Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells

Thomas M. Carlile¹, Maria F. Rojas-Duran¹, Boris Zinshteyn¹, Hakyung Shin¹, Kristen M. Bartoli¹ & Wendy V. Gilbert¹

Post-transcriptional modification of RNA nucleosides occurs in all living organisms. Pseudouridine, the most abundant modified nucleoside in non-coding RNAs, enhances the function of transfer RNA and ribosomal RNA by stabilizing the RNA structure. Messenger RNAs were not known to contain pseudouridine, but artificial pseudouridylation dramatically affects mRNA function—it changes the genetic code by facilitating non-canonical base pairing in the ribosome decoding centre. However, without evidence of naturally occurring mRNA pseudouridylation, its physiological relevance was unclear.

Here we present a comprehensive analysis of pseudouridylation in Saccharomyces cerevisiae and human RNAs using Pseudo-seq, a genome-wide, single-nucleotide-resolution method for pseudouridine identification. Pseudo-seq accurately identifies known modification sites as well as many novel sites in non-coding RNAs, and reveals hundreds of pseudouridylation sites in mRNAs. Genetic analysis allowed us to assign most of the new modification sites to one of seven conserved pseudouridine synthases, Pus1–4, 6, 7 and 9. Notably, the majority of pseudouridines in mRNAs are regulated in response to environmental signals, such as nutrient deprivation in yeast and serum starvation in human cells. These results suggest a mechanism for the rapid and regulated rewiring of the genetic code through inducible mRNA modifications. Our findings reveal unanticipated roles for pseudouridylation and provide a resource for identifying the targets of pseudouridine synthases implicated in human disease.

Although more than 100 classes of RNA modifications have been characterized, primarily in rRNA and tRNA, only three modified nucleotides have been identified within the coding sequences of mRNA, N6-methyladenosine (m6A), 5-methylcytosine (m5C) and inosine (I). To define the global landscape of RNA pseudouridylation in vivo and determine whether mRNAs contain pseudouridine (Ψ), we developed a high-throughput method to identify Ψ in the transcriptome with single-nucleotide resolution. Ψ can be selectively modified with N-cyclohexyl-N′-(2-morpholinoethyl)-carbodiimide metho-p-toluenesulphonate (CMC) to generate a block to reverse transcriptase one nucleotide 3′ to the pseudouridylated site. We exploited this chemistry to determine the locations of Ψ using next-generation sequencing (Fig. 1a; see Methods). Mock-treated (−CMC) RNA fragments were processed in parallel to identify pseudouridine-independent reverse transcription stops.

Using Pseudo-seq and stringent Ψ-calling criteria, we identified 42/51 known Ψs in rRNA and small nuclear RNA (Supplementary Table 1) with an observed false positive rate of 0.1%. The estimated false discovery rate (FDR) ranges from approximately 5% for highly expressed genes per million (RPM). c, A metPsi plot of mean normalized reads (left axis) for +CMC (orange) and −CMC (blue) libraries. The number of Ψs at each position in the metPsi window is indicated (black, right axis). CMC-dependent reverse transcription stops are found 1 nt 3′ of known Ψs.

A receiver operating characteristic curve of the Pseudo-seq signal for all known Ψs at each position in the metPsi window is indicated (black, right axis). CMC-dependent reverse transcription stops are found 1 nt 3′ of known Ψs.
to approximately 12.5% for lowly expressed genes. (Fig. 1b–d and Methods). Six out of nine false negatives were due to ‘shadowing’ from reverse transcription stops 3’ of the Ψ (for example, 25S-Ψ2258 was not detected upstream of Ψ2260). We also identified many Ψs in tRNA, all of which occurred at known positions (Supplementary Table 2). We verified the single-nucleotide resolution of Pseudo-seq by profiling four small nucleolar RNA deletion mutants that eliminate pseudouridylation of nine specific rRNA and snRNA target sites (Extended Data Fig. 6, chrVII: 939458–939671, RNase MRP RNA (NME1), chrXVII: 585585–585925) and an H/ACA snoRNA (snR37, chrX: 228090–228475). Importantly, we ruled out differences in mRNA expression as an explanation for condition-dependent differences in Ψ detection (Extended Data Fig. 5). Thus, the process of mRNA pseudouridylation is regulated in response to environmental cues.

Yeast non-coding RNAs (ncRNA) have been extensively characterized for post-transcriptional modifications. Nevertheless, we identified 74 novel pseudouridylated sites in ncRNAs (Supplementary Table 4). A few, like Ψ274 in the RNase MRP RNA (NME1) (Fig. 2c), were constitutively modified, while most, including the previously described Ψ56 and Ψ93 in U2 snoRNA34, were induced during post-diauxic growth (Extended Data Fig. 6a). snRNAs were notably enriched among ncRNA classes with regulated pseudouridines: 19/29 H/ACA and 14/47 C/D snoRNAs showed one or more sites specifically modified in cells grown to high density (Fig. 2d, Extended Data Fig. 6b, c).

Pseudouridylation of rRNA sites changed very little: only one site, 25S-Ψ2314, changed more than twofold. However, owing to the stability of rRNA and the greatly reduced rate of ribosome synthesis during post-diauxic growth, we cannot rule out production of a minority population of differentially modified ribosomes in dense cultures.

We next sought to define the molecular basis for targeting of novel mRNA and ncRNA sites for pseudouridylation. Ψs in rRNA, snRNA and tRNA are produced by two classes of enzymes with distinct modes of target recognition. The first class, which includes yeast Cbf5 and human dyskerin, associates with H/ACA snoRNAs to direct pseudouridylation of sites that base pair with the snoRNA guide sequences, while the second class recognizes its targets without the aid of an RNA guide. We computationally identified 157 unique sites in mRNAs containing perfect matches to canonical snoRNA targets (Supplementary Table 5). When these potential pseudouridylation sites were considered in aggregate, statistically significant pseudouridylation was detected (Fig. 3a, Extended Data Fig. 7a, b), which increased with the number of base pairs to the snoRNA guide sequence and was specific to post-diauxic growth (Extended Data Fig. 7c, d). However, only three such sites passed our threshold for reproducible Pseudo-seq hits even in transcripts with sparse read coverage, we required reproducibility in 10/14 independent experiments (Extended Data Fig. 3b, c). Notably, we found that many mRNAs contain Ψ (Fig. 2a). In total, we conservatively identified 260 Ψs in 238 protein-coding transcripts (Supplementary Table 3). Relaxing our criteria to include Ψs detected in 9/14 experiments, a category that includes the known Ψ56 in U2 snoRNA, increased the number of candidate mRNA Ψs to 466. We established a rough detectability threshold by determining the lowest observed expression level of genes having sufficient reads for reproducible Ψ calling: 5,278 genes passed the cutoff. Thus, it is unlikely that there are substantially more mRNA Ψs to be discovered in post-diauxic yeast. We conclude that mRNA pseudouridines are relatively scarce. Ψs were found in 5’ transcript leaders (5’ TLs), coding sequences (CDS), and 3’ untranslated regions (3’ UTRs) with an underrepresentation of Ψs in 3’ UTRs (P = 10^-4, hypergeometric test) (Fig. 2b). GUA valine codons were the most frequently modified, suggesting the existence of a sequence-specific mechanism for mRNA pseudouridylation (Extended Data Fig. 4).

We investigated the potential for regulation of mRNA pseudouridylation by comparing two cellular conditions with substantial differences in gene expression and physiology: log phase and post-diauxic growth. Remarkably, most mRNA Ψs were regulated: 42% of the sites modified during post-diauxic growth were not detectably modified in log phase, whereas other sites, such as CDC33 Ψ286, were much more extensively modified during exponential growth. Moreover, of the 150 modified sites detected in both log phase and post-diauxic growth, 62 showed >twofold changes in peak height between conditions indicating growth-state-dependent changes in the extent of mRNA modification (Fig. 2a and Supplementary Table 3). Importantly, we ruled out differences in mRNA expression as an explanation for condition-dependent differences in Ψ detection (Extended Data Fig. 5). Thus, the process of mRNA pseudouridylation is regulated in response to environmental cues.

Figure 2 | Yeast mRNAs and ncRNAs are inducibly pseudouridylated.

a, c, CMC-dependent peaks of reads are indicated with a dashed red line. The median Pseudo-seq peak heights in each condition are given ± s.d.; negative peak values occur when the reads in the −CMC library exceed those in the +CMC library. Traces are representative of four wild-type biological replicates. a, Pseudo-seq reads in RPS28B (chr XII: 673163–673336), MRPS12 (chr XII: 694489–694736) and CDC33 (chrXV: 50560–60875). b, Summary of locations of Ψs within mRNA features. CDS, coding sequence. c, Pseudo-seq reads in U5 snRNA (snR7-L, chr VII: 939458–939671), RNase MRP RNA (NME1, chrXIV: 585585–585925) and an H/ACA snoRNA (snR37, chrX: 228090–228475). d, Summary of novel Ψs identified in ncRNA.
for Ψ calling on their own (Extended Data Fig. 7e). Thus, it is probable that many additional mRNAs are pseudouridylated at a low level and our estimate of 260 mRNA Ψs represents a conservative minimum.

Because most pseudouridylated sites showed no significant complementarity to snoRNA guide sequences, we next investigated whether snoRNA-independent pseudouridine synthases are responsible for modifying sites in mRNAs. Yeast has nine pseudouridine synthase (PUS) genes, all of which are expressed in both log phase and post-diauxic growth. We profiled the eight viable PUS deletion strains (pusΔ) grown to high density and identified mRNA targets for each Pus protein, with the exception of Pus5 whose only known target is the 21S mitochondrial rRNA22 (Fig. 3b, Extended Data Fig. 8a, b and Supplementary Table 6). The largest number of mRNA and novel ncRNA Ψs could be assigned to Pus1, a member of the TruA family that constitutively modifies multiple positions in cytoplasmic tRNAs and one position in U2 snRNA by a mode of target recognition that is incompletely defined. Whereas known Pus1-dependent tRNA targets showed constitutive pseudouridylation as expected, most of the mRNA targets showed increased modification during post-diauxic growth (Extended Data Fig. 8c, Supplementary Table 3). The mRNA targets of Pus1 showed little similarity at the primary sequence level, consistent with the proposed structure-dependent mode of target recognition by this enzyme (Fig. 3c, Extended Data Fig. 8d)31, while Pus2, a close parologue of Pus1, had 14 mRNA targets with a weak sequence consensus distinct from Pus1 (Fig. 3d, Extended Data Fig. 8e). Intriguingly, the Pus1 targets included seven genes encoding five proteins of the large ribosomal subunit, a significant enrichment (P = 0.025). Our comprehensive pseudouridine profiling more than doubles the number of known substrates of Pus1 and Pus2, identifies unanticipated mRNA targets, and provides the first demonstration of regulated pseudouridylation by these enzymes.

Unlike Pus1 and Pus2, the mRNA targets of Pus4 and Pus7 contained clear consensus sites in agreement with the known sequence requirements for these enzymes to modify their canonical tRNA targets, UGΨAR for Pus7 and GUΨCNANNC for Pus4 (Fig. 3e–g, Extended Data Fig. 8f–h, Supplementary Table 3). The mRNA targets of Pus4 and Pus7 show strong similarity to snoRNA guide sequences, we next investigated whether snoRNA-dependent modes of target recognition by these enzymes are responsible for regulating pseudouridylation in mRNAs and the distribution of Ψs among ncRNA types are indicated. NC, noncoding sequence. 

**Figure 3 | Mechanisms of mRNA pseudouridylation.** a, MetaPsi plot of mean normalized reads for computationally predicted snoRNA-dependent targets in mRNA from high density cultures (orange), log phase cultures (blue), + CMC (solid) and − CMC (dashed). Indicated are the predicted snoRNA target site (black, dashed), and the expected peak of CMC-dependent reads (black, dotted). Reads were pooled from four wild-type biological replicate libraries. b, Summary of Ψs identified by Pseudo-seq as PUS-dependent. The few PUS6- and PUS9-dependent Ψs are not shown. The locations of Ψs within mRNAs and the distribution of Ψs among ncRNA types are indicated. NC, noncoding sequence. c–f, Sequence motifs surrounding PUS1 (c), PUS2 (d), PUS7 (e) and PUS4-dependent (f) Ψs in mRNAs, generated by WebLogo 3.3. g, PUS7-dependent Pseudo-seq reads in MRPL36 (chrII: 484301–484400).

**Figure 4 | Regulated pseudouridylation of human RNAs.** Pseudo-seq was performed on HeLa cells grown in the presence or absence of serum for 24 h. a, b, CMC-dependent peaks of reads are indicated with a dashed red line. The median Pseudo-seq peak heights in each condition are given ± s.d. Traces are representative of n = 4 (– serum), and n = 5 (+ serum) biological replicates. Genome browser views represent spliced transcripts. a, Pseudo-seq reads from RPL19 (12–460), and ATP5E (154–437). b, Summary of locations of Ψs within mRNA features. c, Pseudo-seq reads from MALAT1 (5081–5636) and RN7SK (142–307). d, Summary of novel Ψs identified in ncRNA.
As the pseudouridine synthases that modify yeast mRNAs are conserved throughout eukaryotes, we investigated whether regulated mRNA pseudouridylation also occurs in mammalian cells. Human cervical carcinoma (HeLa) cells were profiled during normal proliferation and 24 h after serum starvation. Pseudo-seq detected known pseudouridines with good sensitivity and specificity (Supplementary Table 7, Extended Data Fig. 9a–c). By restricting our analysis to more highly expressed genes and requiring reproducibility in four independent biological replicates, we conservatively identified 96 Ψs in 89 human mRNAs (Supplementary Table 8). As in yeast, some Ψ modifications in human mRNAs were regulated by cellular growth state (Fig. 4a, Extended Data Fig. 10a, b), and modified sites were found throughout the transcript (Fig. 4b). We also discovered novel Ψs in human ncRNAs, including 4 previously unknown sites in rRNA (Extended Data Fig. 10c, Supplementary Table 9) and sites in long non-coding (Inc-) sn- and snoRNAs (Fig. 4c, d). Thus, the Pseudo-seq approach is broadly applicable to diverse organisms and growth states. Moreover, the phenomenon of regulated mRNA pseudouridylation is conserved from yeast to humans.

In summary, Pseudo-seq provides comprehensive analysis of mRNA pseudouridylation with single-nucleotide resolution and reveals that endogenous mRNAs are specifically pseudouridylated in a highly regulated manner in yeast and human cells. Because Ψ stabilizes RNA structure, mRNA pseudouridylation could alter translation initiation efficiency26,27, ribosome pausing28, RNA localization29 and regulation by RNA interference29, to name a few aspects of mRNA metabolism known to be affected by RNA structure, although we cannot exclude the possibility that many instances of mRNA pseudouridylation may be functionally silent. However, given recent evidence that pseudouridine profoundly affects decoding by ribosomes from diverse organisms2, our results also raise the possibility of widespread regulated rewiring of the genetic code. Finally, this work suggests that diseases associated with mutations in pseudouridine synthases, including mitochondrial myopathy and sideroblastic anemia (MLASA)31, dyskeratosis congenita32 and lung cancer33, could be due to misregulation of mRNA targets.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 May; accepted 27 August 2014.

Published online 5 September 2014.

1. Davis, F. F. & Allen, W. F. Ribonucleic acids from yeast which contain a fifth nucleotide. J. Biol. Chem. 227, 907–915 (1957).
2. Arnez, J. G. & Steitz, T. A. Crystal structure of unmodified tRNA23 complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudouridines in stabilization of RNA structure. Biochemistry 33, 7560–7567 (1994).
3. Charette, M. & Gray, M. W. Pseudouridine in RNA: what, where, how, and why. JUBMB Life 49, 341–351 (2000).
4. Davis, D. R. & Poulter, C. D. [4H]-23In NMR studies of Escherichia coli tRNA23 from hisT mutants: a structural role for pseudouridine. Biochemistry 30, 4223–4231 (1991).
5. Davis, D. R., Veltri, C. A. & Nielsen, L. An RNA model system for investigation of pseudouridine stabilization of the codon-anticodon interaction in tRNA23, tRNA23 and tRNA23. Biochemistry 30, 1121–1133 (1991).
6. Hall, K. B. & Mclaughlin, L. W. Properties of a U1/mRNA 5′ splice site duplex containing pseudouridine as measured by thermodynamic and NMR methods. Biochemistry 30, 1795–1801 (1991).
7. Hudson, G. A., Bloomingdale, R. & Znosko, B. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligonucleotides. RNA 19, 1474–1482 (2013).
8. Varian, C. S. et al. Structural and functional roles of N1- and N3-protons of psi at RNA's position 39. Nucleic Acids Res. 27, 3543–3549 (1999).
9. Fernandez, I. S. et al. Unusual base pairing during the decoding of a stop codon by the ribosome. Nature 500, 107–110 (2013).
10. Karjalil, J. & Yu, Y.-T. Convergence of sense codons into sense codons by targeted pseudouridylation. Nature 474, 395–398 (2011).
11. Bykovskaya, Y., Kasas, K., Mengesha, E., Inbal, A. & Fischer-Ghodsian, N. Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). Am. J. Hum. Genet. 74, 1303–1308 (2004).
12. Heiss, N. S. et al. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucelolar functions. Nature Genet. 19, 32–38 (1998).
13. Mei, Y.-P. et al. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. Oncogene 31, 2794–2804 (2012).
14. Cantara, W. A. et al. The RNA Data Modification Database, RNASMB: 2011 update. Nucleic Acids Res. 39, D195–D201 (2011).
15. Chen, L. Characterization and comparison of human nuclear and cytosolic eIF3s. Proc. Natl. Acad. Sci. USA 110, E2741–E2747 (2013).
16. Dominicisini, D. et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 485, 201–206 (2012).
17. Li, J. B. et al. Genome-wide identification of human RNA editing sites by parallel RNA capturing and sequencing. Science 324, 1210–1213 (2009).
18. Meyer, K. D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell 149, 1635–1646 (2012).
19. Squires, J. E. et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. 40, 12503–12532 (2012).
20. Bakin, A. & Ofengand, J. Four newly located pseudouridilate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center; analysis of the application of a new sequencing technique. Biochemistry 32, 9756–9762 (1993).
21. Wu, X., Xiao, M., Yang, C. & Yu, Y.-T. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNPs. EMBO J. 30, 79–89 (2011).
22. Ansmant, I., Massenet, S., Grosjean, H., Motorin, Y. & Branlant, C. Identification of the Saccharomycyes cerevisiae pseudouridine synthase responsible for formation of Ψ2391 in 21S mitochondrial ribosomal RNA. Nucleic Acids Res. 28, 1941–1946 (2000).
23. Arluison, V., Buckle, M. & Grosjean, H. Pseudouridine synthetase Pus1 of Saccharomyces cerevisiae: kinetic characterisation, RNA structural requirement and real-time analysis of its complex with tRNA. J. Mol. Biol. 289, 491–502 (1999).
24. Becker, H. F., Motorin, Y., Sissler, M., Florentz, C. & Grosjean, H. Major identity determinants for enzymatic formation of ribothymidine and pseudouridine in the T-loop of yeast tRNAs. J. Mol. Biol. 274, 505–518 (1997).
25. Behm-Ansmant, I. et al. The Saccharomyces cerevisiae U2 snRNA:pseudouridine-synthase Pus7p is a novel multisite-multiprotein-substrate RNA:Ψ-synthase also acting on tRNAs. RNA 9, 1371–1382 (2003).
26. Kudla, G., Murray, A. W., Tollervey, D. & Potkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. Science 324, 255–258 (2009).
27. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Potkin, J. B. Rate-limiting steps in yeast tRNA translation. Cell 158, 1589–1601 (2014).
28. Somogyi, P., Jenner, A. J., Brierley, I. & Inglis, S. C. Ribosomal pausing during translation of an RNA pseudoknot. Mol. Cell. Biol. 13, 6931–6940 (1993).
29. Jambhekar, A. & Derisi, J. L. cis-acting determinants of asymmetric, cytoplasmic RNA transport. RNA 13, 625–642 (2007).
30. Tan, Y. et al. Tiling genomes of pathogenic viruses identifies potent antiviral shRNAs and reveals a role for secondary structure in shRNA efficacy. Proc. Natl. Acad. Sci. USA 109, 869–874 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank I. Cheeseman, C. Burge, R. RajBhandary, D. Bartel and members of the Gilbert laboratory for comments and discussion. The sequencing was performed at the BioMicro Center under the direction of S. Levine. This work was supported by grants from The American Cancer Society—Robby Sue Mudd Kidney Cancer Research Scholar Grant (RSG-13-396-01-MRC) and the National Institutes of Health (GM094303, GM081399) to W.V.G. T.M.C. was supported by the American Cancer Society. This work was supported in part by the NIH Pre-Doctoral Training Grant T32GM007287.

Author Contributions T.M.C. and W.V.G. conceived and designed the experiments. T.M.C., M.F., D.-H., K.M.B. and W.V.G. performed the experiments. T.M.C., B.Z. and H.S. performed the bioinformatic analyses. T.M.C. and W.V.G. interpreted the results and wrote the paper with input from all authors.

Author Information Data have been deposited in NCBI’s Gene Expression Omnibus (GEO), accession number GSE58200. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.V.G. (wgilbert@mit.edu)
reads were allowed for generation of receiver operating characteristic curves and MetaPsi plots, but were excluded for \( \Psi \) predictions.

**Identification of \( \Psi \).** The yeast transcriptome (downloaded from SGD on 9 February 2011) with annotated 5’ and 3’ UTRs was used to identify new sites of pseudouridylayation\(^3\). Where annotated 5’ and 3’ UTRs were not available, median UTR lengths were used. To identify new sites of pseudouridylation in HeLa cells the human transcriptome described above was used. For a given \(+/-\)/CMC library pair, the \(-\)CMC libraries were first scaled to the size of the \(+\)CMC libraries. Peak values were calculated for each position 1 nt of an U (peak position) in all features with an average per nucleotide read coverage greater than a specified read cutoff:

\[
\text{peak} = ws \times \frac{r}{wt + wr}
\]

Where \( r \) and \( r' \) indicate the number of reads whose 5’ ends map to the position being examined in the \(+\)CMC and \(-\)CMC libraries, respectively, \( wt \) and \( wr \) represent the number of positions whose 5’ ends map to a window centered at the position being examined (exclusive of reads at that position), and \( ws \) specifies the size of this window (exclusive of that position). Sites with peak positions greater than a specified peak cutoff were flagged as potential \( \Psi \). To filter out false positives reproductibility of peak calling in a certain number of libraries was required.

Window size (ws) was set to 150 for all analyses. Only features surpassing an average reads/nt threshold (read cutoff) were considered. For high density yeast, the read cutoff was set to 0.0, the peak cutoff was set to 1.0, and reproducibility was required in 10 of 14 libraries. For log phase yeast, the peak values from log phase data were calculated for the high density identified \( \Psi \). For both serum fed and serum starved HeLa cells, the read cutoff was set to 1.0, the peak cutoff was set to 2.0, and reproducibility was required in 4 of 14 libraries. A subset of called \( \Psi \) in HeLa cells came from very narrow (<20 nt) regions of uniquely mapping reads. These calls were considered unreliable and were removed.

**MetaPsi plots and ROC curves.** For a given \(+/-\)/CMC strain, the reads at each position in a 51-nt window centered at the \( \Psi \) were normalized to the average reads per nucleotide within the window. These windows of normalized reads were then averaged for all known \( \Psi \) in yeast RNA and U2 or human RNA and snoRNA, yielding a metaPsi. Given the close spacing of these features, the number of \( \Psi \) at each position in the metaPsi window was also plotted.

To generate receiver operating characteristic (ROC) curves for a given \(+/-\)/CMC library pair, the \(-\)CMC libraries were first scaled to the size of the \(+\)CMC libraries, and peak values were calculated (see above) for each position 1 nt of a \( \Psi \) or U in the features above. Additionally, a \(-\)CMC peak value was determined:

\[
\text{peak} = ws \times \frac{1}{ws + wr}
\]

Parameters are as defined earlier. A range of 10,000 equally spaced cutoff scores were chosen spanning the range of observed peak values. At each cutoff score, the true positive and false positive rates were calculated, and receiver operating characteristic areas were determined.

**Estimation of false discovery rate.** The lower bound for FDR was estimated from the observed FDR for the RNA. The upper bound of FDR for lowly expressed genes was estimated by randomly down-sampling reads in the rRNA and U2 snRNA to a level of coverage comparable to that of lowly expressed mRNAs with \( \Psi \). These randomizations were performed 14 times followed by \( \Psi \) calling on down-sampled libraries as described above. This number should be considered a rough estimate because the ribosomal RNA may not provide a perfect basis for estimating the FDR of \( \Psi \) in mRNA. The observed number of false positives in the rRNA and snoRNA under the criteria used to call \( \Psi \) in mRNA was two incorrect \( \Psi \) calls in 1905 U residues (0.1%).

**Defining \( \Psi \) regulation and factor-dependence.** To determine if a given \( \Psi \) was condition dependent the median peak values between two conditions were compared. For yeast only wild-type peak values were included. A twofold or greater change in expression was defined as a significant difference.

\( \Psi \)s were identified as Pus-dependent if the peak heights in both biological replicates of a given \( \Psi \) strain were less than 25% of the median peak height for that \( \Psi \) across all libraries. For a given \( \Psi \) to be considered PUS-dependent, we required at least one replicate for a given \( \Psi \) strain to have sufficient reads in the 150-nt window surrounding the \( \Psi \) to be greater than 25% of the median reads in that window for all libraries.

**snoRNA target site predictions and analysis.** To identify potential sites of pseudouridylayation within yeast mRNAs, the yeast transcriptome (described above) was scanned for sites that perfectly match the known target sequences of all yeast Box HAACA snoRNAs, allowing mismatches at bases that are unpaired in known target sites\(^4\).

For the analysis presented in Extended Data Fig. 7a, b, 10,000 randomizations were performed. For each trial, a random \( U \) was chosen for each non-repetitive computationally predicted snoRNA target, and was matched to the same gene, and
transcript feature (5’ UTR, CDS, 3’ UTR). Each randomized set of Us was used to generate a metaPsi from the pooled reads of four libraries, and the differences in mean normalized reads between the +CMC and −CMC pools at the peak position were calculated. The distributions of these values were plotted in a histogram, and compared to the values for the computationally predicted snoRNA target sites.

**Plotting and other analyses.** Sequences and structures of tRNAs were obtained from tRNAdb, and 'Ψ' locations were previously published. Motifs were generated using WebLogo 3.3 using default settings, and the modified position was changed to a 'Ψ' after logo generation. UCSC genome browser was used to generate plots of rpm data, and matplotlib was used to generate the remainder of graphs.

RPKM (reads per kilobase of exon sequence per million exon reads) were calculated from −CMC libraries. GO analysis was performed using the YeastMine feature of SGD (http://yeastmine.yeastgenome.org/). The set of genes whose coverage was sufficient to reproducibly call 'Ψ's was used as a background set. These 5,278 genes had average RPKMs in post-diauxic wild type cultures of ≈9.25, the level of expression of the lowest expressed pseudouridylated mRNA called by our algorithm.

31. Longtine, M. S. et al. Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14**, 953–961 (1998).
32. Winzeler, E. A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
33. Collart, M. A. & Oliviero, S. Preparation of yeast RNA. *Curr. Prot. Mol. Biol.* Ch. 13, Unit–13.12 (2001).
34. Sambrook, J. & Russell, D. W. *Molecular Cloning* (Cold Spring Harbor Laboratory Press, 2001).
35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
36. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
37. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
39. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
40. Piekna-Przybylska, D., Decatur, W. A. & Fournier, M. J. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA* **13**, 305–312 (2007).
41. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
42. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
43. Jühling, F. et al. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **37**, D159–D162 (2009).
44. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
Extended Data Figure 1 | Detection of specific snoRNA target sites by Pseudo-seq. Pseudo-seq was performed on wild-type \((n = 4)\), \(\text{snr}37\Delta (n = 2)\), \(\text{snr}81\Delta (n = 2)\), \(\text{snr}43\Delta (n = 2)\) and \(\text{snr}49\Delta (n = 2)\) yeast strains. Cultures were harvested at high density \((a, b)\) or log phase \((c, d)\). Y\(s\) dependent on the deleted snoRNA are indicated in red. CMC-dependent peaks of reads are indicated with dashed red lines. Traces are representative of indicated number of biological replicates. 

**a**, Pseudo-seq reads in \(\text{RDN}25-1\) (chrXII: 452221–452270) showing \(\text{SNR}37\)-dependence of 25S-\(\Psi\)2944. 

**b**, Pseudo-seq reads in \(\text{RDN}25-1\) (chrXII: 454111–454160, left), and U2 snRNA (LSR1, chrII: 681791–681840, right) showing \(\text{SNR}81\)-dependence of 25S-\(\Psi\)1052 and U2-\(\Psi\)42. 

**c**, Pseudo-seq reads in \(\text{RDN}58-1\) (chrXII: 455466–455515) showing \(\text{SNR}43\)-dependence of 5.8S-\(\Psi\)73. SNR43-dependent 25S-\(\Psi\)960 was not consistently detected in wild type owing to an overlapping CMC-independent reverse transcription stop. 

**d**, Pseudo-seq reads in \(\text{RDN}18-1\) (chrXII: 457361–457610) showing \(\text{SNR}49\)-dependence of 18S-\(\Psi\)302, 18S-\(\Psi\)211, and 18S-\(\Psi\)120. 25S-\(\Psi\)990 was also detected as \(\text{SNR}49\)-dependent (data not shown).
Extended Data Figure 2 | Technical variations of Pseudo-seq give similar results. a–d, MetaPsi plots (left), and ROC curves (right) for various library prep conditions $n = 1$ for each condition. CMC-treated samples (solid) and mock-treated samples (dashed) are indicated. a, Comparison of AMV-RT (orange) and SuperScript III (blue) (0.2 M CMC; 115–130 nt, 100–115 nt fragments respectively). b, Comparison of 115–130 nt (orange), and 130–145 nt (blue) RNA fragment sizes (AMV-RT; 0.2 M CMC). c, Comparison of 0.2 M CMC (blue), and 0.4 M CMC (orange) (AMV-RT; 115–130 nt RNA). d, Comparison of shorter (orange) and longer (blue) truncated reverse transcription fragment sizes (AMV-RT; 115–130 nt RNA; 0.2 M CMC).
Extended Data Figure 3 | Identification of pseudouridines in lowly expressed genes using multiple replicates. a, Growth curves for wild-type yeast were grown in YPD. An $A_{600\text{ nm}}$ of 12 is indicated by the horizontal dotted line.

b, c, Pseudo-seq was performed on polyA+ RNA isolated from high-density wild-type yeast strains. CMC-dependent peaks of reads are indicated with dashed red lines. b, Pseudo-seq reads from $n = 4$ biological replicates in a $\text{CDC39}(\text{chrIII:} 286226–286445, 12.3\text{ average RPKMs})$, and c, $\text{IQG1}(\text{chrXVI:} 90655–90955, 12.4\text{ average RPKMs})$ showing CDC39-$\Psi$6223 and IQG1-$\Psi$4367, respectively.
Extended Data Figure 4 | Codons affected by mRNA pseudouridylation.
Pseudouridylation of mRNA preferentially affects GUA codons. Numbers of pseudouridines observed at the first (dark blue), second (blue), and third positions (light blue) of each codon are indicated.
Extended Data Figure 5 | Expression levels minimally affect identification of yeast mRNAs displaying regulated pseudouridylation. A plot of log-transformed average RPKMs in high-density versus log-phase yeast for all coding genes with a $\Psi$ identified by Pseudo-seq $n = 4$ biological replicates for each condition. All genes (grey), genes with a high density induced $\Psi$ (blue), and genes with a log phase induced $\Psi$ (red) are indicated.
Extended Data Figure 6 | Inducible pseudouridylation of ncRNAs.

**a,** Pseudo-seq was performed on wild-type \((n = 4), snr81A (n = 2), pus1A (n = 2) and pus7A (n = 2) \) yeast strains grown to high density. CMC dependent peaks of reads are indicated with a dashed red line. Traces are representative indicated number of biological replicates. **a,** Pseudo-seq reads in U2 snRNA \((LSR1; \text{chrII: } 681751–681790, \text{left; chrII: } 681769–681818, \text{right})\) showing SNR81-dependence of U2-Ψ93, and PUS1-dependence of U2-Ψ56. Both are dependent on growth to high density. **b,** Pseudo-seq reads in U3a snoRNA \((SNR17A, \text{chrXV: } 780461–780560)\) showing snR17A-Ψ369 \((PUS7\)-dependent), snR17A-Ψ380, snR17A-Ψ391 and snR17A-Ψ425 \((PUS1\)-dependent). **c,** Summaries of the numbers of Ψs called in ncRNAs by Pseudo-seq. Indicated are constitutive Ψs (top), and inducible Ψs (bottom).
Extended Data Figure 7 | Analysis of potential snoRNA targets.
a–d, Pseudo-seq was performed on wild-type yeast in log phase, or grown to high density. Reads from $n=4$ biological replicate libraries for each condition were pooled. b–d, Indicated are the predicted snoRNA target site (black, dashed), and the expected peak of CMC-dependent reads (black, dotted).

a, Results of analysis on sets of random Us. a, A histogram of the differences ($1\text{CMC}-2\text{CMC}$) in mean normalized reads at the $1\text{CMC}$ peak position for 10,000 randomizations for high density (orange) and log phase (blue). The normalized read values for the computationally predicted Ys in exponential and high density samples are indicated by arrows. b, An averaged metaPsi plot for all randomizations. c, d, +CMC (c) and −CMC (d) MetaPsi plots for computationally predicted Ys separated by base pairing. Sites with 8 or more (red), 9 or more (blue), and 10 or more (orange) base pairs are indicated. Data for high density (left), and log phase (right) are indicated. e, Pseudo-seq reads for computationally predicted Ys, CAT2 (chrXII: 193995–19450, left), and AIM6 (chrIV: 31135–31550, right). Traces are representative of six biological replicates.
Extended Data Figure 8 | Mechanisms of Pus-dependent pseudouridylation. a, b, Summaries of the PUS-dependence of called Ψs using higher stringency cut-offs (10/14 libraries) (a) and lower stringency cut-offs (9/14 libraries) (b). c, f, CMC-dependent peaks of reads are indicated with dashed red lines. Traces are representative of \( n = 4 \) (wild type), and \( n = 2 \) (pusA) biological replicates. Pseudo-seq reads for RPL14A (a, chrXI: 431901–432200) and PDI1 (d, chrII: 49401–48760) showing PUS1- and PUS7-dependency, respectively. Both are dependent on growth to high density. d, e, g, h, WebLogo 3.3 was used to generate motifs for PUS1 (d), PUS2 (e), PUS7 (g) and PUS4 (h).
Extended Data Figure 9 | Positive controls for human RNA Pseudo-seq.

a, Pseudo-seq reads for RDN28S (1516–1765) containing five known $\Psi$s (28S-$\Psi$1536, 28S-$\Psi$1582, 28S-$\Psi$1677, 28S-$\Psi$1683 and 28S-$\Psi$1744). CMC-dependent peaks of reads are indicated with dashed red lines. Traces are representative of $n = 5$ biological replicates. b, A metaPsi plot of mean normalized reads (left axis) for +CMC libraries (orange), and −CMC libraries (blue). The number of $\Psi$s at each position in the metaPsi window is indicated (black, right axis). c, A ROC curve of the Pseudo-seq signal for all known $\Psi$s in rRNA and snRNA.
Extended Data Figure 10 | New pseudouridines in human RNAs. a, A Venn diagram showing the overlap of mRNA pseudouridylation events between serum-fed and serum-starved HeLa cells. b, A plot of log-transformed average RPKMs in serum-starved versus serum-fed HeLa for all coding genes with a $\Psi$ identified by Pseudo-seq. All genes with a $\Psi$ (grey), genes with a $\Psi$ induced in plus serum cells (blue), and genes with a $\Psi$ induced in serum-starved cells (red) are indicated. c, Pseudo-seq reads for $RDN18S5$ (184–411) (top, left), $RDN18S5$ (1015–1210) (top, right), $RDN28S5$ (2713–3108) (bottom, left), and $RDN28S5$ (4461–4618) (bottom, right). CMC-dependent peaks of reads are indicated with dashed red lines, and highlighted $\Psi$s are indicated by red boxes. Traces are representative of $n = 4$ biological replicates.