Supplemental Methods:

Supplemental Table 1: Summary of methodologies in ribosome profiling studies:

Study	***Cell line***	***Treatment protocol***	***TIS identification***
Fritsch et al. 2012	THP1	cyclohexamide + puromycin	Novalign + neural network
Lee et al. 2012	HEK293	lactidomycin	Bowtie + threshold on nucleotide resolution read counts
Gao et al. 2014	HEK293	cyclohexamide + lactimidomycin + puromycin	TopHat + ZTNB model

Attributes for uORF classification:

uORF length:
Number of bases in exonic segments of the uORF.
Percentage content of each base (%A, %T, %G, %C):
Number of bases of each variety in the uORF, divided by the total number of bases in the uORF.
Distance between the CDS start codon and the uORF start codon:
D*istance, in number of bases.*
Distance between the CDS start codon and the uORF stop codon:
D*istance, in number of bases.*
Number of rare codons within the uORF (TCG, CGT, ACG, CGA, CTA, GTA):
The number of each of these codons present within the uORF, according to reading frame of the uORF. These are the 6 rarest codons, as defined by their frequency in the human genome. Global frequency data comes from GenScript USA, Inc. http://www.genscript.com/cgi-bin/tools/codon_freq_table
 # Non-cognate start codons :
The number of each non-cognate start codon in the uORF was calculated (10 start codons, giving ten separate features.
Number of ATG/CTG:
Number of uORFs beginning with the canonical start codon ATG, CTG, or both combined.
Number of uORFs, beginning with the same start codon:
Number of uORFs, associated with the same Ensembl transcript ID, that begin with the same start codon.
Optimal DeltaG observed upon folding for the uORF, by segment:
DeltaG is the free energy change expected, when the identified mRNA segment is folded into an optimal secondary structure. This energy change was determined using the Vienna package RNAfold [1]*. Each segement analyzed, is referenced in bases from the uORF start codon: 0-39, 20-59, 40-79, 60-99, 80-119, 100-139; bases from the uORF stop codon: -40- -1, -20-19, 0-39, 20-59, 40-79; or bases from the CDS start codon: -20-19, 0-39, 20-59, 40-79, 60-99, 80-119, 100-139).*
GERP score for the start codon, GERP score for the stop codon, GERP elements percent:

GERP score was used as an assessment of conservation across multiple mammalian species, in a position specific manner [2]*. GERP score was calculated both for the start codon, and the stop codon. The percent overlap between strongly conserved GERP elements and the uORF, was also calculated.*

Length of the 5'UTR:

Length, in n*umber of bases.*

Distance between the 7-methylguanylate cap, and the uORF start codon:

Distance, in number of bases.

Kozak context:

Defined as 'strong' (score = 2) if there is a G immediately following the start codon in the CDS [position 3], and either an A or a G three codons upstream of the start codon [position -3]; defined as 'weak' (score = 1) if there is either a G immediately following the start codon in the CDS [position 3], or either an A or a G three codons upstream of the start codon [position -3]; defined as 'absent' (score = 0) otherwise.

GTEX mRNA expression:

Gene level mRNA expression, according to tissue of origin, was obtained from the GTEX project [3]*. Expression data was clustered according to the following 31 tissue types:*

- *Thyroid*
- *Testis*
- *Cervix Uteri*
- *Adipose Tissue*
- *Breast*
- *Vagina*
- *Ovary*
- *Stomach*
- *Fallopian Tube*
- *Bone Marrow*
- *Spleen Bladder*
- *Blood*
- *Colon*
- *Prostate*
- *Pancreas*
- *Blood Vessel*
- *Liver*
- *Heart*
- *Small*
- *Intestine*
- *Uterus*
- *Pituitary*
- *Muscle*
- *Nerve*
- *Adrenal Gland*
- *Brain*
- *Salivary Gland*
- *Lung*

- *Skin*
- *Esophagus*
- *Kidney*

The mean expression level was calculated across each tissue type. The mRNA expression entropy [4], between tissues, was calculated according to the following formula:

$$H_g = \sum_{t=1}^{N} -p_{tvg} log_2(-p_{tvg})$$

Where H_g is the tissue distribution entropy for gene g, N is the number of tissues, t is a tissue. p_{tvg} is the relative expression of gene g in tissue t, according to the following formula:

$$p_{tvg} = \frac{w_{g,t}}{\sum_{t=1}^{N} w_{g,t}}$$

With $w_{g,t}$ the expression level recorded for gene g in tissue t.

Germline variation – # SNPs, SNPs/length:
The 1000 Genomes annotation indicates the location and frequency of known single nucleotide polymorphisms (SNPs). The intersection of these mutations with uORFs exons was completed using BEDtools. The number of germline variants was also normalized against the uORF length.
Noderer Context:
*Analagous to the Kozak consensus sequence surrounding the ATG start codon, Noderer et al. build a translation efficiency measure, related to the codons surrounding the translation initiation site of AUG codons: UUCAUC**AUG**CA [5]. For each uORF, a translational efficiency score was assigned, according to the codons surround the AUG start codon. This assignment was made according to Supplementary Table 2. of Noderer et al. 2014.*
Heterozygosity Measurement:
Also from the 1000 Genomes annotation, the frequency of an allele may be recovered, in the form AF = AC/AN where AF is the frequency of the allele, AC is the number of times the allele is represented in the population, and AN is the number of individuals in the population.

With knowledge of the frequency of alleles within a uORF, the heterozygosity may be calculated according to:

$$\sum_{i=1}^{n} AF^2 + (1 - AF)^2$$

Where *n* is the total number of mutations within the protein coding regions of a uORF.

Supplemental Table 2: uORF features. Features are listed according to the KS statistic for each attribute, measured between positive and unlabeled uORFs.

Rank	*Attribute*	*KS statistic*	*p value*	*Rank*	*Attribute*	*KS statistic*	*p value*

1	GTEX Bone Marrow	0.54	0.000	46	#AGG	0.20	0.000
2	GTEX Liver	0.50	0.000	47	#CTG	0.20	0.000
3	GTEX Lung	0.49	0.000	48	Kozak context	0.19	0.000
4	GTEX Pituitary	0.49	0.000	49	% GERP elements	0.19	0.000
5	Ribosome profiling uORF start codon frequency	0.48	0.000	50	uORF start codon to CDS start codon distance	0.19	0.000
6	GTEX Nerve	0.48	0.000	51	mRNA ΔG uORF start [20-59]BP	0.18	0.000
7	GTEX Muscle	0.47	0.000	52	#GTG	0.18	0.000
8	GTEX Pancreas	0.47	0.000	53	mRNA ΔG uORF start [40-79]BP	0.18	0.000
9	GTEX Adipose Tissue	0.47	0.000	54	%A	0.17	0.000
10	GTEX Skin	0.47	0.000	55	5' cap to uORF start codon distance	0.16	0.000
11	GTEX Spleen	0.47	0.000	56	mRNA ΔG uORF stop codon [0,39]BP	0.16	0.000
12	GTEX Stomach	0.46	0.000	57	mRNA ΔG uORF stop codon [-20,19]BP	0.16	0.000
13	GTEX Cervix Uteri	0.46	0.000	58	uORF stop codon to CDS start codon distance	0.16	0.000
14	***GTEX (combined)***	***0.46***	***0.000***	59	mRNA ΔG CDS start [-20,19]BP	0.14	0.000
15	GTEX Salivary Gland	0.46	0.000	60	Noderer context	0.13	0.000
16	GTEX Uterus	0.46	0.000	61	%G	0.13	0.000
17	GTEX Small	0.46	0.000	62	mRNA ΔG uORF start	0.12	0.000

	Intestine				[60,99]BP		
18	GTEX Prostate	0.46	0.000	63	mRNA ΔG uORF start [80,119]BP	0.12	0.000
19	GTEX Esophagus	0.46	0.000	64	mRNA ΔG uORF start [-20,19]BP	0.12	0.000
20	GTEX Heart	0.46	0.000	65	mRNA ΔG uORF end [-40,-1]	0.11	0.000
21	GTEX Bladder	0.46	0.000	66	mRNA ΔG uORF start [100,139]	0.11	0.000
22	GTEX Brain	0.45	0.000	67	mRNA ΔG CDS start [20,59]	0.11	0.000
23	GTEX Breast	0.45	0.000	68	mRNA ΔG uORF start [0,39]	0.10	0.000
24	GTEX Blood Vessel	0.45	0.000	69	%C	0.10	0.000
25	GTEX Fallopian Tube	0.45	0.000	70	mRNA ΔG uORF end [20,59]	0.09	0.000
26	GTEX Blood	0.45	0.000	71	mRNA ΔG CDS start [40,79]	0.09	0.000
27	GTEX Thyroid	0.44	0.000	72	mRNA ΔG uORF end [40,79]	0.08	0.000
28	GTEX Vagina	0.44	0.000	73	SNPs/length	0.07	0.001
29	GTEX Colon	0.44	0.000	74	mRNA ΔG CDS start [0,39]	0.06	0.004
30	GTEX Kidney	0.43	0.000	75	#TCG	0.06	0.011
31	GTEX Testis	0.43	0.000	76	mRNA ΔG CDS start 100.139	0.05	0.027
32	GTEX Adrenal Gland	0.42	0.000	77	mRNA ΔG uORF start [80,119]	0.05	0.028
33	GTEX Ovary	0.41	0.000	78	%T	0.05	0.053
34	GTEX Tissue Entropy	0.40	0.000	79	#ACG	0.05	0.056

35	#Same start codon	0.30	0.000	80	#CGA	0.04	0.142
36	#ATG	0.28	0.000	81	#CGT	0.04	0.198
37	#ATA	0.28	0.000	82	mRNA ΔG CDS start [60,99]	0.03	0.309
38	#ATT	0.26	0.000	83	uORF length (BP)	0.03	0.447
39	#ATG + CTG	0.26	0.000	84-89	#ACG		
40	#AAG	0.23	0.000	84-89	#CTA		
41	#ATC	0.22	0.000	84-89	#GTA		
42	Size 5'UTR (%)	0.22	0.000	84-89	Heterozygosity/length		
43	Start codon GERP score	0.22	0.000	84-89	#1000 Genomes SNPs		
44	Stop codon GERP score	0.21	0.000	84-89	Heterozygosity		
45	#TTG	0.21	0.000				

Supplemental Results:

Case study of uORF gain associated with gene EIF5A:

EIF5A has 3 predicted positive uORFs with ATG start codons (GRCh37; ENST00000336458.8). Variant rs9905259, with minor allele frequency 0.05, introduces an additional predicted positive ATG uORF downstream of all other ATG uORFs [Figure S.1]. The overall effect of this additional uORF is an increase in downstream EIF5A protein level ($N_{reference}$=25, μ=-0.241764318; $N_{rs9905259}$=21, μ=-0.17162073; p= 0.044)

It is possible that uORFs upstream of this gained uORF repress this uORF. Furthermore, in addition to introducing an uORF, variant rs9905259 intersects an existing uORF. Thus, with the introduction of an uORF, there is simultaneous alteration of the amino-acid sequence of another uORF. For this reason, there may by multiple uORF related functional consequences resulting from rs9905259. Finally, EIF5A encodes a protein that is involved in translation elongation, and possibly mRNA degradation[6,7]. There is possibility of a regulatory cascade, where uORF mediated regulation of translation affects a gene that itself regulates translation.

GO term annotation of uORFs interrupted in cancer:

Patterns of function in genes affected by mutation of uORFs in cancer was assessed via the GO genome annotation database [8]. Overrepresented GO terms were identified, with overrepresentation

assessed via the hypergeometric statistical test with multiple testing correction via Benjamini & Hochberg's FDR correction [9]. Networks between GO terms were constructed using the Cytoscape package BiNGO [10] [Figure S.2].

Three networks of overrepresented GO terms remain following correction for statistical significance and multiple testing. These are networks associated with cellular functions of probable significance in cancer -- cellular death, immune modulation, and tissue morphogenesis. Lack of response to apoptotic signaling, and immune tolerance, are well known mechanisms by which cancer cells prolong survival. The alteration of genes involved in tissue morphogenesis, may relate to the poor tissue differentiation exhibited by cancer cells.

Analysis of mutation across cancer types:

In order to evaluate the frequency with which uORFs are interrupted by mutation in cancer, the proportion of positive uORFs interrupted by mutation was calculated for each cancer type [Figure S.3]. The proportion of positively scored uORFs to negative scored uORFs varied across cancer types, ranging from a low of 1:12.7 for pilocytic astrocytoma, to a high of 1:4.6 for B-cell lymphoma (chi-square = 45, p-value = <0.001). This suggests that B-cell lymphoma or breast cancer, may depend to a greater extent on altered uORFs for increased cellular fitness, than pilocytic astrocytoma or CLL.

References:

1. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte f Chemie Chem. Mon.* **125,** 167–188 (1994).
2. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–13 (2005).
3. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Publ. Gr.* **45,** (2013).
4. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6,** R33 (2005).
5. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10,** 748–748 (2014).
6. Saini, P., Eyler, D. E., Green, R. & Dever, T. E. Hypusine-containing protein eIF5A promotes translation elongation. *Nature* **459,** 118–121 (2009).
7. Park, M. H., Nishimura, K., Zanelli, C. F. & Valentini, S. R. Functional significance of eIF5A and its hypusine modification in eukaryotes. *Amino Acids* **38,** 491–500 (2010).
8. Gene Ontology Consortium, T. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).
9. Author, T., Benjamini, Y., Hochberg, Y. & Benjaminit, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Source J. R. Stat. Soc. Ser. B J. R. Stat. Soc. Ser. B J. R. Stat. Soc. B* **57,** 289–300 (1995).
10. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21,** 3448–3449 (2005).

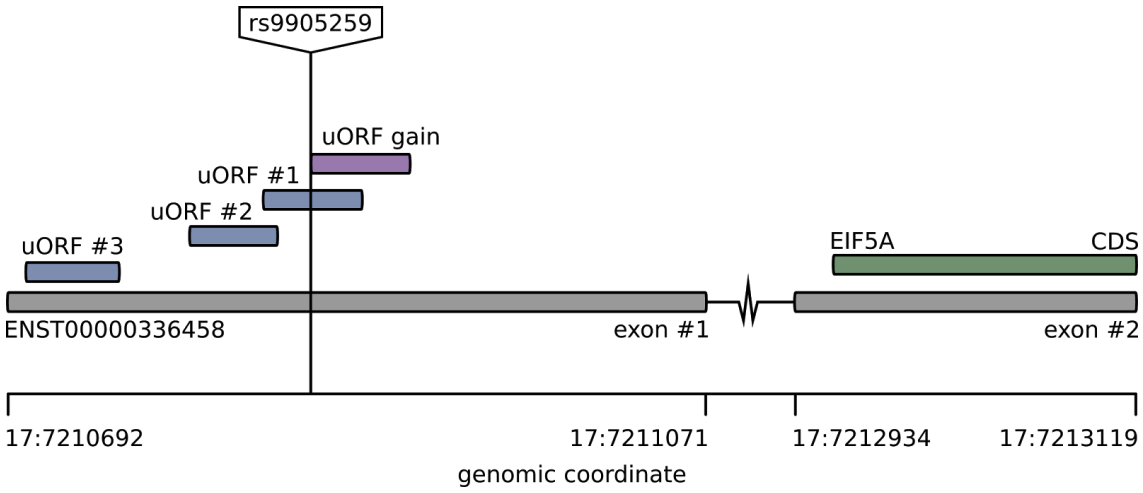

Figure S.1. uORF gain associated with the gene EIF5A and variant rs9905259. *Predicted positive ATG uORFs associated with the reference genome on transcript ENST00000336458.8 are numbered 1-3. The gained predicted positive ATG uORF associated with rs9905259 is positioned downstream of these three other ATG uORFs. Variant rs9905259 also intersects another uORF on the same transcript, altering its amino-acid sequence.*

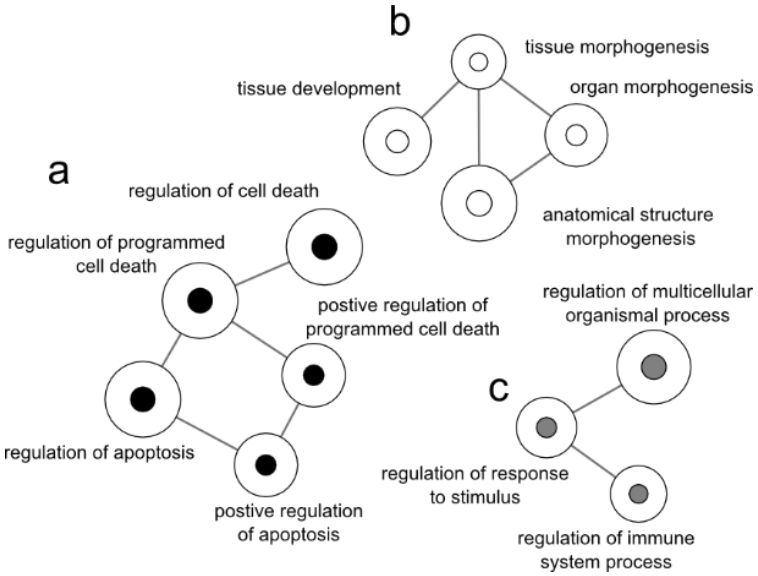

Figure S.2. GO/PANTHER terms, for statistically overrepresented genes with uORF start codons interrupted by somatic variants in tumor samples (Alexandrov et al.). The size of each node corresponds to the number of uORFs associated that GO term. Thresholds were established to eliminate relatively common GO terms (>1250 associated uORFs), and relatively uncommon GO terms (<250 associated uORFs). 3 principle networks emerge: a) tissue morphogenesis, b) immune function, and c) apoptosis. Networks were developed using the statistical package BiNGO and include adjustment for multiple hypothesis testing.

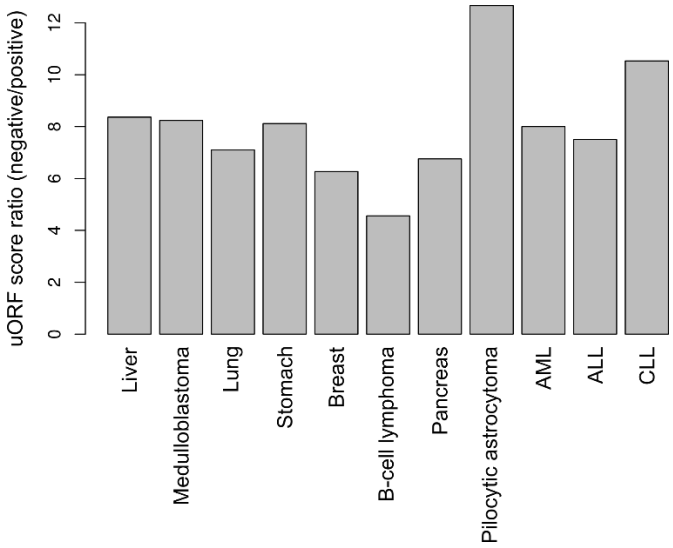

Figure S.3. uORF score ratio (# negative score / # positive score) for uORFs interrupted by mutation in cancer according to cancer type.