SUPPLEMENTAL INFORMATION

Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*

Jenna E. Smith[1], Juan R. Alvarez-Dominguez[2], Nicholas Kline[1], Nathan J. Huynh[1], Sarah Geisler[1,3], Wenqian Hu[2], Jeff Coller[1], and Kristian E. Baker[1]*

[1]Center for RNA Molecular Biology, Case Western Reserve University, Cleveland, OH, 44106 USA

[2]Whitehead Institute for Biomedical Research, Cambridge, MA, 02142 USA

[3]Present address: Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zürich, 4058 Basel, Switzerland

*Correspondence: K. Baker - keb22@case.edu, 216-368-0277, 216-368-2010

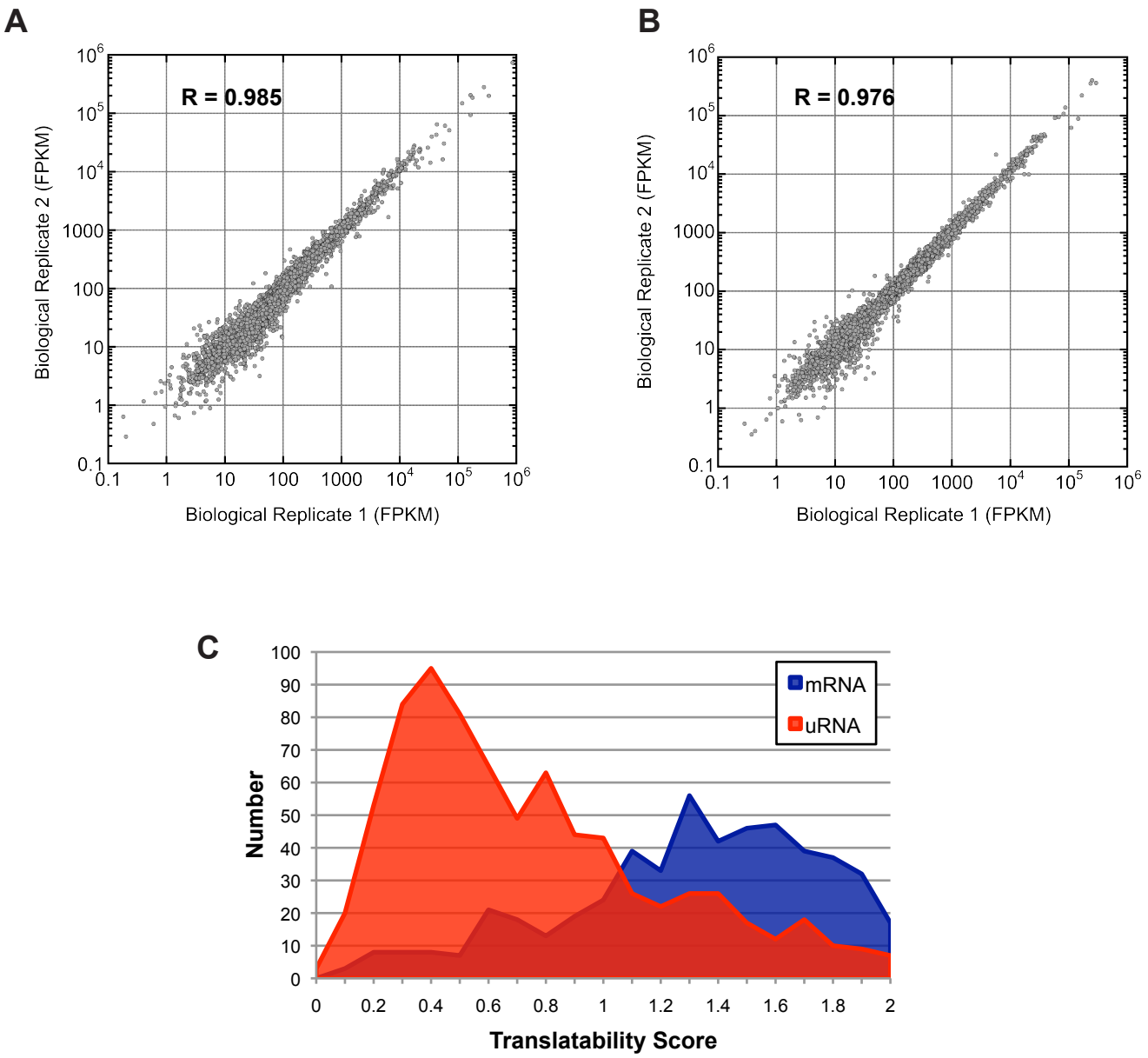

Figure S1, related to Figure 1. RNA-seq provides evidence that uRNAs associate with polyribosomes. (A) Regression analysis for wildtype biological replicates of RNA-seq for total RNA or (B) Polysome-seq for polysome-associated RNA. Correlation coefficients indicated. (C) Translatability score distribution for osubset of mRNAs and uRNAs 200-500 nt in length. mRNA n=571 (11.4% of total analyzed in Figure 1C; blue); uRNA n=824 (71.9% of total analyzed in Figure 1C; red).

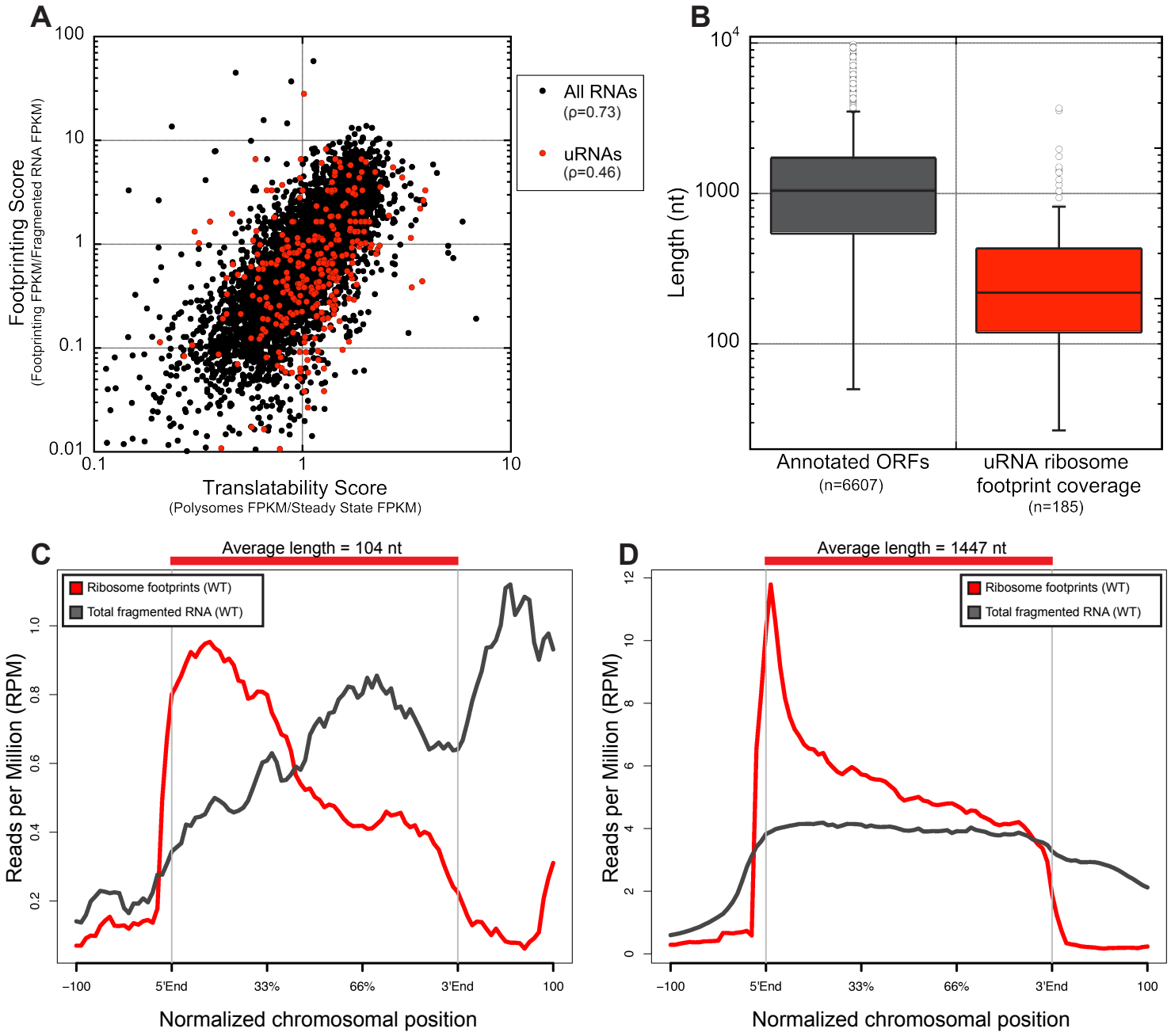

Figure S2, related to Figure 2. Ribosome footprinting predicts short open-reading frames encoded within uRNAs. (A) Comparison of polysome RNA-seq and ribosome profiling. The Translatability Score calculated from polysome RNA-seq and Footprinting Score calculated by ribosome footprinting were compared for all RNAs for which a score could be calculated for both assays. All RNAs (black) and uRNAs (red) are shown. Spearman rank correlation coefficients indicated. (B) Boxplot of the average length of yeast annotated ORFs (including verified, uncharacterized, and dubious ORFs) compared to the average size of the region covered by ribosome footprints for uRNAs. Box includes 25-75 percentiles; whiskers indicate ± 1.5 IQRs, with outliers indicated by circles. (C) and (D) Metagene plot of average sequencing coverage of ribosome footprints (red) or total fragmented RNA (gray) mapped along all predicted sORF coding regions (C) (n=43) or annotated mRNA CDS (D) (n=5017), plus 100 nucleotides flanking either end. 5'End indicates predicted or annotated start codon position; 3'End indicates predicted or annotated stop codon position. Red bar demarcates putative sORF metagene (C) or mRNA ORF metagene (D), with the average size indicated. Data are mean reads per million.

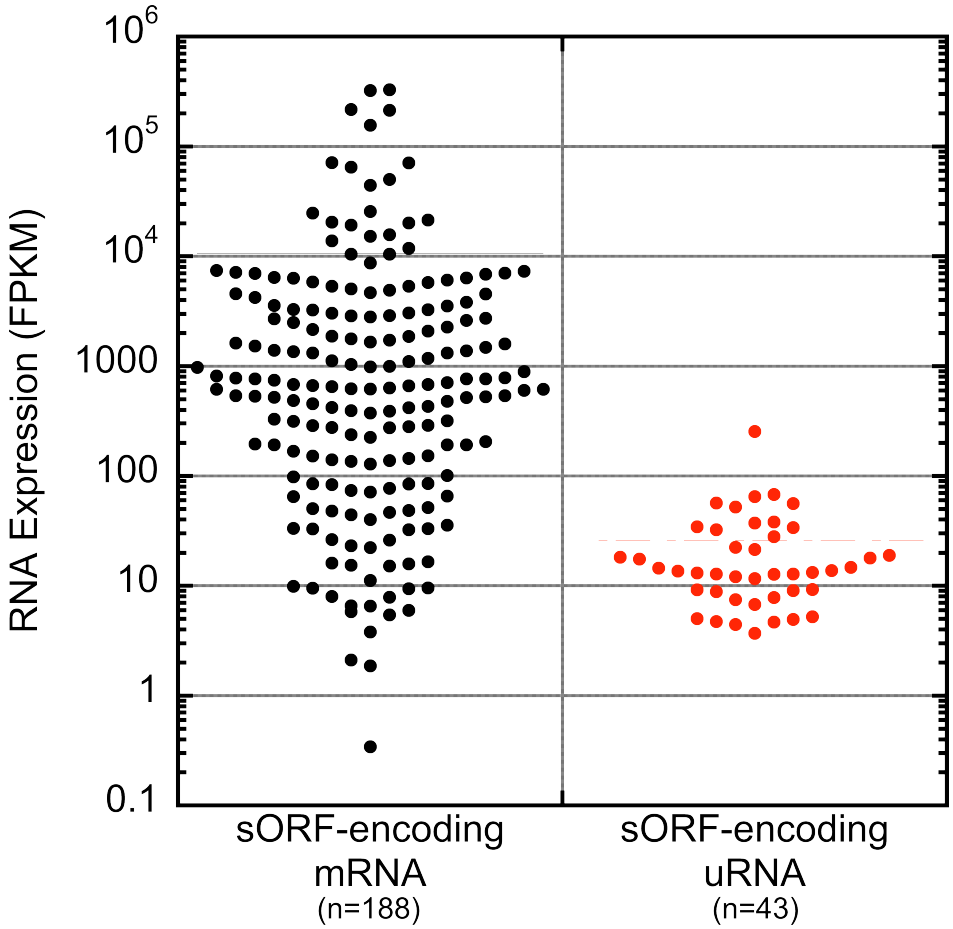

Figure S3, related to Figure 3. sORF-encoding uRNA expression is within range of mRNAs encoding short ORFs. Dot plot displaying expression levels (FPKM) for annotated sORF-encoding mRNAs (black), or uRNAs containing putative sORFs (red). Only mRNAs with an average expression of >0 in WT and >10 in *upf1Δ*, (expression thresholds for all analyses) are included.

Smith et al., Figure S4, related to Figure 4

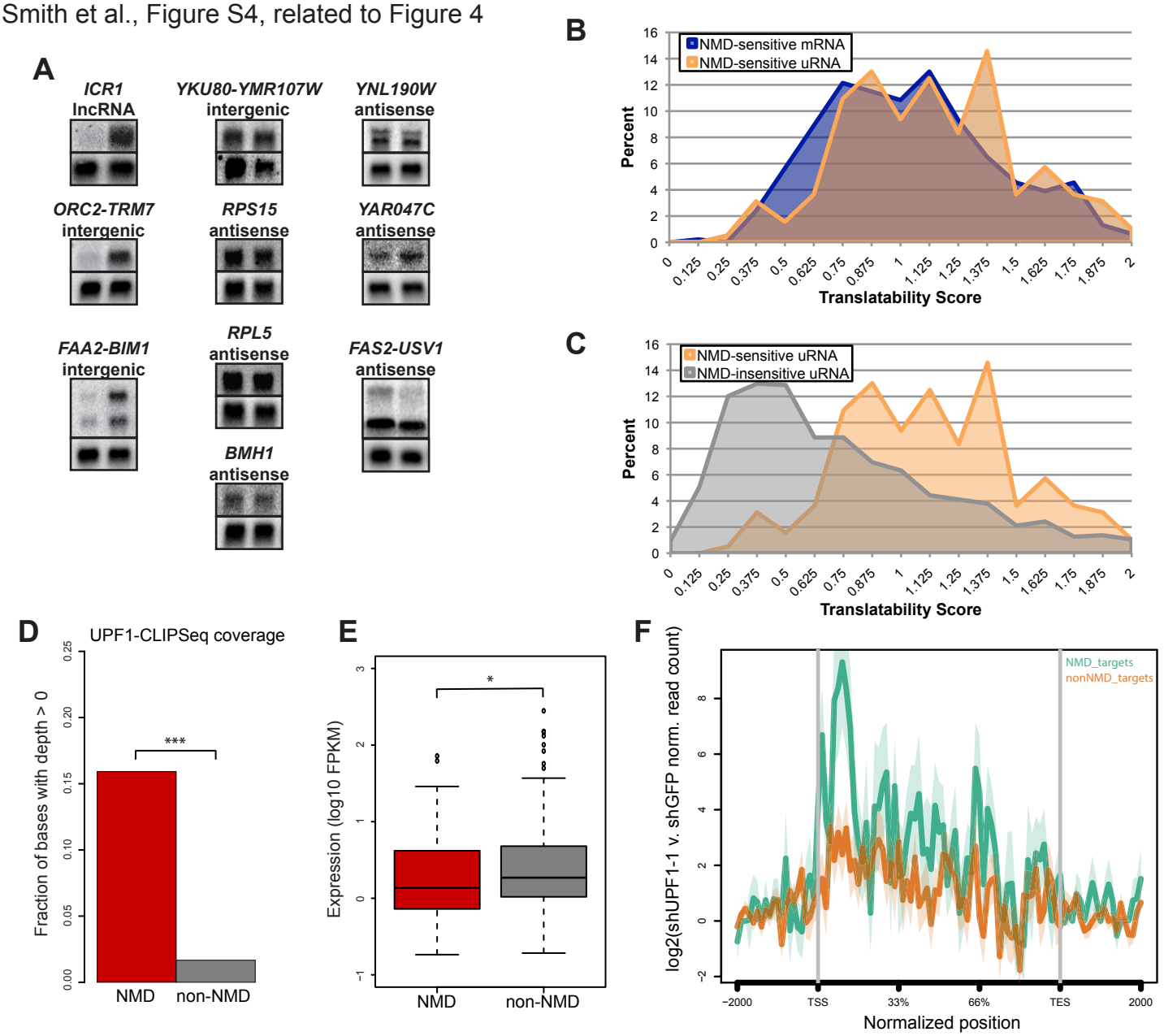

Figure S4, related to Figure 4. uRNAs are sensitive to translation-dependent nonsense-mediated RNA decay. (A) Northern blot analysis of steady state RNA from WT (lane 1) and *upf1Δ* cells (lane 2). Panels are identical to Figure 4C but include their respective *SCR1* loading control (bottom panels). (B) Translatability Score distribution for characterized mRNAs (blue) and uRNAs (orange) sensitive to NMD. (C) Translatability Score distribution for uRNAs sensitive to NMD (orange) vs. insensitive to NMD (gray). (D) Fraction of bases with coverage by UPF1 CLIP-seq tags for lncRNA transcripts sensitive to NMD (red) vs. NMD-insensitive lncRNAs (gray). UPF1 CLIP-seq tags are combined from three independent wild-type experiments. ***p<0.001 (Kolmogorov-Smirnov test). (E) Box-and-whisker plot of the distribution of gene-level expression values (FPKM) for lncRNAs sensitive to NMD (red) vs. those insensitive to NMD (gray). Expression values represent the average expression in wild-type cells from two independent experiments. *p<0.05 (Kolmogorov-Smirnov test). (F) Metagene plot of average change in the density of ribosome profiling reads mapping along the transcript body of intergenic lncRNAs sensitive to NMD (green) vs. those not sensitive to NMD (orange) following UPF1 depletion by shRNA. Changes are shown as log2 ratios of normalized Ribo-Seq read count in the shUPF1-1 experiment to that in the shGFP control. (TSS) transcription start site; (TES) transcription end site. Data are mean +/- SEM.

SUPPLEMENTAL TABLES

Table S1, related to Figures 1, 2, and 4. High-throughput sequencing statistics.

Sample	Replicate	# of Reads	Mapped reads		rRNA reads		non-rRNA mapped reads	
			#	%	#	% (of mapped reads)	#	% (of mapped reads
WT steady state	1	13,803,030	11,096,422	80.4	-	-	-	-
	2	31,213,437	22,617,117	72.5	-	-	-	-
upf1Δ steady state	1	15,855,617	12,523,774	79.0	-	-	-	-
	2	31,154,574	21,890,286	70.3	-	-	-	-
WT polysomes	1	31,342,788	22,141,454	70.6	-	-	-	-
	2	33,619,004	23,239,241	69.1	-	-	-	-
upf1Δ polysomes	1	25,964,168	17,821,788	68.6	-	-	-	-
	2	25,784,511	17,782,505	69.0	-	-	-	-
WT steady state fragmented RNA	-	7,341,479	5,507,344	75.0	526,798	9.6	4,980,546	90.4
upf1Δ steady state fragmented RNA	-	7,912,713	5,759,010	72.8	219,344	2.9	5,539,666	96.2
WT ribosome profiling	1	28,404,464	26,963,688	94.9	20,725,760	75.1	6,237,928	23.1
	2	22,793,695	21,682,577	95.1	17,916,898	80.8	3,765,679	17.4
upf1Δ ribosome profiling	1	21,727,393	20,603,069	94.8	16,440,899	77.9	4,162,170	20.2
	2	23,351,093	21,834,783	93.5	16,885,425	75.6	4,949,358	22.7

Biological replicates of steady-state RNA-seq, Polysome-seq, and ribosome profiling were assayed. Total number of reads before mapping to the sacCer2 yeast reference genome, following mapping (see Supplemental Experimental Procedures), and percentage mapped reads listed for all datasets. Additionally, total number and percentage of rRNA reads, and number and percentage non-rRNA mapped reads listed for ribosome profiling.

Table S2, related to Figure 1. List of uRNAs investigated in this study. All uRNAs identified in this study are listed. For each uRNA, a unique locus identifier, the chromosomal coordinates, strand of expression, and proximal annotated features are listed. uRNAs are referred to in figures based on the "Orientation to reference annotations" column. [a]Coordinates as defined by Cufflinks and based on sacCer2 genome assembly. [b]All coordinates listed in ascending order; for Crick (-) strand, 5' end originates at the second coordinate. (See Excel file)

Table S3, related to Figure 2. List of putative sORFs identified by ribosome profiling.

sORF Number	Encompassing uRNA	sORF Coordinates[ab]	Putative peptide
sORF-1	XLOC_000132-	chrII:169873-169637	MQRVRIGQWVYDMEAIHRSDSHECPKRTCGNQTLNPGSIIKEIQYKKYR YILFPPIAANQFTPGCSEYIPILHDAIKD*
sORF-2	XLOC_000464-	chrIII:309884-309747	MEPPIFIILTILSKLTFDKDDEQTTRRSKALEYCSVQPLSSNIAI*
sORF-3	XLOC_002595+	chrXII:675730-675861	MRSLKCSILMPFFSQIFSVSILAWDALSNKTLLTRSNAKVEVN*
sORF-4	XLOC_002893+	chrXIII:480923-481186	MISMEAINNFIKTAPKHDYLTGGVHHSGNVDVLQLSGNKEDGSLVWNHT FVDVDNNVVAKFEDALEKLESLHRRSSSSTGNEEHANV*
sORF-5	XLOC_000429+	chrIII:242629-242685	MLFHGFVSSKGLSVVPQQ*
sORF-6	XLOC_000768-	chrIV:916135-916022	MNLNRATEKTGRHNKKFISCMTYVTVNYIPNKLYIEK*
sORF-7	XLOC_002334-	chrXI:513546-513430	MMYFTRMERPQESTKQNMMLKYHLFSNIHIASILSYAV*
sORF-8	XLOC_002919-	chrXIII:619370-619098	MTEMTLLPKKKSISIHSSKYSGIQLSWFENLFSSSEEITSPSRLLTTLLTILF RYESSFFSPKILTVLNPIPKCGYEPILMASETFVDSS*
sORF-9	XLOC_003873+	chrXVI:777577-777669	MLVINISQLKEIKLKYSTKTYFRIKFLGGT*
sORF-10	XLOC_001697+	chrVII:902532-902762	MGCISVSHILKLSNSPSKKRDLITRLNQRVTSGCLDLHVSSKKLINPSEKA AAEIKSARVSWSSQTRYCCFFNLFS*
sORF-11	XLOC_002196+	chrXI:66560-66604	MMEMDSGCDCVVKM*
sORF-12	XLOC_002899+	chrXIII:502246-502341	MRCLIPLPKWNANRYAAEPLATHWDHLGIFS*
sORF-13	XLOC_000636+	chrIV:525039-525113	MFKNSKNLSINPGNDADKASFSDP*
sORF-14	XLOC_000636+	chrIV:525082-525150	MRIRRASPTHEHNRSSSVLVVG*
sORF-15	XLOC_003583-	chrXV:969893-969765	MKRKKRNVIIGMMPFVALTSTTSIYDNGNMCDQSELCSCHLH*
sORF-16	XLOC_003728+	chrXVI:286664-286765	MDMHYVGHMTLIVQVQLMYRLKRLRLLIYNAIY*
sORF-17	XLOC_000364+	chrIII:39662-39805	MNIHKIKIICMFLIYRYTIKHFQEFVCRLFARDLIKVLPTKINLYFF*
sORF-18	XLOC_000631+	chrIV:513023-513172	MLVKDYSIGYTEFTRTPPQGYRNLHKGIDISNISSNIVIFLFILCVVSH*
sORF-19	XLOC_003307+	chrXV:38888-39046	MLSLTFISPTLSQIFDHVIYMLFSNQVINFTELKVAAAENSHLIILYIDPTY*
sORF-20	XLOC_003329+	chrXV:96656-96823	MTCTYIVVNVNMSGLPMVQMKEKFVFGCTELLRIEEGLIFHTCFYVLVLA SNHPY*
sORF-21	XLOC_000204+	chrII:362898-362939	MSRQLVTLLLVYT*
sORF-22	XLOC_000631+	chrIV:512693-512740	MYILNVYALNTIDDF*
sORF-23	XLOC_001229+	chrV:285208-285264	MPAVLMVNPCPISLRHFL*
sORF-24	XLOC_001689+	chrVII:875763-875801	MSLVFIQPHFIF*

sORF-25	XLOC_003019+	chrXIII:873056-873103	MPYITNTAEATMSTV*
sORF-26	XLOC_003128+	chrXIV:270232-270270	MKYYKKFINFLN*
sORF-27	XLOC_001405+	chrVII:18942-19028	MFNIFAMIHSRFCPALNFTPTLRVHSVR*
sORF-28	XLOC_000405+	chrIII:169068-169193	YLHSRYRHINIKILNIETSFRLRFGCRKPKMQFKWVNNLNG*
sORF-29	XLOC_000841+	chrIV:1206674-1206796	MLPRLLKHVGIEINYHLLTSIYITSILSYTVLEDDANDEK*
sORF-30	XLOC_001222+	chrV:268931-268963	MYGCARHSSS*
sORF-31	XLOC_003621+	chrXV:1003992-1004036	MLVLNMSIQSLVVH*
sORF-32	XLOC_000144-	chrII:220889-220800	MSFPYEHAQAKNLPEILLYYKKKEMNLSK*
sORF-33	XLOC_000337-	chrII:792024-791974	MSFQRLKLLKTALFVY*
sORF-34	XLOC_000617-	chrIV:471496-471440	MNTVTINKAGLTEHVGSG*
sORF-35	XLOC_000803-	chrIV:1019114-1019028	MSWKLLHFLPFEISSYMYMLTLLTSKMS*
sORF-36	XLOC_001190-	chrV:177560-177402	MVFIRDCVMNSFHYVNEPRSNVNQRTCGQESYNYLFNEPKEITCSRLGV NLF*
sORF-37	XLOC_001277-	chrV:432186-432139	MQSRNKKQIAISSPL*
sORF-38	XLOC_001365-	chrVI:159009-158926	MALAIRTRMHNRQDIIGMQQLKLLLML*
sORF-39	XLOC_001409-	chrVII:17702-17652	MSNSTILENNTIVRCI*
sORF-40	XLOC_001678-	chrVII:811263-811213	MRIYTLHTYYRKHCYF*
sORF-41	XLOC_001678-	chrVII:811200-811162	MANERSSFIPLS*
sORF-42	XLOC_002811-	chrXIII:311954-311892	MHLRDKVKLPSYSCKRNLYF*
sORF-43	XLOC_003248-	chrXV:5661-5560	MIRMSWWPCITNIGSSRGLETLNAVRMDIYIGD*
sORF-44	XLOC_003248-	chrXV:6801-6511	MHRIPCQNVFCYNTFHEANWGYYLHLSGFHICFLDNSFNSSIVVRMAMR IYYGTHRFLRTMSIVKFKRLLGSLCGKKRINDYQRSITFDYSHIRDI*
sORF-45	XLOC_003866-	chrXVI:781588-781517	MVSLDSLLVLLKKNGILFLDNVS*
sORF-46	XLOC_003609-	chrXV:1064363-1064313	MNGRKNCSFEECFSGR*
sORF-47	XLOC_003356-	chrXV:326433-326329	MNLHVFIKLIEMEFSSSSRSVFCRLCSYDAKRLK*

All putative sORFs defined based on phased ribosome footprints (See Supplemental Experimental Procedures), the uRNA by which they are encoded, the chromosomal coordinates for the sORF, and the putative peptide sequence are presented. [a]All chromosomal coordinates based on sacCer2 genome annotation. [b]For Crick strand, coordinates are listed in descending (5'-3') order.

Table S4, related to Figure 3. **Conservation of sORFs in other yeast species.** For each sORF, the phastCons log-odds score for previously identified conserved elements (Siepel et al., 2005), TBLASTN results (percent identical residues relative to full-length putative peptide, E-values, and bit scores), and Ka/Ks ratios (Zhang et al., 2006) are presented. [a]Bold indicates phastCons conserved element completely overlaps sORF. [b]Italics indicates that conserved element may be influenced by gene antisense to sORF uRNA. [c]N/A indicates no conserved element corresponds to sORF locus. [d]Percent identity score of 0 indicates no alignment found at E<10. [e]For all comparisons where BLAST produced no match, "-" is recorded. [f]N.D. = not determined; nucleotide sequences show 100% alignment. [g]N.S. = not significant; value not reported due to Fisher's p-value >0.05. [h]Bold indicates Ka/Ks ratio supports purifying selection. [i]All data reported as for *S. pastorianus*. (See Excel file)

Table S5, related to Experimental Procedures. List of strains, oligos, and plasmids.

Name	Description[a]	Notes	Reference
yKB154	MATa, *ura3*, *leu2*, *his3*, *met15*	Wild-type	EUROSCARF
yKB146	MATa, *ura3*, *leu2*, *his3*, *met15, upf1*::KAN	*upf1Δ*	EUROSCARF
yKB596	MATa, *ura3*, *leu2*, *his3*, *met15,* sORF-4-HA::HIS3	sORF-4 plus 3xHA C-terminal tag	This study
yKB597	MATa, *ura3*, *leu2*, *his3*, *met15,* sORF-4-HA::HIS3	sORF-4 plus out-of-frame 3xHA C-terminal tag	This study
oJC1348	GTCATGCTCCTTTTTATGGGTTCTCGTCGTAATAATCCTG	*ORC2-TRM7* intergenic uRNA oligo probe	This study
oJC1352	ACCTGAAAGAGACGCCTTGTATCTTCTATAGGTCAACTAG	*FAA2-BIM1* intergenic uRNA oligo probe	This study
oJC1917	GTTATTCTATTCTTGAGCAGGCACTTTTAGGGTTGGGCAA	*ICR1* ncRNA oligo probe	This study
oJC1981	GTATGGTTCCATACTAAACTACCATCTTCTTTATTGCCGC	*YKU80-YMR107W* intergenic uRNA oligo probe	This study
oJC306	GTCTAGCCGCGAGGAAGG	*SCR1* ncRNA oligo probe	This study
oJC1984	GAAATGTCCACTGAAGATTTCGTCAAGTTGGCCCC	*RPS15* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB707	CTGGAACAATGATCATGTTTCTCATGTGGGTTCTGACTGGAGC	*RPS15* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB708	AGCTAGAGTTAGAAGAAGATTTGCCCGTG	*RPS15* antisense uRNA asymmetric PCR primer for Northern probe	This study
oJC1989	CAACACAAGGCCAAGTACAACACTCCAAAGTACAGATTGG	*RPL5* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB713	CAACTTCTTCAACACCCTTGTAAGTTTCGTCCAAACC	*RPL5* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB714	CTGTCAAATCATCTCTTCTACCATCACTGGTG	*RPL5* antisense uRNA asymmetric PCR primer for Northern probe	This study
oJC1991	GTCGAGTTGATTTGTTCGTACCGTTCGAAGATTGAGACCG	*BMH1* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB711	AGAGAAGTTAAGAGCCAAACCTAGACGGATTGGGTGAG	*BMH1* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB712	AACTAACTAAGATCTCCGACGATATTTTGTCCG	*BMH1* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB702	CGTAAGAACAATGCCGCCCCTGGTCCATCTAATTTCAACT	*YNL190W* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB717	CACTTTTGCACAAGCACACGTAAACACATAGTAGTCGAAATAG	*YNL190W* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB718	CCATAAAATTGTTTGGTGTTACCGCTGGTAG	*YNL190W* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB700	GTTCCTCGATCGACTAGTGCCATTCAATGAGATAAGGAGT	*YAR047C* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB720	GAGCAGAGGTTAGCTCCGTCTCAACCAATTTTGTAC	*YAR047C* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB721	AGTATAGTAAGATATAATCCCACTAACGATTAGCGAGTG	*YAR047C* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB748	CTTCCAGAGCGCCAGCATCGATCATAGCTG	*FAS2-USV1* antisense uRNA forward PCR primer to make template for asymmetric PCR	This study
oKB750	CTGGTGGGTTTACTATTACTGTCGCTAGAAAATACTTACAAACTCGCTG	*FAS2-USV1* antisense uRNA reverse PCR primer to make template for asymmetric PCR	This study
oKB749	GGACTACCATCTGGTAGACAAGATGGTG	*FAS2-USV1* antisense uRNA asymmetric PCR primer for Northern probe	This study
oKB688	/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGC/iSp18/CACTCA/iSp18/TTCAGACGTGTGCTCTTCCGATCTATTGATGGTGCCTACAG	RT primer for ribosome profiling	Ingolia *et al*., 2012

oKB689	AATGATACGGCGACCACCGAGATCTACAC	PCR amplification of ribosome profiling cDNA libraries, forward primer	Ingolia *et al*., 2012
oKB690	CAAGCAGAAGACGGCATACGAGATTGGTCAG TGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #1	Ingolia *et al*., 2012
oKB691	CAAGCAGAAGACGGCATACGAGATCACTGTG TGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #2	Ingolia *et al*., 2012
oKB692	CAAGCAGAAGACGGCATACGAGATATTGGCG TGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #3	Ingolia *et al*., 2012
oKB693	CAAGCAGAAGACGGCATACGAGATTCAAGTG TGACTGGAGTTCAGACGTGTGCTCTTCGG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #4	Ingolia *et al*., 2012
oKB694	CAAGCAGAAGACGGCATACGAGATCTGATCG TGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #5	Ingolia *et al*., 2012
oKB695	CAAGCAGAAGACGGCATACGAGATTACAAGG TGACTGGAGTTCAGACGTGTGCTCTTCCG	PCR amplification of ribosome profiling cDNA libraries, reverse primer, Index #6	Ingolia *et al*., 2012
oKB769	CTCATCCTCATCCACAGGCAATGAAGAACACG CTAACGTTCGGATCCCCGGGTTAATTAA	Forward PCR primer to amplify 3xHA-His3 product from pFA6a-3HA-His3MX6, with gene-specific sequences for chromosomal tagging of sORF-4	This study
oKB770	CTTATTTCTCACATCATTATGAAGTGACTCCCC TCGGTTAGAATTCGAGCTCGTTTAAAC	Reverse PCR primer to amplify 3xHA-His3 product from pFA6a-3HA-His3MX6, with gene-specific sequences for chromosomal tagging of sORF-4	This study
oKB784	CTCATCCTCATCCACAGGCAATGAAGAACACG CTAACGTCGGATCCCCGGGTTAATTAA	Forward PCR primer to amplify 3xHA-His3 product from pFA6a-3HA-His3MX6, with gene-specific sequences for chromosomal tagging of sORF-4 out-of-frame	This study
oKB789	TTCCTTACGGAACCCAAGTGTG	Forward PCR primer to amplify *YBL027W-YBL026W* intergenic uRNA +/- 500 bp, for generation of pKB561	This study
oKB790	TTACTGTATCTACATCGGGATACTAATAGTAC	Reverse PCR primer to amplify *YBL027W-YBL026W* intergenic uRNA +/- 500 bp, for generation of pKB561	This study
oKB791	ATACCAATTTTACACGATGCCATAAAGGACGAT TATAAAGATGATGATGATAAATAGACAAGCTAC GTTGAAACAAGAACCCGC	Forward PCR primer to insert 1XFLAG tag at C-terminus of sORF-1 in*YBL027W-YBL026W* intergenic uRNA in pKB561, for generation of pKB565	This study
oKB792	GCGGGTTCTTGTTTCAACGTAGCTTGTCTATT TATCATCATCATCTTTATAATCGTCCTTTATGGC ATCGTGTAAAATTGGTAT	Reverse PCR primer to insert 1XFLAG tag at C-terminus of sORF-1 in*YBL027W-YBL026W* intergenic uRNA in pKB561, for generation of pKB565	This study
oKB797	GGTACTTCCGCTAATAGACTACAAAC	Forward PCR primer to amplify *YKU80-YMR107W* intergenic uRNA +/- 500 bp, for generation of pKB562	This study
oKB798	GTTCGTACTTCCTTCTGAGCAG	Reverse PCR primer to amplify *YKU80-YMR107W* intergenic uRNA +/- 500 bp, for generation of pKB562	This study
oKB799	TCCACAGGCAATGAAGAACACGCTAACGTTG ATTATAAAGATGATGATGATAAATAACCGAGGG GAGTCACTTCATAATGATGT	Forward PCR primer to insert 1XFLAG tag at C-terminus of sORF-4 in*YKU80-YMR107W* intergenic uRNA in pKB562, for generation of pKB566	This study
oKB800	ACATCATTATGAAGTGACTCCCCTCGGTTATTT ATCATCATCATCTTTATAATCAACGTTAGCGTG TTCTTCATTGCCTGTGGA	Reverse PCR primer to insert 1XFLAG tag at C-terminus of sORF-4 in*YKU80-YMR107W* intergenic uRNA in pKB562, for generation of pKB566	This study
yEpLac181	2μ, LEU2	Parental vector used to construct pKB561, pKB562, pKB565, pKB566	Gietz and Sugino, 1988
pKB561	*YBL027W-YBL026W* intergenic uRNA +/- 500 bp		This study
pKB562	*YKU80-YMR107W* intergenic uRNA +/- 500 bp		This study
pKB565	*YBL027W-YBL026W* intergenic uRNA +/- 500 bp + C-terminal FLAG		This study
pKB566	*YKU80-YMR107W* intergenic uRNA +/- 500 bp + C-terminal FLAG		This study
pFA6a-3HA -His3MX6		Vector used to construct chromosomal 3xHA-tagged sORF loci	Longtine et al., 1998

All yeast strains, oligonucleotides, and plasmids used to generate data or constructs presented in this study are provided. Description (including nucleotide sequences of oligonucleotides), notes regarding the context of their use, and source for all reagents are provided. [a]All oligonucleotide sequences are listed from 5'-3'.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Yeast Culture

Cells were grown at 30 °C in synthetic medium plus 2% glucose and appropriate amino acids at 250 RPM to mid-log phase, unless otherwise noted. Yeast strains, plasmids, and oligonucleotides (Integrated DNA Technologies) can be found in Table S5.

Total RNA Library Preparation

Whole-cell RNA was isolated using glass-bead cell lysis and phenol extraction as previously described (Geisler et al., 2012). 5 µg of DNase I-treated (Roche 04716728001) whole-cell RNA was depleted of rRNA using the Human/Mouse/Rat RiboZero rRNA removal kit (Epicentre MRZH11124). Small RNAs were excluded using RNA Clean and Concentrator-5 spin columns (Zymo R1015), substituting 26.6% ethanol final concentration at steps 1-2 of the manufacturer's recommended protocol to enhance removal of RNAs <200 nt (data not shown). Strand-specific, random-primed cDNA libraries were generated by the CWRU Genome and Transcriptome Sequencing Core, using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre SSV21106) and ScriptSeq Index PCR Primers. Libraries were prepared for biological replicates of WT and *upf1Δ* strains.

Polyribosome Analysis

Yeast cultures were grown to mid-log phase, treated with 100 µg/mL cycloheximide (CHX), harvested immediately by centrifugation, and cell pellets flash frozen on dry ice. Lysis was carried out at 4 °C. Cell pellets were lysed in polysome lysis buffer (10 mM Tris, pH 7.4, 100 mM NaCl, 30 mM $MgCl_2$, 100 µg/mL CHX, 1 mM DTT) by mechanical disruption using glass beads. Cell debris was removed by centrifugation through an 18 Ga puncture hole for 2 minutes at 2000 RPM, and the resulting lysate was pre-cleared at 29,000 RPM for 10 minutes in a Beckman TLA-120.2 rotor. Lysate was treated on ice with 1% Triton X-100 for 5 minutes. 10 units (OD_{260}) of lysate were added to a 15-45% (w/w) sucrose gradient (buffer 50 mM Tris acetate, pH 7.0, 50 mM NH_4Cl, 12 mM $MgCl_2$, 1 mM DTT) prepared using a Biocomp gradient maker. Gradients were

centrifuged for 2:26 hr at 41,000 RPM in a Beckman Sw-41Ti rotor. Gradients were fractionated, and RNA was precipitated and extracted as described previously (Sweet et al., 2012). 5 µg of RNA was used to prepare polysome-seq libraries as described above for total RNA libraries.

RNA-Seq and Polysome-Seq Sequencing and Analysis

Sequencing and mapping: cDNA libraries prepared from total and polysome-associated RNA were sequenced on the Illumina HiSeq2000 platform at the Institute for Integrative Genome Biology High-Throughput Sequencing Core at the University of California, Riverside on a single-end, 100 cycle flow cell. On the Galaxy platform (usegalaxy.org; Goecks et al., 2010; Blankenberg et al., 2010; Giardine et al., 2005), the sequencing data FASTQ files were run through the "NGS: QC and manipulation/ FASTQ Groomer" tool. "NGS: QC and manipulation/Compute quality statistics" was used to compute quality scores and 1 low-quality nucleotide was trimmed from the right end of all reads during mapping. Reads were mapped to the sacCer2 genome on Galaxy with "NGS: Mapping/Map with Bowtie for Illumina" (Langmead et al., 2009) using a SOAP-like alignment policy to allow 2 mismatches over the entire length of the read (-v 2), and excluding any read that did not map uniquely to the genome (-m 1).

Identification of unannotated RNAs: Reads were assembled into transcripts using Cufflinks v2.1.1 (Trapnell et al., 2010) with bias correction and multi-read correction, using reference annotation-based transcript assembly (RABT; Roberts et al., 2011) to identify unannotated transcripts (-GTF-guide -b -u --library-type ff-firststrand; all other parameters default). The sacCer2 Ensembl Genes annotation downloaded from the UCSC genome table browser was used as a guide during transcript assembly (genome.ucsc.edu/cgi-bin/hgTables?command=start). RABT assembles reads into transcripts, and then compares assembled transcripts to the reference genome annotation to identify transcripts significantly different from transcripts predicted by the annotation (Roberts et al., 2011), allowing the identification of novel transcripts that map to regions of the sacCer2 genome lacking annotated features.

Using the default --overlap-radius option, unique transcripts must be separated by at least 50 basepairs from annotated transcripts to prevent merging either at the Cufflinks step or following Cuffmerge step. Transcripts <200 nucleotides or with a coverage <1 read per million were filtered from the dataset. A master annotation compiling RNAs detected in all datasets was generated with Cuffmerge (Cufflinks v2.1.1). Notably, Cuffmerge includes a step that filters transcripts likely to be artifacts including possible polymerase run-on fragments, or transcripts within 2 kilobases downstream of a reference transcript (Trapnell et al., 2012).

For classification of RNAs: “mRNAs” include any gene annotated with a YXXNNNX systematic name; “known ncRNAs” include C/D box snoRNAs (*snR18, snR65, snR4, snR71, snR76, snR45, snR63, snR128, snR190, snR70*), H/ACA box snoRNAs (*snR46, snR30, snR44, snR34, snR11, snR49, snR81, snR8, snR5, snR161, snR43, snR189, snR84, snR80, snR37, snR42, snR85, snR86, snR191, snR9, snR36, snR35, snR31*), spliceosomal RNAs (*snR14, snR7-L, snR19, LSR1*), U3 snoRNA *snR17b*, telomerase RNA *TLC1*, signal recognition particle 7S RNA *SCR1*, RNase MRP *NME1*, and RNase P component *RPR1*; uRNAs include all assembled transcripts that were not assigned and did not align to a reference annotation, excluding those mapping to mitochondrial DNA. Any assembled transcript spanning more than one annotated chromosomal feature was excluded from all downstream analysis.

Quantification of expression: Expression (FPKM) was calculated using Cuffdiff (Cufflinks v2.1.1; Trapnell et al., 2013) with bias correction and multi-read correction, providing biological replicates for analysis, with the master annotation generated by Cuffmerge above as a reference (-b -u --library-type ff-firststrand). Any RNAs which did not have an average expression in RNA-seq datasets of FPKM ≥10 in *upf1Δ*, and FPKM >10 in wild-type for ncRNAs, were excluded from further analysis.

Comparison to previous ncRNA transcripts: Comparison of uRNAs to previous transcript annotations (Dcp2-sensitive, Geiser et al. 2012; SUTs, CUTs, Xu et al., 2009; or XUTs, van Dijk et al., 2011) was performed by manually comparing the

published chromosomal coordinates from each of these 4 classes of transcripts to the coordinates of uRNAs defined by Cufflinks analysis. If a uRNA overlapped a previously classified ncRNA >50%, or vice versa, the uRNA was categorized as being identical to or overlapping a member of that class. In many cases, ncRNAs have already been previously classified in more than one category. For example, when XUTs were described, 543 were identified as also being SUTs and 183 were identified as also being CUTs (van Dijk et al., 2011); this ambiguity between classes is reflected in the fact that many uRNAs overlap ncRNAs in more than one class. Additionally, in some cases uRNA annotations spanned adjacent but non-overlapping ncRNAs, also resulting in grouping of the uRNA into more than one class; however, in these cases the uRNA transcript isoform described here is likely to have distinct stability characteristics from overlapping ncRNAs.

Translatability Score calculation: For each detected RNA, we calculated the ratio of RNA-Seq reads associated with polysomes (Polysome-seq data) relative to reads from total RNA at steady-state (RNA-seq data), to calculate the Translatability Score ($FPKM_{polysomes}/FPKM_{steady\text{-}state}$). Sequencing datasets were normalized for *PGK1* and *RPL41A* mRNAs to have a translatability score of 1. All graphical representations of translatability score data are presented as histograms of the number of RNAs per bin, generated with 40 bins from scores 0-5.

Identification of NMD-sensitive RNAs: RNAs were identified as upregulated in *upf1Δ* by comparing WT and *upf1Δ* total RNA samples using Cuffdiff with parameters as described above. Upregulated transcripts were required to be statistically significant at an FDR of <0.05 and show a ≥2-fold average increase in expression.

Ribosome Profiling Library Preparation

Isolation and sequencing of ribosome-protected RNA fragments was performed based on the described protocol (Ingolia *et al.*, 2012), with the following modifications. Yeast cultures were grown in synthetic dextrose medium plus amino acids to mid-log phase, treated with 100 µg/mL CHX, harvested immediately by centrifugation, and cell pellets

flash frozen on dry ice. Lysis was carried out at 4 °C. Cell pellets were lysed in polysome lysis buffer (10 mM Tris, pH 7.4, 100 mM NaCl, 30 mM $MgCl_2$, 100 µg/mL CHX, 1 mM DTT) by mechanical disruption using glass beads. Cell debris was removed by centrifugation through an 18 Ga puncture hole for 2 minutes at 2000 RPM, and the resulting lysate was pre-cleared at 14,000 RPM for 10 minutes in tabletop centrifuge. Lysates were treated with 1% Triton X-100 for 5 minutes. 12.5 units (OD_{260}) of lysate were treated with 188 U RNase I (Invitrogen AM2294) in 250 µL at 24 °C for 1hr. Lysates were loaded onto a 15-45% (w/w) sucrose gradient, centrifuged, and fractionated as described for polysome analysis above.

RNA was precipitated from fractions containing the 80S monosome peak with 2 volumes of 95% ethanol at -80 °C overnight, and centrifuged for 30 minutes at 13,200 RPM to collect RNA. RNA was resuspended in LET (25 mM Tris, pH 8.0, 100 mM LiCl, 20 mM EDTA) plus 1% SDS, and extracted once each with an equal volume of phenol/ LET, phenol/chloroform/LET, and chloroform. RNA was precipitated with 300 mM NaCl, 1.5 µL GlycoBlue, and >1 volume isopropanol for 30 minutes on dry ice. RNA was collected by centrifugation at 13,200 RPM for 30 minutes at 4 °C, air dried, and resuspended in 10 mM Tris, pH 8.0. RNA from all monosome fractions for each sample was pooled, and 5 µg aliquots depleted of ribosomal RNA using the Human/Mouse/Rat RiboZero rRNA removal kit (Epicentre MRZH11124). Each rRNA-depleted sample was purified through RNA Clean and Concentrator-5 spin columns (Zymo R1015), substituting 60% ethanol at steps 1-2 of the manufacturer's recommended protocol to facilitate purification of small RNAs.

Size-selection of 26-34 nt fragments of RNA was carried out by electrophoresis on a 15% denaturing polyacrylamide gel, excision, and gel purification as described (Ingolia *et al.*, 2012). 2 aliquots per sample were pooled, and a second ribosomal RNA depletion was performed using the Epicentre Human/Mouse/Rat RiboZero kit (eliminating the 50 °C incubation step) and Zymo RNA Clean and Concentrator-5 spin columns to purify RNA as above. RNA was dephosphorylated, a 3' linker ligated, first-strand cDNA synthesized, and cDNA circularized as in described protocol, (Ingolia *et al*., 2012). cDNA libraries were amplified with 12-14 cycles of PCR with indexed primers (see Table S5).

To generate fragmented RNA control libraries, whole-cell RNA was purified, DNase-treated, and ribosomal RNA removed as described for the RNA-seq library preparation. RNA was fragmented with base as described (Ingolia, 2010) and fragments of 26-34 nt were gel purified and used for library preparation as described above for ribosome footprinting libraries. Libraries were prepared for biological replicates of WT and *upf1Δ* strains for ribosome footprinting, or a single replicate of each strain for the fragmented RNA control.

Ribosome Profiling/Fragmented RNA Sequencing and Analysis

Sequencing and mapping: cDNA libraries prepared for total fragmented RNA or ribosome footprints were sequenced on the Illumina HiSeq2500 platform at the Institute for Integrative Genome Biology High-Throughput Sequencing Core at the University of California, Riverside on a single-end, 50 cycle flow cell. Using the Galaxy platform (usegalaxy.org; Goecks *et al*., 2010; Blankenberg *et al*., 2010; Giardine *et al*., 2005), the sequencing data FASTQ files were run through the "NGS: QC and manipulation/FASTQ Groomer" tool. "NGS: QC and manipulation/Compute quality statistics" was used to compute quality scores which indicated high-quality sequencing across the length of the reads. Data processing was carried out in Galaxy as described (Ingolia *et al*., 2012). Briefly, the sequencing adaptor was clipped from the 3' end of each read with "NGS: QC and manipulation/Clip," and any reads without a clipped adaptor or that were <25nt in length after clipping were discarded. The clipped read was trimmed to nucleotides 2-50 with "NGS:QC and manipulation/Trim sequences." Reads were mapped to the sacCer2 yeast genome on Galaxy with "NGS: Mapping/Map with Bowtie for Illumina" (Langmead *et al*., 2009) using a SOAP-like alignment policy allowing 1 mismatch (-v 1), reporting 1 alignment per read (-k 1), and discarding any reads aligning to more than 16 locations in the genome (-m 16). rRNA reads (any reads mapping to chrXII: 451,000-468,999) were identified and removed using the "Filter and Sort/Select" tool.

Modifying uRNA coordinates: The 5' and 3' termini of all uRNA detectable by total RNA fragmentation were manually demarcated. The most inclusive 5' and 3' terminus among the uRNA boundaries annotated by Cufflinks and the manual annotation of the total fragmented RNA was identified. These updated uRNA transcript boundaries were converted into GTF format, and combined with the sacCer2 Ensembl Genes annotation for use as the reference annotation for quantification of ribosome profiling sequencing data (see below). This adjustment ensured that quantification of ribosome footprinting and total fragmented RNA sequencing was inclusive of the largest isoform of each transcript identified between this sequencing dataset and the RNA-seq sequencing dataset which initially defined uRNA coordinates.

Quantification of ribosome footprint coverage: FPKMs were obtained using Cuffdiff (Cufflinks v2.1.1; Trapnell *et al.*, 2013) with bias correction and multi-read correction, providing biological replicates for analysis where possible, using the reference GTF file described above in "Modifying uRNA coordinates" (-b -u --library-type ff-firststrand). RNAs with poor coverage in the total fragmented RNA datasets (FPKM = 0 in WT or FPKM <10 in *upf1Δ*) were excluded from footprinting score analysis. 331 uRNAs met this filtering cutoff.

Calculation of footprinting score: To calculate the footprinting score, for each RNA we determined the ratio of ribosome footprinting reads relative to reads from total fragmented RNA ($FPKM_{footprints}/FPKM_{fragments}$). Sequencing datasets were normalized for *PGK1* and *RPL41A* mRNAs to have a footprinting score of 1. To compare the translation of RNAs as measured by both the translatability score and footprinting score, for all RNAs with a score >0, a Spearman rank correlation coefficient was calculated.

Because the absence of ribosome footprints could be either due to a true failure to associate with the translation machinery, or insufficient depth of our ribosome profiling, we classify uRNAs showing sufficient evidence of ribosome association (footprinting score in WT > 0 and footprinting score in *upf1Δ* >0.1) as

showing evidence of being ribosome-bound in this assay, and make no conclusions about the absence of ribosome footprinting data. 185 uRNAs (of 331 analyzed) demonstrated ribosome association by these cutoffs.

Demarcation of ribosome-free regions: For uRNAs, the region covered by ribosome footprints was manually demarcated based on visualization of ribosome footprinting sequencing reads in the IGV genome browser (www.broadinstitute.org/igv; Robinson *et al*., 2011). Footprint occupancy regions were annotated to be representative of the ribosome footprint profile and include >75% of ribosome footprint sequencing reads. In cases where ribosome footprints fell marginally outside the uRNA boundaries, the 5' or 3' ribosome-free size was set as "0". Only those uRNAs meeting the expression cutoffs defined above in "Quantification of ribosome footprint coverage" and with sufficient evidence of ribosome association as described in "Calculation of footprinting score" were included in this analysis (n=185).

Assignment of phasing frames for mRNAs: To establish phasing of ribosome footprints along annotated mRNAs, individual sequencing datasets (ribosome profiling or total fragmented RNA) were filtered to include only reads of 27 nucleotides; these reads represent reads that were 28 nucleotides prior to trimming 1 nucleotide from the 5' end during mapping, and represent reads which predictably demonstrate ribosome occupancy (Ingolia *et al*., 2009). Using a custom script, each nucleotide position within all annotated CDS (based on the Ensembl sacCer2 Genes annotation downloaded from the UCSC genome table browser; genome.ucsc.edu/cgi-bin/hgTables?command=start) was assigned a frame as follows: a position -11 of the first nucleotide of the AUG start codon was assigned an in-frame "+1", as well every third nucleotide thereafter through -16 of the last position of the CDS; a position -10 of the first nucleotide of the AUG start codon was assigned "+2", as well as every third nucleotide thereafter through -15 of the last position of the CDS; a position -9 of the first nucleotide of the AUG start codon was assigned "+3", as well as every third nucleotide thereafter through -14 of the last position of the CDS. The sequencing datasets were cross-referenced to this nucleotide frame definition such

that the frame number corresponding to the start position of the sequencing read indicated the frame to which the read aligned; reads not aligning to a CDS were assigned a frame of “0” and not further analyzed. For each dataset, the percentage of reads assigned to each frame was calculated. Graphed data represent the average percentage of reads aligned to each frame for 4 replicates of ribosome footprinting data, and 2 replicates of fragmented RNA data, +/- SEM. Scripts were written using Python.

Assignment of phasing frames for uRNAs: Using a custom script, each nucleotide position within all uRNAs defined in this study was assigned a frame as described above, with the exception that reference points were the transcript start and end position, rather than a CDS start and stop position. Sequencing datasets containing only 27-nucleotide reads (described above) were compared to the uRNA nucleotide frame definition as above, to assign each read to a corresponding frame. The total number of 27-mer reads aligning to each uRNA was determined, combining all 4 ribosome footprinting datasets (two WT biological replicates and two *upf1Δ* biological replicates) or both fragmented RNA datasets (one WT biological replicate and one *upf1Δ* biological replicate); any uRNA with less than 10 combined 27-mer ribosome footprinting reads was discarded from this analysis. For all uRNAs with at least 10 combined ribosomal footprinting reads (n=80), the percentage of reads aligning to each frame was determined. Any uRNAs demonstrating at least 50% of reads aligning to a single frame was considered to show evidence of translation-dependent phasing, and this frame was arbitrarily set to frame +1. Graphed data represent the average percentage of reads aligning to each frame for an individual phased uRNA +/- SEM, and include a total of 61 uRNAs that demonstrated phasing. Scripts were written using Python.

Identification of sORFs: uRNAs demonstrating phasing of ribosome footprints were individually examined to determine if a putative translated ORF could be identified based on the frame to which the ribosome footprinting sequencing reads aligned. This identification required an in-frame canonical AUG start codon near the 5’ end of

ribosome footprints (often centered within the P site of the most 5' footprinting read), and the putative ORF was extended through the first in-frame stop codon following this AUG. All such putative ORFs encoding peptides of at least 10 residues constitute our class of sORFs. In some cases more than one utilized sORF was identified per uRNA. In one case (sORF-28), no canonical start codon could be identified despite strong evidence for phased ribosome footprints throughout the region; in this case, the codon within the P site of the most 5' ribosome footprint was considered the first codon for this sORF.

Data visualization: Snapshots of ribosome profiling read coverage were obtained using the IGV genome browser (www.broadinstitute.org/igv; Robinson *et al*., 2011). Metagene plots of ribosome footprint coverage were generated using ngsplot (https://code.google.com/p/ngsplot/), providing 6-column BED files of sORF or mRNA CDS regions (default parameters and -R bed -FL 30 -SE 0 -L 100).

Northern Analysis of Steady-State RNA

Whole-cell RNA was isolated using glass-bead lysis followed by a phenol/chloroform extraction (Geisler et al., 2012). 40 µg of whole-cell RNA was separated by agarose gel electophoresis on a 1.4% agarose gel with 5.92% formaldehyde. RNA was transferred to a Hybond-N nylon membrane (GE Healthcare RPN303N) and immobilized with UV crosslinking. Membranes were washed in 0.1X SSC/0.1% SDS for 1 hour at 65 °C, incubated for 1 hour in hybridization buffer (10X Denhardt's solution, 6X SSC, 0.1% SDS), and probed overnight in hybridization buffer with either 5' ^{32}P end-labelled DNA oligonucleotides or α-^{32}P CTP probes generated by asymmetric PCR (Rio et al., 2011; see Table S5) at individually optimized temperatures, to detect the RNA of interest. Excess probe was washed from membrane three times for 15 minutes with 6X SSC/ 0.1% SDS at individually optimized temperatures. Membrane was exposed to a storage phosphor screen (Molecular Dynamics), and developed using a GE Typhoon 9400 Variable Mode Imager (Amersham Biosciences).

Generation of HA-tagged sORF Strains

Chromosomal tagging of sORFs at their endogenous loci was performed using standard homologous recombination methods (Longtine et al., 1998). This approach results in incorporation of the 3xHA tag at the C-terminus of the sORF immediately followed by *ADH1* terminator sequences, and incorporation of a downstream selectable marker to facilitate screening of clones. sORFs were selected based on high expression of the uRNA, strong evidence of ribosome footprint phasing, and intergenic genomic location. Yeast genome sequences retrieved from the Saccharomyces Genome Database (www.yeastgenome.org) were used to determine gene-specific sequences to target knock-in of the 3xHA tag and selectable marker to the correct locus. These sequences were designed to insert the 3xHA tag immediately upstream of the predicted stop codon. The 3xHA tag was inserted either in-frame with the putative sORF, or out-of-frame as a control to demonstrate frame-dependent expression of the 3xHA tag. Incorporation of the 3xHA tag was confirmed by Sanger sequencing for each locus. See Table S5 for primers and plasmids used to generate strains.

Generation of FLAG-tagged sORF Plasmids

Based on the yeast genome sequences retrieved from the Saccharomyces Genome Database (www.yeastgenome.org), the genomic region encompassing several uRNAs containing putative sORFs (sORFs were selected based on high expression of the uRNA, strong evidence of ribosome footprint phasing, and intergenic genomic location) plus and minus ~500 bp was amplified by PCR with Phusion High-Fidelity DNA Polymerase (NEB M0530S), which produces a blunt-end PCR product. PCR products were ligated at 16 °C overnight into yEpLac181 previously digested with SmaI blunt-end restriction enzyme (NEB R0141S) using T4 DNA Ligase (Roche 10 481 220 001). Ligated plasmids were transformed into calcium chloride competent XL1-Blue *Escherichia coli*, plated on 2% agar Luria broth plates plus 100 µg/mL ampicillin, and individual clones screened by restriction digest and sequencing to confirm ligation of the appropriate insert. 1X FLAG tag (DYKDDDDK) was added in-frame to the C-terminus of each putative sORF, immediately upstream of the putative stop codon, using a single round of site-directed mutagenesis PCR with Phusion High-Fidelity DNA Polymerase. PCR product was treated with *Dpn*I restriction enzyme (NEB R0176S) to digest any

methylated template. Plasmid was transformed into XL1-Blue *E. coli* as above, and clones screened by sequencing to confirm the in-frame insertion of the FLAG sequence. All plasmids were transformed into WT yeast using a standard lithium acetate transformation protocol. Transformed strains were subsequently maintained and grown in selective media lacking leucine.

Protein Isolation and Western Blot Analysis

WT yeast cultures containing either 1) chromosomal 3xHA-tagged sORFs, or 2) plasmids containing uRNAs encoding a putative sORF with or without a C-terminal FLAG-tag, were grown and treated with proteasome inhibitor MG-132 as described (Liu et al., 2007), and flash frozen on dry ice. Cell pellets were heated in 5*M* urea at 95 °C for 2 minutes, then lysed by mechanical disruption with glass beads by vortexing for 5 minutes. Solution A was added to lysates (125 mM Tris-HCl, pH 6.8, 2% SDS), followed by vortexing for 1 minute and heating to 95 °C for 2 minutes. Glass beads and cellular debris were cleared from lysates by centrifugation at 13,200 RPM for 4 minutes. Equivalent OD units (A_{260}) of lysate in 1X SDS sample buffer (125 mM Tris-HCl, pH 6.8, 2% SDS, 100 mM DTT, 10% glycerol, 0.05% bromphenol blue) were separated on NuPAGE Novex 4-12% Bis-Tris gels (Life Technologies NP0321BOX) by electrophoresis in 1X MOPS SDS running buffer (50 mM MOPS, 50 mM Tris base, 0.1% SDS, 1 mM EDTA, pH 7.7). Proteins were transferred to an Immobilon-P PVDF transfer membrane (Millipore IPVH15150) in 1X western transfer buffer (25 mM Tris base, 192 mM glycine, 20% methanol) at 4 °C by electroblotting at 250 mA for 2 hours. Membrane was blocked in blocking buffer (5% milk powder in 1X TBS/0.1% Tween-20) overnight at 4 °C. Membrane was incubated with primary antibodies (rabbit polyclonal α-FLAG 1:10,000 [Sigma F7425], mouse monoclonal α-HA 1:5,000 [Covance MMS-101P], or mouse monoclonal α-Pgk1p 1:10,000 [Invitrogen 459250]) and secondary antibodies (goat α-rabbit IgG HRP 1:5000 [Pierce 31460] or goat α-mouse IgG HRP 1:5000 [Santa Cruz sc-2005]) in blocking buffer for 1 hour. Between each incubation, membrane was washed with 1X TBS/0.1% Tween-20 3 times for 15 minutes. Signal was detected by chemiluminescence using Blue Ultra Autorad film (GeneMate F-2029).

Conservation of sORF Peptides in Other Fungi

BLAST analysis: A custom database to be used for BLAST search was generated with NCBI BLAST tool formatdb V2.2.29+, with the following genomes: from the *Saccharomyces* Genome Database (yeastgenome.org): *Saccharomyces bayanus* strain S23-6C, *Saccharomyces kudravzevii* strain IFO1802, *Saccharomyces mikatae* strain IFO1815, *Saccharomyces paradoxus* strain NRRL Y-17217, *Saccharomyces pastorianus* strain Weihenstephan 34/70, and 33 *Saccharomyces cerevisiae* strains (standard laboratory strain S228C, AWRI1631, AWRI796, BY4742, BY4741, CBS7960, CEN.PK, CLIB215, CLIB324, CLIB382, EC1118, EC9-8, FL100, FostersB, FostersO, JAY291, Kyokai7, LalvinQA23, M22, PW5, RM11-1a, Sigma1278b, T7, T73, UC5, VIN13, VL3, W303, Y10, YJM269, YJM789, YPS163, ZTW1); from the NCBI Genome database (www.ncbi.nlm.nih.gov/genome): *Naumovozyma castellii* strain CBS 4309 (assembly ASM23723v1), *Candida glabrata* strain CBS138 (assembly ASM253v2), *Kluyveromyces lactis* strain NRRL Y-1140 (assembly ASM251v1), and *Ashbya gossypii* strain ATCC 10895 (assembly ASM9102v4).

Putative sORF peptides were provided as query for TBLASTN against our custom curated database. TBLASTN was run using BLAST v2.2.29+, with the E-value threshold set to 10 and all other parameters default. Results in which the subject sequence contained a termination codon that interrupted the peptide were filtered from the dataset. The number of identical residues relative to the length of the query was used to calculate percent identity. In many cases, local regions of high-identity alignment were reported that did not extend across the entire query length; for these the number of identical residues relative to the full length of the query was used to calculate percent identity. Only the hit with the highest percentage of identical residues relative to the full length of the query for each species is reported. Data is only reported for non-*S. cerevisiae* alignments. See Table S4.

PhastCons conserved elements: Conserved elements across 7 yeast species (*Saccharomyces. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli*, and *S. kluyveri*) have previously been identified using phastCons and

reported (Siepel et al., 2005), and are accessible in the UCSC genome browser (http://genome.ucsc.edu/) using the "Most Conserved" track. Using the sacCer2 *S. cerevisiae* genome assembly, we report the log-odds score of conserved elements that partially or completely overlap each putative sORF. If more than one conserved element overlapped an sORF, the log-odds score for the element displaying the highest degree of overlap is reported. See Table S4.

Calculation of the Ka/Ks ratio: The Ka/Ks ratio (ω; the relative rate of nonsynonymous to synonymous mutations along a conserved sequence), was calculated for putative sORFs using the Ka/Ks_Calculator (Zhang et al., 2006), with the method of model averaging. For each ratio, the sacCer2 reference genome sequence was compared to the nucleotide sequence corresponding to the highest-identity TBLASTN result for each species, as reported in Table S4. Ka/Ks ratios were only calculated if 1) a TBLASTN alignment was reported, and 2) the aligned nucleotide sequence corresponding to the TBLASTN peptide hit did not align 100% with the reference genome nucleotide sequence (marked as "N.D."). Only Ka/Ks ratios with a Fisher's p-value <0.05 are reported.

Analysis of Mammalian Data

Data sources: Ensembl transcript structures and annotations for the mouse July 2007 (NCBI37/mm9) genome assembly were obtained from Ensembl version 67 (http://useast.ensembl.org/info/data/ftp/). Transcript models assembled from poly(A)-selected RNA of mESCs (Guttman et al., 2010) were collected from the Scripture portal (http://www.broadinstitute.org/software/scripture/). RNA-Seq and Ribo-Seq data for shRNA-, cycloheximide- or control-treated mESCs, as well as CLIP-Seq data for UPF1 binding were downloaded from the Gene Expression Omnibus (GSE41785; Hurt et al., 2013).

RNA-Seq analysis: Paired-end directional reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and mapped to mm9 using TopHat v2.0.8 (Trapnell et al., 2009; default parameters and --solexa1.3-quals

--library-type fr-firststrand --min-anchor 5 -r 170). Gene-level expression was estimated as fragments per kilobase of exon model per million mapped fragments (FPKM) using Cufflinks v2.1.1 (Trapnell et al., 2010; default parameters and --min-frags-per-transfrag 0 --compatible-hits-norm --min-isoform-fraction 0.0) considering gene annotations from Ensembl v67 and lincRNA annotations from mESCs (Guttman et al., 2010).

Ribo-Seq analysis: Non-directional reads were quality-checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed from the 3' end using fastx_clipper (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to generate 30 nt fragments. Trimmed fragments were then aligned to rRNA annotations from Ensembl v67 using Bowtie v2.1.0 (Langmead et al., 2009; default parameters and --seedlen=23), and non-rRNA reads were then mapped to mm9 with TopHat (default parameters and --solexa1.3-quals --min-anchor 5 --no-novel-juncs) considering gene annotations from Ensembl v67 and lincRNA annotations from mESCs (Guttman et al., 2010). The fold-change in the normalized density of Ribo-Seq reads mapping along the transcript body of intergenic lincRNAs in shUPF1 v. shGFP libraries was calculated and visualized using ngsplot (https://code.google.com/p/ngsplot/; default parameters and -R genebody -F rnaseq,lincRNA -FL 30).

CLIP-Seq analysis: UPF1 CLIP-Seq reads that were previously processed (trimmed and subtracted from overlapping amplified IgG CLIP-Seq reads) and mapped uniquely to mm9 or to a splice junction database allowing 2-nt mismatches (Hurt et al., 2013) were directly analyzed for their overlap with gene bodies. Data from 3 replicate libraries (two RNAse A- and one RNAse I-treated libraries) were combined prior to analysis. Coverage along gene transcripts was computed using BEDTools (Quinlan and Hall, 2010).

Defining a set of lncRNAs: Starting with Ensembl v67 annotations, we considered only genes annotated as "lincRNA", "non-coding" or "antisense" that had no transcript annotated as "ambiguous_orf". We then incorporated putative lincRNA

transcript models (Guttman et al., 2010) from loci that are reliably active in mESCs (Guttman et al., 2011) provided that they were not annotated in Ensembl v67 already. Finally, we computed the ribosome release score (Guttman et al., 2013) with RRS (http://guttmanlab.caltech.edu/software/RRS.jar; default parameters) based on the shGFP RNA-Seq and Ribo-Seq libraries for both putative lncRNAs curated from Ensembl v67 and from the mESC lincRNA collection. Any gene with a transcript having an RRS score >10 was excluded from further analysis.

Gene expression analysis: To identify mRNAs and lncRNAs reliably expressed in mESCs, we considered only genes that are expressed at FPKM>0.1 in each shRNA-, cycloheximide- or control-treated RNA-Seq library and are expressed at FPKM>1 in at least one of these libraries. This strategy yielded 13043 and 265 mESC-expressed mRNAs and lncRNAs, respectively.

Defining a set of NMD targets: To reliably identify NMD-targeted genes, we assessed the consistency in the response across the three NMD inhibitory treatments (shUPF1-1, shUPF1-2 and CHX). Consistency was determined by taking the geometric mean of the fold-change in FPKM in treated samples versus controls (shUPF1-1 v. shGFP, shUPF1-2 v. shGFP, and CHX v. WT, each averaged over 2 replicates). Genes with a geometric mean >1.5 were designated as consistent NMD targets, based on benchmarking against a set of known mRNA isoforms targeted by NMD.

Additional bioinformatics analyses: Computational analyses were conducted using custom scripts in Python, Perl and R. Statistical tests and plots were implemented in R, and heatmaps were produced using the *gplots* R package (http://CRAN.Rproject.org/package=gplots).

SUPPLEMENTAL REFERENCES

Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S.*, et al.* (2005). The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res *33*, D459-465.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. *89*, 19.10.1-19.10.21.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451-1455.

Gietz, R.D., and Sugino, A. (1988). New yeast-Escherichia coli shuttle vectors constructed with in vitro mutagenized yeast genes lacking six-base pair restriction sites. Gene *74*, 527-534.

Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computation research in the life sciences. Genome Biol. *11*, R86.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L.*, et al.* (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295-300.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C.*, et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol *28*, 503-510.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Liu, C., Apodaca, J., Davis, L.E., and Rao, H. (2007). Proteasome inhibition in wild-type yeast *Saccharomyces cerevisiae* cells. Biotechniques, *42*, 158-162.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rio, D.C., Ares, M. Jr., Hannon, G.J., and Nilsen, T.W. (2010). RNA: A Laboratory Manual (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24-26.

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

Trapnell, C., Hendrickson, D.G., Sauvegeau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*, 46-53.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562-578.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511-515.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184-2185.