Supporting Information
for
"YTHDF2 Binds to 5-Methylcytosine in RNA and Modulates the Maturation of Ribosomal RNA"

Xiaoxia Dai[1,2], Gwendolyn Gonzalez[3], Lin Li[1], Jie Li[4,5], Changjun You[1,2], Weili Miao[1], Junchi Hu[6], Lijuan Fu[3], Yonghui Zhao[7], Ruidong Li[8], Lichao Li[8], Xuemei Chen[7], Yanhui Xu[4,5], Weifeng Gu[8,*], Yinsheng Wang[1,3,*]

[1]Department of Chemistry, University of California Riverside, CA 92521-0403, USA

[2]State Key Laboratory of Chemo/Bio-sensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha, Hunan 410082, China

[3]Environmental Toxicology Graduate Program, University of California Riverside, CA 92521-0403, USA

[4]Fudan University Shanghai Cancer Center, Department of Oncology; and Institutes of Biomedical Sciences and School of Basic Medical Sciences, Shanghai Medical College of Fudan University, Shanghai 200032, China

[5]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China

[6]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

[7]Department of Botany and Plant Sciences, Institute of Integrative Genome Biology, University of California, Riverside, CA 92521-0403, USA.

[8]Department of Cell Biology and Neuroscience, University of California, Riverside, CA 92521-0403, USA

*To whom correspondence should be addressed: yinsheng.wang@ucr.edu and weifeng.gu@ucr.edu

Table of Contents:

Supplementary Materials and Methods:

Western blot

The primary antibodies used in Western blot included mouse anti-YTHDF2 (Sigma, SAB1400554; 1:2,000 dilution), rabbit anti-YTHDF1(Abcam, ab99080; 1:5,000 dilution), mouse anti-YTHDF3 (Santa Cruz Biotechnology, SC-377119; 1:10,000 dilution), rabbit anti-CSTF1 (Sigma, C2872; 1:2,000 dilution), rabbit anti-CSTF2 (Santa Cruz Biotechnology, SC-28201; 1:10,000 dilution), rabbit anti-CSTF3 (Sigma, C9998; 1:10,000 dilution), and rabbit anti-FLAG (Cell Signaling Technology, 2368; 1:30,000 dilution). The secondary antibodies used were anti-mouse-IgG-HRP (Santa Cruz, SC-2005; 1:10,000 dilution) and anti-rabbit-IgG-HRP (Sigma, A0545; 1:10,000 dilution).

Electrophoretic mobility shift assay (EMSA)

EMSA was performed using a previously reported method with some modifications.[1] Briefly, RNA probes were purchased from IDT: 5′-ACUGGCUCCUUCCACGUCUCACXAGGCAGACAGU-3′ (X = C, m^5C, A, or m^6A). The RNA probe (2 pmol) was mixed with 5 μl 10×T4 PNK buffer (NEB), 1 μl T4 PNK (NEB), 40 units ml^{-1} RNase inhibitor (NEB), 40 μl RNase-free water and 1 μl [γ-^{32}P]-ATP at 37°C for 1 h. The probes were then purified by using micro bio-spin P-30 columns (Bio-Rad) and mixed with 2.5 μl 20×SSC (3 M NaCl, 0.3 M sodium citrate). The mixture was incubated at 65°C for 10 min and then cooled to room temperature. The probe (20 fmol) was incubated with increasing amount of YTHDF2 in a binding buffer (10 mM HEPES, pH 8.0, 50 mM KCl, 1 mM EDTA, 0.05% Triton-X-100, 5% glycerol, 10 μg/ml salmon DNA, 1 mM DTT, 40 units ml^{-1} RNase inhibitor) at 4°C for 1 h. The entire 10 μl sample was separated using 8% native polyacrylamide gel and the gel band intensities were quantified using phosphorimager analysis with a Typhoon 9410 Variable Mode Imager and ImageQuant software (GE Healthcare). The dissociation constant (K_d) was calculated using the formula of [Unbound RNA]/[Bound RNA]=$K_d \times 1/$[Protein].

Structure-based docking of RNAs with the YTH domain of YTHDF2

Maestro, version 9.0 (Schrödinger, LLC) was used for molecular docking to gain an understanding of the interaction between the YTH domain of YTHDF2 and m^5C in RNA. The coordinates of the protein were obtained from the X-ray crystal structure of YTHDF2-m^6A complex (PDB NO. 3RDN).[2] The RNA used for docking was from the complex structure of the YTH domain of YTHDC1 with GG(m^6A)CU (PDB NO. 4R3I).[3] The protein was prepared with Protein Preparation Wizard Workflow at pH 7.4 ± 0.0. Other parameters were set as default values. The m^6A-bearing RNA was truncated to a trimer, i.e. G(m^6A)C, to reduce non-specific interactions during the docking study. The truncated trimer RNA was prepared using the LigPrep, version 2.3, to generate a possible protonation state by Epik, version 2.0, at a target pH of 7.4 ± 0.0. A total of 32 conformations of G(m^{6}A)C were then generated. The Glide SP mode[4] in rapid dock package, as integrated in Maestro, was subsequently used to dock the m^{6}A RNA into the YTH domain of YTHDF2, during which the m^{6}A nucleotide was chosen as the docking site. The top ten generated models were examined in PyMol and the m^{6}A nucleotide adopted nearly identical orientation in the aromatic cage. The corresponding m^{5}C-containing RNA, G(m^{5}C)C, was built by substituting the m^{6}A in G(m^{6}A)C with m^{5}C in Maestro. The procedures for the preparation and docking of the m^{5}C RNA were the same as those for m^{6}A-RNA. Out of the top 10 models produced, 7 were with m^{5}C being located in the hydrophobic pocket, and 6 out of 7 exhibited similar orientation.

Bisulfite conversion, next-generation sequencing (NGS) library construction and data analysis

HEK293T cells with the YTHDF2 gene being ablated by CRISPR-Cas9 were generated previously.[5] The bisulfite conversion of mRNA was performed as previously described,[6] and we conducted all experiments in triplicate. Briefly, about 4 μg of mRNA was incubated with 100 μl conversion buffer (40% sodium bisulfite, 600 μM hydroquinone solution, pH 5.1) at 75°C for 4 h, followed by desalting twice using micro bio-spin P-6 columns (Bio-Rad). The samples were then mixed with an equal volume of 1.0 M Tris (pH 9.0) at 75°C for 1 h followed by ethanol precipitation. For conventional bisulfite sequencing, the bisulfite-treated RNA was converted to cDNA using EpiNext Hi-Fi cDNA synthesis kit (EpiGentek) according to manufacturer's instructions. PCR was performed using the primers listed in Supplementary Table S1 and the target products were isolated using gel purification kit (Qiagen). Amplicons were ligated into the pGEM-T vector (Promega) and individual clones were sequenced.

For NGS library construction (Figure S8), the bisulfite-treated poly(A) RNA was hydrolyzed with a buffer containing 0.045 M $NaHCO_3$ and 0.005 M Na_2CO_3 at 95°C for 5 min, and then precipitated using isopropanol and 20 μg glycogen. The sequencing library was constructed using a recently published method with a minor modification, i.e. with the inclusion of T4 polynucleotide kinase (PNK) in the ligation step.[7] The fragmented RNA was cloned with TruSeq LT/V1/V2 linkers using T4 RNA ligase 2. To resolve the 2'-3' cyclic phosphate at the 3' termini of the fragmented RNA, we added 0.5 μM PNK in the 3' ligation reaction which also contained other components necessary for ligation with unmodified 3' termini of RNA. The 3' ligation reaction was incubated at room temperature for 2.5 h and then switched to 37°C with the addition of 0.5 mM ATP to phosphorylate the 5' termini of the fragmented RNA.

The human genome/annotation Ensembl GRCh37 release 71 and pre-rRNA sequence 45SN1 from NCBI RefSeq were used in the analysis.[8] The Generic Genome Browser 1.70[9] was used to visualize the alignments. Single-end high-throughput sequencing reads of 50 nucleotides in length were conducted on a HiSeq 4000 platform (Illumina). A custom PERL script was used to sort the reads according to the indices and remove the 3′ linker sequences, if any. The reads were mapped to the human transcriptome using Bowtie 0.12.7[10] and custom PERL (5.10.1) scripts, which are available at https://github.com/guweifengucr/m5C_analysis.git. Briefly, we first converted all the C's in the transcriptome to U's, and mapped all the reads with at least 30 nucleotides (nts) in length to the converted transcriptome using '-n 3 -e 180 -m 4000', which allowed a maximum of 3 mismatches in the seed region, 6 mismatches for the whole size, and 4000 target hits for each read. For each read, only the best loci, which contained the minimum mismatches, were selected for further analyses. To minimize non-specific matches, we allowed a maximum of 6, 5, 4, 3, 2, and 1 mismatches for reads that are 45-50, 41-44, 38-40, 35-37, 32-34, and 30-31 nts long, respectively. To simplify the analysis, we converted the transcript positions to genomic coordinates; if a sequence matches two or more genomic loci, the read number of this sequence is equally divided among these loci. We calculated the mutation rate for each genomic position. In the bisulfite-treated samples, the nucleotide C is converted to and read as U while the nucleotide m^5C is read as C since it is resistant to the treatment. Therefore, a genomic C, which is read as U, has three mutation rates $A/(A+U+G+m^5C)$, $G/(A+U+G+m^5C)$, and $m^5C/(A+U+G+m^5C)$. Theoretically, these A and G mutations were likely generated by PCR and/or sequencing errors, whereas the m^5C rates are composed of authentic m^5C modifications, and PCR and/or sequencing errors.

We analyzed all the genomic C's covered by at least 100 reads and used the $m^5C/(A+m^5C+G+U)$ mutation rate to define the m^5C-containing sites. We discarded any C (read as U) with more U→A or U→G mutations than U→m^5C because these U's may represent hot spots for frequent PCR/sequencing errors and/or genetic variations. To further minimize the false-positive identification of m^5C sites, we used BisRNA to enrich the authentic sites.[11] For the m^5C analysis of each genomic C site, three replicates were employed to generate the adjusted *p* value,[11] which was employed to define the final list of the 'authentic' m^5C sites, and the median m^5C rate, which was used for comparing the loci-specific m^5C levels in YTHDF2 knockout and parental HEK293T cells.[11]

Northern blot

Northern blots were performed as described previously.[12] Briefly, RNA was isolated using TRI reagent (Sigma). The RNA (2 µg) was separated on a 1% agarose gel in 30 mM triethanolamine, 30 mM tricine, and 1.25% formaldehyde, transferred to a Hybond-N+ membrane (GE Healthcare) using a downward capillary transfer system, and UV cross-linked to the membrane. The membrane was prehybridized at 50°C in 6×SSC (saline-sodium citrate), 5× Denhardt's solution, 0.5% SDS, and 0.9 µg ml^{-1} tRNA. The 5'-radiolabeled oligonucleotide ITS1/ITS2 probe was added after 1 h and incubated at 50°C overnight. Membrane was washed three times in 2×SSC and 0.1% SDS for 10 min, and then exposed.

Table S1. The primers and probes used in the present study.

Primer Name	Primer Sequence
pRK7-YTHDF2-F	5'- AAATCTAGAATGTCGGCCAGCAGCCTCTTG -3'
pRK7-YTHDF2-R	5'- AAAGGATCCTTTCCCACGACCTTGACGTTCC-3'
YTHDF2W432A-F	5'- GTTCCATTAAGTATAATATTGCGTGCAGCACAGAGC -3'
YTHDF2W432A-R	5'-GCTCTGTGCTGCACGCAATATTATACTTAATGGAAC-3'
ITS1	5'-CCTCGCCCTCCGGGCTCCGTTAATGATC-3'
ITS2	5'-CGCACCCCGAGGAGCCCGGAGGCACCCCCGG-3'

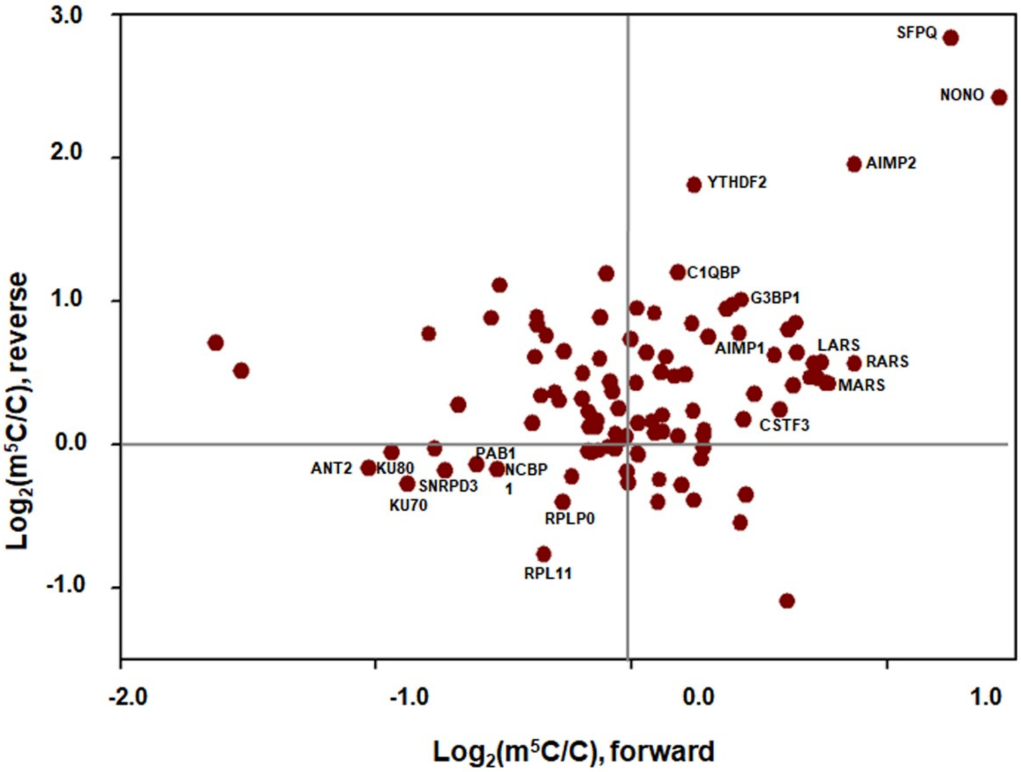

Figure S1. Scatterplot showing the log$_2$(ratio) for the proteins identified in RNA pull-down assay in HEK293T cells. The data were based on results obtained from two forward and two reverse SILAC experiments.

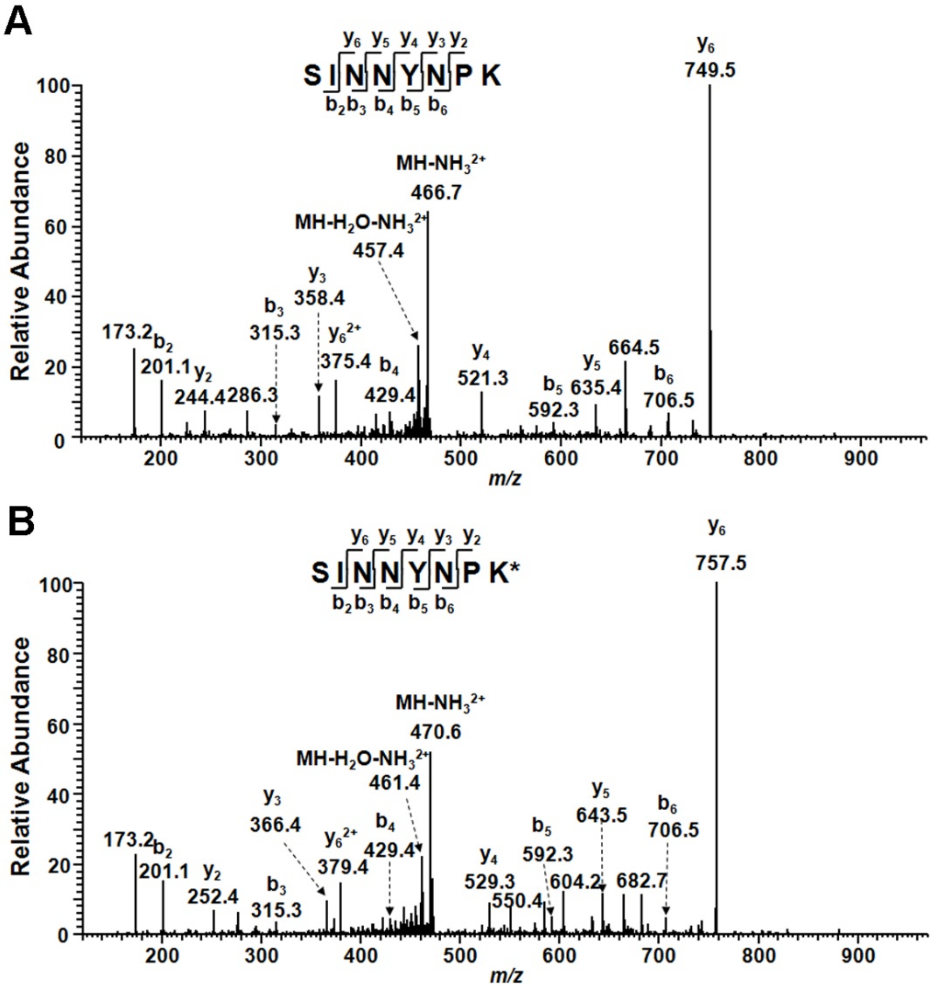

Figure S2. Representative MS/MS data of a tryptic peptide from YTHDF2 in SILAC experiments. Shown are the MS/MS for the $[M+2H]^{2+}$ ions of YTHDF2 peptide SINNYNPK (**A**) and SINNYNPK* (**B**, 'K*' designates the heavy lysine), respectively.

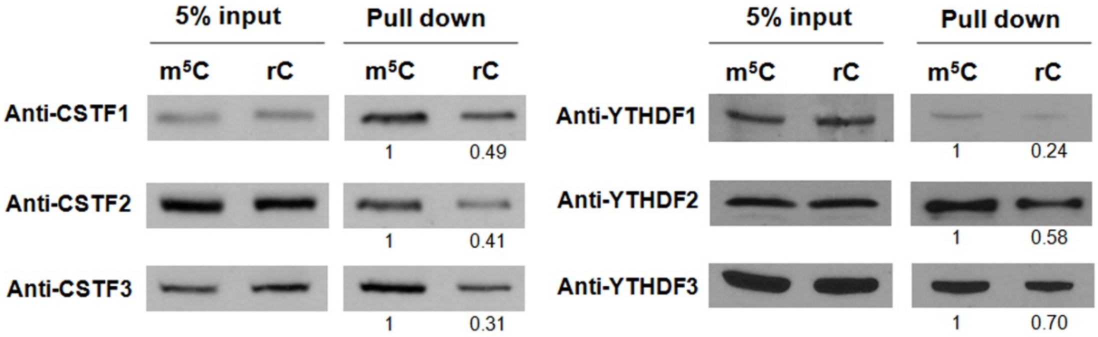

Figure S3. Western blot showing the preferential binding of YTHDF1-3 and CSTF1-3 toward m^{5}C-containing RNA over the corresponding C-containing RNA. The relative band intensities (pull-down/input, and normalized to the results for the m^{5}C probe) for the pull-down proteins are labeled below the gel image.

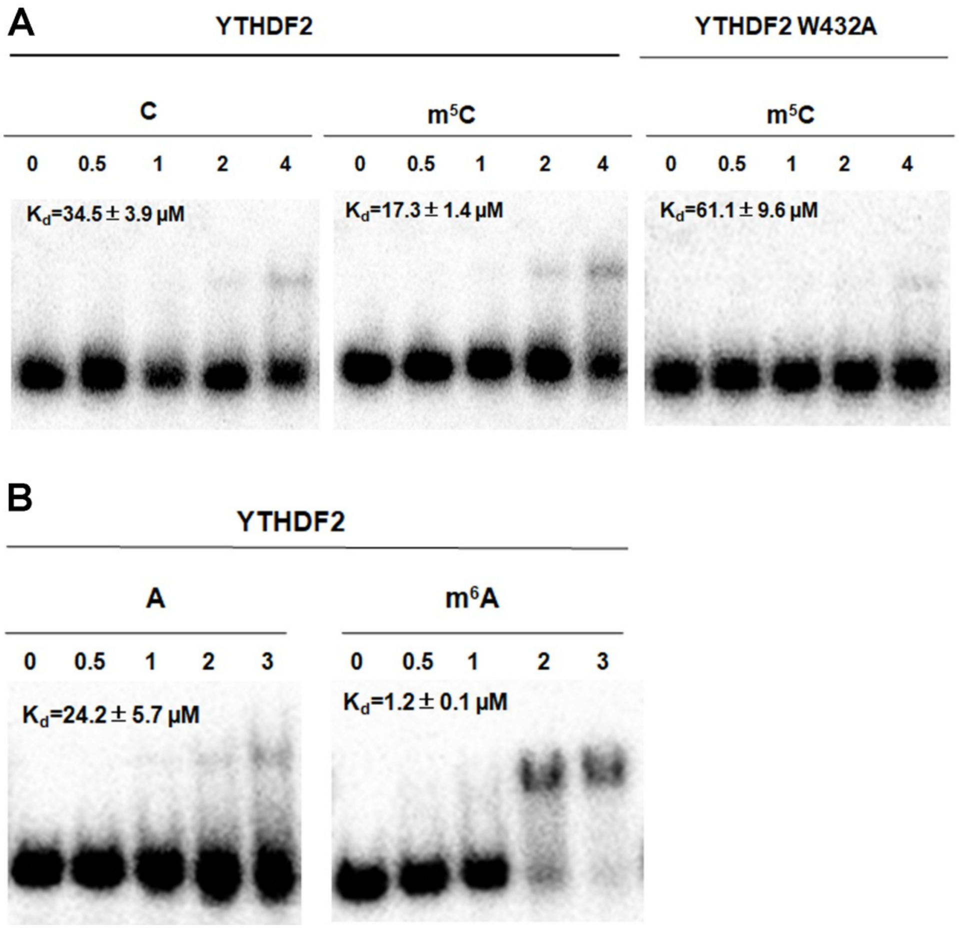

Figure S4. Electrophoretic mobility shift assay for measuring the binding affinities of YTHDF2 and W432A mutant proteins with methylated and unmethylated RNA probes. (**A**) The binding affinity of YTHDF2 and W432A mutant proteins with m^{5}C- and C-containing RNA probes. (**B**) The binding affinity of YTHDF2 with m^{6}A- and A-containing RNA probes. Protein concentrations ranged from 0.5 to 4 μM. The dissociation constants (K_d) are listed in individual figure panels, and the data represent the mean ± S. D. from three separate EMSA experiments.

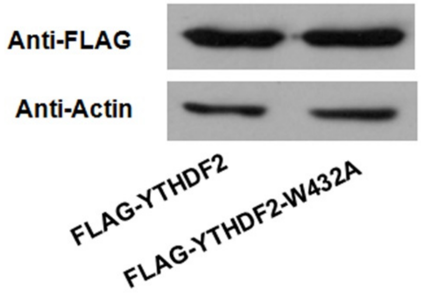

Figure S5. Western blot showing the expression levels of FLAG-YTHDF2 and FLAG-YTHDF2-W432A in HEK293T cells.

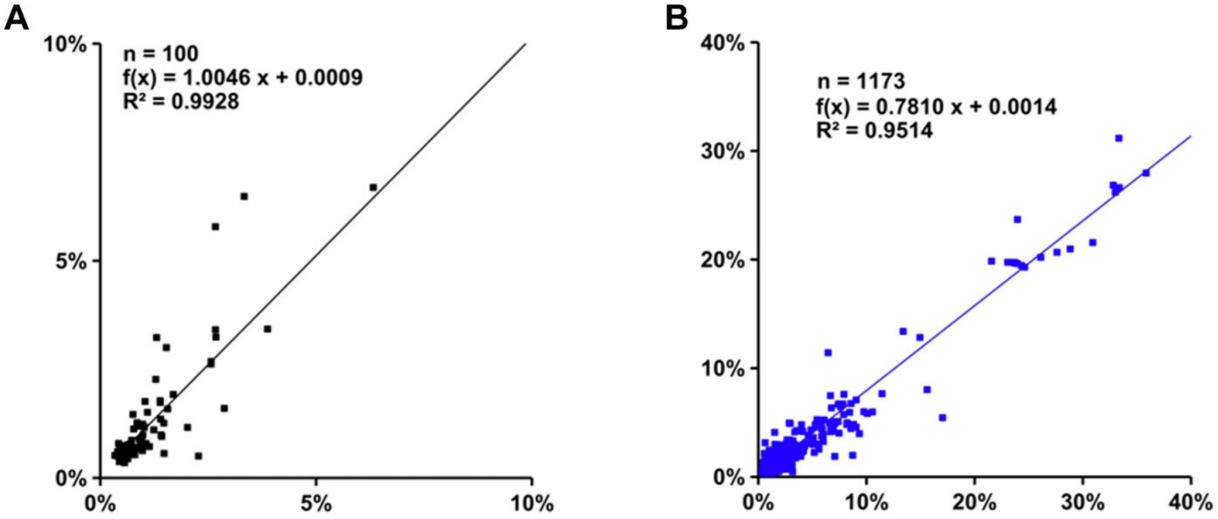

Figure S6. Linear regression analysis to show the overall levels of m^5C in mtRNA (A) and mRNA (B) in the YTHDF2-depeleted cells (Y axis) and the isogenic parental cells (X axis). '*x*' and '*f(x)*' represent m^5C rates in HEK293T and YTHDF2 knockout cells, respectively.

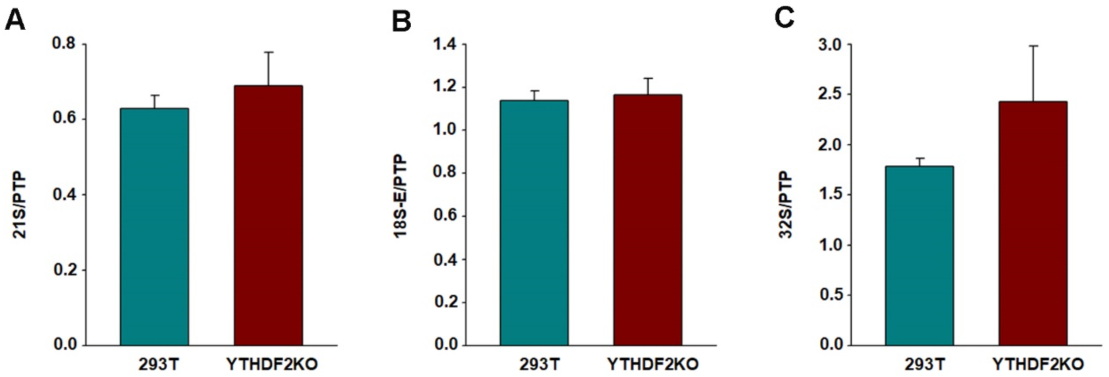

Figure S7. Northern blot quantification analysis of total RNA extracts from HEK293T cells. Northern blots probed with oligonucleotides complementary to the 5′ of the ITS1 (**A**, **B**) and the ITS2 (**C**). PTP, primary transcript plus (47S, 46S, 45S). Error bar represents the S.E. (*n* = 3).

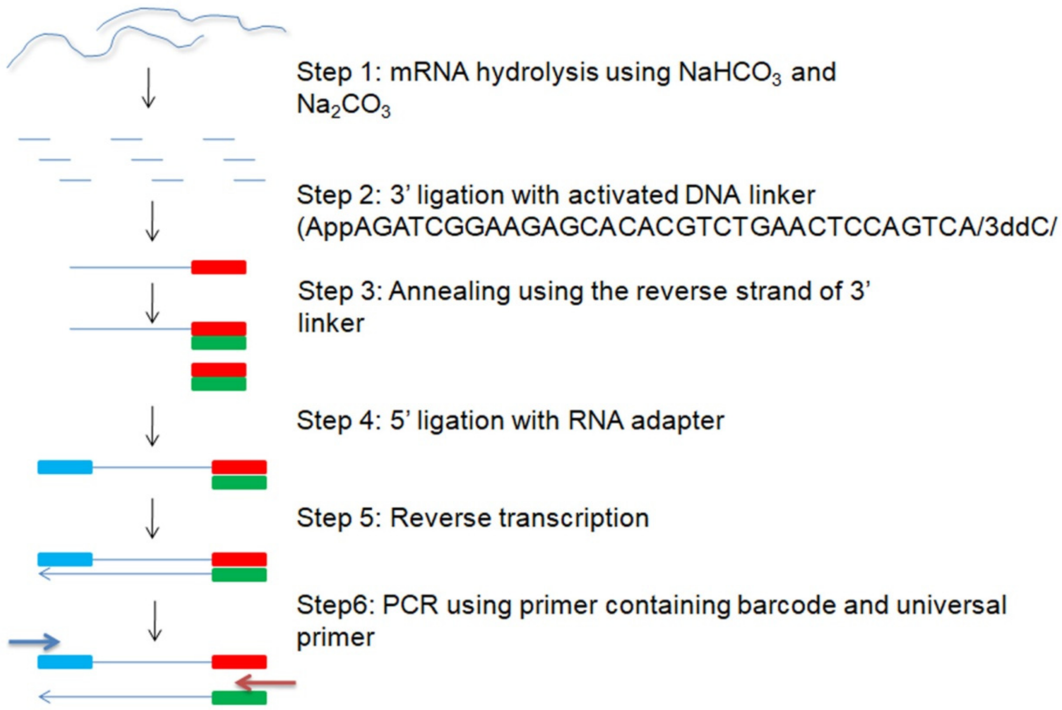
Step 1: mRNA hydrolysis using $NaHCO_3$ and Na_2CO_3
Step 2: 3’ ligation with activated DNA linker (AppAGATCGGAAGAGCACACGTCTGAACTCCAGTCA/3ddC/
Step 3: Annealing using the reverse strand of 3’ linker
Step 4: 5’ ligation with RNA adapter
Step 5: Reverse transcription
Step6: PCR using primer containing barcode and universal primer

Figure S8. A schematic diagram showing the procedures for the preparation of the sequencing library. The bisulfite-treated and ethanol-precipitated RNA was used in Step 1 (see Supplementary Materials and Methods).

References:

(1) Wang, X.; Lu, Z.; Gomez, A.; Hon, G. C.; Yue, Y.; Han, D.; Fu, Y.; Parisien, M.; Dai, Q.; Jia, G.; Ren, B.; Pan, T.; He, C. *Nature* **2014**, *505*, 117-120.
(2) Li, F.; Zhao, D.; Wu, J.; Shi, Y. *Cell Res.* **2014**, *24*, 1490-1492.
(3) Xu, C.; Wang, X.; Liu, K.; Roundtree, I. A.; Tempel, W.; Li, Y.; Lu, Z.; He, C.; Min, J. *Nat. Chem. Biol.* **2014**, *10*, 927-929.
(4) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739-1749.
(5) Miao, W.; Li, L.; Zhao, Y.; Dai, X.; Chen, X.; Wang, Y. *Nat Commun* **2019**, *10*, 3613.
(6) Squires, J. E.; Patel, H. R.; Nousch, M.; Sibbritt, T.; Humphreys, D. T.; Parker, B. J.; Suter, C. M.; Preiss, T. *Nucleic Acids Res.* **2012**, *40*, 5023-5033.
(7) Li, L.; Dai, H.; Nguyen, A. P.; Gu, W. *RNA* **2019**, DOI: 10.1261/rna.071605.071119.
(8) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; Gil, L.; Giron, C. G.; Gordon, L.; Hourlier, T.; Hunt, S.; Johnson, N.; Juettemann, T.; Kahari, A. K.; Keenan, S.; Kulesha, E., et al. *Nucleic Acids Res.* **2014**, *42*, D749-755.
(9) Stein, L. D.; Mungall, C.; Shu, S.; Caudy, M.; Mangone, M.; Day, A.; Nickerson, E.; Stajich, J. E.; Harris, T. W.; Arva, A.; Lewis, S. *Genome Res.* **2002**, *12*, 1599-1610.
(10) Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S. L. *Genome Biol.* **2009**, *10*, R25.
(11) Legrand, C.; Tuorto, F.; Hartmann, M.; Liebers, R.; Jacob, D.; Helm, M.; Lyko, F. *Genome Res.* **2017**, *27*, 1589-1596.
(12) Wang, M.; Pestov, D. G. *Methods Mol. Biol.* **2016**, *1455*, 147-157.