

2016-09-06

A Common Class of Transcripts with 5'-Intron Depletion, Distinct Early Coding Sequence Features, and N1-Methyladenosine Modification

Can Cenik
University of Massachusetts Medical School

Et al.

Let us know how access to this document benefits you.

Follow this and additional works at: https://escholarship.umassmed.edu/faculty_pubs



Part of the [Amino Acids, Peptides, and Proteins Commons](#), [Bioinformatics Commons](#), [Genetic Phenomena Commons](#), and the [Nucleic Acids, Nucleotides, and Nucleosides Commons](#)

Repository Citation

Cenik C, Chua HN, Singh G, Akef A, Snyder MP, Palazzo AF, Moore MJ, Roth FP. (2016). A Common Class of Transcripts with 5'-Intron Depletion, Distinct Early Coding Sequence Features, and N1-Methyladenosine Modification. University of Massachusetts Medical School Faculty Publications. <https://doi.org/10.1101/057455>. Retrieved from https://escholarship.umassmed.edu/faculty_pubs/1561

Creative Commons License



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 License](#). This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts Medical School Faculty Publications by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.

Can Cenik

1 **A Common Class of Transcripts with 5'-Intron Depletion, Distinct Early Coding**
2 **Sequence Features, and N¹-Methyladenosine Modification**

3 Can Cenik^{1,2}, Hon Nian Chua^{3,4}, Guramrit Singh^{2,5,7,8}, Abdalla Akef⁶, Michael P Snyder¹,
4 Alexander F. Palazzo⁶, Melissa J Moore^{2,7,8#}, and Frederick P Roth^{3,9,10#}

5
6 **Affiliations:**

7 ¹ Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

8 ² Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical
9 School, Worcester, MA 01605, USA

10 ³ Donnelly Centre and Departments of Molecular Genetics and Computer Science, University of Toronto
11 and Lunenfeld-Tanenbaum Research Institute, Mt Sinai Hospital, Toronto M5G 1X5, Ontario, Canada

12 ⁴ DataRobot, Inc., 61 Chatham St. Boston MA 02109, USA

13 ⁵ Department of Molecular Genetics, Center for RNA Biology, The Ohio State University, Columbus, OH
14 43210, USA

15 ⁶ Department of Biochemistry, University of Toronto, 1 King's College Circle, MSB Room 5336, Toronto,
16 ON M5S 1A8, Canada

17 ⁷ Howard Hughes Medical Institute, University of Massachusetts Medical School, Worcester, MA 01605,
18 USA

19 ⁸ RNA Therapeutics Institute, University of Massachusetts Medical School, Worcester, MA 01605, USA

20 ⁹ Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston 02215, MA, USA

21 ¹⁰ The Canadian Institute for Advanced Research, Toronto M5G 1Z8, ON, Canada

22

23 #Correspondence to: Melissa.Moore@umassmed.edu , Fritz.Roth@utoronto.ca

24

25 **Short Title:** First intron position defines a transcript class

26 **Keywords:** 5' UTR introns, random forest, N¹ methyladenosine, Exon Junction Complex

27

Can Cenik

1 **Abstract:**

2 Introns are found in 5' untranslated regions (5'UTRs) for 35% of all human transcripts.
3 These 5'UTR introns are not randomly distributed: genes that encode secreted, membrane-
4 bound and mitochondrial proteins are less likely to have them. Curiously, transcripts lacking
5 5'UTR introns tend to harbor specific RNA sequence elements in their early coding regions. To
6 model and understand the connection between coding-region sequence and 5'UTR intron
7 status, we developed a classifier that can predict 5'UTR intron status with >80% accuracy
8 using only sequence features in the early coding region. Thus, the classifier identifies
9 transcripts with 5' proximal-intron-minus-like-coding regions ("5IM" transcripts).
10 Unexpectedly, we found that the early coding sequence features defining 5IM transcripts are
11 widespread, appearing in 21% of all human RefSeq transcripts. The 5IM class of transcripts is
12 enriched for non-AUG start codons, more extensive secondary structure both preceding the
13 start codon and near the 5' cap, greater dependence on eIF4E for translation, and association
14 with ER-proximal ribosomes. 5IM transcripts are bound by the Exon Junction Complex (EJC) at
15 non-canonical 5' proximal positions. Finally, N¹-methyladenosines are specifically enriched in
16 the early coding regions of 5IM transcripts. Taken together, our analyses point to the existence
17 of a distinct 5IM class comprising ~20% of human transcripts. This class is defined by
18 depletion of 5' proximal introns, presence of specific RNA sequence features associated with
19 low translation efficiency, N¹-methyladenosines in the early coding region, and enrichment for
20 non-canonical binding by the Exon Junction Complex.

21

Can Cenik

1 **Introduction:**

2

3 Approximately 35% of all human transcripts harbor introns in their 5' untranslated
4 regions (5'UTRs) (Cenik et al. 2010; Hong et al. 2006). Among genes with 5'UTR introns (5UIs),
5 those annotated as "regulatory" are significantly overrepresented, while there is an under-
6 representation of genes encoding proteins that are targeted to either the endoplasmic
7 reticulum (ER) or mitochondria (Cenik et al. 2011). For transcripts that encode ER- and
8 mitochondria-targeted proteins, 5UI depletion is associated with presence of specific RNA
9 sequences (Cenik et al. 2011; Palazzo et al. 2007, 2013). Specifically, nuclear export of an
10 otherwise inefficiently exported microinjected mRNA or cDNA transcript can be promoted by
11 an ER-targeting signal sequence-containing region (SSCRs) or mitochondrial signal sequence
12 coding region (MSCRs) from a gene lacking 5' UTR introns (Cenik et al. 2011; Lee et al. 2015).
13 However, more recent studies suggest that many SSCRs have little impact on nuclear export for
14 RNAs transcribed *in vivo* (Lee et al. 2015), but rather enhance translation in a RanBP2-
15 dependent manner (Mahadevan et al. 2013).

16 Among SSCR- and MSCR-containing transcripts (referred to hereafter as SSCR and
17 MSCR transcripts), ~75% lack 5' UTR introns ("5UI-" transcripts) and ~25% have them ("5UI+"
18 transcripts). These two groups have markedly different sequence compositions at the 5' ends
19 of their coding sequences. 5UI- transcripts tend to have lower adenine content (Palazzo et al.
20 2007) and use codons with fewer uracils and adenines than 5UI+ transcripts (Cenik et al.
21 2011). Their signal sequences also contain leucine and arginine more often than the
22 biochemically similar amino acids isoleucine and lysine, respectively. Leucine and arginine
23 codons contain fewer adenine and thymine nucleotides, consistent with adenine and thymine

Can Cenik

1 depletion. This depletion is also associated with the presence of a specific GC-rich RNA motif
2 in the early coding region of 5UI⁻ transcripts (Cenik et al. 2011).

3 Despite some knowledge as to their early coding region features, key questions about
4 this class of 5UI⁻ transcripts have remained unanswered: Do the above sequence features
5 extend beyond SSCR- and MSCR-containing transcripts to other 5UI⁻ genes? Do 5UI⁻
6 transcripts having these features share common functional or regulatory features? What
7 binding factor(s) recognize these RNA elements? A more complete model of the relationship of
8 early coding features and 5UI⁻ status would begin to address these questions.

9 Here, to better understand the relationship between early coding region features and
10 5UI status, we undertook an integrative machine learning approach. We reasoned that a
11 machine learning classifier which could identify 5UI⁻ transcripts solely from early coding
12 sequence would potentially provide two types of insight: First, it could systematically identify
13 predictive features. Second, the subset of 5UI⁻ transcripts which could be identified by the
14 classifier might then represent a functionally distinct transcript class. Having developed such
15 a classifier, we found that it identified ~21% of all human transcripts as harboring coding
16 regions characteristic of 5UI⁻ transcripts. While many of these transcripts encode ER- and
17 mitochondrial-targeted proteins, many others encode nuclear and cytoplasmic proteins. This
18 class of transcripts shares characteristics such as a tendency to lack 5' proximal introns, to
19 contain non-canonical Exon Junction Complex binding sites, to have multiple features
20 associated with lower intrinsic translation efficiency, and to have an increased incidence of N¹-
21 methyladenosine modification.

22

23 **Results:**

Can Cenik

1 *A classifier that predicts 5UI status using only early coding sequence information.*

2 To better understand the previously-reported enigmatic relationship between certain early
3 coding region sequences and the absence of a 5UI, we sought to model this relationship.
4 Specifically, we used a random forest classifier (Breiman 2001) to learn the relationship
5 between 5UI absence and a collection of 36 different nucleotide-level features extracted from
6 the first 99 nucleotides of all human coding regions (CDS) (**Figs 1A-C; Table S1; Methods**).
7 We then used all transcripts known to contain an SSCR (a total of 3743 transcripts clusters;
8 **Methods**), regardless of 5UI status, as our training set. This training constraint ensured that all
9 input nucleotide sequences were subject to similar functional constraints at the protein level.
10 Thus, we sought to identify sequence features that differ between 5UI⁻ and 5UI⁺ transcripts at
11 the RNA level.

12 Our classifier assigns to each transcript a “5’UTR-intron-minus-predictor” (5IMP) score
13 between 0 and 10, where higher scores correspond to a higher likelihood of being 5UI⁻ (**Fig**
14 **1C**). Interestingly, preliminary ranking of the 5UI⁻ transcripts by 5IMP score revealed a
15 relationship between the position of the first intron in the coding region and the 5IMP score.
16 5UI⁻ transcripts for which the first intron was more than 85 nts downstream of the start codon
17 had the highest 5IMP scores. Furthermore, the closer the first intron was to the start codon,
18 the lower the 5IMP score (**Fig 1D**). We explored this relationship further by training classifiers
19 that increasingly excluded from the training set 5UI⁻ transcripts according to the distance of
20 the first intron from the 5’ end of the coding region. This revealed that classifier performance,
21 as measured by the area under the precision recall curve (AUPRC), increased as a function of
22 the distance from start codon to first intron distance (**Methods, Fig 1E**). Thus, the RNA
23 sequence features we identified as being predictive of 5UI⁻ transcripts are more accurately
24 described as being predictors of transcripts without 5’-proximal introns.

Can Cenik

1 To minimize the impact of transcripts that may ‘behave’ as though they were 5UI⁺ due
2 to an intron early in the coding region, we eliminated 5UI⁻ SSCR transcripts with a first intron
3 <90 nts downstream of the start codon (Methods) and generated a new classifier.
4 Discriminative motif features were learned independently (Methods), and performance of this
5 new classifier was gauged using 10-fold cross validation. We assessed cross-validation
6 performance in two ways: 1) in terms of the area under the receiver operating curve (AUC)—
7 which can be thought of as a measure of average recall across a range of false positive rates; 2)
8 in terms of area under the precision vs recall curve (AUPRC), which can be thought of as the
9 average precision (fraction of predictions which are correct) across a range of recall values.
10 Specifically, the classifier showed an AUC of 74% and AUPRC of 88% (**Fig 2A, yellow curves**;
11 exceeding AUC 50% and AUPRC 71%, the performance value expected of a naïve predictor).
12 We used this optimized classifier for all subsequent analyses.

13 SSCR transcripts exhibited markedly different 5IMP score distributions for the 5UI⁺ and
14 5UI⁻ subsets (**Fig 2B**). The 5UI⁺ score distribution was unimodal with a peak at ~2.4. In
15 contrast, the 5UI⁻ score distribution was bimodal with one peak at ~3.6 and another at ~9,
16 suggesting the existence of at least two underlying 5UI⁻ transcript classes. The peak at score
17 3.6 resembled the 5UI⁺ peak. Also contributing to the peak at 3.6 is the set of 5UI⁻ transcripts
18 harboring an intron in the first 90 nts of the CDS (55% of all 5UI⁻ transcripts). The other
19 distinct high-scoring 5UI⁻ class (peak at score 9) is composed of transcripts that have specific
20 5UI⁻-predictive RNA sequence elements within the early coding region.

21 We next wished to evaluate whether our classifier was discriminating 5UI⁺ and 5UI⁻
22 SSCR transcripts using signals that appear specifically in the early coding region as opposed to
23 signals that appear broadly across the coding region. To do so, for every transcript we

Can Cenik

1 randomly chose 99 nts from the region downstream of the 3rd exon. The 5IMP score
2 distributions of these '3' proximal exon' sets were identical for 5UI⁺ and 5UI⁻ transcripts (**Figs**
3 **2A, and 2C**), confirming that the sequence features that distinguish 5UI⁺ and 5UI⁻ transcripts
4 are specific to the early coding region.

5

6 *RNA elements associated with 5UI⁻ transcripts are pervasive in the human genome*

7 Having trained the classifier on SSCR transcripts, we wondered how well it would predict the
8 5UI status of other transcripts. Despite having been trained exclusively on SSCR transcripts,
9 the classifier performed remarkably well on MSCR transcripts, achieving an AUC of 86% and
10 AUPRC of 95% (**Fig 2A, purple line; Fig 2D**; as compared with 50% and 77%, respectively,
11 expected by chance). This result suggests that RNA elements within early coding regions of
12 5UI⁻ MSCR-transcripts are similar to those in 5UI⁻ SSCR-transcripts despite distinct functional
13 constraints at the protein level.

14 We next wondered whether the class of 5UI⁻ transcripts that can be predicted on the
15 basis of early coding region features is restricted to transcripts encoding proteins trafficked to
16 the ER or mitochondria, or is instead a more general class of transcripts. We therefore asked
17 whether the classifier could predict 5UI⁻ status in transcripts that contain neither an SSCR nor
18 an MSCR ("S-/M-" transcripts). Because unannotated SSCRs could confound this analysis, we
19 first used SignalP 3.0 to identify S-/M- transcripts most likely to contain an unannotated SSCR
20 (Bendtsen et al. 2004). These 'SignalP+' transcripts had a 5IMP score distribution comparable
21 to those of known SSCR and MSCR transcripts (**Fig 2E**), and the classifier worked well to
22 identify the 5UI⁻ subset of these transcripts (AUC 82% and AUPRC 95%, **Fig 2A, light blue**
23 **line**). While 5UI⁺ SignalP+ transcripts had predominantly low 5IMP scores, 5UI⁻ SignalP+ 5IMP

Can Cenik

1 scores were strongly skewed towards high 5IMP scores (peak at ~9; **Fig 2E**). These results
2 were consistent with the idea that SignalP⁺ transcripts do contain many unannotated SSCRs.

3 Having considered SignalP⁺ transcripts as well as SSCR- and MSCR-containing
4 transcripts, we used the classifier to calculate 5IMP scores for all remaining “S⁻/M⁻/SignalP⁻”
5 transcripts. Although the performance was weaker on this gene set, it was still better than
6 expected of a naive predictor (**Fig 2A, green line**). 5UI⁺ S⁻/M⁻/SignalP⁻ transcripts were
7 strongly skewed toward low 5IMP scores (**Fig 2F**). Surprisingly, however, a significant
8 fraction of 5UI⁻ S⁻/M⁻/SignalP⁻ transcripts had high 5IMP scores (~18%). Thus, our results
9 suggest a broad class of transcripts with early coding regions carrying sequence signals that
10 predict the absence of a 5'proximal intron, or in other words, a class of transcripts with
11 5'proximal-intron-minus-like coding regions. Hereafter we refer to transcripts in this class as
12 “5IM” transcripts.

13 We sought to identify what fraction of transcripts have 5IMP scores that exceed what
14 would be expected in the absence of 5UI⁻-predictive early coding region signals. To establish
15 this expectation, we used the above-described negative control set of equal-length coding
16 sequences from outside of the early coding region. By quantifying the excess of high-scoring
17 sequences in the real distribution relative to this control distribution, we estimate that 21% of
18 all human transcripts are 5IM transcripts (a 5IMP score of 7.41 corresponds to a 5% False
19 Discovery Rate; **Fig 2G**). The set of 5IM transcripts defined by our classifier (**Table S2**)
20 includes many that do not encode ER-targeted or mitochondrial proteins. The distribution of
21 various classes of mRNAs among the 5IM transcripts was: 38% ER-targeted (SSCR or SignalP⁺),
22 9% mitochondrial (MSCR) and 53% other classes (S⁻/M⁻/SignalP⁻) (Figure 2G). These results
23 suggest that RNA-level features prevalent in the early coding regions of 5UI⁻ SSCR and MSCR

Can Cenik

1 transcripts are also found in other transcript types (**Figs 2F-G**), and that 5IM transcripts
2 represent a broad class.

3

4 *Functional characterization of 5IM transcripts*

5 5IM transcripts are defined by mRNA sequence features. Hence, we hypothesized that 5IM
6 transcripts may be functionally related through shared regulatory mechanisms mediated by
7 the presence of these common features. To this end, we collected large-scale datasets
8 representing diverse attributes covering six broad categories (see Table S3 for a complete list):
9 (1) Curated functional annotations -- e.g., Gene Ontology terms, annotation as a 'housekeeping'
10 gene, genes subject to RNA editing; (2) RNA localization -- e.g., to dendrites, to mitochondria;
11 (3) Protein and mRNA half-life, ribosome occupancy and features that decrease stability of one
12 or more mRNA isoforms -- e.g., AU-rich elements (4) Sequence features associated with
13 regulated translation -- e.g., codon optimality, secondary structure near the start codon; (5)
14 Known interactions with RNA-binding proteins or complexes such as Staufen-1, TDP-43, or the
15 Exon Junction Complex (EJC) (6) RNA modifications – i.e., N¹-methyladenosine (m¹A).

16 We adjusted for multiple hypotheses testing at two levels. First, we took a conservative
17 approach (Bonferroni correction) to correct for the number of tested functional
18 characteristics. Second, some of the functional categories were analyzed in more depth and
19 multiple sub-hypotheses were tested within the given category. In this in-depth analysis a false
20 discovery-based correction was adopted. Below, all reported p-values remain significant (p-
21 adjusted < 0.05) after multiple hypothesis test correction.

Can Cenik

1 No associations between 5IM transcripts and features in categories (1), (2) and (3) were
2 found, other than the already-known enrichments for ER- and mitochondrial-targeted mRNAs.
3 However, analyses for the remaining categories yielded the significant results described below.

4

5 *5IM transcripts have features suggesting lower translation efficiency*

6 Translation regulation is a major determinant of protein levels (Vogel and Marcotte 2012). To
7 investigate potential connections between 5IM transcripts and translational regulation, we
8 examined features associated with translation. Features found to be significant were:

9 (I) Secondary structures near the start codon can affect initiation rate by modulating
10 start codon recognition (Parsyan et al. 2011). We observed a positive correlation between
11 5IMP score and the free energy of folding ($-\Delta G$) of the 35 nucleotides immediately preceding
12 the start codon (**Fig 3A**; Spearman $\rho=0.39$; $p < 2.2e-16$). This suggests that 5IM transcripts
13 have a greater tendency for secondary structure near the start codon, presumably making the
14 start codon less accessible.

15 (II) Similarly, secondary structures near the 5' cap can modulate translation by
16 hindering binding by the 43S-preinitiation complex to the mRNA (Babendure et al. 2006). We
17 observed a positive correlation between 5IMP score and the free energy of folding ($-\Delta G$) of the
18 5'most 35 nucleotides (**Fig 3B**; Spearman $\rho=0.18$; $p = 7.9e-130$). This suggests that 5IM
19 transcripts have a greater tendency for secondary structure near the 5' cap, presumably
20 hindering binding by the 43S-preinitiation complex.

21 (III) eIF4E overexpression. The heterotrimeric translation initiation complex eIF4F
22 (made up of eIF4A, eIF4E and eIF4G) is responsible for facilitating the translation of

Can Cenik

1 transcripts with strong 5'UTR secondary structures (Parsyan et al. 2011). The eIF4E subunit
2 binds to the 7mGpppG 'methyl-G' cap and the ATP-dependent helicase eIF4A (scaffolded by
3 eIF4G) destabilizes 5'UTR secondary structure (Marintchev et al. 2009). A previous study
4 identified transcripts that were more actively translated under conditions that promote cap-
5 dependent translation (overexpression of eIF4E) (Larsson et al. 2007). In agreement with the
6 observation that 5IM transcripts have more secondary structure upstream of the start codon
7 and near the 5'cap, transcripts with high 5IMP scores were more likely to be translationally,
8 but not transcriptionally, upregulated upon eIF4E overexpression (**Fig 3C**; Wilcoxon Rank Sum
9 Test $p = 2.05e-22$, and $p = 0.28$, respectively).

10 (IV) Non-AUG start codons. Transcripts with non-AUG start codons also have
11 intrinsically low translation initiation efficiencies (Hinnebusch and Lorsch 2012). These
12 mRNAs were greatly enriched among transcripts with high 5IMP scores (Fisher's Exact Test p
13 = 0.0003; odds ratio = 3.9) and have a median 5IMP score that is 3.57 higher than those with
14 an AUG start (**Fig 3D**).

15 (V) Codon optimality. The efficiency of translation elongation is affected by codon
16 optimality (Hershberg and Petrov 2008). Although some aspects of this remain controversial
17 (Charneski and Hurst 2013; Shah et al. 2013; Zinshteyn and Gilbert 2013; Gerashchenko and
18 Gladyshev 2015), it is clear that decoding of codons by tRNAs with different abundances can
19 affect the translation rate under conditions of cellular stress (reviewed in Gingold and Pilpel
20 2011). We therefore examined the tRNA adaptation index (tAI), which correlates with copy
21 numbers of tRNA genes matching a given codon (dos Reis et al. 2004). Specifically, we
22 calculated the median tAI of the first 99 coding nucleotides of each transcript, and found that
23 5IMP score was negatively correlated with tAI (**Fig S1A**; Spearman Correlation $\rho = -0.23$; $p <$

Can Cenik

1 2e-16 median tAI and 5IMP score). This effect was restricted to the early coding regions as the
2 negative control set of randomly chosen sequences downstream of the 3rd exon from each
3 transcript did not exhibit a relationship between 5IMP score and codon optimality (**Fig S1B-C**).
4 Thus, 5IM transcripts show reduced codon optimality in early coding regions, suggesting that
5 5IM transcripts have decreased translation elongation efficiency.

6 To more precisely determine where the codon optimality phenomenon occurs within
7 the entire early coding region, we grouped transcripts by 5IMP score. For each group, we
8 calculated the mean tAI at codons 2-33 (i.e., nts 4-99). Across this entire region, 5IM
9 transcripts (5IMP >7.41; 5% FDR) had significantly lower tAI values at every codon except
10 codons 24 and 32 (**Fig 3E**; Wilcoxon Rank Sum test Holm-adjusted $p < 0.05$ for all
11 comparisons). To eliminate potential confounding variables, including nucleotide composition,
12 we performed several additional control analyses (Methods); none of these altered the basic
13 conclusion that 5IM transcripts have lower codon optimality than non-5IM transcripts across
14 the entire early coding region.

15 (VI) Ribosomes per mRNA. Finally, we examined the relationship between 5IMP score
16 and translation efficiency, as measured by the steady-state number of ribosomes per mRNA
17 molecule. To this end, we used a large dataset of ribosome profiling and RNA-Seq experiments
18 from human lymphoblastoid cell lines (Cenik et al. 2015). From this, we calculated the average
19 number of ribosomes on each transcript and identified transcripts with high or low ribosome
20 occupancy (respectively defined by occupancy at least one standard deviation above or below
21 the mean; see Methods). 5IM transcripts were slightly but significantly depleted in the high
22 ribosome-occupancy category (**Fig 3F**; Fisher's Exact Test $p = 0.0006$, odds ratio = 1.3).

Can Cenik

1 Moreover, 5IMP scores exhibited a weak but significant negative correlation with the number
2 of ribosomes per mRNA molecule (Spearman $\rho=-0.11$; $p = 5.98e-23$).

3 Taken together, all of the above results reveal that 5IM transcripts have sequence
4 features associated with lower translation efficiency, at the stages of both translation initiation
5 and elongation.

6

7 *Non-ER trafficked 5IM transcripts are enriched in ER-proximal ribosome occupancy*

8 We next investigated the relationship between 5IMP score and the localization of translation
9 within cells. Exploring the subcellular localization of translation at a transcriptome-scale
10 remains a significant challenge. Yet, a recent study described proximity-specific ribosome
11 profiling to identify mRNAs occupied by ER-proximal ribosomes in both yeast and human cells
12 (Jan et al. 2014). In this method, ribosomes are biotinylated based on their proximity to a
13 marker protein such as Sec61, which localizes to the ER membrane (Jan et al. 2014). For each
14 transcript, the enrichment for biotinylated ribosome occupancy yields a measure of ER-
15 proximity of translated mRNAs.

16 We reanalyzed this dataset to explore the relationship between 5IMP scores and ER-
17 proximal ribosome occupancy in HEK-293 cells. As expected, transcripts that exhibit the
18 highest enrichment for ER-proximal ribosomes were SSCR-containing transcripts and
19 transcripts with other ER-targeting signals. Yet, we noticed a surprising positive correlation
20 between ER-proximal ribosome occupancy and 5IM transcripts with no ER-targeting evidence
21 (**Fig 4**). This relationship was true for both mitochondrial genes (**Fig 4**; Spearman $\rho=0.43$; p
22 $< 2.2e-16$), and genes with no evidence for either ER- or mitochondrial-targeting (**Fig 4**;

Can Cenik

1 Spearman $\rho = -0.36$; $p < 2.2e-16$). These results suggest that 5IM transcripts are more likely
2 than non-5IM transcripts to engage with ER-proximal ribosomes.

3
4 *5IM transcripts are strongly enriched in non-canonical EJC occupancy sites*

5 Shared sequence features and functional traits among 5IM transcripts causes one to wonder
6 what common mechanisms might link 5IM sequence features to 5IM traits. For example, 5IM
7 transcripts might share regulation by one or more RNA-binding proteins (RBPs). To
8 investigate this idea further, we tested for enrichment of 5IM transcripts among the
9 experimentally identified targets of 23 different RBPs (including CLIP-Seq, and variants; see
10 Methods). Only one dataset was substantially enriched for high 5IMP scores among targets: a
11 transcriptome-wide map of binding sites of the Exon Junction Complex (EJC) in human cells,
12 obtained via tandem-immunoprecipitation followed by deep sequencing (RIPiT) (Singh et al.
13 2014, 2012). The EJC is a multi-protein complex that is stably deposited upstream of exon-
14 exon junctions as a consequence of pre-mRNA splicing (Le Hir et al. 2000). RIPiT data
15 confirmed that canonical EJC sites (cEJC sites; sites bound by EJC core factors and appearing
16 ~24 nts upstream of exon-exon junctions) occupy ~80% of all possible exon-exon junction
17 sites and are not associated with any sequence motif. Unexpectedly, many EJC-associated
18 footprints outside of the canonical -24 regions were observed (**Fig 5A**) (Singh et al. 2012).
19 These 'non-canonical' EJC occupancy sites (ncEJC sites) were associated with multiple
20 sequence motifs, three of which were similar to known recognition motifs for SR proteins that
21 co-purified with the EJC core subunits (Singh et al. 2012). Interestingly, another motif (**Fig 5B**;
22 top) that was specifically found in first exons is not known to be bound by any known RNA-
23 binding protein (Singh et al. 2012). This motif was CG-rich, a sequence feature that also

Can Cenik

1 defines 5IM transcripts. This similarity presages the possibility of enrichment of first exon
2 ncEJC sites among 5IM transcripts.

3 Position analysis of called EJC peaks revealed that while only 9% of cEJCs reside in first
4 exons, 19% of all ncEJCs are found there. When we investigated the relationship between 5IMP
5 scores and ncEJCs in early coding regions, we found a striking correspondence—the median
6 5IMP score was highest for transcripts with the greatest number of ncEJCs (**Fig 5C**; Wilcoxon
7 Rank Sum Test; $p < 0.0001$). When we repeated this analysis by conditioning on 5UI status, we
8 similarly found that ncEJCs were enriched among transcripts with high 5IMP scores regardless
9 of 5UI status (Fisher's Exact Test, $p < 3.16e-14$, odds ratio > 2.3 ; **Fig 5D**). These results suggest
10 that the striking enrichment of ncEJC peaks in early coding regions was generally applicable to
11 all transcripts with high 5IMP scores regardless of 5UI presence.

12 *Transcripts harboring N¹-methyladenosine (m¹A) have high 5IMP scores*

13 It is increasingly clear that ribonucleotide base modifications in mRNAs are highly prevalent
14 and can be a mechanism for post-transcriptional regulation (Frye et al. 2016). One RNA
15 modification present towards the 5' ends of mRNA transcripts is N¹-methyladenosine (m¹A)
16 (Li et al. 2016; Dominissini et al. 2016), initially identified in total RNA and rRNAs (Dunn 1961;
17 Hall 1963; Klootwijk and Planta 1973). Intriguingly, the position of m¹A modifications has
18 been shown to be more correlated with the position of the first intron than with
19 transcriptional or translational start sites (Figure 2g from Dominissini et al. 2016). When the
20 distance of m¹As to each splice site in a given mRNA was calculated, the first splice site was
21 found to be the nearest for 85% of m¹As (Dominissini et al. 2016). When 5'UTR introns were
22 present, m¹A was found to be near the first splice site regardless of the position of the start
23 codon (Dominissini et al. 2016). Given that 5IM transcripts are also characterized by the

Can Cenik

1 position of the first intron, we investigated the relationship between 5IMP score and m¹A RNA
2 modification marks.

3 We analyzed the union of previously identified m¹A modifications (Dominissini et al.
4 2016) across all cell types and conditions. Although there is some evidence that these marks
5 depend on cell type and growth condition, it is difficult to be confident of the cell type and
6 condition-dependence of any particular mark given experimental variation (see Methods).
7 Nevertheless, we found that mRNAs with m¹A modification early in the coding region (first 99
8 nucleotides) had substantially higher 5IMP scores than mRNAs lacking these marks (**Fig 6A**;
9 Wilcoxon Rank Sum Test $p = 3.4e-265$), and were greatly enriched among 5IM transcripts (**Fig**
10 **6A**; Fisher's Exact Test $p = 1.6e-177$; odds ratio = 3.8). In other words, the sequence features
11 within the early coding region that define 5IMP transcripts also associate with m¹A
12 modification in the early coding region.

13 We next wondered whether 5IMP score was related to m¹A modification generally, or
14 only associated with m¹A modification in the early coding region. Indeed, many of the
15 previously identified m¹A peaks were within the 5'UTRs of mRNAs (Li et al. 2016).
16 Interestingly, 5IMP scores were only associated with m¹A modification in the early coding
17 region, and not with m¹A modification in the 5'UTR (**Fig 6B**). This offers the intriguing
18 possibility that the sequence features that define 5IMP transcripts are co-localized with m¹A
19 modification.

20

21 **Discussion:**

Can Cenik

1 Coordinating the expression of functionally related transcripts can be achieved by post-
2 transcriptional processes such as splicing, RNA export, RNA localization or translation (Moore
3 and Proudfoot 2009). Sets of mRNAs subject to a common regulatory transcriptional process
4 can exhibit common sequence features that define them to be a class. For example, transcripts
5 subject to regulation by particular miRNAs tend to share certain sequences in their 3'UTRs that
6 are complementary to the regulatory miRNA (Ameres and Zamore 2013). Similarly, transcripts
7 that share a 5' terminal oligopyrimidine tract are coordinately regulated by mTOR and
8 ribosomal protein S6 kinase (Meyuhas 2000). Here we quantitatively define '5IM' transcripts
9 as a class that shares common sequence elements and functional properties. We estimate the
10 5IM class to comprise 21% of all human transcripts.

11 Whereas 35% of human transcripts have one or more 5'UTR introns, the majority of
12 5IM transcripts have neither a 5'UTR intron nor an intron in the first 90 nts of the ORF. Other
13 shared features of 5IM transcripts include sequence features associated with low translation
14 initiation rates. These are: (1) a tendency for RNA secondary structure in the region
15 immediately preceding the start codon (**Fig 3A**), and near the 5' cap (**Fig 3B**); (2) translational
16 upregulation upon overexpression of eIF4E (**Fig 3C**); and (3) more frequent use of non-AUG
17 start codons (**Fig 3D**). Also consistent with low intrinsic translation efficiencies, 5IM
18 transcripts additionally tend to depend on less abundant tRNAs to decode the beginning of the
19 open reading frame (**Fig 3E**).

20 We had previously reported that transcripts encoding proteins with ER- and
21 mitochondrial-targeting signal sequences (SSCRs and MSCRs, respectively) are over-
22 represented among the 65% of transcripts lacking 5'UTR introns (Cenik et al. 2011).
23 Transcripts in this set are enriched for the sequence features detected by our 5IM classifier. By

Can Cenik

1 examining these enriched sequence features, we showed that the 5IM class extends beyond
2 mRNAs encoding membrane proteins. Jan et al. recently developed a transcriptome-scale
3 method to identify mRNAs occupied by ER-proximal ribosomes in both yeast and human cells
4 (Jan et al. 2014). As expected, transcripts known to encode ER-trafficked proteins were highly
5 enriched for ER-proximal ribosome occupancy. However, their data also showed many
6 transcripts encoding non-ER trafficked proteins to also be engaged with ER-proximal
7 ribosomes (Reid and Nicchitta 2015b). Similarly, several other studies have suggested a critical
8 role of ER-proximal ribosomes in translating several cytoplasmic proteins (Reid and Nicchitta
9 2015a). Here, we found that 5IM transcripts including those that are not ER-trafficked or
10 mitochondrial were significantly more likely to exhibit binding to ER-proximal ribosomes
11 (**Figs 4A-B**).

12 In addition to ribosomes directly resident on the ER, an interesting possibility is the
13 presence of a pool of peri-ER ribosomes (Jan et al. 2015; Reid and Nicchitta 2015a).
14 Association of 5IM transcripts with such a peri-ER ribosome pool could potentially explain the
15 observed correlation of 5IM status with binding to ER-proximal ribosomes. The ER is
16 physically proximal to mitochondria (Rowland and Voeltz 2012), so peri-ER ribosomes may
17 include those translating mRNAs on mitochondria (i.e., mRNAs with MSCRs) (Sylvestre et al.
18 2003). However, even when transcripts corresponding to ER-trafficked and mitochondrial
19 proteins were excluded from consideration, ER-proximal ribosome enrichment and 5IMP
20 scores were highly correlated (**Fig 4A**). Thus another shared feature of 5IM transcripts is their
21 translation on or near the ER regardless of the ultimate destination of the encoded protein.

22 In an attempt to identify a common factor binding 5IM transcripts, we asked whether
23 5IM transcripts were enriched among the experimentally identified targets of 23 different

Can Cenik

1 RBPs. Only one RBP emerged--the exon junction complex (EJC). Specifically, we observed a
2 dramatic enrichment of non-canonical EJC (ncEJC) binding sites within the early coding region
3 of 5IM transcripts. Further, the CG-rich motif identified for ncEJCs in first exons is strikingly
4 similar to the CG-rich motif enriched in the first exons of 5IM transcripts (**Fig 5B**). Previous
5 work implicated RanBP2, a protein associated with the cytoplasmic face of the nuclear pore, as
6 a binding factor for some SSCRs (Mahadevan et al. 2013). This finding suggests that nuclear
7 pore proteins may influence EJC occupancy on these transcripts.

8 EJC deposition during the process of pre-mRNA splicing enables the nuclear history of
9 an mRNA to influence post-transcriptional processes including mRNA localization, translation
10 efficiency, and nonsense mediated decay (Chang et al. 2007; Kervestin and Jacobson 2012;
11 Choe et al. 2014). While canonical EJC binding occurs at a fixed distance upstream of exon-exon
12 junctions and involves direct contact between the sugar-phosphate backbone and the EJC core
13 anchoring protein eIF4AIII, ncEJC binding sites likely reflect stable engagement between the
14 EJC core and other mRNP proteins (e.g., SR proteins) recognizing nearby sequence motifs.
15 Although some RBPs were identified for ncEJC motifs found in internal exons (Singh et al.
16 2014, 2012), to date no candidate RBP has been identified for the CG-rich ncEJC motif found in
17 the first exon. If this motif does result from an RBP interaction, it is likely to be one or more of
18 the ~70 proteins that stably and specifically bind to the EJC core (Singh et al. 2012).

19 Finally, we observed a dramatic enrichment for m¹A modifications among 5IM
20 transcripts, with specific enrichment for m¹A modifications in the early coding region. Given
21 this striking enrichment perhaps it is perhaps not surprising that m¹A containing mRNAs were
22 also shown to have more structured 5'UTRs that are GC-rich compared to m¹A lacking mRNAs
23 (Dominissini et al. 2016). Similar to 5IM transcripts, m¹A containing mRNAs were found to

Can Cenik

1 decorate start codons that appear in a highly structured context. While ALKBH3 has been
2 identified as a protein that can demethylate m¹A, it is currently unknown whether there are
3 any proteins that can specifically act as “readers” of m¹A. Recent studies have begun to identify
4 such readers for other mRNA modifications such as YTHDF1, YTHDF2, WTAP and HNRNPA2B1
5 (Ping et al. 2014; Liu et al. 2014; Wang et al. 2014, 2015; Alarcón et al. 2015). Our study
6 highlights a possible link between non-canonical EJC binding and m¹A. Hence, our results yield
7 the intriguing hypothesis that one or more of the ~70 proteins that stably and specifically bind
8 to the EJC core can function as an m¹A reader. Future work involving directed experiments
9 would be needed to test this hypothesis.

10 Given that 5IM transcripts are enriched for ER-targeted and mitochondrial proteins, it
11 is plausible that the observed functional characteristics of 5IM transcripts are driven solely by
12 SSCR and MSCR-containing transcripts. Hence, we repeated all analyses for the subclasses of
13 5IM transcripts (MSCR-containing, SSCR-containing, S⁻/M⁻/SignalP⁺, or S⁻/M⁻/SignalP⁻). We
14 found the observed associations remained statistically significant and had the same direction
15 of effect, even after eliminating SSCR- and MSCR-containing transcripts, despite the fact that all
16 training of the 5IM classifier was performed only using SSCR transcripts. We also found that
17 5IMP score was equally or more strongly associated with each of the functional characteristics
18 compared to the 5UI status. In conclusion, the molecular associations we report apply to 5IM
19 transcripts as a whole, and are not driven solely by the subset of 5IM transcripts encoding ER-
20 or mitochondria-targeting signal peptides, and seem to indicate shared features beyond simple
21 lack of a 5'UTR intron.

22 An intriguing possibility is that 5IM transcript features associated with lower intrinsic
23 translation efficiency may together enable greater ‘tunability’ of 5IM transcripts at the

Can Cenik

1 translation stage. Regulated enhancement or repression of translation, for 5IM transcripts,
2 could allow for rapid changes in protein levels. There are analogies to this scenario in
3 transcriptional regulation, wherein highly regulated genes often have promoters with low
4 baseline levels that can be rapidly modulated through the action of regulatory transcription
5 factors. As more ribosome profiling studies are published examining translational responses
6 transcriptome-wide under multiple perturbations, conditions under which 5IM transcripts are
7 translationally regulated may be revealed. Directed experiments will be needed to test
8 translational features of 5IM transcripts hypothesized via this computational analysis.

9 Taken together, our analyses reveal the existence of a distinct '5IM' class comprising
10 21% of human transcripts. This class is defined by depletion of 5' proximal introns, presence
11 of specific RNA sequence features associated with low translation efficiency, non-canonical
12 binding by the Exon Junction Complex and an enrichment for N¹-methyladenosine
13 modification.

14

Can Cenik

1 **Materials and Methods:**

2 *Datasets and Annotations*

3 Human transcript sequences were downloaded from the NCBI human Reference Gene
4 Collection (RefSeq) via the UCSC table browser (hg19) on Jun 25 2010 (Kent et al. 2002; Pruitt
5 et al. 2005). Transcripts with fewer than three coding exons, and transcripts where the first 99
6 coding nucleotides straddle more than two exons were removed from further consideration.
7 The criteria for exclusion of genes with fewer than three coding exons was to ensure that the
8 analysis of downstream regions was possible for all genes that were used in our analysis of
9 early coding regions. Specifically, the downstream regions were selected randomly from
10 downstream of the 3rd exons. Hence, genes with fewer exons would not be able to contribute a
11 downstream region potentially creating a skew in representation. In total there were ~3000
12 genes that were removed from consideration due to this filter. Therefore, our classifier is
13 limited in its ability to assess transcripts from these genes. However, the performance
14 measures reported in our manuscript are robust to exclusion of these genes, in the sense that
15 the same class of transcripts was used in both training and test datasets.

16 Transcripts were clustered based on sequence similarity in the first 99 coding
17 nucleotides. Specifically, each transcript pair was aligned using BLAST with the DUST filter
18 disabled (Altschul et al. 1990). Transcript pairs with BLAST E-values < 1e-25 were grouped
19 into transcript clusters. In total, there were 15576 transcript clusters that were considered
20 further. These clusters that were subsequently assigned to one of four categories: MSCR-
21 containing, SSCR-containing, S⁻/MSCR⁻ SignalP⁺, or S⁻/MSCR⁻ SignalP⁻ as follows:

22 MSCR-containing transcripts were annotated using MitoCarta and other sources as
23 described in (Cenik et al. 2011). SSCR-containing transcripts were the set of transcripts

Can Cenik

1 annotated to contain signal peptides in the Ensembl Gene v.58 annotations, which were
2 downloaded through Biomart on Jun 25 2010. For transcripts without an annotated MSCR or
3 SSCR, the first 70 amino acids were analyzed using SignalP 3.0 (Bendtsen et al. 2004). Using
4 the eukaryotic prediction mode, transcripts were assigned to the S⁻/MSCR⁻ SignalP⁺ category
5 if either the Hidden Markov Model or the Artificial Neural Network classified the sequence as
6 signal peptide containing. All remaining transcript clusters were assigned to the S⁻/MSCR⁻
7 SignalP⁻ category. The number of transcript clusters in each of the four categories was: 3743
8 SSCR, 737 MSCR, 696 S⁻/MSCR⁻ SignalP⁺, 10400 S⁻/MSCR⁻ SignalP⁻.

9 For each transcript cluster, we also constructed matched control sequences. Control
10 sequences were derived from a single randomly chosen in-frame 'window' downstream of the
11 3rd exon from the evaluated transcripts. If an evaluated transcript had fewer than 99
12 nucleotides downstream of the 3rd exon, no control sequence was extracted. 5UI labels and
13 transcript clustering for the control sequences were inherited from the evaluated transcript.
14 The rationale for this decision is that our analysis depends on the position of the first intron;
15 hence genes with fewer than two exons need to be excluded, as these will not have introns. We
16 further required the matched control sequences to fall downstream of the early coding region.
17 In the vast majority of cases the third exon fell outside the first 99 nucleotides of the coding
18 region, making this a convenient criterion by which to choose control regions.

19 *Sequence Features and Motif Discovery*

20 36 sequence features were extracted from each transcript (**Table S1**). The sequence
21 features included the ratio of arginines to lysines, the ratio of leucines to isoleucines, adenine
22 content, length of the longest stretch without adenines, preference against codons that contain
23 adenines or thymines. These features were previously found to be enriched in SSCR-

Can Cenik

1 containing and certain 5UI⁻ transcripts (Cenik et al. 2011; Palazzo et al. 2007). In addition, we
2 extracted ratios between several other amino acid pairs based on having
3 biochemical/evolutionary similarity, i.e. having positive scores, according to the BLOSUM62
4 matrix (Henikoff and Henikoff 1992). To avoid extreme ratios given the relatively short
5 sequence length, pseudo-counts were added to amino acid ratios using their respective
6 genome-wide prevalence.

7 In addition, we used three published motif finding algorithms (AlignACE, DEME, and
8 MoAN (Roth et al. 1998; Redhead and Bailey 2007; Valen et al. 2009)) to discover RNA
9 sequence motifs enriched among 5UI⁻ transcripts. AlignACE implements a Gibbs sampling
10 approach and is one of the pioneering efforts in motif discovery (Roth et al. 1998). We
11 modified the AlignACE source code to restrict motif searches to only the forward strand of the
12 input sequences to enable RNA motif discovery. DEME and MoAN adopt discriminative
13 approaches to motif finding by searching for motifs that are differentially enriched between
14 two sets of sequences (Redhead and Bailey 2007; Valen et al. 2009). MoAN has the additional
15 advantage of discovering variable length motifs, and can identify co-occurring motifs with the
16 highest discriminative power.

17 In total, six motifs were discovered using the three motif finding algorithms (**Table S1**).
18 Position specific scoring matrices for all motifs were used to score the first $99 - l$ positions in
19 each sequence, where l is the length of the motif. We assessed the significance of each motif
20 instance by calculating the p-value of enrichment (Fisher's Exact Test) among 5UI⁻ transcripts
21 considering all transcripts with a motif instance achieving a PSSM score greater than equal to
22 the instance under consideration. The significance score and position of the two best motif
23 instances were used as features for the classifier (**Table S1**).

Can Cenik

1 *5IM Classifier Training and Performance Evaluation*

2 We modified an implementation of the Random Forest classifier (Breiman 2001) to
3 model the relationship between sequence features in the early coding region and the absence
4 of 5'UTR introns (5UIs). This classifier discriminates transcripts with 5'proximal-intron-
5 minus-like-coding regions and hence is named the '5IM' classifier. The training set for the
6 classifier was composed of SSCR transcripts exclusively. There were two reasons to restrict
7 model construction to SSCR transcripts: 1) we expected the presence of specific RNA elements
8 as a function of 5UI presence based on our previous work (Cenik et al. 2011); and 2) we
9 wanted to restrict model building to sequences that have similar functional constraints at the
10 protein level.

11 We observed that 5UI⁻ transcripts with introns proximal to 5' end of the coding region
12 have sequence characteristics similar to 5UI⁺ transcripts (**Fig 1D**). To systematically
13 characterize this relationship, we built different classifiers using training sets that excluded
14 5UI⁻ transcripts with a coding region intron positioned at increasing distances from the start
15 codon. We evaluated the performance of each classifier using 10-fold cross validation.

16 Given that a large number of motif discovery iterations were needed, we sought to
17 reduce the computational burden. We isolated a subset of the training examples to be used
18 exclusively for motif finding. Motif discovery was performed once using this set of sequences,
19 and the same motifs were used in each fold of the cross validation for all the classifiers.
20 Imbalances between the sizes of positive and negative training examples can lead to
21 detrimental classification performance (Wang and Yao 2012). Hence, we balanced the training
22 set size of 5UI⁻ and 5UI⁺ transcripts by randomly sampling from the larger class. We
23 constructed 10 sub-classifiers to reduce sampling bias, and for each test example, the

Can Cenik

1 prediction score from each subclassifier was summed to produce a combined score between 0
2 and 10. For the rest of the analyses, we used the classifier trained using 5UI- transcripts where
3 first coding intron falls outside the first 90 coding nucleotides (**Fig 1E**).

4 We evaluated classifier performance using a 10-fold cross validation strategy for SSCR-
5 containing transcripts (i.e. the training set). In each fold of the cross-validation, the model was
6 trained without any information from the held-out examples, including motif discovery. For all
7 the other transcripts and the control sets (see above), the 5IMP scores were calculated using
8 the classifier trained using SSCR transcripts as described above. 5IMP score distribution for
9 the control set was used to calculate the empirical cumulative null distribution. Using this
10 function, we determined the p-value corresponding to the 5IMP score for all transcripts. We
11 corrected for multiple hypotheses testing using the qvalue R package (Storey 2003). Based on
12 this analysis, we estimate that a 5IMP score of 7.41 corresponds to a 5% False Discovery Rate
13 and suggest that 21% of all human transcripts can be considered as 5IM transcripts.

14 While the theoretical range of 5IMP scores is 0-10, the highest observed 5IMP is 9.855.
15 We note that for all figures that depict 5IMP score distributions, we displayed the entire
16 theoretical range of 5IMP scores (0-10).

17 *Functional Characterization of 5IM Transcripts:*

18 We collected genome-scale dataset from publically available databases and from
19 supplementary information provided in selected articles. For all analyzed datasets, we first
20 converted all gene/transcript identifiers (IDs) into RefSeq transcript IDs using the Synergizer
21 webserver (Berriz and Roth 2008). If a dataset was generated using a non-human species (ex.
22 Targets identified by TDP-43 RNA immunoprecipitation in rat neuronal cells), we used
23 Homologene release 64 (downloaded on Sep 28 2009) to identify the corresponding ortholog

Can Cenik

1 in humans. If at least one member of a transcript cluster was associated with a functional
2 phenotype, we assigned the cluster to the positive set with respect to the functional phenotype.
3 If more than one member of a cluster had the functional phenotype, we only retained one copy
4 of the cluster unless they differed in a quantitative measurement. For example, consider two
5 hypothetical transcripts: NM_1 and NM_2 that were clustered together and have a 5IMP score
6 of 8.5. If NM_1 had an mRNA half-life of 2hr while NM_2's half-life was 1hr than we split the
7 cluster while preserving the 5IMP score for both NM_1 and NM_2.

8 Once the transcripts were partitioned based on the functional phenotype, we ran two
9 statistical tests: 1) Fisher's Exact Test for enrichment of 5IM transcripts within the functional
10 category; 2) Wilcoxon Rank Sum Test to compare 5IMP scores between transcripts partitioned
11 by the functional phenotype. Additionally, for datasets where a quantitative measurement was
12 available (ex. mRNA half-life), we calculated the Spearman rank correlation between 5IMP
13 scores and the quantitative variable. In these analyses, we assumed that the test space was the
14 entire set of RefSeq transcripts. For all phenotypes where we observed a preliminary
15 statistically significant result, we followed up with more detailed analyses described below.

16 *Analysis of Features Associated with Translation:*

17 For each transcript, we predicted the propensity for secondary structure preceding the
18 translation start site and the 5' cap. Specifically, we extracted 35 nucleotides preceding the
19 translation start site or the first 35 nucleotides of the 5'UTR. If a 5'UTR is shorter than 35
20 nucleotides, the transcript was removed from the analysis. hybrid-ss-min utility (UNAFold
21 package version 3.8) with default parameters was used to calculate the minimum folding
22 energy (Markham and Zuker 2008).

Can Cenik

1 Codon optimality was measured using the tRNA Adaptation Index (tAI), which is based
2 on the genomic copy number of each tRNA (dos Reis et al. 2004). tAI for all human codons
3 were downloaded from Tuller et al. 2010 Table S1. tAI profiles for the first 30 amino acids
4 were calculated for all transcripts. Codon optimality profiles were summarized for the first 30
5 amino acids for each transcript or by averaging tAI at each codon.

6 We carried out two control experiments to test whether the association between 5IMP
7 score and tAI could be explained by confounding variables. First, we permuted the nucleotides
8 in the first 90 nts and observed no relationship between 5IMP score and mean tAI when these
9 permuted sequences were used (**Fig S1**). Second, we selected random in-frame 99 nucleotides
10 from 3rd exon to the end of the coding region and found no significant differences in tAI (**Fig**
11 **S1**). These results suggest that the relationship between tAI and 5IMP score is confined to the
12 first 30 amino acids and is not explained by simple differences in nucleotide composition.

13 Ribosome profiling and RNA expression data for human lymphoblastoid cells (LCLs)
14 were downloaded from NCBI GEO database accession number: GSE65912. Translation
15 efficiency was calculated as previously described (Cenik et al. 2015). Median translation
16 efficiency across the different cell-types was used for each transcript.

17 *Analysis of Proximity Specific Ribosome Profiling Data:*

18 We downloaded proximity specific ribosome profiling data for HEK 293 cells from Jan
19 et al. 2014; Table S6. We converted UCSC gene identifiers to HGNC symbols using g:Profiler
20 (Reimand et al. 2011). We retained all genes with an RPKM >5 in either input or pulldown and
21 required that at least 30 reads were mapped in either of the two libraries. We used ER-
22 targeting evidence categories “secretome”, “phobius”, “TMHMM”, “SignalP”, “signalSequence”,
23 “signalAnchor” from (Jan et al. 2014) to annotate genes as having ER-targeting evidence. The

Can Cenik

1 genes that did not have any ER-targeting evidence or “mitoCarta” / “mito.GO” annotations were
2 deemed as the set of genes with no ER-targeting or mitochondrial evidence. We calculated the
3 \log_2 of the ratio between ER-proximal ribosome pulldown RPKM and control RPKM as the
4 measure of enrichment for ER-proximal ribosome occupancy (as in Jan et al. 2014). A moving
5 average of this ratio was calculated for genes grouped by their 5IMP score. For this calculation,
6 we used bins of 30 mitochondrial genes or 100 genes with no evidence of ER- or mitochondrial
7 targeting.

8 *Analysis of Genome-wide Binding Sites of Exon Junction Complex:*

9 Dr. Gene Yeo and Gabriel Pratt generously shared uniformly processed peak calls for
10 experiments identifying human RNA binding protein targets. These datasets include various
11 CLIP-Seq datasets and its variants such as iCLIP. A total 49 datasets from 22 factors were
12 analyzed. These factors were: hnRNPA1, hnRNPF, hnRNPM, hnRNPU, Ago2, hnRNPU, HuR,
13 IGF2BP1, IGF2BP2, IGF2BP3, FMR1, eIF4AIII, PTB, IGF2BP1, Ago3, Ago4, MOV10, Fip1, CF
14 Im68, CF Im59, CF Im25, and hnRNPA2B1. We extracted the 5IMP scores for all targets of each
15 RBP. We calculated the Wilcoxon Rank Sum test statistic comparing the 5IMP score
16 distribution of the targets of each RBP to all other transcripts with 5IMP scores. None of the
17 tested RBP target sets had an adjusted p-value < 0.05 and a median difference in 5IMP score $>$
18 1 when compared to non-target transcripts.

19 In addition, we used RNA:protein immunoprecipitation in tandem (RIPiT) data to
20 determine Exon Junction Complex (EJC) binding sites (Singh et al. 2014, 2012). We analyzed
21 the common peaks from the Y14-Magoh immunoprecipitation configuration (Singh et al. 2012;
22 Kucukural et al. 2013). Canonical EJC binding sites were defined as peaks whose weighted
23 center were 15 to 32 nucleotides upstream of an exon-intron boundary. All remaining peaks

Can Cenik

1 were deemed as “non-canonical” EJC binding sites. We extracted all non-canonical peaks that
2 overlapped the first 99 nucleotides of the coding region and restricted our analysis to
3 transcripts that had an RPKM greater than one in the matched RNA-Seq data.

4 *Analysis of m¹A modified transcripts*

5 We downloaded the list of RNAs observed to contain m¹A from Li et al. (2016) and
6 Dominissini et al. (2016). RefSeq transcript identifiers were converted to HGNC symbols using
7 g:Profiler (Reimand et al. 2011). The overlap between the two datasets was determined using
8 HGNC symbols. m¹A modifications that overlap the first 99 nucleotides of the coding region
9 were determined using bedtools (Quinlan and Hall 2010).

10 Li et al. (2016) identified 600 transcripts with m¹A modification in normal HEK293
11 cells. Of these, 368 transcripts were not found to contain the m¹A modification in HEK293 cells
12 by Dominissini et al. (2016). Yet, 81% of these were found to be m¹A modified in other cell
13 types. Li et al. (2016) also analyzed m¹A upon H₂O₂ treatment and serum starvation in HEK293
14 cells and identified many m¹A modifications that are only found these stress conditions.
15 However, 20% of 371 transcripts harboring stress-induced m¹A modifications were found in
16 normal HEK293 cells by Dominissini et al. (2016). Taken together, these analyses suggest that
17 transcriptome-wide m¹A maps remain incomplete. Hence, we analyzed the 5IMP scores of all
18 mRNAs with m¹A across cell types and conditions. We reported results using the Dominissini
19 et al. dataset but the same conclusions were supported by m¹A modifications from Li et al.

20

Can Cenik

1 **Acknowledgements:**

2 We would like to thank Gene Yeo, and Gabriel Pratt for sharing peak calls for RNA-
3 binding protein target identification experiments (CLIP, PAR-CLIP, iCLIP). We thank Elif
4 Sarinay Cenik, Başar Cenik, and Alper Küçükural for helpful discussions. F.P.R. and H.N.C. were
5 supported by the Canada Excellence Research Chairs Program. This work was supported in
6 part by NIH grants GM50037 (M.J.M.), HG001715 and HG004233 (F.P.R.), the Krembil
7 Foundation, an Ontario Research Fund award, and the Avon Foundation (F.P.R.). A.F.P. was
8 supported by a grant from the Canadian Institutes of Health Research to A.F.P. (FRN 102725).

9 **Author contributions:** CC, MJM, and FPR designed the study. CC and HNC implemented the
10 random forest classifier and carried out all analyses. GS and AA performed initial experiments.
11 MPS and AFP provided critical feedback. MJM and FPR jointly supervised the project. CC, FPR
12 and MJM wrote the manuscript with input from all authors.

13

14

Can Cenik

1 **References:**

- 2 Alarcón CR, Goodarzi H, Lee H, Liu X, Tavazoie S, Tavazoie SF. 2015. HNRNPA2B1 Is a Mediator of
3 m(6)A-Dependent Nuclear RNA Processing Events. *Cell* **162**: 1299–1308.
- 4 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*
5 **215**: 403–410.
- 6 Ameres SL, Zamore PD. 2013. Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol* **14**:
7 475–488.
- 8 Babendure JR, Babendure JL, Ding J-H, Tsien RY. 2006. Control of mammalian translation by mRNA
9 structure near caps. *RNA* **12**: 851–861.
- 10 Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP
11 3.0. *J Mol Biol* **340**: 783–795.
- 12 Berriz GF, Roth FP. 2008. The Synergizer service for translating gene, protein and other biological
13 identifiers. *Bioinformatics* **24**: 2272–2273.
- 14 Breiman L. 2001. Random Forests. *Mach Learn* **45**: 5–32.
- 15 Cenik C, Chua HN, Zhang H, Tarnawsky SP, Akef A, Derti A, Tasan M, Moore MJ, Palazzo AF, Roth FP.
16 2011. Genome analysis reveals interplay between 5'UTR introns and nuclear mRNA export for
17 secretory and mitochondrial genes. *PLoS Genet* **7**: e1001366.
- 18 Cenik C, Derti A, Mellor JC, Berriz GF, Roth FP. 2010. Genome-wide functional analysis of human 5'
19 untranslated region introns. *Genome Biol* **11**: R29.
- 20 Cenik C, Sarinay Cenik E, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL,
21 Tang H, et al. 2015. Integrative analysis of RNA, translation and protein levels reveals distinct
22 regulatory variation across humans. *Genome Res*. <http://dx.doi.org/10.1101/gr.193342.115>.
- 23 Chang Y-F, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway.
24 *Annu Rev Biochem* **76**: 51–74.
- 25 Charneski CA, Hurst LD. 2013. Positively charged residues are the major determinants of ribosomal
26 velocity. *PLoS Biol* **11**: e1001508.
- 27 Choe J, Ryu I, Park OH, Park J, Cho H, Yoo JS, Chi SW, Kim MK, Song HK, Kim YK. 2014. eIF4AIII enhances
28 translation of nuclear cap-binding complex-bound mRNAs by promoting disruption of secondary
29 structures in 5'UTR. *Proc Natl Acad Sci U S A* **111**: E4577–86.
- 30 Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim MS, Dai Q, Di Segni A,
31 Salmon-Divon M, Clark WC, et al. 2016. The dynamic N(1)-methyladenosine methylome in
32 eukaryotic messenger RNA. *Nature* **530**: 441–446.
- 33 dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for
34 translational selection. *Nucleic Acids Res* **32**: 5036–5044.
- 35 Dunn DB. 1961. The occurrence of 1-methyladenine in ribonucleic acid. *Biochim Biophys Acta* **46**: 198–
36 200.

Can Cenik

- 1 Frye M, Jaffrey SR, Pan T, Rechavi G, Suzuki T. 2016. RNA modifications: what have we learned and
2 where are we headed? *Nat Rev Genet* **17**: 365–372.
- 3 Gerashchenko MV, Gladyshev VN. 2015. Translation inhibitors cause abnormalities in ribosome
4 profiling experiments. *Nucleic Acids Res* **42**: e134.
- 5 Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**: 481.
- 6 Hall RH. 1963. Method for isolation of 2'-O-methylribonucleosides and N1-methyladenosine from
7 ribonucleic acid. *Biochimica et Biophysica Acta (BBA) - Specialized Section on Nucleic Acids and*
8 *Related Subjects* **68**: 278–283.
- 9 Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci*
10 *U S A* **89**: 10915–10919.
- 11 Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet* **42**: 287–299.
- 12 Hinnebusch AG, Lorsch JR. 2012. The mechanism of eukaryotic translation initiation: new insights and
13 challenges. *Cold Spring Harb Perspect Biol* **4**. <http://dx.doi.org/10.1101/cshperspect.a011544>.
- 14 Hong X, Scofield DG, Lynch M. 2006. Intron size, abundance, and distribution within untranslated
15 regions of genes. *Mol Biol Evol* **23**: 2392–2404.
- 16 Jan CH, Williams CC, Weissman JS. 2015. LOCAL TRANSLATION. Response to Comment on “Principles of
17 ER cotranslational translocation revealed by proximity-specific ribosome profiling.” *Science* **348**:
18 1217.
- 19 Jan CH, Williams CC, Weissman JS. 2014. Principles of ER cotranslational translocation revealed by
20 proximity-specific ribosome profiling. *Science* **346**: 1257521.
- 21 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human
22 genome browser at UCSC. *Genome Res* **12**: 996–1006.
- 23 Kervestin S, Jacobson A. 2012. NMD: a multifaceted response to premature translational termination.
24 *Nat Rev Mol Cell Biol* **13**: 700–712.
- 25 Klootwijk J, Planta RJ. 1973. Analysis of the methylation sites in yeast ribosomal RNA. *Eur J Biochem* **39**:
26 325–333.
- 27 Kucukural A, Özadam H, Singh G, Moore MJ, Cenik C. 2013. ASPeak: an abundance sensitive peak
28 detection algorithm for RIP-Seq. *Bioinformatics* **29**: 2485–2486.
- 29 Larsson O, Li S, Issaenko OA, Avdulov S, Peterson M, Smith K, Bitterman PB, Polunovsky VA. 2007.
30 Eukaryotic Translation Initiation Factor 4E-Induced Progression of Primary Human Mammary
31 Epithelial Cells along the Cancer Pathway Is Associated with Targeted Translational Deregulation
32 of Oncogenic Drivers and Inhibitors. *Cancer Res* **67**: 6814–6824.
- 33 Lee ES, Akef A, Mahadevan K, Palazzo AF. 2015. The consensus 5' splice site motif inhibits mRNA
34 nuclear export. *PLoS One* **10**: e0122743.
- 35 Le Hir H, Izaurralde E, Maquat LE, Moore MJ. 2000. The spliceosome deposits multiple proteins 20-24
36 nucleotides upstream of mRNA exon-exon junctions. *EMBO J* **19**: 6860–6869.
- 37 Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, et al. 2014. A METTL3-METTL14
38 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat Chem Biol* **10**: 93–95.

Can Cenik

- 1 Li X, Xiong X, Wang K, Wang L, Shu X, Ma S, Yi C. 2016. Transcriptome-wide mapping reveals reversible
2 and dynamic N1-methyladenosine methylome. *Nat Chem Biol*.
3 <http://dx.doi.org/10.1038/nchembio.2040> (Accessed March 29, 2016).
- 4 Mahadevan K, Zhang H, Akef A, Cui XA, Gueroussov S, Cenik C, Roth FP, Palazzo AF. 2013.
5 RanBP2/Nup358 potentiates the translation of a subset of mRNAs encoding secretory proteins.
6 *PLoS Biol* **11**: e1001545.
- 7 Marintchev A, Edmonds KA, Marintcheva B, Hendrickson E, Oberer M, Suzuki C, Herdy B, Sonenberg N,
8 Wagner G. 2009. Topology and regulation of the human eIF4A/4G/4H helicase complex in
9 translation initiation. *Cell* **136**: 447–460.
- 10 Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. *Methods*
11 *Mol Biol* **453**: 3–31.
- 12 Meyuhas O. 2000. Synthesis of the translational apparatus is regulated at the translational level. *Eur J*
13 *Biochem* **267**: 6321–6330.
- 14 Moore MJ, Proudfoot NJ. 2009. Pre-mRNA processing reaches back to transcription and ahead to
15 translation. *Cell* **136**: 688–700.
- 16 Palazzo AF, Mahadevan K, Tarnawsky SP. 2013. ALREX-elements and introns: two identity elements
17 that promote mRNA nuclear export. *Wiley Interdiscip Rev RNA* **4**: 523–533.
- 18 Palazzo AF, Springer M, Shibata Y, Lee C-S, Dias AP, Rapoport TA. 2007. The signal sequence coding
19 region promotes nuclear export of mRNA. *PLoS Biol* **5**: e322.
- 20 Parsyan A, Svitkin Y, Shahbazian D, Gkogkas C, Lasko P, Merrick WC, Sonenberg N. 2011. mRNA
21 helicases: the tacticians of translational control. *Nat Rev Mol Cell Biol* **12**: 235–245.
- 22 Ping X-L, Sun B-F, Wang L, Xiao W, Yang X, Wang W-J, Adhikari S, Shi Y, Lv Y, Chen Y-S, et al. 2014.
23 Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell*
24 *Res* **24**: 177–189.
- 25 Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant
26 sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**: D501–4.
- 27 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
28 *Bioinformatics* **26**: 841–842.
- 29 Redhead E, Bailey TL. 2007. Discriminative motif discovery in DNA and protein sequences using the
30 DEME algorithm. *BMC Bioinformatics* **8**: 385.
- 31 Reid DW, Nicchitta CV. 2015a. Diversity and selectivity in mRNA translation on the endoplasmic
32 reticulum. *Nat Rev Mol Cell Biol* **16**: 221–231.
- 33 Reid DW, Nicchitta CV. 2015b. LOCAL TRANSLATION. Comment on “Principles of ER cotranslational
34 translocation revealed by proximity-specific ribosome profiling.” *Science* **348**: 1217.
- 35 Reimand J, Arak T, Vilo J. 2011. g: Profiler—a web server for functional interpretation of gene lists
36 (2011 update). *Nucleic Acids Res* gkr378.
- 37 Roth FP, Hughes JD, Estep PW, Church GM. 1998. Finding DNA regulatory motifs within unaligned
38 noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**: 939–

Can Cenik

- 1 945.
- 2 Rowland AA, Voeltz GK. 2012. Endoplasmic reticulum–mitochondria contacts: function of the junction.
3 *Nat Rev Mol Cell Biol* **13**: 607–625.
- 4 Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation.
5 *Cell* **153**: 1589–1601.
- 6 Singh G, Kucukural A, Cenik C, Leszyk JD, Shaffer SA, Weng Z, Moore MJ. 2012. The cellular EJC
7 interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell* **151**: 750–764.
- 8 Singh G, Ricci EP, Moore MJ. 2014. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein
9 complexes. *Methods* **65**: 320–332.
- 10 Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat*
11 **31**: 2013–2035.
- 12 Sylvestre J, Vialette S, Corral Debrinski M, Jacq C. 2003. Long mRNAs coding for yeast mitochondrial
13 proteins of prokaryotic origin preferentially localize to the vicinity of mitochondria. *Genome Biol* **4**:
14 R44.
- 15 Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010.
16 An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*
17 **141**: 344–354.
- 18 Valen E, Sandelin A, Winther O, Krogh A. 2009. Discovery of regulatory elements is improved by a
19 discriminatory approach. *PLoS Comput Biol* **5**: e1000562.
- 20 Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and
21 transcriptomic analyses. *Nat Rev Genet* **13**: 227–232.
- 22 Wang S, Yao X. 2012. Multiclass Imbalance Problems: Analysis and Potential Solutions. *IEEE Trans Syst*
23 *Man Cybern B Cybern*. <http://dx.doi.org/10.1109/TSMCB.2012.2187280>.
- 24 Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. 2014. N6-
25 methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**: 117–120.
- 26 Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. 2015. N(6)-
27 methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* **161**: 1388–1399.
- 28 Zinshteyn B, Gilbert WV. 2013. Loss of a conserved tRNA anticodon modification perturbs cellular
29 signaling. *PLoS Genet* **9**: e1003675.
- 30
- 31

Can Cenik

1 **Figure Legends:**

2 **Fig 1 | Modeling the relationship between sequence features in the early coding**

3 **region and the absence of 5'UTR introns (5UIs).** (a) For all human transcripts,

4 information about 36 nucleotide-level features of the early coding region (first 99

5 nucleotides) and 5UI presence was extracted. (b) Transcripts containing a signal sequence

6 coding region (SSCR) were used to train a Random Forest classifier that modeled the

7 relationship between 5UI absence and 36 sequence features. (c) With this classifier, all

8 human transcripts were assigned a score that quantifies the likelihood of 5UI absence

9 based on specific RNA sequence features in the early coding region. Transcripts with high

10 scores are thus considered to have 5'-proximal intron minus-like coding regions (5IMs).

11 (d) "5'UTR-intron-minus-predictor" (5IMP) score distributions for SSCR-containing

12 transcripts shift to higher scores with later-appearing first introns, suggesting that 5IM

13 coding region features not only predict lack of a 5UI, but also lack of early coding region

14 introns. (e) Classifier performance was optimized by excluding 5UI- transcripts with

15 introns appearing early in the coding region. Cross-validation performance (area under the

16 precision recall curve, AUPRC) was examined for a series of alternative 5IM classifiers

17 using different first-intron-position criterion for excluding 5UI- transcripts from the

18 training set (Methods).

19

20 **Fig 2 | Predicting 5UI status accurately using only early coding sequences.** (a) As

21 judged by area under the receiver operating characteristic curve (AUROC) and AUPRC, The

22 5IM classifier performed well for several different transcript classes. (b) The distribution

23 of 5IMP scores reveals clear separation of 5UI⁺ and 5UI⁻ transcripts for SSCR-containing

Can Cenik

1 transcripts, where each SSCR-containing transcript was scored using a classifier that did
2 not use that transcript in training (Methods). **(c)** Coding sequence features that are
3 predictive of 5' proximal intron presence are restricted to the early coding region. This was
4 supported by identical 5IM classifier score distributions with respect to 5UI presence for
5 negative control sequences, each derived from a single randomly chosen 'window'
6 downstream of the 3rd exon from one of the evaluated transcripts. **(d)** MSCR transcripts
7 exhibited a major difference in 5IMP scores based on their 5UI status even though no MSCR
8 transcripts were used in training the classifier. **(e)** Transcripts predicted to contain signal
9 peptides (SignalP+) had a 5IMP score distribution similar to that of SSCR-containing
10 transcripts. **(f)** After eliminating SSCR, MSCR, and SignalP+ transcripts, the remaining S-
11 /MSCR- SignalP- transcripts were still significantly enriched for high 5IM classifier scores
12 among 5UI- transcripts. **(g)** The control set of randomly chosen sequences downstream of
13 the 3rd exon from each transcript was used to calculate an empirical cumulative null
14 distribution of 5IMP scores. Using this function, we determined the p-value corresponding
15 to the 5IMP score for all transcripts. The red dashed line indicates the p-value
16 corresponding to 5% False Discovery Rate. The inset depicts the distribution of various
17 classes of mRNAs among the input set and 5IM transcripts.

18

19 **Fig 3 | 5IM transcripts are more likely to be differentially expressed and have**
20 **sequence features associated with lower translation efficiency (a)** The 5IM classifier
21 score was positively correlated with the propensity for mRNA structure preceding the start
22 codon (-ΔG) (Spearman rho=0.39; p < 2.2e-16). For each transcript, 35 nucleotides
23 immediately upstream of the AUG were used to calculate -ΔG (Methods). **(b)** The 5IM

Can Cenik

1 classifier score was positively correlated with the propensity for mRNA structure near the
2 5'cap (-ΔG) (Spearman rho=0.18; p = 7.9e-130; Methods). **(c)** Transcripts that are
3 translationally upregulated in response to eIF4E overexpression (Larsson et al. 2007)
4 (blue) were enriched for higher 5IMP scores. Light green shading indicates 5IMP scores
5 corresponding to 5% FDR. **(d)** Transcripts with non-AUG start codons (blue) exhibited
6 significantly higher 5IMP scores than transcripts with a canonical ATG start codon (yellow).
7 **(e)** Higher 5IMP scores were associated with less optimal codons (as measured by the
8 tRNA adaptation index, tAI) for the first 33 codons. For all transcripts within each 5IMP
9 score category (blue-high, orange-low), the mean tAI was calculated at each codon position.
10 Start codon was not shown. **(f)** Transcripts with lower translation efficiency were enriched
11 for higher 5IMP scores. Transcripts with translation efficiency one standard deviation
12 below the mean ("LOW" translation, yellow) and one standard deviation higher than the
13 mean ("HIGH" translation, blue) were identified using ribosome profiling and RNA-Seq data
14 from human lymphoblastoid cell lines (Methods).

15

16 **Fig 4 | 5IM transcripts with no evidence of ER-targeting are more likely to exhibit ER-**
17 **proximal ribosome occupancy.** A moving average of ER-proximal ribosome occupancy
18 was calculated by grouping genes by 5IMP score (see Methods). We plotted the moving
19 average of 5IMP scores for transcripts with no evidence of ER- or mitochondrial targeting
20 (green) or for transcripts predicted to be mitochondrial (purple). We plotted a random
21 subsample of transcripts on top of the moving average (circles).

22

Can Cenik

1 **Fig 5 | 5IM transcripts harbor non-canonical Exon Junction Complex (EJC) binding**
2 **sites (a)** Observed EJC binding sites (Singh et al. 2012) are shown for an example 5IM
3 transcript (*LAMC1*). Canonical EJC binding sites (purple) are ~24nt upstream of an exon-
4 intron boundary. The remaining binding sites are considered to be non-canonical (green).
5 **(b)** A CG-rich sequence motif previously identified to be enriched among ncEJC binding
6 sites in first exons (Singh et al. 2012) is shown **(c)** 5IMP score for transcripts with zero,
7 one, two or more non-canonical EJC binding sites in the first 99 coding nucleotides reveals
8 that transcripts with high 5IMP scores frequently harbor non-canonical EJC binding sites.
9 **(d)** Transcripts with high 5IMP scores are enriched for non-canonical EJCs regardless of
10 5UI presence or absence.

11

12 **Fig 6 | 5IM transcripts are enriched for mRNAs with early coding region m¹A**
13 **modifications (a)** Transcripts with m¹A modifications (blue) in the first 99 coding nucleotides
14 exhibit significant enrichment for 5IM transcripts and have higher 5IMPS scores than
15 transcripts without m¹A modifications in the first 99 coding nucleotides (yellow). **(b)**
16 Transcripts with m¹A modifications (blue) in the 5'UTR do not display a similar enrichment.

17

18 **Supporting Information**

19 **Figure S1 | Association between 5IMP scores and codon optimality is restricted to the**
20 **first 30 amino acids and is not explained by nucleotide content. (a)** 5IM transcripts tend
21 to have less optimal codons in their first 30 amino acids as measured by tRNA adaptation index
22 (tAI). The median tAI for each transcript was calculated and transcripts were grouped by their

Can Cenik

- 1 5IMP scores. The distribution of median tAI was plotted as a boxplot. **(b)** We permuted the
- 2 nucleotides of the first 99 nucleotides and found that the relationship between 5IMP and tAI
- 3 was lost. This result suggested that nucleotide composition alone doesn't explain the
- 4 relationship between 5IMP and tAI. **(c)** We used the previously described negative control
- 5 sequences, each derived from a single randomly chosen in-frame 'window' downstream of the
- 6 3rd exon from one of the evaluated transcripts. We found no relationship between 5IMP score
- 7 and tAI in these regions suggesting that the observed association is restricted to the early
- 8 coding region.

- 9 **Table S1-** | List of features used by the 5IM classifier.

- 10 **Table S2-** | 5IMP scores of all human transcripts.

- 11 **Table S3-** | List of functional features tested for association with 5IMP scores.













