Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
ORIGINAL PAPER

OXFORD

Genome Analysis

Lost in Translation: Bioinformatic Analysis of Variations Affecting the Translation Initiation Codon in the Human Genome

Francisco Abad [1], María Eugenia de la Morena-Barrio [2, 3], Jesualdo Tomás Fernández-Breis [1,*] and Javier Corral [2, 3]

[1]Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, Murcia, 30008, Spain,
[2]Servicio de Hematología y Oncología Médica, Hospital Universitario Morales Meseguer, Centro Regional de Hemodonación, Universidad de Murcia, IMIB-Arrixaca, Murcia, Spain and
[3]Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III (ISCIII), Spain.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Translation is a key biological process controlled in eukaryotes by the initiation AUG codon. Variations affecting this codon may have pathological consequences by disturbing the correct initiation of translation. Unfortunately, there is no systematic study describing these variations in the human genome. Moreover, we aimed to develop new tools for *in silico* prediction of the pathogenicity of gene variations affecting AUG codons, because to date, these gene defects have been wrongly classified as missense.
Results: Whole-exome analysis revealed the mean of 12 gene variations per person affecting initiation codons, mostly with high (> 0.01) minor allele frequency (MAF). Moreover, analysis of Ensembl data (December 2017) revealed 11,261 genetic variations affecting the initiation AUG codon of 7,205 genes. Most of these variations (99.5%) have low or unknown MAF, probably reflecting deleterious consequences. Only 62 variations had high MAF. Genetic variations with high MAF had closer alternative AUG downstream codons than did those with low MAF. Besides, the high-MAF group better maintained both the signal peptide and reading frame. These differentiating elements could help to determine the pathogenicity of this kind of variation.
Availability: Data and scripts in Perl and R are freely available at https://github.com/fanavarro/hemodonacion
Contact: jfernand@um.es
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Initiation of translation is a crucial step that is subject to substantial regulation to guarantee correct formation of a protein by an mRNA (Lorsch and Dever, 2010). This is a complex process that involves different elements and steps. Briefly, the pre-initiation complex – formed by the small subunit of the ribosome, by the tRNA that transports methionine, and by several initiation factors – is built. Then, this pre-initiation complex is brought to an mRNA in its 5′-terminal cap to start a scanning process from 5′ to 3′ to find a translation initiation site (TIS), whose main component is the initiation AUG codon. The sequence flanking the initiation codon is important for the pre-initiation complex to recognise AUG as an initiation codon. The consensus sequence GCC(**A/G**)CC**AUGG**, also known as the Kozak sequence (Kozak, 1986), whose most important nucleotides are boldfaced, describes an optimal context for recognition of the underlined AUG as an initiation codon. Thus, the ribosomal large subunit is added to the complex, which proceeds to the next stage of translation: elongation (Preiss and W Hentze, 2003).

Accordingly, most of the genetic variations affecting the initiation codon should have pathological consequences owing to the biological

importance of this codon. Indeed, there are numerous examples of mutations of this codon associated with various diseases (Kozak, 2002; Cazzola and Skoda, 2000). Nonetheless, common polymorphisms affecting the initiation codon without pathological significance have been observed (González-Conejero *et al.*, 2002; Corral *et al.*, 2000, 2010). Therefore, there may be mechanisms that mitigate the effect of this kind of genetic variation, such as the use of an alternative initiation codon located downstream of the wild-type one.

To the best of our knowledge, there is no global study on mutations affecting the initiation codon in the human genome. Therefore, our main objectives are 1) to perform a descriptive analysis by extracting all data about these genetic variations from the Ensembl database; 2) to conduct a comparative analysis between frequent and rare variations, considering the features of the open reading frames (ORFs) provoked by the use of alternative initiation codons, to identify differentiating factors; and 3) to measure the impact of these variations on a concrete individual through analysis of complete exomes of 5 subjects.

2 Methods

2.1 Data collection

We used the Ensembl data (Yates *et al.*, 2016) to create a TSV file containing all variations affecting translation initiation codons in any transcript. Ensembl imports data from different databases such as dbSNP (Sherry *et al.*, 2001), COSMIC (Forbes *et al.*, 2015), or ClinVar (Landrum *et al.*, 2016). The data were retrieved using the Ensembl Perl API (Yates *et al.*, 2015). We provide a full description of the method for identifying variants and for determining their position in Supplementary File 1. The source code of the scripts developed is available at https://github.com/fanavarro/hemodonacion.

In summary, a set of transcript variation allele (TVA) objects was retrieved from Ensembl. These objects are tuples of the form <transcript, variation, allele>, so that each TVA identifies an allele of a variation that affects a concrete transcript. Each row in the TSV file stores information on a TVA. The same variation may appear more than once because it can affect several transcripts and may have several alleles.

Alternative initiation codons were searched for each TVA in the mutated sequences according to 3 approaches:

1. The first AUG codon found in the coding region.
2. The first AUG codon in the coding region whose context has efficiency greater than or equal to 87 according to one study (Noderer *et al.* 2014). That study provides a list with all possible TISs spanning positions -6 to +5 together with an efficiency measure. This efficiency was calculated taking into account dinucleotide interactions.
3. The first AUG found in a strong Kozak context in the coding region. To identify the Kozak context in transcripts, a position weight matrix was compiled by means of 10 nucleotides per side around the initiation codon from 14,160 transcript samples retrieved from Ensembl (see Supplementary File 2). After that, the formed matrix (see Supplementary File 3) was used for scanning the mutated transcripts by means of the R library PWMEnrich.

The consequences of the mutation were also calculated by comparing the length, reading frame, and signal peptide conservation of the wild-type ORF with those caused by the use of the alternative initiation codon found by approaches 1, 2, and 3. The Phobius software (Käll *et al.*, 2004) was employed for predicting the positions relative to the signal peptide in the wild-type transcript. This strategy allowed us to calculate signal peptide conservation according to the alternative initiation codon used. Additionally, methionines in the $5'$ untranslated region (UTR) of the wild-type transcript were analysed. Their positions and reading frames with respect to the wild-type initiation codon were obtained. We also identified existence of a termination codon before the coding region.

The new reading frame of alternative initiation codons was compared with the reading frame of the wild type. This task consists of the following steps:

- Determining the percentage of the wild-type ORF conserved in the mutated ORF. This percentage is obtained by comparing the lengths of the 2 sequences.
  - If the percentage is $\leq 1\%$, then we assume that ORFs produced by the wild-type initiation codon and the alternative initiation codon are not in the same frame.
  - If the percentage is $> 1\%$, then mutated and wild-type ORFs are translated into amino acid sequences.
    - In the case of a deletion, the same number of amino acids that has been deleted is removed from the N terminus of the wild-type sequence.
    - In the case of an insertion, the same number of amino acid residues that appeared because of the insertion plus 1 are removed from the N terminus of the mutated sequence as amino acid residues.
    - If the mutated sequence is contained in the wild-type one or vice versa, then we assume that both are in the same reading frame.

This method can be applied to variations affecting the initiation codon. This method can detect cases like the following deletion, whose deleted region is enclosed in parentheses:

AT(GGAGAGTAA)GGATGAGTAG → M(ESK)DE-

Although the mutated amino acid sequence (MDE-) is not contained in the wild-type sequence (MESKDE-), both are in the same reading frame. The described method will check whether 'DE-' is included in both sequences, thus returning a positive result. This method may provide false positives without length checking when comparing very short sequences with a longer one. These short sequences could be produced when the reading frame is lost and then a premature termination codon is found. The length comparison between ORFs is needed to detect these false positives. A threshold of 1% was selected after analysis of our dataset; this threshold is sufficient to prevent the majority of false positives. Hence, if the length of the mutated ORF is shorter than 1% of the wild-type one, then the method will assume a lost reading frame.

2.2 Data analysis

Firstly, each TVA is classified according to its MAF value: high MAF ($MAF \geq 0.01$) or low MAF ($MAF < 0.01$). Variations without MAF data are excluded from subsequent analysis. Both groups were compared on the following features:

1. The position of the first alternative initiation codon.
2. The percentage of signal peptide conservation.
3. Consequences for the wild-type reading frame.
4. The number and positions of AUG codons found in the $5'$ UTR of the wild-type transcript.

The steps 1, 2, and 3 are carried out according to the ORFs resulting from the use of the alternative initiation codons found by approaches 1, 2, and 3.

Statistical analysis included the non-paired Wilcoxon test (Wilcoxon, 1945) for comparison of the alternative initiation codon position, the percentage of signal peptide conservation, and the number of AUG codons

in the 5′ UTR as well as Yates χ^2 tests (Yates, 1934) for the maintenance of the reading frame. Moreover, Gene Ontology (GO) gene set enrichment analysis of the groups of genes with high and low MAF was carried out using the R Cluster Profiler library (Yu *et al.*, 2012).

2.3 Exome analysis

Complete exomes of 5 patients were analysed to estimate the presence of variations affecting initiation codons in a concrete individual. Ion Proton Platform (Ion Torrent) was employed for sequencing. The software of this platform outputs CSV files with information about genomic variations found in the patient. This information includes the position of a variation in a coding region, which allows researchers to apply a filter to obtain the variations affecting initiation codons. Furthermore, variations with coverage < 20 were removed to avoid possible lecture errors.

3 Results

3.1 Variations affecting initiation codons

The TSV file with all variations affecting initiation codons contains 15,312 TVAs, including 11,261 different variations, which affect 9,591 transcripts of 7,205 genes. Each gene had the mean of 2.12 related TVAs affecting initiation codons. Supplementary File 4 contains the complete dataset, whereas a simplified version is provided in Supplementary File 5. This simplified version contains the minimal information necessary to perform the analysis described in section 2.2. Supplementary File 1 describes the format of these files.

PAX5 was the most affected gene, with 72 TVAs (9 different variations), all present in COSMIC, affecting 8 transcripts of the gene. This gene is a transcription factor whose function is essential for commitment of lymphoid progenitors to the B-lymphocyte lineage (Cobaleda *et al.*, 2007). The second most affected gene is *DTNA*, with 38 TVAs (8 different variations), 1 present in dbSNP and 7 in COSMIC, affecting 8 different transcripts. This gene encodes a protein of the dystrobrevin family playing structural and signaling roles at the plasma membrane of many cell types (Böhm *et al.*, 2008). Table 1 contains the top 10 affected genes.

Table 1. Top 10 genes affected by variations in initiation codons in the order of the number of affected TVAs for each gene.

Gene	TVAs affected in the initiation codon
PAX5	72
DTNA	38
MAX	35
BRCA1	33
EIF3J	30
SDHD	27
SIGLEC7	27
RBMX	26
ELN	24
LEPR	24

Our study identified 87 TVAs (62 variations affecting 75 transcripts) with $MAF > 0.01$, with rs11107 (dbSNP) being the most frequent variation identified ($MAF = 0.48742$). This SNP (ATG/ATA) affects 1 transcript of the *FBXO7* gene, which encodes an F-box protein whose function is involved in ubiquitination processes (Cenciarelli *et al.*, 1999). The second most frequent variation ($MAF = 0.463841$) is rs3764880 (dbSNP). This is another SNP (ATG/GTG) that affects only 1 transcript of *TLR8*, a gene encoding Toll-like receptor 8 involved in the recognition of foreign pathogens (Cervantes *et al.*, 2012). Table 2 shows the top 10 variations ordered by MAF together with the affected genes and transcripts.

Table 2. Top 10 high-MAF variations affecting initiation codons in the order of MAFs (descending order) together with the affected gene and transcripts.

Gene	Transcript	Variation	MAF
FBXO7	ENST00000397426	rs11107	0.4874200
TLR8	ENST00000218032	rs3764880	0.4638410
HIBCH	ENST00000392332; ENST00000359678	rs291466	0.4317090
HRNR	ENST00000368801	rs561299511	0.3881790
ADSSL1	ENST00000332972	rs80097179	0.3761980
ATP6V1B1	ENST00000234396	rs11681642	0.3688100
NFU1	ENST00000303698	rs4453725	0.3170930
PRAM1	ENST00000423345	rs968502	0.3119010
HADHB	ENST00000317799	rs147970487	0.2905350
ZFP62	ENST00000512132	rs705441	0.2863420

On the other hand, 1,446 TVAs (1,008 different variations affecting 1,190 transcripts) had low MAF (< 0.01). One of the rarest variations ($MAF = 0.0002$) is rs139267617 (dbSNP), a single-nucleotide variant [SNV] (ATG/GTG) affecting 1 transcript of *ITLN1*, a gene encoding intelectin-1, which participates in the recognition of microbial glycans (Wesener *et al.*, 2015; Tsuji *et al.*, 2001). Another rare mutation is rs587701888 (dbSNP) ($MAF = 0.0002$), a SNV (ATG/GTG) that affects 3 different transcripts of *SEMA6C*, a gene that encodes a protein of the semaphorin family, which is involved in neural regeneration (Qu *et al.*, 2002). Table 3 shows a sample of 10 variations with the lowest MAF.

Table 3. A sample of 10 low-MAF variations affecting initiation codons together with the affected gene and transcripts.

Gene	Transcript	Variation	MAF
ITLN1	ENST00000326245	rs139267617	0.0002
SEMA6C	ENST00000368914	rs587701888	0.0002
SLAMF8	ENST00000289707	rs149191301	0.0002
PGM1	ENST00000371084	rs200633484	0.0002
RCAN3	ENST00000412742	rs574730150	0.0002
MT1HL1	ENST00000464121	rs546118456	0.0002
CD1E	ENST00000368167	rs201300311	0.0002
CAPZB	ENST00000264203	rs568906973	0.0002
PANK4	ENST00000378466	rs201511268	0.0002
SEMA6C	ENST00000368912	rs587701888	0.0002

3.2 High-MAF versus low-MAF variations, a comparison

The alternative AUG codon found by approach 1 was closer to the wild-type initiation codon in the group of variations with high MAF than in the group with low MAF. Thus, variations with high MAF were associated with an alternative AUG codon at a median distance of 26 bp from the wild-type AUG codon, whereas the first AUG codon appeared at a median distance of 49 bp from the wild-type one in genes affected by variations with low MAF ($p = 5.42e - 6$) (see Figure 1A).

It should be noted that 2 TVAs in the high-MAF group (both affecting the ENST00000391418 transcript of the *KRTAP2-3* gene, which encodes keratin-associated protein 2-3), and 3 TVAs in the group of low MAF (affecting 2 transcripts: ENST00000335123 of the *LEP1* gene, which encodes late cornified envelope 1A protein, and ENST00000600213 of

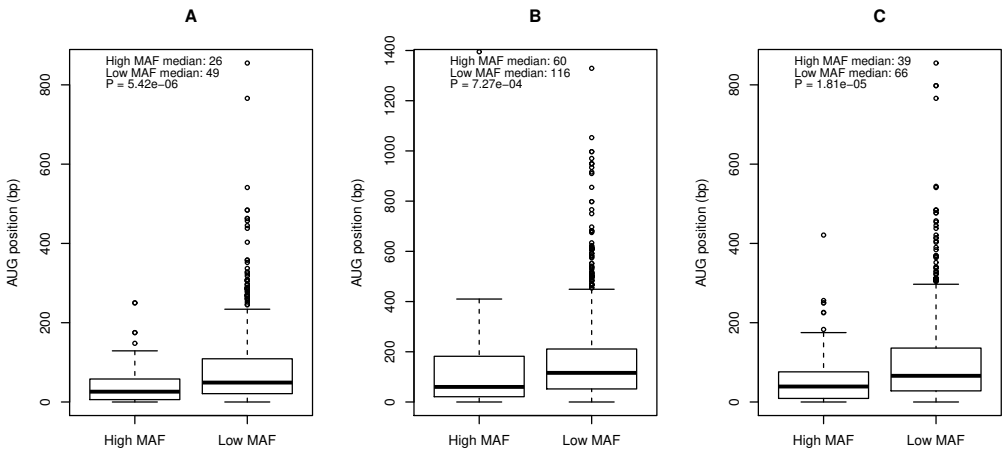

Fig. 1. Position (in base pairs [bp] from the wild-type initiation codon) of the alternative downstream AUG codon found by A) approach 1, B) approach 2, and C) approach 3.

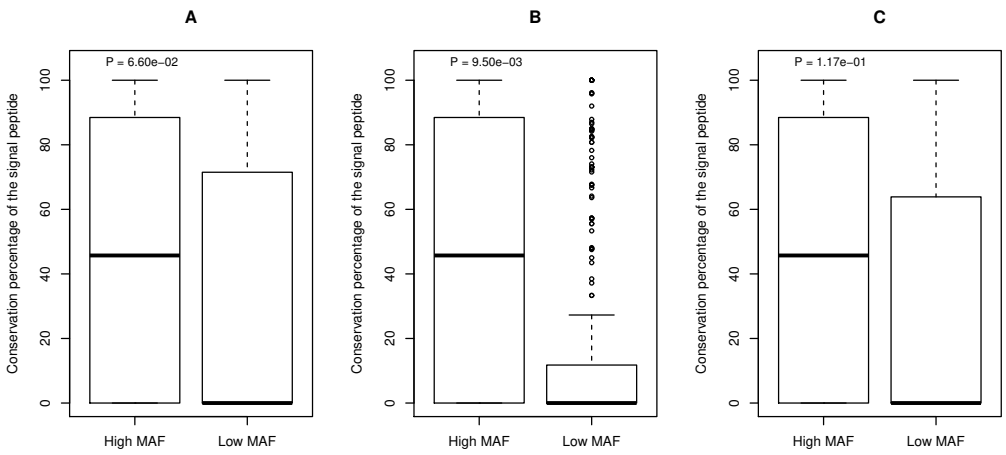

Fig. 2. A comparison of conservation percentages of the signal peptide when an alternative AUG codon is used that is found by A) approach 1, B) approach 2, and C) approach 3.

the *HN2* gene, which codes for late MT-RNR2-like 12 protein) have no downstream alternative initiation codon. All these affected transcripts without alternative AUG codons have a single exon.

The alternative AUG codons found by approaches 2 and 3 were also closer to the wild-type initiation codon for the variations with high MAF than for those with low MAF (60 bp vs 116 bp ;$p = 7.27e - 5$ [Figure 1B] and 39 bp vs 66 bp; $p = 1.81e - 5$ [Figure 1C], respectively.

We next tested by means of Phobius whether the use of the alternative AUG codon might affect the signal peptide. A signal peptide was detected in the wild-type transcript of only 288 TVAs from the group of variations with low MAF and 13 TVAs in the high-MAF group. The use of the alternative AUG codon found by approach 1 in those TVAs resulted in the mean conservation of the signal peptide for variations with low MAF of 32.2% and 51.0% for variations with high MAF (the conservation median is 0% for low MAF and 45.7% for high MAF). The values were similar when we considered the alternative AUGs found by approaches 2 and 3 although only those found by approach 2 reached statistical significance (see Figure 2).

The 3 approaches yield a similar result on the reading frame status (see Table 4): the proportion of alternative initiation codons that maintain the reading frame is higher for high-MAF variations (60.92% vs 48.48% for approach 1, 62.07% vs 41.01% for approach 2, and 67.82% vs 50.97% for approach 3).

We searched for upstream AUG codons in the 5′ UTR of the wild-type transcript. Of note, variations with high MAF had more 5′ UTR AUG codons than variations with low MAF (mean of 1.9 vs 1.7) although these differences did not reach statistical significance ($p = 0.2128$) (see Figure 3).

Finally, the GO gene set enrichment analysis did not yield any statistically significantly enriched molecular function or biological process at p-value and q-value cut-offs of 0.1. This result suggests that these variations are not exclusive for some functional group of genes.

3.3 Exome analysis

The search for variations affecting initiation codons in whole exomes of 5 people revealed the mean of 12 variations with these characteristics from

Table 4. A comparison of the reading-frame status between high-MAF and low-MAF variations using the AUG codons found by approaches 1, 2, and 3.

Alternative initiation codon	Group	TVAs maintaining reading frame	TVAs losing reading frame	TVAs without alternative initiation codon	Total TVAs
Found by approach 1	High MAF	53 (60.92%)	32 (36.68%)	2 (2.3%)	87
	Low MAF	701 (48.48%)	742 (51.31%)	3 (0.2%)	1,446
	p-value = 0.01844				
Found by approach 2	High MAF	54 (62.07%)	29 (33.33%)	4 (4.60%)	87
	Low MAF	593 (41.01%)	812 (56.15%)	41 (2.03%)	1,446
	p-value = 7.27e-05				
Found by approach 3	High MAF	59 (67.82%)	26 (29.88%)	2 (2.3%)	87
	Low MAF	737 (50.97%)	700 (48.41%)	9 (0.62%)	1,446
	p-value = 0.001695				

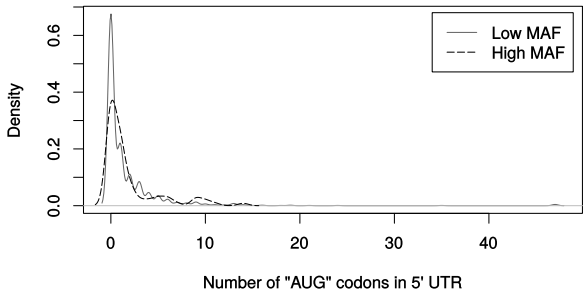

Fig. 3. A density plot of the number of AUG codons found in 5′ UTRs.

the mean of 45,696 variations identified in each patient. Twenty-eight different variations affecting initiation codons were identified in these 5 subjects, and 14 of the variations were identified in more than 1 patient (see Figure 4 and Supplementary File 6). Seven of these variations did not have MAF reported, whereas the remaining 21 mutations had a mean MAF of 0.232. The variation identified in *C6orf7* was detected in homozygosis in all the subjects having a MAF of 0.006. We believe that this is an error in the reference sequence, and consequently, there is no AUG initiation codon at this position. The RefSeq entry of the transcript affected by this variation (refSeqNM_001243308.1) reads 'This RefSeq was removed because currently there is insufficient support for the transcript and the protein', in agreement with our reasoning.

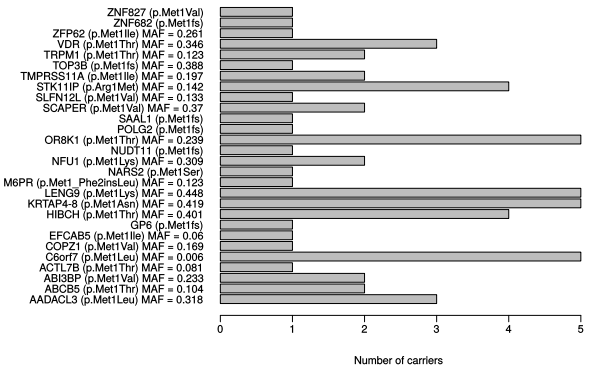

Fig. 4. Variations affecting initiation codons identified by whole-exome analysis of 5 patients.

4 Discussion

Translation is a key biological process that probably has not received due attention, particularly in the context of pathologies. Thus, it is quite normal to find reports of mutations identified in patients with different diseases that affect the initiation codon of many different genes still maintaining the predicted missense change. This phenomenon is due to the simple analysis of the primary sequence, when the consequences of this kind of variations are regarded as a simple nucleotide change instead of considering possible truncations or disruptions of the translational reading frame.

Many widely used variant effect predictors, such as SIFT (Ng and Henikoff, 2003) or PolyPhen (Adzhubei *et al.*, 2010), are mainly based on the amino acid conservation level. These predictors perform a multiple sequence alignment of the family of the affected proteins to estimate the potential amino acid changes that could occur without deleterious effects for each position. Nonetheless, variations affecting initiation codons may prevent translation from the canonical TIS and consequently a change of the first methionine will rarely occur.

A comparison of predictions based on an amino acid residue and our method may produce 2 results: agreement and disagreement. Agreement means that the same prediction was obtained by applying completely different methods. An example of agreement can be found for variation

rs748970009 (ATG/ATA) affecting the canonical transcript of the *RPA2* gene. On the one hand, alternative initiation codons found by our analysis are at positions 2 and 25, thus changing the reading frame. Although MAF is not available, these features seem to indicate a non-viable variation. On the other hand, SIFT and PolyPhen find methionines in the protein family matching the affected one, indicating high conservation. Consequently, they also predict the variation as deleterious (SIFT) and possibly damaging (PolyPhen).

A disagreement means that the predictions of different methods are contradictory, for instance, variation rs11681642 (ATG/ACG) affecting transcript ENST00000234396. This variation is predicted as 'deleterious–low confidence' and 'possibly damaging' by SIFT and PolyPhen, respectively. The reason is that the affected methionine is conserved in the protein family. Nevertheless, the high MAF (0.37) seems to indicate viability. In our method, approaches 1 and 3 (see the TSV file) suggest an alternative initiation codon, located at 6 bp downstream of the wild-type one, maintaining the reading frame and conserving 99.6% of the wild-type protein. Approach 2 proposes an initiation codon 75 bp downstream of the wild-type one, also maintaining the reading frame and conserving 95.1% of the wild-type protein. Based on our hypothesis, these codons could serve as initiators of translation. Due to proximity to the $5'$ region, the AUG codon at position 6 is expected to be more efficient. This codon could produce an isoform of the wild-type protein that loses only 2 amino acid residues, without a significant loss of function. Although our analysis contradicts SIFT and PolyPhen predictions, it explains the high MAF of the variation.

Distortion of the initiation context, particularly if affecting the AUG initiation codon, may result in a different start of translation that might render quite a different protein. In this study, we explored all the variations reported in Ensembl affecting the AUG initiation codon. We found differentiating factors that may facilitate the development of new predictive tools for calculating the deleterious effect of this kind of variation. These factors can be summarised in 2 well-defined situations.

1. The presence of an alternative methionine with a high score of translation and located near the wild-type AUG codon may be considered a protective mechanism against the deleterious consequences of mutations affecting the first AUG codon. These variations may probably behave as benign mutations because of the high MAF seen in most of these cases. We can speculate that these small variants may already contribute to the protein heterogeneity generated by the genome and accordingly will have limited, if any, pathogenic relevance (Chorev *et al.*, 2015; de Klerk and 't Hoen, 2015; Asano, 2014).
2. In contrast, variations in genes with a distant or null alternative initiation methionine will cause a relevant modification of the affected protein, and this action will probably have pathogenic consequences. The low MAF described for these gene variations strongly supports a pathogenic effect of these mutations. Moreover, the resulting protein, if produced, will probably be shorter than the wild type. We may predict 3 consequences for these kinds of mutations:

    - If the alternative methionine maintains the wild-type reading frame, the resulting proteins will lose the N-terminal domain of the wild-type molecule, with the functional consequences of the missing domain in each case. These mutations will follow a classical loss-of-function pathogenic mechanism.
    - If the deleted portion of the molecule includes or affects the signal peptide, the resulting protein will probably have a different cell location, probably intracellular. Thus, a protein that is originally secreted or located at the cell membrane may reach unexpected locations where it may perform completely new and unknown functions. Our group recently characterised an example of this type of mutation. The *SERPINC1* c.3 G>T mutation, which is not included in our TSV file because the Ensembl database does not contain it, affects antithrombin, a key natural anticoagulant serpin that is secreted into blood plasma. On the other hand, mutation of the AUG initiation codon results in the use of 2 potential alternative downstream AUG initiation codons that are located after the signal peptide. The resulting proteins, which are expressed abundantly, are expressed in the cytoplasm and may be involved in the severe clinical phenotype observed in the carriers of this mutation (Navarro-Fernández *et al.*, 2017).
    - If the alternative initiation methionine changes the reading frame of the wild-type molecule, a new protein may be generated. In the last 2 frameworks, the protein generated by the mutated allele may gain completely new and potentially unpredictable functions that may be involved in completely new mechanisms of gain-of-function in unexpected disorders where the wild-type protein will never be involved. Therefore, current enrichment methods used to identify pathogenic mutations by whole-exome or whole-genome analysis associated with various disorders might fail with mutations affecting initiation codons. Further analysis of exome or genome results must be conducted with this new perspective in mind, particularly among cases where the causative mutation has been elusive.

Nevertheless, our study also identified examples that do not fit this classification. Thus, there are examples of very close alternative methionines that however have very low MAF. The best example is SNP rs149191301 (ATG/TTG; MAF = 0.0002) affecting the canonical transcript of the *SLAMF8* gene, which produces a CD2 family protein that may participate in B-lineage commitment and/or modulation of signaling through the B-cell receptor (Kingsbury *et al.*, 2001). According to our TSV file, this transcript has a signal peptide along its first 68 bp of the coding region. It also contains an alternative initiation codon, found by approaches 1 and 3, maintaining the reading frame at 6 bp downstream of the wild-type initiation codon. An isoform could be formed when the alternative initiation codon is employed, maintaining 99.3% of the wild-type protein length and 91.3% of the wild-type signal peptide. Despite these features, the low MAF may indicate a pathological consequence, maybe due to the relevance of the 2 codons lost. Our approach may help to identify key N-terminal domains in proteins. Further experimental analysis must be conducted to check the relevance of the residues eliminated by these mutations.

On the other hand, there are also examples of high-MAF variations affecting AUG initiation codons in transcripts without viable alternative TISs. An example is SNP rs80097179 (ATG/CTG, ATG/GTG) (MAF = 0.376) affecting the canonical transcript of the *ADSSL1* gene, which encodes a muscle-specific enzyme with strong expression in skeletal muscle. The lack of this enzyme may cause severe distal myopathy (Park *et al.*, 2016). The nearest alternative AUG initiation codon is found 250 bp downstream of the wild-type one in a different reading frame. The high MAF of this variation is indicative of the existence of protection mechanisms against variations in the initiation codon. We recently proposed that these mutations might be bypassed if alternative non-AUG codons are used (Navarro-Fernández *et al.*, 2017). Strong Kozak sequences may enable translation initiation from non-AUG codons (Ivanov *et al.*, 2011). Under these conditions, mutations affecting AUG initiation codons may be classified as silent genetic or missense depending on the use of Met-tRNA to initiate translation, regardless of the codon present in the RNA (Peabody, 1989; Starck *et al.*, 2012).

Another interesting example is variation rs11107, discussed in section 3.1; rs11107 is predicted to be non-deleterious by PolyPhen and SIFT.

This variation affects transcript ENST00000397426, which produces isoform 3 of the FBX7 protein. The first methionine of this protein is aligned to an isoleucine at position 114 of Q2T9S7 by the SIFT and PolyPhen algorithms. Therefore, the change M1I is supposed to be benign. Nonetheless, translation will never start at the 'ATA' codon. The features included in our TSV file show an alternative initiation codon at 7 bp downstream of the wild-type one, losing the reading frame, found by the 3 approaches. These characteristics seem to indicate that the variation does not permit the production of isoform 3 of FBX7. The high MAF of this variation may be explained by the existence of several isoforms in which the affected methionine is not the initiation codon; therefore, the change M/I becomes viable in these cases. Moreover, the TSV file includes links to publications that discuss this variation.

In conclusion, the data shown in this manuscript suggest that we are lost in translation. New tools that take into account the features shown in this study are necessary to predict the effect of gene variations affecting AUG initiation codons. This effect may be absolutely unexpected and not related to the functions originally assigned to the wild-type protein.

5 Funding

This work has been supported by Instituto de Salud Carlos III/ Fondo Europeo de Desarrollo Regional [PI15/00079 & CB15/00055]; and Fundación Séneca [19873/GERM/15].

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, **7**(4), 248.

Asano, K. (2014). Why is start codon selection so precise in eukaryotes? *Translation*, **2**(1), e28387.

Böhm, S., Jin, H., Hughes, S. M., Roberts, R. G., and Hinits, Y. (2008). Dystrobrevin and dystrophin family gene expression in zebrafish. *Gene Expression Patterns*, **8**(2), 71–78.

Cazzola, M. and Skoda, R. C. (2000). Translational pathophysiology: a novel molecular mechanism of human disease. *Blood*, **95**(11), 3280–3288.

Cenciarelli, C., Chiaur, D. S., Guardavaccaro, D., Parks, W., Vidal, M., and Pagano, M. (1999). Identification of a family of human F-box proteins. *Current Biology*, **9**(20), 1177–1179.

Cervantes, J. L., Weinerman, B., Basole, C., and Salazar, J. C. (2012). TLR8: the forgotten relative revindicated. *Cell Mol Immunol*, **9**(6), 434–438.

Chorev, D. S., Ben-Nissan, G., and Sharon, M. (2015). Exposing the subunit diversity and modularity of protein complexes by structural mass spectrometry approaches.

Cobaleda, C., Schebesta, A., Delogu, A., and Busslinger, M. (2007). Pax5: the guardian of B cell identity and function. *Nature immunology*, **8**(5), 463–70.

Corral, J., Lozano, M. L., Gonzalez-Conejero, R., Martinez, C., Iniesta, J. A., Rivera, J., and Vicente, V. (2000). A common polymorphism flanking the ATG initiator codon of GPIb does not affect expression and is not a major risk factor for arterial thrombosis. *Thromb Haemost*, **83**, 23–28.

Corral, J., Anton, A. I., Quiroga, T., Gonzalez-Conejero, R., Pereira, J., Roldan, V., Vicente, V., and Mezzano, D. (2010). Influence of the F12-4 C> T polymorphism on hemostatic tests. *Blood Coagulation & Fibrinolysis*, **21**(7), 632–639.

de Klerk, E. and 't Hoen, P. A. (2015). Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., and Others (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, **43**(D1), D805—-D811.

González-Conejero, R., Corral, J., Roldán, V., Martínez, C., Marín, F., Rivera, J., Iniesta, J. A., Lozano, M. L., Marco, P., and Vicente, V. (2002). A common polymorphism in the annexin V Kozak sequence (- 1C> T) increases translation efficiency and plasma levels of annexin V, and decreases the risk of myocardial infarction in young patients. *Blood*, **100**(6), 2081–2086.

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 Research*, **39**(10), 4220–4234.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, **338**(5), 1027–1036.

Kingsbury, G. A., Feeney, L. A., Nong, Y., Calandra, S. A., Murphy, C. J., Corcoran, J. M., Wang, Y., Prabhu Das, M. R., Busfield, S. J., Fraser, C. C., and Villeval, J. L. (2001). Cloning, expression, and function of BLAME, a novel member of the CD2 family. *Journal of immunology*, **166**(9), 5675–5680.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**(2), 283–292.

Kozak, M. (2002). Emerging links between initiation of translation and human diseases.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., and Others (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, **44**(D1), D862—-D868.

Lorsch, J. and Dever, T. (2010). Molecular view of 43S complex formation and start site selection in eukaryotic translation initiation. *The Journal of biological chemistry*, **285**(28), 21203–21207.

Navarro-Fernández, J., Dybedal, I., Águila, S., Bohdam, N., Corrales, F., Miqueo, C., Andresen, M., Ferrer, F., Tjønnfjord, G. E., Martínez-Martínez, I., Heimdal, K., Vicente, V., Corral, J., and Abildgaard, U. (2017). C0380: Clinical and Biochemical Consequences of Met1Ileu Mutation in Serpinc1 Gene: Generation of a Small Non-Inhibitory Antithrombin Variant without the N-Terminal Region by Use of an Alternative Initiation Codon that Has a Strong Gain-Of-Function Associ. *Thrombosis Research*, **133**, S11.

Ng, P. C. and Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, **31**(13), 3812–3814.

Noderer, W. L., Flockhart, R. J., Bhaduri, A., de Arce, A. J. D., Zhang, J., Khavari, P. A., and Wang, C. L. (2014). Quantitative analysis of mammalian translation initiation sites by facs-seq. *Molecular systems biology*, **10**(8), 748.

Park, H. J., Hong, Y. B., Choi, Y. C., Lee, J., Kim, E. J., Lee, J. S., Mo, W. M., Ki, S. M., Kim, H. I., Kim, H. J., Hyun, Y. S., Hong, H. D., Nam, K., Jung, S. C., Kim, S. B., Kim, S. H., Kim, D. H., Oh, K. W., Kim, S. H., Yoo, J. H., Lee, J. E., Chung, K. W., and Choi, B. O. (2016). ADSSL1 mutation relevant to autosomal recessive adolescent onset distal myopathy. *Annals of Neurology*, **79**(2), 231–243.

Peabody, D. S. (1989). Translation initiation at non-AUG triplets in mammalian cells. *Journal of Biological Chemistry*, **264**(9), 5031–5035.

Preiss, T. and W Hentze, M. (2003). Starting the protein synthesis machine: eukaryotic translation initiation. *Bioessays*, **25**(12), 1201–1211.

Qu, X., Wei, H., Zhai, Y., Que, H., Chen, Q., Tang, F., Wu, Y., Xing, G., Zhu, Y., Liu, S., Fan, M., and He, F. (2002). Identification, characterization, and functional study of the two novel human members of the semaphorin gene family. *The Journal of biological chemistry*, **277**(38), 35574–35585.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, **29**(1), 308–311.

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**(6089), 1719 LP – 1723.

Tsuji, S., Uehori, J., Matsumoto, M., Suzuki, Y., Matsuhisa, A., Toyoshima, K., and Seya, T. (2001). Human Intelectin Is a Novel Soluble Lectin That Recognizes Galactofuranose in Carbohydrate Chains of Bacterial Cell Wall. *Journal of Biological Chemistry*, **276**(26), 23456–23463.

Wesener, D. A., Wangkanont, K., McBride, R., Song, X., Kraft, M. B., Hodges, H. L., Zarling, L. C., Splain, R. A., Smith, D. F., Cummings, R. D., Paulson, J. C., Forest, K. T., and Kiessling, L. L. (2015). Recognition of microbial glycans by human intelectin-1. *Nature Structural & Molecular Biology*, **22**(8), 603–610.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**(6), 80.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. (2015). The Ensembl REST API: ensembl data for any language. *Bioinformatics*, **31**(1), 143–145.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., and Others (2016). Ensembl 2016. *Nucleic acids research*, **44**(D1), D710—-D716.

Yates, F. (1934). Contingency Tables Involving Small Numbers and the $\chi 2$ Test. *Supplement to the Journal of the Royal Statistical Society*, **1**(2), 217.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, **16**(5), 284–287.