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.

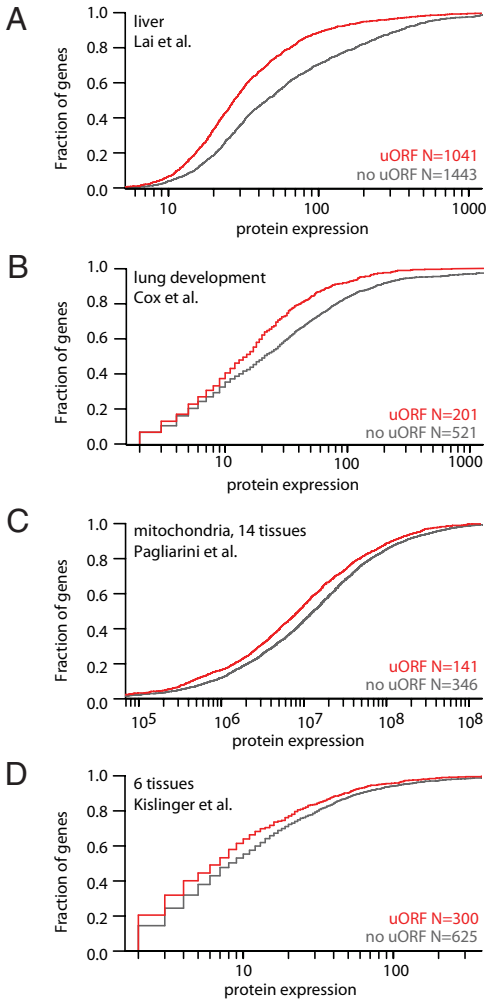
[2]To whom correspondence should be addressed at: Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street CPZN 5–806, Boston, MA 02114. E-mail: vamsi@hms.harvard.edu.

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



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% (0.12)	30% (0.03)
6 tissues	12% (0.361)	-3% (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.

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–Smirnov (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

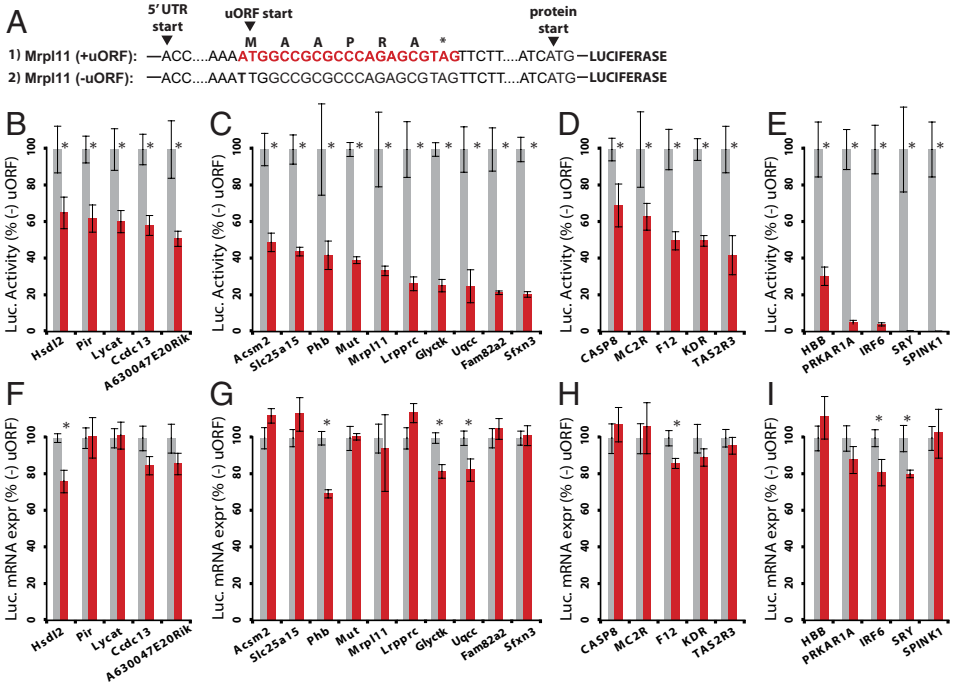


Fig. 3. Luciferase assays of uORF effects on protein and mRNA levels. (*A*) Experimental design of reporter constructs with and without uORFs is shown for example *Mrpl11*. (*B–I*) Normalized luciferase activity (*B–E*) and mRNA expression (*F–I*) are shown for reporter constructs that contain a uORF (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 (*FXII*) 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

Table 1. Notable human variants that create polymorphic uORFs

#	SNP	AvHet, %	Gene	Gene description
1	rs1801020	50	*F12*	Coagulation factor XII (Hageman factor)
2	rs12272467	50	*TRIM6*	Tripartite motif-containing 6
3	rs1108842	50	*GNL3*	Guanine nucleotide-binding protein-like 3 (nucleolar)
4	rs6460054	50	*CLDN3*	Claudin 3
5	rs1046188	50	*SCAMP3*	Secretory carrier membrane protein 3
6	rs13104310	49	*C4orf21*	Chromosome 4 open reading frame 21
7	rs7667298	49	*KDR*	Kinase insert domain receptor
8	rs7331765	49	*RASL11A*	RAS-like, family 11, member A
9	rs2001216	49	*RCCD1*	RCC1 domain containing 1
10	rs12975585	48	*HNRNPUL1*	Heterogeneous nuclear ribonucleoprotein U-like 1
11	rs2838343	46	*HSF2BP*	Heat shock transcription factor 2-binding protein
12	rs765007	46	*TAS2R3*	Taste receptor, type 2, member 3
13	rs17499247	45	*CREM*	cAMP responsive element modulator
14	rs1048371	42	*MUCL1*	Mucin-like 1
15	rs1800070	*	*CFTR*	Cystic fibrosis transmembrane conductance regulator
16	rs34704828	*	*HBB*	Hemoglobin, β
17	rs28926176	0.2	*MC2R*	Melanocortin 2 (ACTH hormone) receptor
18	rs41409645	4	*CCL3*	Chemokine (C-C motif) ligand 3
19	rs2856759	*	*CCR5*	Chemokine (C-C motif) receptor 5
20	rs34819868	*	*HAVCR1*	Hepatitis A virus cellular receptor 1
21	rs41275166	*	*CD59*	CD59 molecule, complement regulatory protein
22	rs6057688	*	*DEFB119*	Defensin, β119
23	rs2234011	*	*TAS2R5*	Taste receptor, type 2, member 5
24	rs1091826	*	*OXTR*	Oxytocin receptor
25	rs6781226	*	*HTR1F*	5-hydroxytryptamine (serotonin) receptor 1F

List contains common SNP variants (#1–14) and genes associated with monogenic disease (#15–17), immune response (#18–22), and receptor activity (#23–25 and 7, 12, 17, 19, 20). Table S3 contains a complete list. AvHet indicates SNP's average heterozygosity.
*Data not available.

in the remaining case (36). To our knowledge, the mechanistic link between the gene mutation and uORFs had not been previously proposed for SRY (32), IRF6 (33), or GCH1 (34).

To assess whether the uORF-altering mutations influenced protein expression, we used luciferase reporter constructs to test patient mutations in 5 genes (*HBB*, *PRKAR1A*, *IRF6*, *SRY*, and *SPINK1*). The uORF-altering mutations in these genes reduced luciferase mRNA levels by <20% and luciferase activity levels by 70–100% (Fig. 3 *E* and *I*). These effects on protein levels were highly significant ($P < 2e{-}12$) and were larger than in the other uORFs experimentally tested ($P = 4e{-}4$). Thus, these uORF-altering mutations cause dramatically reduced protein levels in our reporter assays, suggesting that they may indeed be responsible for the observed disease phenotypes.

Discussion

Our analyses provide an assessment of the widespread impact of uORFs on mammalian protein expression. Many previous studies of individual genes demonstrated that the presence of uORFs can lead to reduced mRNA stability and protein translation. Here, we show that approximately half of human and mouse protein-encoding genes contain uORFs and that uORF presence correlates with reduced protein expression across thousands of mammalian genes in a variety of tissues and conditions (Fig. 2). We quantify uORF effects using mutation experiments on 25 selected 5′ UTRs (Fig. 3), which have typical length, context, position, and conservation features (Fig. S6). These experiments indicate that uORFs typically affect mRNA levels by <30% and reduce protein levels by 30–80%, although complete protein suppression is possible (Fig. 3). Although our mutation experiments focused chiefly on 5′ UTRs containing single uORFs, our MS/MS data suggest that multiple uORFs lead to greater reduction of protein expression (Fig. S5*E*). Collectively, these data suggest that uORFs cause reduced protein levels of thousands of mammalian genes.

Our data provide insight into the mechanism by which uORFs influence protein expression. Without exception, uORF-containing

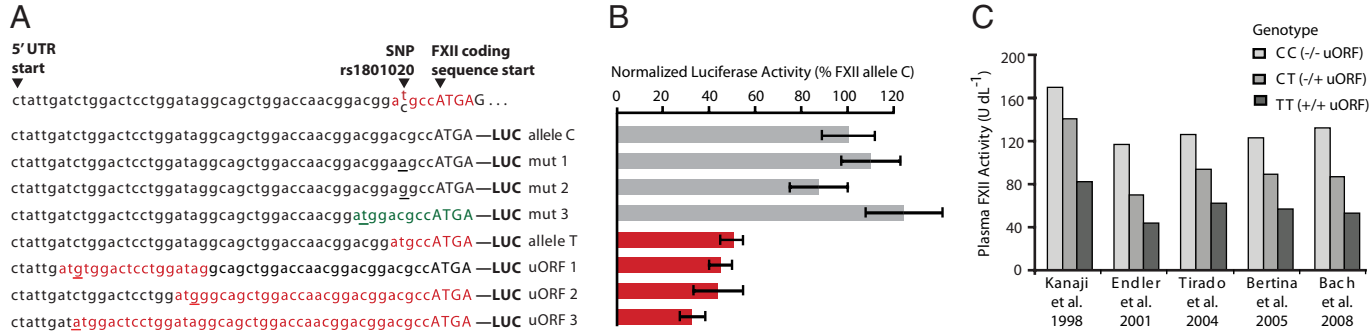


Fig. 4. Polymorphic uORF alters FXII protein expression. (*A*) 5′ UTR sequence of FXII shown with 2 SNP variants, where the T allele creates a uORF (red text). Below are 8 constructs with introduced mutations (underlined text), where colored text indicates a uORF (red) or in-frame alternative start (green). (*B*) Luciferase activity from reporter constructs listed in *A*. Error bars represent ±SD of ≥6 biological replicates. (*C*) Metaanalysis of plasma FXII activity levels measured by 5 independent studies, stratified by genotype of SNP rs1801020.

Table 2. uORF-altering mutations linked to disease

#	Gene	Disease	Mutation	uORF link
1	*THPO*	Thrombocythemia	splice site (3)	Known
2	*CDKN2A*	Melanoma	G-34T (4)	Known
3	*HR*	Marie Unna hereditary hypotrichosis	A-321G (5)	Known
4	*SRY*	Gonadal dysgenesis	G-75A (32)	Novel*
5	*IRF6*	Van der Woude syndrome	A-48T (33)	Novel*
6	*GCH1*	DOPA-responsive dystonia	C-22T (34)	Novel
7	*HAMP*	Juvenile hemochromatosis	G-25A (35)	Predicted
8	*KCNJ11*	Hyperinsulinemic hypoglycemia, 2	C-54T (36)†	Predicted
9	*LDLR*	Familial hypercholesterolemia	delC-22 (37)†	Predicted
10	*PEX7*	Rhizomelic chondrodysplasia punctata	C-45T (38)†	Predicted
11	*POMC*	Proopiomelanocortin deficiency	C-11A (39)	Predicted
12	*PRKAR1A*	Carney complex type 1	G-97A (40)	Predicted*
13	*SPINK1*	Hereditary pancreatitis	C-53T (41)	Predicted*
14	*HBB*	Thalassaemia β	G-29A (42)†	Predicted*

uORF-altering mutations detected in patients but not population controls. Mutation column includes 5′ UTR position relative to translation start and literature reference (in parentheses). The links between mutations and uORFs were previously known, previously predicted, or not previously known (novel).
*Mutations tested experimentally in this study.
†Compound heterozygous mutations.

reporter constructs exhibit more pronounced reduction of protein compared with mRNA levels (Fig. 3), in agreement with the trend observed in large-scale datasets (Fig. 2*E*). This suggests that uORFs act primarily by reducing translational efficiency, and more modestly by affecting mRNA levels. Additionally, because uORF effects do not correlate with the distance between the uORF and CDS (Fig. S5*D*), it is likely that CDS translation generally proceeds from ribosomes that scan through the uORF rather than from ribosomes that reinitiate after uORF translation—at least in genes that contain only a single uORF.

Given that uORFs reduce translation, variants that create or delete uORFs are likely to alter cellular protein levels and in some cases may influence phenotype. uORF-altering variants are likely to be widespread, because each human transcript contains, on average, 28 nt that could be mutated to introduce a uORF. We identified 509 human genes with polymorphic uORFs (puORFs), although more are likely to be identified as genome variation databases expand. Our data suggest that puORFs will typically alter cellular protein levels by 30–80% in cases where the 5′ UTR contains a single uORF. When these puORFs cause physiologically relevant changes in protein levels, as we showed for factor XII, they may cause phenotypic variation. Indeed, the factor XII puORF has been associated with several thromboembolic conditions, although the associations are in contention due to small sample sizes (44). We speculate that other puORFs in our collection (Table S3) may also affect phenotype. For instance, the puORF in chemokine receptor CCR5 might mediate susceptibility to HIV-1 infection, because previous studies showed that variants affecting CCR5 expression alter susceptibility to HIV-1 infection and progression of AIDS (45). Similarly, the puORFs in bitter-taste receptors TAS2R5 and TAS2R3 might lead to common variation in taste perception, and puORFs in receptors for ACTH, serotonin, and oxytocin may modulate neurohormonal response (Table 1).

In addition to common polymorphisms, rare uORF-altering mutations that alter levels of essential proteins can cause human disease. To date, 3 such mutations have been reported. First, a hereditary form of thrombocythaemia is caused by a mutation in THPO mRNA that eliminates a uORF through a splicing defect, and thus causes increased translation of thrombopoetin (3). Second, a mutation introducing a uORF into CDKN2A causes a familial predisposition to melanoma (4). Third, disruption of uORF presence and coding sequence in gene *HR* causes Marie Unna hereditary hypotrichosis (5). Additional uORF-altering mutations detected in patients with 11 diseases have been reported in the literature, although they were not followed up experimentally (Table 2). In each case, the patient mutation was present within a gene known to underlie the disease when disrupted and was the sole mutation detected or was compound heterozygous with a nonsynonymous variant. Using reporter assays, we tested 5 patient mutations in genes associated with disease: Gonadal dysgenesis (*SRY*), Van der Woude syndrome (*IRF6*), Carney complex type 1 (*PRKAR1A*), Hereditary pancreatitis (*SPINK1*), and Thalassaemia-β (*HBB*). We found that the uORF-altering patient mutation caused severely reduced protein levels, and in 2 cases almost no reporter protein was detected (*SRY* and *SPINK1*, Fig. 3*E*). In these 2 cases, the patient mutation created a second uORF within the gene 5′ UTR, rather than creating a single uORF. The strong suppression of protein expression by these 5 patient mutations offers a simple mechanistic basis for their pathogenicity. These cases add to the growing list of uORF-altering mutations linked to disease and highlight the importance of searching for uORF changes in addition to coding changes underlying disease.

In summary, our analyses demonstrate that uORFs have a widespread impact on the expression of human and mouse genes and that the human genome contains hundreds of polymorphic uORFs. With the routine application of newer generation sequencing technologies, an important challenge will be to link variation in genome sequences to physiology and disease—and puORFs may represent an important class of functional variants that can be readily linked to phenotype. Although the current analyses focused on the constitutive effects of uORFs on steady-state protein levels, an important next step is to determine whether the influence of uORFs is widely regulated by environmental conditions or signaling pathways, as has been shown for a handful of examples (2).

Materials and Methods

Human and Mouse uORFs. RefSeq transcripts for human (hg18) and mouse (mm9) were obtained from the University of California, Santa Cruz (UCSC) Genome Browser Database (46) (accessed May 20, 2008), along with 28-vertebrate species alignments (47). Custom Perl scripts annotated uORFs and computed features: uORF context (“strong” indicates a −3 purine and +4 guanine relative to uAUG, otherwise “weak”), cap-to-uORF distance (length between mRNA cap and uAUG), uORF length (including start and stop codon), uORF-to-CDS distance (length between uORF stop codon and CDS start codon), uORF number (number of distinct uORFs in a transcript, where uORFs may overlap but not in the same frame), and conservation (number of species with aligned start codons within 28 species alignments). The first 4 features were analyzed on transcripts containing single uORFs. uORF properties were compared using a Bonferroni-corrected,

1-sided KS test. 5′ UTR trinucleotide conservation was scored by number of identities in 28 species alignments.

Matched mRNA and Protein Datasets. MS/MS protein abundance measurements were obtained from published studies (12–17). Matched mRNA data were available in 3 studies (13–15). For the liver study (12), we used mean mRNA expression from GNF1M liver replicates (21). All data were mapped to Entrez Gene identifiers with the gene inheriting the highest score from any splice form. We excluded genes with poorly quantified mRNA values (expression values <40) and the top 10% most highly expressed genes, based on mean mRNA expression values from the GNF1M atlas. We analyzed mouse genes with annotated 5′ UTRs (>10nt), where all splice forms contained a uORF (6,933 genes) or lacked a uORF (9,343 genes). Differences in median protein expression were measured as percentage reduction from uORF-less genes, using 10,000 permutations of gene uORF labels to assess significance. See *SI Text* for details.

Luciferase Assays. UTR sequences, up to and including the primary ATG initiation codon, were synthesized (IDT), cloned and ligated into the NheI site directly preceding the *Renilla* luciferase gene in the dual-luciferase vector psiCHECK-2 (Promega) (Table S4). Before cloning, the ATG of the *Renilla* luciferase was mutated to TTG so that the *Renilla* luciferase expression would be driven by the primary ATG initiation codon of the gene under investigation. HEK 293A cells were seeded at 6,000 cells per well in 96-well opaque white cell culture plates (Nunc). After overnight incubation, cells were transfected with 20–100 ng of each construct by using Fugene 6 (Roche). Forty-eight hours after transfection, cells were washed with PBS and lysed with Passive Lysis Buffer (Promega). *Renilla* and Firefly luciferase signals were generated by using Promega's Dual-Luciferase Assay System according to the manufacturer's protocol. For each construct, *Renilla* luciferase signal was normalized to the Firefly luciferase internal control signal. Plates were read by using a Victor^3 plate reader (PerkinElmer) and the data analyzed by using Wallac 1420 Workstation software. Two-sided, homoscedastic *t* tests assessed significance.

Real-Time PCR. HEK 293A cells were seeded at 2×10^5 cells per well in 6-well cell culture plates 24 h before transfection. One microgram of each construct was transfected per well by using Fugene 6 as above. Forty-eight hours after transfection, cells were washed with PBS, and RNA was harvested by using a Qiagen RNAeasy kit. First-strand cDNA synthesis was performed by using SuperScript III (Invitrogen) using 1 μg of RNA from each transfection as starting material. Real-time PCR was performed by using custom TaqMan Assays (ABI) designed against *Renilla* luciferase (target) and Firefly luciferase (endogenous control). Two-sided, homoscedastic *t* tests assessed significance.

uORF-Altering Variants. Human dbSNP version 128 (18) was obtained from UCSC (46) and filtered for SNPs (class "single") that mapped to single locations within hg18 and overlapped annotated RefSeq 5′ UTRs, excluding those that overlapped RefSeq CDSs. Custom perl scripts mapped SNPs onto mRNA transcripts and determined those that altered uORF presence. The Human Gene Mutation Database professional release 2008.2 (19), was searched for all noncoding substitutions or microlesions that altered presence of ATG codons and that overlapped 5′ UTRs based on manual inspection of BLAT alignments to hg18.

ACKNOWLEDGMENTS. We thank S. E. Ong for advice on mass spectrometry; M. Garber for advice on alignments; O. Goldberger and Y. Kim for technical assistance; J. Dixon for technical resources; and D. Altshuler, J. Hirshhorn, M. Springer, D. Neafsey, B. Voight, S. Carter, J. Avruch, and E. Lander for comments on the manuscript and project. We thank the Broad Institute Sequencing Platform, Washington University Genome Sequencing Center (St. Louis, MO), and Baylor Human Genome Sequencing Center (Houston, TX) for the vertebrate genome sequences used in comparative analyses. This work was supported by National Institute of General Medical Science Grant GM077465 (to V.K.M.).

1. Kozak M (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 266:19867–19870.
2. Morris DR, Geballe AP (2000) Upstream open reading frames as regulators of mRNA translation. *Mol Cell Biol* 20:8635–8642.
3. Wiestner A, Schlemper RJ, van der Maas AP, Skoda RC (1998) An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat Genet* 18:49–52.
4. Liu L, et al. (1999) Mutation of the CDKN2A 5′ UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat Genet* 21:128–132.
5. Wen Y, et al. (2009) Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat Genet* 41:228–233.
6. Neafsey DE, Galagan JE (2007) Dual modes of natural selection on upstream open reading frames. *Mol Biol Evol* 24:1744–1751.
7. Meijer HA, Thomas AA (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochem J* 367:1–11.
8. Vilela C, McCarthy JE (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. *Mol Microbiol* 49:859–867.
9. Matsui M, Yachie N, Okada Y, Saito R, Tomita M (2007) Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett* 581:4184–4188.
10. Iacono M, Mignone F, Pesole G (2005) uAUG and uORFs in human and rodent 5′ untranslated mRNAs. *Gene* 349:97–105.
11. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, in press.
12. Lai KK, Kolippakkam D, Beretta L (2008) Comprehensive and quantitative proteome profiling of the mouse liver and plasma. *Hepatology* 47:1043–1051.
13. Cox B, et al. (2007) Integrated proteomic and transcriptomic profiling of mouse lung development and Nmyc target genes. *Mol Syst Biol* 3:109.
14. Pagliarini DJ, et al. (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 134:112–123.
15. Kislinger T, et al. (2006) Global survey of organ and organelle protein expression in mouse: Combined proteomic and transcriptomic profiling. *Cell* 125:173–186.
16. Adachi J, Kumar C, Zhang Y, Mann M (2007) In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol Cell Proteomics* 6:1257–1273.
17. Williamson AJ, et al. (2008) Quantitative proteomics analysis demonstrates post-transcriptional regulation of embryonic stem cell differentiation to hematopoiesis. *Mol Cell Proteomics* 7:459–472.
18. Sherry ST, et al. (2001) dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* 29:308–311.
19. Stenson PD, et al. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21:577–581.
20. Kozak M (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Mol Cell Biol* 7:3438–3445.
21. Su AI, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101:6062–6067.
22. Mariani TJ, Reed JJ, Shapiro SD (2002) Expression profiling of the developing mouse lung: Insights into the establishment of the extracellular matrix. *Am J Resp Cell Mol Biol* 26:541–548.
23. Kochetov AV, et al. (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett* 440:351–355.
24. Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44:283–292.
25. Kozak M (1991) An analysis of vertebrate mRNA sequences: Intimations of translational control. *J Cell Biol* 115:887–903.
26. Kozak M (2001) Constraints on reinitiation of translation in mammals. *Nucleic Acids Res* 29:5226–5232.
27. Bach J, et al. (2008) Coagulation factor XII (FXII) activity, activated FXII, distribution of FXII C46T gene polymorphism and coronary risk. *J Thromb Haemost* 6:291–296.
28. Bertina RM, Poort SR, Vos HL, Rosendaal FR (2005) The 46C→T polymorphism in the factor XII gene (F12) and the risk of venous thrombosis. *J Thromb Haemost* 3:597–599.
29. Endler G, et al. (2001) A common C→T polymorphism at nt 46 in the promoter region of coagulation factor XII is associated with decreased factor XII activity. *Thromb Res* 101:255–260.
30. Kanaji T, et al. (1998) A common genetic polymorphism (46 C to T substitution) in the 5′-untranslated region of the coagulation factor XII gene is associated with low translation efficiency and decrease in plasma factor XII level. *Blood* 91:2010–2014.
31. Tirado I, et al. (2004) Association after linkage analysis indicates that homozygosity for the 46C→T polymorphism in the F12 gene is a genetic risk factor for venous thrombosis. *Thromb Haemost* 91:899–904.
32. Poulat F, et al. (1998) Mutation in the 5′ noncoding region of the SRY gene in an XY sex-reversed patient. *Hum Mutat* Suppl 1:S192–S194.
33. Kondo S, et al. (2002) Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 32:285–289.
34. Tassin J, et al. (2000) Levodopa-responsive dystonia. GTP cyclohydrolase I or parkin mutations? *Brain* 123 (Pt 6):1112–1121.
35. Matthes T, et al. (2004) Severe hemochromatosis in a Portuguese family associated with a new mutation in the 5′-UTR of the HAMP gene. *Blood* 104:2181–2183.
36. Huopio H, et al. (2002) Acute insulin response tests for the differential diagnosis of congenital hyperinsulinism. *J Clin Endocrinol Metab* 87:4502–4507.
37. Sozen MM, et al. (2005) The molecular basis of familial hypercholesterolaemia in Turkish patients. *Atherosclerosis* 180:63–71.
38. Braverman N, et al. (2002) Mutation analysis of PEX7 in 60 probands with rhizomelic chondrodysplasia punctata and functional correlations of genotype with phenotype. *Hum Mutat* 20:284–297.
39. Krude H, et al. (1998) Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nat Genet* 19:155–157.
40. Groussin L, et al. (2002) Molecular analysis of the cyclic AMP-dependent protein kinase A (PKA) regulatory subunit 1A (PRKAR1A) gene in patients with Carney complex and primary pigmented nodular adrenocortical disease (PPNAD) reveals novel mutations and clues for pathophysiology: Augmented PKA signaling is associated with adrenal tumorigenesis in PPNAD. *Am J Hum Genet* 71:1433–1442.
41. Witt H, et al. (2000) Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. *Nat Genet* 25:213–216.
42. Oner R, et al. (1991) The G–A mutation at position +22 3′ to the Cap site of the beta-globin gene as a possible cause for a beta-thalassemia. *Hemoglobin* 15:67–76.
43. Cai SP, et al. (1992) Two novel beta-thalassemia mutations in the 5′ and 3′ noncoding regions of the beta-globin gene. *Blood* 79:1342–1346.
44. Bersano A, Ballabio E, Bresolin N, Candelise L (2008) Genetic polymorphisms for the study of multifactorial stroke. *Hum Mutat* 29:776–795.
45. Navratilova Z (2006) Polymorphisms in CCL2&CCL5 chemokines/chemokine receptors genes and their association with diseases. *Biomed Papers Med Faculty Univ Palacky, Olomouc, Czechoslovakia* 150:191–204.
46. Karolchik D, et al. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31:51–54.
47. Miller W, et al. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* 17:1797–1808.