Supplementary Materials and Figures

TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code

Modi Safra[1,*], Ronit Nir[1,*], Daneyal Farouq[2], Ilya Vainberg Slutzkin[3], Schraga Schwartz[1,#]

[1] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

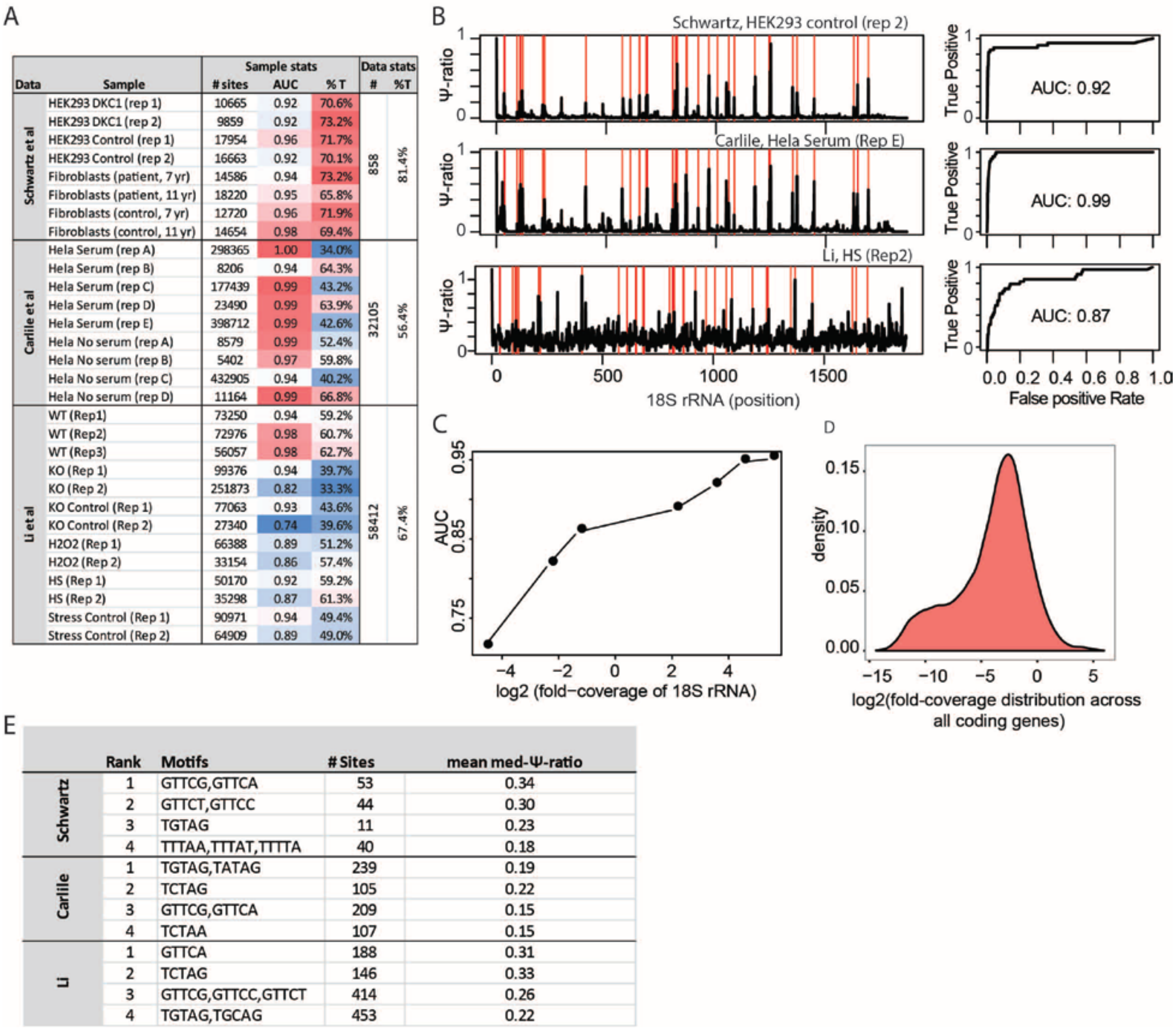
[2] Broad Institute, Cambridge 02142, Cambridge, USA

[3] Department of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot 76100, Israel

* These authors have contributed equally to this manuscript

#Corresponding author. Email: schwartz@weizmann.ac.il

Supplementary Figures:

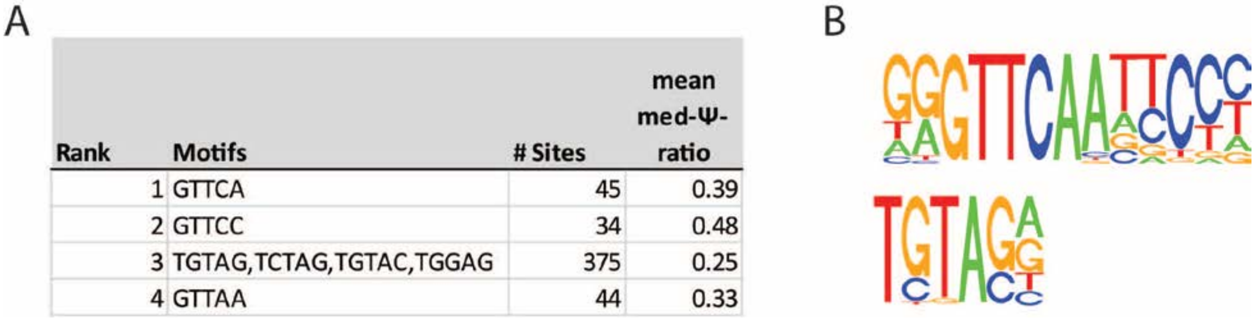


Data	Sample	Sample stats # sites	Sample stats AUC	Sample stats % T	Data stats #	Data stats %T
Schwartz et al	HEK293 DKC1 (rep 1)	10665	0.92	70.6%	858	81.4%
	HEK293 DKC1 (rep 2)	9859	0.92	73.2%		
	HEK293 Control (rep 1)	17954	0.96	71.7%		
	HEK293 Control (rep 2)	16663	0.92	70.1%		
	Fibroblasts (patient, 7 yr)	14586	0.94	73.2%		
	Fibroblasts (patient, 11 yr)	18220	0.95	65.8%		
	Fibroblasts (control, 7 yr)	12720	0.96	71.9%		
	Fibroblasts (control, 11 yr)	14654	0.98	69.4%		
Carlile et al	Hela Serum (rep A)	298365	1.00	34.0%	32105	56.4%
	Hela Serum (rep B)	8206	0.94	64.3%		
	Hela Serum (rep C)	177439	0.99	43.2%		
	Hela Serum (rep D)	23490	0.99	63.9%		
	Hela Serum (rep E)	398712	0.99	42.6%		
	Hela No serum (rep A)	8579	0.99	52.4%		
	Hela No serum (rep B)	5402	0.97	59.8%		
	Hela No serum (rep C)	432905	0.94	40.2%		
	Hela No serum (rep D)	11164	0.99	66.8%		
Li et al	WT (Rep1)	73250	0.94	59.2%	58412	67.4%
	WT (Rep2)	72976	0.98	60.7%		
	WT (Rep3)	56057	0.98	62.7%		
	KO (Rep 1)	99376	0.94	39.7%		
	KO (Rep 2)	251873	0.82	33.3%		
	KO Control (Rep 1)	77063	0.93	43.6%		
	KO Control (Rep 2)	27340	0.74	39.6%		
	H2O2 (Rep 1)	66388	0.89	51.2%		
	H2O2 (Rep 2)	33154	0.86	57.4%		
	HS (Rep 1)	50170	0.92	59.2%		
	HS (Rep 2)	35298	0.87	61.3%		
	Stress Control (Rep 1)	90971	0.94	49.4%		
	Stress Control (Rep 2)	64909	0.89	49.0%		

	Rank	Motifs	# Sites	mean med-Ψ-ratio
Schwartz	1	GTTCG,GTTCA	53	0.34
	2	GTTCT,GTTCC	44	0.30
	3	TGTAG	11	0.23
	4	TTTAA,TTTAT,TTTTA	40	0.18
Carlile	1	TGTAG,TATAG	239	0.19
	2	TCTAG	105	0.22
	3	GTTCG,GTTCA	209	0.15
	4	TCTAA	107	0.15
Li	1	GTTCA	188	0.31
	2	TCTAG	146	0.33
	3	GTTCG,GTTCC,GTTCT	414	0.26
	4	TGTAG,TGCAG	453	0.22

Figure S1: (A) QC metrics and number of sites identified at the sample and dataset levels for 30 samples from three datasets analyzed in this manuscript. **(B)** Ψ ratios for CMC-treated data across 18S rRNA, plotted for one representative sample for each experiment. Known Ψ positions in 18S rRNA are highlighted in red. ROC curves on the basis of the visualized Ψ ratios are plotted to the right of each plot. **(C)** Downsampling analysis, testing the dependency of AUC values (based on known sites in the 18S rRNA) on read depth. For this analysis, we downsampled reads from the HEK293 Control Rep 1 CMC-treated library (from the Schwartz et al dataset) to depths ranging from 2000 to 2,000,0000 reads. At each depth, we aligned the reads against the rRNA and extracted both the number of reads aligning to the

rRNA and the associated AUC. These are plotted, with the X axis depicting a log2 transformation of the number of reads divided by length of 18S rRNA (per base read depth) and the Y axis depicting the AUC. **(D)** The distribution of per-base read depth is plotted for all coding genes in the HEK293 Control Rep 1 CMC-treated library, which comprises >24 million read pairs. The overwhelming majority of genes have per-base read-depths associated with the lower-ranging AUC values in (C). **(E)** Top 4 motifs identified across each of the three datasets, using the described approach for identification of motifs based both on prevalence and pseudouridine levels (Methods).



Rank	Motifs	# Sites	mean med-Ψ-ratio
1	GTTCA	45	0.39
2	GTTCC	34	0.48
3	TGTAG,TCTAG,TGTAC,TGGAG	375	0.25
4	GTTAA	44	0.33

Figure S2: Unbiased motif detection applied to four mouse samples from Li et al dataset. **(A)** Top four motifs obtained following application of unbiased motif-detection approach to 17,548 sites from the Li et al samples. **(B)** Sequence logos depicting motif 1 (TRUB1 motif, top) and motif 3 (Pus7 motif, bottom).

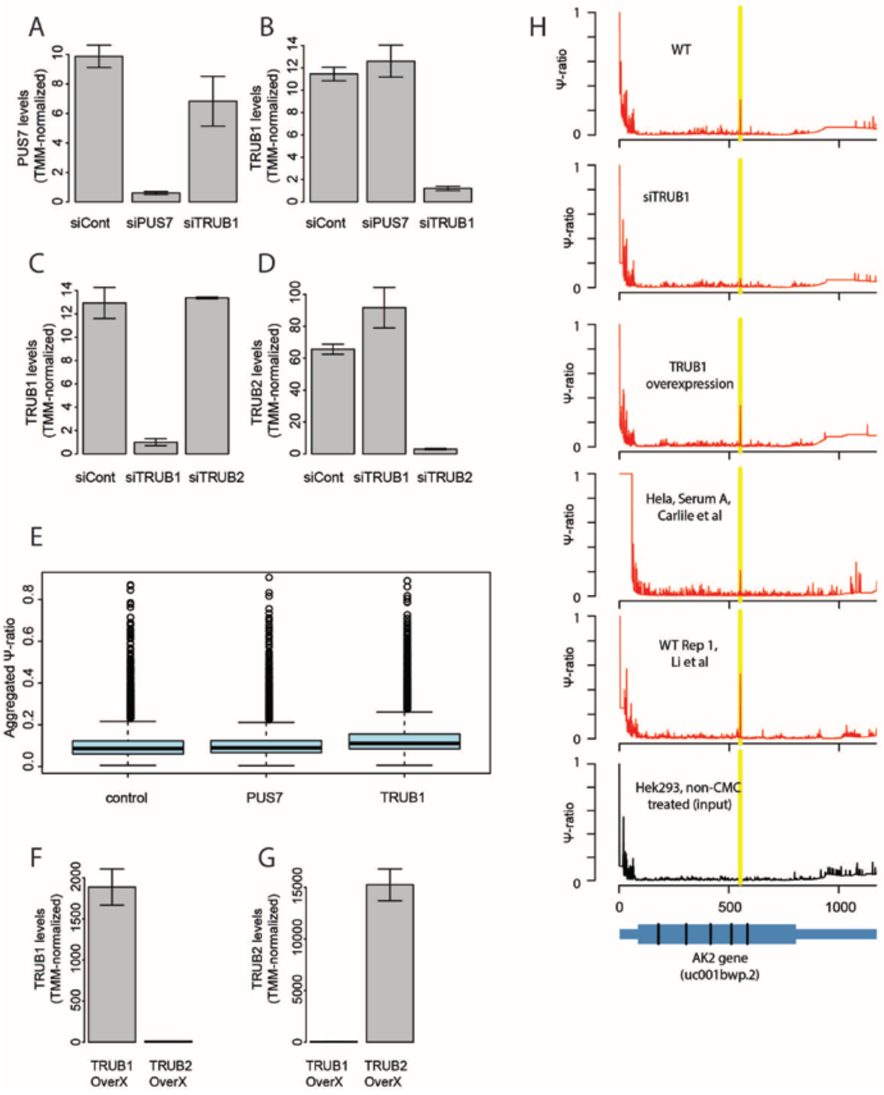


Figure S3: Knockdown and overexpression of selected PUSs. Expression levels of the genes are acquired from the sequenced input samples, and averaged across the replicates. **(A-D)** knockdown efficiencies for indicated genes (Y axis) based on the indicated perturbation (X axis). **(E)** Distribution of Ψ ratios for 5261 sites lacking both a PUS7-like motif and a TRUB1 motif following knockdown of Pus7, TRUB1 or mock knockdown in HEK293 cells. Experiments were performed at least in replicates, putative peaks were identified based on the full dataset (Methods), following which an 'aggregated Ψ-ratio' was calculated for each site, as in **Figure 6B**. **(F-G)** Expression levels, plotted as in A-D, following overexpression of TRUB1 and TRUB2. **(H)** Ψ ratios at a site harboring the TRUB1 consensus in the AK2 gene.

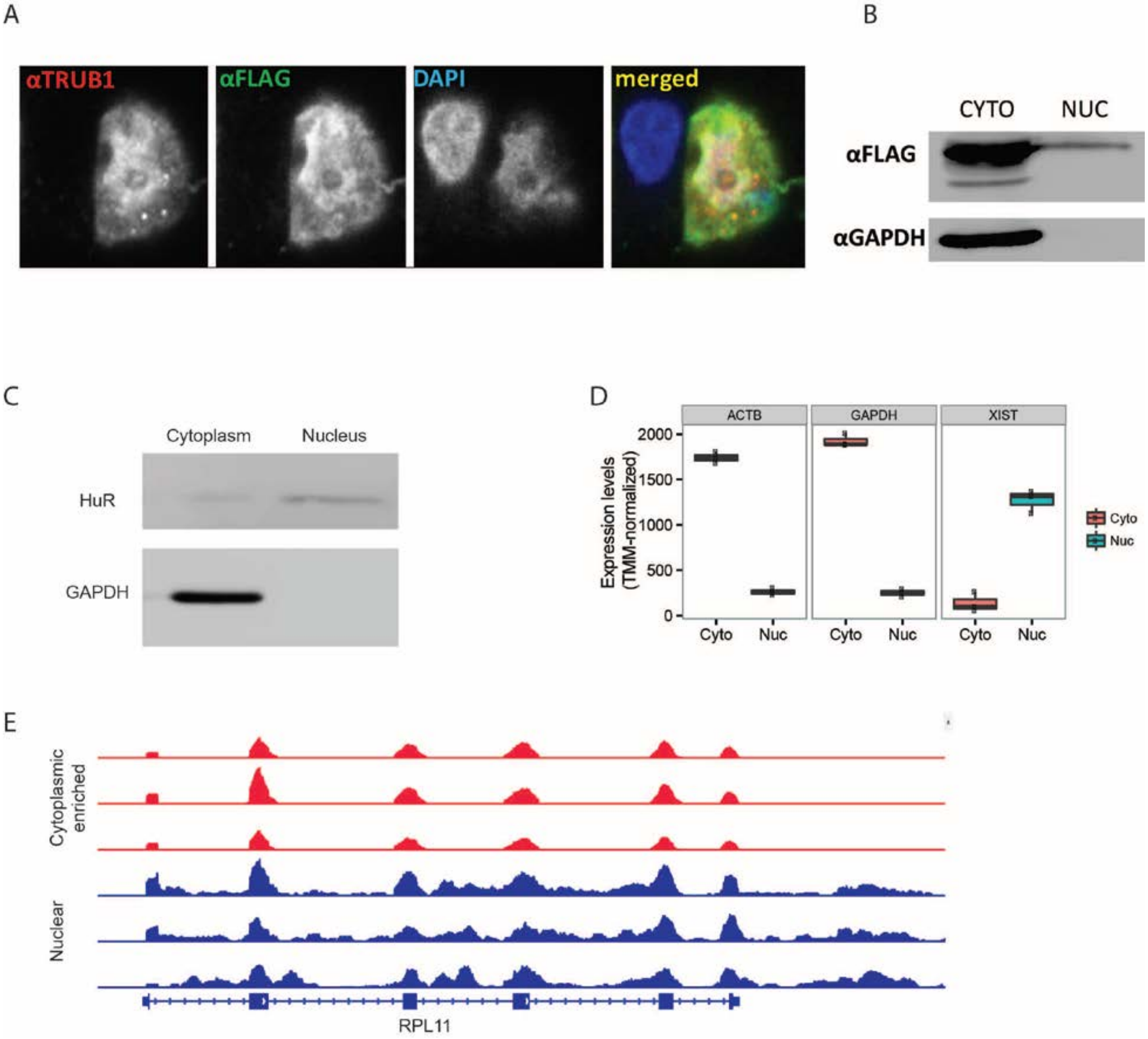


Figure S4: TRUB1 is localized to both nucleus and cytoplasm. **(A)** One slice (1.5μm) of HEK293T cells overexpressing FLAG tagged TRUB1 is shown, photographed using confocal microscope (X630). Cells were stained with αTRUB1 (red), αFLAG (green) and DAPI (blue). **(B)** Western blot using anti FLAG was carried out on nuclear and cytoplasmic fractions of HEK293T overexpressing FLAG tagged TRUB1. Anti GAPDH served as control for purity of nuclear fraction. **(C)** Western blot against cytoplasmic GAPDH and nuclear HuR reveals the former to be completely absent in the nucleus and present only in the cytoplasmic fraction, whereas the latter is enriched in the nucleus but present at low levels also in the cytoplasm. **(D)** Expression levels of nuclear XIST RNA across the two fractions; ACTB and GAPDH are presented in comparison. **(E)** Read coverage across the RPL11 locus in the nuclear and cytosolic fractions - coverage in intronic regions is abundant in the former but not in the latter.

Supplementary Methods

Read mapping

Single-end reads from the Li and Carlile datasets were mapped to the human genome (assembly hg19) using Star with default parameters (Dobin et al. 2013). A script was used to project all reads aligning to the genome to the human transcriptome. Only read pairs fully matching a transcript structure, as defined by the UCSC Known Genes annotation, were retained. For each nucleotide, we recorded the number of 'left' reads initiating at each position (corresponding to the last position traversed by reverse transcriptase) and the number of overall reads covering each position. Paired-end reads from the Schwartz et al dataset were aligned to the genome and cast onto the transcriptome as described in (Schwartz et al. 2014). All samples were further aligned directly to the 18S rRNA using Bowtie (Langmead et al. 2009); For these alignments we sampled 2,000,000 reads from each sample for the Schwartz and Carlile datasets, and 20 million reads from Li et al in which rRNA had been depleted more extensively. Alignments were processed as above to record the number of reads beginning and overlapping at each position.

Detection of putative Ψ sites within an individual sample

Putative Ψ sites were identified using an approach drawing on (Schwartz et al. 2014). Specifically: (1) For each treated or non-treated sample, a Ψ-ratio was calculated, corresponding to the number of reads beginning at the position divided by the overall number of reads covering it. A pseudocount of 1 was added to both the numerator and denominator to stabilize the ratio and avoid division by 0. (2) The Ψ-fold change was calculated as the fold change of Ψ ratios in the treated versus non-treated samples. All positions with a Ψ-ratio ≥0.1, a Ψ-fc ≥3 and with ≥7 reads beginning at the position were considered putative Ψ sites. For each such site we recorded a set of Ψ metrics, consisting of the number of reads beginning and overlapping the position in the treated and untreated sample, the Ψ-ratio in the two samples, and the Ψ-fc. Finally, we recast the transcriptomic mappings to genomic coordinates to filter out redundancies, stemming from single positions being assigned as putative Ψ sites in distinct isoforms of the same gene.

Detection of putative Ψ sites across a dataset

To integrate sites from multiple samples within a given dataset, we first generated a dataset of all sites mapping to unique positions in the transcriptome detected as putative Ψ sites in any of the samples. For each such site, we then recalculated Ψ metrics in each sample. We further calculated a median Ψ-ratio (med-Ψ-ratio) and Ψ-fc (med-Ψ-fc) across all samples. All sites that had (1) ≥7 reads beginning at them

(2) a Ψ-ratio ≥0.1 and (3) a Ψ-fc ≥4, (4) in ≥2 samples within the dataset were flagged as putative pseudouridylated sites at the dataset level.

ROC Curves for Ψ detection in 18S rRNA

For generating Ψ detection ROC curves on the basis of 18S rRNA, we calculated Ψ ratios for each site on the 18S rRNA. We then calculated sensitivity and specificity of detection of Ψ sites at each Ψ ratio threshold. For this analysis, all sites present in the MODOMICs database were considered true (Machnicka et al., 2013), all others false. ROC curves and AUC values were calculated and plotted using the ROCR package (Sing et al. 2005) in R (R Core Team).

Detection and clustering of prevalent and highly pseudouridylated sequence motifs

We developed a motif finding approach that simultaneously takes into consideration both the prevalence of a motif in a dataset and the extent to which sites harboring the motif are pseudouridylated. We implemented it using the following steps: (1) For each putative Ψ site we extracted a 5-bp sequence (pentamer) centered around the modified site, (2) We generate a pentamer frequency table, in which, for each pentamer, we calculate the number of times it occurs in the dataset, and the mean med-Ψ-ratio across all Ψ sites harboring the pentamer; pentamers with fewer than 20 occurrences in the dataset were excluded, (3) For each of the two values calculated above, we calculated the quantile within their respective distributions, and assigned each pentamer a score corresponding to the product of the two quantiles. Thus, the distribution of the score is between 0 and 1, with 1 reflecting the most prevalent and most abundant pentamer and 0 the least. We sort the pentamer frequency table based on this score (4) To allow clustering of similar motifs based on sequence similarity, we extract a position-specific scoring matrix from position -4 to position +4 surrounding the top pentamer in the frequency table, and concatenate the frequencies of each nucleotide at each position into a single vector. We correlate this vector against the corresponding vectors surrounding each other pentamer. (5) We identify all vectors whose Pearson correlation value is > Cor_{thresh}. All vectors with a mean med-Ψ-ratio within Ψ_{dist} of the top motif are collapsed together, and are eliminated from the pentamer frequency table. Cor_{thresh} and Ψ_{dist} allow control over the granularity of the clustering both in terms of sequence and in terms of their associated Ψ levels. We used $Cor_{thresh=}$ 0.7 and Ψ_{dist} =0.04. (6) Steps 4 and 5 are iteratively performed until the pentamer frequency table is emptied. Scripts implementing this clustering are provided in the Supplementary Materials.

Characterization and modelling of TRUB1 dependent pseudouridylation sites

To allow finding differentiating features between sites that acquire, or fail to acquire, TRUB1 dependent pseudouridylation in HEK293 cells, we generated a dataset of 14,189 unique GUUCNANNC sites in the human transcriptome (after filtering redundancies between isoforms). We calculated Ψ metrics for each site, as defined above. All TRUB1 containing sites identified in any of the samples forming part of the Schwartz et al dataset were considered pseudouridylated. All sites which were (1) not identified in any of the samples as putatively pseudouridylated, and (2) had a median coverage of >30 reads overlapping the site across the different samples was classified as non-pseudouridylated. The latter criterion was applied to help ensure that sufficient read coverage at the site was present to have allowed theoretical detection of Ψ. These analyses resulted in a dataset of 92 putatively pseudouridylated sites and 1587 non-pseudouridylated ones. Based on an initial exploration using feature selection approaches, we identified the following features as informative: (1) Total number of bases at positions -3, -4, -5 and -6 with complementary to positions 7, 8, 9 and 10, respectively, (2) Complementarity between positions +7 and -3 (binary variable, 1 if complementary, 0 otherwise), (3) Complementarity between positions +8 and -4, (4) Complementarity between positions +7 and -3 and positions and positions +8 and -4, (5) Predicted free energy of secondary structure for a 24-bp sequence surrounding the Ψ site, (6) presence of a pyrimidine at position +5, (7) Presence of a G at position -3. The coefficients for each of these seven variables in the logistic regression model used in this study are (in the above order) -1.0983,-2.3224,0.8118,-2.3559,-0.1389,-2.5213,-1.77, with an intercept of 9.7878.

Massively Parallel Reporter Assays

Design:

- For design of mutations altering the RNA secondary structure in **Fig. 4G**, the nucleotide opposite nucleotides -2 to -6 was replaced with the nucleotide at position -2 to -6 (given that none of the nucleotides can form base pairs with themselves).
- For disrupting the entire secondary structure in **Fig. 4H**, each of positions +6 to +10 was first assigned the nucleotide at the corresponding opposite sites (-2 to -6), to completely abolish all base-pairing. Complementarity was then systematically restored, first only between the first most loop-proximal positions (-2 and +6), then between the second most loop-proximal positions (-2 and +6; -3 and +7) etc., until all of positions -2 to -7 formed base pairs with positions +6 and +11.
- For decreasing the loop size to 6 nt, the nucleotide present at position +4 was deleted. For increasing the loop to 8 or 9 bp, ‘T’ or ‘TT’ were added following position +5.

Library cloning

The pool of sequences was cloned as 3’UTR downstream of a reporter gene in the pZDonor FC plasmid, essentially as described in (Weingarten-Gabbay et al. 2016). Specifically, the library was amplified in 5 different PCR reactions, each using 30pg as a template and 14 cycles. The primers used for the reactions were: FW TCAGTCGCCGCTGCCAGATCGCGGTACTAGTAGCAATGGGGTTCGGTATGCGC and RV TTGTTCCGCCGCTTCGCTGACTGTGGCGCGCCAACTATCGTCTCGGGGAGCCTT. The reactions were combined, cleaned by QIAquick PCR purification kit (Qiagen), and a total of 540ng was cut by AscI and SpeI restriction enzymes (FastDigest, Thermo Scientific). Following electro-elution from a gel using Midi GeBAflex tubes (GeBA, Kfar Hanagid, Israel), the library was ligated (in 1:1 ratio) to 150ng pZDonor FC plasmid digested by AscI and SpeI, using CloneDirect Rapid Ligation kit (Lucigen Corporation) and transformed into E. coli 10G electrocompetent cells (Lucigen) in 4 cuvettes. The bacteria were grown on 20 14cm plates, reaching in average 230 colonies per each sequence variant.

Cell culture for knockdown and overexpression experiments

Human HEK293 cells were plated in 6-well plates at 20% confluence. siRNAs targeting TRUB1, TRUB2, and PUS7 (Thermo Fisher, catalog numbers: s44502, s25673, s29121) were transfected using Lipofectamine RNAiMAX (Life Technologies) following the manufacturer’s protocols, with two siRNA boosts at 48 and 96 hours following transfection; As negative controls, we used Ambion® In Vivo Negative Control #1 siRNA (catalog number: 4457287). Cells were harvested at 144 hours. For overexpression, plasmids encoding full length TRUB1 and TRUB2 were obtained from DNASU (HsCD00039729 and HsCD00041101, respectively) and cloned into a Gateway destination vector, downstream of an EF1ɑpromoter. The plasmids were transfected into HEK293 cells using either Lipofectamine LTX reagent (Life technologies) or polyJet (SignaGen Laboratories) with one boost of the plasmid at 24 hours. Cells were harvested 48 hours following transfection.

Legend of Supplementary Table

Table S1: Collection of 89,898 sites collectively identified across all experiments analyzed here. Sites are ranked based on the number of samples and experiments in which they were identified, and annotated based on presence of TRUB1 or Pus7 motifs.

Table S2: Tab 1: Set of sequences that were designed for the MPRA Assay. Note that the table contains duplicate rows, to allow annotation of sites forming part of multiple distinct experimental designs. **Tab 2**: Quantification of termination rates at each of the synthetic transcripts using CMC treatment followed by

targeted reverse transcription, ligation and amplification. **Tab 3:** Control ('input') experiment, as in Tab 2 but performed in the absence of CMC treatment.

Table S3: Transcriptome wide prediction of TRUB1 binding across 14,381 non-redundant sites across the human transcriptome. Sites are sorted based on their logistic score.

References

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

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

R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org.

Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, León-Ricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES, et al. 2014. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* **159**: 148–162.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**: 3940–3941.

Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E. 2016. Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**. http://dx.doi.org/10.1126/science.aad4939.