OXFORD JOURNALS
Nucleic Acids Research

Identification of non-AUG initiated, N-terminally extended open reading frames in human genes by an experimentally-informed bioinformatics workflow

Journal:	*Nucleic Acids Research*
Manuscript ID:	Draft
Manuscript Type:	1 Standard Manuscript
Key Words:	translation initiation, eukaryotic initiation factor, non-AUG, open reading frame, alternative initiation codon

Identification of non-AUG initiated, N-terminally extended open reading frames in human genes by an experimentally-informed bioinformatics workflow

Joanne L. Cowan[1], Lisa S. Perry[1], Richard J. Edwards[1,2], David Damerell[3], A. Poppy Roworth[1], Harvey E. Johnston[1] and Mark J. Coldwell[1,*]

[1] Centre for Biological Sciences, University of Southampton, Southampton, SO17 1BJ, UK

[2] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, Sydney, NSW 2052, Australia

[3] School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9QG, UK

* To whom correspondence should be addressed. Tel: +44 023 8059 4342; Fax: +44 023 8059 5159; Email: M.Coldwell@soton.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Present Address: David Damerell, Division of Molecular Biosciences, School of Life Sciences, Imperial College London, London, SW7 2AZ, UK

ABSTRACT

The original hypothesis that a single gene encodes a single protein is a major oversimplification, and the use of alternative promoters or alternative splicing are two possible ways to produce multiple proteins. However, the use of alternative initiation codons to generate N-terminally diverse forms of a protein is less understood. We wished to identify genes where conserved alternative initiation events may lead to more than one protein being synthesised from a single mRNA, and have developed a simple reporter assay coupled with a bioinformatics pipeline to predict candidate genes for in-frame alternative initiation codon usage (i.e. initiation sites that extend or truncate the open reading frame). Following a genome-wide screen of human transcripts, we have successfully identified alternative initiation events other than those assigned by sequence databases in several candidate genes. Here we show that, initiation occurs at non-AUG codons upstream of canonical AUG codons in ST8SIA4 and FBXL3. Following these successful identifications, we have examined the likely consequences for the function of these proteins as determined by their novel extended open reading frames, and the regulation of alternative initiation codon choice by *trans*-acting factors and *cis*-acting sequences within the mRNAs.

INTRODUCTION

In the infancy of molecular biology, it was proposed that one gene gave rise to one protein (1), but it has since become well established that mechanisms to generate a higher diversity of physiologically relevant proteins exist. Multiple proteins can be produced from a single gene by the alternative usage of promoters and splice sites and large scale studies suggest that these are widely used (2,3). However, whether the use of alternative translation initiation codons also generates diversity in the proteome remains a key question in gene expression.

Eukaryotic protein synthesis is comprised of three stages: initiation, elongation, and termination. Initiation refers to the binding of the 40S ribosomal subunit and associated eukaryotic initiation factors (eIFs) to the mRNA, and their positioning at the first codon of the open reading frame (ORF). The “scanning” model postulates that this complex translocates along the mRNA in a 5’ to 3’ direction until it reaches a start codon in an efficient context (GCC[A/G]CCAUGG), where initiation occurs (4). However, many mRNAs contain an upstream AUG (uAUG) in a more favourable (36%) or similar (53%) context (when compared to the consensus) than the presumed AUG (5). Such alternative translation initiation sites control translation in two distinct ways: (a) if the uAUG is not in the correct reading frame it can prevent the main ORF being translated; (b) if the uAUG is in the correct reading frame it may result in alternative isoforms of the protein. Proteins with alternate isoforms produced in this way have been found to (i) be functionally different (e.g. two pore domain potassium (K2P) channels TREK1/TREK2 show altered biophysical properties and ion selectivity (6,7)); (ii) show different subcellular localisation (e.g. BAG-1 (8)) or (iii) retain the same function but with different levels of efficiency (e.g. eIF4GI (9)). In addition to alternative AUG codons, other non-canonical initiation codons can be used for translation (10), indeed we have recently identified a CUG codon that initiates translation of an extended form of initiation factor eIF4GII (11).

Such use of non-canonical near-cognate initiation codons for initiation of translation (i.e. those with one base different within the AUG triplet) was first discovered in viral RNAs (12,13), but it soon became clear that cellular mRNAs could also use near-cognate codons for initiation (14). Until recently, the number of cellular messages known to use non-canonical codons for initiation remained low as they had to be identified on a case-by-case basis. As such, identification of coding sequences in the rapidly expanding genomic and transcriptomic databases mainly relied on computational methods that were limited to solely seeking AUG codons. Recent advances in next generation “deep” sequencing of ribosome-protected sequences stalled at initiation have shown that non-AUG codons are much more commonly used than previously realised (15). While this technology has undergone further refinements and expansion to encompass mammalian systems (16,17), it is still limited in its penetrance of transcriptomic datasets due to (a) only highly expressed transcripts being confidently sequenced, e.g. ~4600 in the first published mammalian dataset (17); (b) tissue/cell specific differences, especially in mammalian systems.

The selection of initiation codons is not solely directed by the sequence immediately surrounding the initiation codon. Firstly, several eIFs have roles in the fidelity of initiation codon selection, with eIF1 and eIF1A causing conformational changes to the scanning 40S ribosomal subunit when an initiation codon is encountered (18,19). In addition, the presence of mRNA secondary structures downstream of an initiation codon in a weak context (or a non-AUG codon) can enhance translation at these sites by inhibiting scanning, therefore pausing the preinitiation complex over the codon (20).

Here we describe a study that combines sequence analysis with further laboratory work to establish instances where alternative translation initiation upstream of the annotated start site is used in human cells to generate more than one protein isoform from a single mRNA. Our approach can use all available transcriptome data and is therefore not limited by tissue/cell type. Furthermore, our criteria for selecting candidates for further study is informed by initial laboratory experiments, thus increasing the stringency of our bioinformatic workflow. This study was initiated independently of, and prior to, publication of a similar study which attempted to identify extended open reading frames by other bioinformatics approaches using a purely *in silico* approach (21). The importance of laboratory verification is exemplified by an N-terminal extension to eIF4GII, which they predicted to be initiated from an AUC codon. However, we independently identified the extension and were able to prove experimentally that it is in fact a CUG codon being used (11).

We show here that our experimentally-informed candidate selection is able to identify cases of alternative translation initiation, which we go on to experimentally verify in cell culture. We further describe the *cis*-acting sequences that can influence non-canonical initiation and demonstrate the effect of *trans*-acting factors that impinge on selection of initiation codons. Finally, we demonstrate the consequences of the presence of the N-terminal extension for our candidates.

MATERIALS AND METHODS

Cell culture, transfection and lysate preparation

All cell culture conditions, transfection and subcellular localisation protocols are as previously published (9,11). Briefly, HeLa cells were maintained in DMEM containing Glutamax (Life Technologies) and 10% FBS and dissociated from culture vessels with Tryple Express (Life Technologies). Transfections were carried out with GeneJuice (Merck), with a 3:1 ratio of GeneJuice to plasmid DNA. Total cellular protein for immunoblotting was extracted using M-PER lysis buffer supplemented with Halt protease and phosphatase inhibitor cocktail (Pierce) as per the manufacturer’s instructions. For protein turnover experiments, cycloheximide at a final concentration of 10µg/ml was added for the times indicated prior to harvest. Antibodies were from the following commercial suppliers: FLAG-M2 affinity purified and α-actin, Sigma-Aldrich (Dorset, UK); Cry1 and β-actin, Proteintech (Manchester, UK); Golgin 97, Life Technologies (Paisley, UK). Immunoblots were visualised by ECL or Licor Odyssey as previously described (22), with quantification undertaken using Licor Image Studio Lite.

Creation of the pICtest2 reporter vector

To obviate any issues with co-transfection efficiency of separate reporter plasmids, we created a dual luciferase vector, which could be assayed quickly and easily. All PCR steps used Pfu Ultra, a high fidelity polymerase, and the integrity of plasmids was checked by restriction enzyme assay and sequencing. All primer sequences are detailed in Supplementary Table S1A. Initially, a backbone was created from the Promega pGL3-control vector, amplified using pICt2 stage1 F and pICt2 stage1 R. This amplicon was digested with XbaI, prior to ligation. This created a backbone plasmid with an SV40 promoter, and SV40 polyadenylation and enhancer sequences. Next, the HSV-Thymidine Kinase promoter from psiCHECK2 (Promega) was amplified with HSV-TK Bam F and HSV-TK Bgl R. This amplicon was digested with BamHI and BglII and ligated into the backbone plasmid digested with BamHI, with the insertion of the promoter in the correct orientation verified by restriction digest. Subsequently, a cassette containing a chimeric intron, the synthetic *Renilla* luciferase (hRluc), and a synthetic polyadenylation signal was amplified from psiCHECK2 with hRen Sal F and hRen Sal R. This cDNA was inserted downstream of the HSV-TK promoter by digesting the vector and amplicon with SalI, with correct orientation again confirmed by restriction digest. Previously, a modified pGL3 control plasmid had been generated where a sequence with near homology to the PEST domain of ornithine decarboxylase (SSGTRHGFPPEVEEQAAGTLPMSCSQESGMDRHPAACASARINV) was fused to the C-terminus of the firefly luciferase (M.Coldwell, unpublished) as others have previously achieved (23). This cDNA was amplified with a common reverse primer (LUCp R) and alternative forward primers (Supplementary table S1A), depending on the initiation codon/context variant that was being tested. These amplicons were digested with NheI and XbaI, then cloned into the complete backbone which had been digested with XbaI. Correct orientation of the insert was confirmed by restriction digestion, prior to automated sequencing.

Reporter assays

At the times indicated in the text below, expression of firefly and *Renilla* luciferase was quantified using the Dual-Luciferase® Reporter Assay system from Promega. Growth media was aspirated from wells and washed with D-PBS; 20 µl of passive lysis buffer was added per well and incubated at room temperature for 15 minutes with agitation, after which 5 µl of each lysate was transferred to an opaque white 96-well plate. A GloMax®-Multi+ Detection System (Promega) was used to assay the expression of the two reporters in each well determined by injection of 25 µl of Luciferase Assay Reagent II, with a 2s delay and 10s luminescence read time, followed by injection of 25 µl of Stop & Glo reagent, a further 2s delay and 10s luminescence read. Each plasmid was tested in triplicate on at least three independent occasions.

Cloning of initiation factors and siRNA plasmids

cDNAs corresponding to the open reading frames of eukaryotic initiation factors were amplified using the primer pairs detailed in Supplementary Table S1B, with each subsequently cloned into pcDNA3.1(+) from Invitrogen or pcDNA myc (24). To knock down expression of the proteins of interest, we employed the same siRNA hairpin-expressing plasmid as previously used (9,11). pSilencer 3.0H1 (Ambion) was linearised by digestion with BamHI and HindIII, and annealed oligonucleotide pairs (Supplementary Table S1B) with compatible overhangs were ligated into the backbone.

Cloning and site directed mutagenesis of candidate genes

cDNAs of interest were purchased from the IMAGE clones collection (Open Biosystems), as vectors containing a cDNA corresponding to the transcripts identified by our bioinformatic searches. Initially, we amplified the whole 5' UTR and a portion of the ORF, with forward primers containing NheI restriction sites, and reverse primers containing XhoI so that the ORF was introduced in frame with a C-terminal 3x FLAG tag, as previously described (11). Primer pairs used to amplify candidate genes are described in Supplementary Tables S1C and D. Further primers were designed to modify predicted and annotated initiation codons to confirm alternative initiation events, and these sequences are described in Supplementary Tables 1C and D, as are primers used to mutate predicted hairpin structures.

Sequence data

All human and murine transcripts were compiled from Ensembl (http://www.ensembl.org). 5' UTR sequences were translated in-frame with the downstream open reading frame using a Perl script, and conservation of human and murine translated 5' UTRs analysed by BLASTp sequence alignment, with further analysis carried out using tBLASTn.

RESULTS

Use of reporter vectors to determine initiation codon efficiency

Previous identification of alternative initiation codons in transcripts have used methods that work on a case-by-case basis. While these have proven successful, it would be difficult to scale up this work when attempting to identify alternative initiation events on a genome-wide scale. While identifying possible in-frame initiation codons upstream and downstream of those annotated in databases is straightforward, we wished to refine our searching methodology using laboratory data. Given the huge number of possible candidates, we thought it important to focus the search on those candidates where the likelihood of identifying such events is increased.

To this end, we created a dual luciferase reporter plasmid that enables quantification of initiation efficiency at either AUG codons in different contexts or non-AUG codons. We constructed a vector, named pICtest2 (plasmid for Initiation Codon testing, Figure 1A) where the luciferase from the sea pansy *Renilla reniformis* would act as an internal control, always being translated from an AUG in the optimal Kozak consensus GCCACCAUGG. Transcription of the *Renilla* reporter mRNA is driven by the HSV Thymidine Kinase promoter. Within the same backbone, so as to obviate any issues which may arise from inefficient cotransfection, luciferase from the firefly (*Photinus pyralis*) is expressed from the SV40 promoter, and it is the initiation codon of this open reading frame which is altered as detailed below. We have also fused the C-terminus of the firefly luciferase ORF to the PEST domain of ornithine carboxylase, so called because it is rich in proline, glutamate, serine and threonine residues (25). The presence of this domain destabilises the luciferase (23) and allows rapid changes in translation to be monitored. In terms of translation efficiency, both untranslated regions are simple, with the 5' UTR being 71nt long, with a G-C content of 42.3%, both of which will allow efficient scanning. The 3' UTR is 170nt upstream with 41.9% G-C content.

The chemistry of the dual luciferase assay means that both luciferases can be rapidly assayed in the same sample, and the sensitivity of the luciferase assay compared to those done with e.g. chloramphenicol acetyl transferase (CAT) will allow us to assay even relatively inefficient initiation. Furthermore, the use of firefly luciferase as our reporter means that N-terminally truncated enzymes generated by leaky scanning are very unlikely to be measured in the assay, so all values truly reflect initiation efficiency at the codon of interest. Loss of the first 10 amino acids reduces the level of firefly luciferase activity to <1% (26) and the next in frame AUG is at the 30th position. Likewise, the only in-frame near-cognate codon with an ability to drive translation (Figure 2) is a CUG at the 17th position. Furthermore, we established "OPP" as a control, where the initiation codon was substituted by UAC as this triplet is the "opposite" to an AUG and would be extremely unlikely to base pair with the UAC anticodon of the initiating methionyl-tRNA. This was used to establish a background threshold, below which activity of the initiation codon was considered nil. In addition, stop codons in this position (UGA, UAG and UAA) also produced similar results (data not shown).

To establish the relative efficiency of initiation codon contexts, our initial experiment varied the context of the simplest minimal Kozak consensus (RCCAUGG). As previous work had established the bases at -3 and +4 as being those which influenced efficiency of translation (4), we sought to confirm these findings. Combinations of each possible nucleotide at these positions were assayed following transfection of our pICtest variants into either HeLa or HEK293 cells, with firefly luciferase activity normalised to *Renilla* luciferase activity (Figure 1B). As well as showing the relative efficiencies of each combination, Figure 1B also shows both the proportion of AUG initiation codons annotated in all human transcripts in the Ensembl database that have the indicated bases at those positions.

As expected, the presence of the G at +4 is the most important nucleotide in determining whether an initiation codon is recognised. In reporters with this context, the presence of a pyrimidine base results in a reduction in the efficiency of translation, which is not statistically significant in the HeLa cell assay, although it is in HEK293 cells. More importantly, in reporters where the +4 base is anything but a G, the presence of a purine at -3 becomes much more essential. For example, with an A at +4 and a purine at -3, translation is reduced to around 40% of the optimal consensus, which implies that 60% of scanning ribosomes do not recognise the initiation codon and instead carry on downstream. If the +4A is coupled with a -3 pyrimidine, then translation of firefly luciferase drops further to around 20% of the optimal consensus. The least efficient combination +4U/-3U, initiation occurs for only 5% of scanning ribosomes.

To simplify our understanding of these results, we can therefore consider an initiation codon within a +4G context to be classified as "strong" (>60% efficiency), and those with -3C/+4C, -3C/+4A, -3U/+4A,

-3U/+4U as “weak” (>30% efficiency), with all other combinations having a “mid” efficiency. When the tally of transcripts with each classification is taken into account (Figure 1C), then it becomes obvious that the majority of initiation is likely to be inefficient or “leaky”, with possible consequences being the production of a vast amount of proteins with N-terminally truncated open reading frames (tORFs). This is further quantified in Figure 1D, where we have used those transcripts classified as “weak” and determined the position of the next likely downstream initiation codon, which may be used as an alternative start site. These may be another AUG, or an efficient near-cognate triplet, based on our below findings (Figure 2A).

We also used our pICtest reporter system to analyse the efficiency of near-cognate non-AUG initiation, in order to help us limit our searches for alternative initiation events to those that are most likely to be used. For this, we also compared the extended full Kozak consensus with that proposed for non-AUG initiation by Wegrzyn et al (27). Their sequence (CGCGUCGCGxxxG) was determined by comparing the context surrounding 43 published non-AUG initiation codons. We compared the two contexts to each other, with initiation codons being AUG or the nine near-cognate initiation codons, which differ by one nucleotide at each position of the triplet. As previously stated, we considered a UAC codon to be a negative control so only luciferase activity greater than that from this reporter was considered as being from a full length protein. In the vast majority of cases, there was barely detectable initiation from the near-cognate codons, in either the Kozak or Wegrzyn consensus (Figure 2A) although all bar AAG exhibit firefly luciferase expression that is significantly higher than the UAC/OPP control in at least one context and one cell line. It can also be seen that CUG and GUG initiation codons do initiate with an efficiency that is similar to, if not better than, values obtained for an AUG in a weak context (compare to Figure 1B). The only initiation codon that benefits from being within the Wegrzyn context is an ACG, with both CUG and GUG codons weaker when in this sequence.

Our data suggest that we can now consider a CUG initiation codon as a particularly viable alternative to an AUG codon for initiating translation, and this codon has previously proven the most likely to be discovered (e.g. 36 of the 43 used in the Wegrzyn study (27)). Therefore, if we now search upstream of annotated initiation codons for in-frame CUG codons, it can be seen that there are a large number of possible candidates which may have an extended CUG-driven form (Figure 2B), as is the case with eIF4GII and BAG-1. Our other most likely candidates for upstream alternative initiation are GUG and ACG, and again we predict that there are large numbers of transcripts that initiate translation at such codons.

Coupled with our ability to measure the efficiency of initiation with our reporter assay, we can use it to assay the effects of *trans*-acting factors upon initiation codon selection. Initiation factors associated with the fidelity of translation initiation are eIF1, which promotes the disassembly of 48S complexes positioned at an incorrect initiation codon, and eIF1A, which is important for initiation codon recognition (28). Two eIF1A genes are present in humans, one on the X-chromosome (eIF1AX) and the other on the Y (eIF1AY), with their respective primary amino acid sequences identical save for a methionine at position 50 in eIF1AX, which is a leucine in eIF1AY.

Vectors were constructed to overexpress either eIF1, eIF1AX or eIF1AY from the pcDNA3.1(+) backbone and these were cotransfected with pICtest vectors to assay how their overexpression altered selection of alternative initiation codons (Figure 2C). Values are expressed compared to control cells cotransfected with the empty backbone, with this value set to 1 for each initiation codon. Overexpression of eIF1 significantly inhibits expression from CUG and GUG initiation codons, while having no effect on AUG. There is a suggestion that UUG-driven translation is more efficient, but this is from such a low base line (see Figure 2A) that even a slight change will be amplified. Likewise, there was no significant difference in alternative initiation codon selection of the reporters when either eIF1AX or eIF1AY were overexpressed, although AUG initiation is slightly impaired in cells overexpressing eIF1AX.

Identification of candidate genes with possible upstream AICs

Having established that CUG, GUG and ACG triplets were the most likely alternative initiation codons that would produce upstream in-frame initiation, we now set out to probe the entire human genome for candidate genes where this phenomenon was possible. Our first bioinformatics screen used the Ensembl genome database (release 48) to identify all human transcripts with 5' UTR sequence data that contain a possible N-terminally extended ORF (eORF). 5'UTRs were translated in frame with the subsequent annotated open reading frame and all those that were >40 amino acids long were retained. This value was chosen purely on the basis that such an extension would be easier to verify by immunoblotting (see later) and more likely to contain motifs/domains which would confer extra function to the extended form, as we have previously seen with eIF4GI isoforms (9).

The subsequent dataset was screened to remove hits resulting from inappropriately annotated entries and splice variants where a known coding sequence had been listed as belonging to a 5' UTR. The remainder (~ 8000) were then searched against a similarly processed translated 5' UTR dataset from murine 5' UTR sequences. This identified a subset of 444 candidates with highly conserved N-terminal extensions, indicating that the eORF may also be functionally important. This dataset is shown in Supplementary Table S2. Many show indications of ORF annotation that has been inferred computationally and have either non-specific sequence-based annotations (e.g. transmembrane or SH3 domains) or are completely unknown/hypothetical proteins. This suggests that there is likely to be a substantial proportion of such identifications/annotations that lack experimental evidence and have been made using rules that neglect the possibility of non-AUG initiation.

The full list of 444 candidates would be impractical to screen in a laboratory setting, so we used tBLASTn to query all translated mRNA sequences with the amino acid sequence of the predicted extensions (Supplementary Table S2). As would be expected, the vast majority have homology to other primate sequences, but in 142 cases, there is conservation with more distantly related mammals, and in some cases homology is predicted with other chordates including *Xenopus laevis*, *Gallus gallus* and *Danio rerio*.

We now wished to proceed by screening for alternative initiation events using the same pcDNA-based 3x FLAG tag-based vector we had previously used (Figure 3A). To begin the process of screening, we now analysed each gene on a case-by-case basis with a Visual Basic-based macro (Supplementary File S3), executed in Microsoft Excel, which could compile EnsEMBL human 5’ UTR and coding sequence data. It annotates positions which have in frame upstream CUG, GUG and ACG initiation codons in a strong Kozak consensus and denotes the relative efficiency of any in-frame AUG codons, identifying possible eORFs and tORFs in all corresponding human transcripts in the Ensembl database. As such, we have nicknamed this macro our ExTATIC Identifier (EXtensions and Truncations from Alternative Translation Initiation Codons).

To aid in screening, and abrogate the need to identify multiple suitable cell lines/tissues for amplification of each of the candidate cDNAs, we chose to only follow up candidates available for purchase as full length cDNA IMAGE clones. For each candidate, primers were designed to amplify the whole 5’UTR and at least 300 nt more than the most 3’ possible alternative initiation codon from the purchased clone. Primers were also designed to allow restriction digestion and ligation into the vector (Figure 3A).

We have now screened a number of candidates and identified novel upstream initiation events, and herein we describe research on two of these candidates while others will be the subject of future publications.

Multiple isoforms of ST8SIA4 arise through alternative initiation, with consequences for subcellular localisation

Firstly, we focused on α-N-acetyl-neuraminide α-2,8-sialyltransferase, encoded by the ST8SIA4 gene. This is a polysialic acid transferase (PST) which catalyses the addition of neuraminic acid derivatives to N-glycan residues on polypeptides (29-31). A number of potential alternative initiation codons were predicted by our searches (Figure 3B) and so the 5’ UTR and part of the annotated coding sequence were subcloned into our reporter vector. In order to fully determine whether alternative forms of ST8SIA4 did indeed arise from the predicted alternative initiation codons, we also created versions where the 5' UTR was truncated at a predicted initiation codon and an AUG introduced in place of the AIC to enhance translation. These versions are shown in Figure 3C, with the expected migrations of peptides translated from the predicted and annotated initiation codons shown therein. Conversely, we also used site-directed mutagenesis to change the predicted CUG codons at -225 and -216 to two UAC codons, thus inhibiting initiation.

The wild type and modified plasmids were transfected into HeLa cells for 48 h and, as can be observed in Figure 3D, a number of FLAG-tagged peptides are detected in total protein extracts from cells transfected with the WT ST8SIA4 plasmid (lanes 2 and 7), suggesting we are indeed observing translation at the upstream alternative initiation codons. Lane 3, where the CUG at -225 has been changed to an AUG, clearly shows that the upper band is enhanced compared to the wild type (WT) in lane 2. The loss of this band in the -225-216 CUG-UAC mutation shown in lane 8 further confirms that novel initiation occurs in this region.

When analysing initiation from the other predicted codons, the upper band in lane 4 (-66 enhanced AIC) is observed to comigrate with the top of the blurred band in the WT lane with the upper bands in the -33 and +1 lanes (5 and 6 respectively) also mapping to this smear in the WT lane. This implied it may be a triplet band. At each intermediate truncation the bands above were lost as would be expected but it is clear from all the truncations that at least one smaller band is observed below, which actually runs closer to the expected weights than the higher, more intense bands. This suggests that each species may be subject to post-translational modification(s), resulting in an addition of approximately 10kDa.

Like the rest of the sialyl transferase (ST) family, ST8SIA4 is a type II transmembrane protein present in the *trans*-Golgi network. Given the importance of N-terminal protein sequences in directing protein targeting to the secretory network *via* the endoplasmic reticulum, we wanted to identify any consequences for the extended protein isoforms. Full length coding sequences corresponding to -225, -66 and +1 ST8SIA4 isoforms were amplified, fused to the C-terminal 3x FLAG tag and expressed in HeLa cells as before. Localisation of the FLAG-tagged proteins was examined by immunofluorescence (Figure 3E), showing that proteins expressed from all alternative initiation codons exhibit a localisation characteristic of the Golgi body (compare staining of Golgin 97 in panel i to FLAG staining in panels ii-iv). However, some localisation of the -66 and -225 form of ST8SIA4 (panels iii and iv) is also markedly more cytoplasmic, suggesting that these forms do not compartmentalise as efficiently, and perhaps also play a role elsewhere in the cell.

Translation of an extended isoform of FBXL3

Another candidate identified in our bioinformatic pipeline was FBXL3. F-box proteins give specificity to the Skp, Cullin, F-box protein (SCF) E3 ubiquitin ligase complexes that target and label proteins with ubiquitin, the first step in protein degradation. FBXL3, originally known as FBLW3, contains an F-box domain as well as several leucine-rich-repeat regions (LRRs). Its function was first identified when two independent mutations in FBXL3 were found to lead to increases in the circadian rhythm period (32-34). In “overtime” mice, the FBXL3 gene has a point mutation that changes isoleucine 364 to a threonine, whereas in “afterhours” mice, cysteine 358 becomes a serine. FBXL3 specifically targets cryptochrome (Cry) proteins, a key component of the negative feedback loop giving accurate periodicity to an organism’s circadian rhythm.

There are two predicted novel upstream initiation codons in the FBXL3 5’ UTR, a GUG and an ACG (Figures 4A, 4B). As before, the 5’ UTR and a portion of the FBXL3 coding region were inserted upstream of a 3x FLAG tag, and expression from this plasmid was examined following transfection into HeLa cells. Figure 4C shows that only two forms of FLAG-tagged protein are detected (lane 3), which correspond to translation from the GUG and annotated AUG only. Given our previous data showing only weak initiation at an ACG codon (Figure 2A), it is perhaps not surprising that in this context, translation from this codon is not observed.

The putative GUG was subjected to site-directed mutagenesis to again enhance or inhibit translation from this site. As expected, mutation to AUG increased translation, and in doing so prevents leaky scanning to the downstream AUG (Figure 4C, lane 4). Interestingly, translation of the upper band was reduced but still observed in the GUG-UAC mutation (lane 5). Given that the triplet upstream of the GUG is also a near-cognate codon AAG, it is possible that limited translation may occur at this point. This is confirmed by two further mutations. The GUG-UAA variant would mean that there was an in-frame stop directly downstream if the AAG was being used, hence why no translation of the upper band is observed (lane 6). Similarly, mutation of both possible near-cognate codons to UAC also results in the loss of the upper band (lane 7).

Given that in our firefly luciferase reporter assay (Figure 2A), expression from an AAG was virtually nil and that from GUG was also very weak compared to an AUG, we found it intriguing that expression of the upper band remained reasonably robust. This suggests that the region surrounding, rather than the codon itself, may play a role in initiation. Furthermore, as previous work had shown that the presence of a hairpin positioned at the end of the leading edge of an initiating ribosome could enhance translation from weak initiation codons (20), we explored the possibility that a similar structure may be influencing translation at the FBXL3 alternative initiation site.

Using five mammalian FBXL3 sequences (*Homo sapiens, Pan troglodytes, Pongo abelii, Mus musculus* and *Bos taurus*), we used the Centroidfold server (35) to predict the consensus secondary structure that may form between the upstream GUG and annotated AUG initiation codons (Figure 4D), which includes a hairpin that would be positioned at a suitable point in the mRNA. This hairpin alone has a predicted minimal free energy of ΔG = -24.4 kcal/mol, which is more stable than the one utilised in Kozak's study of ΔG = -19 kcal/mol (20). Further downstream structures are also predicted to form within the FBXL3 mRNA, which may block ribosomal scanning and suggests why when the mutations which reduce initiation at the AAG/GUG position (Figure 4C, lanes 5-7), there is no concomitant increase in the expression from the downstream AUG.

Relaxing the predicted hairpin that we postulated as being important in enhancing translation from the upstream AAG/GUG position was limited because we could only change nucleotides in silent positions in order to maintain the open reading frame of the extended FBXL3 protein. We created mutations in the left and right stems at the base of the first hairpin (denoted by red and blue arrowheads, respectively), as well as a version where mutations in both sides were combined. The effect on translation of the upstream initiation codons is shown in Figure 4E, with mutation of the left stem and both stems together reducing the expression of the extended form (lanes 3 and 5). The minimal free energy of these forms were respectively -14.6 and -13.7 kcal/mol, while the right stem mutant was -19.2 kcal/mol which perhaps explains why there was no reduction in translation of the upper bands (lane 4).

Having established the importance of this *cis*-acting sequence on the non-AUG driven translation of FBXL3, we wanted to establish whether *trans*-acting factors can also influence the initiation codon choice. As well as vectors overexpressing eIF1 and eIF1AX, we constructed plasmids which would be used to express siRNAs in order to knockdown their expression in HeLa cells over 96 hours, to allow enough time for initiation factor turnover. As can be seen in Figure 4F, and quantified in Figure 4G, knockdown of eIF1 increases relative expression from the non-AUG initiation codons, as would be expected when removing the stringency of selection from the translation machinery. In contrast, overexpression of eIF1 reduces the ratio of non-AUG:AUG translation, a consequence of greater stringency in the system in agreement with our earlier results (Figure 2C).

The extended form of FBXL3 has different degradation kinetics

As FBXL3 plays a key role in the degradation of proteins involved in circadian rhythms, we now wished to establish what the consequences of the two different isoforms were. To this end, the full open reading frames of the extended and annotated forms were fused to the C-terminal 3x FLAG tag (Figure 5A), for expression in HeLa cells. After 72 h, cells were treated with cycloheximide to inhibit translational elongation and allow the turnover of the different FBXL3 proteins and their endogenously expressed cryptochrome targets to be examined. The poor quality of the antibody prevented us interpreting the turnover of Cry2 (data not shown), but the levels of the FLAG-tagged FBXL3 and Cry1 could be assayed and compared to an actin loading control.

A previous study (34) examined transfected FBXL3 levels after transfection of the annotated form in HEK293 cells, and suggested that the protein’s half-life was over 7 h (the duration of their experiment). In our experiments the original, AUG-driven form of FBXL3 indeed appears to have a half-life of around 24 h (Figure 5B, C). However, the extended form appears to be initially more stable, and is unusual in that levels of this protein appear to increase during the first 8 h of cycloheximide treatment, before dropping over the next 16 h to a similar level to that seen in the annotated form.

The most striking results are seen when examining levels of Cry1, one of two cryptochrome proteins with a role in forming complexes with the Period proteins, which go on to inhibit transcription of Clock genes. It is important that accurate degradation of the cryptochrome proteins is carried out effectively, in order to allow reactivation of Clock in the next circadian period. Previous experiments examining the turnover of HA-tagged Cry1 show that in cells lacking FBXL3, the half-life of HA-Cry1 is 6.4 h. Our work shows that endogenous Cry1 in control HeLa cells transfected with the empty pcDNA3.1(+) vector does degrade over time, although we were not able to accurately determine its decay rate (Figure 5Bi, D), as protein levels, normalised to actin, fluctuate over the timecourse.

Here we show that there is a large difference in turnover of Cry1 when the different forms of FBXL3 are overexpressed. In cells transfected with the annotated p52 form of FBXL3, Cry1 appears to be more stable than in untransfected cells. This disagrees with previous experiments expressing the p52 annotated form of FBXL3, which found that Cry1 half-life was reduced, and measured it as 1.7 h (34). This group did also cotransfect HA-ubiquitin, although we would assume that endogenous ubiquitin in the HeLa cells would be used in our experiment. What may instead be happening is that a further component of the SCF complex is not recruited to this form of FBXL3 in HeLa cells. In stark contrast, when cells are transfected with the extended p58 form of FBXL3, accelerated turnover of Cry1 is observed (Figure 5B, D). Taken together, our results suggest that the 64 amino acid extension in the N-terminus of p58 FBXL3 may be involved in both increasing stability of FBXL3 itself and increasing the turnover of its substrate, perhaps by enhancing SCF complex formation. From a mechanistic point of view, use of the upstream initiation codon in preference to the annotated initiation codon could therefore conceivably reduce the period of the circadian rhythm.

DISCUSSION

In this study, we have established an experimentally-informed pipeline to identify novel non-canonical initiation events, and determined aspects of their regulation. Furthermore, we have investigated the consequences of expressing extended forms of some of these proteins.

We first recapitulated work by others to establish the relative efficiency of different initiation codons and contexts using a reporter-based assay. According to our findings, a high number of annotated initiation codons are relatively inefficient at driving translation suggesting there may be a great deal of “leakage,” producing truncated proteins from a downstream initiation codon. Likewise, we confirm that some near-cognate codons can be used to initiate translation with a reasonable efficiency, which makes them likely to be used in cellular mRNAs. Our data with the firefly luciferase reporter mRNA should be considered alongside the knowledge that the presence of secondary structure downstream of an otherwise poorly-recognised initiation codon can increase the efficiency of its use (20). In the firefly luciferase mRNA, there are 85 nt between the two AUGs; the sequence comprises only 52.4% G-C, and is hence unlikely to form a stable structure. A palindromic hairpin of $\Delta G = -19$ kcal/mol inserted 15 nt downstream of the initiation codon of the CAT reporter can enhance translation from a UUG (20), as it is positioned directly at the leading edge of the ribosomal footprint. Unfortunately, there is limited scope for inserting a similar hairpin in the firefly luciferase coding sequence due to a lack of suitable positions for creating silent mutations that ensure maintenance of enzymatic activity, therfore such experiments cannot be pursued further with our reporter system.

We then used our knowledge of the most likely near-cognate initiation codons to inform and narrow down our searches for non-canonical initiation events upstream of annotated initiation codons. Our pipeline began with stringent criteria in our initial dataset, in that we were only aiming to identify extensions of 40 amino acids or more. While this considerably narrowed our field, it does mean we have ignored a number of candidates with possible extensions that are shorter (e.g. 7614 possibly using an upstream CUG codon, as shown in Figure 2C). For example, the upstream CUG which is used for translation of the p67 form of c-Myc is only 15 codons upstream of the AUG responsible for translation of the p64 polypeptide (36). Only considering three of the most efficient near-cognate codons also adds an extra layer of rigour to our searches, rather than purely considering any near-cognate initiation codon, as undertaken in a similar previous study carried out purely *in silico* (21). As mentioned before, that approach meant an AUC was postulated as an alternative upstream initiation codon in eIF4GII, whereas our own work demonstrated that a nearby CUG is responsible for translation of an extended form (11). However, there is a good degree of agreement between our datasets, with 19 of the 42 genes identified by Ivanov *et al* (21) also found in our long-list (Supplementary Table S4).

The advent of ribosomal profiling technology has enabled the identification of ribosome-protected fragments by deep sequencing methods (15), and this has had a profound impact on our knowledge of which open reading frames are translated under the experimental conditions studied. The use of compounds such as harringtonine and emetine to arrest ribosomes at the point of initiation has enabled the large-scale identification of novel initiation codons in mammalian systems (17,37). For example, initiation events have been identified in around 5000 mRNAs from mouse embryonic stem cells (17), demonstrating the presence of N-terminal extensions and truncations (compared to annotated initiation codons) in approximately 15% of mRNAs. When we interrogate this dataset and compare our predictions (Supplementary Table S4), we find that two of our candidates have indeed been found to exhibit initiation from an upstream AUG, with 48 others found to initiate translation from a near-cognate initiation codon. In addition, 114 of our predictions can be found in the dataset from Lee et al. (37), which identified initiation events in regions annotated as 5'-UTRs in HEK293 cells, and 32 of our candidates have been confirmed in both ribosome profiling studies. Most importantly for the work we present here, translation from the upstream GUG in the endogenous FBXL3 transcript is confirmed in the Lee dataset, thus validating our own findings (Figures 4 and 5).

Our experiments have used the construction of truncated variants and site directed mutagenesis to confirm the presence of alternative initiation codons (Figures 3D, 4C), and we can see from the work with FBXL3 that a downstream secondary structure aids in translation from what would otherwise be weak initiation codons (Figures 4D, 4E). We have also examined the consequences for the function of the proteins of interest, with extended forms of ST8SIA4 having an altered subcellular distribution. In FBXL3, the extended form has altered turnover kinetics that influence accelerated degradation of its cryptochrome substrate (Figure 5).

The work presented here also demonstrates that the levels of the *trans*-acting initiation factors eIF1 and eIF1A can affect translation at initiation codons (Figures 2C, 4F, 4G), making it clear that changes in availability/activity of such factors would allow a degree of dynamic fine-tuning to translation depending on cellular conditions. Indeed, the early work with c-Myc showed that there was more substantial expression of the CUG-initiated p67 form in cultures which had been maintained at higher densities (38), perhaps mimicking tumour growth.

As well as eIF1 and eIF1A, several other *trans*-acting factors play a role in translation initiation efficiency, and some of these are considered non-canonical, i.e. not part of the standard AUG-and cap-dependent models of translation initiation. The eukaryotic initiation factors eIF2A and eIF2D (39,40) have been implicated in selection of non-canonical initiation codons due to their ability to bind to tRNAs other than the initiator Met-tRNAi, which is the only tRNA that is present in the canonical ternary complex with eIF2 and GTP. We have also generated siRNA and overexpression plasmids for these factors, although changes in their levels do not appear to influence the use of the GUG initiation codon in either the firefly luciferase reporter or FBXL3 (data not shown). Furthermore, we have also been unable to recapitulate work showing the role of eIF2A in CUG codon initiation (41), although their experimental approach and background was different to our own. However, in work that accompanies and complements this study, we have shown that ABC50 (an ATP-binding cassette protein which interacts with eIF2 (42)) does influence translation at non-AUG codons, in particular increasing translation from the GUG codon in FBXL3 and the firefly luciferase reporter (Stewart, Cowan, Perry, Coldwell and Proud, co-submitted with this manuscript).

It is clear from our work and the work being undertaken in other laboratories, that the identification of alternative initiation codons is an expanding and exciting field. Taken together, our bioinformatics models and case-by-case validations coupled with the profiling data of others suggest a substantial portion of the proteome may be subject to non-canonical initiation. Given the reliance on accurate models of translation initiation for the annotation of protein-coding genes, this finding has wide-ranging implications for genome annotation and proteomics studies. Indeed, others are taking the ribosome profiling work further by pairing it with an N-terminal proteomics approach to confirm that such extensions exist within the cellular proteome (43). These new findings need to be integrated into the annotation of genomes, as even with the latest Ensembl release (74, November 2013), only 38 transcripts have annotated non-AUG initiation codons, versus 7118 with “Weak” AUGs according to our classification (Figure 1C). In contrast, over 19,000 transcripts have a potential in-frame 5’ non-AUG alternative initiation codon (Figures 1D and 2B). Supplementary Table S5 contains the details of these predicted alternative initiation events for each transcript.

As additional novel N-termini are discovered, their functions will be elucidated and our own efforts are currently focussed on several other candidates informed by this study, and how their alternative isoforms determine subcellular localisation and binding partner capacity. As more becomes known about this phenomenon, it is likely that we will discover, along with the use of alternative splice sites and alternative promoters, that alternative translation initiation is also an important mechanism for generating proteome diversity from a limited set of genes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

Supplementary Table S1: Primer sequences

Supplementary Table S2: Extended amino acid sequences with high homology to murine sequences

Supplementary File S3: The ExTATIC identifier Excel macro for case-by-case identification of alternative initiation sites

Supplementary Table S4: Comparison of this dataset with other methods for identifying upstream translation initiation events

Supplementary Table S5: Predicted alternative initiation codons up- and downstream of annotated initiation codons.

ACKNOWLEDGEMENTS

Many have contributed to this wide-ranging project at various stages. Key advisors at the University of Sussex on the initial bioinformatic workflow were Sue Jones and Ruth Spriggs. Lucinda Eaton, Samantha Hodges and Kevin Jones all contributed to construction of vectors for the eIF1 and eIF1A studies. The first generation of pICtest vectors were constructed and tested by Sarah Boreham, Rebecca Brown, Perin Celebi, Stephanie Chatfield, Leanne Conyard, Sophie Flynn, Louise Gaffney-Wilkinson, Katherine Goodman, Mike Langford, Ciaran McVeigh, Rebecca Pook and Sandraneika Young. The long list of candidate genes was narrowed down by Priyanka Bhatt, Tejal Bhatt, Meowea Hezwani, Alice Jamieson, Harvey Johnston, Charlotte Lawrance, Jessica Price, Maria Spyrou and Charlotte Yates. Other candidate genes have been tested by David Barreto Ian, Louise Ellis, Rachel Langford, Ella Mengham, James Pethick, Natalia Pokrovskaia, Aaron Purvey, Emma Sweetman, and Sabrina Safi-Kantanka. Finally, thanks go to Chris Proud for helpful comments on the manuscript.

FUNDING

This work was supported by a New Investigator Research Grant from the Biotechnology and Biological Sciences Research Council (BB/H006834/1 to M.J.C). Preliminary funding for the project was supplied by a University of Southampton Adventure in Research Grant (A2008/06 to M.J.C. and R.J.E). L.S.P. was funded by a BBSRC Doctoral Training Award. Funding for Open Access Charge, BBSRC.

REFERENCES

1. Beadle, G.W. and Tatum, E.L. (1941) Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A*, **27**, 499-506.
2. Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep*, **2**, 388-393.
3. Takeda, J., Suzuki, Y., Nakao, M., Barrero, R.A., Koyanagi, K.O., Jin, L., Motono, C., Hata, H., Isogai, T., Nagai, K. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res*, **34**, 3917-3928.
4. Kozak, M. (1989) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol Cell Biol*, **9**, 5073-5080.

5. Peri, S. and Pandey, A. (2001) A reassessment of the translation initiation codon in vertebrates. *Trends Genet*, **17**, 685-687.

6. Simkin, D., Cavanaugh, E.J. and Kim, D. (2008) Control of the single channel conductance of K2P10.1 (TREK-2) by the amino-terminus: role of alternative translation initiation. *J Physiol*, **586**, 5651-5663.

7. Thomas, D., Plant, L.D., Wilkens, C.M., McCrossan, Z.A. and Goldstein, S.A. (2008) Alternative translation initiation in rat brain yields K2P2.1 potassium channels permeable to sodium. *Neuron*, **58**, 859-870.

8. Packham, G., Brimmell, M. and Cleveland, J.L. (1997) Mammalian cells express two differently localised Bag-1 isoforms generated by alternative translation initiation. *Biochem J*, **328**, 807-813.

9. Coldwell, M.J. and Morley, S.J. (2006) Specific isoforms of translation initiation factor 4GI show differences in translational activity. *Mol Cell Biol*, **26**, 8448-8460.

10. Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A.C. and Vagner, S. (2003) Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol Cell*, **95**, 169-178.

11. Coldwell, M.J., Sack, U., Cowan, J.L., Barrett, R.M., Vlasak, M., Sivakumaran, K. and Morley, S.J. (2012) Multiple isoforms of the translation initiation factor eIF4GII are generated via use of alternative promoters, splice sites and a non-canonical initiation codon. *Biochem J*, **448**, 1-11.

12. Curran, J. and Kolakofsky, D. (1988) Ribosomal initiation from an ACG codon in the Sendai virus P/C mRNA. *EMBO J*, **7**, 245-251.

13. Prats, A.C., De Billy, G., Wang, P. and Darlix, J.L. (1989) CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus. *J Mol Biol*, **205**, 363-372.

14. Peabody, D.S. (1989) Translation initiation at non-AUG triplets in mammalian cells. *J Biol Chem*, **264**, 5031-5035.

15. 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.

16. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. and Weissman, J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc*, **7**, 1534-1550.

17. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789-802.

18. Lorsch, J.R. and Dever, T.E. (2010) Molecular view of 43 S complex formation and start site selection in eukaryotic translation initiation. *J Biol Chem*, **285**, 21203-21207.

19. 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.

20. Kozak, M. (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc Natl Acad Sci U S A*, **87**, 8301-8305.

21. Ivanov, I.P., Firth, A.E., Michel, A.M., Atkins, J.F. and Baranov, P.V. (2011) Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res*.

22. Coldwell, M.J., Cowan, J.L., Vlasak, M., Mead, A., Willett, M., Perry, L.S. and Morley, S.J. (2013) Phosphorylation of eIF4GII and 4E-BP1 in response to nocodazole treatment: A reappraisal of translation initiation during mitosis. *Cell Cycle*, **12**, 3615-3628.

23. Leclerc, G.M., Boockfor, F.R., Faught, W.J. and Frawley, L.S. (2000) Development of a destabilized firefly luciferase enzyme for measurement of gene expression. *Biotechniques*, **29**, 590-601.

24. Coldwell, M.J., Hashemzadeh-Bonehi, L., Hinton, T.M., Morley, S.J. and Pain, V.M. (2004) Expression of fragments of translation initiation factor eIF4GI reveals a nuclear localisation signal within the N-terminal apoptotic cleavage fragment N-FAG. *J Cell Sci*, **117**, 2545-2555.

25. Loetscher, P., Pratt, G. and Rechsteiner, M. (1991) The C terminus of mouse ornithine decarboxylase confers rapid degradation on dihydrofolate reductase. Support for the pest hypothesis. *J Biol Chem*, **266**, 11213-11220.

26. Sung, D. and Kang, H. (1998) The N-terminal amino acid sequences of the firefly luciferase are important for the stability of the enzyme. *Photochemistry and photobiology*, **68**, 749-753.

27. Wegrzyn, J.L., Drudge, T.M., Valafar, F. and Hook, V. (2008) Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. *BMC Bioinformatics*, **9**, 232.

28. 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.

29. Nakata, D., Zhang, L. and Troy, F.A., 2nd. (2006) Molecular basis for polysialylation: a novel polybasic polysialyltransferase domain (PSTD) of 32 amino acids unique to the alpha 2,8-polysialyltransferases is essential for polysialylation. *Glycoconjugate J*, **23**, 423-436.

30. Wang, P.H. (2005) Altered Glycosylation in Cancer: Sialic Acids and Sialyltransferases. *J Cancer Mol*, **1**, 73-81.

31. Zapater, J.L. and Colley, K.J. (2012) Sequences prior to conserved catalytic motifs of polysialyltransferase ST8Sia IV are required for substrate recognition. *J Biol Chem*, **287**, 6441-6453.

32. Busino, L., Bassermann, F., Maiolica, A., Lee, C., Nolan, P.M., Godinho, S.I., Draetta, G.F. and Pagano, M. (2007) SCFFbxl3 controls the oscillation of the circadian clock by directing the degradation of cryptochrome proteins. *Science*, **316**, 900-904.

33. Godinho, S.I., Maywood, E.S., Shaw, L., Tucci, V., Barnard, A.R., Busino, L., Pagano, M., Kendall, R., Quwailid, M.M., Romero, M.R. *et al.* (2007) The after-hours mutant reveals a role for Fbxl3 in determining mammalian circadian period. *Science*, **316**, 897-900.

34. Siepka, S.M., Yoo, S.H., Park, J., Song, W., Kumar, V., Hu, Y., Lee, C. and Takahashi, J.S. (2007) Circadian mutant Overtime reveals F-box protein FBXL3 regulation of cryptochrome and period gene expression. *Cell*, **129**, 1011-1023.

35. Sato, K., Hamada, M., Asai, K. and Mituyama, T. (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res*, **37**, W277-280.

36. Hann, S.R., King, M.W., Bentley, D.L., Anderson, C.W. and Eisenman, R.N. (1988) A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. *Cell*, **52**, 185-195.

37. Lee, S., Liu, B., Huang, S.X., Shen, B. and Qian, S.B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A*, **109**, E2424-2432.

38. Hann, S.R., Sloan-Brown, K. and Spotts, G.D. (1992) Translational activation of the non-AUG-initiated c-myc 1 protein at high cell densities due to methionine deprivation. *Genes Dev*, **6**, 1229-1240.

39. Dmitriev, S.E., Terenin, I.M., Andreev, D.E., Ivanov, P.A., Dunaevsky, J.E., Merrick, W.C. and Shatsky, I.N. (2010) GTP-independent tRNA Delivery to the Ribosomal P-site by a Novel Eukaryotic Translation Factor. *J Biol Chem*, **285**, 26779-26787.

40. Zoll, W.L., Horton, L.E., Komar, A.A., Hensold, J.O. and Merrick, W.C. (2002) Characterization of mammalian eIF2A and identification of the yeast homolog. *J Biol Chem*, **277**, 37079-37087.

41. Starck, S.R., Jiang, V., Pavon-Eternod, M., Prasad, S., McCarthy, B., Pan, T. and Shastri, N. (2012) Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. *Science*, **336**, 1719-1723.

42. Tyzack, J.K., Wang, X., Belsham, G.J. and Proud, C.G. (2000) ABC50 interacts with eukaryotic initiation factor 2 and associates with the ribosome in an ATP-dependent manner. *J Biol Chem*, **275**, 34131-34139.

43. Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappe, J., Gevaert, K. and Van Damme, P. (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics*, **12**, 1780-1790.

FIGURE LEGENDS

Figure 1. Use of a reporter vector to test the efficiency of initiation codon selection.

(A) The pICtest2 reporter plasmid has firefly luciferase (LUCp) under a separate promoter to a Renilla luciferase control (hRluc). Different initiation codon contexts or alternative initiation codons are introduced at the beginning of LUCp, as detailed in the axis labels. (B) Testing of the minimal Kozak consensus surrounding an AUG initiation codon. Following transfection into HeLa or HEK293 cells for 48 hours, the activity of the two luciferases was measured, with firefly luciferase values normalised to Renilla luciferase values, and results then expressed relative to the optimal GCCAUGG consensus sequence. Bars represent the mean of at least three independent experiments, each assayed in triplicate. Error bars represent the standard error of the means and significantly different results (measured versus the GCCAUGG sequence of the relevant cell line) were determined by Student's t-test. Significance levels are denoted throughout this study as *, p>0.05; **, p>0.01; ***, p>0.001. (C) Tally of human AUG initiation codons that may be considered Strong (activity average >70% cf Y/G), Mid (30-70% activity) or Weak (<30% activity). (D) Using the Weak dataset of human transcripts, we predicted the likelihood of "leaking" to the next downstream AUG (in any context), or the next near-cognate CUG, GUG or ACG with a G at the +4 position. The length of the possible truncation of the annotated protein ORF is shown.

Figure 2. Testing the efficiency of non-AUG initiation codon selection.

(A) Efficiency of translation from AUG and near-cognate codons in either the full Kozak consensus sequence versus the non-AUG consensus sequence proposed by Wegrzyn et al., was measured using the pICtest2 vectors as previously described. In this case significance was measured versus the negative control of a UAC codon within a Kozak consensus for each cell line. The lower panel is a duplication of the upper panel, but expanded on the y-axis in order to better show the low levels of translation from the near-cognate codons. (B) Tally of the number of human transcripts where there is a predicted likelihood of upstream initiation from an AUG (in any context), or a near-cognate CUG, GUG or ACG initiation codon with a G at the +4 position. The length of the possible extension of the annotated protein ORF is shown. (C) Open reading frames of different eukaryotic initiation factors (eIFs) were subcloned into the pcDNA3.1(+) vector. The effect of overexpressing eIF1, eIF1AX and eIF1AY on initiation codon selection was measured by assay of luciferase from pICtest2 vectors containing the initiation codons shown, all within the full Kozak consensus.

Figure 3. Identifying non-canonical upstream initiation events following bioinformatic identification.

(A) A bioinformatics pipeline was used to examine 5' UTR sequences for the presence of possible extensions by determining sequence homology with translated 5' UTRs from other species. To detect upstream initiation in a candidate, a cDNA corresponding to the whole 5' UTR, annotated initiation codon and part of the candidate ORF were amplified and inserted upstream of a 3x FLAG tag (maintaining the open reading frame) in a pcDNA vector. (B) The output results from the bioinformatics pipeline for ST8SIA4, detailing the predicted position and codon which may drive alternative initation. eLen/tLen denoted extension length or truncation length in amino acids, compared to the annotated open reading frame. (C) Variants of the initial ST8SIA4-3F vector were created in order to express single isoforms by removing the 5' UTR and changing alternative initiation codons to AUG in a strong context. Predicted migration of eORFs fused to the 3x FLAG tag was determined and is shown alongside the variants (D) SDS-PAGE and immunoblotting with anti-FLAG antibody was used to confirm presence/absence of novel initiation events following transfection of vectors into HeLa cells. Mutation of the CUG codons at -225 and -216 creates a version which would be unable to initiate (lane 8). (E) HeLa cells that had been either untransfected (panel i), or transfected with ST8SIA4 variants (panels ii-iv) were fixed in paraformaldehyde, permeabilised with Triton X-100 and then incubated with murine monoclonal primary antibodies to either Golgin 97 (i) or FLAG (ii-iv). AlexaFluor 488 conjugated secondary antibodies were used to visualise subcellular localisation, with DAPI used to stain DNA.

Figure 4 Identification of upstream initiation in FBXL3.

(A) Two upstream initiation codons are predicted in the FBXL3 5' UTR, and thus a region of FBXL3 cDNA encompassing the entire 5' UTR and a portion of the ORF was amplified from a full length cDNA clone and cloned in-frame with a triple FLAG epitope tag as before. Calculated molecular weights for the annotated and predicted novel forms are shown. (B) The results of the bioinformatics workflow for FBXL3 show the positions of predicted upstream codons. (C). The predicted GUG underwent site-directed mutagenesis from wildtype GUG to AUG, UAC or UAA. Another mutant was generated to mutate AAGGUG to UACUAC. Plasmids were transiently transfected into HeLa cells and proteins were detected by immunoblotting using FLAG or α-actin primary antibodies. (D) Centroidfold (35) was used to predict a possible secondary structure of the GC rich region between the AAG/GUG and AUG codons (outlined in green), using conserved sequences. Site-directed mutagenesis was employed at the positions shown by red and blue arrowheads to reduce the energy required to unwind predicted secondary structure whilst maintaining the encoded amino acids at the indicated nucleotides. (E) Immunoblotting of FBXL3 variants, showing how the predicted hairpin influences translational efficiency. (F) Manipulation of trans-acting factors control isoform expression. Levels of two translation initiation factors, eIF1 and eIF1AX implicated in fidelity of start codon selection were increased by overexpression from pcDNAmyc plasmids or decreased by siRNA (expressed from pSilencer). The effect upon FBXL3 isoform expression was determined by immunoblotting for the proteins shown. (G) Quantification of results from three independent experiments as per panel F showing the change in ratio of extended to annotated forms of FBXL3 in response to alterations in levels of initiation factors.

Figure 5. Examination of turnover of FBXL3 proteins and the Cryptochrome 1 substrate.

(A) A region of FBXL3 cDNA encompassing the extended or annotated ORF was amplified and cloned in-frame with a triple FLAG epitope tag. The initiation codon was mutated in both instances to an AUG codon in the optimal context to enhance translation from the intended start site. (B) Plasmids were transiently transfected into HeLa cells and three days post transfection, cells were treated with cycloheximide to inhibit new protein synthesis and lysed as before at time 0, 2, 8, 16 and 24 hrs. Immunoblotting for the proteins shown was used to establish their levels during the timecourse. (C) Quantification of levels of exogenous isoforms of FBXL3 over the timecourse, normalised to the actin loading control. (D) Quantification of endogenous Cry1 expression, relative to the loading control.

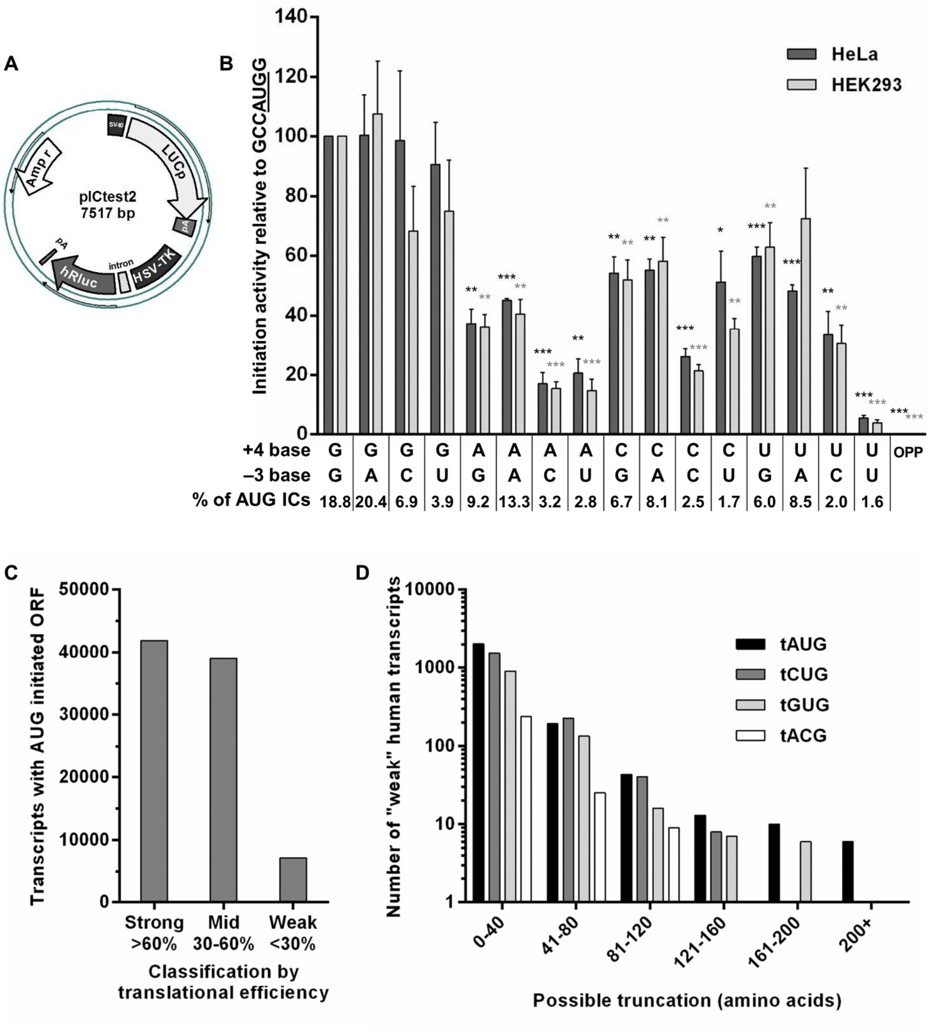


Figure 1. Use of a reporter vector to test the efficiency of initiation codon selection. (A) The pICtest2 reporter plasmid has firefly luciferase (LUCp) under a separate promoter to a Renilla luciferase control (hRluc). Different initiation codon contexts or alternative initiation codons are introduced at the beginning of LUCp, as detailed in the axis labels. (B) Testing of the minimal Kozak consensus surrounding an AUG initiation codon. Following transfection into HeLa or HEK293 cells for 48 hours, the activity of the two luciferases was measured, with firefly luciferase values normalised to Renilla luciferase values, and results then expressed relative to the optimal GCCAUGG consensus sequence. Bars represent the mean of at least three independent experiments, each assayed in triplicate. Error bars represent the standard error of the means and significantly different results (measured versus the GCCAUGG sequence of the relevant cell line) were determined by Student’s t-test. Significance levels are denoted throughout this study as *, p>0.05; **, p>0.01; ***, p>0.001. (C) Tally of human AUG initiation codons that may be considered Strong (activity average >70% cf Y/G), Mid (30-70% activity) or Weak (<30% activity). (D) Using the Weak dataset of human transcripts, we predicted the likelihood of “leaking” to the next

downstream AUG (in any context), or the next near-cognate CUG, GUG or ACG with a G at the +4 position. The length of the possible truncation of the annotated protein ORF is shown.

177x229mm (300 x 300 DPI)

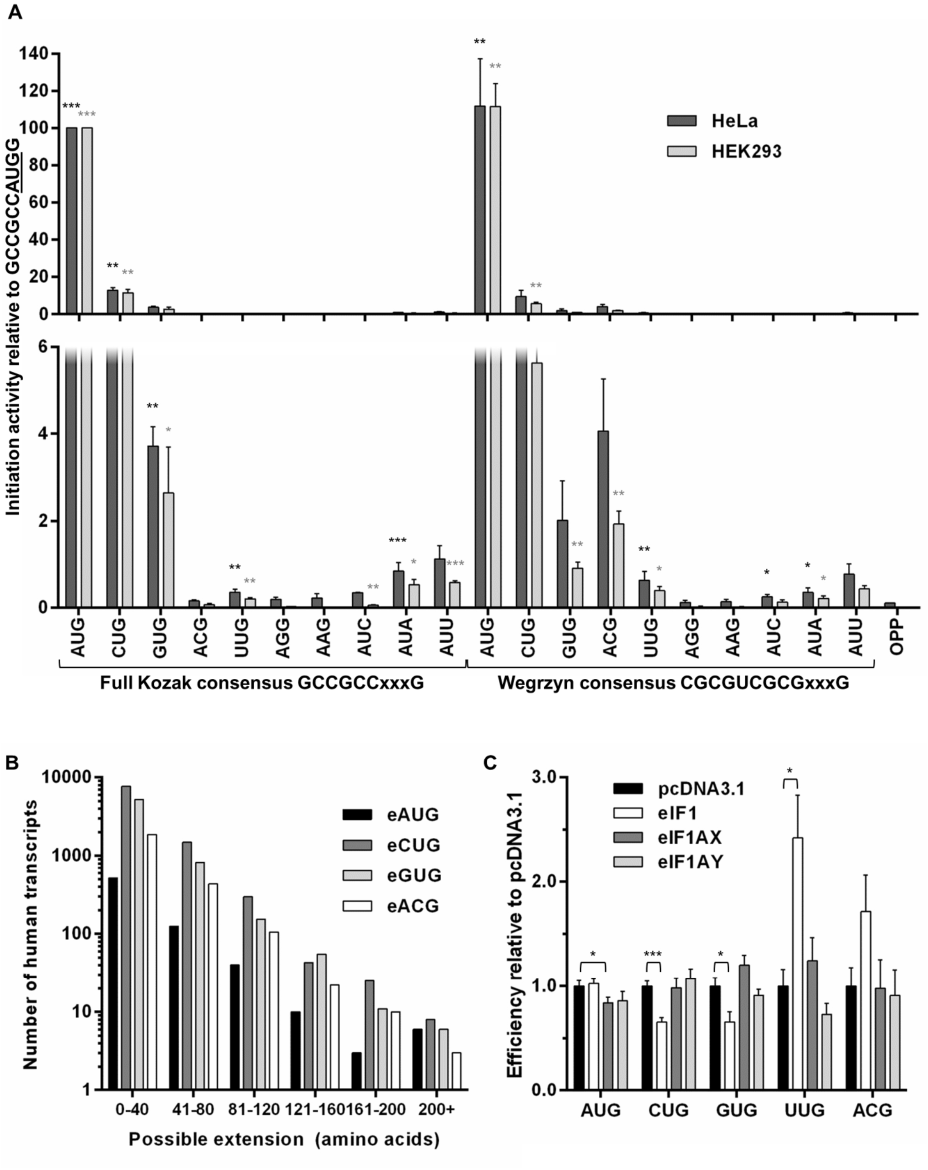
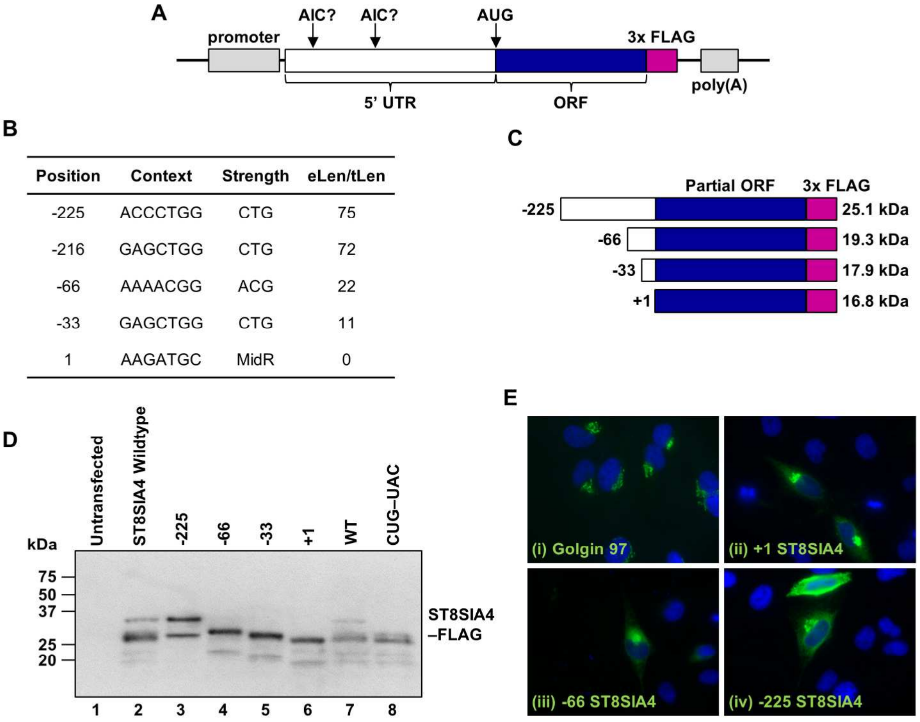


Figure 2. Testing the efficiency of non-AUG initiation codon selection.
(A) Efficiency of translation from AUG and near-cognate codons in either the full Kozak consensus sequence versus the non-AUG consensus sequence proposed by Wegrzyn et al., was measured using the pICtest2 vectors as previously described. In this case significance was measured versus the negative control of a UAC codon within a Kozak consensus for each cell line. The lower panel is a duplication of the upper panel, but expanded on the y-axis in order to better show the low levels of translation from the near-cognate codons. (B) Tally of the number of human transcripts where there is a predicted likelihood of upstream initiation from an AUG (in any context), or a near-cognate CUG, GUG or ACG initiation codon with a G at the +4 position. The length of the possible extension of the annotated protein ORF is shown. (C) Open reading frames of different eukaryotic initiation factors (eIFs) were subcloned into the pcDNA3.1(+) vector. The effect of overexpressing eIF1, eIF1AX and eIF1AY on initiation codon selection was measured by assay of luciferase from pICtest2 vectors containing the initiation codons shown, all within the full Kozak consensus.

177x229mm (300 x 300 DPI)



Position	Context	Strength	eLen/tLen
-225	ACCCTGG	CTG	75
-216	GAGCTGG	CTG	72
-66	AAAACGG	ACG	22
-33	GAGCTGG	CTG	11
1	AAGATGC	MidR	0

Figure 3. Identifying non-canonical upstream initiation events following bioinformatic identification. (A) A bioinformatics pipeline was used to examine 5′ UTR sequences for the presence of possible extensions by determining sequence homology with translated 5′ UTRs from other species. To detect upstream initiation in a candidate, a cDNA corresponding to the whole 5′ UTR, annotated initiation codon and part of the candidate ORF were amplified and inserted upstream of a 3x FLAG tag (maintaining the open reading frame) in a pcDNA vector. (B) The output results from the bioinformatics pipeline for ST8SIA4, detailing the predicted position and codon which may drive alternative initation. eLen/tLen denoted extension length or truncation length in amino acids, compared to the annotated open reading frame. (C) Variants of the initial ST8SIA4-3F vector were created in order to express single isoforms by removing the 5′ UTR and changing alternative initiation codons to AUG in a strong context. Predicted migration of eORFs fused to the 3x FLAG tag was determined and is shown alongside the variants (D) SDS-PAGE and immunoblotting with anti-FLAG antibody was used to confirm presence/absence of novel initiation events following transfection of vectors into HeLa cells. Mutation of the CUG codons at -225 and -216 creates a version which would be unable to

initiate (lane 8). (E) HeLa cells that had been either untransfected (panel i), or transfected with ST8SIA4 variants (panels ii-iv) were fixed in paraformaldehyde, permeabilised with Triton X-100 and then incubated with murine monoclonal primary antibodies to either Golgin 97 (i) or FLAG (ii-iv). AlexaFluor 488 conjugated secondary antibodies were used to visualise subcellular localisation, with DAPI used to stain DNA.

177x229mm (300 x 300 DPI)

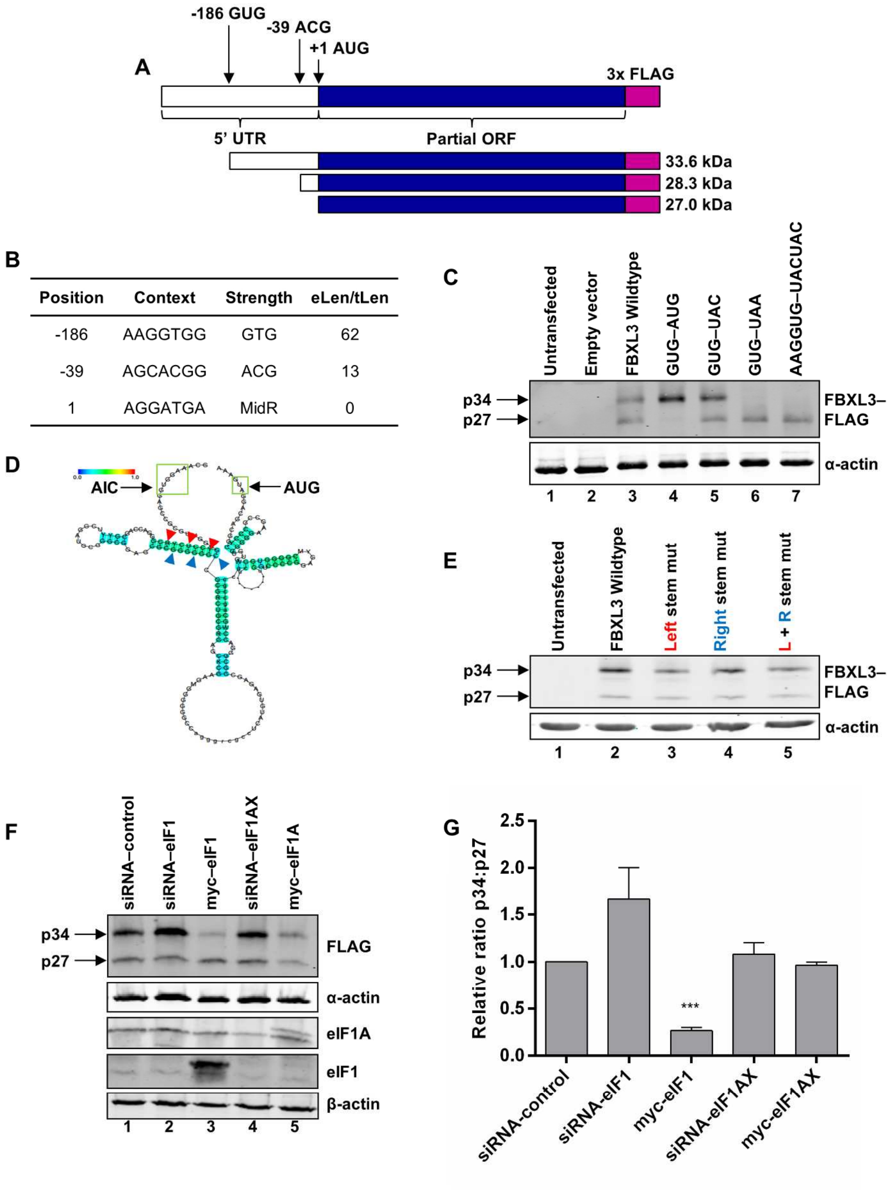


Figure 4 Identification of upstream initiation in FBXL3.
(A) Two upstream initiation codons are predicted in the FBXL3 5' UTR, and thus a region of FBXL3 cDNA encompassing the entire 5' UTR and a portion of the ORF was amplified from a full length cDNA clone and cloned in-frame with a triple FLAG epitope tag as before. Calculated molecular weights for the annotated and predicted novel forms are shown. (B) The results of the bioinformatics workflow for FBXL3 show the positions of predicted upstream codons. (C). The predicted GUG underwent site-directed mutagenesis from wildtype GUG to AUG, UAC or UAA. Another mutant was generated to mutate AAGGUG to UACUAC. Plasmids were transiently transfected into HeLa cells and proteins were detected by immunoblotting using FLAG or α-actin primary antibodies. (D) Centroidfold (35) was used to predict a possible secondary structure of the GC rich region between the AAG/GUG and AUG codons (outlined in green), using conserved sequences. Site-directed mutagenesis was employed at the positions shown by red and blue arrowheads to reduce the energy required to unwind predicted secondary structure whilst maintaining the encoded amino acids at the indicated nucleotides. (E) Immunoblotting of FBXL3 variants, showing how the predicted hairpin influences

translational efficiency. (F) Manipulation of trans-acting factors control isoform expression. Levels of two translation initiation factors, eIF1 and eIF1AX implicated in fidelity of start codon selection were increased by overexpression from pcDNAmyc plasmids or decreased by siRNA (expressed from pSilencer). The effect upon FBXL3 isoform expression was determined by immunoblotting for the proteins shown. (G) Quantification of results from three independent experiments as per panel F showing the change in ratio of extended to annotated forms of FBXL3 in response to alterations in levels of initiation factors.

177x229mm (300 x 300 DPI)

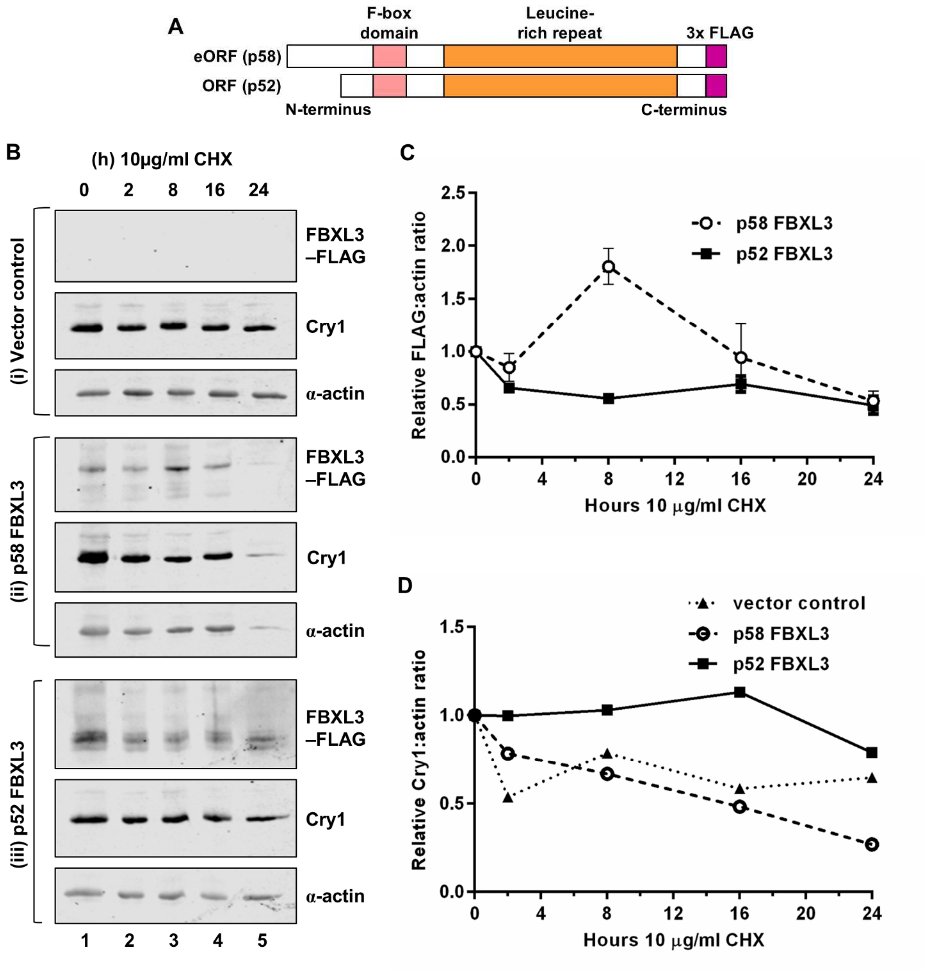


Figure 5. Examination of turnover of FBXL3 proteins and the Cryptochrome 1 substrate. (A) A region of FBXL3 cDNA encompassing the extended or annotated ORF was amplified and cloned in-frame with a triple FLAG epitope tag. The initiation codon was mutated in both instances to an AUG codon in the optimal context to enhance translation from the intended start site. (B) Plasmids were transiently transfected into HeLa cells and three days post transfection, cells were treated with cycloheximide to inhibit new protein synthesis and lysed as before at time 0, 2, 8, 16 and 24 hrs. Immunoblotting for the proteins shown was used to establish their levels during the timecourse. (C) Quantification of levels of exogenous isoforms of FBXL3 over the timecourse, normalised to the actin loading control. (D) Quantification of endogenous Cry1 expression, relative to the loading control.

177x229mm (300 x 300 DPI)