Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans

Sarah E. Calvo[a,b,c,d,1], David J. Pagliarini[a,b,c,1], and Vamsi K. Mootha[a,b,c,2]

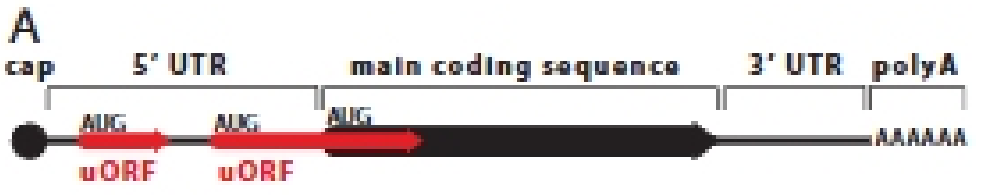
[a]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [b]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114; [c]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; and [d]Division of Health Sciences and Technology, Harvard–MIT, Cambridge, MA 02139

Edited by Jonathan Weissman, University of California, San Francisco, CA, and accepted by the Editorial Board March 18, 2009 (received for review October 29, 2008)

Upstream ORFs (uORFs) are mRNA elements defined by a start codon in the 5′ UTR that is out-of-frame with the main coding sequence. Although uORFs are present in approximately half of human and mouse transcripts, no study has investigated their global impact on protein expression. Here, we report that uORFs correlate with significantly reduced protein expression of the downstream ORF, based on analysis of 11,649 matched mRNA and protein measurements from 4 published mammalian studies. Using reporter constructs to test 25 selected uORFs, we estimate that uORFs typically reduce protein expression by 30–80%, with a modest impact on mRNA levels. We additionally identify polymorphisms that alter uORF presence in 509 human genes. Finally, we report that 5 uORF-altering mutations, detected within genes previously linked to human diseases, dramatically silence expression of the downstream protein. Together, our results suggest that uORFs influence the protein expression of thousands of mammalian genes and that variation in these elements can influence human phenotype and disease.

polymorphism | post-transcriptional control | proteomics | translation | uORF

The regulation of gene expression is controlled at many levels, including transcription, mRNA processing, protein translation, and protein turnover. Posttranscriptional regulation is often controlled by short sequence elements in the UTRs of mRNA. One such 5′ UTR element is the upstream ORF (uORF) depicted in Fig. 1*A*. Because eukaryotic ribosomes usually load on the 5′ cap of mRNA transcripts and scan for the presence of the first AUG start codon, uORFs can disrupt the efficient translation of the downstream coding sequence (1, 2). Previous reports have shown that ribosomes encountering a uORF can (*i*) translate the uORF and stall, triggering mRNA decay, (*ii*) translate the uORF and then, with some probability, reinitiate to translate the downstream ORF, or (*iii*) simply scan through the uORF (2). uORFs have been shown to reduce protein levels in ≈100 eukaryotic genes [supporting information (SI) Table S1]. Additionally, mutations that introduce or disrupt a uORF have found to cause 3 human diseases (3–5). In several interesting cases, the uORF-derived protein is functional; however, in most cases, the mere presence of the uORF is sufficient to reduce expression of the downstream ORF (1, 2, 6–8). Previous genomic analyses suggest that uORFs may be widely functional for several reasons: They correlate with lower mRNA expression levels (9), they are less common in 5′ UTRs than would be expected by chance (6, 10), they are more conserved than expected when present (6), and several hundred have evidence of translation in yeast (11). However, no study has demonstrated that these elements have a widespread impact on cellular protein levels. Moreover, no study has investigated whether uORF presence varies in the human population. Here, we take advantage of recently available datasets of protein abundance (12–17) and genetic variation (18, 19) to assess the impact and natural variation of mammalian uORFs.



B

# Transcripts with:	Human	Mouse
annotated 5′ UTR	23775	18663
≥1 uORF	11670	8253
≥2 uORFs	6268	4197
≥1 uORF fully upstream	9879	6935
≥1 uORF overlapping CDS	4275	2872
Median Length (nt):		
5′ UTR	170	139
uORF	48	48

Fig. 1. uORF definition and prevalence. (*A*) Schematic representation of mRNA transcript with 2 uORFs (red arrows), 1 fully upstream and 1 overlapping the main coding sequence (black arrow). uORFs are defined by a start codon (AUG) in the 5′ UTR, an in-frame stop codon (arrowhead) preceding the end of the main coding sequence, and length ≥9 nt. (*B*) Number and length of uORFs in human and mouse RefSeq transcripts.

Results

uORF Prevalence Within Mammalian Transcripts. We define a uORF as formed by a start codon within a 5′ UTR, an in-frame stop codon preceding the end of the main coding sequence (CDS), and length at least 9 nt including the stop codon. As shown in Fig. 1*A*, this definition includes uORFs both fully upstream and overlapping the CDS, because both types are predicted to be functional (20). We searched for uORFs within all human and mouse RefSeq transcripts with annotated 5′ UTRs >10 nt. Consistent with previous estimates (9, 10), we find that 49% of human and 44% of mouse transcripts contain at least 1 uORF (Fig. 1*B*). Interestingly, human and mouse uORF start codons (uAUGs) are the most conserved 5′ UTR trinucleotide across vertebrate species (Fig. S1), consistent with a widespread functional role.

uORF Impact on Cellular Protein Levels. If uORFs cause widespread reduction in protein expression, as predicted by ribosome scanning

Author contributions: S.E.C., D.J.P., and V.K.M. designed research; S.E.C. and D.J.P. performed research; and S.E.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.W. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]S.E.C. and D.J.P. contributed equally to this work.

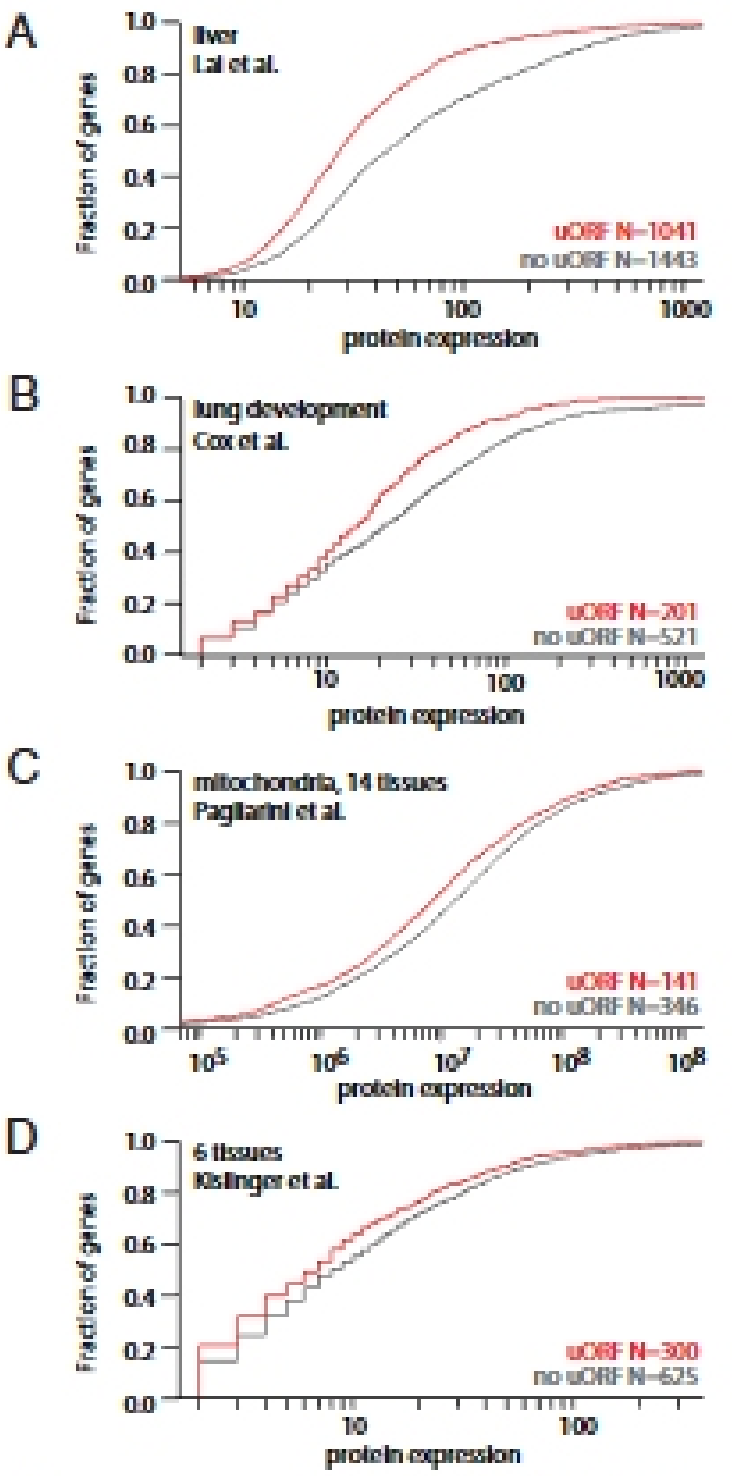
[2]To whom correspondence should be addressed at: Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street CPZN 5–806, Boston, MA 02114. E-mail: vamsi@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0810916106/DCSupplemental.

models, we would expect uORF-containing transcripts to correlate with lower protein levels when compared with uORF-less transcripts. To test this hypothesis, we analyzed a total of 11,649 matched mRNA and protein abundance measurements from 4 published studies across a variety of mouse tissues and developmental stages. These included: 2,484 genes expressed in liver (12), 722 genes expressed in 6 stages of lung development (13), 487 mitochondria-localized gene products expressed in 14 tissues (14), and 925 genes expressed in 6 tissues (15) (see *SI Text* for details). Proteins were detected via tandem mass spectrometry (MS/MS), and abundance was estimated by standard methods using the normalized number (12, 13, 15) or total peak area (14) of matching MS spectra. mRNA abundance in these conditions was measured by microarrays (21, 22). Although neither technology provides absolute quantitation, these large-scale datasets can reveal trends across thousands of genes. Because MS/MS technology cannot reliably distinguish splice variants, we analyzed expression at the gene level and considered only those genes whose collective splice variants either all contain, or all lack, uORFs. Consistent with previous reports (23), we observed that the 10% most highly expressed transcripts based on microarray tissue atlases (21) tend to lack uORFs (Fig. S2 and *SI Text*), and therefore, we conservatively excluded these genes to avoid overestimating uORF effects.

Despite differences in experimental methodology, all 4 independent datasets showed a reduced distribution of protein levels for genes containing versus lacking uORFs (Fig. 2 *A–D*). Median protein levels were reduced, respectively, by 39% ($P = 1e{-}5$), 29% ($P = 0.007$), 34% ($P = 0.008$), and 13% ($P = 0.36$), where significance was determined by empirical permutation testing. mRNA levels were reduced to a lesser extent with only the liver dataset (12) showing a statistically significant median reduction (Fig. 2*E* and Fig. S3). Importantly, the ratio of protein to mRNA was significantly reduced for uORF-containing genes in 3 of 4 datasets (Fig. 2*E* and Fig. S3), suggesting that uORF presence likely inhibits translation of the main coding sequence. We observed the same trends when we modified the definition of a uORF by altering length and overlap criteria, and when we included the 10% most highly expressed genes (Fig. S4). Analysis of 2 additional MS/MS studies of mouse adipocyte cells (16) and differentiating embryonic stem cells (17) also showed reduced protein levels for uORF-containing genes, although matched mRNA data were not available (Fig. S3). Collectively, these analyses across 3,297 mouse genes demonstrated the first large-scale correlation of uORF presence with reduced protein levels.

To determine whether uORFs play a causal role in reducing protein levels, and to more accurately quantify their effect size, we performed a series of experiments on 15 uORF-containing genes using dual-luciferase reporter constructs (see *Materials and Methods*). Five genes were chosen randomly from the set of all mouse transcripts containing single uORFs and where, for technical ease, 5′ UTR length exceeded 100 nt (Fig. 3 *B* and *F*). An additional 10 were selected from our mitochondrial study (14) where MS/MS and conservation data suggested functionality (Fig. 3 *C* and *G*). We cloned the 5′ UTR of each selected gene upstream of a luciferase reporter (Fig. 3*A*). HEK 293A cells were then transfected with uORF-containing luciferase constructs or control constructs where the uORF's start codon (ATG) was mutated to TTG. After 48 h, cells were assayed for luciferase transcript levels by quantitative PCR and for luciferase activity by luminometry. These experiments showed that, on average, uORFs cause a 58% decrease in protein levels (Fig. 3 *B* and *C*) and a 5% decrease in transcript levels (Fig. 3 *F* and *G*). All individual protein differences and 4 mRNA differences were statistically significant (Fig. 3), and all protein/mRNA ratio differences were statistically significant except for gene *Hsdl2* (Table S2). The constructs with randomly selected uORFs showed higher protein levels compared with the uORFs selected with evidence of functionality ($P = 1e{-}5$ based on t test). Similar results were obtained using HEK 293T cells. Together, the



E

median reduction of expression for genes with vs without uORFs

	protein	mRNA	protein/mRNA
liver	39% (1e-5)	18% (1e-5)	21% (1e-4)
lung dev.	29% (0.007)	17% (0.08)	9% (0.26)
mitochondria	34% (0.008)	8% (0.13)	30% (0.03)
6 tissues	12% (0.361)	-5% (0.63)	24% (0.03)

Fig. 2. Protein expression of uORF-containing genes. (*A–D*) Cumulative distribution of protein expression for mouse genes containing uORFs (red curve) or lacking uORFs (gray curve) in each of 4 independent MS/MS studies (12–15). N indicates the number of unique genes in each set. (*E*) Median reduction of protein and mRNA expression for genes containing uORFs compared with genes lacking uORFs, with *P* values (in parentheses) computed by empirical permutation testing.

large-scale correlations and validation experiments demonstrate that uORFs cause blunted protein expression of downstream coding sequences.

Influence of uORF Context, Position, and Conservation. We next investigated whether specific uORF properties were associated with stronger translational inhibition. We analyzed uORF length, number, conservation, position relative to the cap, position relative to the CDS, and uAUG context (also called "Kozak sequence") (see *Materials and Methods*). We quantified uORF effects using the Kolmogorov–Smirnow (KS) D statistic within the largest dataset (liver), which offered statistical power for these analyses. All tested subsets of uORFs showed reduced protein levels compared with uORF-less genes ($P < 0.05$), although certain properties modified

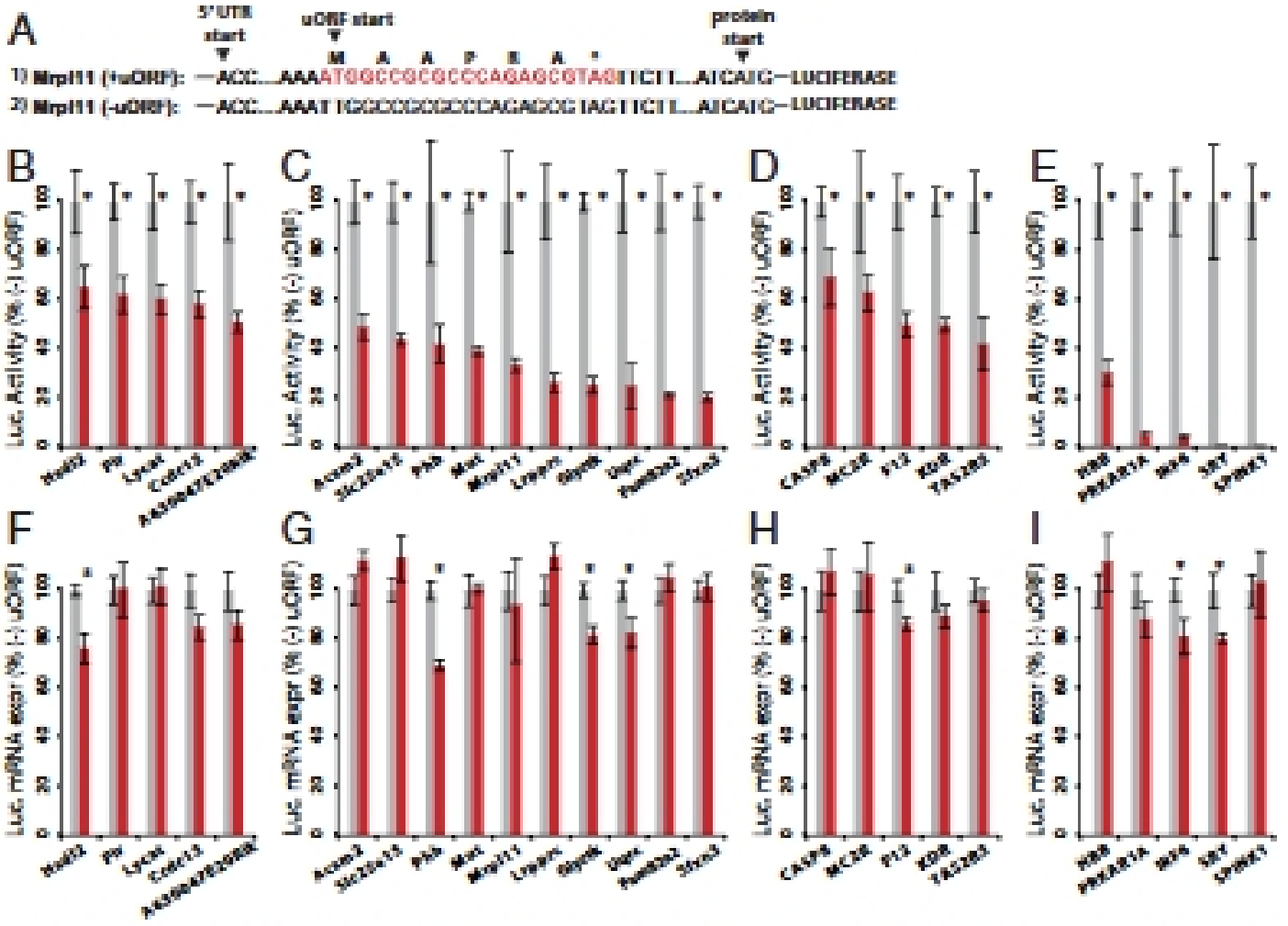


Fig. 3. Luciferase assays of uORF effects on protein and mRNA levels. (*A*) Experimental design of reporter constructs with and without uORFs is shown for example *Mrpl11*. (*B–I*) Normalized luciferase activity (*B–E*) and mRNA expression (*F–I*) are shown for reporter constructs that contain a uORF (red) or lack a uORF (gray) due to a mutation that disrupts the uORF start codon. The constructs contain 5′ UTRs from: 5 mouse genes chosen randomly (*B* and *F*), 10 mouse genes with proteomic and conservation signatures of functional uORFs (*C* and *G*), 5 human genes with polymorphic uORFs (*D* and *H*), and 5 human disease genes with uORF-altering mutations detected in patients (*E* and *I*). Error bars represent ±SE of ≥6 biological replicates (*B–E*) and ≥4 technical replicates (*F–I*). Asterisks indicate significant difference ($P < 0.01$).

the effect size (Fig. S5). As predicted by Kozak's classic experiments (1, 20, 24–26), increased inhibition correlated with strong versus weak uAUG context ($P = 0.04$), long versus short cap-to-uORF distance ($P = 0.009$ to 4e−4), presence of multiple uORFs in the 5′ UTR ($P = 8e{-}6$), and increased conservation ($P = 1e{-}6$) (Fig. S5). Surprisingly, we observed no significant difference between uORFs fully upstream versus overlapping the CDS ($P = 0.9$), between uORFs of different proximity to the CDS ($P = 0.6$ to 0.5) or between uORFs of different lengths ($P = 0.3$). These comparisons over hundreds of liver genes indicate that although all types of uORFs can reduce protein expression, 4 uORF properties are associated with greater inhibition: strong uAUG context, evolutionary conservation, increased distance from the cap, and multiple uORFs in the 5′ UTR.

Polymorphic uORFs in Humans. Given that uORFs reduce protein expression, polymorphisms that create or delete uORFs could influence human phenotypes. Therefore, we searched for uORF-altering variants within the 12 million SNPs in the human dbSNP database (18). We coin the term polymorphic uORF (puORF) to indicate a uORF that is created or deleted by a polymorphism. We identified puORFs in 509 unique genes (Table S3), of which 366 genes had multiple uORFs, and 143 genes had single uORFs (Table 1). Using the cellular reporter constructs described above, we tested the functionality of 5 puORFs. In all cases, the constructs with uORFs produced 30–60% less protein than those with the uORF-less SNP variant, with an average 3% decrease in mRNA levels (Fig. 3 *D* and *H*). All individual protein and protein/mRNA reductions were statistically significant (Table S2). The impact of the puORFs was comparable with all other uORFs that were tested experimentally (Fig. 3). Thus, naturally occurring uORF-altering polymorphisms are likely to alter cellular expression of the downstream protein.

puORF-Mediated Differences in Factor XII Protein Levels. One of the human uORF-altering SNPs (rs1801020) has previously been associated with differences in circulating plasma levels of clotting factor XII (*F12*) in 5 independent studies (27–31) (Fig. 4). This SNP represents a common T/C polymorphism with prevalence of the T allele estimated at 20% in Caucasian and 70% in Asian populations (27–31). Kanaji and colleagues demonstrated that the T allele reduces protein levels, and proposed that the mechanism could be due to disruption of the Kozak consensus sequence or to the introduction of a uORF, although these hypotheses were not tested (30). To experimentally test the uORF hypothesis, we created 8 reporter constructs that included all 4 possible nucleotide variants at the SNP site, 3 artificial uORF-generating mutations, and 1 mutation creating an alternate in-frame start site (Fig. 4*A*). All 4 uORF-containing UTR constructs showed >50% reduction in protein levels ($P < 2e{-}6$), whereas the 4 constructs lacking uORFs did not show strong differences in protein levels (Fig. 4*B*). mRNA levels were altered by <30% (Table S2). These results strongly suggest that the presence of a puORF is responsible for the observed variation in human factor XII protein levels.

uORF-Altering Mutations Related to Human Disease. In addition to common puORFs, rare mutations that alter uORFs may cause disease, as has been shown for 3 genes (Table 2). To systematically identify additional cases, we searched the Human Gene Mutation Database (19) for mutations that introduce or eliminate uORFs. We found 11 additional mutations (Table 2) that were detected by resequencing in known disease-related genes in affected patients (32–42). These uORF-altering mutations were not present in population controls (32–42), and were either the sole mutation detected in the sequenced exons, or were compound heterozygous with a missense/nonsense mutation (Table 2). The patient presentation was consistent with a recessive phenotype in 3 of the 4 compound heterozygous cases (37, 38, 42, 43), and was ambiguous