derive from genes with detectable homologs in mouse. This limitation was needed to ensure sufficient depth of phylogenetic analysis required for MLOGD and K_a/K_s calculations. Therefore all instances of non-AUG initiation utilization in human genes with no orthologs in mouse would be missed by this study. The second limitation was the length of P5EC. All sequences with in-frame stop codons closer than 50 codons to the annotated CDS start site were removed by our pipeline. In addition, all sequences that do not produce an ungapped alignment of mouse-human P5ECs of at least 25 codons in length were removed. The minimal alignment length was set to ensure a sufficient number of substitutions for K_a/K_s estimation. Therefore it is likely that we missed some cases where a non-AUG initiator is located close to the annotated CDS start. In other cases, alternative splicing or alternative transcription initiation sites create mRNA species that are missing the first annotated AUG codon. It is conceivable that some of these transcripts could be translated from a non-AUG codon prior to the next available in-frame AUG. Our search deliberately excluded these potential cases. The currently available RefSeq sequences are often missing significant portions of the 5′-terminal regions of the corresponding mRNAs. This became evident during our manual analysis of the 742 automatically selected candidates. In fact, in one case of a candidate that made it among the 42 finalists (NM_025160, version from 26 June 2007), the originally aligned sequence did have an upstream in-frame AUG codon. However, after the analysis of available ESTs for the locus we built a gene model with an extended 5′-end beyond this AUG that itself turned out to have a conserved non-AUG-initiated extension. The extension appeared to have a candidate non-AUG start, and the protein sequence generated by conceptual translation starting from this non-AUG codon is phylogenetically conserved. In this context, it should be noted that it is conceivable that the extensions of some of the candidates identified here are after all initiated by AUG codons resulting from rare transcript variants for which there is currently no evidence in existing EST databases. Another recently identified case of non-AUG initiation [CUG in thioredoxin reductase 3, TXNRD3 (67)] escaped our analysis for a similar reason—the corresponding Refseq entry XM_001130163 lacks the 5′ region where the CUG codon is situated. Cases of N-terminal extensions under positive selection would also have been missed. Finally, only extensions that can be traced near the root of the mammalian tree were scored as candidates in our analysis but it is likely that smaller phylogenetic branches also have conserved and physiologically active non-AUG-initiated extensions. For example our analysis indicates that the known non-AUG extension of CDKN2B is conserved in primates but not in more distantly related mammals. For these and probably other reasons the cases of physiologically important non-AUG initiation are likely to be more numerous than those identified in the current and previous studies. Furthermore, there could be non-AUG-initiated coding regions in the 5′-UTRs of mRNAs that are in a different reading frame or separated by a stop codon from the main CDS. Examples include a uORF that regulates expression of AZIN1 in vertebrates (30), and others found by 'wet' proteomics analysis (68). Such cases of non-AUG initiation would have been completely undetected by our search. Despite all these limitations, our analysis suggests the existence of at least 59 human genes, ~0.3% of the total, with conserved non-AUG-initiated N-terminal extensions.

Previous examples of non-AUG initiation were almost always found serendipitously after experimental data showed mismatches between expected and observed protein masses or sequences. As our analysis has shown, often such additional protein products were treated as artifacts or simply ignored because identification of their nature was a problem remote from the primary goals of these studies. This study has addressed the issue systematically by employing the power of comparative genomics to predict completely overlooked conserved potential N-terminal extensions.

A very exciting possible reason for utilization of non-AUG initiation is alternative regulation of initiation at AUG and non-AUG start codons. Eukaryotes have developed elaborate mechanisms for recognition of the correct initiation codon. Mutations are known that reduce the fidelity of initiation in eukaryotes—specifically the 'Sui$^-$' mutants in *Saccharomyces cerevisiae* (69–70). Many of these mutations occur in SUI1, also known as eukaryotic initiation factor 1 (eIF1). Within the ribosomal machinery ensuring fidelity of initiation in eukaryotes, eIF1 seems to play a central role (13,71). eIF1 has been implicated in all three proofreading processes during initiation: discrimination between AUG and non-AUG codons; discrimination between AUG codons in good and poor (Kozak) context and discrimination between AUG codons close (<20 nts) to the 5′ cap and AUG codons that are further away from the 5′ cap (the latter being preferred for initiation in eukaryotes). eIF1 appears to confer conformational changes to the scanning 40S ribosome (72). When bound to it, eIF1 induces an open conformation conducive to scanning (13). Upon eIF1 release, the 40S subunit is believed to take a closed conformation that precludes further scanning and promotes initiation of translation. If non-AUG initiation is widely used for regulation, it seems likely that eIF1 plays a major role in the mechanism. Remarkably, eIF1 in most eukaryotes appears to be under translation autoregulation utilizing its ability to inhibit initiation at poor start sites (32). In this case, an AUG codon in a poor context is used instead of non-AUG initiation, but experiments in the same study demonstrated that the amplitude of repression associated with high levels of eIF1 is even more

pronounced at non-AUG start codons. Although the existence of a translational regulation mechanism that exploits variability in the stringency of start codon selection is now proven, little is known about the extent of its utilization in humans. Tripling the number of known conserved human non-AUG initiation sites would certainly aid in addressing this issue.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Drs Nick Ingolia and Jonathan Weismann for introducing us to the details of the ribosomal profiling technique.

FUNDING

Science Foundation Ireland (grant number 06/IN.1/B81 to P.V.B., 08/IN.1/B1889 to J.F.A.); Wellcome Trust (grant number 088789 to A.E.F.); National Institute of Health (grant number GM079523 to J.F.A.). Funding for open access charge: Science Foundation Ireland.

Conflict of interest statement. None declared.

REFERENCES

1. Ramakrishnan,V. (2002) Ribosome structure and the mechanism of translation. *Cell*, **108**, 557–572.
2. Simonetti,A., Marzi,S., Myasnikov,A.G., Fabbretti,A., Yusupov,M., Gualerzi,C.O. and Klaholz,B.P. (2008) Structure of the 30S translation initiation complex. *Nature*, **455**, 416–420.
3. Potapov,A.P., Triana-Alonso,F.J. and Nierhaus,K.H. (1995) Ribosomal decoding processes at codons in the A or P sites depend differently on 2′-OH groups. *J. Biol. Chem.*, **270**, 17680–17684.
4. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2004) P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA*, **10**, 221–230.
5. Ogle,J.M., Brodersen,D.E., Clemons,W.M. Jr, Tarry,M.J., Carter,A.P. and Ramakrishnan,V. (2001) Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, **292**, 897–902.
6. Baumann,L., Thao,M.L., Hess,J.M., Johnson,M.W. and Baumann,P. (2002) The genetic properties of the primary endosymbionts of mealybugs differ from those of other endosymbionts of plant sap-sucking insects. *Appl. Environ. Microbiol.*, **68**, 3198–3205.
7. Shine,J. and Dalgarno,L. (1975) Determinant of cistron specificity in bacterial ribosomes. *Nature*, **254**, 34–38.
8. Kozak,M. (1980) Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. *Cell*, **22**, 7–8.
9. Kozak,M. (2001) New ways of initiating translation in eukaryotes? *Mol. Cell. Biol.*, **21**, 1899–1907.
10. Schneider,R., Agol,V.I., Andino,R., Bayard,F., Cavener,D.R., Chappell,S.A., Chen,J.J., Darlix,J.L., Dasgupta,A., Donze,O. *et al.* (2001) New ways of initiating translation in eukaryotes. *Mol. Cell. Biol.*, **21**, 8238–8246.
11. Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.
12. Mitchell,S.F. and Lorsch,J.R. (2008) Should I stay or should I go? Eukaryotic translation initiation factors 1 and 1A control start codon recognition. *J. Biol. Chem.*, **283**, 27345–27349.
13. Jackson,R.J., Hellen,C.U. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.
14. Kozak,M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.
15. Kozak,M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283–292.
16. Peabody,D.S. (1989) Translation initiation at non-AUG triplets in mammalian cells. *J. Biol. Chem.*, **264**, 5031–5035.
17. Chen,S.J., Lin,G., Chang,K.J., Yeh,L.S. and Wang,C.C. (2008) Translational efficiency of a non-AUG initiation codon is significantly affected by its sequence context in yeast. *J. Biol. Chem.*, **283**, 3173–3180.
18. Kozak,M. (1989) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol.*, **9**, 5073–5080.
19. Portis,J.L., Spangrude,G.J. and McAtee,F.J. (1994) Identification of a sequence in the unique 5′ open reading frame of the gene encoding glycosylated Gag which influences the incubation period of neurodegenerative disease induced by a murine retrovirus. *J. Virol.*, **68**, 3879–3887.
20. Kozak,M. (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc. Natl Acad. Sci. USA*, **87**, 8301–8305.
21. Tikole,S. and Sankararamakrishnan,R. (2006) A survey of mRNA sequences with a non-AUG start codon in RefSeq database. *J. Biomol. Struct. Dyn.*, **24**, 33–42.
22. Kozak,M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, **299**, 1–34.
23. Natsoulis,G., Hilger,F. and Fink,G.R. (1986) The HTS1 gene encodes both the cytoplasmic and mitochondrial histidine tRNA synthetases of *S. cerevisiae*. *Cell*, **46**, 235–243.
24. Chatton,B., Walter,P., Ebel,J.P., Lacroute,F. and Fasiolo,F. (1988) The yeast VAS1 gene encodes both mitochondrial and cytoplasmic valyl-tRNA synthetases. *J. Biol. Chem.*, **263**, 52–57.
25. Souciet,G., Menand,B., Ovesna,J., Cosset,A., Dietrich,A. and Wintz,H. (1999) Characterization of two bifunctional *Arabidopsis thaliana* genes coding for mitochondrial and cytosolic forms of valyl-tRNA synthetase and threonyl-tRNA synthetase by alternative use of two in-frame AUGs. *Eur. J. Biochem.*, **266**, 848–854.
26. Chang,K.J. and Wang,C.C. (2004) Translation initiation from a naturally occurring non-AUG codon in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **279**, 13778–13785.
27. Imataka,H., Olsen,H.S. and Sonenberg,N. (1997) A new translational regulator with homology to eukaryotic translation initiation factor 4G. *EMBO J.*, **16**, 817–825.
28. Rajkowitsch,L., Vilela,C., Berthelot,K., Ramirez,C.V. and McCarthy,J.E. (2004) Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. *J. Mol. Biol.*, **335**, 71–85.
29. Hinnebusch,A.G. (2005) Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.*, **59**, 407–450.
30. Ivanov,I.P., Loughran,G. and Atkins,J.F. (2008) uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. *Proc. Natl Acad. Sci. USA*, **105**, 10079–10084.
31. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
32. Ivanov,I.P., Loughran,G., Sachs,M.S. and Atkins,J.F. (2010) Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc. Natl Acad. Sci. USA*, **107**, 18056–18060.
33. Shabalina,S.A., Ogurtsov,A.Y., Rogozin,I.B., Koonin,E.V. and Lipman,D.J. (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.*, **32**, 1774–1782.
34. Kochetov,A.V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays*, **30**, 683–691.

35. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
36. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
37. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
38. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
39. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
40. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
41. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
42. Firth,A.E. and Brown,C.M. (2006) Detecting overlapping coding sequences in virus genomes. *BMC Bioinformatics*, **7**, 75.
43. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.
44. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
45. Nekrutenko,A., Makova,K.D. and Li,W.H. (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
46. Ivanov,I.P. and Atkins,J.F. (2007) Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation. *Nucleic Acids Res.*, **35**, 1842–1858.
47. Bekaert,M., Firth,A.E., Zhang,Y., Gladyshev,V.N., Atkins,J.F. and Baranov,P.V. (2010) Recode-2: new design, new search tools, and many more genes. *Nucleic Acids Res.*, **38**, D69–D74.
48. Lin,M.F., Carlson,J.W., Crosby,M.A., Matthews,B.B., Yu,C., Park,S., Wan,K.H., Schroeder,A.J., Gramates,L.S., St Pierre,S.E. *et al.* (2007) Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res.*, **17**, 1823–1836.
49. Zougman,A., Ziolkowski,P., Mann,M. and Wisniewski,J.R. (2008) Evidence for insertional RNA editing in humans. *Curr. Biol.*, **18**, 1760–1765.
50. Stewart,A.F., Richard,C.W. 3rd, Suzow,J., Stephan,D., Weremowicz,S., Morton,C.C. and Adra,C.N. (1996) Cloning of human RTEF-1, a transcriptional enhancer factor-1-related gene preferentially expressed in skeletal muscle: evidence for an ancient multigene family. *Genomics*, **37**, 68–76.
51. Takahashi,K., Maruyama,M., Tokuzawa,Y., Murakami,M., Oda,Y., Yoshikane,N., Makabe,K.W., Ichisaka,T. and Yamanaka,S. (2005) Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2). *Genomics*, **85**, 360–371.
52. McLysaght,A., Hokamp,K. and Wolfe,K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.*, **31**, 200–204.
53. Xiao,J.H., Davidson,I., Matthes,H., Garnier,J.M. and Chambon,P. (1991) Cloning, expression, and transcriptional properties of the human enhancer factor TEF-1. *Cell*, **65**, 551–568.
54. Uhlmann-Schiffler,H., Rossler,O.G. and Stahl,H. (2002) The mRNA of DEAD box protein p72 is alternatively translated into an 82-kDa RNA helicase. *J. Biol. Chem.*, **277**, 1066–1075.
55. Browne,G.J., Fardilha,M., Oxenham,S.K., Wu,W., Helps,N.R., da Cruz,E.S.O.A., Cohen,P.T. and da Cruz,E.S.E.F. (2007) SARP, a new alternatively spliced protein phosphatase 1 and DNA interacting protein. *Biochem. J.*, **402**, 187–196.
56. Everett,A.D. (2001) Identification, cloning, and developmental expression of hepatoma-derived growth factor in the developing rat heart. *Dev. Dyn.*, **222**, 450–458.
57. Everett,A.D., Narron,J.V., Stoops,T., Nakamura,H. and Tucker,A. (2004) Hepatoma-derived growth factor is a pulmonary endothelial cell-expressed angiogenic factor. *Am. J. Physiol. Lung Cell Mol. Physiol.*, **286**, L1194–L1201.
58. Gradi,A., Imataka,H., Svitkin,Y.V., Rom,E., Raught,B., Morino,S. and Sonenberg,N. (1998) A novel functional human eukaryotic translation initiation factor 4G. *Mol. Cell. Biol.*, **18**, 334–342.
59. Spano,D., Branchi,I., Rosica,A., Pirro,M.T., Riccio,A., Mithbaokar,P., Affuso,A., Arra,C., Campolongo,P., Terracciano,D. *et al.* (2004) Rhes is involved in striatal function. *Mol. Cell. Biol.*, **24**, 5788–5796.
60. Zhou,X.F. and Rush,R. (1995) Sympathetic neurons in neonatal rats require endogenous neurotrophin-3 for survival. *J. Neurosci.*, **15**, 6521–6530.
61. Zhu,Y., Wang,Y., Xia,C., Li,D., Li,Y., Zeng,W., Yuan,W., Liu,H., Zhu,C., Wu,X. *et al.* (2004) WDR26: a novel Gbeta-like protein, suppresses MAPK signaling pathway. *J. Cell. Biochem.*, **93**, 579–587.
62. Wei,X., Song,L., Jiang,L., Wang,G., Luo,X., Zhang,B. and Xiao. (2010) Overexpression of MIP2, a novel WD-repeat protein, promotes proliferation of H9c2 cells. *Biochem. Biophys. Res. Commun.*, **393**, 860–863.
63. Chueh,P.J., Kim,C., Cho,N., Morre,D.M. and Morre,D.J. (2002) Molecular cloning and characterization of a tumor-associated, growth-related, and time-keeping hydroquinone (NADH) oxidase (tNOX) of the HeLa cell surface. *Biochemistry*, **41**, 3732–3741.
64. Cho,N., Chueh,P.J., Kim,C., Caldwell,S., Morre,D.M. and Morre,D.J. (2002) Monoclonal antibody to a cancer-specific and drug-responsive hydroquinone (NADH) oxidase from the sera of cancer patients. *Cancer Immunol. Immunother.*, **51**, 121–129.
65. Kang,S.H., Kim,G.R., Seong,M., Baek,S.H., Seol,J.H., Bang,O.S., Ovaa,H., Tatsumi,K., Komatsu,M., Tanaka,K. *et al.* (2007) Two novel ubiquitin-fold modifier 1 (Ufm1)-specific proteases, UfSP1 and UfSP2. *J. Biol. Chem.*, **282**, 5256–5262.
66. Ha,B.H., Ahn,H.C., Kang,S.H., Tanaka,K., Chung,C.H. and Kim,E.E. (2008) Structural basis for Ufm1 processing by UfSP1. *J. Biol. Chem.*, **283**, 14893–14900.
67. Gerashchenko,M.V., Su,D. and Gladyshev,V.N. (2010) CUG start codon generates thioredoxin/glutathione reductase isoforms in mouse testes. *J. Biol. Chem.*, **285**, 4595–4602.
68. Oyama,M., Kozuka-Hata,H., Suzuki,Y., Semba,K., Yamamoto,T. and Sugano,S. (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol. Cell Proteomics*, **6**, 1000–1006.
69. Donahue,T.F., Cigan,A.M., Pabich,E.K. and Valavicius,B.C. (1988) Mutations at a Zn(II) finger motif in the yeast eIF-2 beta gene alter ribosomal start-site selection during the scanning process. *Cell*, **54**, 621–632.
70. Donahue,T.F. (2000) Genetic approaches to translation initiation in *Saccharomyces cerevisiae*. In Sonenberg,N., Hershey,J.W.B. and Mathews,M.B. (eds), *Translational Control of Gene Expression*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 595–614.
71. Nanda,J.S., Cheung,Y.N., Takacs,J.E., Martin-Marcos,P., Saini,A.K., Hinnebusch,A.G. and Lorsch,J.R. (2009) eIF1 controls multiple steps in start codon recognition during eukaryotic translation initiation. *J. Mol. Biol.*, **394**, 268–285.
72. Passmore,L.A., Schmeing,T.M., Maag,D., Applefield,D.J., Acker,M.G., Algire,M.A., Lorsch,J.R. and Ramakrishnan,V. (2007) The eukaryotic translation initiation factors eIF1 and eIF1A induce an open conformation of the 40S ribosome. *Mol. Cell*, **26**, 41–50.
73. Jacobs,G.H., Chen,A., Stevens,S.G., Stockwell,P.A., Black,M.A., Tate,W.P. and Brown,C.M. (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, **37**, D72–D76.