Codon Optimality Confers GC-Rich-Induced Stability to mRNAs in Humans

Fabian Hia[1], Sheng Fan Yang[1], Yuichi Shichino[2], Masanori Yoshinaga[1], Yasuhiro Murakawa[3], Alexis Vandenbon[4], Akira Fukao[5], Toshinobu Fujiwara[5], Markus Landthaler[6], Tohru Natsume[7], Shungo Adachi[7], Shintaro Iwasaki[2,8], and Osamu Takeuchi[1,*]

[1] Department of Medical Chemistry, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan
[2] RNA Systems Biochemistry Laboratory, RIKEN Cluster for Pioneering Research, Wako, Saitama 351-0198, Japan
[3] Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa 230-0045, Japan; RIKEN Preventive Medicine and Diagnosis Innovation Program, Yokohama, Kanagawa 230-0045, Japan.
[4] Laboratory of Infection and Prevention, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, 606-8507, Japan.
[5] Laboratory of Biochemistry, Department of Pharmacy, Kindai University, Higashiosaka City, Osaka, 577-8502 Japan
[6] RNA Biology and Posttranscriptional Regulation, Max Delbrück Center for Molecular Medicine Berlin, Berlin Institute for Molecular Systems Biology, 13125 Berlin, Germany; IRI Life Sciences, Institut für Biologie, Humboldt-Universität zu Berlin, 10115 Berlin, Germany.
[7] Molecular Profiling Research Center for Drug Discovery (molprof), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan
[8]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan
* To whom correspondence should be addressed. Tel: +81-75-753-9500; Fax: +81-75-753-9502; Email: otake@mfour.med.kyoto-u.ac.jp

ABSTRACT

Codon optimality has been implicated as one of the major factors contributing to mRNA stability in yeast. However, the presence of codon-optimality-mediated decay has been unclear in humans. Here we show that human cells possess a mechanism to modulate RNA stability via codon optimality with a unique codon bias different from that of yeast. We performed dimensionality reduction analysis of genome-wide codon frequencies and found that codons could be clustered into two distinct groups – codons with A or T at the third base position (AT3) and codons with either G or C at the third base position (GC3). Quantifying codon bias and subsequently gene optimality showed that increased GC3-content entails proportionately higher GC-content which in turn confers stability to mRNA transcripts. Agreement of our codon optimality-derived metric and ribosome occupancies across mRNAs determined from ribosome profiling suggests that codon optimality affects ribosome occupancy. This system was verified by measuring the stabilities of codon optimized and deoptimized reporter transcripts. Employing an immunoprecipitation-based strategy, we identified ILF2 as an RNA binding protein that regulates global mRNA abundances via AU-content. Our results demonstrate that codon-optimality-mediated decay is a highly conserved system which in the course of evolution has seen changes in codon usage.

INTRODUCTION

Messenger RNA (mRNA) regulation represents an essential part of regulating a myriad of physiological processes in cells, being indicated in the maintenance of cellular homeostasis to immune responses (1–3). In addition to transcription regulation, post-transcriptional regulation of mRNA stability is vital to the fine-tuning of mRNA abundance. To date, several mRNA-intrinsic properties, often in 5′ or 3′ untranslated regions (UTR), have been shown to affect mRNA stability (4, 5). Due to the recent advances in technology, the contribution of mRNA stability to gene expression has been suggested (6). However, the regulation of mRNA stability, which is possibly governed by mRNA intrinsic features, has not been fully elucidated.

One of the most crucial mRNA-intrinsic features is codon usage. To scrutinize the bias in usage of redundant codons, several metrics of codon optimality have been proposed. In a classical metric called the codon Adaptation Index (cAI), gene optimality is calculated by comparison between codon usage bias of a target gene and reference genes which are highly expressed (7, 8). Another index termed the tRNA Adaption Index (tAI) gauges how efficiently tRNA is utilized by the translating ribosome (9, 10). More recently, the normalized translation efficiency (nTE), which takes into consideration not only the availability of tRNA but also demand, was also proposed (11).

Recently, Presnyak and colleagues showed that mRNA half-lives are correlated with optimal codon content based on a metric, the Codon Stabilization Coefficient (CSC) which was calculated from the correlations between the codon frequencies in mRNAs and stabilities of mRNAs. Additionally, they showed that the substitutions of codons with their synonymous optimal and non-optimal counterparts resulted in significant increases and decreases in mRNA stability in yeast (12). This effect was brought by an RNA binding protein (RBP) Dhh1p (mammalian ortholog DDX6), which senses ribosome elongation speed (12–14). In yeast, these differences in ribosome elongation speed in turn are influenced by tRNA availability and demand (11, 15, 16). Taken together, codons can be designated into optimal and non-optimal categories; the former hypothesized to be decoded efficiently and accurately (17, 18) while the latter slow ribosome elongation resulting in decreased mRNA stability (12–14). It is also important to make the distinction that common and rare codons do not necessarily imply optimal and non-optimal codons.

At present, codon optimality-mediated decay has been extensively studied and established particularly in *Saccharomyces cerevisiae* as well as other model organisms such as *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, and *Neurospora crassa* (19–23). Nevertheless to date, this system of codon optimality has been inadequately scrutinized in humans.

In this study, we show that a system of codon optimality-mediated decay exists in humans. Principal component analysis (PCA) showed that codons could be clustered into two distinct groups; codons with A or T at the third base position (AT3) and codons with either G or C at the third base position (GC3). This clustering was associated with mRNA half-lives enabling us to determine GC3 and AT3 codons as optimal and non-optimal codons respectively. We then developed an algorithm to quantify the optimality of genes based on gene codon bias. We show that the use of GC3 codons inevitably

increases GC-content, which in turn confers stability to mRNA transcripts. With ribosome profiling, we show that codon optimality-derived occupancy scores agreed with ribosome occupancy. Finally, employing a ribonucleoprotein immunoprecipitation strategy, we identified RNA binding proteins which were bound to transcripts with low or high optimality scores. We propose that interleukin enhancer-binding factor 2 (ILF2) mediates mRNA stability of low optimality transcripts via recognition of AU-rich sequences.

MATERIAL AND METHODS

Cell Cultures, Growth, and Transfection Conditions

HEK293T cells were maintained in Dulbecco's modified eagle medium (DMEM) (Nacalai Tesque), supplemented with 10% (v/v) fetal bovine serum. HEK293 Tet-off cells were maintained in Minimum Essential Medium Eagle - Alpha Modification (α-MEM) (Nacalai Tesque), supplemented with 10% (v/v) Tet-system approved fetal bovine serum (Takara Bio) and 100 µg/ml of G418 (Nacalai Tesque). For REL and IL6 overexpression experiments, plasmids were transfected using PEI MAX (Polysciences Inc). For co-transfection of ILF2 siRNA with REL plasmids, Lipofectamine 2000 was used as per manufacturer's protocol. ILF2 siRNA which targeted ILF2 at exons 8 and 9 were Silencer Select siRNA, S7399 (Ambion, Life Technologies).

Plasmid Construction

Codon optimized-*REL* (*REL-OPT*), *IL6* (*IL6-OPT*) and codon deoptimized *IL6* (*IL6-DE*) sequences were synthesized as gBlocks Gene Fragments (Integrated DNA Technologies) (**Supplementary Table 1**). These sequences and corresponding WT sequences were polymerase chain reaction (PCR) amplified (with the inclusion of a FLAG tag for *REL* sequences) and inserted into the pcDNA3.1(+) vector (Invitrogen) and pTRE-TIGHT vector (Takara Bio). The sequences were confirmed via restriction enzyme digest and sequencing.

Tet-Off Assay

HEK293 Tet-off cells (Clontech) were transfected with pTRE-TIGHT plasmids bearing the (de)optimized and WT sequences and incubated overnight at 37ºC. Transcriptional shut-off for the indicated plasmids was achieved by the addition of doxycycline (LKT Laboratories Inc.) to a final concentration of 1 µg/ml. Samples were harvested at the indicated timepoints after the addition of doxycycline.

RNA Extraction, Reverse Transcription PCR, and Quantitative Real-time PCR

Total RNA was isolated from cells using TRIzol reagent (Invitrogen) as per manufacturer's instructions. Reverse transcription was performed using the ReverTra Ace qPCR RT Master Mix with gDNA remover kit (Toyobo) as per manufacturer's instructions. cDNA was amplified with PowerUp SYBR Green Master Mix (Applied Biosystems) and quantitative real-time PCR (qPCR) was performed on the StepOne Real-Time PCR System (Applied Biosystems). Human GAPDH abundance was used for normalization. The list of qPCR primers can be found in **Supplementary Table 1**.

Sucrose Gradient Centrifugation (Polysome Profiling)

HEK293T were transfected with equal concentrations of *REL-OPT* and *REL-WT* plasmids. Cells were lysed the next day in polysome buffer [20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES-KOH) (pH 7.5], 100 mM KCl, 5 mM $MgCl_2$, 0.25% (v/v) Nonidet P-40, 10 µg/ml cycloheximide, 100 units/ml RNase inhibitor, and protease inhibitor cocktail (Roche)]. Lysates were loaded on top of a linear 15%–60% sucrose gradient [15%–60% sucrose, 20 mM HEPES-KOH [pH 7.5], 100 mM KCl, 5 mM $MgCl_2$, 10 µg/ml cycloheximide, 100 units/ml RNase inhibitor, and protease inhibitor cocktail (Roche)]. After ultracentrifugation at 38,000 rpm for 2.5 h at 4ºC in a HITACHI P40ST rotor, fractions were collected from the top of the gradient and subjected to UV-densitometric analysis. The absorbance profiles of the gradients were determined at 254 nm. For disassociation of ribosome and polysome, EDTA was added to Mg^{2+}-free polysome buffer and 15%–60% sucrose gradient at concentrations of 50 mM and 20 mM, respectively. For RNA analysis, RNA from each fraction was extracted via the High Pure RNA Isolation Kit (Roche) and subject to reverse transcription and qPCR.

Immunoblot Analysis

Samples were lysed in RIPA buffer (20 mM Tris-HCl [pH 8], 150 mM NaCl, 10 mM EDTA, 1% Nonidet-P40, 0.1% SDS, 1% sodium deoxycholate, and cOmplete Mini EDTA-free Protease Inhibitor Cocktail [Roche]). Protein concentration was determined by the BCA Protein Assay (Thermo Fisher). Whole cell lysates were resolved by SDS-PAGE and transferred onto PVDF membranes (Bio-Rad). The following antibodies were used for immunoblot analysis: mouse monoclonal anti-FLAG (F3165, Sigma), mouse monoclonal anti-ILF2 (sc-365283, Santa Cruz Biotechnology), mouse anti-β-actin (sc-47778, Santa Cruz), and mouse IgG HRP linked $F(ab')_2$ fragment (NA9310, GE Healthcare). Luminescence was detected with a luminescent image analyser (Amersham Imager 600; GE Healthcare).

ELISA

HEK293T cells were transfected with pcDNA3.1(+) plasmids bearing the (de)optimized and WT sequences and incubated overnight at 37ºC. Cell supernatant was aspirated and the cell monolayer washed with 1x PBS (pre-warmed at 37ºC). Pre-warmed DMEM was added to the monolayer and the cells incubated for 2 hr at 37ºC. Thereafter, the cell supernatant was harvested and centrifuged at 300 x g to pellet residual cells. The resulting supernatant was decanted and the concentration of secreted IL6 was measured by the human IL6 ELISA kit (Invitrogen) according to the manufacturer's instructions.

ISRIM (*In vitro* Specificity based RNA Regulatory protein Identification Method)

Preparation of bait RNAs. T7-tagged cDNA template was PCR amplified and subjected to *in vitro* transcription using a MEGAscript T7 kit (Applied Biosystems). Amplified cRNA was purified with an RNeasy Mini Kit (Qiagen) and then subjected to FLAG conjugation as described (10) with some modifications. Briefly, 60 µl of freshly prepared 0.1 M $NaIO_4$ was added to 60 µl of 250 pmol cRNA, and the mixture was incubated at 0ºC for 10 min. The 3′ dialdehyde RNA was precipitated with 1 ml of 2% $LiClO_4$ in acetone followed by washing with 1 ml acetone. The pellet was dissolved in 10 µl of 0.1 M sodium acetate, pH 5.2 and then mixed with 12 µl of 30 mM hydrazide–FLAG peptide. The reaction

solution was mixed at room temperature for 30 min. The resulting imine-moiety of the cRNA was reduced by adding 12 µl of 1 M $NaCNBH_3$, and then incubated at room temperature for 30 min. The RNA was purified with an RNeasy Mini Kit (Qiagen).

Purification and analysis of RNA-binding proteins. Purification and analysis of RNA-binding protein (RBP) were carried out as described (24) with some modifications. Briefly, HEK293T cells were lysed with lysis buffer [10 mM HEPES (pH 7.5), 150 mM NaCl, 50 mM NaF, 1 mM Na_3VO_4, 5 µg/ml leupeptin, 5 µg ml aprotinin, 3 µg/ml pepstatin A, 1 mM phenylmethylsulfonyl fluoride (PMSF), and 1 mg/ml digitonin] and cleared by centrifugation. The cleared lysate was incubated with indicated amounts of FLAG-tagged bait RNA, antisense oligos and FLAG-M2-conjugated agarose for 1 hr. The agarose resin was then washed three times with wash buffer [10 mM HEPES (pH 7.5), 150 mM NaCl, and 0.1% Triton X-100] and co-immunoprecipitated RNA and proteins were eluted with FLAG elution buffer [0.5 mg/ml FLAG peptide, 10 mM HEPES (pH 7.5), 150 mM NaCl, and 0.05% Triton X-100]. The bait RNA associated proteins were digested with lysyl endopeptidase and trypsin. Digested peptide mixture was applied to a Mightysil-PR-18 (Kanto Chemical) frit-less column (45 3 0.150 mm ID) and separated using a 0–40% gradient of acetonitrile containing 0.1% formic acid for 80 min at a flow rate of 100 nl/min. Eluted peptides were sprayed directly into a mass spectrometer (Triple TOF 5600+; AB Sciex). MS and MS/MS spectra were obtained using the information-dependent mode. Up to 25 precursor ions above an intensity threshold of 50 counts/s were selected for MS/MS analyses from each survey scan. All MS/MS spectra were searched against protein sequences of RefSeq (NCBI) human protein database using the Protein Pilot software package (AB Sciex) and its decoy sequences then selected the peptides FDR was <1%. Ion intensity of peptide peaks ware obtained using Progenesis QI for proteomics software (version 3 Nonlinear Dynamics, UK) according to the manufacturer's instructions.

Ribosome profiling and RNA-Seq

Ribosome profiling was performed according to the method previously described with following modifications (25). RNA concentration of naïve HEK293T lysate was measured by Qubit RNA BR Assay Kit (Thermo Fisher Scientific). The lysate containing 10 µg RNA was treated with 20 U of RNase I (Lucigen) for 45 min at 25ºC. After ribosomes were recovered by ultracentrifugation, RNA fragments corresponding to 26-34 nt were excised from footprint fragment purification gel. Library length distribution was checked using a microchip electrophoresis system (MultiNA, MCE-202, Shimadzu).

For RNA-seq, total RNA was extracted from the lysate using TRIzol LS reagent (Thermo Fisher Scientific) and Direct-zol RNA Kit (Zymo research). Ribosomal RNA was depleted using the Ribo-Zero Gold rRNA Removal Kit (Human/Mouse/Rat) (Illumina) and the RNA-seq library was prepared using TruSeq Stranded mRNA Library Prep Kit (Illumina) according to the manufacturer's instructions.

The libraries were sequenced on a HiSeq 4000 (Illumina) with a single-end 50 bp sequencing run. Reads were aligned to human hg38 genome as described (25, 26). The offsets of A-site from the 5′ end of ribosome footprints were determined empirically as 15 for 25-30 nt, 16 for 31-32 nt, and 17 for 33 nt. For RNA-seq, offsets were set to 15 for all mRNA fragments. For calculation of the ribosome

occupancies, mRNAs with lower than one footprint per codon were excluded. For calculation of the translation efficiencies (TEs), we counted the number of reads within each CDS, and ribosome profiling counts were normalized by RNA-seq counts using the DESeq package (27). Reads corresponding to the first and last five codons of each CDS were omitted from the analysis of TEs. The Custom R scripts will be available upon requests.

Bioinformatics and Computational Analyses

Principal component analysis. CDS data (Human genes, GRCh38p12) was obtained from the Ensembl Biomart Database and the average human codon frequencies per transcript were calculated for all transcripts. The resulting data was subject to PCA analysis using the Python 3.4 environment via the factoextra program. Finally, the data was trimmed to remove truncated sequences as well as sequences with non-canonical start codons to a final of 9898 genes.

Hierarchical clustering analysis. mRNA transcripts ranked in order of their half-lives, divided equally into 4 groups and their average half-lives within each group was calculated. The corresponding codon frequencies of transcripts within each group were averaged. Hierarchical clustering was performed using the complete linkage method to cluster the codon frequencies in R.

Quantification of gene optimality scores. To quantify gene optimality, we summed up the codon frequencies of GC3 codons and expressed the frequencies on a percentage scale.

Binning of ribosomal occupancy frequencies and calculation of codon optimality-derived occupancy scores. To quantify codon optimality, the factor loading scores of the codons from the first principal component were normalized linearly on a percentage scale from 0 to 1 where 0 corresponded to the codon with the lowest optimality (AAT) and 1 for the codon with the highest optimality (GCC) (**Supplementary Figure 2A**). Binning of the ribosome occupancies were performed in the R environment via a custom script. To calculate the corresponding codon optimality-derived occupancy scores, we substituted the codon sequences of mRNA transcripts with their respective codon scores and in a similar fashion, binned the data into 25 bins. As the optimality scores of codons should inversely reflect the ribosome occupancy (i.e. higher ribosome occupancy associated with lower codon optimality scores), we calculated the inverse of the binned codon optimality scores within each bin for all 25 bins to derive the codon optimality-derived occupancy scores. Both ribosome occupancy and codon optimality-derived occupancy scores were normalized on a linear scale and a Pearson correlation performed on each transcript. To exclude the possibility that the correlations were due to chance, we shuffled the bins for the codon optimality-derived occupancy scores within each individual transcripts and calculated the Pearson correlation between shuffled and ribosomal occupancy data.

Statistical Analyses.

Unless mentioned, statistical significance was determined by performing a two-way student's T-test between samples. A two-way ANOVA with multiple comparisons (Holm-Sidak) was performed for the comparison between the HEK293 Tet-off degradation curves of *REL-OPT/WT* transcripts as well as the *IL6-OPT/WT/DE* samples. P-values are denoted as follows: $p < 0.05$ (*), $p<0.01$ (**), and $p<0.001$ (***).

RESULTS

Optimal and non-optimal codons in *Homo sapiens* can be categorized into GC3 and AT3 codons

To examine whether the codon-optimality-mediated decay exists in human, we first compared codon frequencies in *Homo sapiens* and other model organisms. Hierarchical clustering analysis of codon frequency data obtained from Ensembl database (28) showed a difference between lower eukaryotes such as *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, and higher eukaryotes such as *Homo sapiens* and *Mus musculus* **(Figure 1A)**. To investigate the codon bias in humans, we then calculated the codon frequencies of individual genes from *H. sapiens* and performed a principal component analysis (PCA) on the data. The first principal component (PC1) of the PCA which accounted for 22.85% of the total variance, divided codons into two clusters: codons with either G or C at the third base position (GC3) and codons with either A or T at the third base position (AT3) **(Figure 1B)**. Interestingly, the division within the second principal component (PC2) appeared to be split along the number of G/C or A/T bases in codons. We repeated our analysis on the CDS sequences from *S. cerevisiae* and found no such clustering **(Supplementary Figure 1A).** However, we discovered that the factor loading scores of the codons along the first principal component of our analysis in yeast corresponded to the CSC metric (12), albeit differences in the order **(Supplementary Figure 1B)**. The above-mentioned results therefore raised the possibility that GC3 and AT3 codons in humans may have a valid effect on mRNA stability.

We tested the link between mRNA stability and GC3-AT3 codons using published datasets of global mRNA decay rates in physiologically growing HEK293 cells (GSE69153) (29). Briefly, we divided the transcripts equally into quartiles based on their half-lives and averaged the codon frequencies within the quartiles. Strikingly, genes with short half-lives were associated with AT3 codons while genes with longer half-lives were associated with GC3 codons **(Figure 1C)**, suggesting a connection between third base of codons and the stability of mRNAs.

Broadly, the codon bias in mRNA can predict the stability of the mRNA. By summing the GC3 frequencies of CDS sequences, we could determine the GC3-content of a gene which we hereby refer to as gene optimality **(Supplementary Table 2)**. We then visualized the gene optimality landscape by plotting the gene optimality values via a histogram (**Figure 1D**). Gene optimality was represented as a bimodal distribution with a range of values from the minimum of 24.1% to the maximum of 100%. A Pearson correlation analysis (R^2 = 0.87) between GC-content and gene optimality **(Figure 1E)** reflected an enrichment of GC-content with increased gene optimality. Indeed, higher gene optimality scores were generally associated with better stability **(Figure 1F, Supplementary Figure 1C)**. We noted that, despite possessing a system of codon optimality in mRNA stability, the codon bias *per se* was different between yeast and humans **(Figure 1B and Supplementary Figure 1A)** (12)**.** Taken together, our analysis allowed us to designate GC3 and AT3 codons as optimal and non-optimal codons respectively. Additionally, high gene optimality in humans results in high GC-content, which is a feature of stable mRNAs.

We then asked about the biological relevance associated with gene optimality. Taking the 5% of lowest and highest ranked genes into account, we observed that genes with high optimality percentages were enriched in developmental processes while genes with low optimality percentages were enriched in cellular division processes **(Supplementary Figure 1D and 1E)**, suggesting the importance of codon-mediated mRNA decay across dynamic cellular processes in humans.

Ribosome profiling reveals that ribosome occupancy is correlated with codon optimality

It has been proposed that slower ribosome elongation rate modulated by low codon optimality affects the stability of mRNAs in yeast (12). This led us to examine whether decelerated ribosomes could be observed especially in regions where codon optimality was low. Because we speculated that a single codon would be insufficient in eliciting any noticeable effects on the speed of the ribosome, we divided each CDS into 25 bins from start codon to stop codon and summed the codon occupancy within each bin. From the PCA, PC1 factor loadings of the codons were indicative of how much a particular codon contributed to the AT3-GC3 grouping i.e. instability-stability (**Supplementary Figure 2A**). Therefore as a measure of estimating ribosome occupancy, the factor loading scores of the codons from the first principal component were utilized to derive codon optimality-derived scores. We then compared the codon optimality-derived scores with corresponding ribosome occupancies derived from ribosome profiling (25). Ribosome occupancy obtained from HEK293 cells growing under physiological conditions generally coincided with codon optimality-derived occupancy (**Figure 2A**). These measurements were highly reproducible between replicates of ribosome profiling experiments across the transcriptome (R^2 = 0.75, 16,423 transcripts) (**Supplementary Figure 2B**). We observed a significantly better prediction of ribosome occupancy by codon optimality-derived occupancy scores than that derived from scrambled codon optimality-derived occupancy scores (**Figure 2B**). Indeed, representative transcripts showed a good correlation between our scoring codon optimality-derived occupancy scores and ribosome occupancy as exemplified by EIF2B2, DYNC1LI2, and IDH3G transcripts (**Figure 2C**). Therefore, codon optimality determines ribosome occupancy in humans.

Although translation elongation and initiation are distinct steps, previous literature has suggested that optimal codons are also enriched in mRNAs with high translation (30). Ribosome footprint reads normalized by mRNA abundances from RNA-Seq enables the calculation of translation efficiency which in turn is also generally regarded as the translation initiation rate (31). Therefore, to investigate the link between translation status and codon optimality, we calculated the translation efficiency (TE)—ribosome footprints normalized by mRNA abundance. Indeed, our results showed that mRNAs with high codon optimality generally possessed high TE (**Figure 2D**).

Codon optimality affects mRNA stability

We then experimentally validated our bioinformatics observations of codon optimality-mediated decay in human cells. We developed a scheme based on the PCA PC1 factor loadings in which codons could be substituted with their optimized and non-optimized counterparts within their codon boxes *i.e.*

synonymous substitutions (**Supplementary Figure 3A)**. Single box codons such as TGG (Trp) and ATG (Met) would remain unchanged. We synthesized two independent genes (*REL* and *IL6*) with differential codon optimalities (**Supplementary Figure 3B, Supplementary Table 1**) and examined the stability of these reporter RNA in HEK293 cells utilizing the Tet-off system (**Figure 3A**). As expected, the optimized transcripts of *REL* and *IL6* were more stable than their wild-type counterparts. Additionally, the decay rate of the de-optimized *IL6* reporter was faster, confirming that low optimality transcripts were unstable.

In addition to the RNA stability, higher codon optimality was also associated with higher translation efficiency (**Figure 2D**), thereby increasing protein production. Indeed, the protein abundance of the optimized *REL* reporter was higher than *REL-WT* even after normalization of protein abundance by mRNA levels (**Figure 3B, Supplementary Figure 3C**). Using enzyme-linked immunosorbent assay (ELISA), we observed that expression of *IL6-OPT* resulted in a 1.5-fold and 2-fold significantly higher level of IL6 compared to its WT and *IL6-DE,* respectively (**Figure 3C**). In a similar fashion, normalization of IL6 protein abundance by mRNA levels revealed that translation efficiency of the optimized *IL6* reporter was higher than its WT and deoptimized reporter counterparts (**Supplementary Figure 3D**). We tested our REL reporters in HeLa cells and show that the high protein abundance of *REL-OPT* could also be observed (**Supplementary Figure 3E**). Moreover, polysome fractionation and subsequent qPCR analysis revealed that within the polysome fractions, *REL-OPT* transcript amounts were approximately 6-fold higher than *REL-WT* transcripts, suggesting that *REL-OPT* was translated more efficiently than *REL-WT* (**Figure 3D**). Taken together, our results validate bioinformatics analyses and show that mRNA stability is a function of codon optimality in humans.

RNA binding proteins can regulate mRNA stability via GC- or AU-content

Having shown that high optimality content inevitably accords high GC-content which in turn promotes mRNA stability, we wondered if there were RNA binding proteins (RBPs) which scrutinize, discriminate or even affect an mRNA's fate. To identify RBPs which were either bound to transcripts bearing high or low optimality, we performed a ribonucleoprotein immunoprecipitation-based approach termed ISRIM (*In vitro* Specificity based RNA Regulatory protein Identification Method) (24). Lysates of HEK293 cells were mixed with FLAG peptide-conjugated *REL* and *IL6* transcripts of high and low optimality and their interacting proteins were determined using mass spectrometry. We then selected RBP candidates which were specifically enriched with either low or high optimality transcripts common to both REL and IL6 ISRIM experiments (**Figure 4A, Supplementary Table 3**). As IL6 transcripts possessed three levels of optimality (OPT, WT, DE), we defined high optimality-binding RBPs based on the RBP enrichment of *IL6-OPT* compared *to IL6-WT* as well as *IL6-WT* to *IL6-DE* (**Supplementary Figure 4A, Supplementary Table 3**). Similarly, we defined low optimality-binding RBPs based on the RBP enrichment of *IL6-DE* compared to *IL6-WT* as well as *IL6-WT* to *IL6-OPT*. Of interest were ILF2 and ILF3, RPBs identified from the list of RBPs interacting exclusively with low optimality transcripts. ILF2 and ILF3, also known as NF45 and NF90, respectively, are well known to be able to function dominantly as heterodimers which bind double stranded RNA. ILF3 has been

extensively studied, having shown to bind to AU-rich sequences in 3′ UTR of target RNA to repress its translation (32). We hypothesize that the binding of ILF2 and ILF3 to their targets occur as low optimality transcripts are inadvertently AU-rich. Here we focused on the effects of ILF2 on low optimality transcripts.

We investigated ILF2's role in codon optimality-mediated RNA decay. Firstly, using published RIP-seq data of ILF2 in two multiple myeloma cell lines, H929 and JJN3, we observed that ILF2 interacts with low optimality transcripts (**Supplementary Figure 4B**) (33). Additionally, we analysed RNA-Seq data obtained from the ENCODE project (ENCSR073QLQ) of K562 cells treated by CRISPR interference targeting ILF2 (34). Strikingly, we observed that transcripts that possessed low optimality scores were upregulated whereas transcripts that possessed high optimality scores were downregulated (**Figure 4B, Supplementary Figure 4C**). The abundance changes of representative mRNAs by ILF2 knockdown were antiparallel to their codon optimalities (**Supplementary Figure 4D**).

To confirm our observations, we examined the stability of FLAG-tagged versions of *REL-OPT* and *REL-WT* in the Tet-off system after ILF2 knockdown via siRNA (**Figure 4C**). Indeed, we observed that the optimized reporter was more unstable whereas the low optimality WT reporter was more stable. We then expressed FLAG-tagged versions of *REL-OPT* and *REL-WT*, along with the two isoforms of ILF2. A significant decrease in band intensity was observed for the REL-WT bands when both isoforms of ILF2 were expressed, whereas the amount of REL-OPT was not changed (**Figure 4D and Supplementary Figure 4E**). We also found a significant increase in protein levels of REL-WT when cells were treated with ILF2-targeting siRNA (**Figure 4E and Supplementary Figure 4F**). Taken together, our results suggest that ILF2 targets low optimality mRNA transcripts to induce their decay.

DISCUSSION

This study provides a framework describing codon optimality-mediated RNA decay in humans. We first show that GC3 codons are optimal codons which are associated with stability. We quantified gene optimality by calculating the GC3 content within the CDS of genes and showed that gene optimality is strongly correlated with RNA stability and amount of protein expressed. In general, the use of optimal GC3 codons correlated with higher GC-content at a genome-wide level. We then show a modest agreement between codon optimality-derived scores and ribosome occupancy as determined by ribosome profiling. Screening of RNA binding proteins revealed that ILF2, which interacts with transcripts with low optimality, induces their degradation. Taken together, we conclude that gene expression can be shaped by codon optimality and inevitably by GC/AU-content through the modulation of mRNA stability in human cells.

Establishing the Presence of a System of Codon Optimality in Humans

Since translation elongation is affected by tRNA availability, the tRNA adaptation index (tAI), which is based on genomic tRNA copy number, has been used as a surrogate for codon optimality. However, in contrast to yeast, tRNA copy number in genome is not always correlated with tRNA abundance in higher eukaryotes (35). Hence, this metric is less suitable for quantifying codon optimality in humans.

Independent of tRNA-based metrics, we addressed these challenges by utilizing an unsupervised learning algorithm, PCA, to identify features in that were mRNA-intrinsic. In the PCA of both yeast and humans, we demonstrated that the first principal component mirrored optimal/non-optimal assignments. We also show that the codon bias is different between these two organisms (**Figure 1B and Supplementary Figure 1A**). In humans, the classification of codons into AT3 and GC3 groups was striking, but the percentage by which it accounts for its variation however was modest. From our PCA, the first and second principal components only explain a quarter of total variance in codon frequencies (**Figure 1B**), implying that other factors that explain bias of codon frequency possibly remains in human cells.

Assuming that evolution drives the selection of codons, synonymous codon usage in different organisms must be fine-tuned over time to achieve precise expression levels of mRNA and eventually proteins in essential physiological process. It is therefore plausible that high GC- or AT-content in mRNA is selected for, but subject to constraints by amino sequence, to modulate transcript stability in essential physiological processes. Indeed, we show that transcripts with top and bottom 5% of gene optimality were linked to particular physiological and cellular processes (**Supplementary figure 1D and E**). In a particular study, Gingold and colleagues argue that tRNA abundances vary in proliferating and differentiating cell types (36). Interestingly, they showed that codons preferred by cell cycling genes were AT3 codons while pattern-specification preferred codons tended to be GC3 codons—in agreement with our GO analyses. In *Drosophila*, the correlation between codon optimality and mRNA stability has been demonstrated to be attenuated in neural development, possibly allowing the effect of trans-acting factors to dominate development (22). c-Rel, a protein encoded by the *REL* gene and a canonical nuclear factor κB (NF-κB) subunit, is expressed abundantly in differentiated lymphoid cells and has been shown to be vital in thymic regulatory T cell development in addition to controlling cancer via activated regulatory T cells (37, 38). Given the inherent low optimality and associated instability of *REL* in its WT form (**Figure 3A**, **Supplementary Figures 3B**), we wonder if besides transcriptional control of *REL*, could there be other post-transcriptional regulation systems at play. Further studies would be necessary to investigate if codon optimality or codon optimality-associated RBPs modulate *REL* gene expression.

Our results show that codon optimality affects ribosome occupancy to a significant extent (**Figure 2B**). Our study along with others' suggests that slower elongation of ribosome is a key feature of mRNA stability. Although codon optimality is a dominant factor in general, other factors may also be involved in decelerated ribosomes, such as secondary structures (39, 40). These obstacles for ribosome elongation are reversible and dynamically regulated by RNA helicases (41, 42). Further studies will be required to elucidate the role of RNA secondary structures and helicases and their relevance to codon optimality decay.

The stability of mRNA can be modulated by RBPs which bind AU-rich sequences

Whereas AU-rich elements (AREs) in the 3′ UTR have been traditionally targeted by RBPs, we found that coding regions are also targeted by ARE-recognizing RBPs. The identification of the

heterodimeric complex consisting of ILF2 and ILF3 among others shows that a wide array of RBPs recognizes low optimality (AU-rich) sequences (**Figure 4 and Supplementary Figure 4**). Interestingly, Kuwano and colleagues show that NF90, the shorter isoform of ILF3, specifically targets AU-rich sequences in mRNA 3'UTRs and represses their translation, not stability (32). Taking into consideration reports that ILF2 and ILF3 can function independently of each other (42–44), it is possible that ILF2 and ILF3 regulate the fate of mRNA differently. Our screens also detected HNRNPD/AUF1, which destabilizes transcripts via recognition of AU-rich motifs (46), binding to low optimality mRNAs (Supplementary Table 3). These observations emphasize the importance of AU-content, which is strongly connected with low optimality, in RNA destabilization. However, we could not exclude the possibility that these factors induced the degradation of AU-rich transcripts independent of the model proposed by Presnyak and Radhakrishnan (12, 13) as our RBP identification method was not reflective of the active translational status required for co-translational degradation of mRNA transcripts. Further studies would be necessary to elucidate if these or other factors act as sensors of codon optimality during translation.

In conclusion, in human cells, the redundancy of the genetic code allows the choice between alternative codons for the same amino acid which may exert dramatic effects on the process of translation and mRNA stability. This system potentially confers freedom for calibrating protein and mRNA abundances without altering protein sequence. Beginning from our exploratory analysis, we have developed an approach to quantify optimality and demonstrate that beneath the redundancy of codons, exists a system which modulates mRNA and consequently, protein abundance.

AVAILABILITY

- Ensembl is a genome browser for vertebrate genomes which contains tools such as BioMart
  - http://www.ensembl.org/index.html
- The ENCODE Portal contains the data generated by the ENCODE Consortium
  - https://www.encodeproject.org/
- R is free software environment for statistical computing and analysis; associated tools are listed below
  - https://www.r-project.org/
  - Differential gene expression analysis, edgeR and DESeq:
    https://bioconductor.org/packages/release/bioc/html/edgeR.html
    http://bioconductor.org/packages/release/bioc/html/DESeq.html
  - Principal component analysis, factoextra:
    https://www.rdocumentation.org/packages/factoextra/versions/1.0.3
  - Hierarchical clustering analysis, gplots
    https://cran.r-project.org/web/packages/gplots/index.html

ACCESSION NUMBERS

Global mRNA decay rates of mRNA from HEK293 cells were accessed at the NCBI GEO database under the dataset GSE69153. RNA-seq data derived from RNA-seq on K562 cells treated by CRISPR interference targeting ILF2 were accessed at the ENCODE database under the dataset ENCSR073QLQ. Ribosome profiling and RNA-Seq results of HEK293 cells have been deposited at GEO and can be accessed under dataset GSE126298.

ACKNOWLEDGEMENT

The authors express their gratitude to all members of the laboratory of Medical Chemistry, Kyoto University, for their kind advice and discussions. DNA libraries were sequenced by the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. Computations were supported by Manabu Ishii, Itoshi Nikaido, and the Bioinformatics Analysis Environment Service on RIKEN Cloud at RIKEN ACCC.

FUNDING

This work was supported by the JSPS KAKENHI (18H05278), AMED-CREST from Japan Agency for Medical Research and Development and the JSPS through Core-to-Core Program.

This work was supported by Joint Usage/Research Center program of Institute for Frontier Life and Medical Sciences, Takeda Science Foundation, the Uehara Memorial Foundation.

S.I. was supported by Grant-in-Aid for Scientific Research on Innovative Areas “nascent chain biology” (JP17H05679) and Grant-in-Aid for Young Scientists (A) (JP17H04998) from JSPS, the Pioneering Projects ("Cellular Evolution") and the Aging Project from RIKEN, and Takeda Science Foundation.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

REFERENCES

1. Huang,L., Lou,C.-H., Chan,W., Shum,E.Y., Shao,A., Stone,E., Karam,R., Song,H.-W. and Wilkinson,M.F. (2011) RNA Homeostasis Governed by Cell Type-Specific and Branched Feedback Loops Acting on NMD. *Molecular Cell*, **43**, 950–961.

2. Mino,T., Murakawa,Y., Fukao,A., Vandenbon,A., Wessels,H.-H., Ori,D., Uehata,T., Tartey,S., Akira,S., Suzuki,Y., *et al.* (2015) Regnase-1 and Roquin Regulate a Common Element in Inflammatory mRNAs by Spatiotemporally Distinct Mechanisms. *Cell*, **161**, 1058–1073.

3. Yoshinaga,M., Nakatsuka,Y., Vandenbon,A., Ori,D., Uehata,T., Tsujimura,T., Suzuki,Y., Mino,T. and Takeuchi,O. (2017) Regnase-1 Maintains Iron Homeostasis via the Degradation of Transferrin Receptor 1 and Prolyl-Hydroxylase-Domain-Containing Protein 3 mRNAs. *Cell Rep*, **19**, 1614–1630.

4. Leppek,K., Das,R. and Barna,M. (2018) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol*, **19**, 158–174.

5. Cheng,J., Maier,K.C., Avsec,Ž., Rus,P. and Gagneur,J. (2017) Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA*, **23**, 1648–1659.

6. Vogel,C. and Marcotte,E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, **13**, 227–232.

7. Zhou,T., Weems,M. and Wilke,C.O. (2009) Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. *Mol Biol Evol*, **26**, 1571–1580.

8. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, **15**, 1281–1295.

9. Reis,M. dos, Savva,R. and Wernisch,L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*, **32**, 5036–5044.

10. dos Reis,M., Wernisch,L. and Savva,R. (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res*, **31**, 6976–6985.

11. Pechmann,S. and Frydman,J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of co-translational folding. *Nat Struct Mol Biol*, **20**, 237–243.

12. Presnyak,V., Alhusaini,N., Chen,Y.-H., Martin,S., Morris,N., Kline,N., Olson,S., Weinberg,D., Baker,K.E., Graveley,B.R., *et al.* (2015) Codon optimality is a major determinant of mRNA stability. *Cell*, **160**, 1111–1124.

13. Radhakrishnan,A., Chen,Y.-H., Martin,S., Alhusaini,N., Green,R. and Coller,J. (2016) The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell*, **167**, 122-132.e9.

14. Sweet,T., Kovalak,C. and Coller,J. (2012) The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement. *PLoS Biol.*, **10**, e1001342.

15. Dana,A. and Tuller,T. (2014) The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res*, **42**, 9171–9181.

16. Gardin,J., Yeasmin,R., Yurovsky,A., Cai,Y., Skiena,S. and Futcher,B. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, **3**.

17. Drummond,D.A. and Wilke,C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**, 341–352.

18. Akashi,H. (1994) Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. *Genetics*, **136**, 927–935.

19. Harigaya,Y. and Parker,R. (2016) Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. *BMC Genomics*, **17**, 895.

20. Lee,Y., Zhou,T., Tartaglia,G.G., Vendruscolo,M. and Wilke,C.O. (2010) Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics*, **10**, 4163–4171.

21. Mishima,Y. and Tomari,Y. (2016) Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular Cell*, **61**, 874–885.

22. Burow,D.A., Martin,S., Quail,J.F., Alhusaini,N., Coller,J. and Cleary,M.D. (2018) Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. *Cell Rep*, **24**, 1704–1712.

23. Boël,G., Letso,R., Neely,H., Price,W.N., Wong,K.-H., Su,M., Luff,J., Valecha,M., Everett,J.K., Acton,T.B., *et al.* (2016) Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, **529**, 358–363.

24. Adachi,S., Homoto,M., Tanaka,R., Hioki,Y., Murakami,H., Suga,H., Matsumoto,M., Nakayama,K.I., Hatta,T., Iemura,S., *et al.* (2014) ZFP36L1 and ZFP36L2 control LDLR mRNA stability via the ERK–RSK pathway. *Nucleic Acids Res*, **42**, 10037–10049.

25. McGlincy,N.J. and Ingolia,N.T. (2017) Transcriptome-wide measurement of translation by ribosome profiling. *Methods*, **126**, 112–129.

26. Cap-specific terminal N6-methylation of RNA by an RNA polymerase II–associated methyltransferase | Science.

27. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

28. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res*, **46**, D754–D761.

29. Murakawa,Y., Hinz,M., Mothes,J., Schuetz,A., Uhl,M., Wyler,E., Yasuda,T., Mastrobuoni,G., Friedel,C.C., Dölken,L., *et al.* (2015) RC3H1 post-transcriptionally regulates A20 mRNA and modulates the activity of the IKK/NF-κB pathway. *Nat Commun*, **6**, 7367.

30. Tuller,T., Kupiec,M. and Ruppin,E. (2007) Determinants of Protein Abundance and Translation Efficiency in S. cerevisiae. *PLOS Computational Biology*, **3**, e248.

31. Iwasaki,S. and Ingolia,N.T. (2016) Seeing translation. *Science*, **352**, 1391–1392.

32. Kuwano,Y., Pullmann,R., Marasa,B.S., Abdelmohsen,K., Lee,E.K., Yang,X., Martindale,J.L., Zhan,M. and Gorospe,M. (2010) NF90 selectively represses the translation of target mRNAs bearing an AU-rich signature motif. *Nucleic Acids Res*, **38**, 225–238.

33. Marchesini,M., Ogoti,Y., Fiorini,E., Aktas Samur,A., Nezi,L., D'Anca,M., Storti,P., Samur,M.K., Ganan-Gomez,I., Fulciniti,M.T., *et al.* (2017) ILF2 Is a Regulator of RNA Splicing and DNA Damage Response in 1q21-Amplified Multiple Myeloma. *Cancer Cell*, **32**, 88-100.e6.

34. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

35. Zheng,G., Qin,Y., Clark,W.C., Dai,Q., Yi,C., He,C., Lambowitz,A.M. and Pan,T. (2015) Efficient and quantitative high-throughput transfer RNA sequencing. *Nat Methods*, **12**, 835–837.

36. Gingold,H., Tehler,D., Christoffersen,N.R., Nielsen,M.M., Asmar,F., Kooistra,S.M., Christophersen,N.S., Christensen,L.L., Borre,M., Sørensen,K.D., *et al.* (2014) A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. *Cell*, **158**, 1281–1292.

37. Grinberg-Bleyer,Y., Oh,H., Desrichard,A., Bhatt,D.M., Caron,R., Chan,T.A., Schmid,R.M., Hayden,M.S., Klein,U. and Ghosh,S. (2017) NF-κB c-Rel Is Crucial for the Regulatory T Cell Immune Checkpoint in Cancer. *Cell*, **170**, 1096-1108.e13.

38. Oh,H., Grinberg-Bleyer,Y., Liao,W., Maloney,D., Wang,P., Wu,Z., Wang,J., Bhatt,D.M., Heise,N., Schmid,R.M., *et al.* (2017) An NF-κB Transcription-Factor-Dependent Lineage-Specific Transcriptional Program Promotes Regulatory T Cell Identity and Function. *Immunity*, **47**, 450-465.e5.

39. Pop,C., Rouskin,S., Ingolia,N.T., Han,L., Phizicky,E.M., Weissman,J.S. and Koller,D. (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.*, **10**, 770.

40. Endoh,T. and Sugimoto,N. (2016) Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells. *Sci Rep*, **6**, 22719.

41. Thandapani,P., Song,J., Gandin,V., Cai,Y., Rouleau,S.G., Garant,J.-M., Boisvert,F.-M., Yu,Z., Perreault,J.-P., Topisirovic,I., *et al.* (2015) Aven recognition of RNA G-quadruplexes regulates translation of the mixed lineage leukemia protooncogenes. *Elife*, **4**.

42. Pan,L., Li,Y., Zhang,H.-Y., Zheng,Y., Liu,X.-L., Hu,Z., Wang,Y., Wang,J., Cai,Y.-H., Liu,Q., *et al.* (2017) DHX15 is associated with poor prognosis in acute myeloid leukemia (AML) and regulates cell apoptosis via the NF-kB signaling pathway. *Oncotarget*, **8**, 89643–89654.

43. Harashima,A., Guettouche,T. and Barber,G.N. (2010) Phosphorylation of the NFAR proteins by the dsRNA-dependent protein kinase PKR constitutes a novel mechanism of translational regulation and cellular defense. *Genes Dev.*, **24**, 2640–2653.

44. Wolkowicz,U.M. and Cook,A.G. (2012) NF45 dimerizes with NF90, Zfr and SPNR via a conserved domain that has a nucleotidyltransferase fold. *Nucleic Acids Res*, **40**, 9356–9368.

45. Graber,T., Baird,S., Kao,P., Mathews,M. and Holcik,M. (2010) NF45 functions as an IRES trans-acting factor that is required for translation of cIAP1 during the unfolded protein response. *Cell Death Differ*, **17**, 719–729.

46. Gratacós,F.M. and Brewer,G. (2010) The role of AUF1 in regulated mRNA decay. *Wiley Interdiscip Rev RNA*, **1**, 457–473.

TABLE AND FIGURES LEGENDS

Figure 1. Bioinformatics analysis reveals that optimal and non-optimal codons can be categorized into GC3 and AT3 codons respectively.

(A) Hierarchical clustering analysis of model organisms and their average CDS codon frequencies.
(B) Principal component analysis of the CDS codon frequencies of 20176 protein-coding genes. F1 and F2 indicate the first and second principal components.
(C) Hierarchical clustering analysis of the half-lives of mRNA and their CDS codon frequencies. The transcripts were ranked according to their half-lives and divided equally into quartiles. The respective codon frequencies of each group were averaged and the hierarchical clustering performed.
(D) Histogram illustrating the distribution of genes and their respective gene optimality scores.
(E) Pearson correlation between GC-content and gene optimality.
(F) Comparison of average transcript mRNA half-lives across their respective gene optimality ranges. Number of transcripts within each gene optimality range is indicated above their respective points. Error bars represent the 95% confidence intervals.

Figure 2. Ribosome profiling reveals that ribosome occupancy is moderately correlated with codon optimality.

(A) Average ribosome occupancy and their respective codon optimality-derived occupancy scores across the CDS of transcript (in 25 bins). Ribosome occupancy for 16,423 transcripts and their respective codon optimality-derived occupancy were firstly binned into 25 bins and the mean occupancy was calculated for each bin. Error bars represents the 95% confidence intervals.

(B) Cumulative distribution plots showing the distributions of correlations between ribosome occupancy and codon optimality-derived occupancy. Correlations obtained from the ribosome occupancy and scrambled codon optimality-derived occupancy served as the control. A Kolmogorov–Smirnov test was performed between the codon optimality-derived occupancy and the control group.
(C) Three example transcripts (EIF2B2, DYNC1LI2 and IDH3G) which demonstrate high correlation between ribosome occupancy and codon optimality-derived scores (left) as well as their corresponding Pearson correlations over 25 bins (right).
(D) Comparison of average transcript translation efficiencies (TEs) across their respective gene optimality ranges. Number of transcripts within each gene optimality range is indicated above their respective points. Error bars represent the 95% confidence intervals.

Figure 3. Optimality of transcripts determine their fate.

(A) HEK293 Tet-off experiments showing the degradation of *REL-OPT* and *REL-WT* transcripts (left) as well as *IL6-OPT*, *IL6-WT* and *IL6-DE* transcripts (right), post-doxycycline addition. Data is representative of 3 independent experiments in which the data represents the mean ± SD for 3 biological replicates. A two-way ANOVA with Holm-Sidak multiple comparisons was performed. P-values are denoted as follows: $p < 0.05$ (*), $p<0.01$ (**) and $p<0.001$ (***).
(B) Representative immunoblot of FLAG-tagged REL-OPT and REL-WT in HEK293T cells transfected with either empty plasmids, plasmids bearing *REL-OPT* or *REL-WT*. The immunoblot is representative of 3 independent experiments. ACTB was shown as loading controls.
(C) ELISA of secreted IL6 concentrations of IL6-OPT, IL6-WT and IL6-DE from HEK293T cells transfected with plasmids bearing *IL6-OPT*, *IL6-WT* and *IL6-DE*. The data is representative of 3 independent experiments. A one-way ANOVA with Tukey's multiple comparisons was performed between samples where, $p<0.01$ (**) and $p<0.001$ (***).
(D) Fold changes of *REL-OPT* relative to *REL-WT* transcript levels (top) as detected by qPCR across polysome fractions (below). Data represents the mean ± SD for 3 biological replicates.

Figure 4. RNA binding proteins can regulate mRNA stability via GC- or AU-content.

(A) Venn diagram indicating the number of RBPs identified from the REL and IL6 ISRIM experiments.
(B) Cumulative distribution plots showing the difference in distribution of transcript optimality between upregulated and downregulated transcripts in K562 cells subject to ILF2 CRISPR interference targeting ILF2. Kolmogorov–Smirnov tests were performed on the upregulated and downregulated groups against the control group. P-values are denoted (right).
(C) HEK293 Tet-off experiments showing the degradation of *REL-OPT* and *REL-WT* transcripts with ILF2 siRNA and Control (CTR) siRNA treatment, post-doxycycline addition. Data is representative of 3 independent experiments in which the data represents the mean ± SD for 3 biological replicates. Unpaired t-tests were performed between samples where, $p < 0.05$ (*).
(D) Representative immunoblot of FLAG-tagged REL-OPT and REL-WT in HEK293T cells co-expressed with two different isoforms of ILF2. The immunoblot is representative of 3 independent experiments. ACTB was shown as loading controls.
(E) Representative immunoblot of FLAG-tagged REL-OPT and REL-WT expressed in HEK293T cells under ILF2 siRNA treatment. The immunoblot is representative of 3 independent experiments. ACTB was shown as loading controls.

Supplementary Figure 1 | Related to Figure 1.

(A) Principal component analysis of the CDS codon frequencies of protein-coding genes in S. cerevisiae. F1 and F2 indicate the first and second principal components.
(B) F1 factor loadings of codons from the yeast dataset ranked from the highest to the lowest. The optimal and non-optimal designation at the bottom of the figure refers to the designation according to Presnyak and colleagues (12).
(C) Violin plots visualizing the distribution of mRNA half-lives across their respective gene optimality brackets.

(D) Gene ontology analysis (biological processes) of the top 5% ranked genes in terms of gene optimality
(E) Gene ontology analysis (biological processes) of the bottom 5% ranked genes in terms of gene optimality

Supplementary Figure 2 | Related to Figure 2.

(A) PC1 factor loadings of codons from the human dataset ranked from the highest to the lowest (bottom) and their corresponding percentage optimality scores after linear normalization onto a percentage scale (top).
(B) Pearson correlation between the correlations of derived from comparison of ribosome occupancy and codon optimality-derived scores for two ribosome profiling sample replicates.

Supplementary Figure 3 | Related to Figure 3.

(A) Example of how transcript optimization and deoptimization was performed to generate optimized and deoptimized versions of *REL* and *IL6* transcripts.
(B) The gene optimality scores of *REL-OPT/WT* and *IL6-OPT/WT/DE*.
(C) Protein abundance of immunoblot of FLAG-tagged REL-OPT and REL-WT in HEK293T cells transfected with either empty plasmids, plasmids bearing *REL-OPT* or *REL-WT* (corresponding to Figure 3D). The protein abundance was normalized by respective mRNA levels. The densitometry data is representative of 3 independent experiments. Unpaired t-tests were performed within the REL-OPT and REL-WT samples, $p<0.05$ (*).
(D) Protein abundance as determined by ELISA of IL6-OPT, IL6-WT and IL6-DE in HEK293T cells transfected with either empty plasmids, plasmids bearing *IL6-OPT*, *IL6-WT* or *IL6-DE* (corresponding to Figure 3E). The protein abundance was normalized by respective mRNA levels. The ELISA quantification is representative of 3 independent experiments. A one-way ANOVA with Tukey's multiple comparisons was performed between samples where, $p<0.01$ (**) and $p<0.001$ (***).
(E) Representative immunoblot of FLAG-tagged *REL-OPT* and *REL-WT* expressed in HeLa cells (left). The immunoblot is representative of 3 independent experiments.

Supplementary Figure 4 | Related to Figure 4.

(A) Venn diagram indicating the number of RBPs identified from the IL6 ISRIM experiments.
(B) Cumulative distribution plots showing the optimality score distribution of transcripts bound to by ILF2 in H929 (top) and JJN3 cells (bottom). Kolmogorov–Smirnov tests were performed on the ILF2 RIP group against the control group.
(C) Scatterplot of the RPKM values of mRNA transcripts in K562 cells subject to ILF2 CRISPR interference and its corresponding WT control. mRNA transcripts were coloured according to their gene optimality scores.
(D) Fold changes of example mRNA representing low, average and high optimality transcripts from the RPKM values of mRNA transcripts in K562 cells subject to ILF2 CRISPR interference. Data represents the mean ± SD.
(E) Densitometric analysis of immunoblot of FLAG-tagged REL-OPT and REL-WT expressed in HEK293T cells co-expressed with two different isoforms of ILF2 (corresponding to Figure 4D). The densitometry data is representative of 3 independent experiments. Unpaired t-tests were performed within the REL-OPT and REL-WT samples, $p<0.01$ (**).
(F) Densitometric analysis of immunoblot of FLAG-tagged REL-OPT and REL-WT expressed in HEK293T cells under ILF2 siRNA treatment (corresponding to Figure 4E). The densitometry data is representative of 3 independent experiments. Unpaired t-tests were performed within the REL-OPT and REL-WT samples. P-values are denoted as follows, $p < 0.05$ (*).

Supplementary Table 1. List of qPCR primers and their corresponding sequences.

Supplementary Table 2. List of genes and their corresponding GC3 content / gene optimality.

Supplementary Table 3. List of RBPs identified in ISRIM experiments.

Figure 1

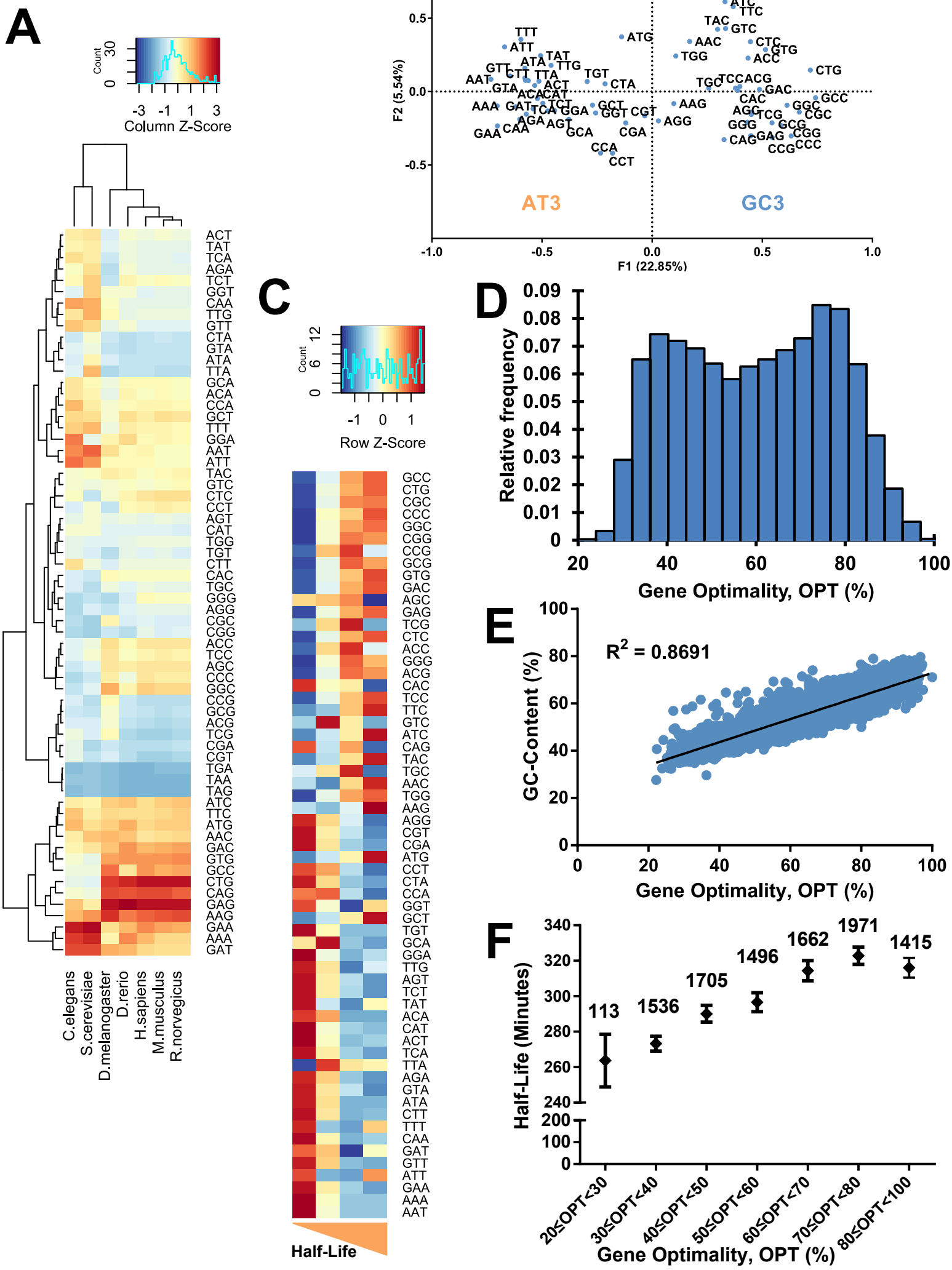

A
Count
Column Z-Score
C.elegans
S.cerevisiae
D.melanogaster
D.rerio
H.sapiens
M.musculus
R.norvegicus
B
F2 (5.54%)
F1 (22.85%)
AT3
GC3
C
Row Z-Score
Half-Life
D
Relative frequency
Gene Optimality, OPT (%)
E
R^2 = 0.8691
GC-Content (%)
Gene Optimality, OPT (%)
F
Half-Life (Minutes)
113
1536
1705
1496
1662
1971
1415
20≤OPT<30
30≤OPT<40
40≤OPT<50
50≤OPT<60
60≤OPT<70
70≤OPT<80
80≤OPT<100
Gene Optimality, OPT (%)

Figure 2

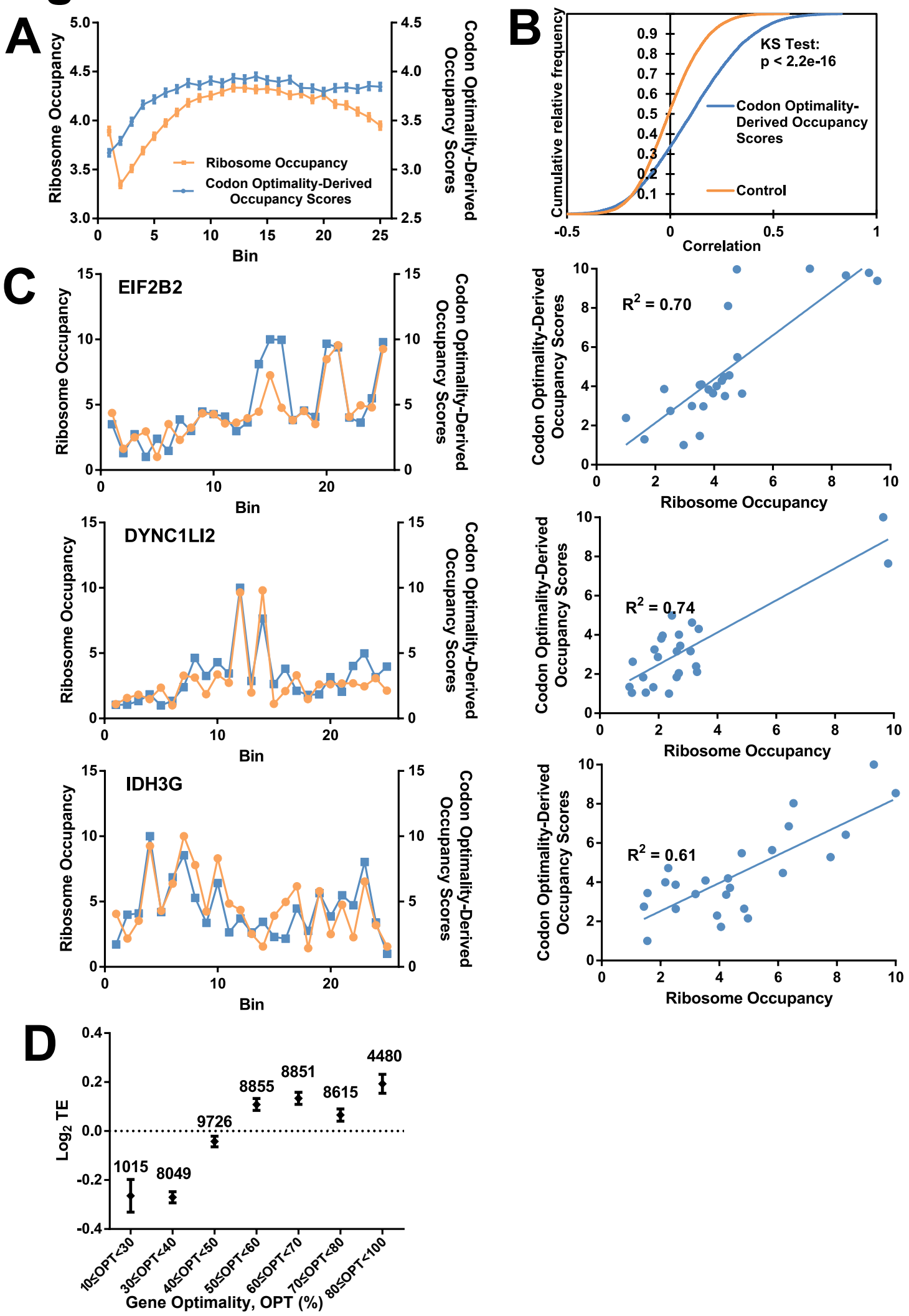
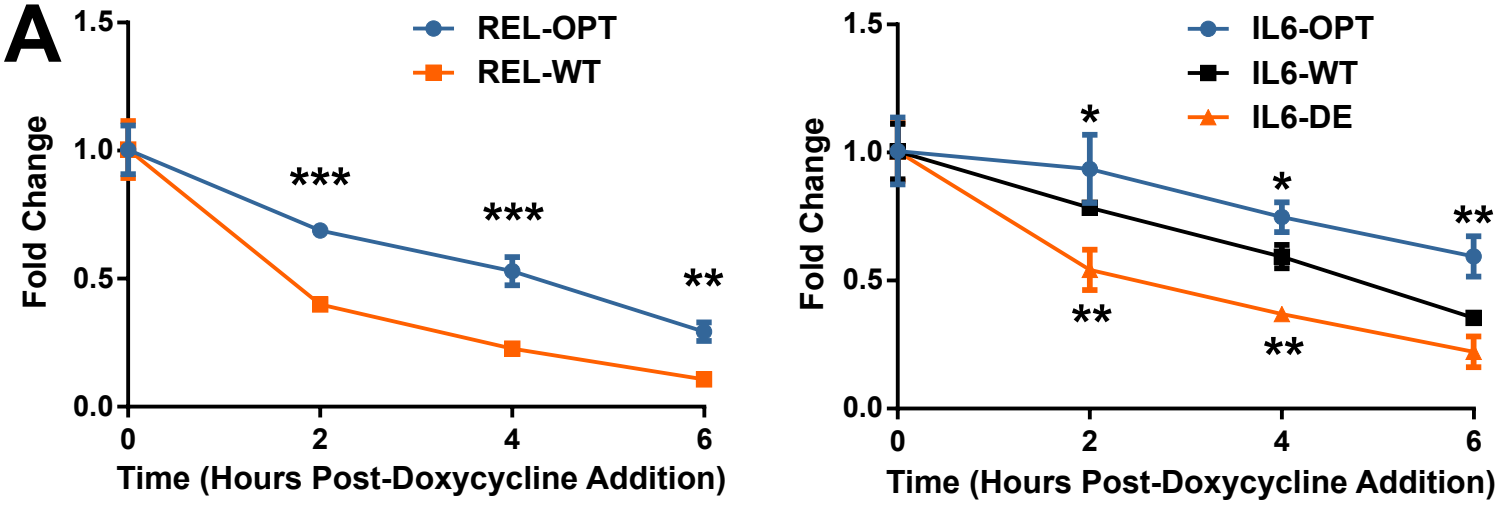
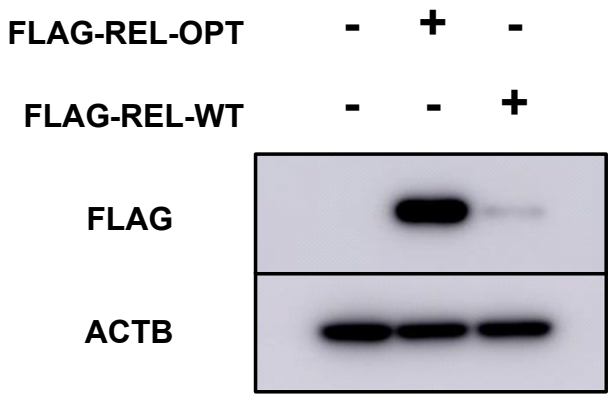
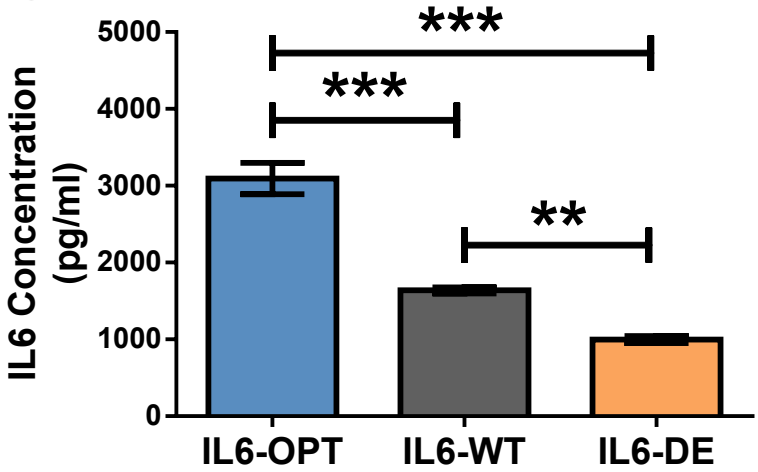

A
Ribosome Occupancy
Codon Optimality-Derived Occupancy Scores
Bin
Ribosome Occupancy
Codon Optimality-Derived Occupancy Scores
B
Cumulative relative frequency
KS Test:
p < 2.2e-16
Codon Optimality-Derived Occupancy Scores
Control
Correlation
C
EIF2B2
DYNC1LI2
IDH3G
Ribosome Occupancy
Codon Optimality-Derived Occupancy Scores
Bin
R2 = 0.70
R2 = 0.74
R2 = 0.61
D
Log2 TE
1015
8049
9726
8855
8851
8615
4480
10≤OPT<30
30≤OPT<40
40≤OPT<50
50≤OPT<60
60≤OPT<70
70≤OPT<80
80≤OPT<100
Gene Optimality, OPT (%)

Figure 3


A
REL-OPT
REL-WT
Fold Change
1.5
1.0
0.5
0.0
***
***
**
0
2
4
6
Time (Hours Post-Doxycycline Addition)
IL6-OPT
IL6-WT
IL6-DE
*
*
**
**
**


B
FLAG-REL-OPT - + -
FLAG-REL-WT - - +
FLAG
ACTB


C
IL6 Concentration (pg/ml)
5000
4000
3000
2000
1000
0
***
***
**
IL6-OPT
IL6-WT
IL6-DE

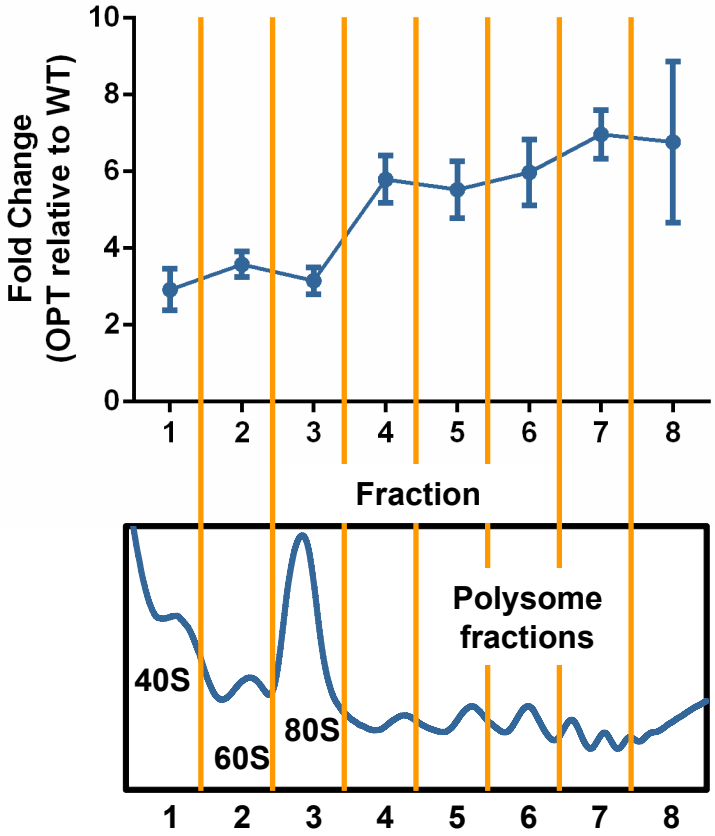
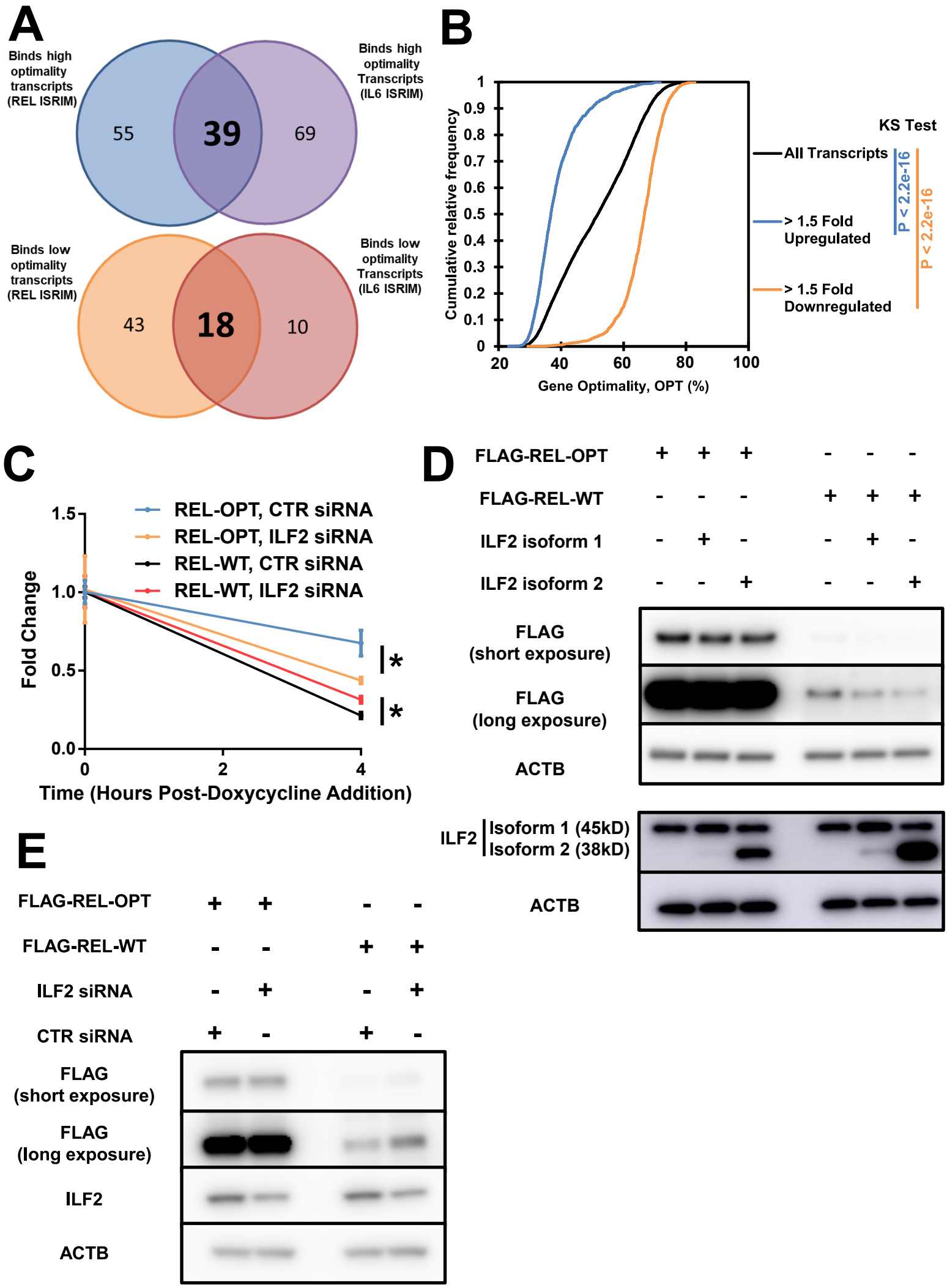

D
Fold Change (OPT relative to WT)
10
8
6
4
2
0
1
2
3
4
5
6
7
8
Fraction
Polysome fractions
40S
60S
80S

Figure 4


A
Binds high optimality transcripts (REL ISRIM)
Binds high optimality Transcripts (IL6 ISRIM)
55
39
69
Binds low optimality transcripts (REL ISRIM)
Binds low optimality Transcripts (IL6 ISRIM)
43
18
10
B
Cumulative relative frequency
Gene Optimality, OPT (%)
KS Test
All Transcripts
> 1.5 Fold Upregulated
> 1.5 Fold Downregulated
P < 2.2e-16
P < 2.2e-16
C
REL-OPT, CTR siRNA
REL-OPT, ILF2 siRNA
REL-WT, CTR siRNA
REL-WT, ILF2 siRNA
Fold Change
Time (Hours Post-Doxycycline Addition)
D
FLAG-REL-OPT
FLAG-REL-WT
ILF2 isoform 1
ILF2 isoform 2
FLAG (short exposure)
FLAG (long exposure)
ACTB
ILF2
Isoform 1 (45kD)
Isoform 2 (38kD)
ACTB
E
FLAG-REL-OPT
FLAG-REL-WT
ILF2 siRNA
CTR siRNA
FLAG (short exposure)
FLAG (long exposure)
ILF2
ACTB