THE UNIVERSITY OF CHICAGO

CHEMICAL APPROACHES TO DECIPHER TRANSCRIPTION AND RNA METABOLISM

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

TONG WU

CHICAGO, ILLINOIS

DECEMBER 2020

Table of Contents

List of Figures

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Prof. Chuan He. I have always been encouraged by his continuous passion for a variety of fundamental scientific questions and his genuine care to students. When I joined the lab, I was scared because I was trained as an organic chemist, and I barely knew anything about biology. Therefore, I am especially grateful that Chuan suggested I use small molecules to study biological questions, which fits my aptitude. I was able to make N_3-kethoxal soon after I joined the lab, but it took me two years to figure out its applications. During this time, Chuan offered a lot of ideas, patience, and encouragement. He also provides freedom for me to try new projects and directions in our lab, as well as opportunities to collaborate with other labs with different expertise. These experiences broadened my scientific horizon, which is beneficial to my own projects and my career in the long term. This dissertation will be impossible without consistent support from Chuan.

I would like to thank all the members from the He lab, as colleagues, teachers, and friends. The interdisciplinary environment contributed by every lab member enabled the development and applications of new chemical biology tools. Dr. Xiaocheng Weng guided me through my first year in the lab, and we collaborated on the synthesis of N_3-kethoxal and its application in RNA secondary structure profiling. Dr. Ruitu Lyu worked closely with me on the KAS-seq project. Dr. Ziyang Hao and Dr. Boxuan Zhao taught me many basic knowledge and techniques about DNA and RNA methylation. Dr. Qiancheng You provided critical suggestions for the KAS-seq and RNA-RNA interaction projects. Dr. Pingluan Wang made a number of kethoxal derivatives, which broadened the application scope of kethoxal and deepened my understanding of their structure-activity relationship. I also thank Dr. Xiaolong Cui, Dr. Xiaoyang Dou, Dr. Qing Dai, Qinzhe Liu, and Xinran Feng for assisting experiments or data analysis. Dr. Jun Liu, Dr. Bryan Harada, and

Dr. Hailing Shi provided valuable discussions. I thank Dr. Jordi Tauler for maintaining all the facilities and other daily operations in the lab. It is my great fortune to work in this fantastic group for the past five years of my life.

I also want to thank my collaborators. Prof. Zhengqing Ouyang, Prof. Brenton Graveley, and Anthony Cheng from the University of Connecticut provided crucial help on analyzing RNA-RNA interaction data. Prof. Qiangfeng Cliff Zhang and Jing Gong from Tsinghua University analyzed Keth-seq data. I worked with Dr. Huilin Huang in Prof. Jianjun Chen's lab from the City of Hope to study the effect of H3K36me3 on m^6A deposition. I have also been studying the roles of RNA modifications on various biological systems with Prof. Yi Liu at the University of Texas Southwestern medical center, Prof. Robert Ho, and Prof. Yun Fang from the University of Chicago.

I am grateful for the help from Prof. Bryan Dickinson and Prof. Yamuna Krishnan for their time and suggestions as my dissertation committee members.

I cannot finish my research without the support from my family and friends. Unconditional care and love from my parents throughout my life have always been my source of comfort and motivation. Their education since my early years plays a determinant role in shaping my characters that are essential for research. I enjoy every moment I spent with friends. Their company made my life much more colorful.

Abstract

Gene expression involves complex processes and determines cell fate and physiological functions. Dysfunction of gene expression regulation is associated with human diseases. Coupled with the next-generation sequencing, small-molecule probes and nucleic acid analogs have been widely involved in interrogating gene expression and its regulatory mechanisms in a high-throughput manner. Although powerful, existing techniques have different limitations, and the current chemical toolbox still needs to be expanded. My doctoral work focus on the development and applications of a novel chemical probe, N_3-kethoxal, which reacts specifically with guanine in single-stranded regions of DNA and RNA. I showed that N_3-kethoxal serves as a versatile tool that helps to profile RNA secondary structures, RNA-RNA interactions, and transcription dynamics. I also show that METTL4-mediated DNA N^6-methyldeoxyadenosine (6mA) modification accumulates in mammalian mitochondrial DNA (mtDNA). 6mA attenuates mitochondrial activity by repressing mtDNA transcription and replication.

List of Publications Based on Work Presented in this Thesis†

1. Qiancheng You*, Anthony Cheng*, Xi Gu*, Bryan Harada, Miao Yu, **Tong Wu**, Bing Ren, Zhengqing Ouyang, and Chuan He. Chemical-crosslinking assisted proximity capture (CAP-C) uncovers transcription-dependent chromatin organization at high resolution, *Nat. Biotechnol.*, in press.
2. **Tong Wu***, Ruitu Lyu*, Qiancheng You, and Chuan He. Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ, *Nat. Methods*, 17, 515-523 (2020)
3. Ziyang Hao*, **Tong Wu***, Xiaolong Cui, Pingping Zhu, Caiping Tan, Xiaoyang Dou, Kai-Wen Hsu, Yueh-Te Lin, Pei-Hua Peng, Li-Sheng Zhang, Yawei Gao, Lulu Hu, Hui-Lung Sun, Allen Zhu, Jianzhao Liu, Kou-Juey Wu, and Chuan He. N^6-deoxyadenosine methylation in mammalian mitochondrial DNA, *Mol. Cell*, 78, 382-395 (2020)
4. Xiaocheng Weng*, Jing Gong*, Yi Chen*, **Tong Wu***, Fang Wang, Shixi Yang, Yushu Yuan, Guanzheng Luo, Kai Chen, Lulu Hu, Honghui Ma, Pingluan Wang, Qiangfeng Cliff Zhang, Xiang Zhou, and Chuan He. Keth-seq for transcriptome-wide RNA structure mapping, *Nat. Chem. Bio.*, 16, 489-492 (2020)
5. Olga Karginova*, Claire M. Weekley*, Akila Raoul, Alhareth Alsayed, **Tong Wu**, Steve Seung-Young Lee, Chuan He, and Olufunmilayo I. Olopade. Inhibition of Copper Transport Induces Apoptosis in Triple-Negative Breast Cancer Cells and Suppresses Tumor Angiogenesis, *Mol. Cancer. Ther.*, 18 (5), 873-885 (2019)
6. Huilin Huang*, Hengyou Weng*, Keren Zhou*, **Tong Wu***, Boxuan Simen Zhao*, Mingli Sun, Zhenhua Chen, Xiaolan Deng, Gang Xiao, Franziska Auer, Lars Klemm, Huizhe Wu, Zhixiang Zuo, Xi Qin, Yunzhu Dong, Yile Zhou, Hanjun Qin, Shu Tao, Juan Du, Jun Liu,

Zhike Lu, Hang Yin, Ana Mesquita, Celvie L. Yuan, Yueh-Chiang Hu, Wenju Sun, Rui Su, Lei Dong, Chao Shen, Chenying Li, Ying Qing, Xi Jiang, Xiwei Wu, Miao Sun, Jun-Lin Guan, Lianghu Qu, Minjie Wei, Markus Müschen, Gang Huang, Chuan He, Jianhua Yang, and Jianjun Chen. Histone H3 trimethylation at lysine 36 guides m^6A RNA modification co-transcriptionally, *Nature*, 567, 414-419 (2019)

7. Rui Su*, Lei Dong*, Chenying Li*, Sigrid Nachtergaele*, Mark Wunderlich, Ying Qing, Xiaolan Deng, Yungui Wang, Xiaocheng Weng, Chao Hu, Mengxia Yu, Jennifer Skibbe, Qing Dai, Dongling Zou, **Tong Wu**, Kangkang Yu, Hengyou Weng, Huilin Huang, Kyle Ferchen, Xi Qin, Bin Zhang, Jun Qi, Atsuo T. Sasaki, David P. Plas, James E. Bradner, Minjie Wei, Guido Marcucci, Xi Jiang, James C. Mulloy, Jie Jin, Chuan He, and Jianjun Chen. R-2HG exhibits anti-tumor activity by targeting FTO/m6A/MYC/CEBPA signaling, *Cell*, 172, 90-105 (2018)
8. Xiao Shu*, Qing Dai*, **Tong Wu***, Ian R. Bothwell, Yanan Yue, Zezhou Zhang, Jie Cao, Qili Fei, Minkui Luo, Chuan He, and Jianzhao Liu. N^6-allyladenosine: a new small molecule for RNA labeling identified by mutation assay. *J. Am. Chem. Soc.*, 139 (48), 12713-12716 (2017)

* Co-authors contributed equally.

† The following chapters of the dissertation contain sections and figures adopted from the listed pubilications with modifications. Chapter 2: publication 4; chapter 3: publication 2; chapter 5: publication 1; chapter 6: publication 3.

Chapter 1

Introduction: Transcription and RNA Metabolism

1.1 Transcription and its regulation

Transcription is the first step that a cell makes proteins from DNA. During transcription, RNA polymerases catalyze the formation of phosphodiester bonds between four types nucleotides in 5'-to-3' direction to make RNA chains ranging from tens to thousands of nucleotides in length. In the nucleus of eukaryotes, different RNA species are transcribed by three different RNA polymerases, namely RNA polymerase (Pol) I, II, and III. Pol I generates large ribosomal RNA (rRNA) precursors (45S)[1]; Pol II transcribes messenger RNAs (mRNA) and some non-coding RNAs[2]; Pol III produces transfer RNAs (tRNA), some small RNAs, and small ribosomal RNA (5S)[3]. POLRMT (DNA-directed RNA polymerase, mitochondrial) transcribes mammalian mitochondrial DNA (mtDNA) into a nascent polycistronic precursor, which was subsequently excised into mature mitochondrial mRNA and tRNAs[4].

Eukaryotic transcription is a highly-regulated process with many proteins involved[5,6]. The transcription reaction is initiated by the binding of transcription initiation factors to gene promotors to form a pre-initiation complex (PIC). RNA polymerases are then recruited by PIC and enter the elongation phase, during which an elongation complex is formed to regulate the processivity of the polymerases. In metazoans, elongating Pol II complexes frequently pause at certain DNA sequences, especially at promoter-proximal regions. Paused Pol II complexes are stabilized by DRB (5,6-Dichlorobenzimidazole 1-β-D-ribofuranoside) sensitive inducing factor (DSIF) and negative elongation factor (NELF). The pausing effect can be released by positive elongation factor b (P-TEFb)[7]. The elongation complex and the RNA product move continuously along the

DNA template until reaching the termination sequence, where the polymerases fall off, and nascent RNAs are released.

Epigenetics factors, including histone and DNA modifications, were known to regulate transcription (Figure 1.1). Eukaryotic DNA is packed into nucleosomes, which are composed of DNA and five histone proteins, including H1, H2A, H2B, H3, and H4[8]. These histones process disordered N-terminal tails that are post-translationally modified. These modifications are highly-dynamics, and their association with gene transcription is rigorously explored[9]. Covalent modifications on DNA, especially 5-methylcytosine (5mC) and its oxidative derivatives, are also associated with transcription activities[10]. *N*^{6}-methyldeoxyadenosine (6mA) was reported to play roles in transcription regulation[11-13], but its level in mammalian genomes tends to be extremely low[14].

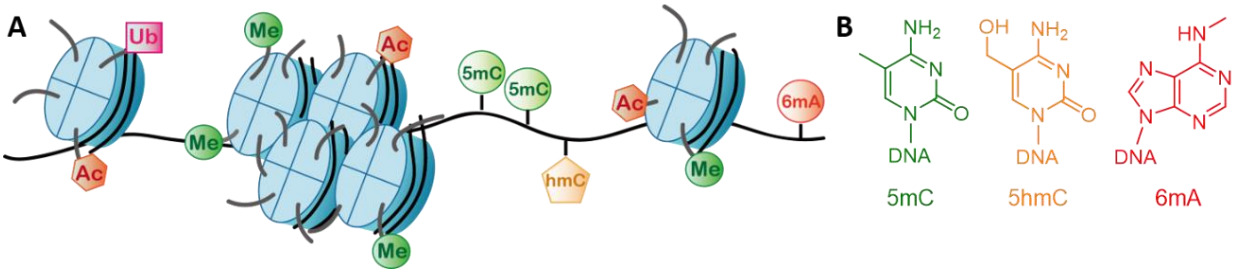

Figure 1.1 Distribution and structure of epigenetic factors that affect transcription

(**A**) The distribution of histone and DNA modifications that are associated with transcription. (**B**) The chemical structures of DNA 5mC, 5hmC, and 6mA modifications.

1.2 RNA secondary structures: information beyond the sequence

After being transcribed, RNA starts to play a critical role in gene expression and its regulation[15,16]. The functions of RNAs are closely related to their structures, which can be decomposed into primary structure, secondary structure, and tertiary structure. The primary structure of RNA refers to the atomic composition and the chemical bond connection, which is determined during transcription by its template DNA sequence. RNA chains intrinsically forms

base-pairing structures through hydrogen bonding, which is referred to as secondary structures. The tertiary (3D) structure is the shape of RNA molecules in three-dimensional space (Figure 1.2).

RNA secondary and tertiary structures can change dynamically in response to cellular environments such as temperature, pH, ligands, and protein binding. The correct folding of RNAs is essential for their proper functions. For instance, riboswitch RNAs usually contain unique structure segments, which is vital for its conformation change upon the binding of specific ligands[17]. The cloverleaf secondary structure and the L-shaped tertiary structure of tRNAs[18] facilitates their fitting into the P and A sites of the ribosomes. The double-stranded region on the precursor of microRNAs (pre-miRNAs) is indispensable for its recognition and processing by the Dicer protein. The Argonaute (AGO) protein has to recognize duplex structures on mature miRNAs to target mRNAs with complementary sequences[19-21].

With the recent advent of transcriptome-wide RNA secondary structure mapping techniques[22-31], RNA structure was found to play roles in almost every step of mRNA post-transcriptional regulations, such as polyadenylation, translation, and modifications. Regions near mRNA start codon are usually less structured, which could facilitate ribosome binding and translation initiation[30]. Guanine-enriched RNAs with unique sequence motifs tend to fold into G-quadruplex (rG4), which was reported to affect miRNA targeting[32]. Regions from -15 to -2 nt upstream of the alternative polyadenylation cleavage site are highly-structured, while regions from -1 to +5 nt are usually single-stranded, indicating the importance of unique secondary structure pattern in maintaining adequate polyadenylation. A human N^6-methyladenosine (m^6A) methyltransferase, METTL16, was reported to preferentially methylated structured RNA such as U6 snRNA and the conserved hairpin structure in the 3' untranslated region (UTR) of *MAT2A*[33,34]. On mRNAs, however, most m^6A-enriched regions tend to be more single-stranded[30]. mRNA m^6A

modifications alter the stability of RNA duplexes, and partially melted duplexes at selected regions facilitate the binding of a group of "m^{6}A reader" proteins such as HNRNPC[35].

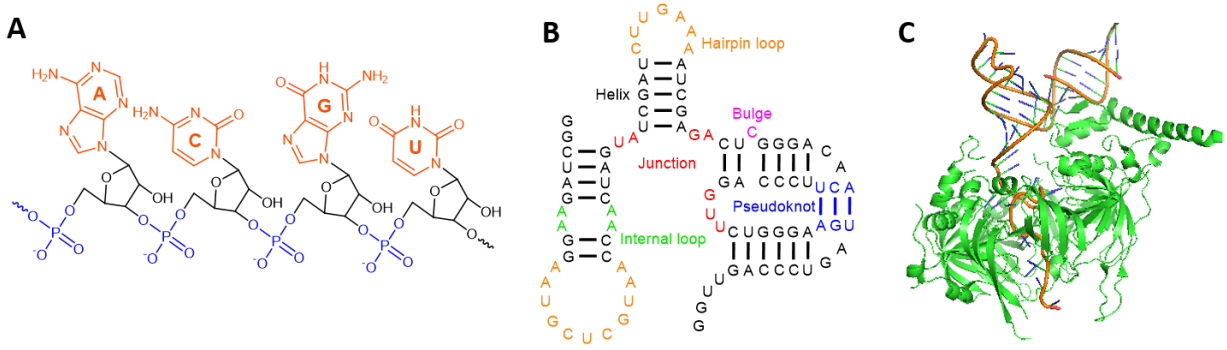

Figure 1.2 Illustrations for RNA primary, secondary, and tertiary structures

(A), (B), and (C) denoted RNA primary, secondary, and tertiary structures, respectively. (C) was generated from published human minimal U1 snRNP crystal structure data (PDB:4PJO) using Pymol.

1.3 Clothes on mRNA: mRNP and higher-order structures

In cells, RNAs are not "naked" molecules. Instead, they interact with other RNA molecules and are bound by RNA-binding proteins (RBPs)[36,37]. RBPs further pack RNA into higher-order structures with different constitutions and topologies, namely, ribonucleoprotein (RNP) complexes. In high eukaryotes, mRNA-containing RNPs (mRNPs) involve tens of thousands of different RNA sequences and hundreds of RBPs, with both of them play essential roles in mRNP assembly and functions[36]. Further aggregation of RNPs results in the formation of different types of RNA granules, with each type containing unique RNA and RBP species[38]. Most of these granules are dynamics, controlling the localization, stability, and translation of their RNA cargo in response to different cellular stresses[39,40].

Since the 1970s, when heterogeneous nuclear ribonucleoproteins (HNRNPs), Y-box proteins (YBXs), and polyadenosine binding protein (PABP) were discovered as the first several RBPs, functions of RBPs were diversely investigated[41-43]. In recent years, the advanced next-

generation sequencing and mass spectrometry techniques provide more insights into their composition, function, and working mechanisms[43,44]. For example, the combination of UV crosslinking and proteins mass spectrometry enables the analysis of RBP atlas, and can discover new RBPs globally[45,37,46], or for specific RNAs-of-interest[47]. UV crosslinking and immunoprecipitation (CLIP) type assays with high-throughput sequencing have been used to study the transcriptome-wide binding position of more than a hundred RBPs at high resolution[48].

Compare to RBPs, the composition and structure of RNA in mRNP complexes are much less understood, potentially due to the following reasons[49,50]. 1) most mRNPs and RNA granules are dynamic, with the compositions and interactions responsive to cellular environments; 2) RNA chains are highly flexible; some of the conformations can be transient and are hard to be captured; 3) RNA tertiary structures can be more heterogeneous than their secondary structures; many different conformations can co-exist in the same cells. Therefore, although methods have been developed to study local RNA folding and RNA-RNA interactions, very few studies have been done for understanding RNA conformation in 3D space within the context of RNPs.

1.4 Probing DNA and RNA structures with small molecules

Various approaches have been developed to understand the topology of nucleic acids. Biophysical characterizations were used to study the exact structure and the structural dynamics at atomic resolution. For instance, the DNA double helix structure was proposed in the 1950s based on results from X-ray diffraction[51], which was then widely applied to study the 3D structure of short RNAs. Nuclear magnetic resonance (NMR) can "visualize" RNA structures in 3D[52], and can detect transient RNA conformation changes[53]. However, these techniques usually require pure DNA/RNA molecules in systems with relatively simple compositions, which hampers their

applications to study RNA structure and its heterogeneity in complex biological contexts. Moreover, it is hard to apply biophysical tools in high-throughput manners.

Biochemical approaches complement biophysical characterizations by offering two major advantages. First, many chemical and enzymatic reactions can happen in vivo, which traps the real-time information in live cells or even live animals. Second, biochemical approaches are compatible with high-throughput detection techniques, such as mass spectrometry and high-throughput sequencing.

Small molecule probes are critical tools for biochemical DNA/RNA structure profiling, which has been applied in both gene-specific or high-throughput manners. In 1980, Peattie and Gilbert reported dimethyl sulfate (DMS) and its derivative, diethylpyrocarbonate, as the first chemical probes for RNA structure mapping in vitro[54]. DMS methylates the Watson-Crick interface (N1 position of adenosine, and N3 position of cytosine) and the major and minor groove (N7 position of guanine, N3 position of adenosine) of RNAs. It is also the first RNA structure probe used in live cells[55,56]. Glyoxal, kethoxal, as well as other molecules containing the α,β-dicarbonyl structure, were reported to react with the Watson-Crick interface (N1, and N2 positions) of guanine[57]. 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) attacks the N3 position of cytosine[58]. The 2′-hydroxyl (2′-OH) acylation (SHAPE) chemistry happens at 2'-OH of sugar[59]. Hydroxyl radicals[60] and the light-activated structural examination of RNA (LASER) probes[61] modified the C8 position of purines (Figure 1.3). Single-stranded and/or physically exposed RNA regions can be modified by these chemical probes, while Watson-Crick base-pairing or interactions with other macromolecules block the labeling reactions. Modified bases cause stops and/or mutations during reverse transcription (RT), which can be detected by loci-specific analysis and high-throughput sequencing.

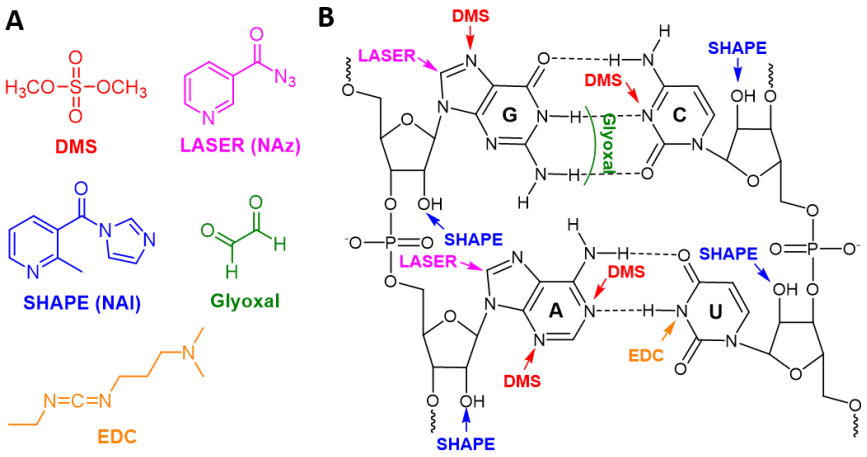

Figure 1.3 Commonly-used small molecule RNA structure probes and the positions they label

(**A**) The chemical structures of DMS, NAz (as an example of LASER probes), NAI (as an example of SHAPE probes), glyoxal, and EDC. (**B**) The reactive sites for each probe on RNA.

1.5 Scope of this thesis

My thesis will focus on developing new chemical biology techniques to monitor the dynamics of transcription, RNA secondary structures, and higher-order RNA interactions. The development of these techniques enables the discovery and understanding of factors that regulated these processes.

Chapter 2 describes the synthesis and reactivity of a new chemical probe, N_3-kethoxal, and its utility in high-throughput RNA secondary structure profiling.

Chapter 3 showed the genome-wide profiling of single-stranded DNA (ssDNA) by using N_3-kethoxal, which monitors transcription dynamics, enhancer activities, and non-B form DNA structures simultaneously.

Chapter 4 presents the application of N_3-kethoxal in mapping RNA-RNA interaction.

Chapter 5 describes the enrichment of METTL4-mediated 6mA modification on mammalian mtDNA and its regulatory role in mitochondria.

Chapter 6 summarizes the applications of the N_3-kethoxal molecule in high-throughput studies. This chapter also envisions potential applications of techniques developed in this thesis, as well as further questions that need to be addressed.

Chapter 2

N_3-kethoxal and its application in RNA structure mapping

2.1 Introduction: general approaches for transcriptome-wide RNA secondary structure mapping

The secondary structure of RNAs plays a critical role in RNA metabolism. The recent advent of next-generation sequencing led RNA secondary structure study to a high-throughput era. The first generation of transcriptome-wide RNA secondary profiling methods relies on structure-specific nucleases[24,25,28]. For instance, RNase V1 preferentially cleaves the phosphodiesterase bond at the 3' of double-stranded RNA, while S1 nuclease prefers to cleave single-stranded RNA.

In recent years, small-molecule probes lead to an increasing number of methods for high throughput RNA structure profiling[26,27,29-31]. While nucleases are usually bulky, small-molecule probes were tailored to be cell-permeable to probe RNA folding in live cells. Moreover, these probes can access to regions that are not accessible by nucleases. As mentioned in section 1.4, many chemicals were reported to preferentially modify the single-stranded regions of RNA versus the double-stranded ones in live cells to "encode" the structure information in vivo. Modified nucleotides cause stops and/or mutations during reverse transcription, which can be detected by high-throughput sequencing. During data analysis, RT-stop/mutation signals are used to estimate the reactivity of the probe on each nucleotide, and higher reactivity reflects less possibility to form secondary structures. Therefore, the in vivo secondary structure of RNA in the whole transcriptome can be inferred at single-nucleotide resolution by selective chemical labeling.

Among all reported RNA labeling molecules, DMS and SHAPE reagents are widely used for high-throughput studies[26,27,29-31]. Combining DMS/SHAPE labeling with different library

preparation procedures creates varieties of methods to detect either RT stop or mutation signatures. Although powerful, DMS and SHAPE labeling also have limitations. For instance, DMS is toxic at high concentrations and does not provide a handle to enrich labeled products. Most SHAPE molecules are hydrolytically unstable and mark the 2'-OH of sugar instead of the bases. New RNA labeling reagents are needed to expand to the toolbox of RNA secondary structure probing.

2.2 Results

2.2.1 The synthesis and the chemical reactivity of N_3-kethoxal

Kethoxal and its analogues were first reported to react with and inactivate an RNA virus since the 1950s[62]. However, the synthesis of kethoxal derivatives were rarely reported. A review of the literature shows that kethoxal preparation was mostly based on oxidation by selenium dioxide (SeO_2) under reflux conditions, following purification by vacuum distillation[63,64]. This method has several limitations. First, the oxidation reaction by SeO_2 always results in a series of byproducts. Second, many kethoxal derivatives, especially the ones with reductive functional groups, cannot tolerate oxidative conditions at high temperatures. Third, vacuum distillation is not suitable for kethoxal derivatives with high boiling points. We tried to apply the SeO_2-based method to synthesis N_3-kethoxal with no success.

Dimethyl-dioxirane (DMD) is a mild oxidant that can quantitatively transform diazoketone to α-keto aldehyde at room temperature. The solvent and excess DMD can be removed by rotation evaporation and yield product with high purity without the need for distillation or column chromatography[65]. Therefore, by applying DMD oxidation, we prepared N_3-kethoxal through a simple three-step synthesis (Figure 2.1). Starting from ethyl 2-bromopropionate, N_3-kethoxal was made in a total yield of 8.5%. The structure and purity of the product were validated by NMR and

high-resolution mass spectrometry (HR-MS). This strategy can also be applied for the preparation of other kethoxal derivatives with various functional groups.

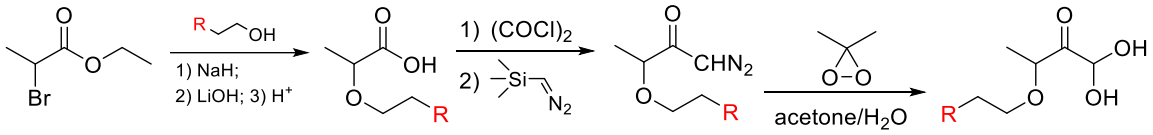

Figure 2.1 The general synthetic route for kethoxal derivatives.

The "R" group denotes azide for N_3-kethoxal. A detailed synthetic procedure is included in the Methods section.

We tested the reactivity of N_3-kethoxal in vitro. We first incubated N_3-kethoxal with guanine at 37 ℃ for 10 min under natural pH, and we detected the N_3-kethoxal-guanine product by HR-MS. In Figure 2.2A, product I was the major product, because the nucleophilic addition between the N1 position of guanine and the aldehyde group of N_3-kethoxal is the first step[57,66]. We then tested whether N_3-kethoxal reacts specifically with G bases by incubating N_3-kethoxal with different nucleic bases, respectively, and monitored the reaction by high-performance liquid chromatography (HPLC). We showed that N_3-kethoxal only reacts with G and N^7-methylguanosine (m^7G) and is inert to A, C, U, and modified G bases with the Watson-Crick interface blocked by methylation (m^1G and m^2G) (Figure 2.2B).

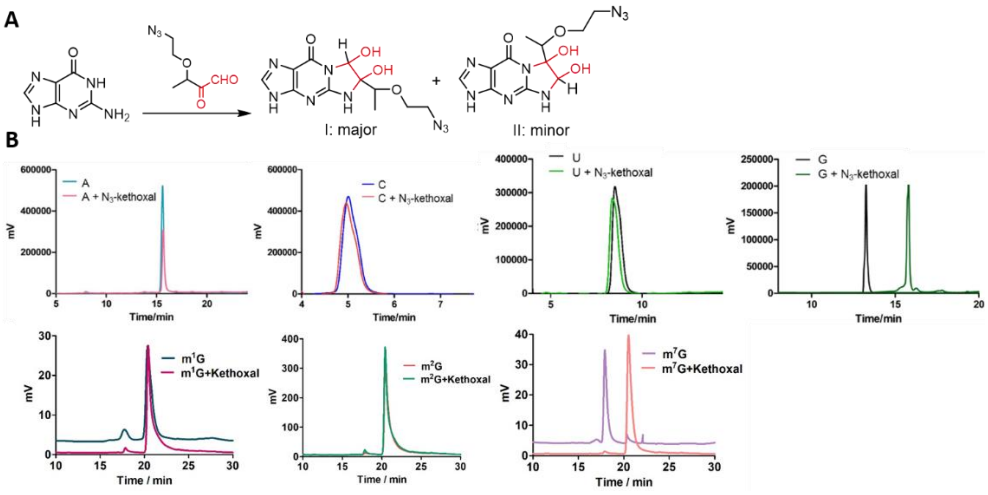

Figure 2.2 The reactivity of N_3-kethoxal on RNA nucleosides

(**Figure 2.2, continued**) (**A**) The scheme of N_3-kethoxal reacting with guanine bases. (**B**) N_3-kethoxal is inert to A, C, and U. For guanine and its derivatives, only G and m^7G can react with N_3-kethoxal. Both m^1G and m^2G, with methylation on N1 or N2 positions respectively, cannot be labeled by N_3-kethoxal. Experiments were repeated twice with similar results obtained.

We tested the reactivity of N_3-kethoxal with an RNA oligo and compared its reactivity with other small molecules probes for RNA labeling. We incubated different probes with an RNA oligo containing four G bases, respectively. We monitored the reaction by matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry. All four guanine bases in the oligo were labeled by N_3-kethoxal, indicating its high reactivity. Moreover, N_3-kethoxal exhibits higher activity than other probes, including glyoxal, NAI-N_3, DMS, and EDC (Figure 2.3).

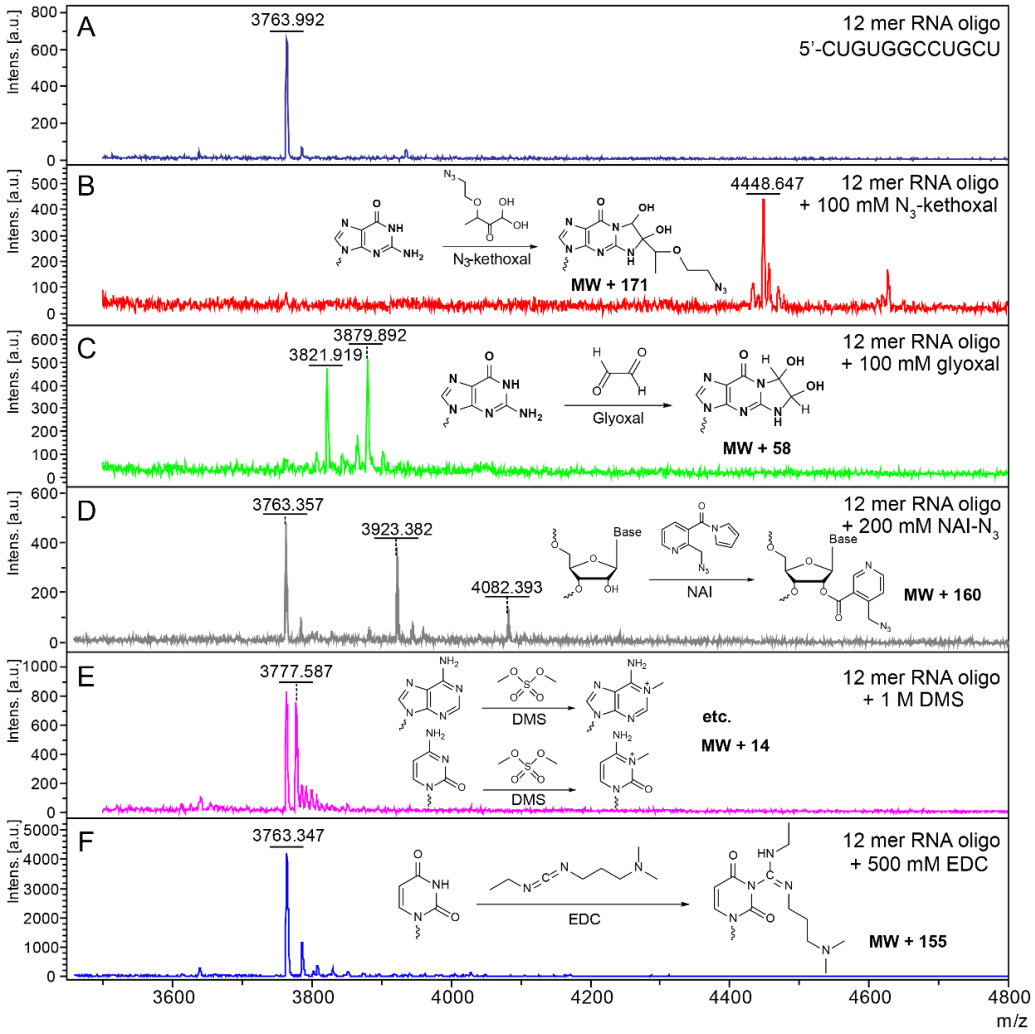

Figure 2.3 The RNA labeling reactivity of different RNA secondary structure probes

(**Figure 2.3, continued**) (**A**) MALDI–MS analysis of 12 mer RNA oligo (5'-CUGUGGCCUGCU). (**B**) MALDI–MS analysis of 12 mer RNA oligo after reacting with N_3-kethoxal (100 mM). All guanines were modified by N_3-kethoxal with the increment of molecular weight (MW) by 684. (**C**) MALDI–MS analysis of 12 mer RNA oligo after reacting with glyoxal (100 mM). The MWs were 3821 and 3879, which means one or two guanines were labeled by glyoxal. (**D**) MALDI–MS analysis of 12 mer RNA oligo after reacting with icSHAPE probe NAI-N_3 (200 mM). The MWs were 3923 and 4082, which represent the addition of one or two NAI-N_3 modifications in RNA, respectively. (**E**) MALDI–MS analysis of 12 mer RNA oligo after reacting with DMS (1 M). The MW of RNA oligo increased to 3777, which means DMS labeled only one base. (**F**) MALDI–MS analysis of 12 mer RNA oligo after reacting with EDC (500 mM). No prominent MW increment was observed. Experiments were repeated twice with similar results obtained.

The reactivity and specificity of N_3-kethoxal were also monitored by gel electrophoresis. As expected, N_3-kethoxal only reacts with single-stranded RNA (ssRNA) but not with double-stranded RNA (dsRNA) (Figure 2.4A). N_3-kethoxal labeled RNA was incubated with biotin-DBCO at 37 ℃ and was then subjected to gel electrophoresis and dot blot. Both assays show that N_3-kethoxal-modified RNA can be efficiently biotinylated through click chemistry (Figure 2.4B).

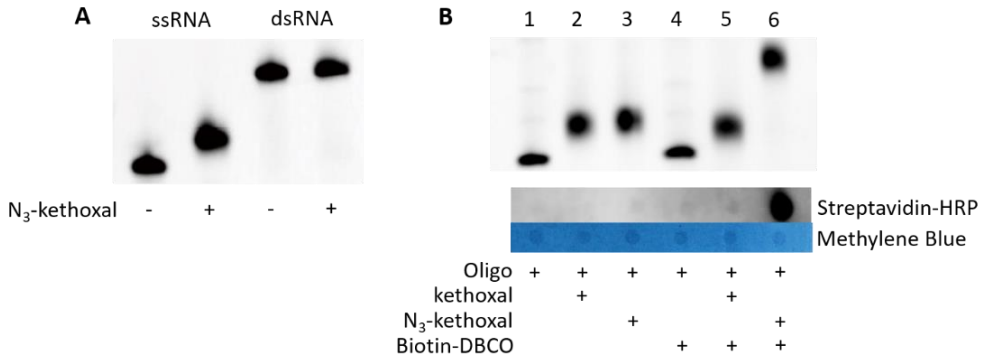

Figure 2.4 N_3-kethoxal selectively labels ssRNA and enables biotinylation

(**A**) Denaturing gel electrophoresis of the reaction between N_3-kethoxal and ssRNA or dsRNA. (**B**) Top: denaturing gel electrophoresis analysis of kethoxal and N_3-kethoxal-mediated labeling reaction (lanes 1–3) and biotinylation reaction (lanes 4-6). Bottom: dot blot analysis of kethoxal- and N_3-kethoxal-modified RNAs before and after the biotinylation reaction.

The kethoxal-guanine adduct was reported to be unstable[67]. The kethoxal labels can be eliminated to yield unmodified RNA in a neutral pH in a short period by adding excessive guanine monomers to trap dissociated kethoxal (Figure 2.5A). We examined this reversibility with the same RNA oligo used for the in vitro activity test and MALDI-TOF analysis. We showed that the

modifications continuously dissociated as time went by, and were almost completely removed after 8 h at 37 ℃ in the presence of 10 mM GTP (Figure 2.5B). As previously reported, borate buffer can stabilize the kethoxal-guanine adduct by interacting with vicinal diols, providing flexibility to manipulate the stability of the N_3-kethoxal-RNA adducts[67] (Figure 2.5A).

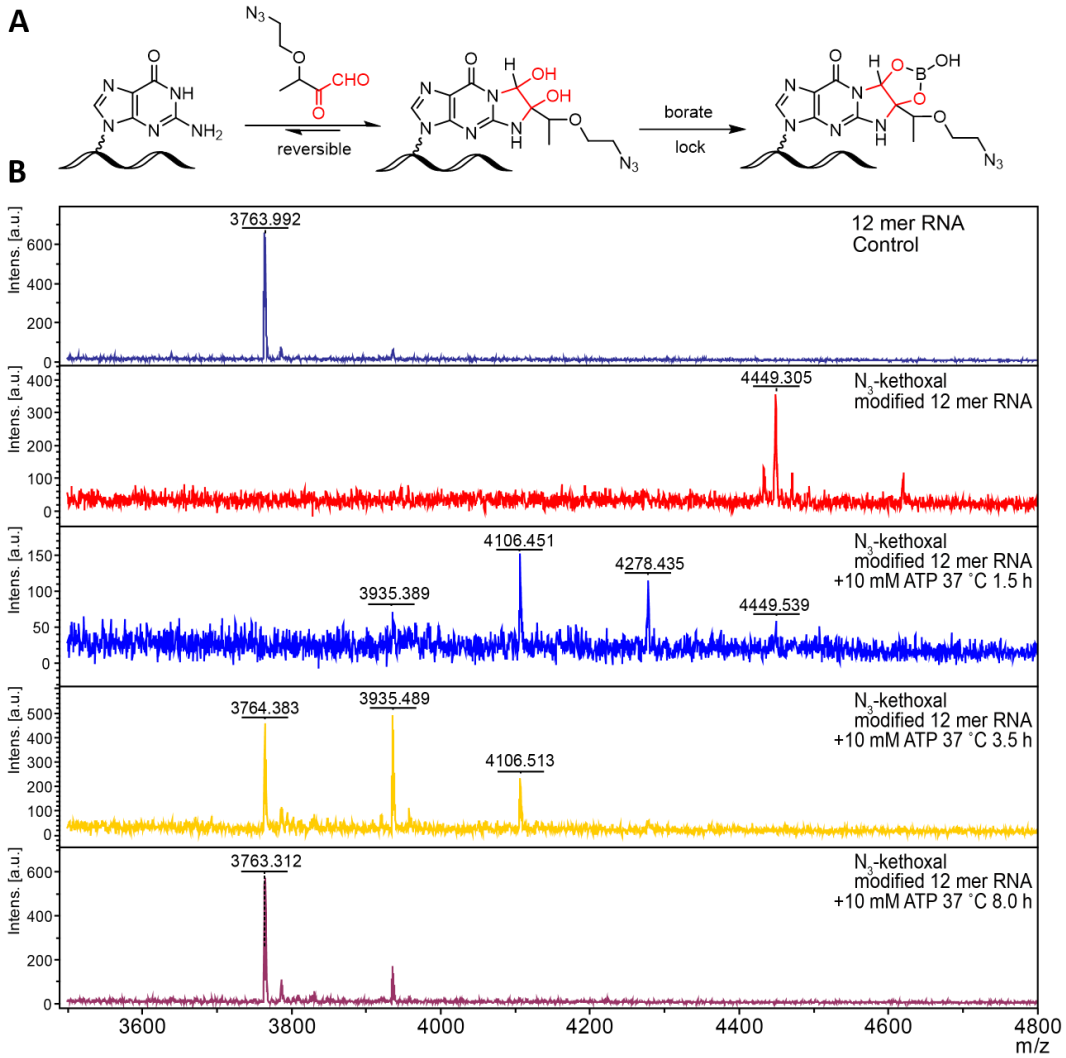

Figure 2.5 N_3-kethoxal labeling is reversible

(**A**) The scheme showing the equilibrium of the N_3-kethoxal-guanine reaction. (**B**) The reversibility of N_3-kethoxal-labeled RNA monitored by MALDI-TOF. The N_3-kethoxal modifications were completely removed after 8 h incubation with 10 mM GTP in neutral buffer at 37 ℃. Experiments were repeated twice with similar results obtained.

2.2.2 The activity of N_3-kethoxal in live cells and the development of Keth-seq

We next evaluated the labeling efficiency in live cells by directly adding N_3-kethoxal into the culture media of mouse embryonic stem cells (mESCs). We isolated total RNA, performed biotinylation by click chemistry, and checked the biotin signal by dot blot assay. We detected the biotin labels within 1 min after adding N_3-kethoxal into the culture media, with the signal saturated after 5 min incubation, suggesting fast cell penetration and labeling reaction (Figure 2.6A). We treated biotinylated mRNA from mESCs at 95 ℃ in the presence of 50 mM GTP for different periods. Dot blot showed that the labels can be removed after 10 min treatment, validating the reversibility of N_3-kethoxal labeling observed by using RNA oligos in vitro (Figure 2.6B).

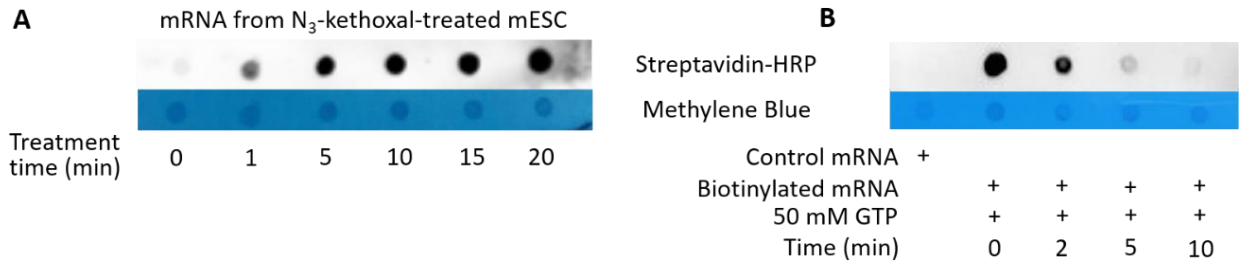

Figure 2.6 N_3-kethoxal labeling kinetics in live cells and its reversibility on mRNA

(**A**) Dot blot assaying the in vivo labeling efficiency of N_3-kethoxal. Biotinylated mRNAs isolated from N_3-kethoxal-treated mESCs were used. (**B**) Analysing the reversibility of N_3-kethoxal labeling by using mRNA isolated from N_3-kethoxal-labeled mESCs. Experiments were repeated twice with similar results obtained.

As kethoxal labeling was known to cause RT-stops, and the azide group provides a handle to enrich modified RNA fragments, we next combined N_3-kethoxal probing with high-throughput sequencing to examine RNA secondary structures in vivo. We name this approach as kethoxal-assisted RNA structure sequencing (Keth-seq). In each experiment, we constructed three different RNA libraries, including an N_3-kethoxal-modified RNA sample, a no-labeling control sample, and an N_3-kethoxal-removal sample (Figure 2.7). For the N_3-kethoxal labeled sample, 3' adapter was ligated onto biotinylated mRNA fragments, and the ligation product was subjected to reverse

transcription. The mRNA-cDNA hybrid was then used for pull-down by using streptavidin beads. Enriched cDNA was circularized and amplified by PCR. In the N_3-kethoxal-removal sample, N_3-kethoxal labels were removed before reverse transcription. No streptavidin pull-down was performed for the no-labeling control sample and the N_3-kethoxal-removal sample (Figure 2.7).

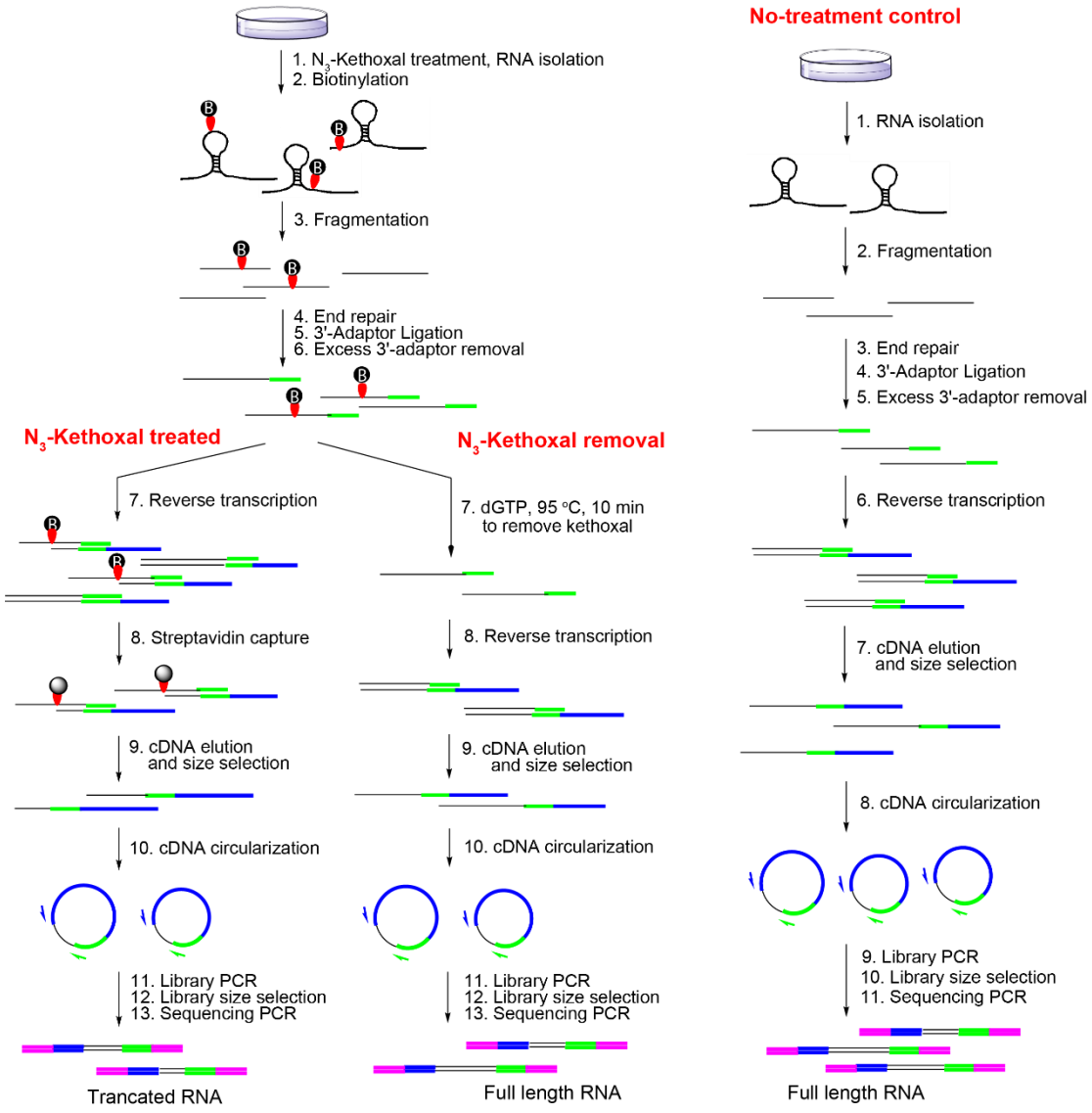

Figure 2.7 Library preparation procedures for Keth-seq and corresponding controls

We observed a high correlation at both RPKM (Figure 2.8A) and RT-stop level (Figure 2.8B) between Keth-seq replicates, indicating the robustness of Keth-seq. We performed Keth-seq after labeling the cells with N_3-kethoxal for different periods, and calculated the percentage of

sequencing reads that stop at guanine. In the N_3-kethoxal samples, the portion of G-stops saturated at 5 min, with more than 80% of the reads stop at guanine, confirming that N_3-kethoxal labeling is fast and is highly selective (Figure 2.8C). This result is also consistent with the dot blot assay that shows the saturation of the biotin signal at 5 min (Figure 2.6A). The ratio of G-stop in N_3-kethoxal-removal samples is similar to that in the no-treatment controls (around 25%), validating that N_3-kethoxal modification was completely removed (Figure 2.8D). RT-stop distributes evenly across all four bases in the no-treatment control sample (Figure 2.8D).

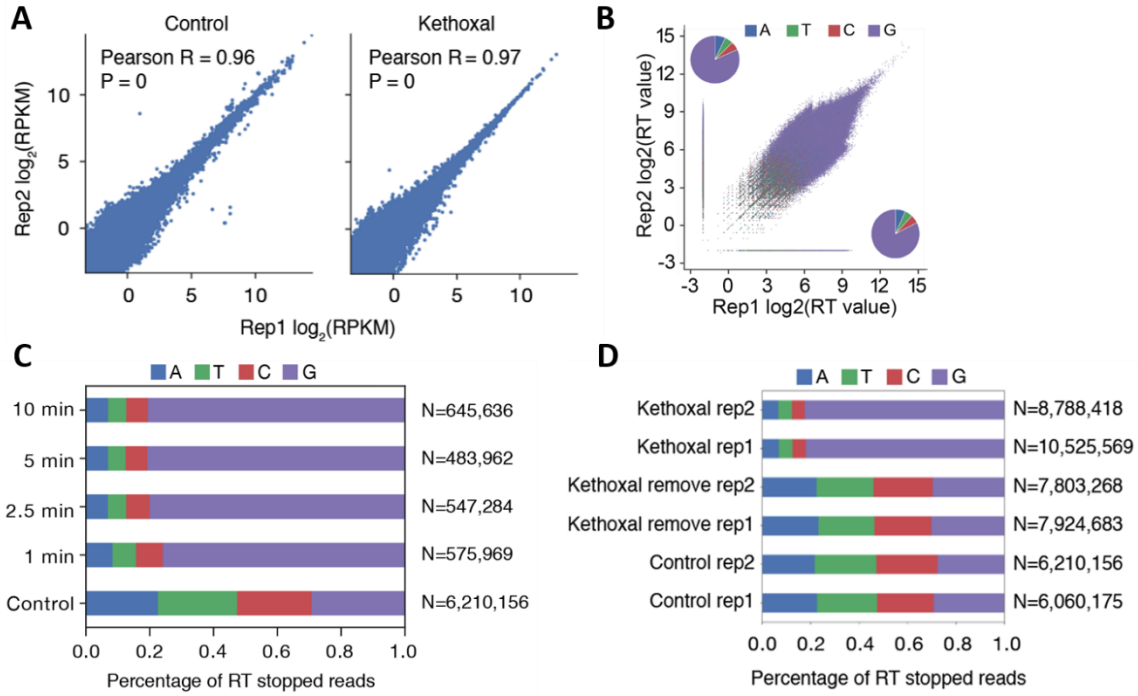

Figure 2.8 Quality control of Keth-seq data

(**A**) RPKM correlation between replicates for N_3-kethoxal (right) and no-treat control (left) sample (n = 69,594 transcripts). The R and P denote Pearson's correlation coefficient and two-tailed p-value. (**B**) A scatter plot of the RT-stop reads distribution between two N_3-kethoxal-treated replicates. 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). (**C**) The distribution of reads that stop at different bases in N_3-kethoxal samples generated from mESCs treated with N_3-kethoxal for 0, 1, 2.5, 5, and 10 min, respectively. (**D**) The distribution of reads that stop at different bases in N_3-kethoxal, N_3-kethoxal remove samples, and no-treat control samples. In (**C**) and (**D**), the N on the right denotes the total number of reads with RT-stops for each sample.

2.2.3 Keth-seq detects transcriptome-wide RNA secondary structures

We next compared Keth-seq with existing RNA secondary structure techniques to evaluate the performance of Keth-seq on RNA secondary structure determination. We first analyzed RT-stops on guanines from Keth-seq data and compared them with icSHAPE data both globally and at the transcript level[30]. For every 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. About 80% of the transcripts show a positive correlation (Pearson correlation coefficient $R \geq 0.4$, Figure 2.9A), 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 guanine reactivity profile of Keth-seq and icSHAPE on mouse 18S ribosomal RNA, whose secondary structure was previously established (RNA STRAND database, ID: CRW_00356)[68]. Keth-seq reactivity profile achieves a higher area under the curve (AUC) than icSHAPE in fitting the 18S ribosomal RNA model (Keth-seq = 0.81, icSHAPE = 0.71) (Figure 2.9B). Both methods agree well with the 18S model on its double-stranded regions (Figure 2.9B). Meanwhile, Keth-seq shows higher reactivity scores than icSHAPE on single-stranded G nucleotides and thus is more accurately revealing unpaired G bases (Figure 2.9C). We then expanded the comparison to 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 (Figure 2.9B).

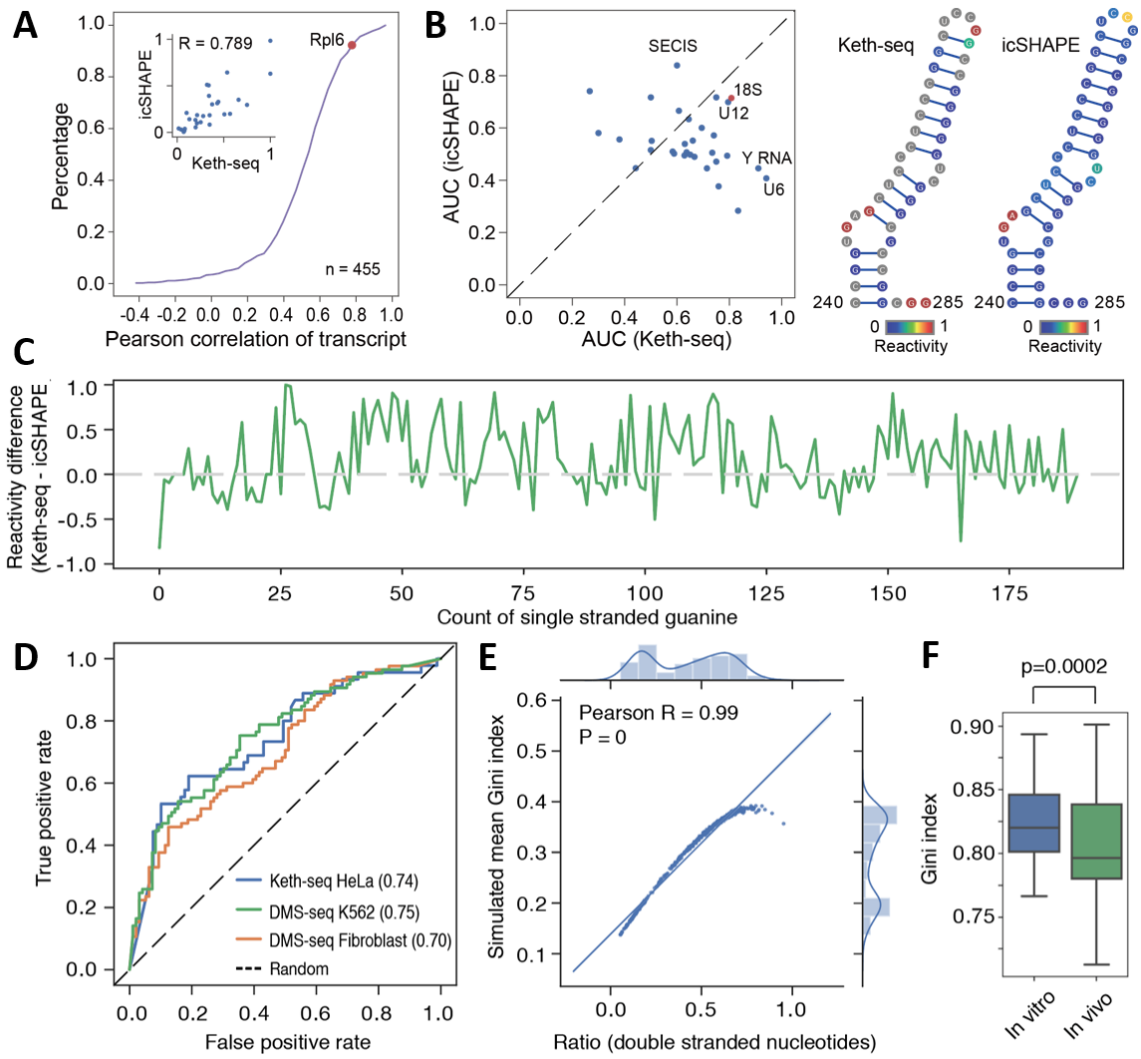

Figure 2.9 The comparison between Keth-seq, icSHAPE, and DMS-seq

(**A**) The accumulative curve of the correlation coefficient between Keth-seq and icSHAPE for all transcripts. For each common transcript, we calculated the Pearson correlation coefficient by using the reactivities on all guanine bases. For example, the inset plot shows the reactivities of all guanines revealed by Keth-seq and icSHAPE, respectively, on the Rpl6 transcript (Pearson correlation coefficient R = 0.789). (**B**) Left, scatter plot showing AUC numbers of Keth-seq and icSHAPE, respectively, on 32 individual RNAs with a known structure (18S ribosomal RNA from RNA STRAND database, and the others from Rfam database). Right, a region (240–285) of 18S ribosomal RNA with both Keth-seq and icSHAPE reactivities filled in the structure model. (**C**) Distribution of reactivity difference between Keth-seq and icSHAPE on single-stranded guanines in mouse 18S ribosomal RNA. (**D**) Comparison between Keth-seq and DMS-seq on human 18S ribosomal RNA. AUCs were shown as the numbers in the brackets. (**E**) Correlation between the ratio of double-stranded regions and calculated Gini index in all mouse RNAs with known secondary structures in the Rfam database (n = 614). The R and P denote Pearson's correlation coefficient and two-tailed p-value, respectively. (**F**) Gini index distribution of all transcripts

(**Figure 2.9, continued**) calculated by using *in vivo* and *in vitro* mESC Keth-seq data (n = 78 transcripts). The p-value is calculated using a two-sided t-test. The box spans first to last quartiles. The centerline denotes median, and whiskers represent 1.5 × the interquartile range.

We also compared Keth-seq and DMS-seq[26] by evaluating their performance on human 18S RNA (ID: CRW_00347) and showed that Keth-seq achieves comparable accuracy to DMS-seq with similar AUCs obtained (Figure 2.9D). Furthermore, we applied Keth-seq to probe RNA structure both in vivo and in vitro for mESCs and calculated the Gini index[26] to measure the structural evenness of RNAs. Consistent with previous findings, the Gini index is positively correlated with the ratio of double-stranded nucleotides (Figure 2.9E). We also observed that RNAs in vitro showed a higher Gini index than that in vivo (Figure 2.9F), suggesting the folding complexity of cellular RNAs and the feasibility of Keth-seq for in vivo detection.

2.2.4 Keth-seq identifies RNA G-quadruplexes (rG4) in vitro and in vivo

The formation of RNA G-quadruplexes (rG4) in purified RNAs has been shown in different studies. However, the in vivo detection of rG4 remains challenging, with controversy remains[32,69]. N_3-kethoxal specifically reacts with N1 and N2 positions of guanine, and the formation of rG4 blocks the reaction. As pyridostain (PDS) is known to stabilize G-quadruplexes structures inside cells[70], the kethoxal reactivity differences before and after PDS induction can serve as evidence for the presence of the rG4 structure in live cells.

We conducted Keth-seq using isolated HeLa RNA or in live HeLa cells in the presence or absence of PDS, respectively. We explored the structure landscapes of previously identified rG4 regions by rG4-seq under PDS treatment in vitro[32]. 95 regions were detected by Keth-seq RT-stops under both native and PDS treatment conditions (Figure 2.10A). In the PDS-treated samples, these regions show a higher Gini index than the control sample, validating the accumulation of rG4 structures under PDS treatment (Figure 2.10B). We also looked at the distribution pattern and potential functional relevance of these rG4 regions. Consistent with previous observations[32], these

rG4 regions preferentially occur at UTRs (Figure 2.10C) and are associated with particular biological pathways (Figure 2.10D) such as translation, transcription, and metabolism.

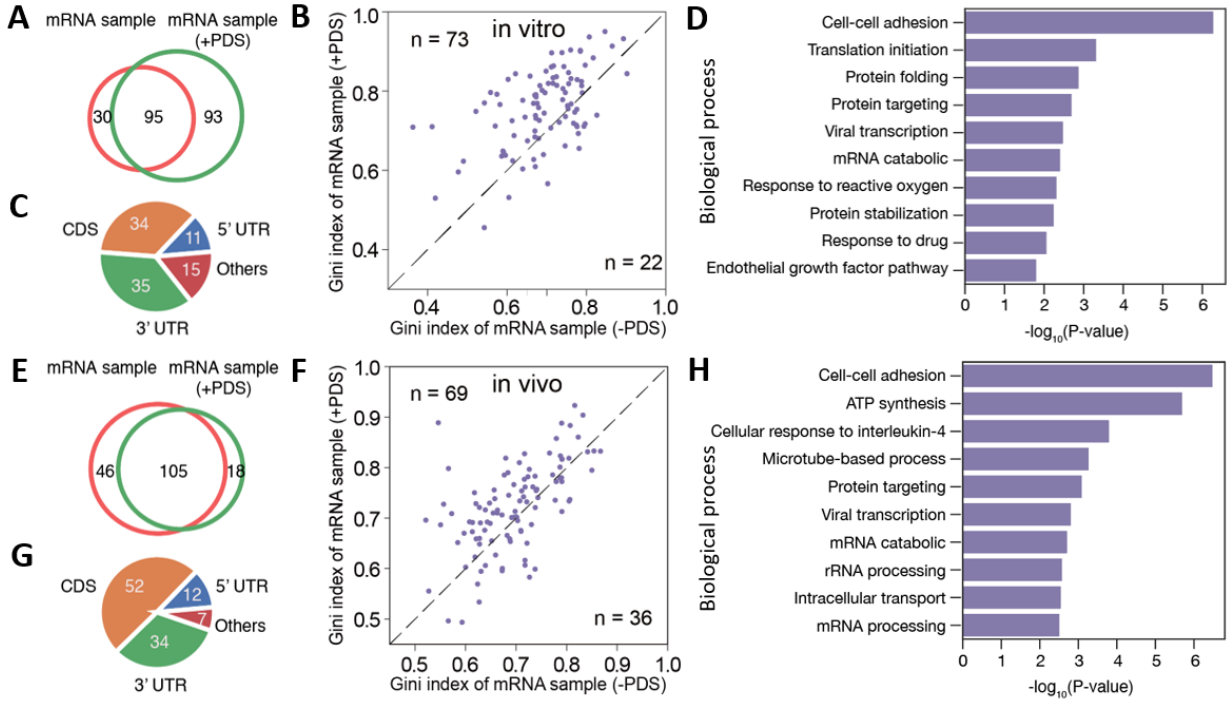

Figure 2.10 Keth-seq detects potential rG4 formation in vitro and in vivo

(**A**) and (**E**) Overlap of rG4 regions between native and PDS treatment conditions revealed by in vitro (**A**) and in vivo (**E**) Keth-seq data. (**B**) and (**F**) Comparison of Gini indexes of rG4 regions under native and PDS treatment conditions. rG4 regions were defined by using in vitro Keth-seq data in and in vivo in (**F**). (**C**) and (**G**) Genomic context distribution of rG4 regions revealed by in vitro (**C**) and in vivo (**G**) Keth-seq data. (**D**) and (**H**) Functional analysis of rG4-containing genes by using in vitro (**D**) and in vivo (**H**) rG4 information. For (**B**) – (**D**) and (**F**) – (**H**), only rG4 regions showing up in both native and PDS treatment conditions were used for analysis.

To further explore whether rG4 can fold in live cells, we performed a similar analysis using in vivo Keth-seq data. We detected 105 previously identified rG4 regions under both native and PDS treatment conditions (Figure 2.10E). Of these 105 regions, 69 showed a higher Gini index under PDS treatment compared with the control (Figure 2.10F), indicating that the rG4 structure could potentially form at these regions in live cells. The genomic context distribution and enriched biological pathways of these regions are both similar to those in vitro (Figure 2.10G-H).

2.3 Discussion and conclusion

In this study, we synthesized N_3-kethoxal under a mild oxidative condition, which can be potentially applied to make a broad range of kethoxal derivatives with different functional groups. N_3-kethoxal exhibit higher reactivity than other RNA labeling probes. We showed that N_3-kethoxal specifically labels single-stranded RNA in vivo and in vitro, and established Keth-seq as an effective method for transcriptome-wide RNA secondary structure mapping in live cells. We envision that N_3-kethoxal labeling can be utilized in RNA enrichment, RNA targeting, and RNA proximity studies in the future.

Keth-seq validates the presence of rG4 in live cells and suggests their potential functions. We noticed that Keth-seq detected only a small subset of rG4s possibilities from the rG4-seq dataset. This could be due to insufficient sequencing depth. It is also possible that chemical labeling only identifies kinetically stable rG4s and miss highly dynamic ones[71]. Although rG4s detected in vitro may not always fold in vivo, our study does suggest that a portion of rG4s can exist in cells.

2.4 Methods

2.4.1 Materials

All chemicals used for N_3-kethoxal synthesis were purchased from Sigma-Aldrich. RNA oligos were purchased from Integrated DNA Technologies, Inc. (IDT), or Takara Biomedical Technology Co., Ltd. DBCO-Biotin was purchase from Click Chemistry Tools LLC (A116-10). All RNase-free solutions were prepared from DEPC-treated water.

mESCs were purchased from ATCC (CRL-1821) and were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1 mM L-glutamine (Gibco), 0.1 mM β-mercaptoethanol (Gibco), 1% (v/v) nonessential amino acid stock (100×, Gibco), 1%

penicillin/streptomycin stock (100×, Gibco), and 1,000 U/mL LIF (Millipore). HeLa cells were purchased from ATCC (CCL-2) and were cultured in DMEM (Gibco 11965) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1% penicillin and streptomycin (Gibco) and grown at 37 ℃ with 5% CO_2.

2.4.2 The synthesis of N_3-kethoxal

2.4.2.1 The synthesis of 2-(2-azidoethoxy)propanoic acid

Sodium hydride (60 % dispersion in mineral oil, 6 g, 0.15 mol) was added to a 250 mL two-necked flask. 50 mL anhydrous THF was then added under N_2 protection. The suspension was vigorously stirred and cooled to 0 ℃. 2-Azidoenthanol (8.7 g, 0.1 mol) in 20 mL anhydrous THF was added dropwise over 20 minutes. The solution was stirred at ambient temperature for 15 min and then cooled to 0 ℃. Ethyl 2-bromopropionate (27.15 g, 0.15 mol) in 10 mL THF was added dropwise. The reaction mixture was warmed to room temperature and stirred overnight under a N_2 atmosphere. 100 mL water was used to quench the reaction, and the resulted mixture was washed by 100 mL diethyl ether three times. The combined organic layers were dried over anhydrous Na_2SO_4, and the solvent was removed.

The crude product was dissolved in 50 mL THF and was added to 40 mL 1 M LiOH solution. The mixture was stirred for 16 h at room temperature before THF was removed, and then 2 M HCl was used to adjust pH to around 2. The mixture was extracted by 100 mL diethyl ether three times. The combined organic layers were dried over anhydrous Na_2SO_4, concentrated, and subjected to silica gel chromatography. The product was collected as a colorless oil (6.67 g, 26 %). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 4.09 (q, *J* = 6.9 Hz, 1H), 3.85 (ddd, *J* = 9.8, 5.9, 3.4 Hz, 1H), 3.66 – 3.58 (m, 1H), 3.55 – 3.46 (m, 1H), 3.42 – 3.33 (m, 1H), 1.49 (t, *J* = 9.4 Hz, 3H). ^{13}C NMR

(101 MHz, $CDCl_3$): δ = 178.5, 74.9, 69.1, 50.7, 18.5. HRMS $C_3H_9N_3O_3^+$ $[M+H]^+$ calculated 160.0717, found 160.0709.

2.4.2.2 The synthesis of 3-(2-azidoethoxy)-1-diazopentane-2-one

Under N_2 condition, 1.59 g (10 mmol) 2-(2-azidoethoxy)propanoic acid was dissolved in 15 mL anhydrous CH_2Cl_2 with one drop of DMF. Oxalyl chloride (926 μL, 15 mmol) was slowly added to the solution and stirred at room temperature for 2 h. The solvent and excess oxalyl chloride were then removed. The residue was dissolved in 50 mL anhydrous CH_3CN and cooled to 0 ℃. 4 mL 2 M (Trimethylsilyl)diazomethane solution in diethyl ether (10 mmol) was added to the mixture dropwise. The reaction was stirred at 0 ℃ overnight. The solvent was removed and the mixture was subjected to silica gel chromatography to afford the product as a yellow oil (620 mg, 33.8 %). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 5.82 (s, 1H), 4.00 – 3.85 (m, 1H), 3.72 – 3.60 (m, 2H), 3.48 – 3.35 (m, 2H), 1.38 (d, *J* = 6.8 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$): δ = 196.9, 80.9, 68.7, 52.3, 50.9, 18.6. HRMS $C_6H_9N_5O_2^+$ $[M+H]^+$ calculated 184.0829, found 184.0822.

2.4.2.3 The synthesis 3-(2-azidoethoxy)-1,1-dihydroxybutan-2-one (N_3-kethoxal)

Dimethyldioxirane (DMD) in acetone solution was prepared as previously described[65]. 11 mL DMD-acetone solution was added into 183 mg 3-(2-azidoethoxy)-1-diazopentane-2-one (1 mmol) in several portions. The reaction mixture was stirred at room temperature until the reaction was complete as monitored by TLC. The solvent was removed to yield the product as a yellow oil. ^{1}H NMR (400 MHz, C_6D_6): δ = [9.5 (m) + 5.5 (m), 1H], 4.75 – 3.40 (m, 1H), 3.26-3.18 (m, 2H), 2.81 – 2.71 (m, 2H), 1.36 – 1.12 (m, 3H). ^{13}C NMR (101 MHz, C_6D_6): δ = 197.7, 102.7, 77.1, 68.7, 50.6, 14.2. HRMS $C_6H_9N_3O_3^+$ $[M+Na]^+$ calculated 194.0536, found 194.0555.

2.4.3 The reaction between N_3-kethoxal and guanine bases

Guanine (100 μM, 2 μL), N_3-kethoxal (1 M in DMSO, 1 μL), sodium cacodylate buffer (500 mM, pH = 7.0, 2 μL), and 6 μL nuclease-free water were mixed and incubated at 37 ℃ for 10 min. The reaction was analyzed by HRMS directly. HRMS $C_{11}H_{14}N_8O_4^+$ $[M+H]^+$ calculated 323.1216, found 323.1203.

2.4.4 The reaction between N_3-kethoxal and other nucleobases

N_3-kethoxal (1 M in DMSO, 1 μL), sodium cacodylate buffer (500 mM, pH = 7.0, 2 μL), and 6 μL nuclease-free water were mixed with different nucleobases (100 μM, 2 μL), respectively. The mixtures were incubated at 37 ℃ for 10 min. The reaction was analyzed by using LC-6AD (Shimadzu, Japan) HPLC instrument equipped with an Inertsil ODS-SP column (5 μm, 250×4.6 mm) (GL Science Inc. Japan). 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 at 35 ℃. The gradient was set as the followings: 0 – 5 min: 5% phase B; 5 – 30 min: 5% - 30% phase B.

2.4.5 Testing the reactivity of different probes by using RNA oligos

In a 10 μL volume, 100 pmol RNA oligo (12 mer, 5'-CUGUGGCCUGCU) was incubated with small-molecule probes in the reaction buffer (100 mM sodium cacodylate, 10 mM $MgCl_2$, pH 7.0) at 37 ℃ for 10 min. The modified RNA was purified by Micro Bio-Spin™ P-6 Gel Columns (Biorad, 7326222). The purified labeled RNA was used for mass spectrometry, gel electrophoresis, and copper-free click reaction with biotin-DBCO. For MALDI-TOF analysis, 1 μL reaction product solution was mixed with 1 μL matrix, which includes 2'4'6'-trihydroxyacetophenone (THAP, 10 mg/mL in 50% CH_3CN/H_2O) and ammonium citrate (50 mg/mL in H_2O) at 8:1 ratio (v/v). The mixture was then spotted on the MALDI sample plate, air dried, and analyzed by Bruker Ultraflextreme MALDI-TOF-TOF Mass Spectrometers.

2.4.6 The reaction between N_3-kethoxal and ssRNA or dsRNA

Gel electrophoresis was applied to study the selectivity of N_3-kethoxal to ssRNA. A fluorescent RNA oligo (5'-FAM-GAGCAGCUUUAGUUUAGAUCGAGUGUA) was used as the ssRNA. It was annealed with its complementary sequence (5- UACACUCGAUCUAAACUA-AAGCUGCUC) to form the dsRNA. The reaction was performed and the product was purified as described in 2.4.5. The purified products were analyzed by denaturing gel electrophoresis (Novex™ TBE-Urea Gels, 15%, Invitrogen, EC6885BOX). The gel was imaged by using Pharos FX Molecular imager (Bio-Rad, USA).

2.4.7 N_3-kethoxal labeling in live cells and RNA isolation

Cells with around 80% confluence were labeled in their culture media supplemented with 10 mM N_3-kethoxal at 37 ℃ in a CO_2 incubator for denoted periods. Cells were then collected and washed once in PBS. Total RNA was isolated from the labeled cells by using RNeasy plus mini kit (Qiagen, 74134). mRNA was isolated from total RNA by using the Dynabeads® mRNA direct purification kit (Thermo, 61011) following the manufacture's protocol.

For in vivo rG4 detection, HeLa cells were treated with 2 μM PDS for 24 h before N_3-kethoxal labeling.

2.4.8 Biotinylation

Purified N_3-kethoxal-labelled RNA oligos or biological RNA samples were resuspended in 78 μL nuclease-free water. 10 μL PBS, 5 μL 500 mM K_3BO_3 (pH 7.0), 5 μL 20 mM WS DBCO-biotin (Click Chemistry Tools, A116-10, DMSO solution), and 2 μL SuperRNase Inhibitor (Thermo, AM2696) were then added. The mixture was incubated at 37 ℃ for 2 h. For RNA oligos, the biotinylated product was purified by Micro Bio-Spin™ P-6 Gel Columns (Biorad, 7326222) and used for gel electrophoresis or MALDI-TOF-MS analysis. For biological samples, RNA was recovered by using the RNeasy MinElute kit (Qiagen, 74204). 350 μL buffer RLT and 900 μL

100 % ethanol was added to the 100 μL reaction solution. The mixture was loaded on the column. The column was washed and the RNA was eluted by following the manufacturer's protocol. Purified biological RNA was used for dot blot and Keth-seq.

2.4.9 Dot blot

1 μL 100 ng/μL RNA was loaded onto the Amersham Hybond-N+ membrane (GE Healthcare, RPN119B). Membranes were air-dried and were crosslinked by UV stratalinker 2400 at 150 mJ/cm^2 twice. The membranes were then blocked overnight in 5% fatty-acid free BSA in PBST (0.1% Tween-20). The second day, the membrane was washed and incubated in streptavidin-HRP (Thermo, S-911) in PBST supplemented with 3% fatty-acid free BSA. The membrane was washed in PBST for 5 times before developed by SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo, 34577).

2.4.10 Testing the reversibility of N_3-kethoxal reaction

The reversibility of the reaction was analyzed by using both RNA oligos and mESC mRNA. For RNA oligos, the N_3-kethoxal-modified oligos was made and purified as described in 2.4.5. 1 μL 100 mM ATP solution was then mixed with 9 μL purified product, and the mixture was incubated at 37 ℃ for 1.5 h, 3.5 h, and 8 h, respectively. After the incubation, the mixtures were purified by running through Micro Bio-Spin™ P-6 Gel Columns (Biorad, 7326222) and were subjected to MALDI-TOF-MS analysis.

For mESC mRNA, biotinylated mRNA was made as described in 2.4.8 and was incubated at 95 ℃ for 10 min. The RNA was recovered by the RNeasy MinElute kit (Qiagen, 74204) following the procedures in 2.4.8 and was used for dot blot.

2.4.11 Keth-seq library preparation

Labeling live cells with N_3-kethoxal, RNA isolation, and biotinylation were performed as described in 2.4.7 and 2.4.8. For both control (no N_3-kethoxal label) and N_3-kethoxal labeled samples, biotinylated mRNA was suspended into 100 μL water and was fragmented to 150-350 bp size by using Bioruptor Pico at the 30s-on/30s-off setting for 30 cycles. The fragmented RNA was lyophilized to 3 μL, to which 1 μL 10 × T4 PNK buffer, 1 μL SUPERase inhibitor (Thermo, AM2694), 1 μL 500 mM K_3BO_3, 1 μL 10 mM ATP, 1 μL FastAP thermosensitive alkaline phosphatase (Thermo, EF0654), and 1 μL T4 PNK (NEB, M0201S) was added. The mixture was incubated at 37 ℃ for 1 h for end repair. After the incubation, 1 μL 20 μM 3' adapter (/5rApp/TGGAATTCTCGGGTGCCAAGG/3ddC/, from IDT), 1 μL 10 × T4RNL2tr buffer, 1 μL 100 mM DTT, 6 μL 50% PEG8000, and 1.5 μL T4RNL2tr K227Q (M0351S) were added to the mixture, which was then incubated at 16 ℃ for 16 h. The reaction was then mixed with 29.5 μL H_2O and was subsequently subjected to purification by RNA clean & concentrator-5 kit (Zymo, R1013) following the manufacturer's protocol. The excess 3' adapters were removed by incubating 13 μL eluted RNA sample with 2 μL NEB buffer 2 (B7002S), 2 μL 5'-deadenylase (M0331S), and 1 μL 500 mM K_3BO_3 at 30 ℃ for 30 min, followed by adding 2 μL $RecJ_f$ (NEB, M0264S) and incubating at 37 ℃ for 1 h. The RNA was recovered by using RNA clean & concentrator-5 kit (Zymo, R1013) following the manufacturer's protocol. For the N_3-kethoxal labeled sample, RNA was eluted in 12 μL nuclease-free water, with 10.5 μL used for the next step, and 1.5 μL saved to make the N_3-kethoxal-remove sample. For the control (no N_3-kethoxal label) sample, RNA was eluted in 10.5 μL volume and was directly used for the next step.

To remove N_3-kethoxal, 1.5 μL RNA solution was mixed with 1 μL SUPERNase inhibitor (Thermo, AM2694), 5 μL 100 mM GTP (Thermo, R0461), and 2.5 μL nuclease-free water, and was incubated at 95 ℃ for 10 min before cleaned up by RNA clean & concentrator-5 kit.

Reverse transcription was then performed for all three samples. For each sample, 10.5 μL RNA was mixed with 1 μL 1 μM RT primer (/5Phos/DDDNNAACCNNNNGAT-CGTCGGACTGTAGAACTCTGAACAT/iSp18/GGATCC/iSp18/TACCTTGGCACCC, from IDT) and 1 μL 500 mM K_3BO_3. The mixture was incubated at 70 ℃ for 5 min followed by a ramp to 25 ℃ in 45 seconds, and was then incubated at 25 ℃ for 1 min. 4 μL 5× first strand buffer, 1 μL 10 mM dNTP mix (Thermo, 18427013), 1 μL Superscript III reverse transcriptase (Thermo 18080044), 1 μL 100 mM DTT, and 0.5 μL SUPERNase inhibitor (Thermo, AM2694) was added to the RNA. The mixture was then incubated at 25 ℃ for 3 min, followed by 42 ℃ for 7 min, and 52 ℃ for 30 min.

After reverse transcription, products from the control and the kethoxal-remove samples were kept at 4 ℃, while the one from the N_3-kethoxal sample was subjected to streptavidin pull-down. For each sample, 20 μL Dynabeads MyOne streptavidin C1 (Thermo, 65001) was washed twice with 1 mL binding buffer (100 mM Tris-HCl pH 7.0, 10 mM EDTA, 1 M NaCl), and was then resuspended into 10 μL binding buffer. The beads were then mixed with the reverse transcription mixture and incubated at room temperature for 45 min with rotation. The supernatant was then removed and the beads were washed by using 500 μL wash buffer (10 mM Tris-HCl pH 7.0, 1 mM EDTA, 4 M NaCl, 0.2% Tween-20) 5 times followed 500 μL PBS twice. The beads were then incubated at 37 ℃ for 30 min in a mixture of 30.5 μL nuclease-free water, 12.5 μL D-biotin, 5 μL 10× RNase H buffer, 1 μL RNase H (NEB, M0297S), and 1 μL RNaseA/T1 (Thermo, EN0551). 1 μL DMSO was then added into the sample, which was then incubated at 95 ℃ for 4 min. The supernatant was collected and subjected to clean-up. In the meantime, the 20 μL cDNA solution from the control (no N_3-kethoxal label) and the kethoxal-remove samples were incubated with 21 μL nuclease-free water, 5 μL 10× RNase H buffer, 2 μL RNase H (NEB, M0297S), and 2

μL RNaseA/T1 (Thermo, EN0551) at 37 ℃ for 30 min. The reactions were then subjected to cleanup together with the supernatant from the N_3-kethoxal samples. The cleanup was done by using the DNA clean and concentrator-5 kit (Zymo, R1013). For DNA binding, 350 μL DNA binding buffer and 350 μL 100% ethanol were mixed with the 50 μL reaction mixture.

Purified cDNA was mixed with 2 × TBU sample buffer (Thermo, LC6876) and denatured at 95 ℃ for 2 min, and was the separated by running on a 6% TBE-urea gel (Thermo, EC6865BOX) at 180 V for 40 min. The gel was stained by using Sybr gold and region with DNA size between 70 – 500 nt was cut out. The gel was sliced into small pieces and incubated in a mixture of 400 μL nuclease-free water, 40 μL 5 M ammonium acetate, and 2 μL 10% SDS at 50 ℃ for 3 h. The liquid was then collected and the DNA was purified from it by ethanol precipitation.

The recovered cDNA pellet was suspended in a mixture of 14 μL nuclease-free water, 2 μL 10 × CircLigase buffer, 1 μL 1 mM ATP, 1 μL $MnCl_2$, and 2 μL CircLigase II (Epicenter, CL9025K), and was incubated at 60 ℃ for 2 h. The circulated cDNA was then mixed with 30 μL nuclease-free water and was purified by using the DNA clean and concentrator-5 kit (Zymo, R1013), with 350 μL DNA binding buffer and 350 μL 100% ethanol added for DNA binding.

The DNA was then amplified by two rounds of PCR. The actual PCR cycle numbers were determined by qPCR. The PCR reaction mix was assembled by mixing eluted DNA (24 μL) with 1 μL 5 μM F/R short primer mix (F: 5'-TGGCACCCGAGAATTCCA; R: 5'-TTCAGAGTTCTACAGTCCGA, from IDT) and 25 μL Phusion high-fidelity PCR master mix (NEB, M0531S). cDNA was amplified in a thermos cycler with the following setting: 98 ℃ 30 s, then 98 ℃ 10s, 60 ℃ 30 s, 72 ℃ 30s for several cycles. The PCR product was subjected to size selection by running on a 6% TBE-urea gel. The region with DNA size between 80 – 300 nt was cut out, and DNA was extracted from the gel as described above. The purified pre-amplified DNA

in 23 µL water was mixed with 25 µL Phusion high-fidelity PCR master mix (NEB, M0531S) and 1 µL forward primer, and 1 µL reverse primer from the TruSeq® Small RNA Sample Prep Kit (Illumina, RS-200-0012). The index PCR was performed following the same PCR program as described above for 3-5 cycles. The final library was then purified by cutting a 3% low melting point agarose gel followed by recovery by QIAquick gel extraction kit (Qiagen, 27804). All libraries were sequenced on Illumina platforms.

2.4.12 Keth-seq data processing

First, readCollapse.pl was used to collapse the reads with the default parameter. A barcode of random hexamer (NNNNNN) ligated to the fragments during library construction was used to identify PCR duplicates from real different fragments with identical sequences. Reads with identical barcode and the insert sequences were marked as PCR duplicates and were filtered before subsequent analysis. But reads with different barcodes were retained, even they may contain the identical insert fragments.

We then used trimming.pl to cut potential adapter sequences (-l 13 -t 0 -c phred33 –a adapter.fa -m 0, adapter sequence: ATGGAATTCTCGGGTGCCAAGGAACTCCAG-TCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA). Next, we mapped the trimmed reads to the mouse (mm10) or human (hg38) transcriptome using Bowtie with default parameters. We calculated RT stop signals using the script calcRT.pl. After evaluating the correlation between different replicates (correlationRT.pl). RT stop signals showed up in both replicates were combined (combineRTreplicates.pl). Signals in both control and kethoxal-treated samples were normalized (normalizeRTfile.pl -m mean:vigintile2 -d 32 -l 32) for subsequent analysis.

A kethoxal reactivity score for each nucleotide was calculated by comparing the RT signal in the N_3-kethoxal sample versus those in the control sample (background) by using the script calcEnrich.pl (-w factor5:scaling1 –x 0.25). The calculation is based on the following formula: A*(RT[kethoxal] – B*RT[control])/BD[control], where RT means the RT stop count of the nucleotide, BD means the base density of the nucleotide, A and B are scaling factors. These scaling factors have been tuned to maximize the correlation between calculated reactivity scores and the structure of mouse 5S ribosomal RNA determined by low-throughput RT-stop gel analysis. We found that the scaling factors perform well at A = 10 and B = 0.25. Finally, to obtain high-quality scores, only nucleotides with adequate sequencing coverage were kept for analysis. We used the script: filterEnrich.pl –T 2 -t 200 -s 5 -e 30. "-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. "-s 5 -e 30" trims away the first 5 and the last 30 nucleotides as they tend to have low sequencing quality scores.

2.4.13 Compare Keth-seq with icSHAPE or DMS-seq

To compare the performance of Keth-seq and icSHAPE, we collected known RNA secondary structure models from public databases, including the mouse 18S ribosomal RNA structure from the RNA STRAND database[68] (CRW_00356) and the other 614 RNA structure models from the Rfam database[72]. Both Keth-seq and icSHAPE sequencing reads were remapped to these specific RNAs, and the reactivity score profiles were calculated. 18S rRNA and 31 other RNAs with structure information available in both techniques were retained for comparison. We applied the receiver operator characteristic (ROC) curves to measure the accuracy of reactivity scores on fitting reference structure models. Under certain cutoffs, each nucleotide can be classified as single-stranded or double-stranded according to the reactivity scores. 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 double-stranded base with a reactivity score lower than the cutoff. AUC is calculated using the signals of guanine nucleotides for Keth-seq, while the signals of all four nucleotides were considered for icSHAPE.

Similarly, to compare Keth-seq with DMS-seq, we collected DMS-seq data for fibroblast and K562 samples and evaluate their performance on human 18S ribosomal RNA (RNA STRAND id: CRW_00347). For DMS-seq, AUC is calculated by using the signals of the adenine and cytosine nucleotides.

2.4.14 Calculation of Gini index

We followed a previously reported method to calculate the Gini index[26]. We assume the reactivity profile of a region is: $(x_1, x_2, x_3, \dots, x_n)$, where x_n is the reactivity score for base n. The reactivity values of the regions were arranged in ascending order and the summation ($Sum = \sum_{j=1}^{n} x_j$) and accumulation ($Acc_j = \sum_{i=1}^{j} x_i$) were calculated. We then calculate the accumulating area by using $Cumulating_{area} = \sum_{j=1}^{n} \left(Acc_j - \frac{x_j}{2} \right)$ and the fair area by following $Fair_{area} = \frac{Sum * n}{2}$. The Gini index value can then calculated as $Gini = \frac{Fair_{area} - Cumulating_{area}}{Fair_{area}}$.

2.4.15 Keth-seq signals on potential rG4 regions

We first converted the genomic coordinates of previously reported rG4 regions in HeLa cells into transcriptome coordinates. The converted regions with >=60% NULL value of structure scores from Keth-seq data were used for subsequent analysis.

2.4.16 Data availability

All sequencing data are available at NCBI Gene Expression Omnibus with the accession number: GSE122096.

Chapter 3

Single-stranded DNA profiling by KAS-seq

3.1 Introduction: genome-wide approaches to transcription and its regulation

Transcription and its regulation determine cell fate and physiological functions, with dysfunctions in transcriptional regulation associated with various human diseases[73]. To understand global transcription regulation, genome-wide sequencing approaches have been developed to analyze the occupancy of RNA polymerases (ChIP-seq)[74], or to detect the presence and level of nascent RNA. Nascent RNA analysis is usually based on run-on assays[75,76], metabolic labeling[77,78], and Pol II-associated or chromatin-associated RNA enrichment[79-83]. Although powerful, these methods also have limitations. Run-on-based methods and Pol II-associated RNA enrichment typically require millions of cells as starting materials. Pol II ChIP-seq could not distinguish whether RNA polymerases are simply bound or are actively engaged in transcription[75]. Metabolic labeling may not be able to measure transient and low-abundant RNA species accurately, such as enhancer RNAs (eRNAs), especially when using limited materials with modest sequencing depth. As most RNAs undergo post-transcriptional processing, their levels are indirect readouts that may not accurately reflect transcription dynamics in situ.

Transcriptionally engaged RNA polymerases resolve DNA double helices and generate single-stranded DNA bubbles. Therefore, we envision that mapping single-stranded DNA (ssDNA) throughout the genome provides a readout of the activity and dynamics of transcriptionally engaged RNA polymerases. Permanganate was previously reported to preferentially oxidize single-stranded thymidine residues[84]. It was subsequently used to reveal Pol II-induced promotor melting in both loci-specific[84,85] and genome-wide manners[86]. The combination of permanganate treatment and S1 nuclease digestion allows the genome-wide identification of non-B form DNA

structures[87]. However, this method requires tens of millions of cells and shows low sensitivity when detecting relatively weak and broad signals derived from Pol II elongation at gene bodies.

3.2 Results

3.2.1 The principle of KAS-seq and its validation

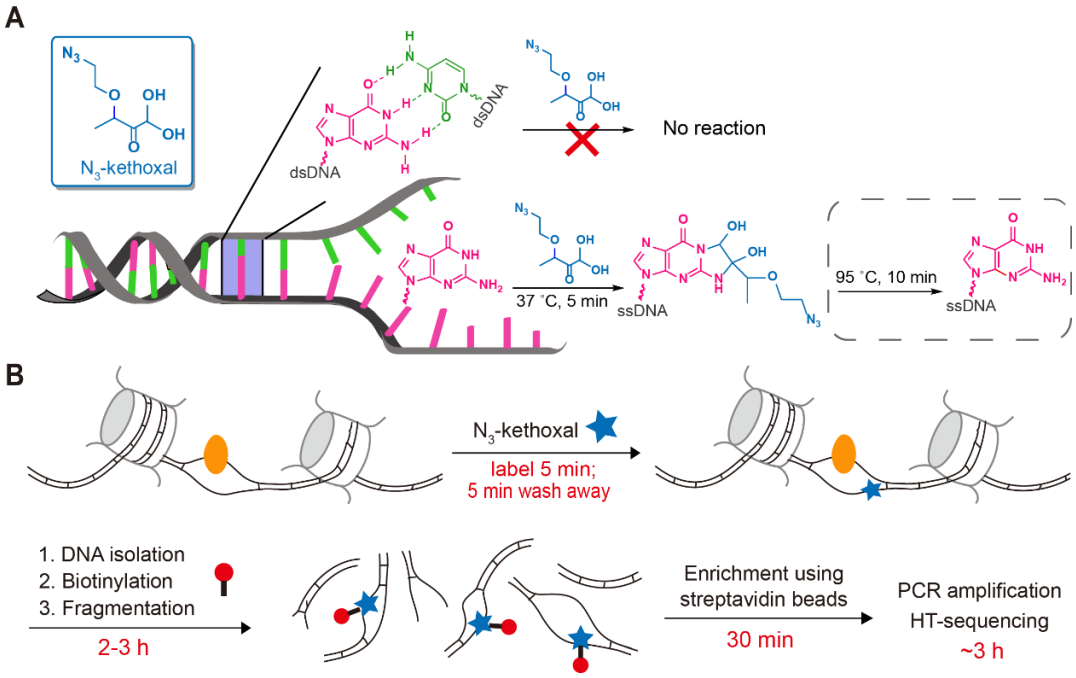

Figure 3.1 Probing single-stranded DNA regions in the genome by using KAS-seq

(**A**) The molecular structure of N_3-kethoxal and how N_3-kethoxal labels guanines in single-stranded DNA but not in double-stranded DNA. (**B**) The scheme of KAS-seq. N_3-kethoxal (blue star) reacts with single-stranded guanines in the genome (resolved by DNA-binding proteins, such as Pol II, as shown in yellow), which can be further biotinylated (red) and enriched for sequencing. The whole process takes 6-7 h in total, from live-cell labeling to finish library preparation.

In Chapter 2, I showed that N_3-kethoxal provides an effective way to map RNA secondary structures by labeling guanines in single-stranded RNAs in live cells under mild conditions[88]. I reasoned that N_3-kethoxal should also enable specific ssDNA labeling and profiling, because the formation of Watson-Crick base-pairing in dsDNA blocks the labeling reaction (Figure 3.1A). Based on this principle, we developed kethoxal-assisted single-stranded DNA sequencing (KAS-seq). After labeling live cells with N_3-kethoxal, genomic DNA (gDNA) can be isolated and

subjected to biotinylation through 'click' chemistry before being fragmented. The single-stranded fragments can then be enriched through the biotin-streptavidin interaction and subjected to library construction (Figure 3.1B). N_3-kethoxal labels can be removed by short heating at 95 ℃ to avoid affecting PCR amplification[89].

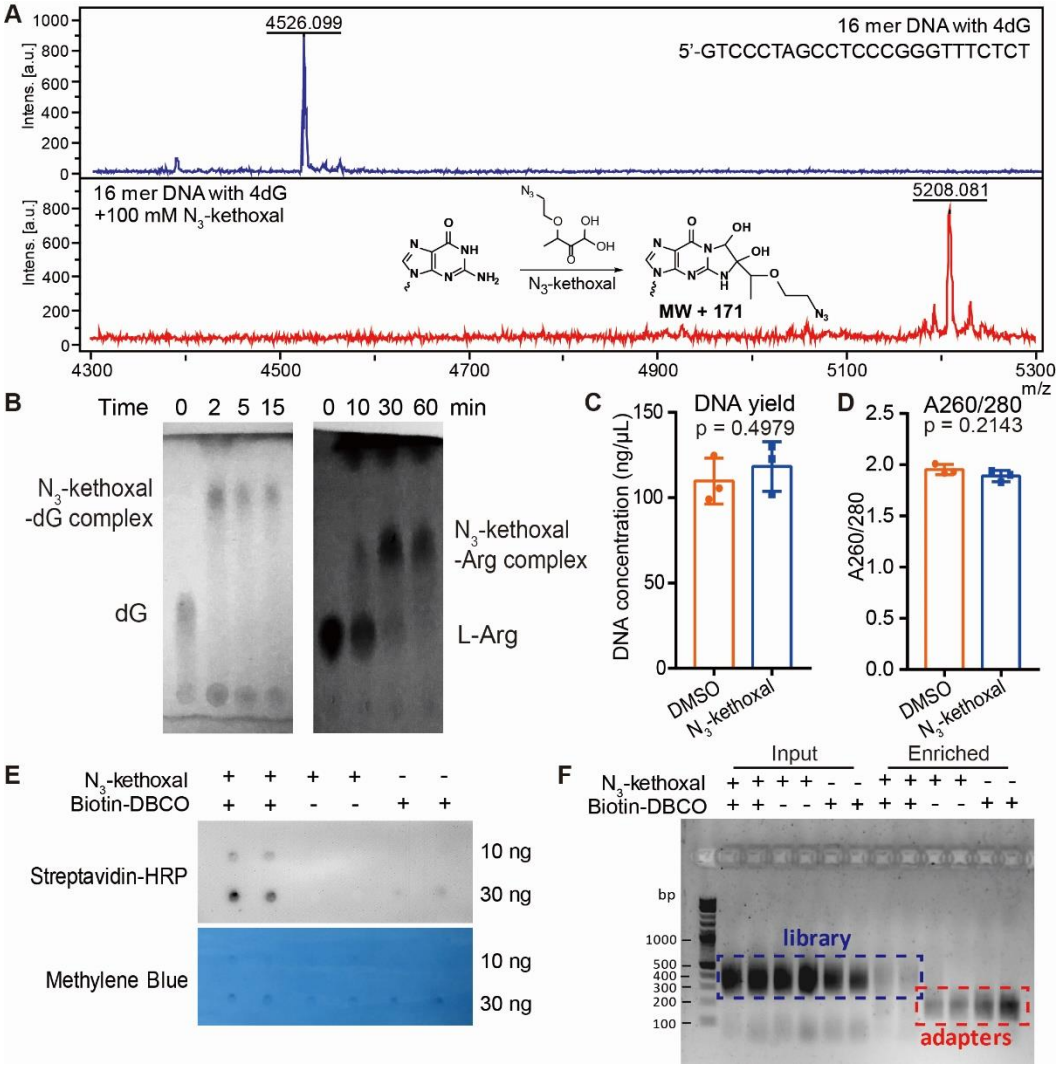

Figure 3.2 Characterization of N_3-kethoxal-based labeling

(**A**) MALDI-TOF analysis of the reaction between a 16-mer DNA oligo and N_3-kethoxal. The experiment was performed in duplicates with similar results obtained. (**B**) TLC analysis of the reaction between N_3-kethoxal and deoxyguanosine (dG, left) or L-arginine (L-Arg, right) after different time intervals. The N_3-kethoxal-dG results were visualized by 254 nm UV light. The N_3-kethoxal-L-Arg results were visualized by ninhydrin staining. The experiment was performed in duplicates with similar results obtained. (**C**) - (**D**) The DNA yield (**C**) and the A260/280 ratio (**D**) of gDNA isolated from N_3-kethoxal-treated and control cells. P values were calculated by using

(**Figure 3.2, continued**) two-sided unpaired Student's t-test (n = 3 independent experiments). (**E**) Dot blot showing biotin signals of the DNA after the biotinylation reaction in the presence or absence of N_3-kethoxal or biotin-DBCO. Results from two replicates were shown for each condition. The experiment was performed in duplicates with similar results obtained. (**F**) Agarose gel image showing the profile of libraries constructed by using input and enriched DNA samples made in the presence or absence of N_3-kethoxal or biotin-DBCO. Results from two replicates were shown for each condition. The experiment was performed in duplicates with similar results obtained.

To prove the principle, we first performed an *in vitro* labeling assay using a synthetic DNA oligo probe containing four deoxyguanosine bases. After incubating the oligo with N_3-kethoxal at 37 ℃ for 5 min, all four deoxyguanosine bases on all oligo molecules were labeled (Figure 3.2A), suggesting a high labeling reactivity of N_3-kethoxal on ssDNA *in vitro*. While N_3-kethoxal reacts with deoxyguanosine bases under neutral conditions within 2 min, very few L-arginine could be labeled within 10 min under the same conditions (Figure 3.2B), indicating that protein labeling could be minimized under the labeling conditions of KAS-seq.

We then performed KAS-seq starting from one million live HEK293T cells and mouse embryonic stem cells (mESCs). N_3-kethoxal labeling does not affect gDNA isolation yield and purity (Figure 3.2C-D). KAS-seq performed in the absence of N_3-kethoxal or the biotinylation reagent (biotin-DBCO) resulted in negligible biotin signals shown by dot blot (Figure 3.2E), nor sufficient enriched DNA for library construction (Figure 3.2F), suggesting minimum background of KAS-seq.

3.2.2 Quality control of KAS-seq data

KAS-seq data show high enrichment efficiency (Figure 3.3A) along with high correlation (r = 0.99, Figure 3.3B) and high peak overlap (Figure 3.3C) between replicates. KAS-seq signals exhibit a similar distribution pattern as Pol II ChIP-seq signals along regions with different G/C contents (Figure 3.3D), suggesting that the G-specific labeling does not notably induce bias, although G/C content effect should be considered for more specified applications of KAS-seq.

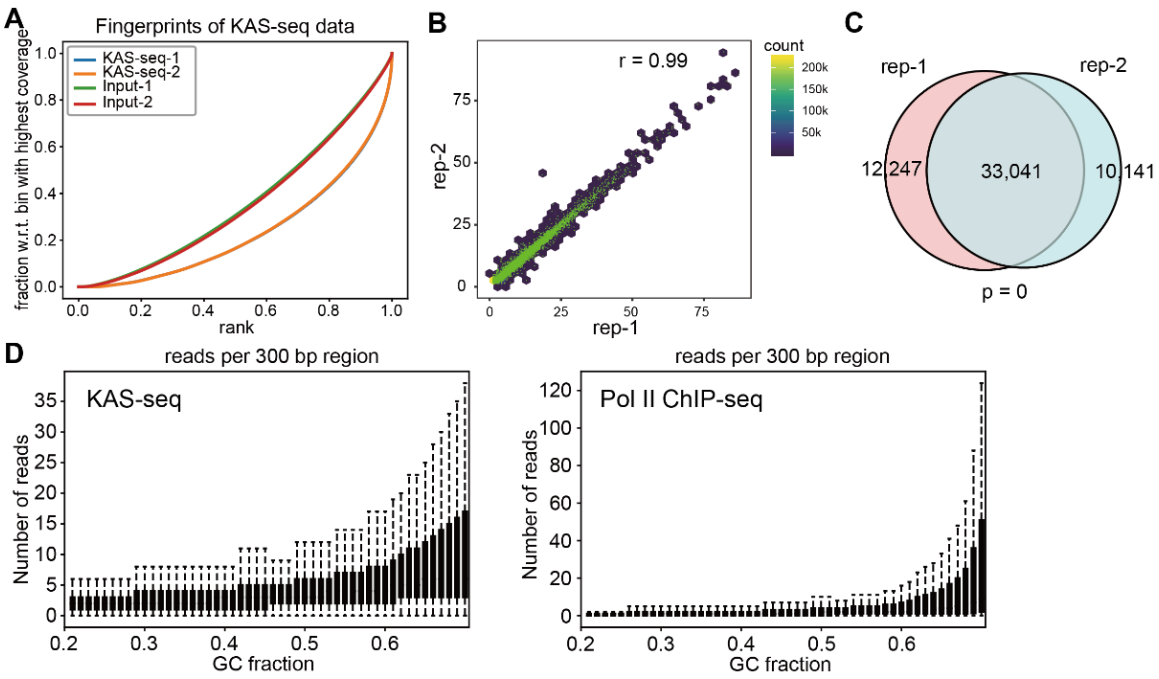

Figure 3.3 Quality control of KAS-seq data

(**A**) Fingerprint plot of KAS-seq libraries and the corresponding inputs in HEK293T cells. (**B**) Pearson correlation scatterplot between two independent KAS-seq replicates (r = 0.99) in HEK293T cells (n = 287,970 10 kb bins in the hg19 genome). (**C**) Peak overlaps between two independent KAS-seq replicate in HEK293T cells. The p-value was calculated using two-sided Fisher's exact test. (**D**) Reads distributions of KAS-seq (left) and Pol II ChIP-seq (right) signals respect to different GC fractions.

3.2.3 KAS-seq signals mark active transcription

KAS-seq reads are considerably enriched at gene-coding regions, especially at gene promotors and transcription termination areas, while depleted at intergenic regions (Figure 3.4A). KAS-seq profile on gene-coding regions revealed a strong and sharp peak around transcription start site (TSS), relatively weak and broad signals that cover the entire gene body, and a strong but broad peak starting from transcription end site (TES) to its downstream regions (Figure 3.4B-C).

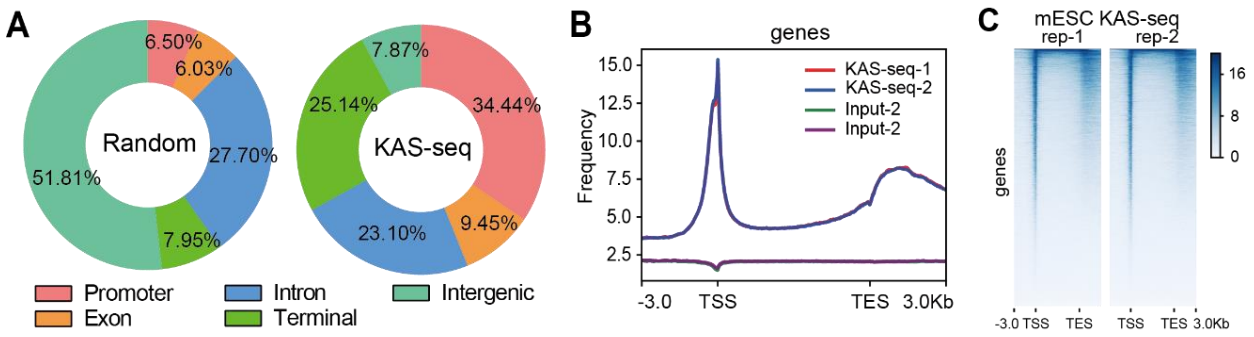

Figure 3.4 An overview of KAS-seq in HEK293T cells and mESCs

(**A**) Genome-wide distribution of KAS-seq peaks in HEK293T cells. "KAS-seq" denotes the percentage overlap of KAS-seq peaks with different genomic features. "Random" denotes the percentage overlap of randomly generated regions with the same number and length of real peaks with different genomic features. (**B**) The distribution of KAS-seq signals at gene-coding regions in HEK293T cells, with 3 kb upstream of TSS and 3 kb downstream of TES shown. (**C**) Heatmap showing reads distribution of two independent KAS-seq replicates at gene-coding regions in mESCs.

KAS-seq signals show positive correlations with histone modifications that mark active transcription, such as H3K4me3, H3K27ac, and H3K36me3, and are negatively correlated with inactive chromatin markers such as H3K27me3 and H3K9me3 (Figure 3.5A). Notably, KAS-seq signals correlate better with H3K36me3 than ATAC-seq results do, indicating that while ATAC-seq serves as a powerful tool to probe chromatin accessibility[90], KAS-seq directly measures transcription activities. KAS-seq signals at TSS overlap with H3K4me3 and H3K27ac, and KAS-seq signals at gene body overlap with H3K36me3 (Figure 3.5B-C). These results collectively suggest that KAS-seq signals are derived from Pol II-mediated transcription.

We also compared KAS-seq with permanganate/S1 footprinting. Both methods show similar sensitivity in detecting the strong “promotor melting” signals. But KAS-seq is much more sensitive on detecting the weaker and broad ssDNA signals on the gene bodies and terminal regions (Figure 3.5D).

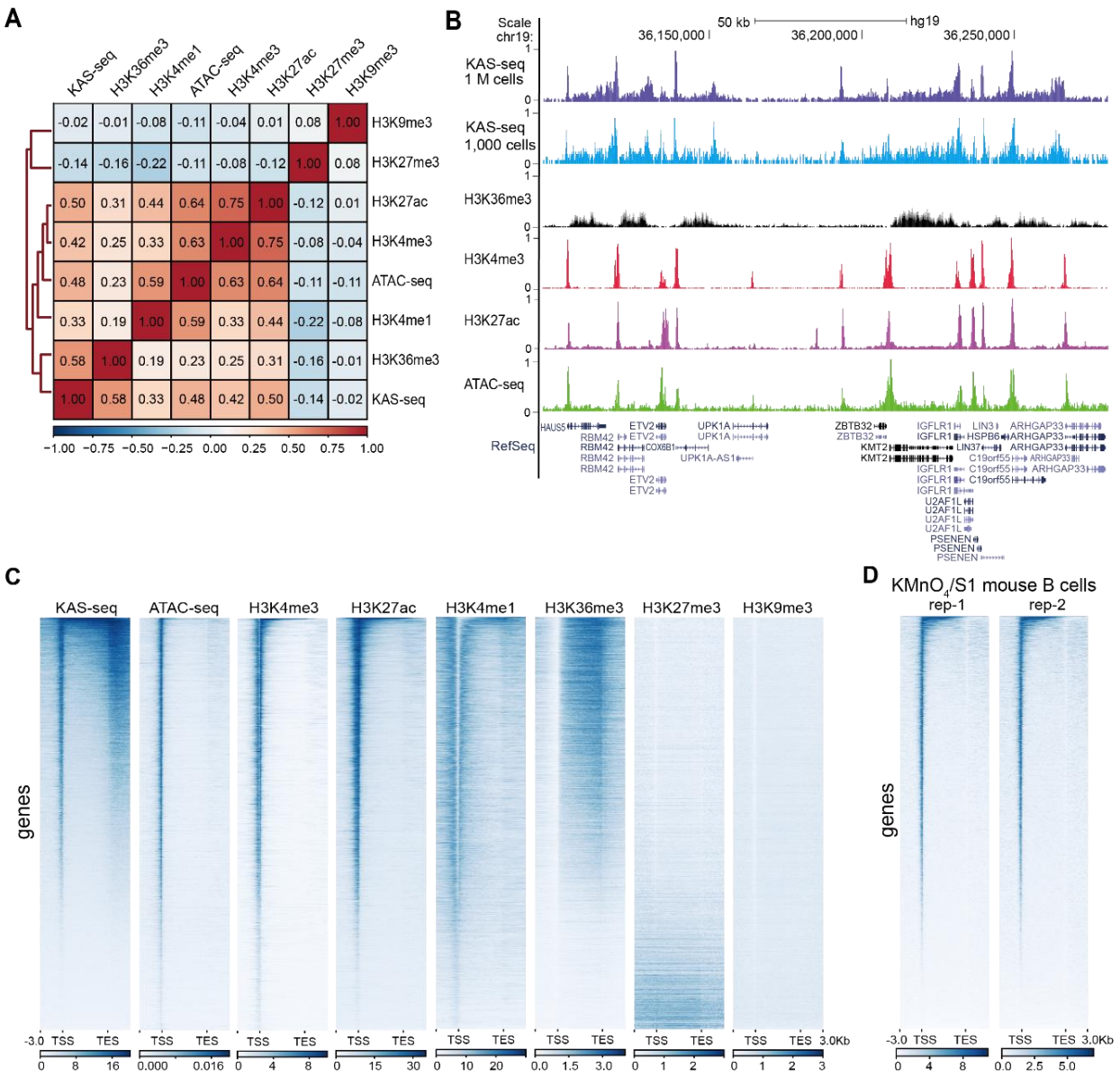

Figure 3.5 KAS-seq correlates with histone modifications that mark active transcription

(**A**) The genome-wide Pearson correlation heatmap among averaged KAS-seq signals, selected histone modifications, and ATAC-seq reads density in HEK293T cells. Heatmap was clustered using hierarchical clustering, with pairwise correlation coefficients noted in each square (n = 302,755 10 kb bins in the hg19 genome). (**B**) A snapshot from UCSC Genome Browser, showing the relationship between KAS-seq peaks, selected histone modifications, and ATAC-seq peaks at a highlighted locus. (**C**) The distribution of KAS-seq signals, ATAC-seq signals, and selected histone modifications at gene-coding regions in HEK293T cells. (**D**) Heatmap showing the distribution of the reads of two $KMnO_4$/S1 footprinting replicates (activated mouse B cells) at gene-coding areas.

3.2.4 KAS-seq works by using 1,000 cells and mouse liver samples

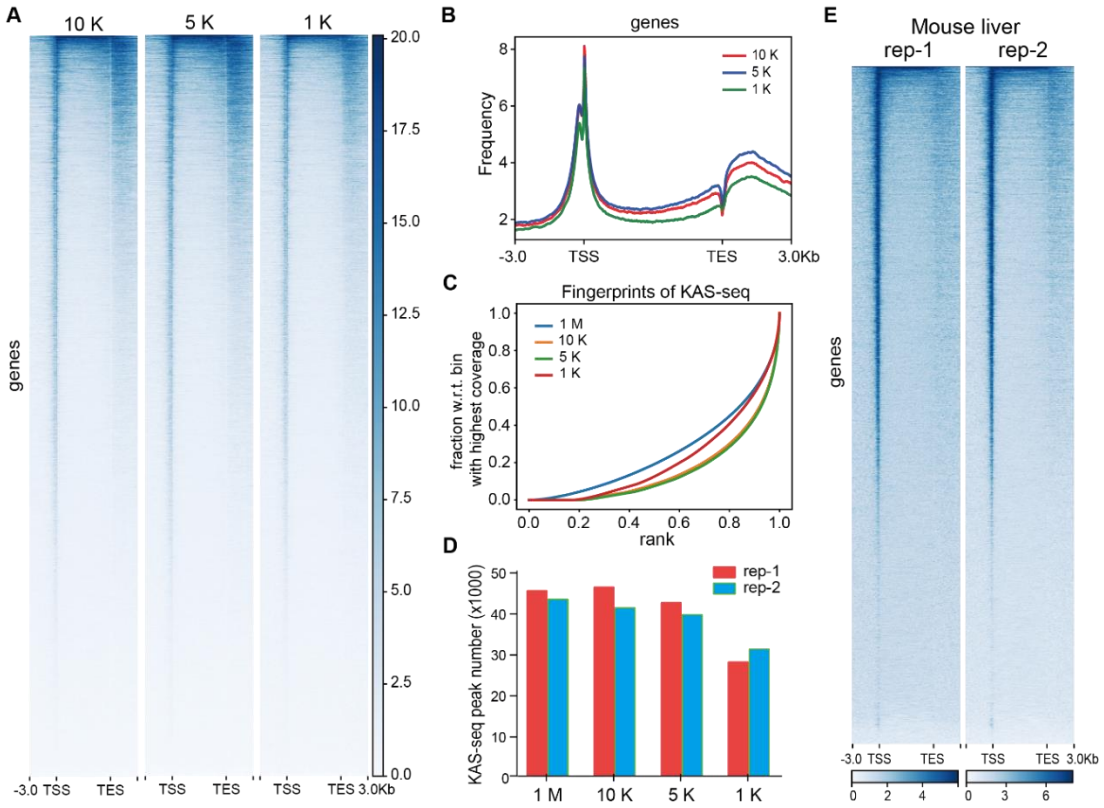

Figure 3.6 KAS-seq using low input cells and mouse liver

(**A**) KAS-seq signal distribution at gene-coding regions revealed by using different numbers of HEK293T cells (n = 26,910 genes). (**B**) Profiles of KAS-seq data at gene-coding regions using different numbers of HEK293T cells. (**C**) Fingerprint plot of low-input KAS-seq libraries. (**D**) Numbers of KAS-seq peaks detected by using different amounts of HEK293T cells. (**E**) Heatmap showing reads distribution of two independent KAS-seq replicates at gene-coding regions generated by using livers from two mice. 1 M: 1 million; 10 K: 10 thousand; 5 K: 5 thousand; 1 K: 1 thousand.

Because of the high guanine labeling reactivity of N_3-kethoxal and the high affinity between biotin and streptavidin, KAS-seq is expected to maintain its sensitivity when using low-input starting materials or primary tissue samples. Indeed, the distribution of KAS-seq signals at gene-coding regions and the overlap with histone modifications remain unchanged by using 10,000, 5,000, or even 1,000 HEK293T cells (Figure 3.5B, Figure 3.6A-B). KAS-seq results with low input cells showed similar enrichment efficiency and captured similar numbers of peaks compared

with KAS-seq libraries generated from 1 million cells (Figure 3.6C-D). KAS-seq performed by using mice liver tissues also show strong signals at TSS, with modest signals on the gene bodies and at TES regions (Figure 3.6E). Thus, KAS-seq is a method suitable for a wide range of potential applications to study rare cell samples and clinical samples in the future.

3.2.5 KAS-seq data correlates well with transcription activity and gene expression level

We next compared KAS-seq in HEK293T cells with GRO-seq and Pol II ChIP-seq in the same cell line. KAS-seq results correlate well with results from these assays (Figure 3.7A). In mESCs, ~95% of KAS-seq peaks on promotors overlap with Pol II ChIP-seq peaks (Figure 3.7B). Reads density of KAS-seq and Pol II ChIP-seq on the gene bodies show a strong positive correlation (Pearson r = 0.81, Figure 3.7C). We then ranked all genes into four groups according to their expression levels based on RNA-seq data (Figure 3.7D), and showed that the strength of KAS-seq signals drops notably in genes with low expression levels (Figure 3.7E).

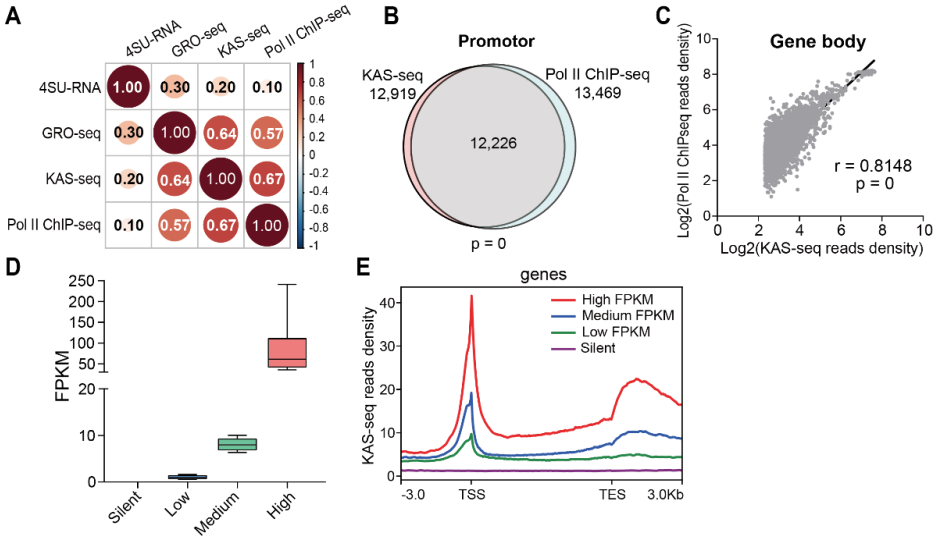

Figure 3.7 The correlation between KAS-seq, Pol II ChIP-seq, GRO-seq, and RNA-seq

(**A**) Genome-wide Pearson correlation heatmap between KAS-seq, Pol II ChIP-seq, GRO-seq, and nascent RNA-seq (4SU-seq) reads density on gene-coding regions in HEK293T cells. Pairwise correlation coefficients are noted in each square (n = 839,684 1 kb bins in the hg19 genome). (**B**) Venn diagram showing the overlap between KAS-seq peaks and Pol II ChIP-seq peaks at promotor in mESCs. The p-value was calculated using two-sided Fisher's exact test. (**C**) Pearson correlation

(**Figure 3.7, continued**) scatterplot (n = 24,359 genes) between KAS-seq and Pol II ChIP-seq at gene bodies in mESCs. The r-value was calculated as a two-tailed probability. (**D**) Genes were grouped according to different expression levels based on RNA-seq. 10-90 percentile of data points are shown, with the centerline showing the median, and the box limits showing the upper and lower quartiles. (**E**) KAS-seq reads density at gene-coding regions of genes with different expression levels (defined by RNA-seq) in HEK293T cells.

3.2.6 KAS-seq reveals Pol II dynamics and defines gene transcription states.

To further validate that transcriptionally engaged Pol II is the primary source of detected ssDNA signals, we treated HEK293T cells with 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) and triptolide, respectively, and then performed KAS-seq. DRB inhibits Pol II release from pausing at TSS, and triptolide inhibits recruitment and loading of Pol II to promoters[6]. While the majority of peaks overlap with those at the native state, after DRB and triptolide treatment, KAS-seq peak numbers decreased by 57% and 93%, respectively (Figure 3.8A). As expected, DRB severely diminished ssDNA signals at gene body and TES regions with increased signals at TSS; triptolide almost completely erased all signals at the entire gene-coding areas (Figure 3.8B-D). These observations confirm that the strong KAS-seq peaks on gene promotors reflect transcription initiation and Pol II pausing near TSS[75,85] and that KAS-seq signals at gene bodies are derived from transcription elongation.

Comparing KAS-seq signals at promotor-proximal and gene body regions enabled us to sort genes into four classes with distinct transcription states: class I, paused and active; class II, paused and inactive; class III, not paused and active; class IV, not paused and inactive (Figure 3.8E). In HEK293T cells, 60% (11,715 out of 19,279) of all genes showed significant pausing signals around TSS, and the majority of these paused genes (10,204 out of 11,715) also showed active Pol II elongation, which is consistent with results obtained previously from GRO-seq[75].

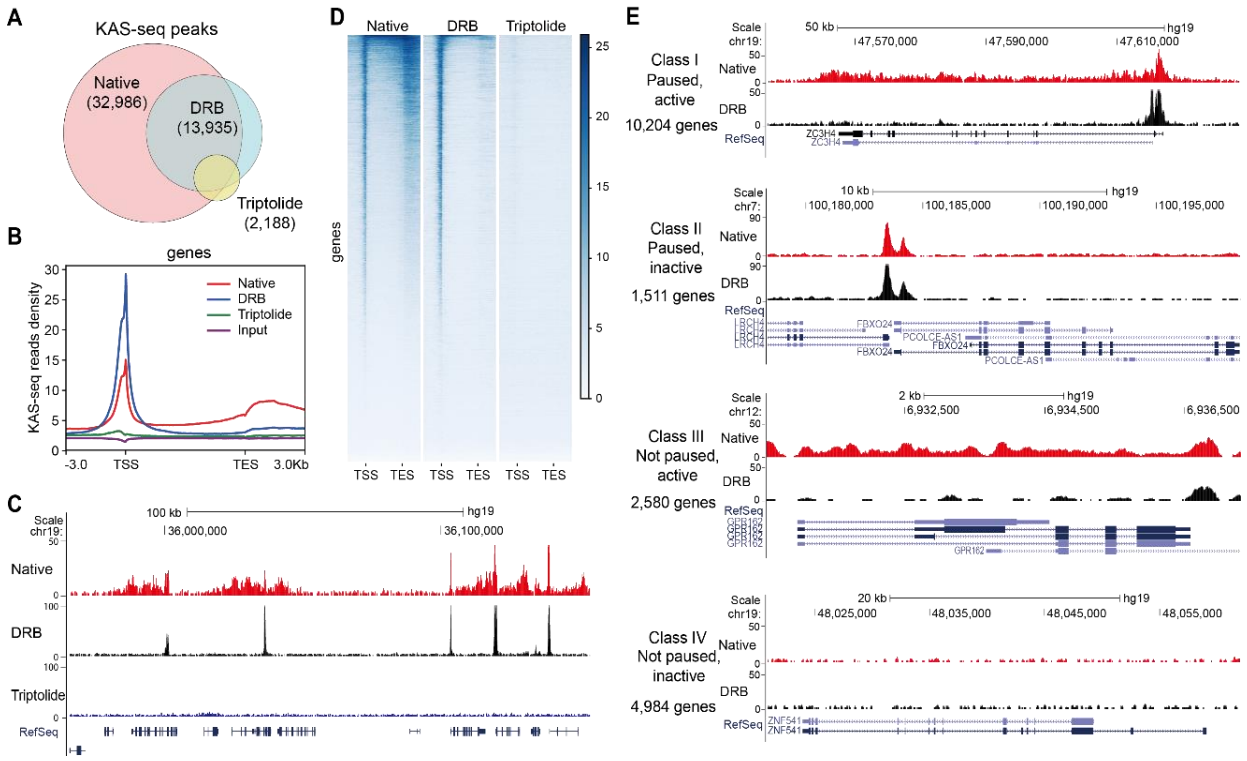

Figure 3.8 KAS-seq reveals Pol II dynamics and defines gene transcription states

(**A**) Venn diagram showing the overlap of KAS-seq peaks in HEK293T cells under native, DRB treatment, and triptolide treatment conditions. The number of overlapped peaks between the two replicates was used in each case. (**B**) Metagene profile of KAS-seq signals at gene-coding regions under control, DRB treatment, and triptolide treatment conditions. (**C**) A snapshot of KAS-seq profiles from UCSC Genome Browser under control, DRB treatment, and triptolide treatment conditions. (**D**) Heatmap showing KAS-seq signal distribution at gene-coding regions under native, DRB treatment, and triptolide treatment conditions. Regions of 3 kb upstream of TSS and 3 kb downstream of TES were shown. (**E**) Defining four groups of genes with different transcription states based on KAS-seq results. In each group, one gene is shown as an example by using the snapshot of KAS-seq signals under native and DRB-treated conditions.

3.2.7 KAS-seq enrich signals at transcription termination regions

Apart from signals on promotors and gene bodies, we also found KAS-seq signals are enriched at transcription termination regions (Figure 3.4A-C). These signals were removed by DRB treatment (Figure 3.9A), indicating that they are derived from Pol II elongation (and pausing) at the termination window. We sorted all genes with KAS-seq signals at this region into three groups according to the length of their termination signals (Figure 3.9B). We then analyzed the

averaged KAS-seq reads density on the entire terminal region of the three groups without observing notable differences (Figure 3.9C), suggesting that KAS-seq does not exhibit length-dependent bias. We calculated the 'termination index' as the ratio of reads density at TES-downstream regions relative to the density in the promotor-proximal regions (Figure 3.9D). KAS-seq revealed a higher termination index than Pol II ChIP-seq and GRO-seq do in the same cell line (Figure 3.9E), suggesting that Pol II accumulation at TES-downstream regions can be more than previously expected.

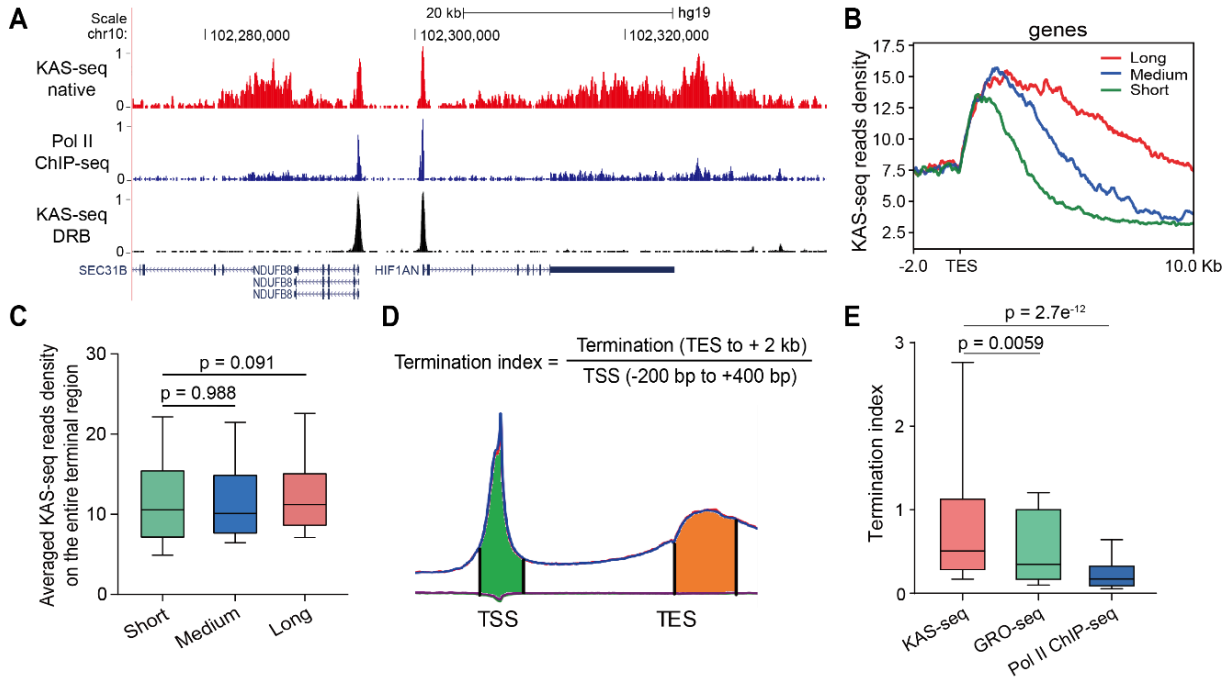

Figure 3.9 KAS-seq shows no length-dependent bias with strong signals around TES regions

(**A**) A snapshot from UCSC Genome Browser showing KAS-seq and Pol II ChIP-seq profiles at the native state, and KAS-seq profile at the DRB-treated state, indicating that KAS-seq signals around TES are derived from Pol II. The autoscale setting is used for all tracks. (**B**) KAS-seq reads densities of three groups of genes with different lengths of termination signals. (**C**) Averaged KAS-seq reads density in the entire terminal regions in the three groups of genes defined in (**B**). n = 660 genes for all three groups. (**D**) Termination index for each gene was calculated as the ratio of KAS-seq reads density on TES to its downstream 2 kb region, versus reads density on the – 200 bp to +400 bp region around TSS. (**E**) The distribution of termination index for all genes in KAS-seq, GRO-seq, and Pol II ChIP-seq (n = 29,160 genes). For (**C**) and (**E**), 10 - 90 percentile of data points are shown, with the centerline indicating the median, and the box limits showing the upper and lower quartiles. P values were calculated using a two-sided unpaired Student's t-test.

3.2.8 KAS-seq detects Pol I- and Pol III-mediated transcription events and non-B form ssDNA structures in the same assay

RNA polymerase I (Pol I) transcribes 5.8S, 18S, and 28S ribosomal RNAs (rRNAs); RNA polymerase III (Pol III) synthesize 5S rRNAs, transfer RNAs (tRNAs) and some small RNAs[1,3]. As expected, apart from detecting Pol II activities, KAS-seq simultaneously detects transcription events mediated by Pol I and Pol III, which do not respond to DRB and triptolide (Figure 3.10A-C). Note that only a portion of tRNAs are actively transcribed (411/606) (Figure 3.10B), which may suggest a transcription level regulation of codon usage. KAS-seq can thus monitor the transcription activity dynamics of all RNA polymerases in one assay.

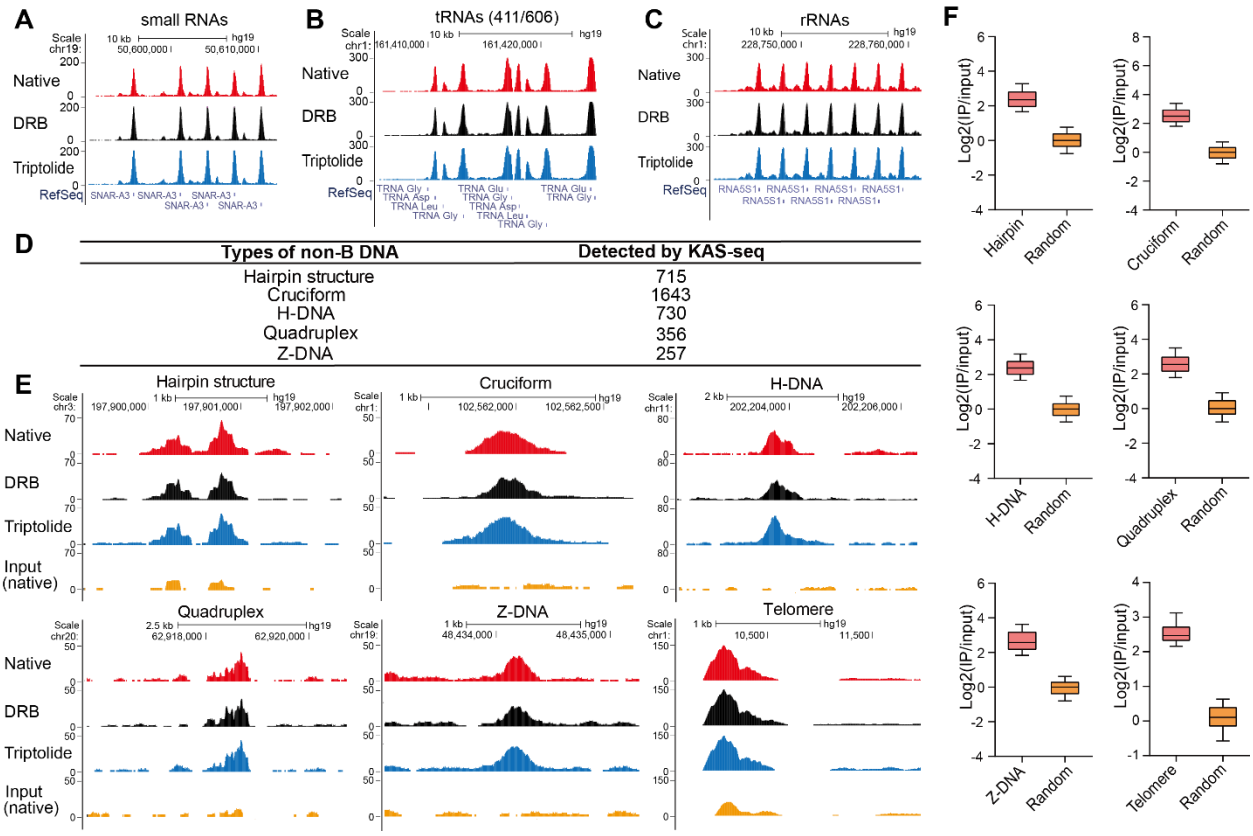

Figure 3.10 KAS-seq detects Pol I and Pol III-mediated transcription, non-B form DNA structures, and telomeric DNA

(**A**) - (**C**) Snapshots of KAS-seq signals at selected small RNA, tRNA, and rRNA loci in HEK293T cells under native, DRB treatment, and triptolide treatment conditions. (**D**) A summary of different

(**Figure 3.10, continued**) types of non-B form DNA structures and the number of KAS-seq peaks (under triptolide condition) detected at each type of predicted non-B form DNA regions. (**E**) Snapshots from the UCSC genome browser showing examples of KAS-seq signals under native, DRB, and triptolide-treatment conditions at different non-B form DNA regions and telomeric DNA regions. (**F**) Enrichment of KAS-seq signals at different non-B form DNA and telomeric DNA regions showed in (**D**). n = 715 regions for hairpin, n = 1,643 regions for cruciform, n = 730 regions for H-DNA, n = 356 regions for quadruplex, n = 256 regions for Z-DNA, n = 29 regions for telomere.

We also noticed many KAS-seq peaks that are not derived from Pol I or Pol III-mediated transcription under triptolide-treatment condition; these peaks could be derived from other DNA forms and telomeric DNAs. We followed a previous report[91] to predict potential genomic locations of different non-B form DNA species, including cruciform, quadruplex, H-DNA, Z-DNA, and hairpin structures. We found many KAS-seq signals under triptolide-treatment condition overlap with these non-B DNA and telomere regions (Figure 3.10D-E) with significant enrichment (Figure 3.10F), suggesting potential applications of KAS-seq to study other ssDNA-involved biological processes.

3.2.9 KAS-seq detects transcribing enhancers

We defined enhancers with KAS-seq peaks as ssDNA-containing enhancers (SSEs). We used the KAS-seq data under DRB-treatment conditions to annotate SSEs, because some enhancers are located at gene bodies that can form ssDNA upon transcription elongation. Around 25% of all annotated enhancers were defined as SSEs in mESCs, with the majority of enhancers showing no KAS-seq signal (Figure 3.11A-B). Note that the cutoff we used for peak-calling filters off some weak KAS-seq signals, which may appear in the defined double-stranded enhancers.

ssDNA-containing enhancers include two sub-types, with one type showing KAS-seq signals spanning over the entire enhancer, and the other type showing KAS-seq signals more localized when comparing with H3K27ac signals (Figure 3.11B). KAS-seq signals at ssDNA-containing enhancers tend to increase upon DRB treatment (Figure 3.11C), supporting the

presence of enhancer transcription pausing and elongation[92]. ssDNA-containing enhancers include 94% of super-enhancers[93], suggesting most of the super-enhancers are actively transcribed (Figure 3.11D). Genes associated with SSEs show higher expression levels (Figure 3.11E), and these enhancers possess much more long-range interactions mediated by both CTCF and Pol II (Figure 3.11F). These results collectively indicate that these transcribing enhancers may possess a stronger capability to activate their target genes.

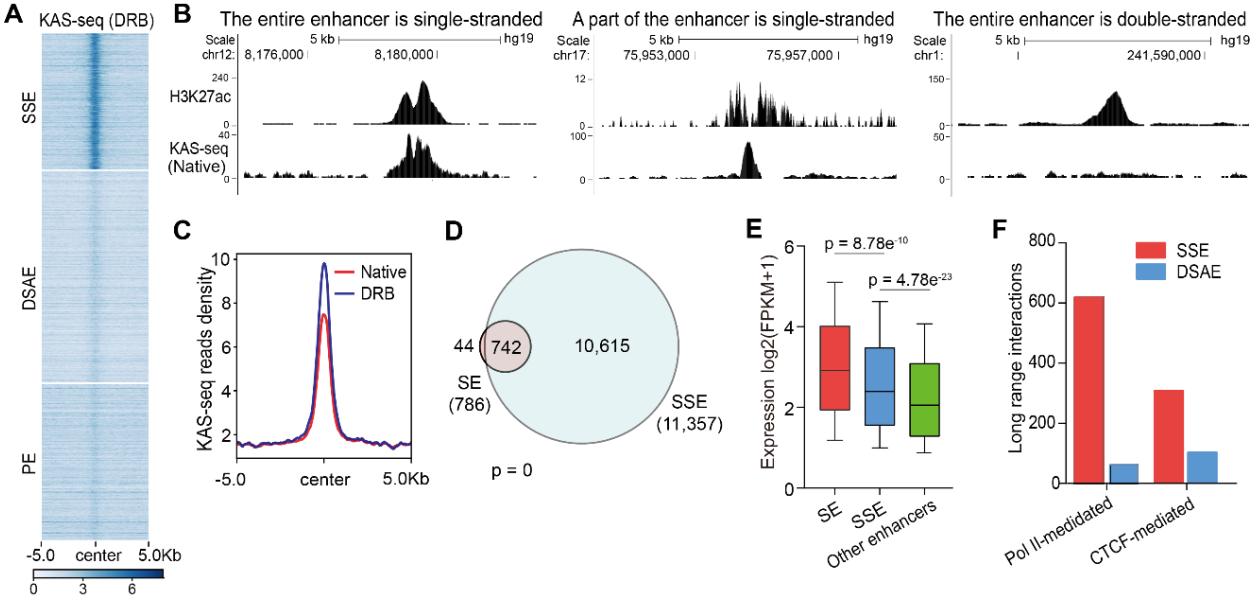

Figure 3.11 Single-stranded enhancers in mESCs

(**A**) Heatmap of KAS-seq reads density at all enhancer regions in mESCs. Active and poised enhancer regions are defined by distal H3K27ac and H3K4me1 signals. Active enhancers are sub-grouped into SSEs and DSAEs. More than 40% of active enhancers (25% of all enhancers) are single-stranded. (**B**) Snapshots of HEK293T KAS-seq under the native condition and H3K27ac signals from UCSC Genome Browser, showing examples that the entire enhancer is single-stranded, a part of the enhancer is single-stranded, or the entire enhancer is not single-stranded, respectively. (**C**) KAS-seq reads densities on ssDNA-containing enhancers in mESCs under native and DRB-treatment conditions. (**D**) The numbers of ssDNA-containing enhancers and super-enhancers in mESCs and their overlap. The p-value was calculated using two-sided Fisher's exact test. (**E**) Boxplot showing the expression levels of genes regulated by denoted enhancers. 10 - 90 percentile of data points are shown, with the centerline showing the median, and the box limits showing the upper and lower quartiles. P values were calculated using two-sided unpaired Student's t-test (n = 617 genes for SEs, n = 3,262 genes for SSEs, n = 3,367 genes for other enhancers). (**F**) ssDNA-containing enhancers possess more long-range interactions mediated by both Pol II and CTCF than those from double-stranded active enhancers. Both Pol II-mediated and

(**Figure 3.11, continued**) CTCF-mediated long-range interactions were defined from public ChIA-PET data in mESCs.

To compare SSEs with enhancers that simply possess high TF-binding signals, we sorted all ATAC-seq-positive enhancers into two groups according to whether they possess KAS-seq signals or not. We found 50% of ATAC-seq-positive enhancers show no (or very weak) KAS-seq signals in mESCs (Figure 3.12A). The averaged intensities of ATAC-seq signals on these two groups are similar (Figure 3.12B), but genes associated with the KAS-seq-positive group show a higher expression level (Figure 3.12C). ssDNA-containing enhancers appear to enrich unique sequence motifs (Figure 3.12D), suggesting their distinct sequence features and potential binding by specific transcription factors (TFs). Sequence motifs enriched in ATAC-seq-positive but KAS-seq-negative enhancers are different from those in SSEs (Figure 3.12E).

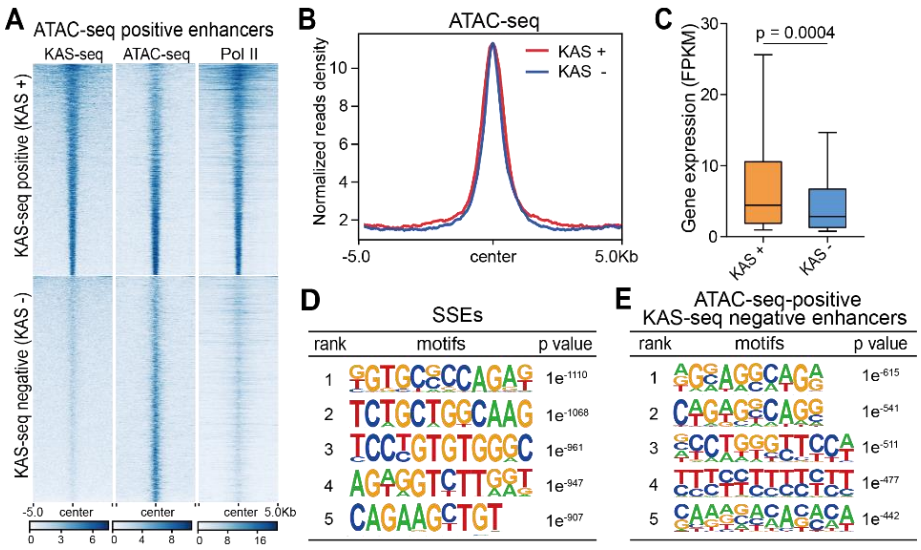

Figure 3.12 ssDNA-containing enhancers are distinct from enhancers with high TF binding

(**A**) All ATAC-seq-positive enhancers were sorted into two groups based on whether they are KAS-seq-positive or not. Heatmaps of KAS-seq, ATAC-seq, and Pol II ChIP-seq signals on these two groups are shown. (**B**) A metagene profile showing ATAC-seq reads density on the two groups of enhancers defined in (**A**). (**C**) Expression levels of genes associated with KAS-seq positive (n = 3,080 genes) and KAS-seq negative (n = 1,544 genes) enhancers defined in (**A**). 10 - 90 percentile of data points are shown, with the centerline showing the median, and the box limits showing the upper and lower quartiles. The p-value was calculated using a two-sided unpaired Student's t-test. (**D**) Sequence motifs enriched in ssDNA-containing enhancers in mESCs. P-values were

(**Figure 3.12, continued**) calculated by using a two-sided binomial test (n = 786 SSEs). (**E**) Sequence motifs enriched in ATAC-seq-positive but KAS-seq-negative enhancers in mESCs. The p values were calculated by a two-sided binomial test (n = 6,082 enhancers).

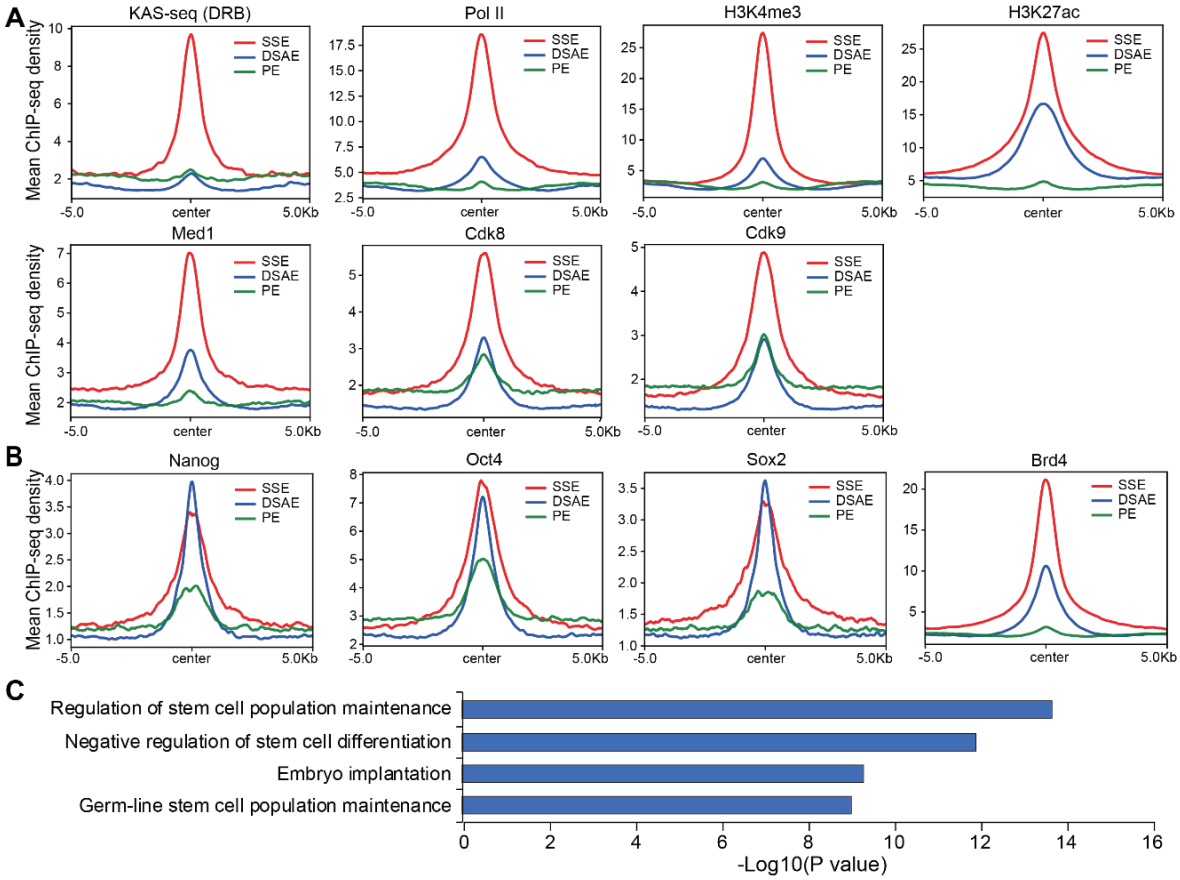

Figure 3.13 ssDNA-containing enhancers are associated with critical functions

(**A**) Metagene profiles of KAS-seq (DRB), Pol II, H3K4me3, H3K27ac, Med1, Cdk8, and Cdk9 ChIP-seq reads densities at denoted enhancers in mESCs. Regions within 10 kb around the enhancer centers are shown. (**B**) Metagene profiles of Nanog, Oct4, Sox2, and Brd4 ChIP-seq read densities at denoted enhancers in mESCs. Regions within 10 kb around the enhancer centers are shown. (**C**) GREAT analysis of genes regulated by ssDNA-containing enhancers in mESCs. P-values were calculated by using a two-sided binomial test (n = 786 SSEs). SSE: ssDNA-containing enhancers; DSAE: double-stranded active enhancers; PE: poised enhancers.

We then examined the occupancy of Pol II, histone modifications, and other transcription regulatory proteins on the ssDNA-containing enhancers. Consistent with them being transcribed, the occupancy of Pol II, H3K4me3, H3K27ac, Med1, Cdk8, and Cdk9 on these enhancers are considerably higher than those double-stranded enhancers (Figure 3.13A). Moreover, while the

binding of Oct4, Nanog, and Sox2 showed no significant difference in SSEs comparing with double-stranded ones, Brd4 is considerably enriched in SSEs (Figure 3.13B). Brd4 was previously reported to regulate the expression of pluripotency factors such as *Pou5f1 (Oct4)* and *Nanog* in mESCs and mouse embryos[94-96], indicating potential roles of these transcribing enhancers on regulating mESC differentiation. Moreover, gene ontology analysis[97] revealed critical biological processes enriched in genes regulated by ssDNA-containing enhancers, including the regulation of stem cell population maintenance, differentiation, and embryo implantation (Figure 3.13C).

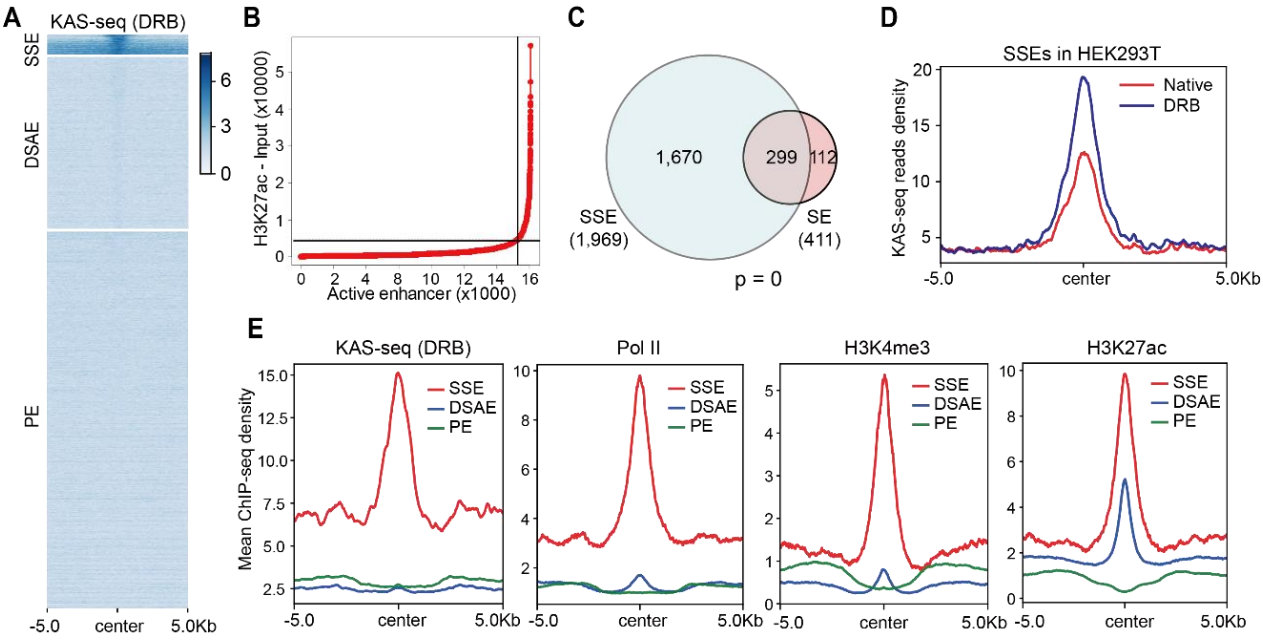

Figure 3.14 ssDNA-containing enhancers in HEK293T cells

(**A**) A group of enhancers are single-stranded in HEK293T cells. Heatmap of KAS-seq reads densities at denoted enhancers in HEK293T cells. Active and poised enhancer regions are defined by distal H3K27ac and H3K4me1 signals. Active enhancers are sub-grouped into SSEs and DSAEs. (**B**) Distribution of H3K27ac ChIP-seq signal across all HEK293T enhancers. Super-enhancers are defined as containing exceptionally high amounts of H3K27ac. (**C**) The number of ssDNA-containing enhancers and super-enhancers in HEK293T cells and the overlap. The p-value was calculated by two-sided Fisher's exact test. (**D**) KAS-seq reads densities on SSEs in HEK293T cells under native and DRB-treatment conditions. (**E**) Metagene profiles of KAS-seq, Pol II, H3K4me3, and H3K27ac ChIP-seq read densities at denoted enhancers in HEK293T cells. Regions within 10 kb around the enhancer centers are shown. SSE: ssDNA-containing enhancers; DSAE: double-stranded active enhancers; PE: poised enhancers.

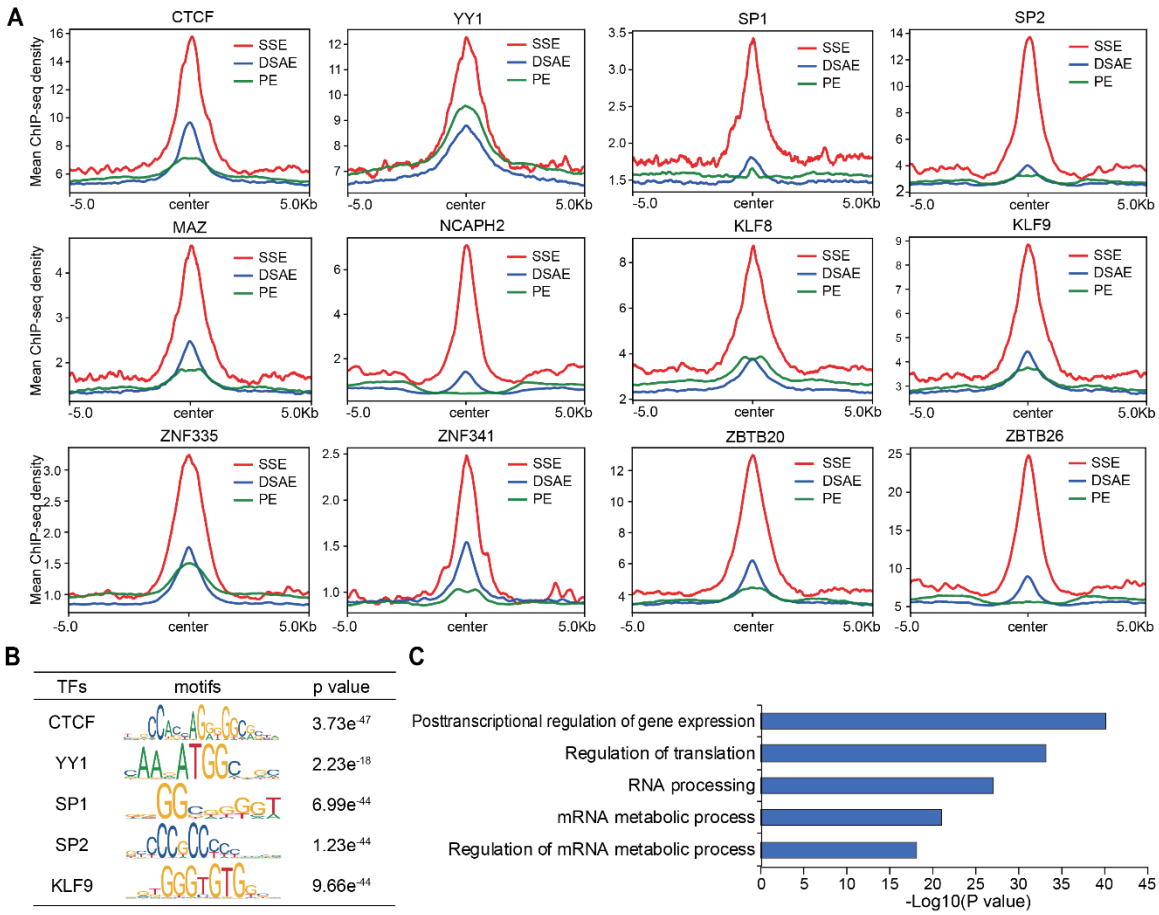

Figure 3.15 Transcription factors that preferentially bind at SSEs in HEK293T cells

(**A**) Metagene profiles of CTCF, YY1, SP1, SP2, MAZ, NCAPH2, KLF8, KLF9, ZNF335, ZNF341, ZBTB20, and ZBTB26 ChIP-seq reads densities at denoted enhancers in HEK293T cells. Regions within 10 kb around the enhancer centers are shown. (**B**) Transcription factor binding motifs enriched at ssDNA-containing enhancers in HEK293T cells (n = 1,969 enhancers) with corresponding p values by using the genome as background. Only TFs with motif information in the TRANSFAC vertebrates library were analyzed. P values were calculated by a two-sided binomial test. (**C**) GREAT analysis of genes regulated by ssDNA-containing enhancers (n = 1,969 enhancers) in HEK293T cells. P values were calculated by a two-sided binomial test. SSE: ssDNA-containing enhancers; DSAE: double-stranded active enhancers; PE: poised enhancers.

In HEK293T cells, though the ratio of SSEs over total enhancers is lower than that in mESCs (Figure 3.14A), SSEs still show high overlap with super-enhancers, can respond to DRB treatment, and enrich Pol II, H3K4me3 and H3K27ac signals (Figure 3.14B-E). Chromatin regulatory factors such as CTCF and YY1, as well as transcription factors such as SP1, SP2, MAZ, NCAPH2, KLF8, and KLF9, showed high occupancy on these ssDNA-containing enhancers

(Figure 3.15A), with their binding motifs enriched at these regions (Figure 3.15B). Several other zinc-finger-domain-containing TFs were also shown enriched on these SSEs (Figure 3.15A). mRNA processing, translation regulation, and several other essential pathways are enriched in genes regulated by these enhancers (Figure 3.15C). Enriched TFs and gene sets in HEK293T cells are different from those in mESCs, suggesting potential regulatory functions by these transcribing enhancers in cell-type-specific manners.

Collectively, KAS-seq can detect SSEs as transcribing enhancers, which appear to possess distinct genomic features and unique TF-binding footprints. Consistent with previous observations[98-102], these enhancers are associated with higher enhancer activity and can be cell-type specific.

3.2.10 ssDNA dynamics upon the inhibition of protein condensates

Considering the fast reaction between N_3-kethoxal and ssDNA as well as the high sensitivity of KAS-seq, we speculated that KAS-seq could detect transcription dynamics in transient events. Protein condensates are highly dynamic structures formed through interactions between mediators, TFs, and other transcription coactivators, and were shown to incorporate Pol II to activate transcription[103-107]. 1,6-hexanediol is widely used to dissociate these condensates *in vivo*, reducing the occupancy of BRD4, MED1, and Pol II on many genes and enhancers[105]. However, how transcription (Pol II) is perturbed dynamically during this process has not been fully elucidated.

To probe protein condensation dynamics taking advantage of the superb sensitivity of KAS-seq, we performed KAS-seq in HEK293T cells treated with 1.5% 1,6-hexanediol for 0 min (no treatment), 5 min, 15 min, 30 min, and 60 min, respectively. PCA analysis showed that KAS-seq profiles at each time point are distinct from the others (Figure 3.16A), indicating dynamic

transcription changes happening from 5 min to 60 min. Consistent with previous results[105], total KAS-seq signals on the gene body gradually decrease from 15 min to 60 min (Figure 3.16B), supporting the role of protein condensate formation on transcription activation. However, after 5 min treatment, we observed a previously unnoticed increase of ssDNA clustered in a ~4 kb window around TSS, which resulted in a slightly increased ssDNA signal on gene body, accompanied by a decreased ssDNA signal at TSS (Figure 3.16B-E). These ssDNA clusters form at both directions of TSS at bi-directional promotors (Figure 3.16C), while they were only observed at TSS downstream regions for uni-directionally transcribed genes (Figure 3.16D). As time went by, these clustered ssDNA signals moved continuously towards TES and gradually diminished, accompanied by increased ssDNA signals at promotor-proximal regions (Figure 3.16C-E).

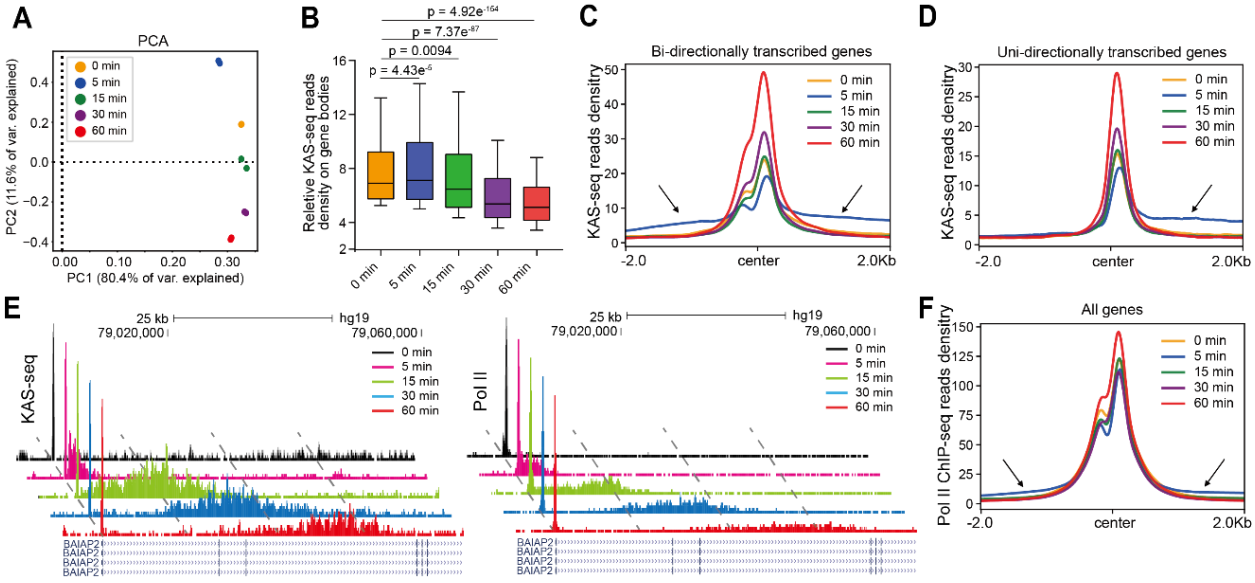

Figure 3.16 Transcription dynamics upon protein condensation inhibition

(**A**) PCA analysis of KAS-seq data at different time points after 1,6-hexanediol treatment (n = 3,122,843 1 kb bins). (**B**) Box plots showing normalized KAS-seq reads densities on gene bodies (from 0.5 kb downstream TSS to TES) of the genes defined as responsive to 1,6-hexanediol treatment. 10-90 percentile of data points are shown, with the centerline showing the median, and the box limits showing the upper and lower quartiles. P values were calculated by using a two-sided unpaired Student's t-test. (**C**) - (**D**) KAS-seq read densities around TSS on uni-directional (**C**) and bi-directional (**D**) transcribed genes after HEK293T cells were treated with 1,6-hexanediol for denoted time intervals. Arrows indicate the upstream and downstream "released" KAS-seq

(**Figure 3.16, continued**) signals at the 5 min time point. (**E**) Snapshots of KAS-seq and Pol II ChIP-seq signals on the *BAIAP2* gene after cells were treated with 1,6-hexanediol for denoted time intervals. Snapshots at different time points for each data set are staggered to show differences clearly. The autoscale setting was used for all tracks. The genomic coordinates and the Refseq tracks are aligned to the 60 min time point. (**F**) Pol II ChIP-seq read densities around TSS after cells were treated by 1,6-hexanediol for denoted time intervals.

We next performed Pol II ChIP-seq at corresponding time points to validate the observations revealed by KAS-seq. The change of Pol II binding generally followed the changes observed by KAS-seq, with a portion of clustered Pol II released from TSS and subsequently moved towards TES at a similar speed as ssDNA clusters (Figure 3.16E-F). Notably, the moving speed of these released Pol II is notably slower (~40 kb per hour, Figure 3.16E) than the rate of Pol II elongation under native condition (>200 kb per hour), perhaps due to a lack of certain regulatory components under 1,6-hexanediol treatment.

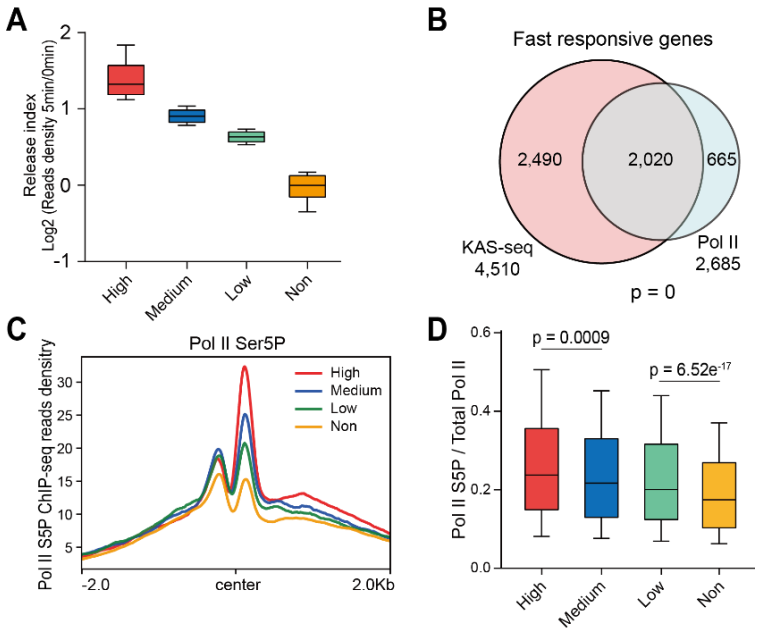

Figure 3.17 Definition and features of fast responsive genes

(**A**) Boxplot showing the calculated release index of high (n = 1,730 genes), medium (n =1,730 genes), low (n = 1,730 genes) and non-responsive (n = 1,188 genes) genes. (**B**) Numbers of fast responsive genes defined by KAS-seq and Pol II ChIP-seq, and the overlap. The p-value was calculated by two-sided Fisher's exact test. (**C**) Pol II CTD S5P densities on four groups of genes that respond to 1,6-hexanediol to different extents. (**D**) Boxplot showing the ratio of Pol II S5P over total Pol II on TSS in high-, medium-, low-, and non-responsive genes (n = 1,730 for high

(**Figure 3.17, continued**) responsive genes, n = 1,730 for medium responsive genes, n = 1,730 for low responsive genes, and n = 1,188 for non-responsive genes). For (**A**) and (**D**), 10 - 90 percentile of data points are shown, with the centerline showing the median, and the box limits showing the upper and lower quartiles. P values were calculated using a two-sided unpaired Student's t-test.

We defined and ranked genes with the aforementioned 'release' feature by calculating a 'release index', as the ratio of KAS-seq reads density at 0.5–2.5 kb downstream TSS at 5 min versus that under the native state. We defined 4,510 genes as 'fast responsive genes', with significant ssDNA cluster formation in this region (Figure 3.17A). We then performed a similar analysis by using Pol II ChIP-seq. 75% (2,020/2,685) of fast responsive genes defined by Pol II ChIP-seq overlap with those detected by KAS-seq, but KAS-seq identified considerably more genes (Figure 3.17B). This number difference and the metagene profiles (Figure 3.16C-F) showed that KAS-seq exhibits higher sensitivity than Pol II ChIP-seq on revealing transcription dynamics during the early stage of inhibition. The extent of the Pol II release correlates with the Pol II CTD serine 5 phosphorylation (S5P) level at TSS at the native state (Figure 3.17C-D), supporting Pol II phosphorylation as a mechanism to tune transcription regulated through condensate formation[6, 108].

3.3 Discussion and conclusion

KAS-seq simultaneously detects the dynamics of transcriptionally engaged Pol II, transcribing enhancers, potential non-B form ssDNA structures, and the activities of Pol I and Pol III with high sensitivity and low input materials. ssDNA hotspots may also form during DNA damage[109], DNA replication, and meiotic/mitotic double-strand break (DSB)[110,111]. Although we focused on transcription in this work, KAS-seq can be useful for understanding all of these processes. The robust and tissue-friendly nature, coupled with low input material requirement, makes KAS-seq a method that can be broadly applied to profile transcription dynamics and other ssDNA-involving processes in rare samples, such as primary cells and patient samples.

ssDNA-containing enhancers show unique sequence features, correlates with more active transcription of downstream genes, and enrich certain functions. Although we observed two different types of ssDNA-containing enhancers, our current analysis did not distinguish these two types. Current KAS-seq has a similar resolution as ChIP-seq, which is commonly used to study and define enhancers. Other techniques with higher resolution, or a high-resolution version of KAS-seq may be applied to differentiate the two types of enhancers and explore their unique properties.

KAS-seq revealed a previously unnoticed phosphorylation-dependent Pol II releasing from promotors to elongation at an early stage of protein condensate inhibition, suggesting that protein condensates at promotors may store pre-phosphorylated Pol II[6] and facilitate fast initiation-elongation transition. Released Pol II move continuously at a relatively slow speed upon condensate inhibition, while the new Pol II recruited to the promotor does not release, potentially due to the dissociation of a series of critical TFs, coactivators, and kinases required for elongation. A similar process may exist during cell response to other stresses. The nature of the Pol II complexes that are released from the promoter and elongated at a slow rate is unclear at this moment, nor is its potential physiological relevance or functional roles. Future characterization of this process and the complexes involved may reveal new insights into transcription regulation.

3.4 Methods

3.4.1 Labeling DNA oligos with N_3-kethoaxl *in vitro*

1 μL 100 μM synthetic DNA oligo (IDT) was mixed with 5 μL nuclease-free water, 2 μL 5× reaction buffer (0.5 M sodium cacodylate, 50 mM $MgCl_2$, pH 7.0) and 2 μL 500 mM N_3-kethoxal (DMSO solution). The mixture was incubated at 37 ℃ for 10 min. The reaction product was purified by Micro Bio-Spin™ P-6 Gel Columns (Biorad, 7326222) and then used for MALDI-

TOF analysis directly. 2'4'6'-trihydroxyacetophenone (10 mg/mL in 50% CH_3CN/H_2O) and ammonium citrate (50 mg/mL in H_2O) was mixed in 1:8 (v/v) ratio as the matrix for MALDI-TOF. 1 μL purified reaction product was mixed with 1 μL matrix on the MALDI sample plate and analyzed by Bruker Ultraflextreme MALDI-TOF-TOF.

3.4.2 Dot blot

1 μL DNA was loaded onto the Amersham Hybond-N+ membrane (GE Healthcare, RPN119B). Membranes were air-dried and were crosslinked by UV stratalinker 2400 at 150 mJ/cm^2 twice. The membranes were then blocked overnight in 5% fatty-acid free BSA in PBST (0.1% Tween-20). The second day, the membrane was washed and incubated in streptavidin-HRP (Thermo, S-911) in PBST supplemented with 3% fatty-acid free BSA. The membrane was washed in PBST for 5 times before developed by SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo, 34577).

3.4.3 Comparing the labeling reactivity of N_3-kethoxal on deoxyguanosine and L-arginine

2 mM deoxyguanosine or 2 mM L-arginine were mixed with 4 mM N_3-kethoxal in neutral reaction buffer (0.1 M sodium cacodylate, 10 mM $MgCl_2$, pH 7.0), respectively. The reactions were performed at 37 ℃ and were monitored by thin-layer chromatography (TLC). The reaction between N_3-kethoxal and deoxyguanosine was developed in a mixture of dichloromethane and methanol (2:1, v/v), and was visualized by 254 nm UV light. The reaction between N_3-kethoxal and L-arginine was developed in 1:1 (v/v) ratio of acetonitrile and ammonium hydroxide and was visualized by ninhydrin staining.

3.4.4 Cell culture

HEK293T cells were purchased from ATCC (CRL11268) and were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1% penicillin and

streptomycin (Gibco) and grown at 37 ℃ with 5% CO_2. Murine embryonic stem (ES) cells were purchased from ATCC (CRL-1821) and were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1 mM L-glutamine (Gibco), 0.1 mM β-mercaptoethanol (Gibco), 1% (v/v) nonessential amino acid stock (100×, Gibco), 1% penicillin/streptomycin stock (100×, Gibco), and 1,000 U/mL LIF (Millipore).

Cell lines used in this study were examined for mycoplasma contamination test using LookOut Mycoplasma PCR Kit (Sigma, MP0035).

3.4.5 KAS-seq

Cells were incubated in a completed culture medium containing 5 mM N_3-kethoxal and for 5-10 min at 37 ℃, 5% CO_2. For transcription inhibition experiments, cells were treated for 2 h under 100 μM DRB (Sigma, D1916) or 1 μM triptolide (Sigma, T3652) before incubated in the N_3-kethoxal-containing medium. For 1,6-hexanediol treatment experiments, cells were treated with 1.5% (v/v) 1,6-hexanediol (Sigma, 240117) in the culture medium for 0 min, 5 min, 15 min, 30 min, and 60 min, respectively, before subjected to N_3-kethoxal labeling. Cells were harvested and genomic DNA (gDNA) was isolated from cells by PureLink genomic DNA mini kit (Thermo, K182002). 1 μg genomic DNA was then suspended in 95 μL DNA elution buffer supplemented with 5 μL 20 mM DBCO-PEG_4-biotin (DMSO solution, Sigma, 760749), 25 mM K_3BO_3, and incubated at 37 ℃ for 1.5 h with a gentle shake. 5 μL RNase A (Thermo, 12091039) was added into the reaction mixture, followed by incubation at 37 ℃ for 5 min. Biotinylated gDNA was then recovered by DNA Clean & Concentrator-5 kit (Zymo, D4013). gDNA was suspended into 100 μL water and was fragmented to 150-350 bp size by using Bioruptor Pico at the 30s-on/30s-off setting for 30 cycles. 5% of the fragmented DNA was saved as input, and the rest 95% was used to enrich biotin-tagged DNA by incubating with 10 μL pre-washed Dynabeads MyOne

Streptavidin C1 (Thermo, 65001) at room temperature for 15 min. The beads were washed, and DNA was eluted by heating the beads in 15 μL H_2O at 95 ℃ for 10 min. Eluted DNA and its corresponding input were used for library construction by using the Accel-NGS Methyl-seq DNA library kit (Swift, 30024). The libraries were sequenced on Illumina Nextseq500 platform with single-end 80 bp mode, aiming to get 30 million reads per library.

3.4.6 KAS-seq using mice liver

Male B6 mice were purchase from the Jackson Laboratory (catalog No: C57BL/6J). All mice were used at 6-12 weeks of age. Mice were housed under pathogen-free conditions per the NIH Guide for the Care and Use of Laboratory Animals. All animal care and experiments were approved by the University of Chicago Institutional Animal Care and Use Committee (IACUC) and are compliant with all relevant ethical regulations regarding animal research.

For KAS-seq performed by using mice liver, the tissue was first homogenized to cell suspension by using a Dounce homogenizer or a pestle grinder. The suspended cells were then washed and subjected to typical KAS-seq procedures.

3.4.7 Low-input KAS-seq

KAS-seq protocol was applied to 1,000, 5,000, and 10,000 HEK293T cells with the following changes. gDNA was isolated from denoted numbers of N_3-kethoxal-labeled cells by using the Quick gDNA mini plus kit (Zymo, D4068). After biotinylation, gDNA was fragmented by Tn5 transposase (Illumina, 10527865, 1.5 μL for 1,000 cells, 2 μL for 5,000 cells, 5 μL for 10,000 cells) in a 50 μL volume at 37 ℃ for 30 min, followed by a clean-up by DNA Clean & Concentrator-5 kit (Zymo, D4013). After immunoprecipitation using 5 μL pre-washed Dynabeads MyOne Streptavidin C1, DNA-conjugated beads and corresponding inputs were directly used for library PCR by using i5 and i7 index primers (Illumina, 20027213) and NEBNext Ultra II Q5

Master Mix (NEB, M0544S). The PCR reactions were heated at 5 min at 72 ℃ followed by 10 min at 95 ℃ and were then amplified by 15 cycles (10 sec at 98 ℃, 30 sec at 60 ℃, 1 min at 72 ℃). The libraries were then cleaned-up by using the MinElute PCR purification kit (Qiagen, 28804).

3.4.8 ChIP-seq

Cells were crosslinked in 1% formaldehyde diluted in culture medium for 10 min and then quenched with 125 mM glycine for 5 min. 5 million cells were used for all ChIP reactions. Crosslinked cells were resuspended in ice-cold lysis buffer (50 mM HEPES, pH 7.9, 5 mM $MgCl_2$, 0.2% Triton X-100, 20% glycerol, 300 mM NaCl) and incubated on ice for 10 min before centrifuged at 500 g for 5 min. The pellets were resuspended in 0.1% SDS lysis buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 150 mM NaCl) and incubate on ice for 10 min and then sheared by using Bioruptor Pico at the 30s-on/30s-off setting for 20 cycles. 5% of sheared chromatin was saved as input, and the rest was subjected to pre-clear and then mixed with 30 μL protein A/G bead coated with 5-10 μg antibodies. Immunoprecipitation was performed overnight. The beads were then washed twice with 0.1% SDS lysis buffer, high salt buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate; 0.1% SDS, 350 mM NaCl), LiCl wash buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.5% NP-40, 0.5% sodium deoxycholate, 250 mM LiCl), and once with TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.2% Triton X-100). Enriched chromatin was eluted from the beads by incubating with elution buffer (50 mM $NaHCO_3$, 10 mM EDTA, 1% SDS) at room temperature for 1 h. The eluent and the input were subjected to reverse crosslink and proteinase K digestion before DNA was purified from the mixture by using DNA Clean & Concentrator-5 kit

(Zymo, D4013). Recovered DNA was used for library construction by Kapa HyperPlus kit (Kapa, KK8515).

3.4.9 RNA-seq

Total RNA was extracted from cells by using Trizol reagent (Thermo, 15596026). PolyA selection was then performed by using the Dynabeads mRNA purification kit (Thermo, 61006). 20 ng polyA RNA was used for library construction by using the SMARTer Stranded RNA-seq kit (Takara, 634839).

3.4.10 KAS-seq data processing and peak calling

Low-quality and adapter-containing reads were trimmed from KAS-seq raw data using trim-galore package[112] under single-end mode. Reads shorter than 50 bp were removed. Trimmed reads were aligned to the reference genome (hg19 for HEK293T cells or mm10 for mESC) using bowtie2 (v2.3.3.1)[113] under default parameters. Mapped sam files were subsequently converted and sorted to bam files using samtools sort (v1.9)[114]. Duplicates were removed using samtools rmdup (v1.9) to get the unique mapped reads. Unique mapped reads were extended to 150 bp to match the average length of insert DNA fragments of the KAS-seq libraries. Bam files were converted to bed files and bedGraph files using bedtools. BedGraph files were then converted to bigWig files using bedgraphtobigwig from UCSC pre-compiled utilities. BedGraph files were used for visualization at the UCSC genome browser, and bigWig files were used to calculate tag density under 50-bp resolution.

We used MACS[115] to call all reported KAS-seq peaks in this manuscript (macs2 callpeak -t KAS-seq_IP.bed -c KAS-seq_Input.bed -n KAS-seq_peaks.bed --broad -g hs --broad-cutoff 0.01 -q 0.01). Because most KAS-seq peaks on gene bodeis are very broad, MACS2 was run using

broad peaks-call mode under default parameters, except for '--broad-cutoff = 0.1' and "–qvalue = 0.01".

3.4.11 Genome-wide distribution of KAS-seq peaks

A group of regions that have the same number and length of KAS-seq peaks were randomly generated from the hg19 genome by using bedtools shuffle. These random regions and real KAS-seq peaks were subsequently overlapped with different genomic features retrieved from the hg19 Refseq annotation in the order of promoters (TSS +/-2 kb), exons, introns, and terminal regions (TES to +2 kb of TES). If a peak overlaps with promoters, it is regarded as a promoter peak and removed from the peak list. The remaining peaks were then subjected to similar overlap analysis with exons, introns, and terminal regions. Peaks do not have overlap with these genomic features are regarded as intergenic peaks.

3.4.12 RNA-seq data processing

Low-quality and adapter-containing reads were trimmed using the trim-galore package under the paired-end mode. Any reads shorter than 50 bp were removed. The remaining trimmed sequences were mapped to the reference genome (hg19) with hisat2[116] under default settings. The expression level of each gene was quantified with normalized FPKM with FPKM_count.pl in the RSeQC[117] software. Genes with FPKM higher than 0.5 were defined as expressed genes. To arbitrarily define genes with high, medium, and low expression levels, expressed genes were ranked and sorted into 3 groups based on their FPKM values. The top 2,000 genes were defined as highly-expressed; 2,000 genes in the middle were defined as mediumly expressed; the bottom 2,000 genes were defined as lowly-expressed. 2,000 genes with FPKM lower than 0.5 were randomly selected and defined as silent genes.

3.4.13 ChIP-seq data processing and peak calling

ChIP-seq data processing and peak calling generally follow the procedure used for KAS-seq data processing and peak calling.

3.4.14 Correlation analysis

Correlation calculations between KAS-seq, histone modification ChIP-seq, and ATAC-seq were performed using deeptools[118] package. First, multiBigwigSummary was used to calculate the averaged read coverage within equally sized 10 kb bins of the entire genome. Regions in the human genome blacklist were excluded from the calculation. PlotCorrelation was subsequently used to calculate pairwise Pearson correlation coefficients by using the output of multiBigwigSummary. Outliers were defined using the median absolute deviation (MAD) method by applying a threshold of 200, and were removed for correlation analysis. Heatmaps were generated with pairwise Pearson correlation coefficients depicted by varying color intensities and were clustered using hierarchical clustering. Correlation calculations between KAS-seq, Pol II ChIP-seq, GRO-seq, and 4SU-seq were performed using a similar approach but based on gene-coding regions.

3.4.15 Definition of four transcription states

To define transcription states, we calculated the KAS-seq tag density on the promoters (from -200 bp to +400 bp from TSS) and gene bodies (from +400 bp downstream TSS to TES) of protein-coding genes. Gene promoters with KAS-seq tag densities more than 20× as the density on average were considered to be paused. Similarly, gene bodies with KAS-seq tag densities more than 10× as the density on average were defined as actively transcribed.

3.4.16 Defining genes with long, medium and short terminal regions

Genes that do not overlap with other genes within a 10 kb range downstream TES were used for analysis. The 10 kb region downstream TES were divided into 20 bins of the same length. We calculated the averaged KAS-seq reads density on each bin, and bins with averaged KAS-seq

reads density equal to or greater than 5 were defined as positive bins. We ranked all genes according to their number of positive bins, from highest to lowest. Genes ranked among top 1/3 were defined as long terminal genes; genes ranked among bottom 1/3 were defined as short terminal genes; the rest genes were defined as with medium-length termination signals.

3.4.17 Calculation of the termination index

We calculated the termination index for KAS-seq, Pol II ChIP-seq, and GRO-seq as the log2 ratio of reads density on terminal regions (from TES to +2 kb from TES) over that around TSS (from -200 bp to +400 bp from TSS). Only genes with KAS-seq tag density on promoters more than 50× as the density on average were included in the calculation.

3.4.18 Identification of predicted non-B form DNA with KAS-seq peaks

The positions of all the non-B form DNA motifs were downloaded from non-B DB v2.0[91]. To obliterate the effect of Pol II-induced KAS-seq signals, we used KAS-seq peaks identified in triplotide-treated HEK293T cells. KAS-seq peaks related to tRNA, rRNA, small NF90-associated RNAs, and U6 spliceosomal RNA, which were generated by Pol I and Pol III, were excluded from the analysis.

Enrichment of KAS-seq signals on non-B form DNAs was determined by calculating log2(IP reads density/input reads density) on each KAS-seq positive non-B form DNA region. The distributions of enrichment for each non-B form DNA type were shown in box plots, comparing with the same number of regions randomly found in the genome. To calculate the enrichment of the KAS-seq signal on telomeres, we used the KAS-seq signals the 15 kb rightmost and 15 kb leftmost regions of all chromosomes on the hg38 reference genome.

3.4.19 Defining single-stranded-DNA-containing enhancers and super-enhancers

We used H3K27ac and H3K4me1 peaks distal from genes promoters (based on mm10 and hg19 on NCBI Refseq) to define active and poised enhancers. H3K27ac enriched regions were defined as active enhancers; regions with enriched H3K4me1 but not H3K27ac were defined as poised enhancers. We found that very few poised enhancers are single-stranded, so only active enhancers with KAS-seq peaks were defined as single-strand-DNA-containing enhancers. In addition, some active enhancers are located on the gene body. Thus KAS-seq signals on these enhancers may derive from Pol II elongation. Therefore, KAS-seq under DRB treatment was used to define single-stranded-DNA-containing enhancers. Enhancers with KAS-seq peaks observed in both DRB replicates were defined as single-stranded-DNA-containing enhancers.

Super enhancers were defined using the ROSE package as previously described[93].

3.4.20 Motif analysis

Sequence motifs enriched by ssDNA-containing enhancers and ATAC-seq-positive but KAS-seq-negative enhancers were analyzed by using HOMER[119].

The sequences of ssDNA-containing enhancers were extracted and used as input for TRAP[120] using TRANSFAC vertebrates as the comparison library, promoter sequences as the background, and Benjamini-Hochberg as the correction. P-values were displayed in figures corresponding to the 'corrected p' in the output.

3.4.21 Assigning enhancers to their regulated genes

We assigned enhancers to their regulated genes based on the NCBI RefSeq gene annotations. We calculated the distance from the center of enhancers to the TSS of genes. Within 50 kb, the gene closest to the enhancer is assigned as the gene regulated by this enhancer.

Pol II and CTCF ChIA-PET data were used to define the long-range interactions.

3.4.22 Calculate the release index to define genes responsive to 1,6-hexanediol

We calculated Pol II or ssDNA release index as the log2 ratio of Pol II or KAS-seq reads density at a region from +0.5 kb to +2.5 kb downstream TSS at 5 min versus that with no treatment (0 min). As some genes have very short gene bodies, only genes with gene bodies longer than 5 kb were included in the calculation. Genes with Pol II or ssDNA release index higher than 0.5 were defined as genes affected by protein condensation inhibition, which were sorted into high-, medium- and low-affected genes groups, with the number of genes in each group the same. Genes with Pol II or ssDNA release index lower than 0.2 were defined as non-affected genes.

3.4.23 Definition of bidirectional and uni-directional promoters

We defined bidirectional and unidirectional promoters by reanalyzing published NET-seq data in HEK293T cells. Promoter-proximal regions were carefully defined to ensure minimal signal contamination from genes nearby. Genes shorter than 5 kb were excluded from the analysis. Genes with TSS located within 2.5 kb upstream of the TSS of another gene, or 2.5 kb downstream of the polyA cleavage site of another gene, were excluded from the analysis. In cases of conflicting isoform annotations, the most upstream annotated TSS and the most downstream annotated polyA cleavage sites were used. Within a 4 kb region around TSS, promoters with more than 40 NET-seq signals covering both sense and antisense directions were defined as bidirectional. In contrast, promotors with 40 NET-seq signals covering only sense but not antisense direction were defined as uni-directional.

3.4.24 Data availability

All sequencing data are available at NCBI Gene Expression Omnibus with the accession number: GSE139420.

Chapter 4

N_3-kethoxal-mediated profiling of the RNA-RNA interactions

4.1 Introduction: transcriptome-wide detection of RNA-RNA interactions

Ribonucleoprotein (RNP) are dynamic complexes composed of highly-structured RNAs and RNA-binding proteins (RBPs). Both RNA and RBP play critical roles for RNP functions[36]. Extensive efforts have been made to understand the association between RBPs and their RNA targets[48]. However, how RNA interacts with each other globally has not been fully understood. To address this question, transcriptome-wide approaches have been developed to capture double-stranded RNA duplexes or transcripts in close proximity. Double-stranded RNA mapping usually takes advantage of psoralen crosslinking, which captures direct RNA-RNA base pairing[121-123]. RNA proximity profiling typically crosslinks RNA with proteins, followed by RNA fragmentation and proximity ligation, to acquire contact frequency as a read-out for physical distances[124-126]. Both types of approaches provide new insights in RNA-RNA interactions globally, but challenges still remain. Psoralen crosslinking capture one steady-state of RNA-RNA interactions, but miss the distance information between transcripts that are physically close but not forming base pairs. Formaldehyde-mediated approaches usually fail to enrich crosslinked and/or chimeric products, resulting in limited sensitivity for transcripts with modest expression levels. This problem was partly solved by a recent work that incorporates a pCp-biotin moiety during ligation, which was subsequently used for enriching ligated products[126]. However, this method may introduce bias caused by performing RNA fragmentation and proximity ligation inside crosslinked cells, where the local protein and RNA concentration is spatially heterogeneous. Proteins bound to RNA with high occupancy can mask RNA fragmentation enzymes, and RNA proximity ligation

preferentially happens at condensed subcellular compartments (such as chromatin). The accuracy of in situ proximity ligation has also been questioned[127].

In Chapter 2, I have shown the usage of N_3-kethoxal as an effective way to modify RNA with azido (-N_3) groups, which can be subsequently functionalized by click chemistry. Previous work from our lab demonstrated the utility of multifunctional poly(amidoamine) (PAMAM) dendrimer for capturing proximal chromatin contact. Here I present a kethoxal-assisted RNA-RNA interaction mapping (KARR-seq), which captures physically proximal RNA through PAMAM dendrimers and N_3-kethoxal-mediated click chemistry, instead of formaldehyde-mediated RNA-protein crosslinking. KARR-seq detects intra- and inter-molecular RNA-RNA interactions with high sensitivity and resolution. The frequency of KARR-eq chimeras accurately reveals the physical distances in both rRNA and mRNA, as validated by known RNA cryo-EM structures. Based on KARR-seq data, we propose that apart from stable RNA helixes, physically close RNAs can form transient domain structures. Compared with RNA-RNA interactions formed in vitro, most interactions are diminished in vivo by both translation machinery and potentially RBP binding.

4.2 Results

4.2.1 The development of KARR-seq

The diameters of the dendrimers can be precisely tuned by controlling the branching cycles during synthesis, which provides us the flexibility to choose the correct size to fit the crosslinker into RBP complexes. We purchased commercially available PAMAM dendrimers with different sizes, namely G1 (22 Å diameter), G3 (36 Å diameter), G5 (54 Å diameter), and G7 (81 Å diameter). We decorated them with multiple DBCO (dibenzocyclooctane) and biotin moieties (Figure 4.1A). DBCO crosslinks proximal N_3-kethoxal modified RNAs via click chemistry, and

biotin enables the enrichment of crosslinked products. The decorated dendrimers were characterized by measuring the characteristic UV absorption of DBCO at 295 nm (Figure 4.1B), and were termed G1-DBCO-biotin, G3-DBCO-biotin, G5-DBCO-biotin, and G7-DBCO-biotin, respectively.

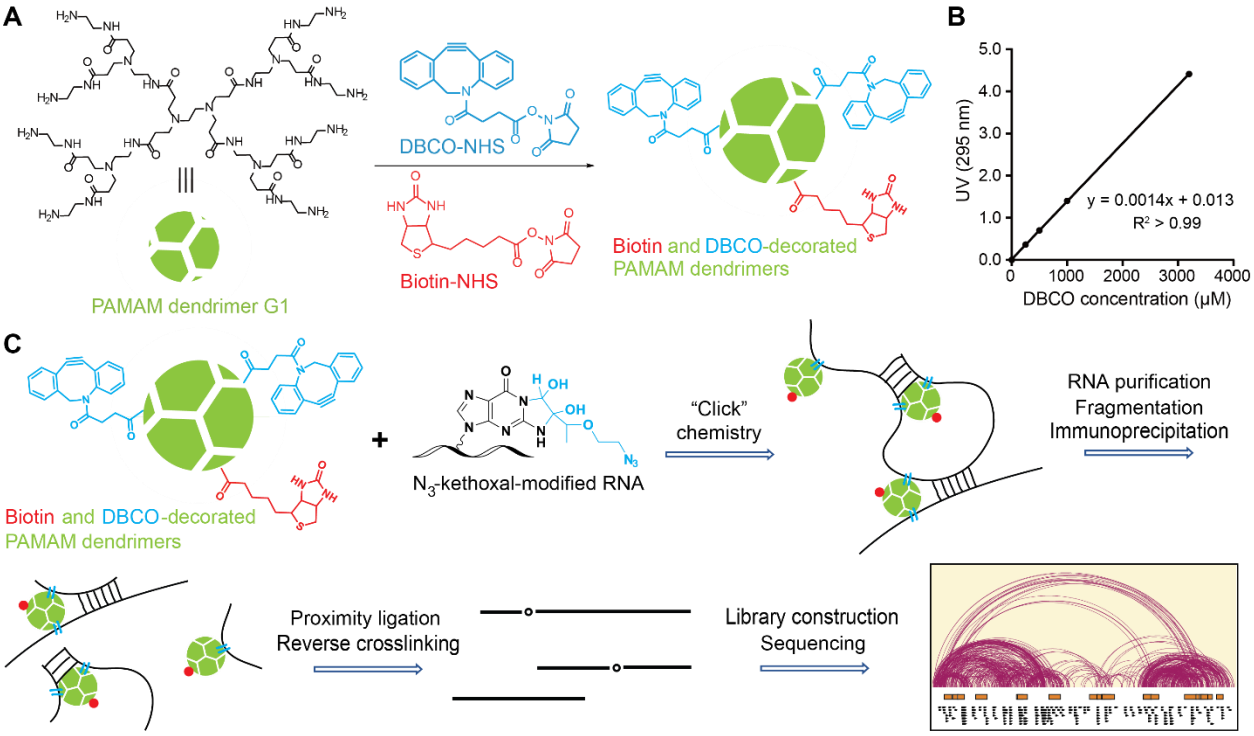

Figure 4.1 The design of KARR-seq

(**A**) The scheme illustrating the synthesis of biotin and DBCO-decorated PAMAM dendrimers, using G1 as an example. A detailed synthetic procedure was described in the Methods section. (**B**) The calibration curved used for the quantification of decorated dendrimers. DBCO-NHS monomers at different concentrations were used as standards, whose UV adsorption at 295 nm was plotted versus concentrations. R denotes the Pearson correlation coefficient. (**C**) The scheme of KARR-seq. The dendrimer was shown in green. Dendrimer crosslinked RNA was fragmented and captured by streptavidin beads (omitted). Enriched RNAs were subjected to proximity ligation and library preparation.

To capture interacting RNAs, we first labeled two million cells with N_3-kethoxal and then diffused dendrimers into permeabilized cells. The cells were subjected to click chemistry reaction for crosslinking, after which proteins were digestion and RNA was purified. Crosslinked RNA was then fragmented and immobilized on streptavidin-coated beads. After end repair, on-beads

proximity ligation was performed under ultra-diluted conditions, which tends to decrease the background caused by random molecular collisions. The RNA was then eluted from the beads for library construction and pair-ended sequencing, with chimeric sequencing reads stands for interacting transcripts (Figure 4.1C). Both gel electrophoresis (Figure 4.2A) and dot blot assays (Figure 4.2B) show that the combination of N_3-kethoxal and decorated dendrimers successfully crosslinks RNA. Control experiments performed in the absence of N_3-kethoxal or decorated dendrimer resulted in very weak or invisible signals in dot blot (Figure 4.2B), suggesting the low background of click chemistry-mediated crosslinking.

4.2.2 KARR-seq detects known RNA-RNA interactions with physical distance information

We performed KARR-seq by using G3-DBCO-biotin in mouse embryonic stem cells (mESCs). To estimate potential background noise, we also performed KARR-seq in the absence of N_3-kethoxal, dendrimer, ligation, or biotin enrichment, respectively. We found that chimeric reads in KARR-seq reconstitute known RNA-RNA interactions (18S rRNA, for example, Figure 4.2C). In the meantime, few chimeric reads were detected in the negative control libraries, with very weak interactions observed in the heatmap (Figure 4.2C).

We next performed KARR-seq in mESCs by using decorated dendrimers with different diameters and use chimeric reads for analysis. We found a high correlation between individual replicates for each dendrimer, suggesting the robustness of the KARR-seq protocol (Figure 4.2D). Interestingly, when we analyzed correlations between different dendrimers, we found that results from G1 display a high correlation with those from G3 but lower correlations with those derived from G5 and G7 (Figure 4.2E), suggesting a potential importance of dendrimer size in interaction capture. While the ligation rates among all libraries are similar (Figure 4.3A), small dendrimers (G1, G3) captures many more transcripts with valid interactions than larger ones (G5, G7) (Figure

4.3B). Therefore, we envisioned that smaller dendrimers are more accessible to RNA, especially in the context of condensed RNPs. We used G1-DBCO-biotin for all the following KARR-seq experiments and analysis.

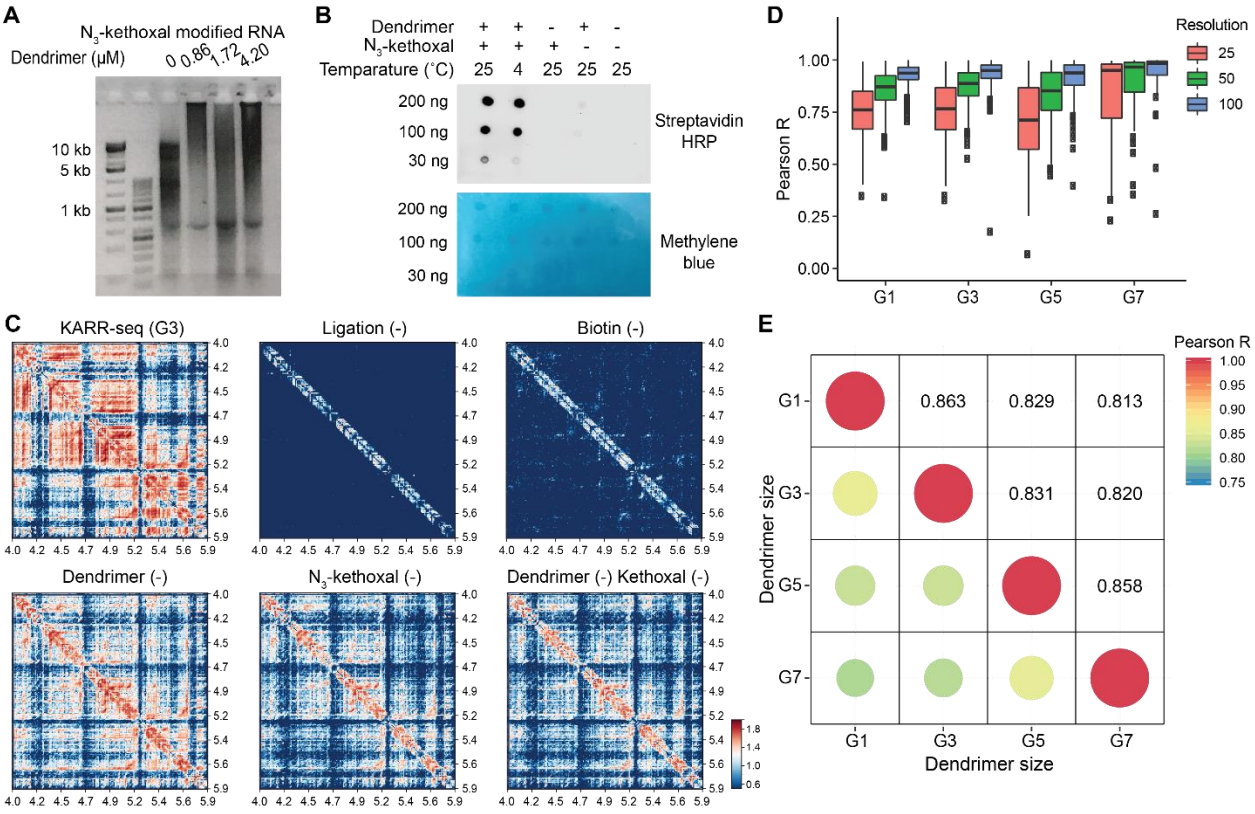

Figure 4.2 Quality control of KARR-seq data

(**A**) Agarose gel electrophoresis of crosslinked and uncrosslinked RNAs. (**B**) Dot blot analysis of RNAs crosslinked at different temperatures and negative controls that dendrimer and/or N3-kethoxal was omitted for crosslinking. (**C**) The interaction map of mouse 18S rRNA produced by G3-DBCO-biotin-mediated KARR-seq data in comparison with several control samples. (**D**) Pearson correlation coefficients between interaction maps based on individual KARR-seq replicates. Data generated by different dendrimers and heatmaps under different resolutions were analyzed, respectively. (**E**) Pearson correlation between KARR-seq data generated by different dendrimers.

While numerous approaches have been developed to map RNA secondary structures transcriptome-wide, very few methods can be applied to effectively determine RNA tertiary structure, namely the physical distance between nearby RNA regions. For instance, psoralen-based methods are not feasible for distance measurement, because psoralen only crosslinks duplexes with

a fixed physical distance. We speculated that the physical distance should correlate with the frequency of proximity ligation, which can be read out by the abundance of KARR-seq chimeric reads. We plotted an interaction map of mouse 18S rRNA based on KARR-seq data in mESCs and compared it with the 18S rRNA physical distance map revealed by Cryo-EM[128]. We found that the contact frequency measured by KARR-seq matches very well with the real physical distance among different regions of 18S rRNA (Figure 4.3C). To evaluate this correlation quantitatively, we aligned KARR-seq chimeric reads onto the 18S CryoEM coordination map. We calculated the Euclidean distances between regions and plotted all the distances as a distribution. We found that the distribution estimated by KARR-seq overlaps well with the exact distribution revealed by Cryo-EM (Figure 4.3D). In comparison, RIC-seq and PARIS enrich interactions within short physical distances (Figure 4.3D), suggesting that KARR-seq outperforms these methods in physical distance estimation.

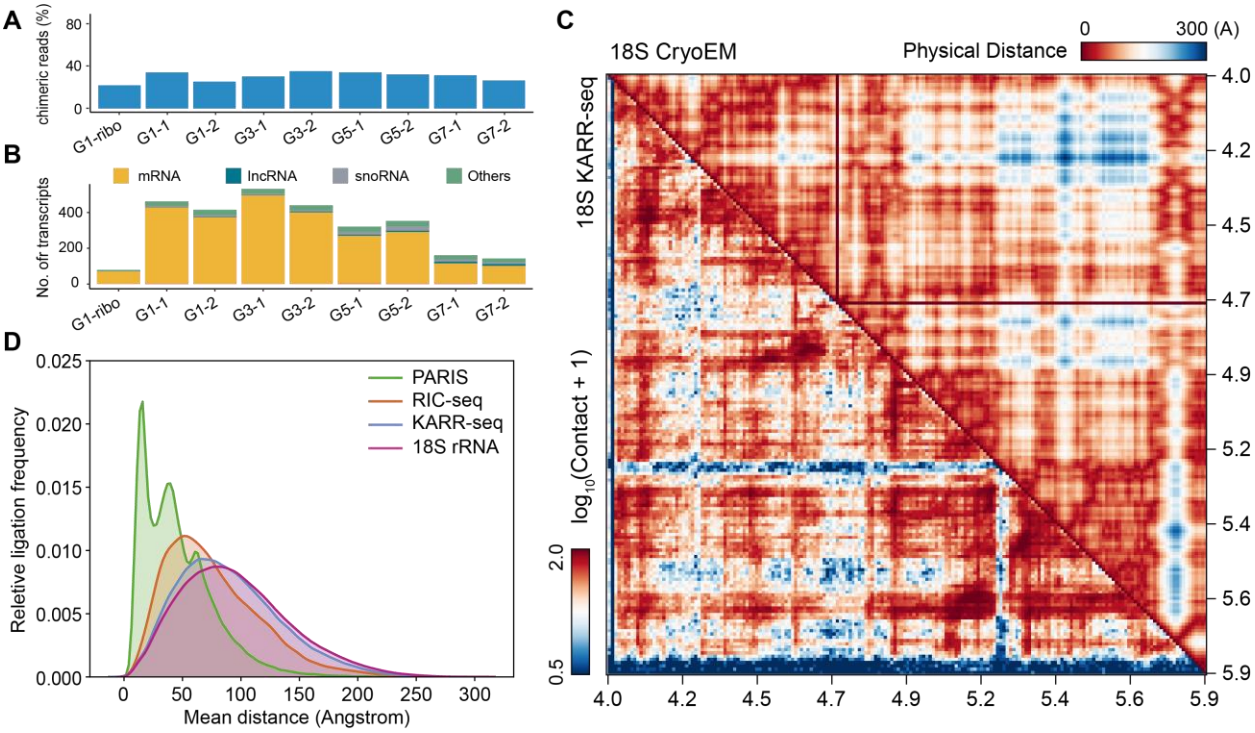

Figure 4.3 KARR-seq reveals the physical distance between physically close RNAs

(**Figure 4.3, continued**) (**A**) The ratio of chimeric read over other mapped reads among KARR-seq data produced by different dendrimers. (**B**) The number of transcripts captured by different dendrimers with valid interactions (chimeric reads). Different RNA species were shown in different colors. (**C**) Bottom right: the interaction map of mouse 18S rRNA produced by KARR-seq data in mESCs. Top right: the physical distance map based on the CryoEM structure of 18S rRNA. (**D**) The distribution of all distances among 18S rRNA estimated by KARR-seq, RIC-seq, and PARIS, respectively, in comparison with the actual distance distribution revealed by cryoEM.

Compared with rRNA and relatively abundant non-coding RNAs, mapping the 3D structure of mRNA is particularly challenging, because of their relatively low abundance and more dynamic structures. We evaluated the behavior of KARR-seq in 3D structure mapping of mRNA in both human and mouse cell lines. We set up 3D structure models of mRNAs based on RNA folding information and compared them with the KARR-seq interaction maps. We found that the many long-range interactions are conserved between human and mouse (Figure 4.4A and C), confirming a critical role of 1D sequences in shaping 3D folding. KARR-seq interaction maps recapitulate the main features in the physical distance map (Figure 4.4B and D), which demonstrates the power of KARR-seq in 3D structure detection.

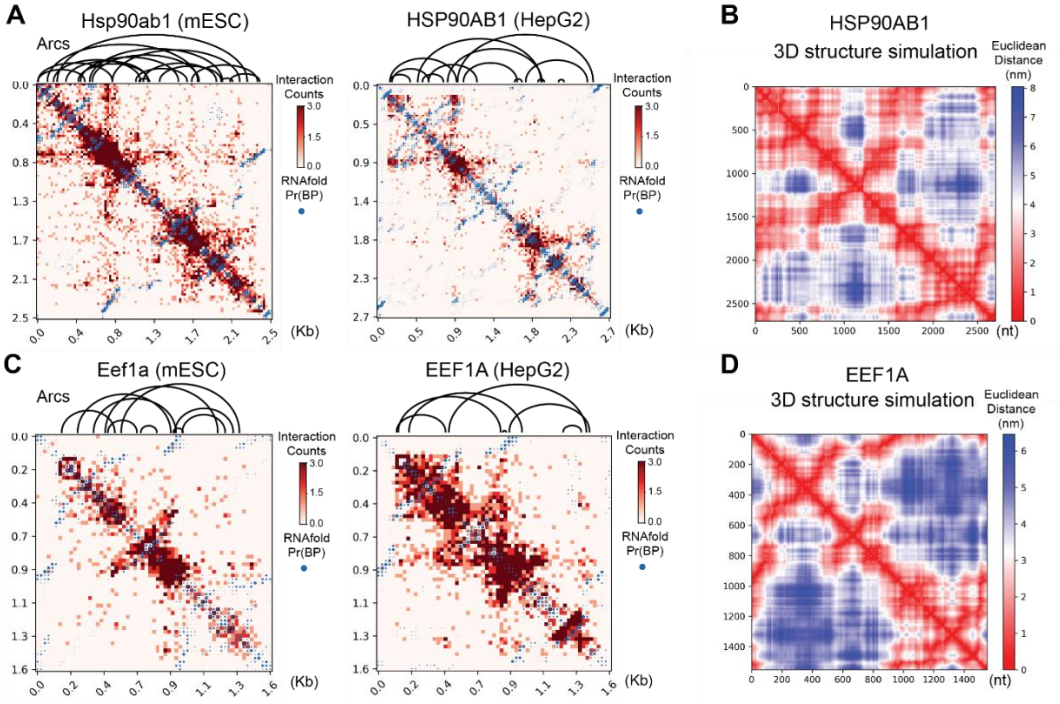

Figure 4.4 KARR-seq detects mRNA 3D structures

(**Figure 4.4, continued**) (**A**) and (**C**) KARR-seq interaction maps of HSP90AB1 (**A**) and EEF1A (**C**) transcripts in mESC (left) and HepG2 cells (right). (**B**) and (**D**) Physical distance maps of HSP90AB1 (**B**) and EEF1A (**D**) transcripts based on the 3D structure simulation.

4.2.3 Benchmarking KARR-seq with other RNA-RNA interaction mapping methods

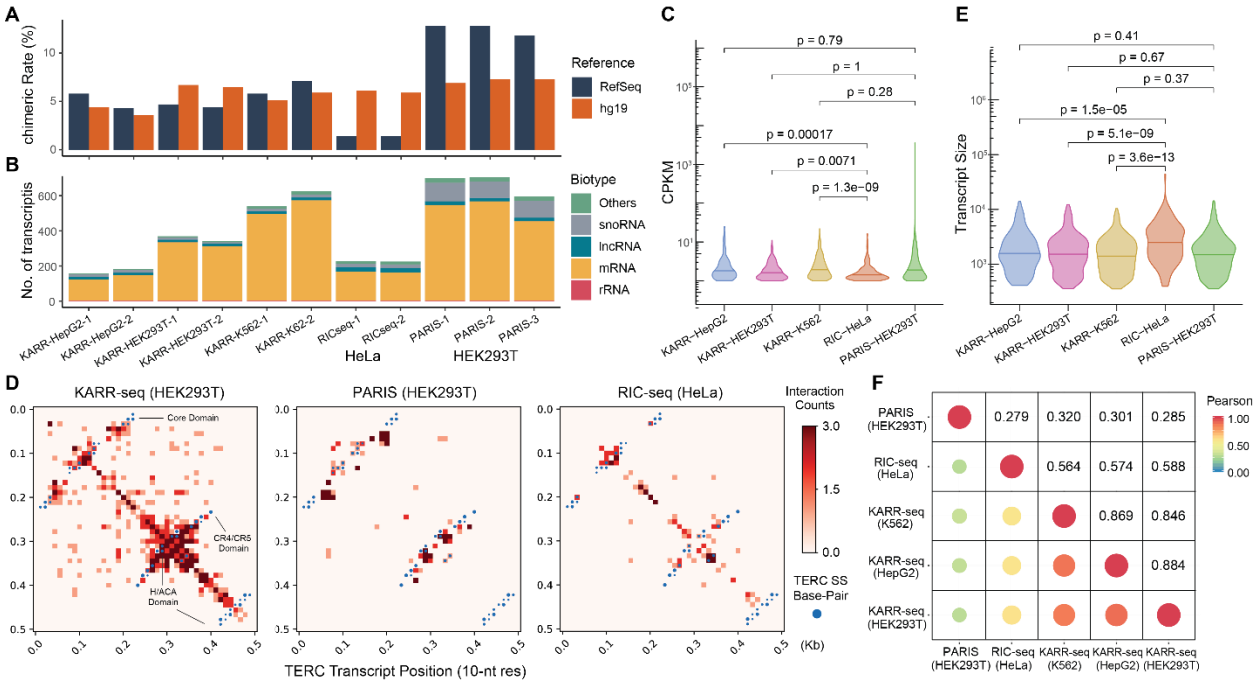

Figure 4.5 Comparing KARR-seq with RIC-seq and PARIS technology

(**A**) The chimeric reads ratio for KARR-seq, RIC-seq, and PARIS in different cell lines when mapped to Refseq and hg19, respectively. (**B**) The number of transcripts with valid interactions captured by KARR-seq, RIC-seq, and PARIS in different cell lines, respectively. (**C**) The chimeric reads abundance for transcripts with interactions captured by different technologies. (**D**) The interaction map of the TERC transcript produced from KARR-seq (HEK293T), PARIS (HEK293T), and RIC-seq (HeLa) data. (**E**) The size of transcripts with interactions captured by different technologies. (**F**) Pearson correlation between KARR-seq, RIC-seq, and PARIS. P values were calculated by one-sided Mann-Whitney test for (**C**) and two-sided Mann-Whitney test for (**E**).

We next performed KARR-seq in K562, HEK293T, and HepG2 cells (with rRNA depletion) and compared KARR-seq with RIC-seq and PARIS, which represent formaldehyde- and psoralen-mediated approaches, respectively. KARR-seq shows similar chimeric reads ratios when taking either Refseq or the hg19 genome as the reference for mapping. The chimeric reads ratios among all three methods are comparable when taking hg19 as the reference. However, RIC-

seq shows very low chimeric reads ratios when it maps to Refseq (Figure 4.5A), consistent with the observation that RIC-seq significantly enriches introns. The number of transcripts with valid interactions capture by KARR-seq varies among cell types, with only around 200 transcripts detected in HepG2 (comparable with RIC-seq performed in HeLa) but more than 500 transcripts in K562 (similar with PARIS performed in HEK293T) (Figure 4.5B). The abundance of chimeric reads in KARR-seq and PARIS are higher than RIC-seq (Figure 4.5 C), and higher abundance usually results in more structure information (Figure 4.5D). Interacting transcripts detected in KARR-seq share a similar size distribution as that in PARIS, while RIC-seq notably enriches longer transcripts (Figure 4.5E). Interactions revealed by KARR-seq show a higher correlation with RIC-seq than PARIS (Figure 4.5 F), because both KARR-seq and RIC-seq capture physically proximal transcripts while PARIS detects RNA duplexes.

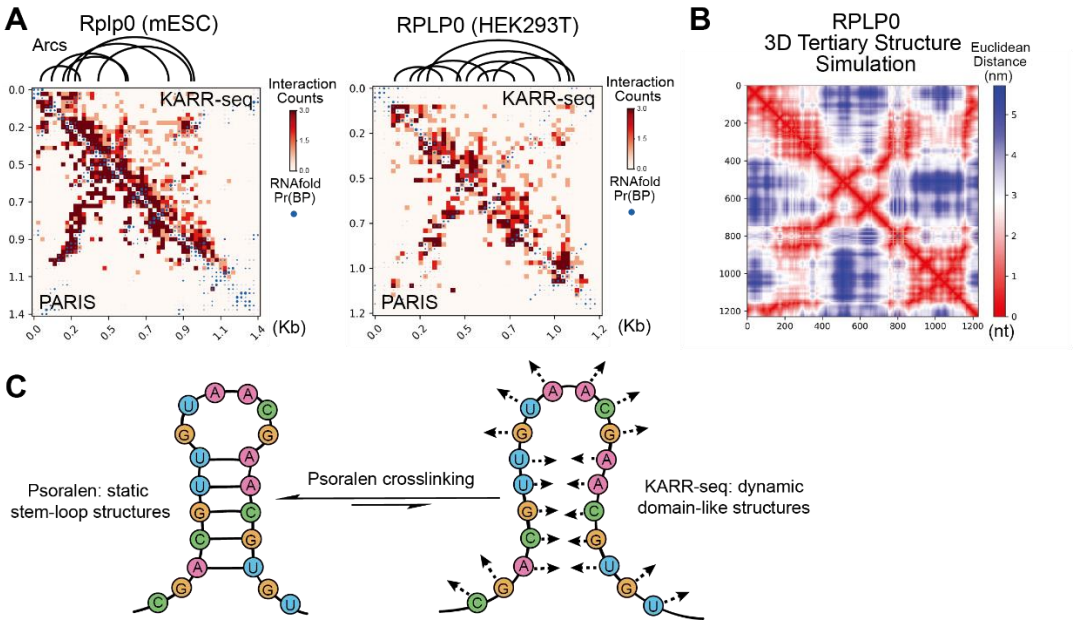

Figure 4.6 KARR-seq reveals domain-like RNA 3D structures

(**A**) The interaction heatmap of RPLP0 transcript in mESCs and HEK293T cells revealed by KARR-seq (top right) and PARIS (bottom left), respectively. (**B**) The physical distance map of the RPLP0 transcript based on the 3D structure simulation. (**C**) Proposed RNA conformations captured by psoralen crosslinking and KARR-seq.

On the 2D heatmap, PARIS data are presented as segments that are perpendicular to the diagonal, which is consistent with the fact that PARIS directly captures RNA duplexes (Figure 4.6A). KARR-seq data, however, were shown as a series of right triangles. The RNA regions included in KARR-seq triangles overlap with the regions that PARIS maps (Figure 4.6A). Therefore, while psoralen crosslinking specifically enriches RNA duplexes, KARR-seq reveals all possible conformations that an RNA polymer can exist as, which is essential to 3D structure mapping (Figure 4.6B). The triangles shown by KARR-seq heatmap suggested the existence of dynamic domain-like structures in RNA (Figure 4.6C). For each RNA transcript, many conformations can co-exist in a cell population or in each cell as an equilibrium. The stem-loop structures revealed by psoralen crosslinking is one of the possible conformations at steady-state, but KARR-seq detects all these possibilities at the same time. As psoralen crosslinking is not reversible under cellular conditions, the crosslinking reaction may also shift the equilibrium to the direction of duplex formation.

4.2.4 In vivo long-range interactions are suppressed by translation and are associated with RBP binding

RNA structure and interactions were dynamic and can respond to perturbations of cellular environments. RNA secondary structure has been shown to shape RBP binding and play critical roles in stress adaptation. However, how long-range RNA-RNA interactions interplay with different cellular contexts has not been fully understood. To study these cellular effects systematically, we designed an in vitro KARR-seq, in which N_3-kethoxal labeling and dendrimer crosslinking were performed on refolded RNA in solutions. Interactions captured by in vitro KARR-seq should reveal the intrinsic ability of RNA to form long-range interaction, which is determined by the RNA sequences but independent of cellular environments.

When we compare KARR-seq performed in vitro and in vivo, many more long-range interactions were detected in vitro, as shown by both interaction heatmaps (Figure 4.7A) and 1D arcs (Figure 4.7B). Statistically, for all transcripts with valid self-interactions, we plotted the correlation between coordinate distance and contact frequency. The contact frequency decreases almost linearly with respect to the coordinate distance, and the slope of the line was defined as the beta coefficient. The beta coefficient for in vitro KARR-seq is around -1.5 (Figure 4.7B), which fits the free-jointed chain polymer model, suggesting the validity of in vitro KARR-seq. The beta coefficient for in vivo KARR-seq is around -2.4, showing less long-range interactions than the in vivo sample (Figure 4.7B). We also calculated the beta coefficient for individual transcripts and plotted the distribution of all beta coefficients. In vivo KARR-seq shows a much broader distribution than that in vitro (Figure 4.7B), suggesting the complexity of RNA-RNA interactions in cells. The median of the in vivo distribution is smaller (Figure 4.7B), validating that many interactions were sequestered in cells.

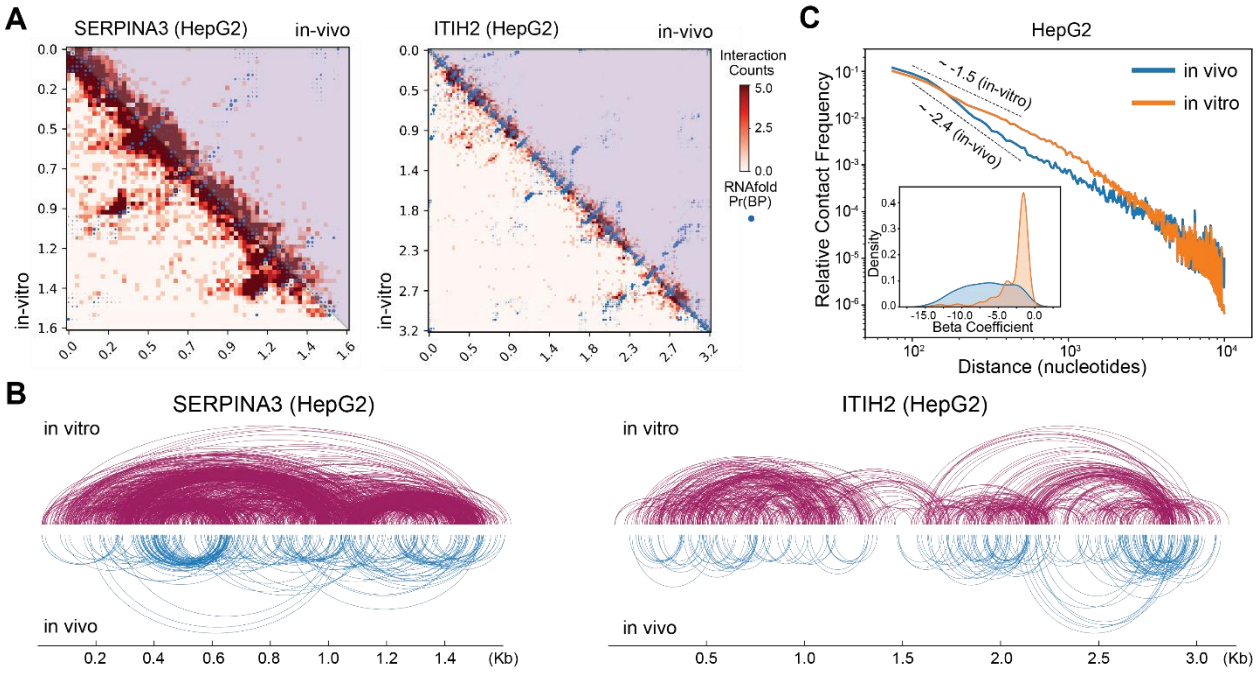

Figure 4.7 RNA-RNA interactions are relatively sequestered in vivo

(**Figure 4.7, continued**) (**A**) 2D heatmap for SERPINA3 and ITIH2 transcripts in HepG2 cells (top right) and re-folded HepG2 RNA in vitro (bottom left). (**B**) 1D arcs for data presented in (**A**). (**C**) Outer: the correlation between coordinate distance and contact frequency in HepG2 cells and re-folded HepG2 RNA. All transcripts with valid interactions were used for analysis. The estimated slope of the lines denotes the beta coefficient. Inner: the distribution of beta coefficients for all individual transcripts.

We next asked what cellular factors suppress long-range interactions. We first speculated that RNA helicase could unwind interaction duplexes. We knocked down DEAD-box protein 3, X isoform (DDX3X), an ATP-dependent RNA helicase in HepG2 cells, and performed KARR-seq. Surprisingly, we did not see notable differences between the knockdown and the control samples (data not shown). We also envision that RNA can be unwound by ribosome translocation during translation. We performed KARR-seq by using harringtonine and cycloheximide-treated HepG2 cells, respectively. Harringtonine depletes elongation ribosomes from mRNAs, and cycloheximide freezes elongating ribosomes[129]. After translation inhibitor treatment, the level of long-range interactions increased to a level that is similar to that in vitro (Figure 4.8A-B). Statistically, after translation inhibition, the distribution of beta coefficients in vivo right-shifted to the direction of in vitro (Figure 4.8C). Metagene plot shows that the increased interactions are mainly located in the ORF region of the mRNA (Figure 4.8D). These results collectively suggest that ribosome elongation suppresses the long-range interactions of mRNA. Moreover, we defined interaction anchor as the region with maximal base-pairing strength within the duplex group. We checked the level of RBP binding on these anchors by using all eCLIP data in HepG2 and K562 produced by ENCODE. We observed a higher RBP binding on these anchors in both cell lines, suggesting the potential association of RBP with RNA-RNA interactions. Additional future studies are required to reveal effects from individual RBPs.

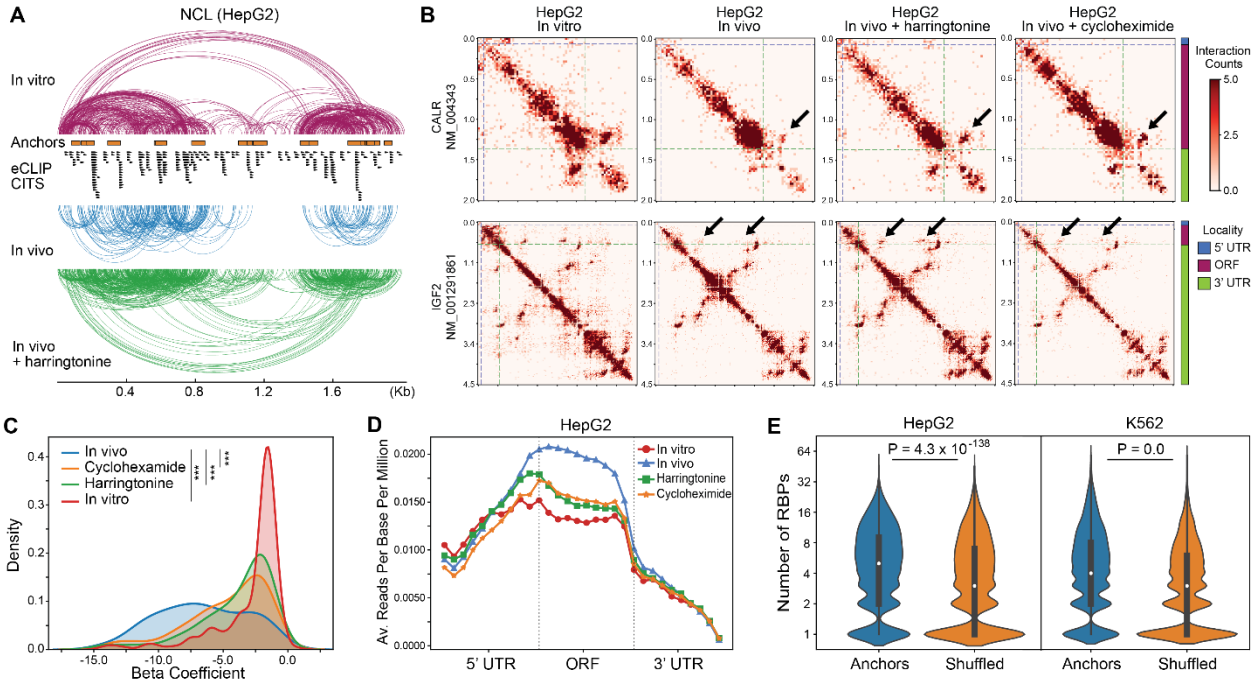

Figure 4.8 Translation machinery and RBP binding suppress RNA-RNA interactions

(**A**) 1D interaction arcs for the NCL transcript in HepG2 RNA, native HepG2 cells, and harringtonine-treated HepG2 cells. (**B**) 2D interaction maps for CALR and IGF2 transcripts in HepG2 RNA, native HepG2 cells, harringtonine-treated HepG2 cells, and cycloheximide-treated HepG2 cells. (**C**) The distribution of beta coefficients in different conditions. (**D**) The metagene plot of mRNA long-range interactions in different conditions. (**E**) The number of RBP binding sites at interaction anchor regions compared with the average in HepG2 and K562 cells, respectively.

4.3 Discussion and conclusion

By employing multi-functional RNA crosslinkers, I describe the usage of N_3-kethoxal in a two-dimension manner. The combination of N_3-kethoxal and DBCO-biotin-decorated PAMAM dendrimers facilitates mapping RNA-RNA interactions. KARR-seq detected known and new, intramolecular and intermolecular interactions. KARR-seq outperforms other existing technologies in measuring physical distances, therefore enables accurate determination of 3D structure for both mRNA and non-coding RNAs. KARR-seq is distinguished from RIC-seq, which seems to enrich nuclear transcripts, and psoralen-based methods, which directly captures RNA duplexes. We show that RNA tends to form more long-range self-interactions in vitro. And many

of these interactions are suppressed by translation and are related to the binding of RBPs in cells. RBPs may mediate the formation of some interactions; they may also preferentially bind to duplex RNAs. The details behind their associations will be explored in the future.

4.4 Methods

3.4.1 Cell culture

HEK293T cells (ATCC, CRL11268) and HepG2 cells (ATCC, HB8065) were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1% penicillin and streptomycin (Gibco). Murine embryonic stem (ES) cells (ATCC, CRL-1821) were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1 mM L-glutamine (Gibco), 0.1 mM β-mercaptoethanol (Gibco), 1% (v/v) nonessential amino acid stock (100×, Gibco), 1% penicillin/streptomycin stock (100×, Gibco), and 1,000 U/mL LIF (Millipore). K562 cells (ATCC, CCL243) were cultured in RPMI 1640 (Gibco 11875) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1% penicillin and streptomycin (Gibco). All cells were grown at 37 ℃ with 5% CO_2.

Cell lines used in this study were examined for mycoplasma contamination test using LookOut Mycoplasma PCR Kit (Sigma, MP0035).

3.4.2 Dot blot

Purified crosslinked RNA was subjected to dot blot analysis. 1 μL RNA was loaded onto the Amersham Hybond-N+ membrane (GE Healthcare, RPN119B). Membranes were air-dried and were crosslinked by UV stratalinker 2400 at 150 mJ/cm^2 twice. The membranes were then blocked overnight in 5% fatty-acid free BSA in PBST (0.1% Tween-20). The second day, the membrane was washed and incubated in streptavidin-HRP (Thermo, S-911) in PBST

supplemented with 3% fatty-acid free BSA. The membrane was washed in PBST for 5 times before developed by SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo, 34577).

3.4.3 The synthesis of G1-DBCO-biotin dendrimer

1.53 μmol of PAMAM dendrimer G1 (Sigma, 412384) was dissolved in 2 mL methanol. 3.06 μmol DBCO-NHS (Sigma, 761524), 1.53 μmol biotin-NHS (Sigma, 203112), and 5 μL triethylamine (Sigma, 471283) was added in the solution. The reaction mixture was stirred overnight at room temperature before the addition of 100 μL acetic anhydride (Sigma, 320102) and 100 μL triethylamine (Sigma, 471283) to modify unreacted amine branches. The reaction mixture was stirred for another 24 h at room temperature before the addition of 2 mL water to neutralize the reaction. The dendrimer solution was purified and concentrated with Microsep Advance Centrifugal Devices with Omega Membrane 1 K (Pall Corporation, MCP001C41) by series centrifugation at 5,000 g at 4 ℃.

Characterization and quantification of G1-DBCO-biotin dendrimer were performed by measuring the characteristic UV absorbance of the DBCO moiety at 295 nm. A series of DBCO-NHS solutions with known concentrations were prepared as the standard solutions. UV absorbance at 295 nm of each standard solution was measured by nanodrop. The calibration curve was plotted with the values of UV absorbance (295 nm) on the y-axis, and the concentrations of standard solutions on the x-axis. UV absorbance at 295 nm for the dendrimer was then measured. The concentration of the dendrimer was then determined by using the calibration curve.

3.4.4 KARR-seq

3.4.4.1 In vivo crosslinking and RNA purification

Cells were crosslinked in 1% formaldehyde diluted in culture medium for 10 min and then quenched with 125 mM glycine for 5 min. For each reaction, 2 million cells were resuspend cells

into 500 μL lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630, 5 mM EDTA) supplemented with 20 mM N_3-kethoxal, proteinase inhibitor (Sigma, 11873580011), and SUPERNase inhibitor (Thermo, AM2696). The cells were rotated at room temperature for 30 min and were collected by centrifuging at 2,500 g for 5 min. The cell pellet was washed by resuspending into 500 μL lysis buffer and was collected by centrifugation. Then resuspended the cells in 500 μL lysis buffer supplemented with 12.5 μM G1-DBCO-biotin dendrimer and shook the cell suspension at 1,000 rpm at 37 ℃ for 1 h. After the reaction, cells were collected and washed once as above. To isolate RNA, resuspended cell pellets in 410 μL 25 mM K_3BO_3, with 50 μL 10% SDS, 30 μL proteinase K (Thermo, 25530049), 10 μL SUPERNase inhibitor. The mixture was shaken at 55 ℃ for 2 h and was subjected to phenol-chloroform (Thermo, AM9722) extraction and ethanol precipitation.

The next day, dissolve RNA pellet in 104 μL 25 mM K_3BO_3. Add 12 μL 10× DNase I buffer (Thermo, AM8170G), 2 μL DNase I (Thermo, 18047019), and 2 μL SUPERNase inhibitor. The mixture was shaken at 1,000 rpm at 37 ℃ for 30 min. 130 μL 2× proteinase K buffer (100 mM Tris-HCl pH 7.5, 200 mM NaCl, 2 mM EDTA, 1% SDS) and 10 μL proteinase K (Thermo, 25530049) was then added. The mixture was shaken at 1,000 rpm at 55 ℃ for another 30 min before phenol-chloroform extraction and ethanol precipitation was performed.

3.4.4.2 RNA fragmentation and immunoprecipitation

Precipitated RNA was dissolved in a mixture of 61 μL 25 mM K_3BO_3, 7 μL 10× RNA fragmentation buffer (Thermo, AM8740), and 2 μL SUPERNase inhibitor. The mixture was heated at 70 ℃ for 15 min before 8 μL fragmentation stop buffer (Thermo, AM8740) was added. The mixture was then put on ice immediately.

Prepare 30 μL Dynabeads Myone Streptavidin C1 (Thermo, 65001) by washing once with 100 μL 1×binding/wash buffer (5 mM Tris-HCl pH 7.4, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20), once with 100 μL buffer A (100 mM NaOH, 50 mM NaCl), once with 100 μL buffer B (100 mM NaCl). Then block the beads with 100 μL binding/wash buffer containing 1 μg/μL BSA (NEB, B9000S) and 1 μg/μL salmon sperm DNA (Thermo, 15632011) by rotating at room temperature for 30 min. Wash beads once with 100 μL 1× binding and wash buffer. Beads were then resuspended in 80 μL 2×binding/wash buffer (10 mM Tris-HCl pH 7.4, 1 mM EDTA, 2 M NaCl, 0.1% Tween-20) and were mixed with the fragmented RNA. The beads-RNA mixture was rotated at room temperature for 20 min before washed twice with 100 μL 1×binding/wash buffer, and once with 100 μL 1×PNK buffer (diluted from 10×PNK buffer from NEB, M0201L).

3.4.4.3 RNA end repair and proximity ligation

Resuspend the beads in 41 μL 25 mM K_3BO_3. Add 5 μL 10×T4 PNK buffer, 3 μL T4 PNK (NEB, M0201L), and 1 μL SUPERNase inhibitor. Shake the beads at 1000 rpm at 37 ℃ for 30 min. Then add another 1 μL 10×T4 PNK buffer, 3 μL T4 PNK (NEB, M0201L), and 6 μL 10 mM ATP into the reaction mixture. Shake the tube at 37 ℃ for another 30 min. After the reaction, the beads were washed twice with 100 μL 1× binding/wash buffer and once with 100 μL 1× ligation buffer (diluted from 10×T4 RNA ligase buffer from NEB, M0437M). For ligation reaction, resuspend beads in 660 μL 25 mM K_3BO_3, add 100 μL 10×T4 RNA ligase buffer, 2 μL 10 mM ATP, 200 μL 50% PEG 8000, 20 μL T4 RNA ligase I (NEB, M0437M), and 10 μL SUPERNase inhibitor. Shake the reaction mixture at 1000 rpm (or rotate the tube) at 16 ℃ for 16 h.

3.4.4.4 RNA purification, library preparation, and sequencing

After ligation, collect and wash beads three times with 1×binding/wash buffer. Elute RNA by heating the beads in 50 μL H_2O at 95 ℃ for 10 min. Purify the RNA by using RNA Clean &

Concentrator kit (Zymo, R1014) following the manufacturer's protocol. Elute RNA with 30 μL H_2O. Measure the concentration of the RNA and take 10 ng RNA for library construction by using the SMARTer Stranded Total RNA-seq kit v2 – pico input mammalian (Takara, 634413). Libraries were sequenced on the Illumina Novaseq4000 platform, PE150 mode, with around 80 million reads per sample.

Chapter 5

N^6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA

5.1 Introduction: N^6-Deoxyadenosine Methylation (6mA) in eukaryotes

DNA methylation is a widespread epigenetic mechanism that plays critical roles in a wide range of biological processes[130]. 5-methylcytosine (5mC) is predominantly found in higher eukaryotes[131], especially in plants[132] and mammals[133]. In contrast, N^6-methyldeoxyadenosine (6mA) is widespread in prokaryotes and functions primarily in restriction-modification (R-M) systems[134]. Prokaryotes use 6mA to discriminate against the host genomic DNA from foreign pathogenic DNA to protect the host genome. As there is no equivalent bacterial restriction-modification system found in eukaryotes, the distributions and biological functions of 6mA in eukaryotic genomic DNA (gDNA) had, until recently, remained elusive. By taking advantage of more sensitive detection techniques and high-throughput sequencing, recent studies uncovered the genome-wide distribution and potential regulatory roles of 6mA in the genomes of *Chlamydomonas reinhardtii*[12], ciliates[135,136], *Caenorhabditis elegans*[137], *Drosophila*[138], and fungi[139].

While the levels of 6mA in genomes of certain invertebrates are relatively high, 6mA levels in the genomes of mammals tend to be low, ranging from a few to tens of ppm of the total deoxyadenosines to even lower under normal growth conditions[13,14,140]. Nevertheless, the potential roles of DNA 6mA in vertebrates and mammalian cells have been proposed[140-143]. For instance, in mouse embryonic stem cells, 6mA was enriched in young and active LINE1 transposons on the X chromosome at the frequency of 25–30 ppm of deoxyadenosine[13]. In glioblastoma cancer stem cells and patient tumors, 6mA was found as a repressive maker, which associates with repressive histone mark H3K9me3 in genes related to neurogenesis and neuronal development[141]. 6mA was

also shown to play roles in stress response. For instance, 6mA was identified in the mouse brain in response to chronic restraint stress[142]. *C. elegans* also adopts DNA 6mA as a transgenerational epigenetic modification that confers mitochondrial stress adaptation[143].

5.2 Results

5.2.1 The enrichment of 6mA in human mitochondrial DNA (mtDNA)

Although 6mA is not abundant in the nuclear genomes of most mammalian cells, we hypothesized that 6mA could be enriched in specific subcellular locations and exert certain functions. To test this hypothesis, we quantified the abundance of 6mA in nuclear genomic DNA (gDNA) and in purified mtDNA (isolated from the same HepG2 cells) via ultra-performance liquid chromatography mass spectrometry **(**UHPLC-QQQ-MS/MS). Consistent with a previous study, we detected a very low level (up to ~0.3 ppm) of 6mA/dA in HepG2 gDNA (Figure 5.1A-B). Crude mtDNA was extracted from intact mitochondria, which were isolated by utilizing a commercially available kit. This method yields a ~6-fold enrichment of mtDNA over total DNA by RT-qPCR (Figure 5.1A), resulting in a 6mA/dA level at ~20 ppm, a more than 60-fold enrichment over that in HepG2 gDNA (Figure 5.1B). To further eliminate gDNA contamination and enrich mtDNA, a circular DNA-safe DNase was applied to digest most linear chromosomal DNA but not circular mtDNA. After digestion, mtDNA was more than 100-fold enriched over total DNA by RT-qPCR with the 6mA/dA level reaching ~400 ppm (Figure 5.1A-B), representing a more than 1,300-fold enrichment than that in gDNA and corresponding to approximately four 6mA modifications per mtDNA molecule.

Note that even after DNase digestion, there is still residual nuclear DNA contamination left. The real content of 6mA in mtDNA should be even higher if future approaches are developed to further enrich mtDNA. The same approach was applied to quantify 6mA level in gDNA and

mtDNA in other cells and tissues such as MDA-MB-231 cells, 143B cells, mouse primary fibroblast cells, testes, and spleen. 6mA was shown to be enriched in mtDNA for all these tested cell lines and tissues (Figure 5.1C-D).

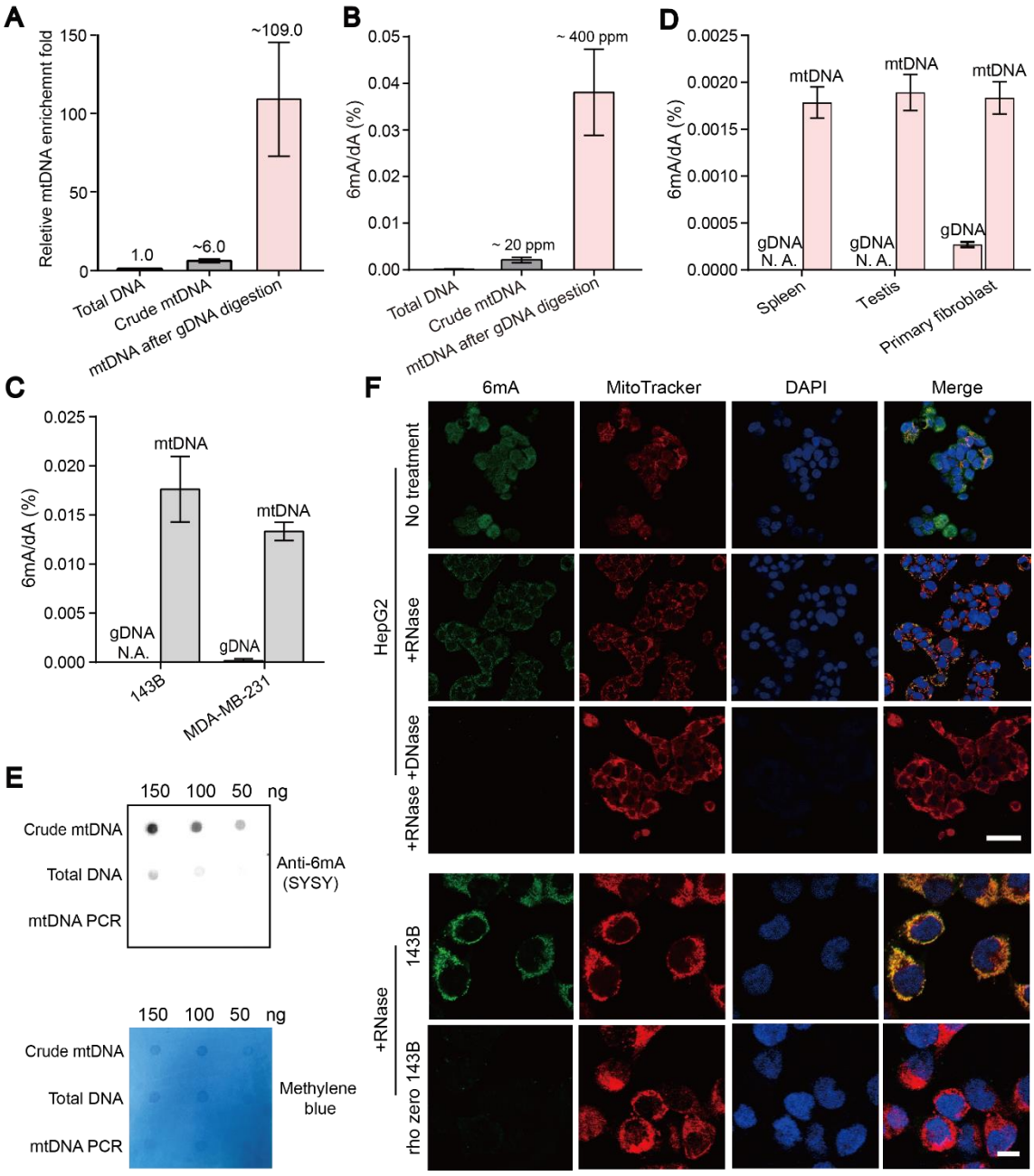

Figure 5.1 The presence of N^6-deoxyadenosine methylation (6mA) in human mtDNA

(**A**) Relative mtDNA enrichment folds in total DNA, crude mtDNA, and mtDNA after gDNA digestion from HepG2 cells. (n = 3, mean ± SEM). (**B**) UHPLC-QQQ-MS/MS showing 6mA/dA

(**Figure 5.1, continued**) levels in total DNA, crude mtDNA, and DNase digested mtDNA in HepG2 cells (n = 3, mean ± SEM). (**C**) UHPLC-QQQ-MS/MS showing 6mA levels in gDNA and mtDNA in 143B cells and MDA-MB-231 cells (n = 2, mean ± SEM). (**D**) UHPLC-QQQ-MS/MS showed 6mA levels in gDNA and mtDNA in mouse spleen, testis, and primary fibroblast cells derived from mice directly (n = 2, mean ± SEM). (**E**) 6mA dot blot of total DNA, crude mtDNA, and PCR amplified mtDNA. (**F**) Top: 6mA signals (green) and their co-localization with mitochondria marker (red) in HepG2 cells with no treatment, RNase treatment, and DNase + RNase treatment; scale bar: 100 μm. Bottom: 6mA signals (green) and their co-localization with mitochondria marker (red) in normal and rho zero 143B cells after RNase treatment; scale bar: 10 μm.

We then performed dot blot analysis using an anti-6mA antibody and detected 6mA signals only with purified endogenous mtDNA, but not with total DNA or PCR-amplified mtDNA (Figure 5.1E), confirming the presence of 6mA in mitochondria. The existence of 6mA in mtDNA was further supported by immunofluorescence staining. As the 6mA antibody also recognizes RNA m^6A modification, stringent RNase treatment was applied to eradicate signals derived from endogenous RNA. After RNase treatment, 6mA signals (green) in HepG2 cells co-localized with MitoTracker (red), with weak to no visible signal detected in the nucleus (Figure 5.1F, top). Besides, we did not detect 6mA signals in DNase- and RNase-treated HepG2 cells. We then repeated the imaging experiment by using typical and rho zero 143B cells, in which mtDNA is depleted[144]. Upon RNase treatment, 6mA signals (green) overlap well with MitoTracker (red) in typical 143B cells; but we failed to detect 6mA signals in the rho zero cells, indicating that the signals detected in typical 143B cells were indeed derived from mtDNA (Figure 5.1F, bottom). Altogether, these data strongly suggest that human mtDNA contains enriched 6mA.

5.2.2 Mapping 6mA location in human DNA

We next sought to map 6mA distribution in mtDNA by using the anti-6mA antibody to enrich 6mA-containing DNA fragments and to identify potential consensus methylation motifs. Considering the small size of the human mitochondrial genome (16,569 base-pair length), we employed a photo-crosslinking, enzyme digestion, and immunoprecipitation-based 6mA-mapping

approach (6mA ChIP-exo)[145] to map 6mA in mtDNA at a relatively high resolution (Figure 5.2A). Two different anti-6mA antibodies from two different companies were applied to validate the results obtained by each other.

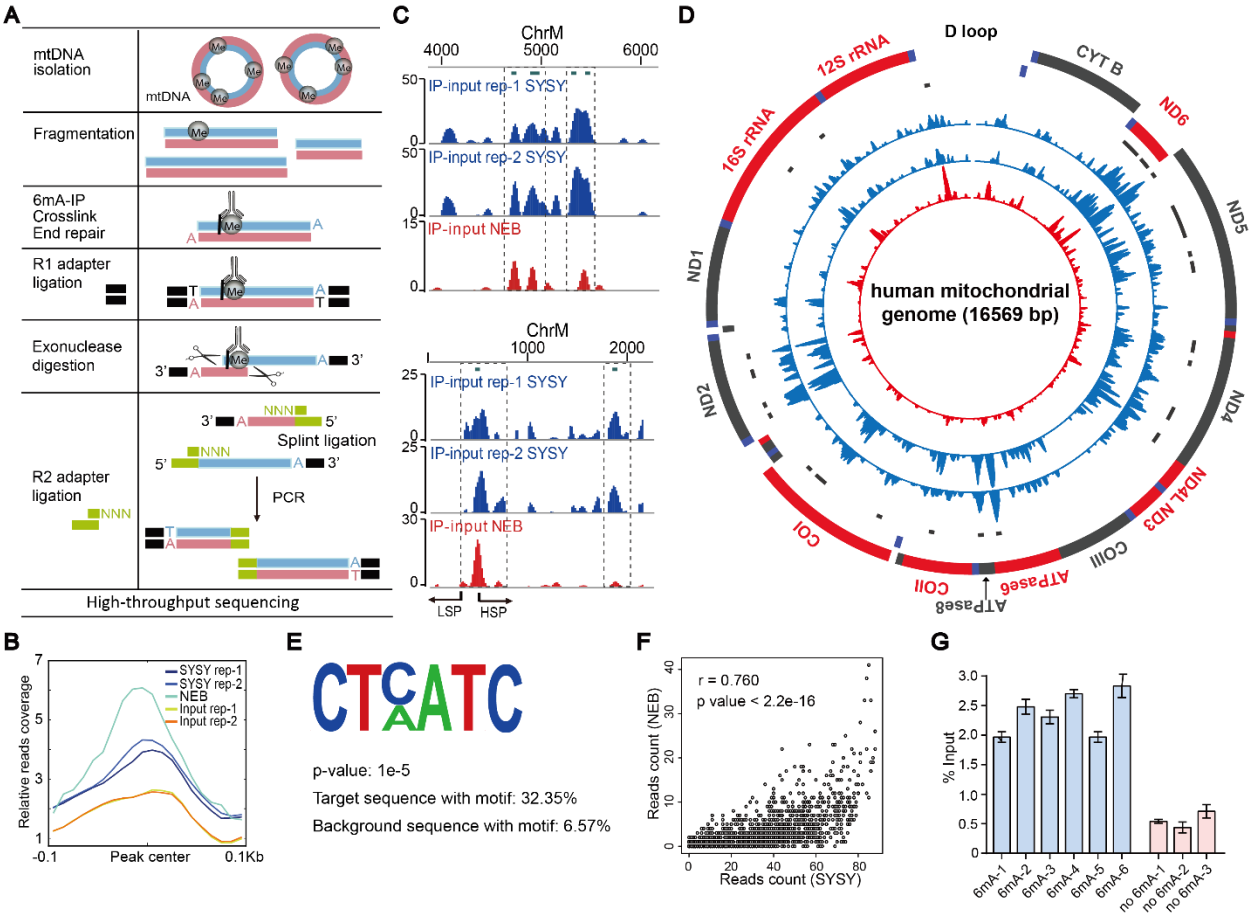

Figure 5.2 The distribution of 6mA in human mtDNA

(**A**) The workflow of 6mA mapping in human mtDNA using high-throughput sequencing. (**B**) A metapeak profile shows the relative reads coverage of input and IP samples near identified 6mA peaks. (**C**) Two examples of 6mA peaks (IP - input) detected by both SYSY and NEB antibodies. Reads from input libraries have been subtracted. The bottom penal showed the peak detected in the promotor region. (**D**) The circos plot showing the 6mA distributions in mtDNA. The mitochondrial genome is shown as the outermost circle in red/gray/blue with gene annotations. 6mA profiles (IP - input) revealed by SYSY-IP rep-1 (blue), SYSY-IP rep-2 (blue), and NEB-IP (red) were shown in three tracks, respectively, from outside to inside. Gray dots and bars indicate 6mA peaks. Reads from input libraries have been subtracted. (**E**) Consensus motifs identified from 6mA-containing regions. (**F**) Spearman correlation analysis of 6mA-IP profiles generated by using NEB and SYSY anti-6mA antibodies, respectively. Spearman r = 0.760. (**G**) 6mA-IP qPCR validation of 6mA-positive and 6mA-negative sites revealed by 6mA mapping (n = 2, mean ± SEM).

Metagene profile depicts a successful 6mA enrichment by 6mA ChIP-exo (Figure 5.2B). 6mA ChIP-exo revealed 23 high-confidence 6mA modification sites ($p < 0.01$) that showed up in the promoter region of mtDNA and clustered in ND2, COI, and ND4-6 regions (Figure 5.2C-D). A consensus sequence of CTCATC was discovered in these 6mA-enriched sites (Figure 5.2E), suggesting a potential sequence preference of the mtDNA 6mA modification. Libraries generated by two different antibodies show a strong positive correlation (spearman r = 0.76) (Figure 5.2F), and 6mA peaks discovered by these two antibodies overlapped very well (Figure 5.2C-D).

6mA-IP qPCR was performed to validate 6mA sites identified by 6mA-mapping. We examined 6 sites among 23 identified 6mA sites and three 6mA-negative sites. The qPCR results showed ~3 to 5-fold enrichment on all selected 6mA-containing sites compared to 6mA-negative sites (Figure 5.2G).

5.2.3 METTL4 protein accumulates in mitochondria

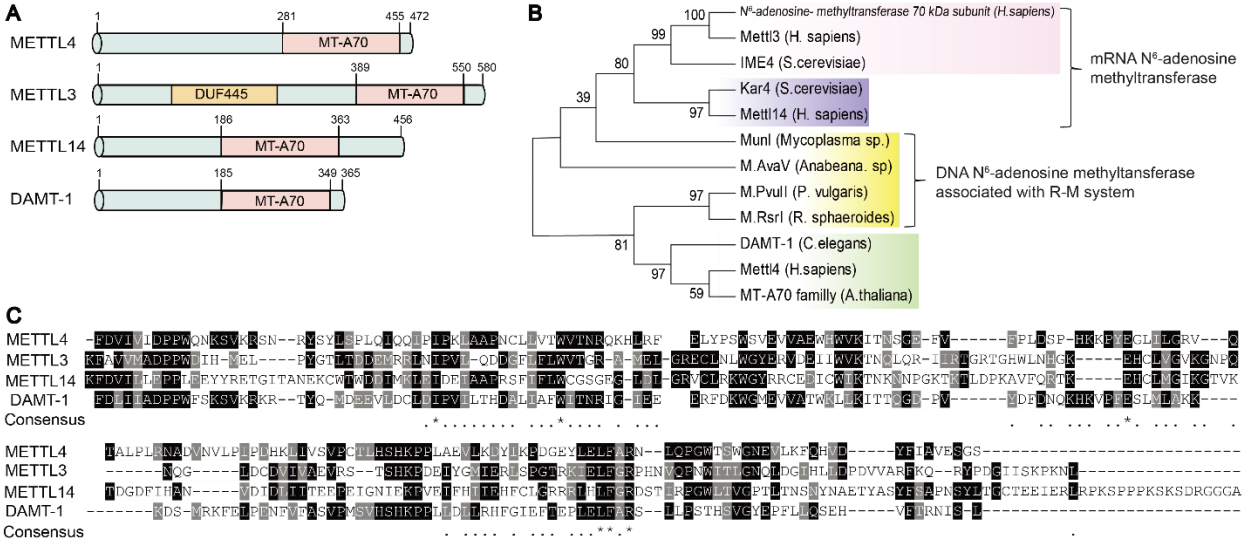

Figure 5.3 Relationship between METTL4 and its homologues

(**A**) Multiple alignments of METTL4 to RNA m^{6}A methyltransferases METTL3/METTL14 and *C. elegans* DNA 6mA methyltransferase DAMT-1. (**B**) A neighbor-joining consensus tree illustrates the phylogenetic relationship of MTA70 family members in eukaryotes and their closest prokaryotic homologues. Each color represents the one lineage of the MTA70 family. (**C**) Multiple sequence alignment of the methyltransferase MTA70 domain of METTL4 with those from

(**Figure 5.3, continued**) METTL3/METTL14 and DAMT-1 using T-coffee and Boxshade programs. Identical and semi-conserved residues are highlighted in black and grey, respectively.

There are three close homologous in the MTA70 family methyltransferases in mammals, namely, METTL3, METTL14, and METTL4 (Figures 5.3A-B), which are thought to have evolved from the MunI-like bacterial DNA 6mA methyltransferases[146]. METTL3 and METTL14 form a heterodimer complex to mediate the mammalian m^{6}A mRNA methylation[147]. The third member, METTL4, is also a homologue of the DNA 6mA methyltransferase DAMT-1 found in *C. elegans* (Figures 5.3B-C) and was proposed as a potential candidate for mammalian DNA 6mA methylation[148].

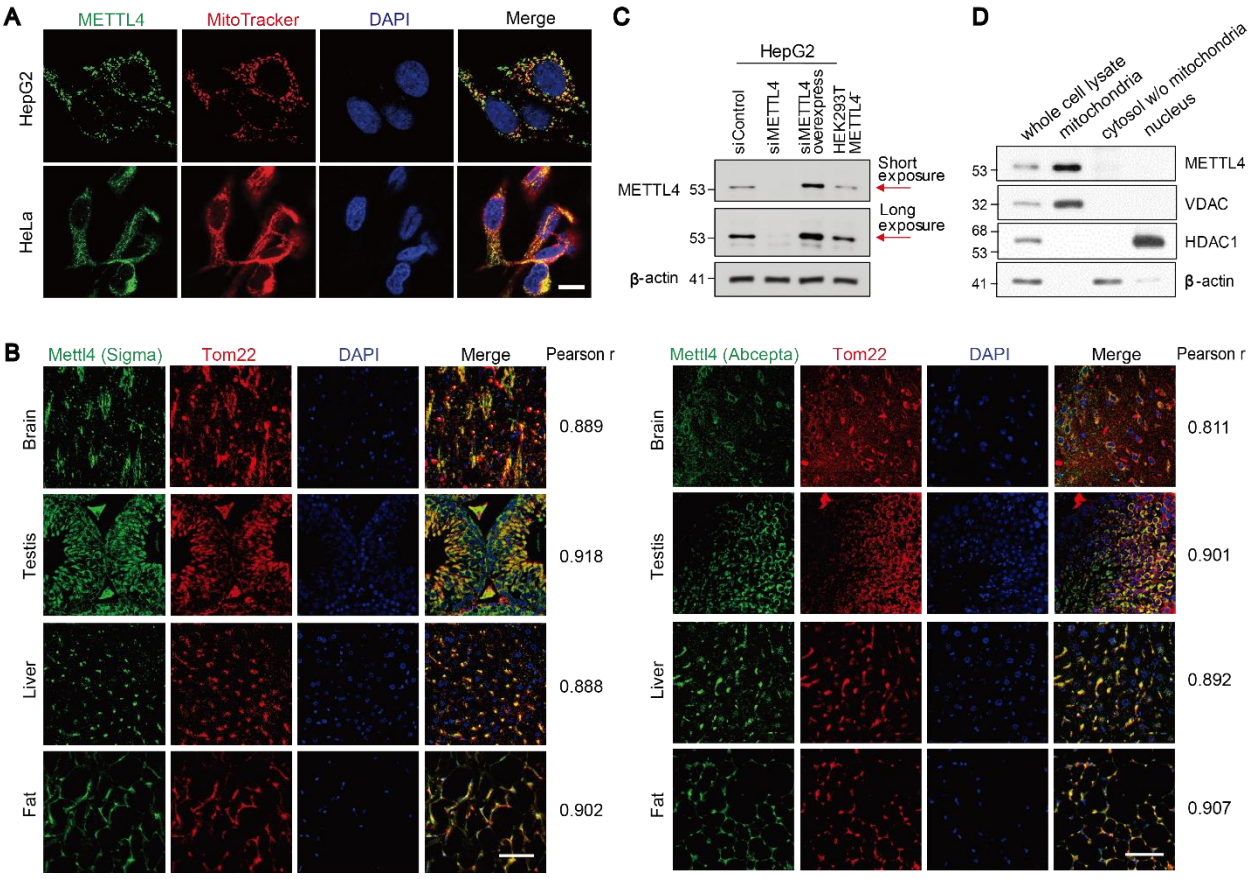

Figure 5.4 Subcellular localization of METTL4 protein

(**A**) Immunofluorescence displayed the co-localization of METTL4 with mitochondria in HepG2 and HeLa cells. Scale bar: 20 μm. The co-localization of METTL4 with mitochondria was evaluated by the Pearson correlation coefficient, with Rr = 0.60 for HepG2 and 0.74 for HeLa. (**B**)

(**Figure 5.4, continued**) Immunofluorescence using two anti-METTL4 antibodies from Sigma (left panel) and Abcepta (right panel), respectively, displaying the co-localization of METTL4 with mitochondria in the mouse brain, testis, liver and fat tissues. Scale bar: 100 μm. The co-localization of METTL4 with mitochondria was evaluated by Pearson correlation analysis with the r-values shown on the right side of the images. (**C**) The anti-METTL4 antibody was validated by using control, METTL4 knockdown, METTL4 knockdown followed by METTL4 overexpression in HepG2 cells, as well as purified METTL4 protein expressed from HEK293T cells. (**D**) Western blot revealed the uneven distribution of the METTL4 protein in different subcellular fractions and the enrichment of METTL4 in mitochondria. VDAC (mitochondrial), β-actin (cytosolic), and HDAC1 (nuclear) were chosen as compartment-specific markers demonstrating the purity of each subcellular fraction.

We performed immunofluorescence with an anti-METTL4 antibody to determine the localization of METTL4 in mammalian cells. We observed the co-localization of METTL4 with MitoTracker in HepG2 and HeLa cells, indicating a substantial accumulation of METTL4 in mitochondria (Figure 5.4A). We also performed immunofluorescence in the mouse brain, testis, liver, and fat, with two METTL4 antibodies produced by using the N-terminal and C-terminal immunogen sequences, respectively. The results generated by both antibodies showed that METTL4 co-localizes with mitochondria in all tested tissues (Figure 5.4B). We further examined METTL4 distribution in different subcellular fractions by western blot. We found that METTL4 is significantly enriched in the mitochondrial fraction, with a small portion located in fractions of the cytosol and nucleus. (Figures 5.4C-D).

We next employed mitochondrial localization signal searching tools, MitoProt II and TargetP, to analyze the METTL4 sequence, which showed that METTL4 has a low score in N-terminal mitochondrial targeting sequence (MTS) prediction. However, a recently developed tool, Integrated Mitochondrial Protein Index (IMPI)[149], which integrates the multiple datasets and machine learning algorithm, shows a high possibility of METTL4 to be a mitochondrial protein (IMPI score: 0.748).

5.2.4 METTL4 mediates mtDNA 6mA methylation

Next, we evaluated the relationship between METTL4 and 6mA. We knocked down METTL4 in HepG2 cells and observed a 30% decrease of the 6mA/dA ratio in the mtDNA compared to the control (Figure 5.5A). Consistent with published results[147,150], we noticed that METTL4 knockdown has no significant effect on m^6A nor m^6Am levels for mitochondrial rRNAs, mRNAs, and small RNAs (Figure 5.5B-C), suggesting that METTL4 acts on mtDNA rather than on mitochondrial RNAs.

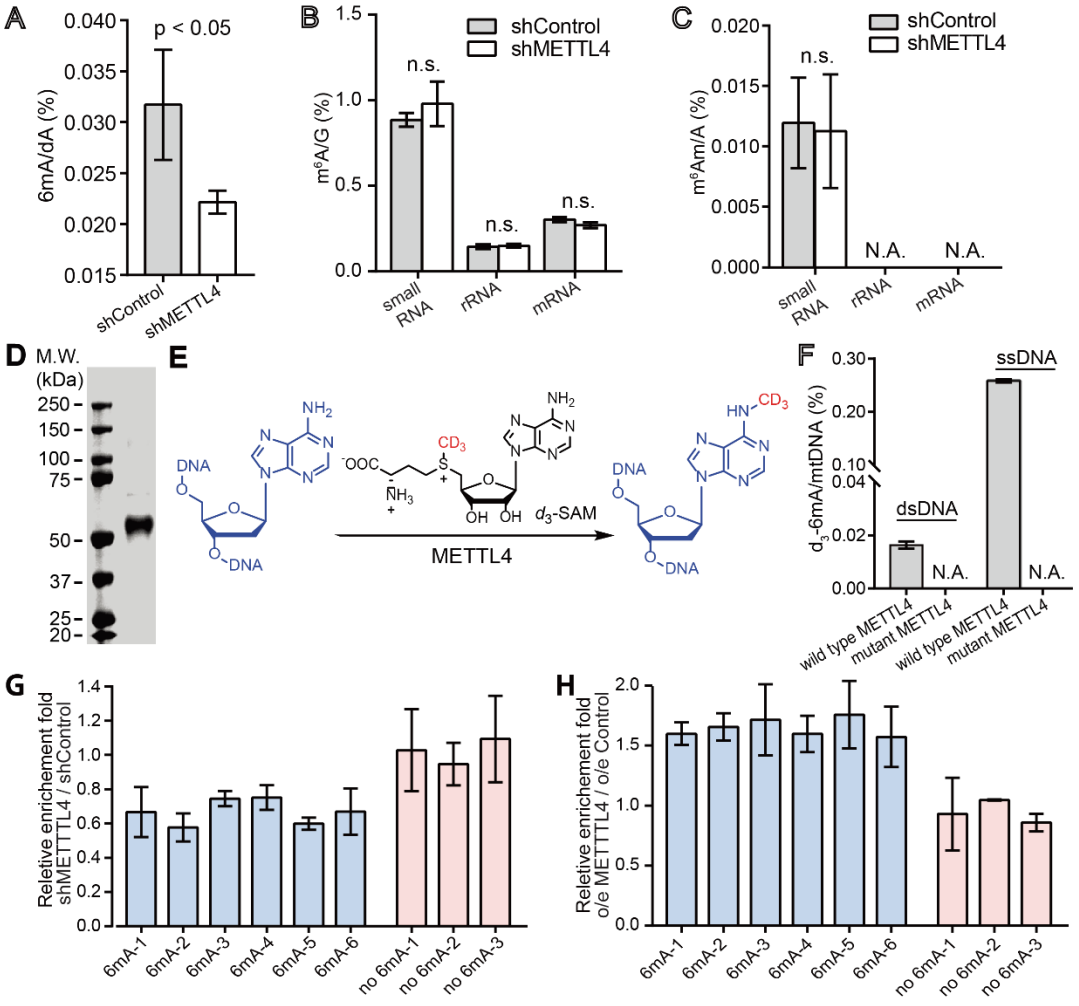

Figure 5.5 METTL4 mediates mtDNA 6mA methylation

(**A**) The knockdown of METTL4 resulted in a decrease in the total 6mA level in HepG2 mtDNA (n = 4, $p < 0.05$, t-test, mean ± SEM). (**B**) Knockdown METTL4 did not reduce the m^6A level in mtRNA species (n = 4, mean ± SEM). (**C**) METTL4 knockdown did not reduce the m^6Am level in mtRNA species (n = 4, mean ± SEM). (**D**) SDS-PAGE showed the purity of the recombinant

(**Figure 5.5, continued**) METTL4 protein. (**E**) A schematic illustration of N^6-deoxyadenosine methylation of DNA in the presence of METTL4 and isotope-labeled cofactor *S*-(5'-Adenosyl)-L-methionine-d_3 (d_3-SAM)**.** d_3-SAM was used to achieve accurate mass spectrometry quantification instead of typical SAM to exclude potential contaminations during the protein purification step. (**F**) The *in vitro* methylation activity of METTL4 and its inactive mutant on single-stranded and double-stranded mtDNA (n = 2, mean ± SEM). (**G**) 6mA-IP qPCR of 6mA-positive and 6mA-negative sites revealed by 6mA mapping in stable METTL4 knockdown and control HepG2 cells (n = 2, mean ± SEM). (**H**) 6mA-IP qPCR of 6mA-positive and 6mA-negative sites revealed by 6mA mapping in METTL4 overexpression and control HepG2 cells (n = 2, mean ± SEM).

We further performed *in vitro* biochemical studies to confirm the DNA methylation activity of METTL4. We cloned, expressed, and purified full-length human METTL4 (~53 kDa, Figure 5.5D) and a mutant METTL4 from HEK293T cells with the conserved methylation signature motif DPPW mutated to APPA. We incubated recombinant METTL4 proteins with isolated mtDNA in the presence of *S*-(5'-adenosyl)-L-methionine-d_3 (d_3-SAM), a deuterium-substituted methyl donor cofactor, digested the mtDNA and quantified the formation of d_3-6mA using UHPLC-QQQ-MS/MS. We detected 6mA methyltransferase activity on both single-stranded and double-stranded mtDNA using wild-type METTL4 but not for the mutant (Figure 5.5E-F), with the activity on single-stranded DNA much higher.

We then performed 6mA-IP by using METTL4 knockdown and METTL4 overexpression cell lines, respectively, and performed qPCR to check the enrichment efficiency of the 6mA-containing and 6mA-negative sites mentioned above. The enrichment fold of 6mA-containing sites decreased by ~30 to 40% in METTL4 knockdown cells compared with control cells. These sites were ~1.5-fold more enriched in METTL4 overexpressed cells, with no such change observed on 6mA-negative sites (Figure 5.5G-H). Taken together, we concluded that METTL4 could act as a methyltransferase to perform the 6mA methylation on mtDNA in mammalian cells.

5.2.5 METTL4 Affects Mitochondrial Activity

The identification of METTL4 as a 6mA methyltransferase allowed us to probe the roles of 6mA in mammalian mitochondria. Western blot showed no significant difference in SDHA level between wild-type and METTL4 knockdown cells (Figure 5.6A), indicating METTL4 knockdown does not change mitochondria number. We performed a cell energy phenotype test. Oxygen consumption rate (OCR), which reflects the mitochondrial respiration activity, and extracellular acidification rate (ECAR), which indicates the glycolytic activity, were measured under basal and stressed conditions, respectively. Compared to wild-type cells, the knockdown of METTL4 exhibited elevated baseline activity as well as increased mitochondrial respiration and glycolysis in response to mitochondrial stressors (Figure 5.6B). Thus, METTL4 appears to affect mitochondrial activity negatively under both basal and stressed (high-energy consuming) conditions.

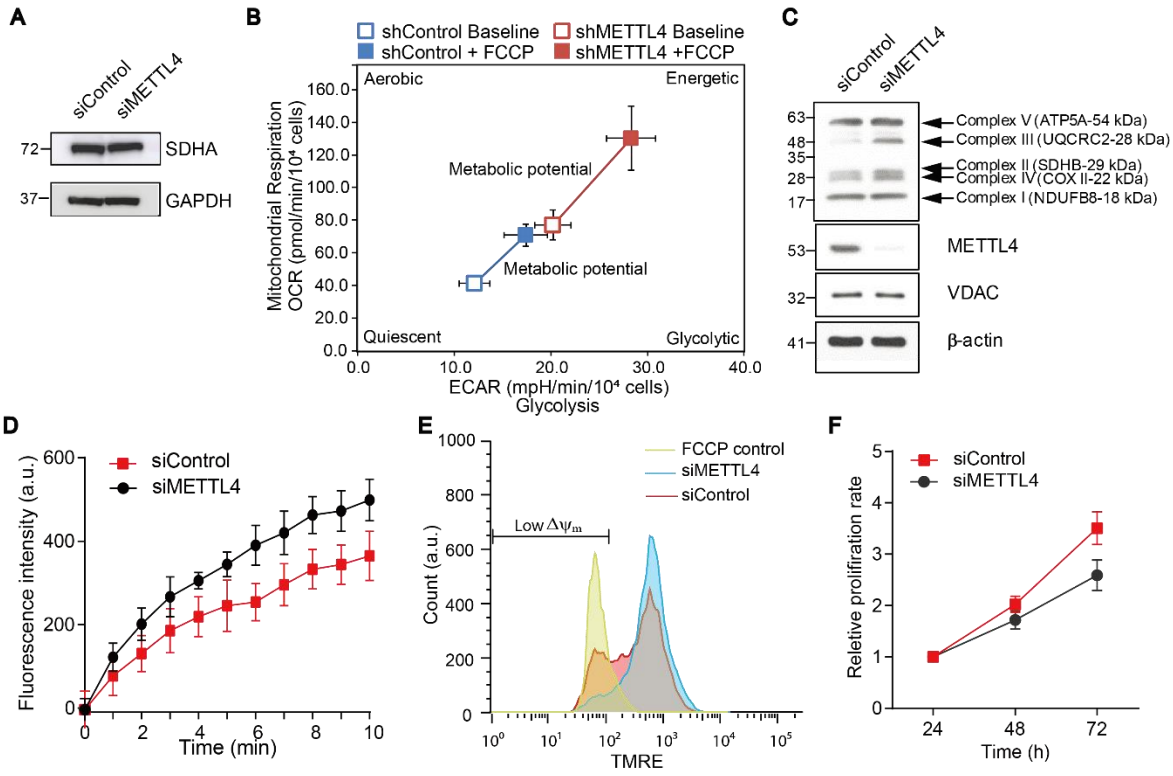

Figure 5.6 METTL4-mediated 6mA methylation affects mitochondrial functions

(**Figure 5.6, continued**) (**A**) SDHA level remained unchanged in METTL4 knockdown cells. (**B**) Metabolic phenotype plot of oxygen consumption rate (OCR) vs. extracellular acidification rate (ECAR). METTL4 knockdown increased mitochondrial activity under both basal and stressed conditions (n = 4, mean ± SEM). (**C**) METTL4 knockdown significantly increased the OXPHOS complex III assembly. The level of complex IV was also shown to increase slightly. (**D**) METTL4 knockdown increased ROS generation in mitochondria (n = 8, mean ± SEM). (**E**) METTL4 knockdown increased the overall membrane potential of mitochondria. (**F**) METTL4 knockdown decreased cell proliferation rates (n = 5, mean ± SEM).

To study the mechanism of increased mitochondrial activity upon METTL4 knockdown, we examined the levels of OXPHOS complex components in control and METTL4 knockdown cells. We observed an increase in the OXPHOS complex III component and a slight increase of complex IV component in METTL4 knockdown cells compared to the control (Figure 5.6C), which explains the elevated OCR observed upon METTL4 knockdown (Figure 5.6B).

As high mitochondrial OXPHOS activity may increase the level of cellular reactive oxygen species (ROS)[151,152], we also monitored the effect of METTL4 on mitochondrial ROS levels. A time-lapse of fluorescence revealed increased ROS levels in the METTL4 knockdown cells compared to the control (Figure 5.6D). We then analyzed the effect of METTL4 on mitochondrial membrane potential ($\Delta\psi m$) by flow cytometry. The knockdown of METTL4 induced the hyperpolarization of $\Delta\psi_m$ (Figure 5.6E), which is consistent with its effect on ROS generation and mitochondrial respiration[152]. The imbalanced mitochondrial activity induced by METTL4 knockdown leads to mitochondrial dysfunction and reduced cell proliferation (Figure 5.6F).

5.2.6 METTL4 modulates mitochondrial transcripts level and mtDNA level

To uncover the molecular mechanism underlying the role of METTL4 in mitochondrial function, we investigated the relationship between 6mA and mitochondrial gene expression. We performed RNA-seq and RT-qPCR under METTL4 knockdown and control conditions. We observed an upregulation of mitochondrial rRNAs and most mRNAs upon METTL4 knockdown (Figures 5.7A-B), which could contribute to increased expression of OXPHOS complex

components and upregulated mitochondrial activity. Both TFAM protein and mRNA level in the wild-type and METTL4 knockdown cells are similar (Figures 5.7C), indicating that the increased transcription in the METTL4 knockdown cells is not due to altered TFAM level.

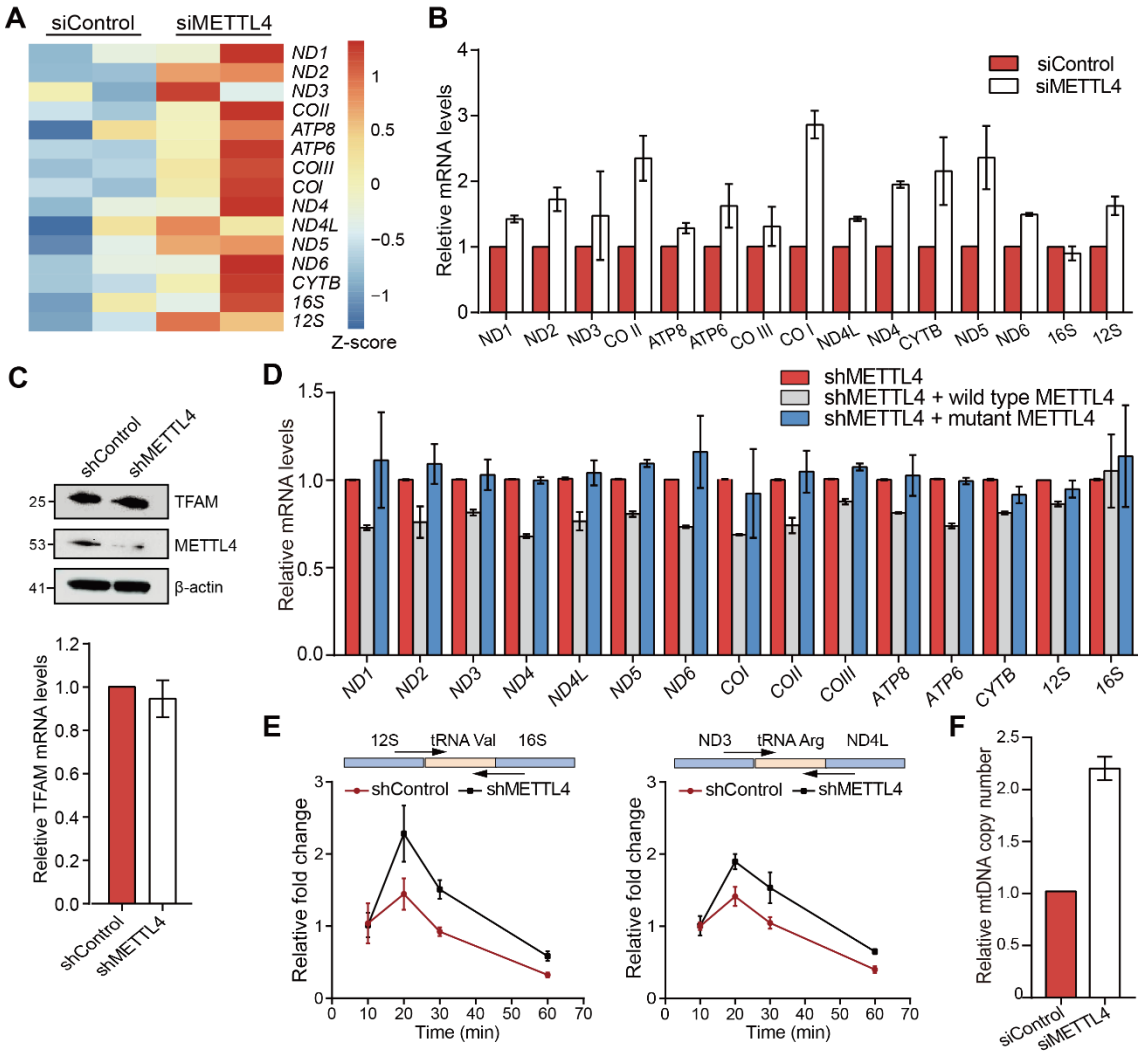

Figure 5.7 METTL4 and 6mA affect the levels of mtDNA and mitochondrial transcripts

(**A**) A heatmap summarizing the mtDNA encoded transcripts from RNA-seq analysis. METTL4 knockdown increased the expression level of most mtDNA encoded mRNAs and rRNA. (**B**) METTL4 knockdown increased the expression levels of most mitochondria-encoded transcripts (n = 2, mean ± SEM). (**C**) TFAM protein (top) and mRNA (bottom) levels in METTL4 knockdown and control cells. (**D**) METTL4 knockdown and complementation showed that the methylation activity of METTL4 contributes to mitochondrial gene expression regulation (n = 3, mean ± SEM). (**E**) Fold change of mitochondrial precursor RNA transcripts in the wild-type and METTL4 knockdown cells at different time points after EU labeling. Primers for qPCR were designed for mt-tRNA^{Val} (left) and mt-tRNA^{Arg} (right) junction sites (n = 4, mean ± SEM). METTL4

(**Figure 5.7, continued**) knockdown increases the formation rate of mitochondrial precursor RNA. (**F**) METTL4 knockdown doubled the relative mtDNA copy number (n = 3, mean ± SEM).

We further performed rescue experiments to validate the effect of 6mA on mitochondrial gene expression. In METTL4 stable knockdown cells, overexpression of the wild-type METTL4, but not of the inactive METTL4 mutant (D292A, W294A), decreased mitochondrial mRNA levels (Figures 5.7D).

We also analyzed levels of mitochondrial nascent polycistronic precursor RNAs instead of steady-state RNAs. Both the control and METTL4 knockdown HepG2 cells were labeled with 5'-ethynyl-uridine (5-EU) for 10, 20, 30, and 60 min, respectively. 5-EU-labeled RNAs were enriched for RT-qPCR analysis with primers designed to span the mitochondrial splice junctions, ensuring that only polycistronic precursor RNAs can be amplified. The normalized time curve showed a faster accumulation rate of polycistronic transcripts upon METTL4 knockdown (Figures 5.7E), indicating that METTL4 is directly involved in mtDNA transcription regulation. We further discovered that the mtDNA copy number was elevated by ~2-fold in METTL4 knockdown cells relative to control cells (Figures 5.7F). Therefore, we hypothesized that the presence of mtDNA 6mA could affect mitochondrial transcription and replication, which are coupled in mitochondria[153,154].

5.2.7 6mA suppresses mitochondrial transcription in vitro

Based on the 6mA-mapping results, we observed a notable 6mA peak locating at the promoter region (HSP and LSP) (Figure 5.2C-D), which was further validated by RT-qPCR (Figure 5.2G). We performed an in vitro transcription assay to investigate the effect of 6mA at the promotor region on transcription initiation. We used linear DNAs which include the HSP or LSP sequences, with or without one 6mA modification inserted at a specific site (Figure 5.8A). The reaction was initiated by assembling the templates with transcription complex components,

including the mitochondrial RNA polymerase (POLRMT), the mitochondrial transcription factor B2 (TFB2M), and the mitochondrial transcription factor A (TFAM). No visible transcript was observed when TFAM was not added (Figure 5.8B), confirming the specificity of the *in vitro* transcription assay. 6mA modifications at the heavy-strand promoter (HSP) were found to attenuate transcription by ~60% (Figure 5.8B-C). Transcripts generated from the light-strand promoter (LSP) were also decreased by ~30% in the 6mA-containing template (Figure 5.8B-C). These results support a transcription suppression role of 6mA at the promoter region.

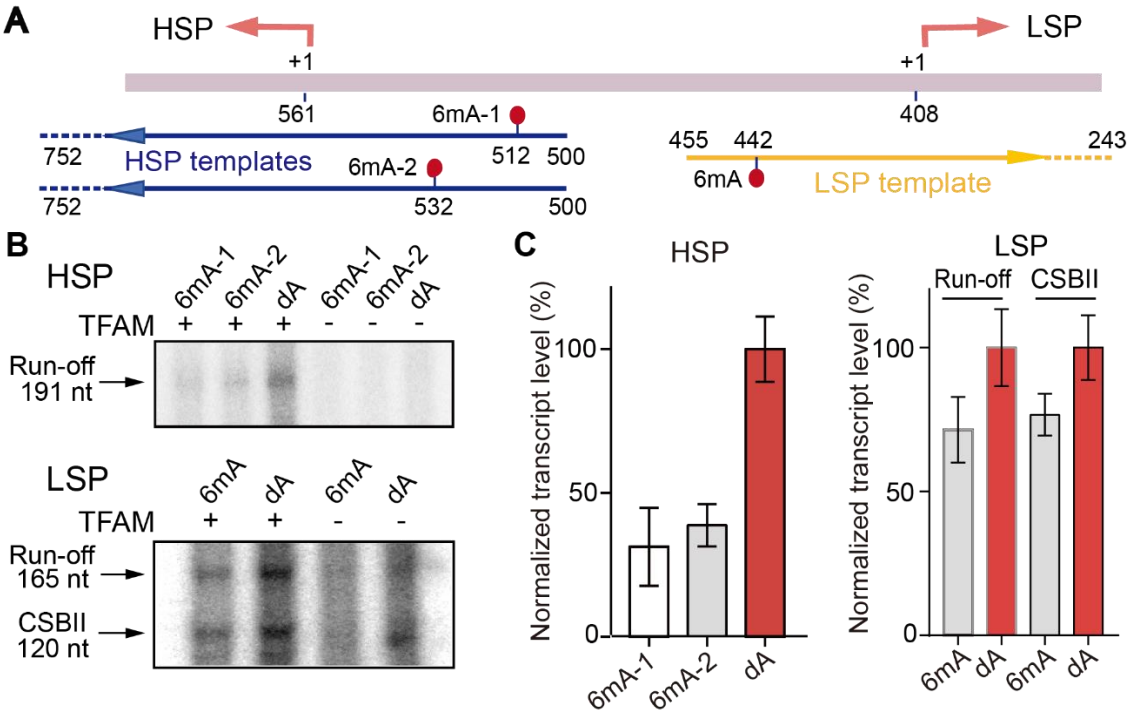

Figure 5.8 DNA 6mA methylation attenuates mitochondrial transcription in vitro

(**A**) Scheme of the mtDNA transcriptional control region showing the HSP and LSP. The HSP and LSP templates used for *in vitro* transcription were denoted below as blue and yellow lines, respectively. One single 6mA modification was inserted into the HSP and LSP templates. (**B**) Autoradiography imaging of the products from *in vitro* transcription reaction using HSP (top) and LSP (bottom) templates. LSP templates generate a 165-nt full-length transcript and a 120-nt truncated transcript, which corresponds to termination at the CSB II site. (**C**) Quantification of the amount of transcript generated in the in-vitro transcription assay by using the images shown in (**B**). The levels of transcripts produced from the 6mA-modified probes were normalized to that from the unmodified probes (n = 2, mean ± s.d.).

5.2.8 6mA affects DNA binding and bending by TFAM

Previous studies revealed that 6mA methylation could affect DNA bending[155,156] and DNA-protein recognition[157]. The critical mitochondrial transcription unit TFAM is known to bind and bend mtDNA, and it bends mtDNA differently at LSP and HSP, which is thought to result in different in vitro transcription outcomes[158,159].

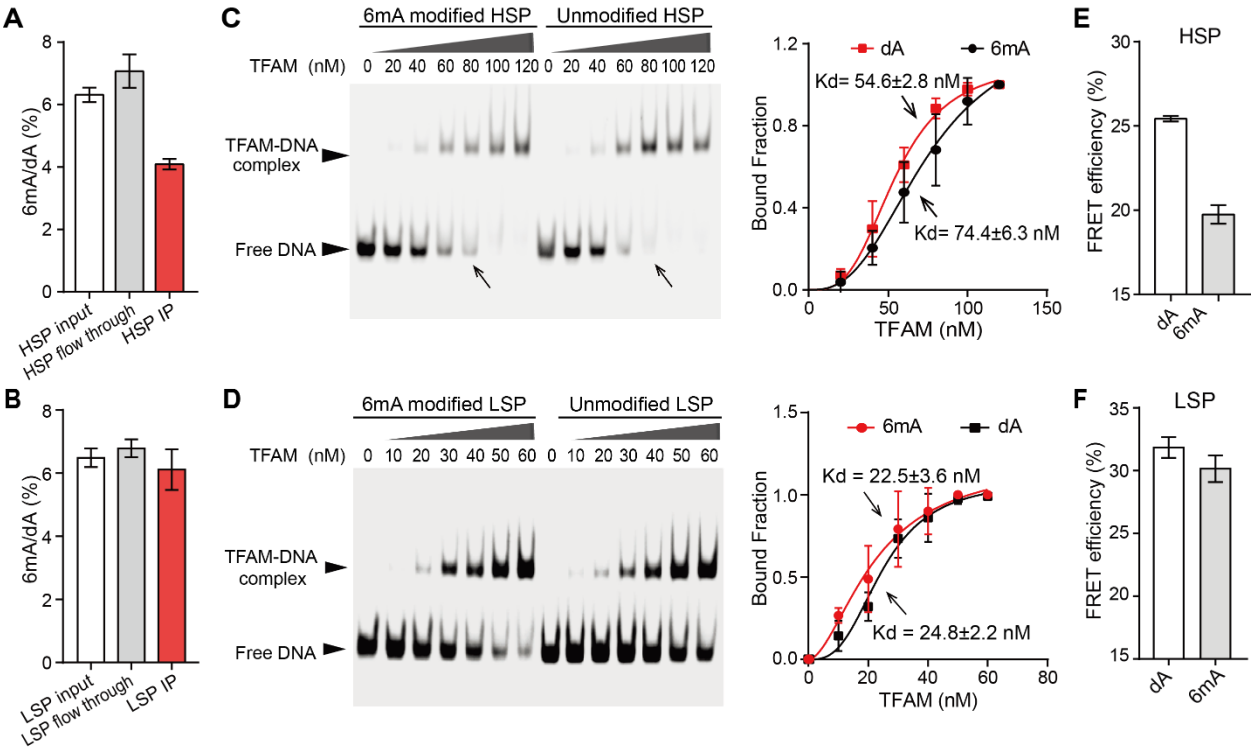

Figure 5.9 6mA affects DNA binding and bending by TFAM

(**A**)-(**B**) In vitro competition assays performed by using HSP (**A**) or LSP (**B**) probes. The ratio of 6mA/dA in the input, TFAM-IP, and the flow-through fractions were shown (n = 2, mean ± s.d.). (**C**)-(**D**) EMSA performed using the 6mA-modified and unmodified HSP (**C**) or LSP (**D**) probes. The dissociation constants of TFAM to the 6mA-modified and unmodified probes were calculated from the gel plot by using bound fraction as a function of the TFAM concentration (n = 2, mean ± s.d.). (**E**)-(**F**) FRET efficiency of the 6mA-modified and unmodified HSP (**E**) or LSP (**F**) probes upon TFAM binding (n = 2, mean ± s.d.).

To test the effect of 6mA on TFAM binding, we performed an in vitro pull-down by incubating purified TFAM protein with a mixture of 6mA-modified and unmodified HSP probes. Free and TFAM-bound DNAs were purified and subjected to LC-MS/MS analysis. 6mA/dA level

in the IP fraction was lower than that in input, indicating TFAM preferentially binds to unmodified HSP probe (Figure 5.9A) Consistently, EMSA assays also showed that the presence of 6mA decreased the binding affinity of TFAM to HSP DNA (Figure 5.9C). This attenuated binding effect was not detected when the LSP probe was applied (Figure 5.9B and D), which is in line with the in vitro transcription results.

To test if 6mA could interrupt the bending of DNA by TFAM, we employed a Förster resonance energy transfer (FRET) assay by using short HSP and LSP probes with Cy3 and Cy5 labeled respectively on two opposite ends. The addition of 200 nM TFAM (binding saturates at 100 nM, as shown by EMSA) led to the bending of DNA probes and strong FRET signal. Comparing with unmodified probes, the presence of 6mA in DNA decreases the FRET efficiency by ~22% and ~8% when using HSP and LSP probes, respectively (Figure 5.9E-F).

Taken together, 6mA attenuates TFAM binding and bending of its cognate DNA. Because TFAM is also known to bend DNA outside the promoter regions[160], this attenuation effect is not only relevant to transcription activation, but also mtDNA packaging and stability[161,162].

5.2.9 mtDNA 6mA level is significantly elevated under hypoxic stress

As the presence of 6mA could attenuate gene expression in mitochondria, we hypothesized that cells might keep the endogenous levels of 6mA low to maintain proper mitochondrial function. We suspected that the 6mA level in mtDNA could be elevated under certain stress conditions to control mitochondrial activity. After treating HepG2 cells to hypoxia for 24 and 48 hours, we discovered that the 6mA level in mtDNA gradually increased and reached to 0.1% 6mA/dA (~3 fold higher compared with that in normoxia) after 48 h treatment, which corresponds to ~12 or more 6mA modifications per mtDNA molecule (Figure 5.10A). RT-qPCR and western blot showed an upregulation of METTL4 transcripts and protein levels under hypoxia (Figure 5.10B).

Therefore, hypoxia could activate the expression of METTL4 within mitochondria to facilitate 6mA methylation. We exposed METTL4 knockdown cells and control cells to hypoxia for 48 h. The 6mA level was ~70% lower in METTL4 knockdown cells than that in control cells (Figure 5.10C), confirming an essential role of METTL4 in tuning 6mA level in mtDNA under hypoxia.

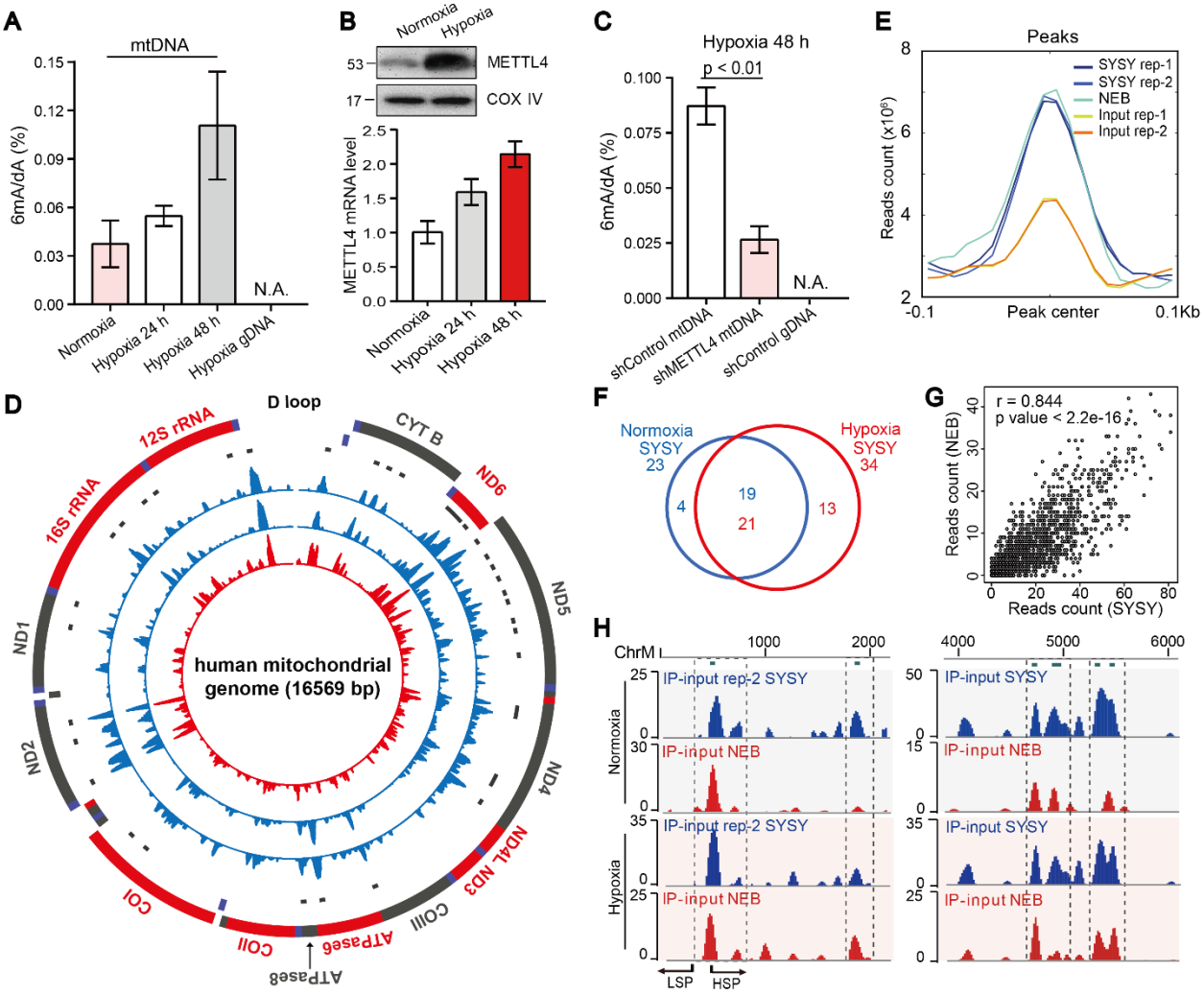

Figure 5.10 METTL4 and 6mA methylation accumulates in mtDNA under hypoxic stress

(**A**) The 6mA level in mtDNA increased during hypoxia treatment. Low 6mA was observed in gDNA under hypoxia (n = 3, p < 0.01, t-test, mean ± SEM). (**B**) The METTL4 mRNA and protein level in mitochondria was elevated under hypoxia. (**C**) The knockdown of METTL4 under hypoxia resulted in a substantially reduced 6mA level in mtDNA (n = 2, p < 0.01, t-test, mean ± SEM). (**D**) The distribution of 6mA in mtDNA under hypoxia revealed by 6mA mapping. The mitochondrial genome is shown as the outermost circle in red/gray/blue with gene annotations. 6mA profiles revealed by SYSY-IP rep-1 (blue), SYSY-IP rep-2 (blue), and NEB-IP (red) were shown in three

(**Figure 5.10, continued**) tracks, respectively, from outside to inside. Gray dots and bars indicate 6mA peaks. Reads from input libraries have been subtracted. (**E**) Metapeak profile shows the relative reads coverage of input and IP samples near identified 6mA peaks under hypoxia. (**F**) Peak overlap between 6mA mapping results under normoxia and hypoxia. (**G**) Spearman correlation analysis of 6mA peaks mapped by using NEB and SYSY anti-6mA antibody under hypoxia. Spearman r = 0.844. (**H**) Two examples showing 6mA peaks (IP – input) under normoxia and hypoxia overlap well. Peaks revealed by SYSY and NEB antibodies overlap well. Reads from input libraries have been subtracted. The left penal showed the peak detected in the promotor region.

We next performed 6mA-mapping in hypoxia-treated (48 h) cells and uncovered 34 potential 6mA sites, with 21 sites (62%) overlapping with 6mA sites under normoxia (Figure 5.10 D-F). Libraries generated by two different antibodies show a strong positive correlation and peak overlap (Figure 5.10G-H). Although many 6mA sites appear to be conserved, some 6mA sites change under hypoxia, suggesting potential dynamics of 6mA methylation under different physiological conditions to affect gene expression. The introduction of 6mA into mtDNA could be a mechanism for cells to cope with certain types of stresses, which will be important to explore in future studies.

5.3 Discussion and conclusion

Compared to 5mC, 6mA is prevalent in bacteria, playing critical roles in R-M systems, DNA replication, DNA repair, and transcription regulation. We discovered enriched 6mA in human mitochondria DNA and its function in repressing the mitochondrial gene expression. Mitochondria are thought to have bacterial ancestry[163], most likely from the α-proteobacteria, in which the CcrM family of DNA 6mA methyltransferases are widespread and play critical roles beyond restriction-modification[164]. Although a variety of pathways and mechanisms exist to tune mitochondrial functions, our current work adds adenosine-based DNA methylation as a regulatory mark in the mammalian mitochondrial genome. This could trace back to bacterial origins and is distinct from the predominant 5mC methylation in mammalian nuclear genomes.

Mechanistically, we concluded that the physiological levels of 6mA could elicit a notable inhibition effect on the binding or processing ability of key mitochondrial transcription machinery components. This inhibitory effect can happen directly at the promoter region, but can also occur along the gene body by affecting mtDNA compaction. More detailed future studies are required to elucidate the exact roles of 6mA.

We found METTL4 as a methyltransferase that can install 6mA in mtDNA, thereby repressing mitochondrial gene expression by reducing mtDNA copy number and attenuating transcription. We found neither obvious N-terminal MTS nor cleavable pre-sequence in METTL4. It will be interesting to elucidate how METTL4 translocates into mitochondria in the future.

A recent study reported that METTL4 acts as a methyltransferase to install internal m^6Am methylation on U2 small nuclear RNA (snRNA) in HEK293T cells, suggesting that METTL4 could play a nuclear role on RNA splicing regulation[165]. The function of METTL4 seems to be context dependent; we could not exclude that METTL4 has other cellular functions.

5.4 Methods

5.4.1 Cell culture

HeLa, HepG2, MDA-MB-231, and HEK293T cells were purchased from ATCC. HeLa cells were cultured in DMEM (Gibco 11965) supplemented with 10% (v/v) fetal bovine serum (Gibco), penicillin and streptomycin (Gibco) and grown at 37 ℃ with 5% CO_2. HepG2, HEK293T cells were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), penicillin and streptomycin (Gibco) and grown at 37 ℃ with 5% CO_2. 143B cells, Rho zero 143B cells are gifts from Dr. Y. H. Wei. Hypoxia treated cells were incubated at 37 ℃, 5% CO_2, and 2.3% O_2. Primary fibroblasts were isolated from postnatal 0-to 4-day-old wild-type mice by cutting skin tissue into small pieces in PBS with 100 IU/ml penicillin and streptomycin. For

each mouse, skin pieces were attached to one FBS-coated 10 cm dish and cultured in DMEM supplemented with 10% FBS, 1 mM L-glutamine, 100 IU/ml penicillin and streptomycin. All cell lines used in this study were examined by mycoplasma contamination test using LookOut Mycoplasma PCR Kit (Sigma, MP0035).

5.4.2 Plasmid construction

Recombinant METTL4 (Isoform 1, 472 aa) was cloned from commercial cDNA (Open Biosystems) into the pCDH-CMV-MCS-T2A copGFP vector (EcoRI and NotI) with flag-tag fused at the C-terminus of the protein. The human TFAM gene was synthesized with codon optimization (Genscript) and cloned into the pET28a expression vector (XhoI and BamHI). This construct encodes residues 43–246, which eliminates the N-terminal mitochondrial leader sequence (residues 1–42). Plasmids were prepared with the QIAGEN Plasmid Maxi kit (Qiagen).

5.4.3 Expression and purification of recombinant human METTL4 protein

Plasmid harboring METTL4 was transfected into HEK293T cells, which were harvested after 48 h. To purify METTL4 protein, cell pellets were suspended in 2 volumes of lysis buffer (150 mM KCl, 10 mM HEPES pH 7.6, 2 mM EDTA, 0.5% NP-40, 0.5 mM DTT, protease inhibitor cocktail) and rotated at 4 ℃ for 30 min. The lysate was cleared by centrifuging at 13,000 rpm for 30 min at 4 ℃, followed by passing through a 0.22 μm syringe filter. The lysate was then mixed with anti-Flag M2 magnetic beads (Sigma-Aldrich) and rotated at 4 ℃ for 3 h. The beads-protein complex was then washed with ice-cold wash buffer (200 mM NaCl, 50 mM HEPES pH 7.6, 2 mM EDTA, 0.05% NP-40, 0.5 mM DTT, protease inhibitor cocktail). METTL4 protein was eluted by incubating beads in 300 μL elution buffer (200 mM KCl, 50mM Tris-HCl, pH 8.0, 1.5 mM $MgCl_2$, 0.5 mM DTT, proteinase inhibitor cocktail, 0.5 mg/mL 3× Flag peptide (Genscript)) for

30 min at 4 ℃ with gentle rotation. The protein was washed and concentrated by using the Amicon Ultra 10K centrifugal filter Device (Millipore).

5.4.4 Expression and purification of TFAM protein

Plasmid harboring the TFAM gene was transformed into BL21 (DE3) Escherichia coli (Invitrogen). A single colony was amplified in 20 mL LB medium containing 50 μg/mL kanamycin and was then diluted into 4 L LB medium. When OD600 reached 0.6, 1 mM isopropyl β-d-1-thiogalactopyranoside was added to the culture, which was subsequently grown overnight at 25 ℃. Bacteria were harvested and lysed in 60 mL lysis buffer (20 mM Tris-HCl, 500 mM NaCl, pH 7.5) by sonication for 15 min (10 s on and 20 s off) on ice. His-tagged TFAM was purified from the cleared lysate with the Histrap HP column (GE healthcare). The protein was eluted in elution buffer (20 mM Tris-HCl, 500 mM NaCl, 300 mM imidazole, pH 7.5) and further purified by gel filtration chromatography using a Hi-Load Superdex 200 16/60 column (GE Healthcare) in an FPLC (GE healthcare). The peak fraction was collected and concentrated using Amicon Ultra 10K centrifugal filter Device (Millipore).

5.4.5 Mitochondria isolation and DNA extraction

Cells were harvested and washed once in isolation buffer (225 mM mannitol, 75 mM sucrose, 20 mM MOPS pH 7.2, 1 mM EGTA, 0.1% BSA). Cell pellets were then resuspended and incubated in lysis buffer (100 mM sucrose, 10 mM MOPS pH 7.2, 1 mM EGTA, 0.1% BSA) for 5 min at 4 ℃, and then homogenized with Omni homogenizer for 1 min at low speed. One volume of 1.25 M sucrose was then added, and the mixture was centrifuged at 1,000 g for 10 min at 4 ℃. The supernatant was transferred to a new tube, which was subjected to centrifuge at 12,000 g for 5 min. The mitochondria-containing pellet was further purified by the Mitochondria Isolation kit (Miltenyi Biotec) according to the manufacturer's protocol. The eluted mitochondria were washed

with high salt buffer (10 mM Tris-HCl pH 7.6, 10 mM KCl, 10 mM $MgCl_2$, 0.4 M NaCl and 2 mM EDTA) and resuspended in plasmid-safe reaction buffer (33 mM Tris-acetate pH 7.5, 66 mM potassium acetate, 10 mM magnesium acetate, 1 mM DTT). The mixture was vigorously vortexed and passed through 27 ½ G needle for 2-3 strokes. Plasmid-safe ATP-dependent DNase (Epicentre) was then added into the mixture to digested contaminated nuclear DNA according to the manufacturer's protocol. After the reaction, the mtDNA was purified by the DNA Clean & Concentrator kit (Zymo).

5.4.6 Immunofluorescence microscopy

Around 2×10^4 HepG2, 143B, or HeLa cells were seeded in the Lab-Tec 8-well chamber (Thermo) and grown overnight. The next day, cells were stained with 300 nM MitoTracker Deep Red FM (Thermo Fisher) in DMEM medium for 30 min at 37 ℃. Cells were washed with cold PBS, fixed in 4% paraformaldehyde, and permeabilized by 0.3% Triton X-100 dissolved in PBS at room temperature. For 6mA imaging, the permeabilized cells were washed with PBS and then blocked for 1 h in blocking buffer (1% BSA in PBS containing 0.1% Tween) containing either 50 μg/mL RNase A (Invitrogen) or 50 μg/mL RNase A plus 2 U/μL DNase I (Invitrogen). Cells were then washed and incubated in 6mA antibody (1: 1,000, Synaptic Systems) at 4 ℃ overnight followed by incubation in the secondary antibody (goat anti-rabbit IgG Alexa Fluor 488, Thermo Fisher) at a 1:2000 dilution. For METTL4 staining, nuclease-treatment was omitted, and cells were incubated with anti-METTL4 antibody (Sigma, HPA040061, 1 μg/mL) instead of 6mA antibody. All cells were stained with 1:1000 diluted DAPI (Thermo) before subjected to Leica SP5 II laser scanning confocal microscopy. The images were processed using ImageJ.

5.4.7 Western Blot

Cells were lysed, and protein concentration in the cleared lysate was measured by using Quick Start Bradford Protein Assay Kit (Bio-Rad). Equal amounts of protein (20-30 μg) were subjected to SDS-PAGE and transferred to nitrocellulose membrane (Bio-Rad). The membranes were then blocked and subject to overnight incubation with primary antibodies in 5% milk in PBST (0.1% Tween-20) at 4 ℃. Membranes were then washed and incubated with secondary antibodies for 1 h at room temperature. Membranes were washed before applied with ECL and developed.

5.4.8 Dot blot

Samples were heated at 95 ℃ for 10 min and then put on ice immediately before loading on membranes. 50 ng, 100 ng, 150 ng, or 200 ng DNA were loaded on Hybond N+ membranes (GE Healthcare Life Sciences). Membranes were air-dried and were crosslinked by UV stratalinker 2400 at 150 mJ/cm^2 twice. 6mA antibody (Synaptic Systems, 1: 1000) and HRP-conjugated anti-rabbit IgG (Cell Signaling) were then sequentially applied to the blocked membrane as did for Western blot. Membranes were washed before applied with ECL and developed.

5.4.9 RNA extraction and quantitative RT-qPCR

Total RNA was isolated by using the RNeasy Plus Mini kit (Qiagen). 750 ng total RNA were reverse transcribed into cDNA with PrimeScript™ RT reagent Kit (Takara), and then subjected to qPCR analysis with FastStart SYBR Green Master Mix (Roche) in a Roche LightCycler 96. Primers for qPCR are listed as the following.

METTL4: F: GGGAAGAGGCTACTTTGTCCT; R: GCCAAAGGGGGTTAAAACCTG

ND1: F: ATACCCCCGATTCCGCTACGAC; R: GTTTGAGGGGGAATGCTGGAGA

ND2: F: GGCCTAGAAATAAACATGCTA; R: GGGCTATTCCTAGTTTTATT

ND3: F: GCGGCTTCGACCCTATATCC; R: AGGGCTCATGGTAGGGGTAA

ND4: F: ACAAGCTCCATCTGCCTACG; R: GAAGCTTCAGGGGGTTTGGA

ND4L: F: ACTCCCACTAATAGCTTTTTGATG; R: AGGGCTGTGACTAGTATGTTGAG

ND5: F: CAAAACCTGCCCCTACTCCT; R: GGGTTGAGGTGATGATGGAG

ND6: F: ACACTCACCAAGACCTCAACC; R: TAGTTTTTTTAATTTATTTAGGGGGAAT

COI: F: CGATGCATACACCACATGAA; R: AGCGAAGGCTTCTCAAATCA

COII: F: GCTGTCCCCACATTAGGCTT; R: ACCGTAGTATACCCCCGGTC

COIII: F: AAAAGGCCTTCGATACGGGA; R: ATTTAGCGGGGTGATGCCTG

ATP8: F: CCACCTACCTCCCTCACCAAA; R: GATTGTGGGGGCAATGAATGA

ATP6: F: CGCCACCCTAGCAATATCAA; R: TTAAGGCGACAGCGATTTCT

CYTB: F: AATTCTCCGATCCGTCCCTA; R: GGAGGATGGGGATTATTGCT

GAPDH: F: GTCTCCTCTGACTTCAACAGCG; R: ACCACCCTGTTGCTGTAGCCAA

B2M: F: TGCTGTCTCCATGTTTGATGTATCT; R: TCTCTGCTCCCCACCTCTAAGT

ACTB: F: CTGGAACGGTGAAGGTGACA; R: AAGGGACTTCCTGTAACAATGCA

12S: F: GGTTGGTCAATTTCGTGCCAGC; R: GGGGTGATCTAAAACACTCTTTACGC

16S: F: AGACTTCACCAGTCAAAGCGA; R: ACATCGAGGTCGTAAACCCT

mt-tRNA^{Val}: F: CTTGGACGAACCAGAGTGTAG; R: GCTAGGTTTAGCTCAGAGCGGT

mt-tRNA^{Arg}: F: GGATTAGACTGAGCCGAATTGGTA; R: TGTAAATGAGGGGCATTTGGTAAAT

5.4.10 mtDNA copy number quantification

mtDNA copy number was calculated based on the RT-qPCR method by using primers targeting mtDNA with genomic DNA as control. Total DNA was extracted from cells by using the PureLink Genomic DNA kit (Invitrogen) and was subjected to qPCR analysis with FastStart SYBR Green Master Mix (Roche) in a Roche LightCycler 96. Primers for qPCR are listed as the following.

mtDNA: F: CACCCAAGAACAGGGTTTGT; R: TGGCCATGGGTATGTTGTTA

gDNA: F: CGAGTAAGAGACCATTGTGGCAG; R: GCACTGGCTTAGGAGTTGGACT

5.4.11 siRNA knockdown and plasmid transfection

METTL4 siRNA was purchased (targeting sequence: AAGCCCTACGAAGGTATTATA, Qiagen SI04136671) and transcribed by using Lipofectamine RNAiMAX (Invitrogen). Plasmids were transcribed by using Lipofectamine 2000 (Invitrogen). Knockdown efficiency was checked by using RT-qPCR and/or Western blot.

5.4.12 Lentivirus siRNA experiments

Lentivirus containing short hairpin RNAs (shRNAs) expressed in the pLKO.1-puro vector were generated in HEK293T cells as previously described. These plasmids and the packaging plasmid pCMVΔR8.91 were provided by the National RNAi Core Facility of Academia Sinica (Taipei, Taiwan). HEK293T cells were transfected with 5 μg lentiviral vectors expressing individual shRNA along with 0.5 μg envelope plasmid pMD.G and 5 μg packaging plasmid pCMVΔR8.91. The virus was collected 48 h after transfection. To prepare METTL4 knockdown cells, cells were infected with lentivirus for 24 h. Stable clones were then generated by selection with corresponding antibiotics.

5.4.13 RNA-seq

mRNA was isolated from total RNA by using Dynabeads mRNA DIRECT kit (Ambion). For each sample, 50 ng mRNA was used for library construction by using the TruSeq Stranded mRNA sample preparation kit (Illumina). Libraries were sequenced on Illumina Hiseq 4000 platform.

5.4.14 6mA-ChIP-exo

6mA ChIP-exo 5.0 was performed according to the previously reported procedure[145]. Briefly, 1 μg mtDNA was sonicated to 200-400 bp and immunoprecipitated using anti-6mA antibodies (Synaptic Systems, 202003 or NEB, E1610S) at 4 ℃ overnight. The antibody-DNA

complex was then split into a 96-well plate with 50 μL per well, irradiated by UV 254 nm with 0.15 mJ/cm^2 energy for 5 times. The crosslinked samples were then incubated with 80 μL pre-blocked Dynabeads Protein A (for SYSY) or Protein G (for NEB) slurry for 2 h at 4 ℃. A tailing, the first adapter ligation, kinase reaction, fill-in reaction, and lambda exonuclease digestion were performed on beads as previously described. DNA was then released from beads, and the purified DNA was subjected to splint ligation. The ligation product was amplified by PCR, and the libraries were subjected to high-throughput sequencing by using Illumina Nextseq500.

5.4.15 Quantification of 6mA in DNA and m^{6}A in RNA by UHPLC–QQQ–MS/MS

Denatured gDNA (150 ng) or mtDNA (50 ng) in 20 mM NH_4OAc (pH 5.3) was digested using 1 μL nuclease P1 (1 U/μl, Wako USA, 145-08221) at 42 ℃ overnight. 3 μL 1 M NH_4HCO_3 solution and 1 μL of phosphodiesterase I (0.001 U, Sigma, P3243-1VL) were then added. The mixture was incubated at 37 °C for 2 h. 1 U FastAP Thermosensitive Alkaline Phosphatase and 3 μL 10 × FastAP Buffer (Thermo, EF0651) were then added, and the mixture was incubated at 37 °C for 4 h. For RNA m^{6}A quantification, 50 ng RNA was digested by using 1 U nuclease P1 (Wako USA, 145-08221) in 20 μL reaction containing 20 mM NH_4OAc at 42 ℃ for 2 h. 1 μL FastAP Thermosensitive Alkaline Phosphatase and 2.3 μL 10 × FastAP Buffer (Thermo, EF0651) were then added. The reaction was incubated at 37 ℃ for 2 h. Digested DNA or RNA was diluted to 60 μL, with 10 μL sample used for each injection. Nucleosides were separated by reverse-phase C18 column (Agilent, 927,700-092) followed by MS detection using Agilent 6460 QQQ–MS/MS. The mass of precursor ion and product ion are 266.1 and 150.0 for 6mA, 252.1 and 136.0 for dA, 282.1 and 150.1 for m^{6}A, 268.1 and 136.1 for A. The concentration of nucleosides was quantified using the calibration curves generate from nucleoside standards running at the same condition. The final

6mA/dA and m^6A/A ratios were calculated by using the calibration curve after subtracting the background (mock control) derived from digestion enzymes.

5.4.16 In vitro DNA methylation assay

In a 50 μL volume, 150 ng ds-mtDNA or ss-mtDNA, 10 μM METTL4 protein and 0.8 mM d_3-SAM were mixed in reaction buffer (80 mM KCl, 1.5 mM $MgCl_2$, 5 mM DTT, 4% glycerol and 15 mM HEPES pH 7.9). The reaction was performed at 16 ℃ for 14 h. After incubation, the DNA was purified by DNA Clean & Concentrator kit (Zymo Research) before digested for quantification.

5.4.17 Inner mitochondrial membrane potential (ΔΨm) detection

Control cells and METTL4 knockdown cells were stained with 300 nM TMRE at 37 ℃ for 15 min. 20 μM carbonyl cyanide p-(trifluoromethoxy) phenylhydrazone (FCCP) was added to cell culture media 10 minutes before staining with TMRE as a negative control. After TMRE staining, cells were incubated in PBS containing 0.2% BSA on ice in the dark. The fluorescent signal of TMRE was immediately measured via flow cytometry LSR-Fortessa 4-15 HTS (BD digital instrument) with 10,000 ungated cells acquired. TMRE is excited by the 561 nm yellow laser equipped in LSR-Fortessa, and the emission at 610 ± 20 nm was detected. Data was analyzed by FlowJo software version 10.0.8.

5.4.18 Mitochondria ROS detection

ROS production was measured using MitoTracker Red CM-H_2XROS (Molecular Probes). METTL4 knockdown cell and control cells were suspended in 1 × HBSS buffer in 96-well white polystyrene microplates (Thermo). The time-course changes of fluorescence were recorded upon the addition of 2 μM MitoTracker Red CM-H2XROS using the Synergy HT (Bioteck, Winooski, VT) plate reader, with excitation at 560 ± 10 nm, and emission at 620 ± 20 nm.

5.4.19 Cell proliferation assay

3,000 METTL4 knockdown and control cells were seeded in each well in a 96-well plate. The cells were incubated in 5% CO_2 at 37 ℃ for 24 h, 48 h, or 72 h after transfection. At each time point, cell numbers were measured by adding CellTiter 96 Aqueous One reagent (Promega) followed by incubation at 37 ℃ for 2 h. The absorbance intensity at 490 nm was then measured using Synergy HT (Bioteck, Winooski, VT) plate reader. The signals were normalized to the value observed at the 24 h time point.

5.4.20 Cellular bioenergetics analysis using XFe96 extracellular flux analyzer

The metabolic phenotype was measured with a Seahorse Bioscience XF96 analyzer (Seahorse Bioscience Inc.) by using Seahorse XFp Cell Energy Phenotype Test Kit (Agilent). Briefly, 2×10^4 control or METTL4 knockdown cells were plated in Seahorse XF96 cell culture microplates (Agilent). Cells were grown overnight and incubated with XFp media for 1 h in a non-CO_2 incubator before measurement. The 96-well sensor cartridge was hydrated in 200 μL water overnight before an 1 h incubation in XF calibrant solution (Seahorse Bioscience Inc.) at 37 ℃. During the measurement, cells were treated with 1 μM of oligomycin followed by three serial injections of FCCP to achieve the final concentration of 0.5, 1, and 2 μM. Seahorse Wave software (version 2.6) was used for data processing to remove outliers and calibrate cell numbers. The final reports for the OCR and ECAR signals were generated using the Seahorse XF Cell Energy Phenotype report generator.

5.4.21 In vitro transcription

Recombinant POLRMT (134kDa), human TFAM (26.6 kDa, aa 43–246), and TFB2M (45.8 kDa, aa 20–396) were purchased from Enzymax. Templates were generated by ligating DNA fragment 1 with fragment 2 using T4 DNA ligase. The ligation products were purified and size-

selected twice by using AMPure XP beads (Beckman Coulter). Only DNA longer than 200 bp was retained. The transcription reactions were assembled as described previously with 10 mM HEPES pH 7.5, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 0.1 μg/uL BSA, 0.4 mM ATP, 0.4 mM GTP, 0.4 mM CTP, 10 μM UTP, 0.3 μCi [α-32P] UTP (800 Ci/mmol), 100 nM DNA template, 500 nM TFAM, 200 nM TFB2M, 150 nM POLRMT and 4 U/μL RNaseOUT (Thermo) Inhibitor. Reactions were carried out at 32 ℃ for 45 min and were quenched by adding 10 μL TBE-Urea gel loading buffer (Thermo Fisher) supplemented with formaldehyde (80%) and 10 mM EDTA. The run-off products were resolved by using 6 % Novex TBE-Urea gel at 120 V for 1 h and visualized by autoradiography (Amersham Typhoon 5 Biomolecular Imager). Quantification was performed using GelQuant software based on the results from two independent experiments.

The sequences of the DNA fragments were listed as the following.

HSP fragment 1: F: CCCATCCTACCCA(or 6mA)GCACACACACACCGCTGCTAA(or 6mA)-CCCCATACCCCGAACCAAC; R: /5Phos/GTTCGGGGTATGGGGTTAGCAGCGGTGTGT-GTGTGCTGGGTAGGATGGG

HSP fragment 2: F: /5Phos/CAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCT-CCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACCCCATAAACAAA TAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCC CGTTCCAGTGAGTTCACCCTCTAAATCACCACGATC; R: GATCGTGGTGATTTAGAG-GGTGAACTCACTGGAACGGGGATGCTTGCATGTGTAATCTTACTAAGAGCTAATAGA AAGGCTAGGACCAAACCTATTTGTTTATGGGGTGATGTGAGCCCGTCTAAACATTTT CAGTGTATTGCTTTGAGGAGGTAAGCTACATAAACTGTGGGGGGTGTCTTTGGGGTT TGGTTG

LSP fragment 1: F: AAAATAATGTGTTAGTTGGGGGGTGA(or 6mA)CTGTTAAAAGTGC-AT; R: /5Phos/CTTTTAACAGTCACCCCCCAACTAACACATTATTTT

LSP fragment 2: F: /5Phos/ACCGCCAAAAGATAAAATTTGAAATCTGGTTAGGCTGGTG-TTAGGGTTCTTTGTTTTTGGGGTTTGGCAGAGATGTGTTTAAGTGCTGTGGCCAGAA GCGGGGGAGGGGGGGTTTGGTGGAAATTTTTTGTTATGATGTCTGTGTGGAAAGTGG CTGTGCAGACATTCAATT; R: AATTGAATGTCTGCACAGCCACTTTCCACACAGACA-TCATAACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTA AACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACACCAGCCTAACCAGAT TTCAAATTTTATCTTTTGGCGGTATGCA

5.4.22 Evaluate the level of nascent mitochondrial polycistronic precursor RNA

5-ethynyl uridine (5-EU) labeling was performed using the Click-iT® Nascent RNA Capture Kit (Life Technologies). Control and METTL4 knockdown HepG2 cells were treated with 0.5 mM 5'-ethynyl-uridine (5-EU) for 10, 20, 30, and 60 min respectively before harvested. Total RNA was extracted as mentioned above and was subjected to biotinylation, purification, and pull-down according to manufacturer's protocol. Enriched RNA was reverse transcribed using the Recombinant HIV Reverse Transcriptase (Worthington Biochemical Corporation). RT-qPCR analysis was performed using primers that span the splice junctions of pre-mtRNA (listed in 5.4.9) to ensure that only transcribed polycistronic pre-RNAs can be amplified.

5.4.23 In vitro pull-down assay

In 20 μL volume, 1 μM 6mA modified probe, 1 μM unmodified probe, and 10 μM recombinant TFAM protein were mixed and incubated in TFAM binding buffer (50 mM KCl, 5 mM $MgCl_2$, 10% glycerol and 10 mM Tris-HCl, pH 7.8) at room temperature 20 min. 20 μL His-affinity magnetic beads (Invitrogen) were washed four times with 200 μL TFAM binding buffer.

The beads were then re-suspended in 180 μL TFAM binding buffer and were then incubated with the 20 μL binding reaction for 1 h at 4 °C with gentle rotation. “Flow-through” DNA was purified from the supernatant by ethanol precipitation. The beads were washed four times with 300 μL TFAM binding buffer, and the TFAM–DNA complex was eluted from beads by using 100 μL TFAM binding buffer supplemented with 300 mM imidazole. 30 μg proteinase K was added into the eluent to digest TFAM. DNA was then recovered from the eluent by ethanol precipitation as the “IP” fraction. Input, flow-through, and IP DNAs were subjected to mass spectrum analysis as described above. DNA probes used for pull-down are listed as the following.

LSP: F: TGTTA(or 6mA)GTTGGGGGGTGACTGTTAAAAGT; R: ACTTTTAACAGTCACC-CCCCAACTAACA

HSP: F: GGTTGGTTCGGGGTA(or 6mA)TGGGGTT; R: AACCCCATACCCCGAACCAACC

5.4.24 EMSA assay

DNA probes used for the EMSA assay are the same as those used for the in vitro pull-down (5.4.23). In a 20 μL volume, 20 nM annealed dsDNA was mixed with recombinant TFAM protein (10-60 nM for LSP, 20-120 nM for HSP) in the binding buffer (50 mM KCl, 5 mM $MgCl_2$, 10% glycerol and 10 mM Tris-HCl, pH 7.8). The binding reactions were incubated at room temperature for 10 min before separated on the 4-12% TBE gel (Thermo Fisher) with 0.5 × TB buffer (diluted from 5 × TB buffer: 50 mM Tris-HCl, 41.5 mM borate, pH 7.8) at 110 V for 1 h at 4 ℃. Gels were stained in SYBR Gold nucleic acid staining solution before visualized with UV light at 254 nm.

5.4.25 FRET assay

We use Cy3-Cy5 double-labeled DNA to measure FRET efficiency. We use DNAs labeled with a single Cy3 or Cy5 to obtain correction factors for FRET efficiency calculation. The FRET measurements were recorded on a FluoroMax-3 spectrofluorimeter (Jobin–Yvon–Horiba) at room

temperature. 200 nM TFAM protein was added into a cuvette containing 40 nM 6mA modified or unmodified FRET probes (Cy3-only, Cy5-only, or Cy3-Cy5-labeled DNA) in binding buffer (20 mM Tris pH 7.5, 150 mM NaCl, 1 mM DTT). Raw FRET signals were observed by measuring emission from 640 to 750 nm with excitation at 525 nm and denoted as F for Cy3-Cy5-labeled DNA, F_D for Cy3-only DNA, F_A for Cy5-only DNA. For Cy3-only DNA, emission from 550 to 580 nm was measured with excitation at 525 nm in the absence (Du) or presence (Db) of TFAM. For Cy5-only DNA, emission from 640 to 750 nm was measured with excitation at 625 nm in the absence (Au) or presence (Ab) of TFAM. Slit widths for excitation and emission were 4 nm and 10 nm, respectively.

The correction factors for donor bleed-through (χ_D) and acceptor bleed-through (χ_A) were calculated as: $\chi_D = F_D / D$, $\chi_A = F_A / A$. To account for fluorescence change caused by FRET-independent TFAM interaction, two additional correction factors were determined as: $\sigma D = Db / Du$, $\sigma A = Ab / Au$. Corrected FRET and donor florescence intensities were then calculated as: $Fcorr = [F - (\chi_D) - (\chi_A)] / \sigma_A$, $Dcorr = Du / \sigma_D$. The FRET efficiency (E) was calculated as: $E = Fcorr / (Fcorr + Dcorr)$. DNA probes used for the FRET assay were listed as the following.

FRET LSP probe: F: /Cy3/-TGTTA(or 6mA)GTTGGGGGGTGACTGTTAAAAGT; R: /Cy5/-ACTTTTAACAGTCACCCCCCAACTAACA

Cy3-only LSP probe: F: /Cy3/-TGTTA(or 6mA)GTTGGGGGGTGACTGTTAAAAGT; R: ACT-TTTAACAGTCACCCCCCAACTAACA

Cy5-only LSP probe: F: TGTTA(or 6mA)GTTGGGGGGTGACTGTTAAAAGT; R: /Cy5/-ACT-TTTAACAGTCACCCCCCAACTAACA

FRET HSP probe: F: /Cy3/-GGTTGGTTCGGGGTA(or 6mA)TGGGGTT; R: /Cy5/-AACCCC-ATACCCCGAACCAACC

Cy3-only HSP probe: F: /Cy3/-GGTTGGTTCGGGGTA(or 6mA)TGGGGTT; R: AACCCCAT-ACCCCGAACCAACC

Cy5-only HSP probe: F: GGTTGGTTCGGGGTA(or 6mA)TGGGGTT; R: /Cy5/-AACCCCAT-ACCCCGAACCAACC

5.4.26 RNA-seq data analysis

Raw single-end reads were first trimmed to remove adaptor sequences and low-quality nucleotides and were then aligned to the hg19 reference genome and transcriptome using Hisat[116]. Reads mapped to multiple loci in the genome or mapped to tRNA/rRNA regions were excluded from downstream analysis. RPKM values were calculated by Cuffnorm using the geometric library for normalization. Expression heatmaps were generated by the pheatmap R package with z-scores calculated for each row.

5.4.27 6mA ChIP-exo data analysis

Illumina sequencing reads were first trimmed to discard adapter sequences and low-quality nucleotides and then mapped to the hg19 reference genome using Bowtie. To avoid NUMTs-induced false positives, only reads that exclusively mapped to chrM were kept for downstream analysis. Normalized BigWig files were generated based on reads aligned to chrM only. For peak calling, reads were counted on consecutive 10-bp windows on mtDNA by FeatureCounts. Normalization was conducted using DESeq2. Considering the small size of the mitochondrial genome, a cutoff with log2(fold change) > 0.1 and normalized reads larger than 30 were applied.

5.4.28 Data and software availability

Raw and analyzed data for all sequencing experiments have been deposited at NCBI Gene Expression Omnibus under accession number GSE102670.

Chapter 6

Summary and perspectives

6.1 N_3-kethoxal: from chemical structures to biological applications

In this thesis, I have shown how a single small molecule facilitates the development of a series of biological tools. Conjugated with next-generation sequencing, in vivo N_3-kethoxal labeling enables the profiling of RNA secondary structure, transcription dynamics, and RNA-RNA interactions. From a chemical perspective, I keep asking myself how N_3-kethoxal is structurally unique, what is the relationship between its structure and its chemical properties, and how these chemical properties affect applications.

N_3-kethoxal, as well as other kethoxal derivatives, share an α-keto aldehyde structure, which is crucial for the reaction at N1 and N2 position of guanines. However, the reactivities of these α-keto aldehyde structures are not the same. For instance, glyoxal, the simplest compound which contains this structure, shows a lower reactivity than N_3-kethoxal in vitro (Figure 2.3). The nature of the kethoxal-guanine reaction is a two-step nucleophilic addition reaction between amines and carbonyls, which can be regulated by both electronic and steric effects. With more similar compounds synthesized recently, we realized the electronic effect, especially the oxygen atom at the α position of the carbonyl, is crucial for high reactivity. The high reactivity enables the fast and efficient labeling of DNA and RNA in live cells, as well as high crosslinking efficiency for RNA-RNA interaction applications.

However, high reactivity is not always ideal. For example, for chemical-based RNA secondary structure profiling, one modification per ~300 nt region (single-hit kinetics) is preferred[23], because the modification of the first nucleotide may induce conformational changes of other nucleotides, which may produce artifacts. Moreover, as N_3-kethoxal modification can

cause RT stop, dense modifications on nearby nucleotides may hamper reverse transcription and result in very short sequencing reads. The concentration of N_3-kethoxal used for Keth-seq can be further optimized, and other kethoxal derivatives with fine-tuned reactivity can be applied in the future.

For RNA secondary structure profiling, the behavior of N_3-kethoxal is similar to other RNA labeling probes such as DMS and NAI-N_3. However, two unique properties of N_3-kethoxal make it almost irreplaceable for ssDNA profiling. 1) N_3-kethoxal reacts specifically with single-stranded guanine bases, and the formation of Watson-Crick base pairs blocks the reaction. 2) N_3-kethoxal modification can be removed after DNA enrichment, and therefore, does not block PCR reaction. Both features are indispensable for the development of KAS-seq. Moreover, the specificity to unpaired guanines also makes KARR-seq a proximity-ligation based approach, which is distinguished from PARIS, a method which only captures RNA duplexes.

While the reversibility is required for KAS-seq, it can be detrimental for RNA secondary structure and RNA-RNA interaction profiling. The dissociation of kethoxal-guanine adduct during library preparation can reduce RT-stop signals; in other words, the sensitivity of Keth-seq. It may also result in the decomposition of RNA-RNA crosslinking products in KARR-seq. The dissociation is minimized under low temperature and borate fixation conditions during Keth-seq and KARR-seq, but developing similar probes with reduced reversibility can be beneficial for certain applications. The reversibility may be related to reactivity, more detailed study into their relationship is ongoing.

6.2 Expanding the spectrum of kethoxal derivatives

Chemically modifying DNA and RNA in vivo with desired functional groups provide great flexibility to study their structures and functions. However, only limited approaches were

developed. Our new synthetic strategy for N_3-kethoxal takes advantage of DMD oxidation, which is relatively moderate and show high functional group compatibility. Although my work presented in this thesis focus on the applications of N_3-kethoxal, kethoxal can serve as a platform that decorates DNA and RNA with different functional groups other than $–N_3$.

We are currently synthesizing a series of kethoxal derivatives and testing their activities. A number of them may have significant applications in biological studies. For instance, kethoxals conjugated with biotin can be used to simplify KAS-seq and Keth-seq procedures. Kethoxals with tetrazine or other click chemistry moieties provide flexibility for Keth-seq and KAS-seq when azide group is not compatible. Kethoxal-diazirine conjugates may enhance RNA-protein crosslinking; kethoxal with tetrazoles groups potentially enables light-controlled RNA labeling (Figure 6.1).

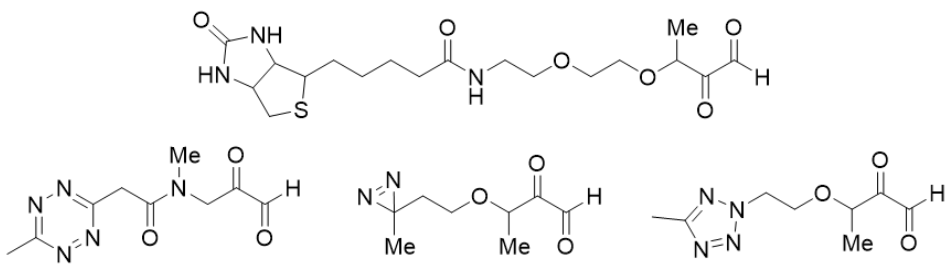

Figure 6.1 Examples of synthesized kethoxal derivatives with various functional groups

Proximity labeling of RNA by using ascorbate peroxidase (APEX-seq) produces a nanometer-resolution spatial map of the transcriptome[166]. In this method, genetically encoded APEX2 enzyme localize to subcellular regions of interest and generate biotin-phenol radicals nearby upon H_2O_2 treatment. The biotin-phenol radicals add a biotin tag to local RNA by attacking the 8- positions of guanines. As the radical only diffuse within a very short distance, transcripts that spatially close to the APEX2 enzyme can be enriched for sequencing.

APEX-seq can be potentially applied to many systems. However, the labeling efficiency on RNA is low, because most of the biotin-phenol radicals were trapped by tyrosine residues on proteins, which are better radical stabilizers than guanines. To increase the RNA labeling efficiency by APEX-seq, we thought to decorate RNAs with phenol groups to mimic tyrosines by using phenol-substituted kethoxals.

We made phenol-kethoxal and diphenol-kethoxal with click chemistry by using N_3-kethoxal as the starting material (Figure 6.2A). As expected, both molecules label RNA oligos with high efficiency. We then applied the probes to cells and isolated phenol-decorated RNAs for peroxidase-mediated labeling reaction in vitro (Figure 6.2B). We used biotin-phenol as the substrate to mimic the APEX-seq condition and checked the labeling efficiency by dot blot assay using streptavidin-HRP. Comparing with the control RNA with no phenol decorations, both phenol and diphenol labeled RNA show more evident biotin signals, confirming that phenol decoration could enhance peroxidase-mediated labeling (Figure 6.2C). Although further in vivo study is needed, our preliminary data shows the power of kethoxal as a mediator for modifying RNA with non-natural functional groups.

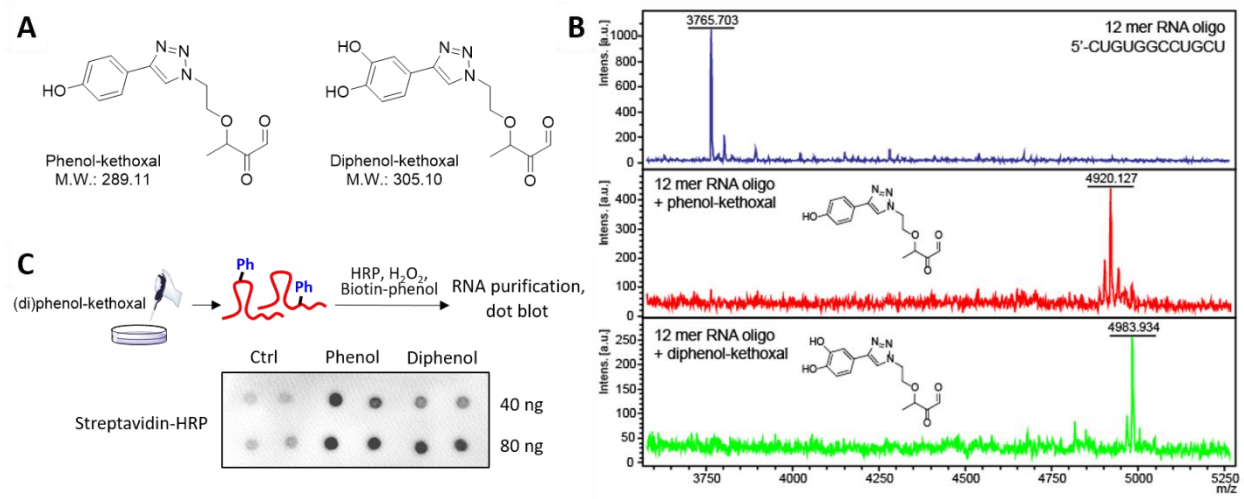

Figure 6.2 Enhance peroxidase-mediated RNA labeling by kethoxal derivatives

(**Figure 6.2, continued**) (**A**) The structures of phenol-kethoaxl and diphenol-kethoxal. (**B**) MALDI-TOF analysis of the reaction between a 12-mer RNA oligo and phenol-kethoxal and diphenol-kethoxal. (**C**) Testing the reactivity of phenol and diphenol decorated RNA in vitro. Top: the experiment scheme. Bottom: a dot blot assay showing biotin signals on the control, phenol decorated, and diphenol decorated RNA after horseradish peroxidase (HRP) mediated radical reaction with biotin-phenol as the substrate.

6.3 The future directions of ssDNA and RNA structure studies

Despite current state-of-the-art facilitate a better understanding of transcription, ssDNA, and RNA structure regulation, many important questions remain unanswered. New experimental and analytical tools are still in need.

6.3.1 Deconvolution of ssDNA signals.

As N_3-kethoxal labels all single-stranded DNA, KAS-seq is not able to distinguish ssDNA generated from different biological processes. In this thesis, I showed that more than 90% of KAS-seq signals are the outcome of gene transcription, with some signals deriving from enhancers and non-B form ssDNA structures. To exclude the potential "contamination" from transcription signals, non-B form DNA structures shown in Figure 3.10 are defined by KAS-seq data under triptolide condition. Therefore, most of the non-B form DNA regions presented in this thesis are intergenic. However, many non-B form DNA structures can be located in the coding region, and their formation may correlate with transcription activities. For instance, DNA G-quadruplexes (G4) were shown to be enriched at gene promoter and 5' untranslated regions (UTR), and could potentially control the transcription of certain genes[167]. Some enhancers are also located in the gene coding regions. Moreover, although KAS-seq can be done with low input materials, the profile generated is still an average of a cell population. Therefore, ssDNA produced from DNA replication, which exists almost evenly throughout the whole genome, were averaged out and blocked by transcription signals.

Therefore, in the mixture of ssDNA signals, it is important to sort out and quantify the portion of ssDNA contributed by different biological events. This may require efforts from both experiment and analysis sides. Comparing KAS-seq signals with epigenetic marks, TF binding, chromatin accessibility, and DNA sequence features could help the deconvolution process from a perspective of bioinformatics. KAS-seq with higher resolution, strand specificity, or in a single-cell manner will reveal unique features about different sorts of ssDNA species.

6.3.2 Regulatory roles of certain ssDNA structures

Although the majority of ssDNA are transcription bubbles, it is interesting to see that ssDNA containing enhancers (SSEs) contains specific sequence motifs (Figure 3.12D, Figure 3.15B) and genes associated with these enhancers are associated with particular functions (Figure 3.13C, Figure 3.15C). We show that the motifs enriched by SSEs are different from motifs enriched by enhancers with high ATAC-seq signals (Figure 3.12E), indicating that a portion of TFs are more associated with SSEs. We also expanded the comparison between ATAC-seq and KAS-seq to promoters, with similar results obtained.

Therefore, a group of proteins are associated with the feature of getting single-stranded. What are these proteins? Do they help to open the bubble, or do they simply preferentially bind to ssDNA structures? These are all questions that remain to be studied.

Some proteins were known to bind specific ssDNA structures. For instance, in breast cancer and ovarian cancer cells, BRCA1 protein interacts with helicase protein SETX to suppress R-loops at gene termination sites, thereby reducing DNA damage and increasing the genome stability[168]. BRCA2 depletion also increases R-loop levels and causes genome instability[169]. Moreover, Myc-associated zinc finger (MAZ) and poly(ADP-ribose) polymerase 1 (PARP-1) proteins bind to G-quadruplex structure upstream of the TSS of some genes, suggesting these

proteins may affect transcription via G4-associated mechanisms[170]. KAS-seq data should enable the systematic discovery of these ssDNA binder/effector proteins. It will be more exciting if KAS-seq can differentiate different ssDNA species after deconvolution and find the binding proteins for each type.

6.3.3 More accurate determination of RNA-RNA interactions

Although with many methods developed, detecting RNA-RNA interactions with high accuracy is still challenging, especially for transcripts with low abundance. There are several difficulties to profile RNA-RNA interactions accurately and efficiently. 1) some interactions can be transient and highly dynamic; 2) it is hard to estimate and remove the background in proximity-ligation based approaches; random ligation can happen in solution through disordered molecular collisions, even under ultra-dilute conditions; 3) the efficiency of crosslinking and ligation in existing methods tend to be low, with only a small portion of transcripts detected with structure information even with deep sequencing; 4) some transcripts have multiple copied in each cell, making it challenging to distinguish intermolecular from intramolecular interactions.

Steps have been taken to solve these problems. For instance, in an optimized PARIS protocol, amotosalen was used for RNA crosslinking to replace traditional psoralen AMT (aminomethyl trioxalen). Amotosalen has a higher water solubility than AMT, therefore increasing the crosslinking efficiency. Coupled with a denature-denature 2D gel system to purify crosslinked RNA, this protocol claims to result in more than 4,000-fold efficiency than the traditional PARIS method[171]. Proximity RNA-seq was recently developed to detect physically proximal transcripts by reading out different barcodes. It is based on massive-throughput RNA barcoding of subnuclear particles in water-in-oil emulsion droplets, and therefore, bypass proximity ligation[172].

Developing new methods should provide a more accurate and efficient determination of RNA interactomes and promise new insights into RNA biology.

List of references

1. Paule, M.R. & White, R.J. SURVEY AND SUMMARY Transcription by RNA polymerases I and III. *Nucleic Acids Research* **28**, 1283-1298 (2000).
2. Young, R.A. RNA POLYMERASE II. *Annual Review of Biochemistry* **60**, 689-715 (1991).
3. Borchert, G.M., Lanier, W. & Davidson, B.L. RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology* **13**, 1097-1101 (2006).
4. Tiranti, V. *et al.* Identification of the Gene Encoding the Human Mitochondrial RNA Polymerase (h-mtRPOL) by Cyberscreening of the Expressed Sequence Tags Database. *Human Molecular Genetics* **6**, 615-625 (1997).
5. Kornberg, R.D. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences* **104**, 12955 (2007).
6. Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45-54 (2019).
7. Chen, F.X., Smith, E.R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* **19**, 464-478 (2018).
8. Nelson, D.L., Lehninger, A.L. & Cox, M.M. *Lehninger principles of biochemistry*. (Macmillan, 2008).
9. Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics* **32**, 42-56 (2016).
10. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484-492 (2012).
11. Xiao, C.-L. *et al.* N6-Methyladenine DNA Modification in the Human Genome. *Molecular Cell* **71**, 306-318.e307 (2018).
12. Fu, Y. *et al.* N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. *Cell* **161**, 879-892 (2015).
13. Wu, T.P. *et al.* DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature* **532**, 329-333 (2016).
14. Schiffers, S. *et al.* Quantitative LC–MS Provides No Evidence for m6dA or m4dC in the Genome of Mouse Embryonic Stem Cells and Tissues. *Angewandte Chemie International Edition* **56**, 11268-11271 (2017).
15. Dethoff, E.A., Chugh, J., Mustoe, A.M. & Al-Hashimi, H.M. Functional complexity and regulation through RNA dynamics. *Nature* **482**, 322-330 (2012).
16. Mortimer, S.A., Kidwell, M.A. & Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* **15**, 469-479 (2014).
17. Henkin, T.M. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes & Development* **22**, 3383-3390 (2008).
18. Holley, R.W. *et al.* Structure of a Ribonucleic Acid. *Science* **147**, 1462 (1965).

19. Bernstein, E., Caudy, A.A., Hammond, S.M. & Hannon, G.J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363-366 (2001).

20. MacRae, I.J. *et al.* Structural Basis for Double-Stranded RNA Processing by Dicer. *Science* **311**, 195 (2006).

21. Park, J.-E. *et al.* Dicer recognizes the 5′ end of RNA for efficient and accurate processing. *Nature* **475**, 201-205 (2011).

22. Bevilacqua, P.C. & Assmann, S.M. Technique Development for Probing RNA Structure In Vivo and Genome-Wide. *Cold Spring Harbor Perspectives in Biology* **10**, a032250 (2018).

23. Strobel, E.J., Yu, A.M. & Lucks, J.B. High-throughput determination of RNA structures. *Nature Reviews Genetics* **19**, 615-634 (2018).

24. Underwood, J.G. *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods* **7**, 995-1001 (2010).

25. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103-107 (2010).

26. 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).

27. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696-700 (2014).

28. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706-709 (2014).

29. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E. & Weeks, K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods* **11**, 959-965 (2014).

30. Spitale, R.C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486-490 (2015).

31. Zubradt, M. *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods* **14**, 75-82 (2017).

32. 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. *Nature Methods* **13**, 841-844 (2016).

33. Pendleton, K.E. *et al.* The U6 snRNA m6A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell* **169**, 824-835.e814 (2017).

34. Mendel, M. *et al.* Methylation of Structured RNA by the m6A Writer METTL16 Is Essential for Mouse Embryonic Development. *Molecular Cell* **71**, 986-1000.e1011 (2018).

35. Liu, N. *et al.* N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* **518**, 560-564 (2015).

36. Singh, G., Pratt, G., Yeo, G.W. & Moore, M.J. The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. *Annual Review of Biochemistry* **84**, 325-354 (2015).

37. Trendel, J. *et al.* The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell* **176**, 391-403.e319 (2019).

38. Anderson, P. & Kedersha, N. RNA granules. *Journal of Cell Biology* **172**, 803-808 (2006).

39. Anderson, P. & Kedersha, N. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nature Reviews Molecular Cell Biology* **10**, 430-436 (2009).

40. Roden, C. & Gladfelter, A.S. RNA contributions to the form and function of biomolecular condensates. *Nature Reviews Molecular Cell Biology* (2020).

41. Dreyfuss, G., Kim, V.N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology* **3**, 195-205 (2002).

42. Lunde, B.M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology* **8**, 479-490 (2007).

43. Hentze, M.W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology* **19**, 327-341 (2018).

44. Ramanathan, M., Porter, D.F. & Khavari, P.A. Methods to study RNA–protein interactions. *Nature Methods* **16**, 225-234 (2019).

45. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149**, 1393-1406 (2012).

46. Urdaneta, E.C. *et al.* Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nature Communications* **10**, 990 (2019).

47. Chu, C. *et al.* Systematic Discovery of Xist RNA Binding Proteins. *Cell* **161**, 404-416 (2015).

48. Lee, F.C.Y. & Ule, J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell* **69**, 354-369 (2018).

49. Kubota, M., Tran, C. & Spitale, R.C. Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology* **11**, 933-941 (2015).

50. Gong, J., Ju, Y., Shao, D. & Zhang, Q.C. Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale. *Quantitative Biology* **6**, 239-252 (2018).

51. Watson, J.D. & Crick, F.H.C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738 (1953).

52. Fürtig, B., Richter, C., Wöhnert, J. & Schwalbe, H. NMR Spectroscopy of RNA. *ChemBioChem* **4**, 936-962 (2003).

53. Zhang, Q., Stelzer, A.C., Fisher, C.K. & Al-Hashimi, H.M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450**, 1263-1267 (2007).

54. Peattie, D.A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences* **77**, 4679-4682 (1980).

55. Moazed, D., Robertson, J.M. & Noller, H.F. Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23S RNA. *Nature* **334**, 362-364 (1988).

56. Climie, S.C. & Friesen, J.D. In vivo and in vitro structural analysis of the rplJ mRNA leader of Escherichia coli. Protection by bound L10-L7/L12. *Journal of Biological Chemistry* **263**, 15166-15175 (1988).

57. Mitchell, D. *et al.* Glyoxals as in vivo RNA structural probes of guanine base-pairing. *RNA* **24**, 114-124 (2018).

58. 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).

59. 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). *Journal of the American Chemical Society* **127**, 4223-4231 (2005).

60. Latham, J.A. & Cech, T.R. Defining the inside and outside of a catalytic RNA molecule. *Science* **245**, 276 (1989).

61. Feng, C. *et al.* Light-activated chemical probing of nucleobase solvent accessibility inside cells. *Nature Chemical Biology* **14**, 276-283 (2018).

62. Staehelin, M. Inactivation of virus nucleic acid with glyoxal derivatives. *Biochimica et Biophysica Acta* **31**, 448-454 (1959).

63. Brewer, L.A. & Noller, H.F. Ribonucleic acid-protein cross-linking within the intact Escherichia coli ribosome, utilizing ethylene glycol bis[3-(2-ketobutyraldehyde) ether], a reversible, bifunctional reagent: identification of 30S proteins. *Biochemistry* **22**, 4310-4315 (1983).

64. Tiffany, B.D. *et al.* Antiviral Compounds. I. Aliphatic Glyoxals, α-Hydroxyaldehydes and Related Compounds. *Journal of the American Chemical Society* **79**, 1682-1687 (1957).

65. Jiang, B., Liu, J.-F. & Zhao, S.-Y. Enantioselective Synthesis for the Antipodes of Slagenins B and C: Establishment of Absolute Stereochemistry. *Organic Letters* **3**, 4011-4013 (2001).

66. Lai, C., Lin, G., Wang, W. & Luo, H. Absolute configurations and stability of cyclic guanosine mono-adducts with glyoxal and methylglyoxal. *Chirality* **23**, 487-494 (2011).

67. Schroeder, S.J. & Turner, D.H. (Academic Press, 2009).

68. Andronescu, M., Bereg, V., Hoos, H.H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC bioinformatics* **9**, 340 (2008).

69. Guo, J.U. & Bartel, D.P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **353**, aaf5371 (2016).

70. 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. *Nature Chemistry* **6**, 75-80 (2014).

71. Kwok, C.K., Marsico, G. & Balasubramanian, S. Detecting RNA G-Quadruplexes (rG4s) in the Transcriptome. *Cold Spring Harbor Perspectives in Biology* **10**, a032284 (2018).

72. Kalvari, I. *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research* **46**, D335-D342 (2017).

73. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).

74. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).

75. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**, 1845-1848 (2008).

76. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339**, 950-953 (2013).

77. Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biology* **15**, R69 (2014).

78. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225-1228 (2016).

79. Churchman, L.S. & Weissman, J.S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368-373 (2011).

80. Christopher, Ramachandran, S. & Henikoff, S. Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase. *Molecular Cell* **53**, 819-830 (2014).

81. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526-540 (2015).

82. Mayer, A. *et al.* Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell* **161**, 541-554 (2015).

83. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nature Genetics* **51**, 1369-1379 (2019).

84. Mirkovitch, J. & Darnell, J.E. Mapping of RNA polymerase on mammalian genes in cells and nuclei. *Molecular Biology of the Cell* **3**, 1085-1094 (1992).

85. Muse, G.W. *et al.* RNA polymerase is poised for activation across the genome. *Nature Genetics* **39**, 1507-1511 (2007).

86. Kouzine, F. *et al.* Global Regulation of Promoter Melting in Naive Lymphocytes. *Cell* **153**, 988-999 (2013).

87. Kouzine, F. *et al.* Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. *Cell Systems* **4**, 344-356.e347 (2017).

88. Shapiro, R. & Hachmann, J. The Reaction of Guanine Derivatives with 1,2-Dicarbonyl Compounds*. *Biochemistry* **5**, 2799-2807 (1966).

89. Weng, X. *et al.* Keth-seq for transcriptome-wide RNA structure mapping. *Nature Chemical Biology* **16**, 489-492 (2020).

90. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213-1218 (2013).

91. Cer, R.Z. *et al.* Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Proceedings of the National Academy of Sciences* **41**, D94-D100 (2013).

92. Henriques, T. *et al.* Widespread transcriptional pausing and elongation control at enhancers. *Genes & Development* (2018).

93. Warren *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307-319 (2013).

94. Raffaella *et al.* Control of Embryonic Stem Cell Identity by BRD4-Dependent Transcriptional Elongation of Super-Enhancer-Associated Pluripotency Genes. *Cell Reports* **9**, 234-247 (2014).

95. Liu, W. *et al.* BRD4 regulates Nanog expression in mouse embryonic stem cells and preimplantation embryos. *Cell Death & Differentiation* **21**, 1950-1960 (2014).

96. Wu, T., Kamikawa, Y.F. & Donohoe, M.E. Brd4's Bromodomains Mediate Histone H3 Acetylation and Chromatin Remodeling in Pluripotent Cells through P300 and Brg1. *Cell Reports* **25**, 1756-1771 (2018).

97. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495-501 (2010).

98. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394 (2011).

99. Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520 (2013).

100. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

101. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010-1014 (2015).

102. Li, W., Notani, D. & Rosenfeld, M.G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics* **17**, 207-223 (2016).

103. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. & Sharp, P.A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23 (2017).

104. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855.e1816 (2018).

105. Sabari, B.R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).

106. Cho, W.-K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412-415 (2018).

107. Chong, S. *et al.* Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**, eaar2555 (2018).

108. Guo, Y.E. *et al.* Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572**, 543-548 (2019).

109. Zhou, Z.X. *et al.* Mapping genomic hotspots of DNA damage by a single-strand-DNA-compatible and strand-specific ChIP-seq method. *Genome Research* **23**, 705-715 (2013).

110. Khil, P.P., Smagulova, F., Brick, K.M., Camerini-Otero, R.D. & Petukhova, G.V. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. *Genome Research* **22**, 957-965 (2012).

111. Lydall, D., Nikolsky, Y., Bishop, D.K. & Weinert, T. A meiotic recombination checkpoint controlled by mitotic checkpoint genes. *Nature* **383**, 840-843 (1996).

112. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

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

114. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

115. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

116. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357-360 (2015).

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

118. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**, W187-W191 (2014).

119. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576-589 (2010).

120. Thomas-Chollier, M. *et al.* Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols* **6**, 1860-1869 (2011).

121. Lu, Z. *et al.* RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**, 1267-1279 (2016).

122. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. & Blencowe, Benjamin J. Global Mapping of Human RNA-RNA Interactions. *Molecular Cell* **62**, 618-626 (2016).

123. Aw, Jong Ghut A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular Cell* **62**, 603-617 (2016).

124. Ramani, V., Qiu, R. & Shendure, J. High-throughput determination of RNA structure by proximity ligation. *Nature Biotechnology* **33**, 980-984 (2015).

125. Nguyen, T.C. *et al.* Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. *Nature Communications* **7**, 12023 (2016).

126. Cai, Z. *et al.* RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature* **582**, 432-437 (2020).

127. Alsemarz, A., Lasko, P. & Fagotto, F. Limited significance of the in situ proximity ligation assay. *bioRxiv*, 411355 (2018).

128. Anger, A.M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80-85 (2013).

129. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences* **109**, 14728 (2012).

130. Jones, P.A. & Takai, D. The Role of DNA Methylation in Mammalian Epigenetics. *Science* **293**, 1068 (2001).

131. Ehrlich, M. & Wang, R.Y. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350 (1981).

132. Zhang, X. *et al.* Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell* **126**, 1189-1201 (2006).

133. Chen, Z.-X. & Riggs, A.D. DNA Methylation and Demethylation in Mammals. *Journal of Biological Chemistry* **286**, 18347-18353 (2011).

134. Luo, G.-Z., Blanco, M.A., Greer, E.L., He, C. & Shi, Y. DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nature Reviews Molecular Cell Biology* **16**, 705-710 (2015).

135. Luo, G.-Z. *et al.* N6-methyldeoxyadenosine directs nucleosome positioning in Tetrahymena DNA. *Genome Biology* **19**, 200 (2018).

136. Beh, L.Y. *et al.* Identification of a DNA N6-Adenine Methyltransferase Complex and Its Impact on Chromatin Organization. *Cell* **177**, 1781-1796.e1725 (2019).

137. Greer, Eric L. *et al.* DNA Methylation on N6-Adenine in C. elegans. *Cell* **161**, 868-878 (2015).

138. Zhang, G. *et al.* N6-Methyladenine DNA Modification in Drosophila. *Cell* **161**, 893-906 (2015).

139. Mondo, S.J. *et al.* Widespread adenine N6-methylation of active genes in fungi. *Nature Genetics* **49**, 964-968 (2017).

140. Koziol, M.J. *et al.* Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature Structural & Molecular Biology* **23**, 24-30 (2016).

141. Xie, Q. *et al.* N6-methyladenine DNA Modification in Glioblastoma. *Cell* **175**, 1228-1243.e1220 (2018).

142. Yao, B. *et al.* DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nature Communications* **8**, 1122 (2017).

143. Ma, C. *et al.* N6-methyldeoxyadenine is a transgenerational epigenetic signal for mitochondrial stress adaptation. *Nature Cell Biology* **21**, 319-327 (2019).

144. King, M.P. & Attardi, G. Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation. *Science* **246**, 500 (1989).

145. Rossi, M.J., Lai, W.K.M. & Pugh, B.F. Simplified ChIP-exo assays. *Nature Communications* **9**, 2842 (2018).

146. Iyer, L.M., Zhang, D. & Aravind, L. Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. *BioEssays* **38**, 27-40 (2016).

147. Liu, J. *et al.* A METTL3–METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nature Chemical Biology* **10**, 93-95 (2014).

148. Luo, G.-Z. & He, C. DNA N6-methyladenine in metazoans: functional epigenetic mark or bystander? *Nature Structural & Molecular Biology* **24**, 503-506 (2017).

149. Smith, A.C. & Robinson, A.J. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. *Nucleic Acids Research* **47**, D1225-D1228 (2018).

150. Wang, Y. *et al.* N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology* **16**, 191-198 (2014).

151. Guzy, R.D. *et al.* Mitochondrial complex III is required for hypoxia-induced ROS production and cellular oxygen sensing. *Cell Metabolism* **1**, 401-408 (2005).

152. Murphy, Michael P. How mitochondria produce reactive oxygen species. *Biochemical Journal* **417**, 1-13 (2008).

153. Bonawitz, N.D., Clayton, D.A. & Shadel, G.S. Initiation and Beyond: Multiple Functions of the Human Mitochondrial Transcription Machinery. *Molecular Cell* **24**, 813-825 (2006).

154. Agaronyan, K., Morozov, Y.I., Anikin, M. & Temiakov, D. Replication-transcription switch in human mitochondria. *Science* **347**, 548 (2015).

155. Fazakerley, G.V., Gabarro-Arpa, J., Lebret, M., Guy, A. & Guschlbauer, W. The GTm 6 AC sequence is overwound and bent. *Nucleic Acids Research* **17**, 2541-2556 (1989).

156. Bang, J., Bae, S.-H., Park, C.-J., Lee, J.-H. & Choi, B.-S. Structural and Dynamics Study of DNA Dodecamer Duplexes That Contain Un-, Hemi-, or Fully Methylated GATC Sites. *Journal of the American Chemical Society* **130**, 17688-17696 (2008).

157. Polaczek, P., Kwan, K. & Campbell, J.L. GATC motifs may alter the conformation of DNA depending on sequence context and N6-adenine methylation status: possible implications for DNA-protein recognition. *Molecular and General Genetics MGG* **258**, 488-493 (1998).

158. Bestwick, M.L. & Shadel, G.S. Accessorizing the human mitochondrial transcription machinery. *Trends in Biochemical Sciences* **38**, 283-291 (2013).

159. Uchida, A. *et al.* Unexpected sequences and structures of mtDNA required for efficient transcription from the first heavy-strand promoter. *Elife* **6**, e27283 (2017).

160. Ngo, H.B., Kaiser, J.T. & Chan, D.C. The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. *Nature Structural & Molecular Biology* **18**, 1290-1296 (2011).

161. Campbell, C.T., Kolesar, J.E. & Kaufman, B.A. Mitochondrial transcription factor A regulates mitochondrial transcription initiation, DNA packaging, and genome copy number. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1819**, 921-929 (2012).

162. Lezza, A.M.S. Mitochondrial transcription factor A (TFAM): one actor for different roles. *Frontiers in Biology* **7**, 30-39 (2012).

163. Gray, M.W., Burger, G. & Lang, B.F. Mitochondrial Evolution. *Science* **283**, 1476 (1999).

164. Wright, R., Stephens, C. & Shapiro, L. The CcrM DNA methyltransferase is widespread in the alpha subdivision of proteobacteria, and its essential functions are conserved in Rhizobium meliloti and Caulobacter crescentus. *Journal of Bacteriology* **179**, 5869 (1997).

165. Chen, H. *et al.* METTL4 is an snRNA m6Am methyltransferase that regulates RNA splicing. *Cell Research* **30**, 544-547 (2020).

166. Fazal, F.M. *et al.* Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473-490.e426 (2019).

167. Spiegel, J., Adhikari, S. & Balasubramanian, S. The Structure and Function of DNA G-Quadruplexes. *Trends in Chemistry* **2**, 123-136 (2020).

168. Hatchi, E. *et al.* BRCA1 Recruitment to Transcriptional Pause Sites Is Required for R-Loop-Driven DNA Damage Repair. *Molecular Cell* **57**, 636-647 (2015).

169. Bhatia, V. *et al.* BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511**, 362-365 (2014).

170. Brázda, V., Hároníková, L., Liao, J.C.C. & Fojta, M. DNA and RNA quadruplex-binding proteins. *International journal of molecular sciences* **15**, 17493-17517 (2014).

171. Zhang, M. *et al.* Optimized photochemistry and enzymology enable efficient analysis of RNA structures and interactions in cells and virus infections. *bioRxiv*, 2020.2004.2030.071167 (2020).

172. Morf, J. *et al.* RNA proximity sequencing reveals the spatial organization of the transcriptome in the nucleus. *Nature Biotechnology* **37**, 793-802 (2019).