nature chemical biology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41589-019-0459-3

Keth-seq for transcriptome-wide RNA structure mapping

Xiaocheng Weng[1,2,5], Jing Gong[3,5], Yi Chen[2,5], Tong Wu[1,5], Fang Wang[1,4], Shixi Yang[2], Yushu Yuan[2], Guanzheng Luo[1], Kai Chen[1], Lulu Hu[1], Honghui Ma[1], Pingluan Wang[1], Qiangfeng Cliff Zhang[3]*, Xiang Zhou[2]* and Chuan He[1]*

RNA secondary structure is critical to RNA regulation and function. We report a new N_3-kethoxal reagent that allows fast and reversible labeling of single-stranded guanine bases in live cells. This N_3-kethoxal-based chemistry allows efficient RNA labeling under mild conditions and transcriptome-wide RNA secondary structure mapping.

Knowledge of RNA folding is critical to the understanding of the function of various RNA species[1]. Chemical probes have played key roles in transcriptome-wide RNA secondary structure studies[2]. An increasing number of methods have been developed in recent years for high-throughput RNA structure mapping[3–11]. Two notable classes of chemical probes, dimethyl sulfate (DMS) and SHAPE, enable transcriptome-wide in vivo RNA structurome mapping[12]. Both methods are effective, but with sufficient space for improvement. DMS is toxic at high concentration and mostly methylates the Hoogsteen face of bases; SHAPE molecules are hydrolytically unstable and label the 2'-OH of sugar instead of the bases[13,14]. A more specific, nontoxic reagent that rapidly labels the Watson–Crick interface under mild conditions will offer additional advantages for in vivo RNA labeling and RNA secondary structure probing.

Ethyl-3-(3-dimethylaminopropyl)carbodiimide) (EDC), nicotinoyl azide (NAz), glyoxal and its derivatives were recently developed to expand the toolbox of probing RNA secondary structures in a low-throughput manner[15–18]. Kethoxal (1,1-dihydroxy-3-ethoxy-2-butanone) is known to react with guanines in single-stranded RNA (ssRNA) under mild conditions, which induces reverse transcription (RT) stops[19]. It could also react with inosine to form an unstable hemiacetal adduct[20]. However, lack of synthetic routes to modified kethoxal hampered its use for transcriptome-wide studies. Here, with a new synthetic design (Supplementary Note 1), we report the preparation of azido-kethoxal (N_3-kethoxal, **1**) for the specific labeling of the N1 and N2 positions at the Watson–Crick interface of guanines in ssRNA (Fig. 1a). The azido group offers a bioorthogonal handle that can be modified with a biotin or dyes for enrichment or other applications[10]. In addition, the reversibility of the kethoxal-guanine reaction under alkaline or heating conditions[19] provides an additional advantage in the RT-stop-based RNA structure mapping by producing read-through controls after removing the kethoxal labels.

N_3-kethoxal only reacts with guanine in ssRNA and is inert with other nucleic bases (Supplementary Fig. 1 and Fig. 1b). Among chemically modified guanines, N_3-kethoxal does not react with m^1G and m^2G but can label m^7G, verifying that N_3-kethoxal specifically modifies the N1 and N2 positions of guanine (Supplementary Fig. 1). In a synthetic RNA oligo, all guanines in the guanine-containing oligo were labeled by N_3-kethoxal, while the oligo without guanine showed no reaction (Supplementary Fig. 2). N_3-kethoxal exhibits higher RNA labeling activity compared with other reported RNA secondary structure probes, including DMS, NAI, glyoxal and EDC (Supplementary Fig. 3). As shown by gel electrophoresis, dot blot and mass spectrum analysis (Fig. 1c and Supplementary Fig. 4), N_3-kethoxal-modified RNAs can be successfully biotinylated, which can then be enriched by streptavidin-conjugated beads, to increase the signal-to-noise ratio in biological applications.

We evaluated cell-based labeling efficiency by adding N_3-kethoxal into the culture medium of mouse embryonic stem cells (mESCs) directly. Dot blotting of biotinylated RNA indicated that N_3-kethoxal could permeate into living cells efficiently in 1 min, with the signal saturated in 5 min, suggesting a quick cell penetration and high labeling efficiency of N_3-kethoxal (Fig. 1d). The fast labeling is also confirmed by high-throughput sequencing results where the G-stop ratio increased from 1 min and reached the maximum after 2.5 min (Supplementary Fig. 5). The rapid labeling property enables N_3-kethoxal to be used in transient events such as stress response, signaling and so on.

The kethoxal-guanine adduct is unstable under alkaline conditions[19]. By adding excessive guanine monomers to trap dissociated N_3-kethoxal, the labeling can be removed to yield unmodified RNAs in a neutral buffer in a short period of time (Supplementary Fig. 6a). The excessive GTP almost completely removes the N_3-kethoxal modification on the labeled RNA in 8 h at 37 °C (Supplementary Fig. 6b) or in 10 min at 95 °C (Fig. 1e and Supplementary Fig. 6c). The labeling adduct of kethoxal-guanine could be stabilized in borate buffer as previously reported[19], providing flexibility to manipulate the N_3-kethoxal adduct on RNA.

We next combined N_3-kethoxal probing with deep sequencing (Keth-seq) to probe RNA secondary structures in mESCs. In each experiment, we constructed three different RNA libraries, including an N_3-kethoxal-modified RNA sample, a no-treatment control sample and an N_3-kethoxal-removal sample made by erasing the N_3-kethoxal labeling before the RT (Supplementary Fig. 7). We observed a high correlation at both RPKM (Supplementary Fig. 8a)

[1]Department of Chemistry, Department of Biochemistry and Molecular Biology, Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, Chicago, IL, USA. [2]College of Chemistry and Molecular Sciences, Key Laboratory of Biomedical Polymers of Ministry of Education, Wuhan University, Wuhan, China. [3]MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China. [4]Wuhan University School of Pharmaceutical Sciences, Wuhan, China. [5]These authors contributed equally: Xiaocheng Weng, Jing Gong, Yi Chen, Tong Wu. *e-mail: qczhang@tsinghua.edu.cn; xzhou@whu.edu.cn; chuanhe@uchicago.edu

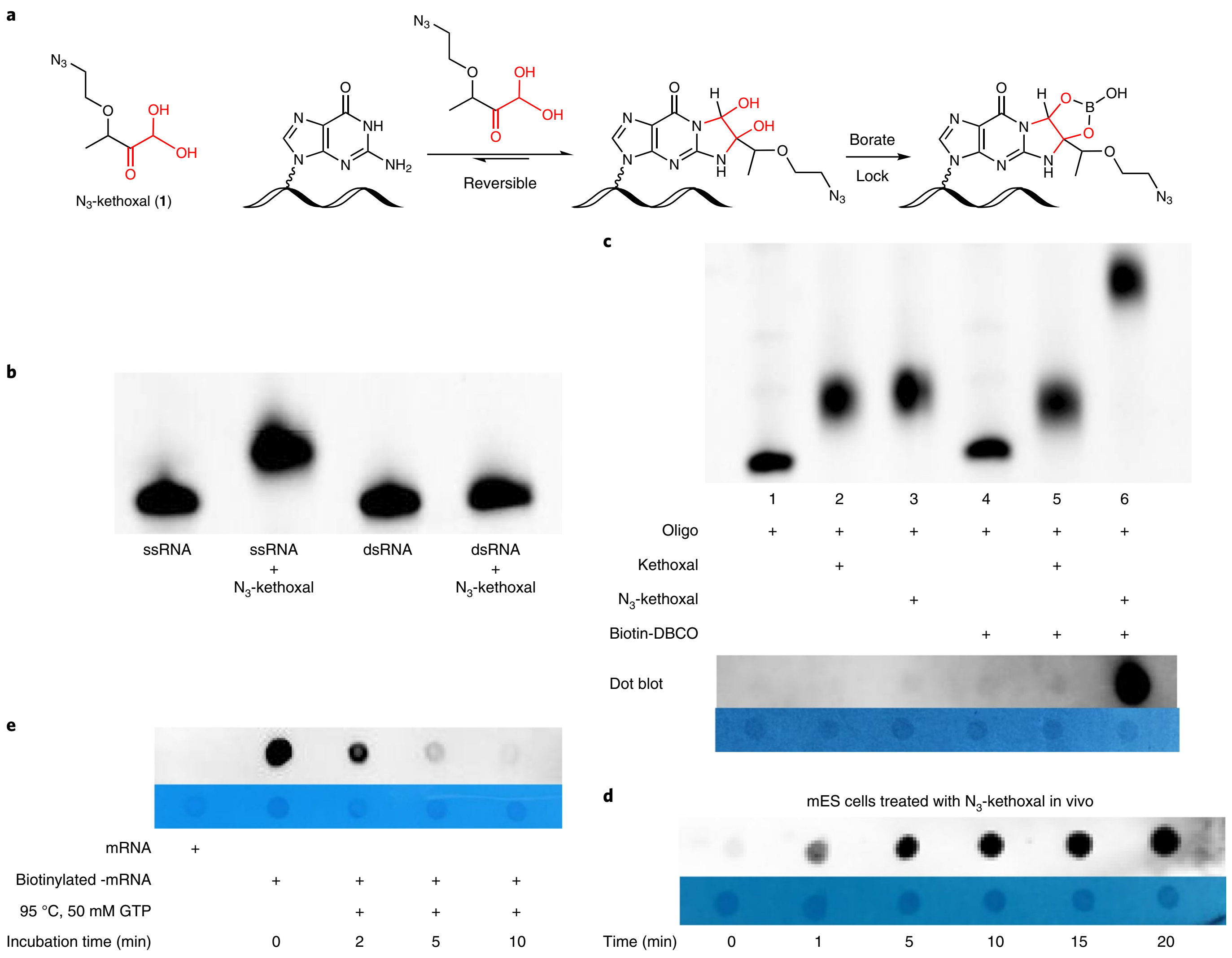

Fig. 1 | N_3-kethoxal and experimental evaluation of its selectivity, cell permeability and reversibility. a, The structure of N_3-kethoxal and the reaction with guanine. **b**, Denaturing gel electrophoresis demonstrating N_3-kethoxal only reacts with single-strand RNA (ssRNA). **c**, Upper, denaturing gel electrophoresis analysis of the labeling reaction of kethoxal and N_3-kethoxal with FAM-RNA oligo (5′-FAM-GAGCAGCUUUAGUUUAGAUCGAGUGUA, lanes 1–3) and biotinylation with biotin-DBCO (lanes 5 and 6). Only N_3-kethoxal labeled RNA can be biotinylated (lane 6). Bottom, dot blot of RNA after labeling and biotinylation reactions. Methylene blue dot results are listed as a control. **d**, Dot blot of isolated total RNA from mES cells that were treated by N_3-kethoxal with different periods, 1, 5, 10, 15 and 20 min. **e**, Dot blot analysis of the reversibility of N_3-kethoxal labeled messenger RNA in the presence of 50 mM GTP at 95 °C. The N_3-kethoxal modification in mRNA was removed thoroughly after 10 min incubation. Experiments were independently repeated twice with similar results obtained. Uncropped scans for **b**–**e** are provided in Supplementary Fig. 15.

and RT-stop level between Keth-seq replicates (Fig. 2a), indicating that Keth-seq is highly reproducible. Additionally, in the N_3-kethoxal sample, guanine (>80%) dominates the RT-stopped sites among all reads, with no RT-stop bias across all four bases in the no-treatment control sample (Supplementary Fig. 8b), confirming that N_3-kethoxal is highly selective to guanine. RT-stopping sites in N_3-kethoxal-removal samples decreased dramatically to a similar level to the no-treatment control, indicating that N_3-kethoxal modification was almost completely removed during the reversal process (Supplementary Fig. 8b). In mRNA *mt-Atp8*, for instance, we observed more full-length RNA fragments in the N_3-kethoxal-removal sample than in the no-treatment control sample, suggesting that the RT-stopped sites could be more confidently identified using the N_3-kethoxal-removal sample as the 'background' (Supplementary Fig. 9).

To validate Keth-seq, we analyzed guanine signals from Keth-seq and compared with icSHAPE both globally and at the transcript level[10]. For every common transcript ($n = 455$), we calculated a correlation coefficient between Keth-seq and icSHAPE by using their reactivity profile on all guanines, and plotted the whole distribution as an accumulative curve (Fig. 2b). About 80% of the transcripts show a positive correlation (Pearson correlation coefficient $R \geq 0.4$, Fig. 2b), indicating that Keth-seq agrees well with the established icSHAPE technology. To directly evaluate the accuracy of Keth-seq in determining RNA secondary structure, we compared the reactivity profile of Keth-seq with icSHAPE on all guanines in the mouse 18S ribosomal RNA with the known structure model from the RNA STRAND database (ID: CRW_00356)[21]. Keth-seq reactivity profile achieves a higher area under curve (AUC) than icSHAPE in fitting the 18S ribosomal RNA model (Keth-seq = 0.81, icSHAPE = 0.71) (Fig. 2c). More specifically, Keth-seq shows a higher reactivity score than icSHAPE for single-stranded G nucleotides and thus more accurately revealing unpaired Gs (Supplementary Fig. 10a), although both methods

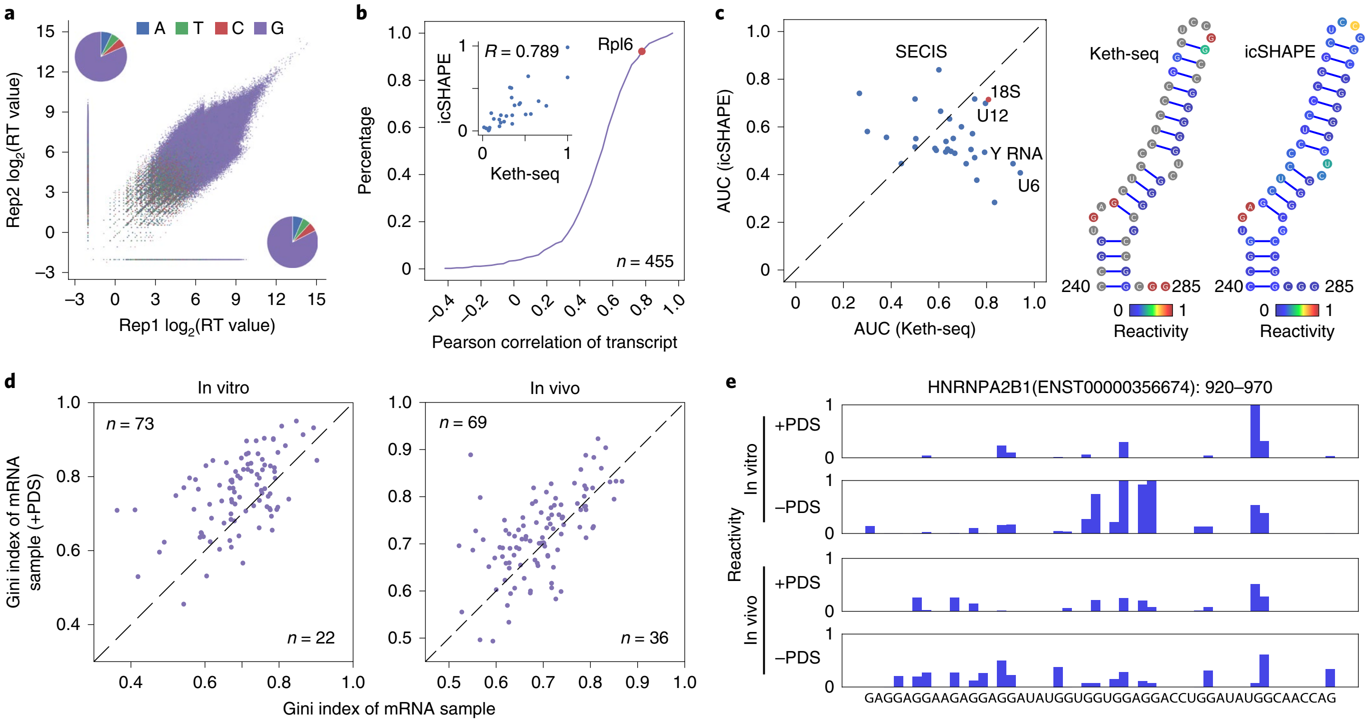

Fig. 2 | Keth-seq method and the profile around rG4 regions. a, Scatter plot of the RT-stop reads distribution between replicates for the N_3-kethoxal sample. The inset pie plots show RT-stopped base distribution for replicate 2 (upper left: A, 604,222; T, 497,602; C, 481,596; G, 7,204,998) and replicate 1 (bottom right: A, 703,486; T, 586,297; C, 551,962; G, 8,683,824). **b**, Accumulation plot of correlation coefficient between Keth-seq and icSHAPE for all transcripts. For each common transcript, we calculated the Pearson correlation coefficient for structural signal of guanine bases. The inset plot shows all guanine reactivity between Keth-seq and icSHAPE for *Rpl6* (a gene encoding ribosomal protein) transcript with a high correlation (Pearson correlation coefficient $R = 0.789$). **c**, Left, scatter plot of AUC between Keth-seq and icSHAPE for RNAs with a known structure model (18S ribosomal RNA from RNA STRAND database and others from Rfam database, 32 RNAs in total). Right, a fragment (240–285) of 18S ribosomal RNA with both Keth-seq and icSHAPE reactivity filled in the structure model. **d**, Gini index of known rG4 regions (based on those previously identified by Kwok et al.[22]) between +PDS treatment and native sample for in vitro (left) and in vivo (right). Only regions with structural information in both +PDS treatment and native conditions are retained for plotting (extended to 50-nucleotides long). **e**, An example of a Keth-seq profile around previously identified in vitro rG4 regions.

similarly agree well with the 18S model on its double-stranded regions (Fig. 2c).

We then extended the comparison by using all available mouse RNA secondary structure models from the Rfam database, and found that Keth-seq achieves a higher AUC than icSHAPE for most of these RNAs (Fig. 2c). In addition, we compared Keth-seq and DMS-seq[6] by evaluating their performance on human 18S RNA (ID: CRW_00347) and showed that Keth-seq achieves a comparable accuracy to DMS-seq with similar AUCs obtained (Supplementary Fig. 10b). Furthermore, we applied Keth-seq to probe RNA structure both in vivo and in vitro for mESCs and used the Gini index to measure the structural evenness of RNAs[6]. Consistent with previous findings, we observed that RNAs in vitro showed a higher Gini index than in vivo (Supplementary Fig. 10c,d), validating the folding complexity of cellular RNAs and the feasibility of Keth-seq for in vivo detection.

Formation of RNA G-quadruplexes (rG4) from isolated RNAs has been shown in different studies. However, the in vivo detection of rG4 remains challenging[22,23]. As N_3-kethoxal is highly specific toward labeling N1 and N2 positions of guanine and can enrich labeled RNA, Keth-seq could be sensitive to probe the potential presence of the rG4 structure in live cells. After demonstrating that N_3-kethoxal can detect rG4 in vitro (Supplementary Fig. 11a–c), we conducted Keth-seq using isolated HeLa RNA or in live HeLa cell in the presence or absence of PDS, which has been shown to induce rG4 formation inside cells[24]. We first explored the structure landscape of previously identified rG4 regions by rG4-seq under PDS treatment in vitro[22], and detected 95 regions with the structure information detected by Keth-seq under both native and PDS treatment conditions (Supplementary Fig. 12a). In the PDS-treated samples, these regions show a higher Gini index than the control sample, suggesting the formation of rG4 under PDS treatment (Fig. 2d). Consistent with previous observations[22], these rG4 regions preferentially occur at untranslated regions (Supplementary Fig. 12b) and are associated with biological pathways (Supplementary Fig. 12c) including translation, transcription and metabolism, suggesting potential regulatory roles of rG4s.

To further explore whether rG4 can fold in live cells, we performed similar analysis using in vivo Keth-seq data and detected 105 previously identified rG4 regions under both native and PDS treatment conditions (Supplementary Fig. 12d). Of these 105 regions, 69 showed a higher Gini index under PDS treatment compared with the control, indicating that rG4 structure could potentially form at these regions in live cells. The genomic context distribution and top enriched biological pathways of these regions are both similar to those in vitro (Supplementary Fig. 12e,f). We included examples where the signal in the defined rG4 regions in the PDS sample is lower than that in the control sample (Fig. 2e and Supplementary Fig. 13), indicating that PDS treatment induces rG4 formation both in vitro and in live cells.

We noted that only a small subset of rG4s possibilities from the rG4 dataset[22] are detected by Keth-seq. This could be due to insufficient sequencing depth or the possibility that chemical labeling only detects kinetically stable structures and may miss highly dynamic rG4s[25]. Although rG4s detected in vitro may not fold in vivo[23], our study does suggest that a portion of rG4s could exist in situ.

In summary, we showed that N_3-kethoxal readily labels RNA and established Keth-seq as an effective method for transcriptome-wide RNA secondary structure mapping in live cells. Because of the high selectivity and reactivity of N_3-kethoxal labeling of guanines in ssRNA, Keth-seq is able to map secondary structures such as rG4 under mild conditions. The efficient live cell RNA labeling by N_3-kethoxal provides an approach that could be expanded for RNA enrichment, RNA targeting and RNA proximity studies in the future.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-019-0459-3.

Received: 19 January 2019; Accepted: 20 December 2019;
Published online: 03 February 2020

References

1. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* **12**, 641–655 (2011).
2. Kubota, M., Tran, C. & Spitale, R. C. Progress and challenges for chemical probing of RNA structure inside living cells. *Nat. Chem. Biol.* **11**, 933–941 (2015).
3. Kertesz, M. et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).
4. Underwood, J. G. et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* **7**, 995–1001 (2010).
5. Lucks, J. B. et al. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl Acad. Sci. USA* **108**, 11063–11068 (2011).
6. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
7. Ding, Y. et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
8. Talkish, J., May, G., Lin, Y., Woolford, J. L. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20**, 713–720 (2014).
9. Wan, Y. et al. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
10. Spitale, R. C. et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
11. Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2016).
12. Lu, Z. & Chang, H. Y. Decoding the RNA structurome. *Curr. Opin. Struct. Biol.* **36**, 142–148 (2016).
13. National Toxicology Program. Dimethyl sulfate. *Rep. Carcinog.* **12**, 174–175 (2011).
14. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* **127**, 4223–4231 (2005).
15. Mitchell, D. et al. Glyoxals as in vivo RNA structural probes of guanine base-pairing. *RNA* **24**, 114–124 (2018).
16. Mitchell, D. et al. In vivo RNA structural probing of uracil and guanine base pairing by 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC). *RNA* **25**, 147–157 (2019).
17. Wang, P. Y., Sexton, A. N., Culligan, W. J. & Simon, M. D. Carbodiimide reagents for the chemical probing of RNA structure in cells. *RNA* **25**, 135–146 (2019).
18. Feng, C. et al. Light-activated chemical probing of nucleobase solvent accessibility inside cells. *Nat. Chem. Biol.* **14**, 276–283 (2018).
19. Xu, Z. & Culver, G.M. In *Methods in Enzymology; Biophysical, Chemical, and Functional Probes of Rna Structure, Interactions and Folding, Pt A* (ed. Herschalag, D.) Vol 468, 47–165 (Academic Press, 2009).
20. Morse, D. P. & Bass, B. L. Detection of inosine in messenger RNA by inosine-specific cleavage. *Biochemistry* **36**, 8429–8434 (1997).
21. Andronescu, M., Bereg, V., Hoos, H. H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinforma.* **9**, 340 (2008).
22. Kwok, C. K., Marsico, G., Sahakyan, A. B., Chambers, V. S. & Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods* **13**, 841–844 (2016).
23. Guo, J. U. & Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **353**, aaf5371 (2016).
24. Biffi, G., Di Antonio, M., Tannahill, D. & Balasubramanian, S. Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat. Chem.* **6**, 75–80 (2014).
25. Kwok, C. K., Marsico, G. & Balasubramanian, S. Detecting RNA G-quadruplexes (rG4s) in the transcriptome. *Cold Spring Harb. Perspect. Biol.* **10**, a032284 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Synthesis of N_3-kethoxal. The synthesis of N_3-kethoxal and the characterization of compounds (^{1}H nuclear magnetic resonance (NMR), ^{13}C NMR and high-resolution mass spectrometry) are included in Supplementary Note 1.

General chemical and biological materials. All chemical reagents for N_3-kethoxal synthesis were purchased from commercial sources. RNA oligos were purchased from Integrated DNA Technologies (IDT) and Takara Biomedical Technology. Buffer salts and chemical reagents for N_3-kethoxal synthesis were purchased from commercial sources. Superscript III, Dynabeads MyOne Streptavidin C1 was purchased from Life Technologies. T4 PNK, T4 RNL2tr K227Q, 5′-Deadenylase, RecJ$_f$ were purchased from New England Biolabs. CircLigaseII was purchase from Epicentre (an Illumina company). DBCO-Biotin was purchase from Click Chemistry Tools LLC (A116-10). All RNase-free solutions were prepared from DEPC-treated MilliQ-water.

The reaction of N_3-kethoxal and RNA oligo. The reaction was generally performed with the following protocol: 100 pmol RNA oligo and 1 µmol N_3-kethoxal was incubated in a total 10-µl solution in kethoxal reaction buffer (0.1 M sodium cacodylate, 10 mM $MgCl_2$, pH 7.0) at 37 °C for 10 min. To induce rG4 folding in vitro, RNA were denatured at 95 °C for 5 min then cooled to 4 °C for 5 min, before 1 M KCl (2 µl), 0.1 M sodium cacodylate buffer (pH 7.0) and PDS (final concentration of 5 µM) were added. The mixture was incubated at 37 °C for 10 min to achieve equilibration. N_3-kethoxal was then added to the reaction mixture to react with folded RNA. The final reaction volume was 10 µl. The modified RNA was purified by Micro Bio-Spin P-6 Gel Columns (Biorad, 7326222). The purified labeled RNA can be used for further used for mass spectrometry, gel electrophoresis, primer extension assay and copper-free click reaction with biotin-DBCO.

Remove N_3-kethoxal modification from N_3-kethoxal labeled RNA. The detailed protocol of N_3-kethoxal modification erasing is listed in step 5 'N_3-kethoxal-remove sample preparation' of the Keth-seq protocol in the Supplementary Information. Generally, the purified N_3-kethoxal modified RNA was incubated with a high concentration of GTP (1/2 volume of the reaction solution, with final concentration 50 mM) at 37 °C for 6 h or at 95 °C for 10 min. Higher temperature benefits the removal of the N_3-kethoxal modification.

Fixation of N_3-kethoxal modification in RNA. The N_3-kethoxal modification in RNA can be fixed in the presence of borate buffer. The solution of N_3-kethoxal labeled RNA was mixed with 1/10 volume of stock borate buffer (final concentration 50 mM; stock borate buffer: 500 mM potassium borate, pH 7.0; pH was monitored while adding potassium hydroxide pellets to 500 mM boric acid). The borate buffer fixation was used in steps 2, 4 and 6 of the Keth-seq protocol in the Supplementary Information.

Matrix-assisted laser desorption/ionization–time of flight–mass spectometry (MALDI–TOF–MS) analysis of N_3-kethoxal labeled RNA oligo. The N_3-kethoxal labeled RNA was purified by Micro Bio-Spin P-6 Gel Columns. Meanwhile the buffer exchange occurred from the kethoxal reaction buffer to the Tris buffer that can be directly used in MALDI–TOF–MS experiments without an extra desalt step. One microliter of product solution was mixed with 1 µl of matrix, which included an 8:1 volume ratio of 2'4'6'-trihydroxyacetophenone (10 mg ml^{-1} in 50% CH_3CN/H_2O):ammonium citrate (50 mg ml^{-1} in H_2O). Then, the mixture was spotted on the MALDI sample plate, dried and analyzed by Bruker Ultraflextreme MALDI–TOF–TOF Mass Spectrometers.

The selectivity of N_3-kethoxal to ssRNA by gel electrophoresis. The complementary RNA oligos FS1 (Fluorescent RNA oligo) and S2 were hybridized to double-stranded RNA (dsRNA) in the ratio of FS1:S2 = 1.2:1 to ensure all FS1 was involved in the formation of dsRNA. After the reaction with N_3-kethoxal, the purified product by Micro Bio-Spin P-6 Gel Columns was analyzed by denaturing gel electrophoresis (Novex TBE-Urea Gels, 15%, Invitrogen, EC6885BOX). Gel Imaging was collected in a Pharos FX Molecular imager (Biorad).

RNA sequence:

FS1: 5′-FAM-GAGCAGCUUUAGUUUAGAUCGAGUGUA
S2: UACACUCGAUCUAAACUAAAGCUGCUC.

High-performance liquid chromatography condition. The product of N_3-kethoxal for RNA nucleic bases was analyzed using an LC-6AD (Shimadzu) HPLC instrument, which was equipped with an Inertsil ODS-SP column (5 µm, 250 × 4.6 mm^2) (GL Science). The phase A (100 mM TEAA buffer, pH = 7.0) and phase B (CH_3CN) were used as eluents with a flow rate of 1 ml min^{-1} at 35 °C (B concentration: 5–5–30%/0–5–30 min).

The biotinylation of N_3-kethoxal labeled RNA and dot blot assay. For the in vitro study, the purified N_3-kethoxal RNA was incubated with DBCO-biotin at 37 °C for 2 h in the presence of RNase inhibitor, borate buffer (step 2 'biotinylation' of the Keth-seq protocol in the Supplementary Information). For RNA oligo analysis, the biotinylated product was purified by Micro Bio-Spin P-6 Gel Columns and subject to dot blot assay and MALDI–TOF–MS detection; for total RNA or mRNA, the product was purified by RNA Clean & Concentrator 5 (Zymo Research, R1015) and subject to further experiments.

For the in vivo study, 10 µl of N_3-kethoxal was added to the cell culture medium in a 100-mm cell culture dish with nearly 80% confluent mES cells. After incubation at 37 °C in a CO_2 incubator for a specific time, the medium was aspirated and the cells were washed three times with PBS. The total RNA was isolated by Trizol reagent (Invitrogen, 15596026) or Qiagen RNeasy plus mini kit (Qiagen, 74134). mRNA was isolated by Dynabeads mRNA DIRECT Purification Kit (Invitrogen, 61011). The biotinylation step was the same as for the in vitro study. The biotinylated RNA was purified by RNA Clean & Concentrator 5.

For the dot blot assay, 1 µl of RNA (100 ng µl^{-1}) of the sample was spotted onto the Amersham Hybond-N+ membrane (RPN119B, GE Healthcare) and ultravioletly crosslinked to the membrane by an UVP HL-2000 hybriLinker. The membrane was washed using 1× PBST (0.1% tween-20) and blocked with 5% nonfat dry milk in 1× PBST overnight at 4 °C. After four washes using 1× PBST at 10-min intervals, the streptavidin-horseradish peroxidase (1:15,000 dilutions, streptavidin-HRP, Life Technologies, S-911) in 1× PBST with 3% BSA was added and incubated at room temperature for 40 min. Then, the membrane was washed again using 1× PBST at 10-min intervals and developed using SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Scientific, 34577). The membrane was washed again by 1× PBST and stained with methylene blue solution (0.02% methylene blue in 0.3 M sodium acetate pH 5.2).

Primer extension assay. RNA templates (N_3-kethoxal treated or not) were dissolved in 13.5 µl of nuclease-free water. Then, 1 µl of 10 µM FAM-labeled DNA primer, 2 µl of RT buffer, 1 µl of 0.1 M DTT, 1 µl 5 mM dNTPs and 1.5 µl of RevertAid Reverse Transcriptase (200 U µl^{-1}) were added (total volume 20 µl). The G ladder was made by dideoxy sequencing method, with 2 µl of 10 mM corresponding ddNTP added to replace 2 µl of nuclease-free water. The RT was performed at 37 °C for 30 min, and then 20 µl of deionized formamide was added. The reaction mixture was immediately heated up to 95 °C for 10 min, then cooled down to 4 °C. The complementary DNAs were size fractionated by 20% denaturing polyacrylamide gel containing 8 M urea. The gel was scanned with Pharos FX Molecular imager (BioRad) operated in the fluorescence mode (excitation wavelength (λ_{ex}) = 488 nm).

Keth-seq library preparation. The library was prepared following a previously reported procedure with slight changes[10]. The detailed protocol was included in Supplementary Note 2. For in vitro library preparation, RNA was isolated and refolded in RNA folding mix buffer (100 mM HEPES, pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$). The refolded RNA was treated with N_3-kethoxal and then used for library construction. For the in vivo study, N_3-kethoxal was added into the culture medium of mES cell or HeLa cell and the RNA was isolated to be used for library construction.

Isolated N_3-kethoxal labeled RNA was biotinylated with water-soluble DBCO-biotin (Click Chemistry tool, A116) by a copper-free click reaction, then fragmented by sonication. The RNA Fragmentation Reagent is not suitable for the fragmentation step of Keth-seq because high temperature will affect N_3-kethoxal labeling in RNA. In addition, the borate buffer is also necessary in each step before cDNA production except for the kethoxal-remove experiment. Fragmented RNA was subjected to end repair by T4 PNK, 3′-adapter ligation (3′ adapter, /5rApp/TGGAATTCTCGGGTGCCAAGG/3ddC/) followed by the 3′ adapter removing by 5′-deadenylase and RecJf digestion. The ligation products were separated into two fractions, with 90% used to produce the N_3-kethoxal library and the remaining 10% used for the N_3-kethoxal-remove sample library.

For the N_3-kethoxal sample, RT primer, dNTPs, SuperScript III, borate solution and RNase inhibitor were mixed with RNA to perform RT (RT primer, /5Phos/DDDNNAACCNNNNGATCGTCGGACTGTAGAACTCTGAACAT/iSp18/GGATCC/iSp18/TACCTTGGCACCC). After cDNA synthesis, the cDNA–RNA was kept cool to avoid denaturing and was immediately subjected to immunoprecipitation by Dynabeads MyOne Streptavidin C1. Beads were washed and the truncated cDNA was eluted by RNase A/T1 and RNase H digestion. For the N_3-kethoxal-remove sample, RNA was incubated with GTP in nuclease-free water at 95 °C for 10 min to remove any N_3-kethoxal modification. The RNA was then purified and RT was performed similar to the N_3-kethoxal sample.

cDNA from the N_3-kethoxal sample and the N_3-kethoxal-remove sample were subjected to size selection by gel electrophoresis separation, which can remove the excess RT primers and the self-ligation product of primers. The purified cDNA was used for cDNA cyclization by CircLigaseII to obtain the circDNA. The circDNA were then amplified by PCR with short primers (F, 5′-TGGCACCCGAGAATTCCA; R, 5′-TTCAGAGTTCTACAGTCCGA). In this step, quantitative PCR was performed to evaluate the cycle numbers of each samples to avoid over-amplification. After purification and size selection, a final library PCR amplification was performed using the full sequencing primers from TruSeq Small RNA Sample Prep Kits of (Illumina). The products were purified by low melting point agarose gel and used for deep sequencing.

For the no-treatment control sample, RNA was isolated from cells without any N_3-kethoxal treatment, followed by fragmentation, adapter ligation, RT, cDNA cyclization and PCR amplification as described above to construct the library for deep sequencing.

Sequencing data processing. As the library structure is similar to that of icSHAPE, we used the same strategy to process the sequencing reads by using the icSHAPE scripts at (https://github.com/qczhang/icSHAPE)[10]. First, readCollapse.pl was used to collapse the reads with default parameter. Note that we include a barcode of random hexamer (NNNNNN) ligated to the fragments during library construction (Supplementary Fig. 14). These random barcodes serve to identify PCR duplicates from real different fragments with the identical sequences. Reads with fully identical sequence (including the barcode and the insert fragment) were marked as PCR duplicates and filtered before subsequent analysis. However, reads with different barcodes were retained, even though they contained the identical insert fragments and were subsequently mapped to the same start and end positions, as they actually represent different fragments in the library.

Then we used trimming.pl to cut potential adapter sequences (-l 13 -t 0 -c phred33 –a adapter.fa -m 0, adapter sequence: ATGGAATTCTCGGGTGCCAAG GAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAA AAAA). Next, we mapped the clean reads to ribosomal RNAs, and the unmapped reads were mapped to the transcriptome (Gencode, mm10 for mouse and hg38 for human) using Bowtie with default parameters. We calculated RT-stop signals using the script calcRT.pl. After evaluating the correlation between different replicates (correlationRT.pl), we combined RT signals of replicates (combineRTreplicates.pl) for subsequent analysis, and then normalized them for both the kethoxal and the control samples, respectively (normalizeRTfile.pl -m mean:vigintile2 -d 32 -l 32). A structure reactivity score for each nucleotide in each transcript was calculated by comparing the kethoxal sample (foreground) versus the control sample (background) using the script calcEnrich.pl (-w factor5:scaling1 –x 0.25). The calculation is based on following formula: $A \times (\mathrm{RT[kethoxal]} - B \times \mathrm{RT[control]})/\mathrm{BD[control]}$, where RT represents the RT-stop count of the nucleotide in the sample, BD represents the base density of the nucleotide, and A and B are scaling factors to control the effects of subtraction. Our previous work on icSHAPE technology development trained the two scaling factors on mouse 5S ribosomal RNA, the structure of which has been determined by both high-throughput sequencing and low-throughput RT-stop gel analysis[10,26], by maximizing the correlation between calculated reactivity scores and gel-based results. We found that although the scaling factors perform best around $A = 10$ and $B = 0.25$, they are relatively insensitive. We thus used the same parameters in Keth-seq and we did observe high accuracy on known structures. Finally, to obtain high-quality scores, we only kept nucleotides with adequate sequencing coverage: filterEnrich.pl –T 2 -t 200 -s 5 -e 30. Here, '-T 2' requires the minimal average number of RT stops over the whole transcript to be no less than 2; '-t 200' requires the base density of a nucleotide with reactivity to be no less than 200; and '-s 5 -e 30' is to trim away the first five and the last 30 nucleotides as they tend to have low sequencing quality scores.

For rG4 probing, we first converted the genomic coordinated of previously reported rG4 regions in HeLa cells[22] to transcriptome coordinates[27]. The converted regions with ≥60% NULL value of structure scores from our Keth-seq experiments were filtered from subsequent analysis. The retained regions were used for comparison between +PDS and –PDS Keth-seq samples.

Comparison between Keth-seq and icSHAPE/DMS-seq. To compare the performance of Keth-seq and icSHAPE, we collected known RNA secondary structure models from different sources, including the mouse 18S ribosomal RNA structure from the RNA STRAND database[21] (ID: CRW_00356) and the other 614 RNA structure models from the Rfam database[28]. Both Keth-seq and icSHAPE[10] sequencing reads were remapped to these specific RNAs and the reactivity score profiles were calculated (18S rRNA and 31 other RNAs with structure information are common in the two experiments and retained). We calculated receiver operator characteristic curves to measure to what degree the structural probing reactivity scores fits the reference structure model. Using different reactivity score cutoffs, each nucleotide can be predicted (classified) as single-stranded or double-stranded. A true positive is defined as a single-stranded base with a reactivity score higher than the cutoff. A true negative is defined as a paired base with a reactivity score lower than the cutoff. AUC is calculated using the signals of guanine nucleotides for Keth-seq while considering the signals of all four bases for icSHAPE. To compare Keth-seq with DMS-seq, we collected DMS-seq data[6] for Fibroblast and the K562 sample, and evaluated the performance on human 18S ribosomal RNA (RNA STRAND ID: CRW_00347). AUC is calculated using the signals of the adenine and cytosine nucleotides for DMS-seq. For 18S rRNA, we first parsed the known three-dimensional ribosome structure (PDB ID 4V6X) and derived an accessibility score for each nucleotide. Only nucleotides with accessibility score greater than three were retained for evaluation.

Calculating Gini index for regions. We followed a previously reported method to calculate the Gini index to measure the level of RNA structure formation[6]. In principle, a Gini index represents data evenness, and a completely unfolded and a completely base-paired RNA would both have a low Gini index. However, in reality, a RNA usually has a comparable number of single-stranded and double-stranded nucleotides, and in a normal range, the Gini index of a RNA well correlates with its structure formation. To explore the correlation between RNA structure and Gini index, we collected all mouse RNAs from the Rfam database ($n = 614$), the secondary structures of which have been determined. We performed a simulation analysis, where a random reactivity score is assigned to each nucleotide in the RNA, with a score in the range 0.5–1.0 for a single-stranded nucleotide, and a score in the range 0.0–0.5 for a double-stranded nucleotide. Then a Gini index is calculated from the simulated reactivity profile. The random simulation was repeated 100 times. We observed that the higher the Gini index, the more structured (higher double-stranded nucleotide ratio) the RNA is (Supplementary Fig. 8c). Also note that another measurement of RNA structure level is the average of reactivity scores, and we have repeated and confirmed all conclusions in the main text concerning Gini index with reactivity score average.

Assume the reactivity profile of a region is $(x_1, x_2, x_3, \ldots, x_n)$, where x_n is the reactivity score for base n. We calculate a Gini index value as follows: (1) sort the reactivity value of the region in ascending order and take the summation $\left(\mathrm{Sum} = \sum_{j=1}^{n} x_j\right)$ and accumulation $\left(\mathrm{Acc}_j = \sum_{i=1}^{j} x_i\right)$; (2) calculate the accumulating area $\left(\mathrm{Cumulating}_{\mathrm{area}} = \sum_{j=1}^{n} \left(\mathrm{Acc}_j - \frac{x_j}{2}\right)\right)$ and fair area $\left(\mathrm{Fair}_{\mathrm{area}} = \frac{\mathrm{Sum}^{*} n}{2}\right)$ and (3) calculate the Gini index value $\left(\mathrm{Gini} = \frac{\mathrm{Fair}_{\mathrm{area}} - \mathrm{Cumulating}_{\mathrm{area}}}{\mathrm{Fair}_{\mathrm{area}}}\right)$.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All genomic data sets have been deposited in the Gene Expression Omnibus under accession number GSE122096. Other data and materials are available from the authors upon reasonable request.

Code availability

All custom codes used in this study are available at https://github.com/Tsinghua-gongjing/Keth-seq.

References

26. Spitale, R. C. et al. RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9**, 18–20 (2013).
27. Lu, Z. et al. RNA Duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**, 1267–1279 (2016).
28. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D442 (2018).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant nos. 21572172, 21778040 and 21822704 to X.W.; 21432008, 91753201 and 21721005 to X.Z.; 31671355, 91740204 and 31761163007 to Q.C.Z.) and National Institutes of Health grant no. HG008935 (C.H.). C.H. is an investigator of the Howard Hughes Medical Institute. X.W. was supported by China Scholarship Council (CSC) during his visit to the University of Chicago. We acknowledge S. Frank, who edited the manuscript.

Author contributions

X.W., Q.C.Z., X.Z. and C.H. conceived the project, designed the experiments and wrote the manuscript. X.W., Y.C. and T.W. performed the experiments with the help of F.W., S.Y., Y.Y., G.L., K.C., L.H., H.M. and P.W. J.G. and Q.C.Z. designed and performed the bioinformatics analysis.

Competing interests

C.H. is a scientific founder and a member of the scientific advisory board of Accent Therapeutics, Inc., and a shareholder of Epican Genetech.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-019-0459-3.

Correspondence and requests for materials should be addressed to Q.C.Z., X.Z. or C.H.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Chuan He

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Flexcontrol (3.4, Bruker) was used for MALDI-TOF data collection. Image lab (5.2.1) was used for gel imaging. HCS (3.4.0) was used for sequencing data collection on the Hiseq4000 platform. Novaseq Control Software was used for sequencing data collection in the Novaseq platform.
Data analysis	Topspin; Python (2.7.15); Bowtie (1.1.2); icSHAPE pipeline (https://github.com/qczhang/icSHAPE): readCollapse.pl; trimming.pl; calcRT.pl; correlationRT.pl; combineRTreplicates.pl; normalizeRTfile.pl; calcEnrich.pl; filterEnrich.pl. All codes are available at https://github.com/Tsinghua-gongjing/Keth-seq.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data have been deposited in the NCBI Gene Expression Omnibus (GEO) and are accessible through GEO series accession number GSE122096.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Two biological replicates were performed for all sequencing experiments, which is generally accepted by the field, as variations between replicates are usually not big within the same cell line.
Data exclusions	No data were excluded from the analysis.
Replication	Results were confirmed in two or three biological replicates for each experiment unless otherwise stated. Detailed replication information is stated in the legends of each figures. All attempts to replicate data are successful.
Randomization	The experiments were not randomized. Controlling for covariates was unnecessary because all assays were performed in pairs.
Blinding	The investigators were not blinded to allocation during experiments and outcome assessment due to feasibility.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Unique biological materials
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	No antibody used used in this study except for Streptavidin-HRP (Thermo, catalog number 21130).
Validation	Validation statements of the Streptavidin-HRP used in this study are available on the manufacturer's websites.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa cell used in this study was purchased from ATCC (catalog number CCL-2). mESC used in this study was purchased from ATCC (catalog number CRL-1821).

Authentication	None of the cell lines used were authenticated.
Mycoplasma contamination	All cell lines used in this study were tested negative of mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified line was used.