RESOURCE
https://doi.org/10.1038/s41594-019-0218-x

nature structural & molecular biology

Genome-wide identification of mRNA 5-methylcytosine in mammals

Tao Huang[1,3], Wanying Chen[1,3], Jianheng Liu[1,3], Nannan Gu[1] and Rui Zhang[1,2*]

Accurate and systematic transcriptome-wide detection of 5-methylcytosine (m^5C) has proved challenging, and there are conflicting views about the prevalence of this modification in mRNAs. Here we report an experimental and computational framework that robustly identified mRNA m^5C sites and determined sequence motifs and structural features associated with the modification using a set of high-confidence sites. We developed a quantitative atlas of RNA m^5C sites in human and mouse tissues based on our framework. In a given tissue, we typically identified several hundred exonic m^5C sites. About 62–70% of the sites had low methylation levels (<20% methylation), while 8–10% of the sites were moderately or highly methylated (>40% methylation). Cross-species analysis revealed that species, rather than tissue type, was the primary determinant of methylation levels, indicating strong *cis*-directed regulation of RNA methylation. Combined, these data provide a valuable resource for identifying the regulation and functions of RNA methylation.

Cellular RNAs can be chemically modified in over a hundred different ways, and such modifications have been recently deemed important post-transcriptional regulatory features[1–3]. Despite the fact that so many RNA modifications have been documented[4], our understanding about their regulation and function is still limited, especially for messenger RNAs (mRNAs), due to the technical limitations in accurately locating most modifications genome-wide[5]. For example, we and other groups have shown the difficulty in removing false positives and identifying genuine adenosine-to-inosine (A-to-I) modification in RNA-sequencing data[6–10]. 5-methylcytosine (m^5C) is one of the longest-known RNA modifications and is mediated by the DNMT2 and NSUN methyltransferase family[11]. Previous studies showed that m^5C is present in diverse RNA species. m^5C in transfer RNAs (tRNAs) is involved in translational regulation by affecting tRNA stability and translational fidelity[11,12], as well as by controlling formation of the tRNA fragment that regulates protein synthesis[13]. m^5C also occurs in ribosomal RNA (rRNA) and is involved in the quality control of ribosome biogenesis[14,15]. More importantly, the regulatory role of m^5C in mRNAs is beginning to be revealed. Recent studies have indicated that mRNA m^5C regulates the structure, stability and translation of mRNAs[16–19]. The fate of m^5C-modified mRNAs can be regulated by the reader protein of m^5C[20]. Moreover, the mRNA m^5C level has been shown to be mis-regulated in pathological contexts[21].

Several methods have been applied in the identification of mRNA m^5C sites. m^5C RNA immunoprecipitation (RIP) could provide a global view of m^5C map, and Aza-IP[22] and miCLIP[23] have been used to identify the m^5C map of NSUN2; however, these methods cannot reach single-base resolution. Additionally, the ability of these methods to identify mRNA m^5C may have been limited because the mRNA was not enriched before sequencing. Similar to m^5C in DNA[24], RNA m^5C persists to sulfonation and can also be determined by bisulfite treatment. Thus, RNA bisulfite sequencing (BS-seq) on enriched mRNA is a better choice to identify and quantify mRNA m^5C[25]. A few studies have used RNA BS-seq and developed computational approaches to identify mRNA m^5C at the transcriptome level[20,26–30]. However, variable results have been observed, and the number of putative mRNA m^5C sites varies 1,000-fold between studies. Among them, three studies mapped up to 10,000 sites in one tissue or cell type in mammals[20,26,29]. However, these studies could not define a common set of substrate mRNAs or consensus methylation target sequences, indicating that some of the results were influenced by incomplete conversion of structural RNAs or improper analysis of the sequencing data. A recent study, instead, suggested that mRNAs were found to be very sparsely methylated or not methylated at all[28]. Thus, a major challenge in mRNA m^5C studies is to develop an experimental and computational framework that can distinguish true mRNA m^5C events from noise and reveal the landscape and feature of mRNA m^5C.

Results

Evaluation of RNA bisulfite treatment conditions for cytosine conversion and m^5C level quantification. As m^5C is generally believed to be less abundant in mRNAs than in rRNAs or tRNAs, it is important to achieve a high conversion rate to remove false-positive sites due to incomplete deamination. To develop a robust BS-seq library construction protocol, we selected the EZ RNA methylation kit as our starting point and modified the conversion steps to achieve a high conversion rate (Methods). We tested three different conversion conditions (low-, medium- and high-stringency conditions, see Supplementary Table 1 and Methods) using poly(A)-selected RNAs. The unmethylated RNA control mixes developed by the External RNA Controls Consortium (ERCC) were spiked into the poly(A)-selected RNAs to assess the conversion rate for transcripts of dynamic range. ERCC mixes contain pre-formulated blends of 92 transcripts of 250–2,000 nucleotides (nt) in length, which mimic natural eukaryotic mRNAs with secondary structures. Moreover, they span an approximately 10^6-fold concentration range. Next, we characterized our experimental method by analyzing both the ERCC mixes and poly(A)-selected RNAs. Reads were mapped to the genome by HISAT2 and were subsequently mapped to the transcriptome by Bowtie2 (Supplementary Fig. 1a and Supplementary Note 1). First, we found that the

[1]MOE Key Laboratory of Gene Function and Regulation, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China. [2]RNA Biomedical Institute, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China. [3]These authors contributed equally: Tao Huang, Wanying Chen, Jianheng Liu. *e-mail: zhangrui3@mail.sysu.edu.cn

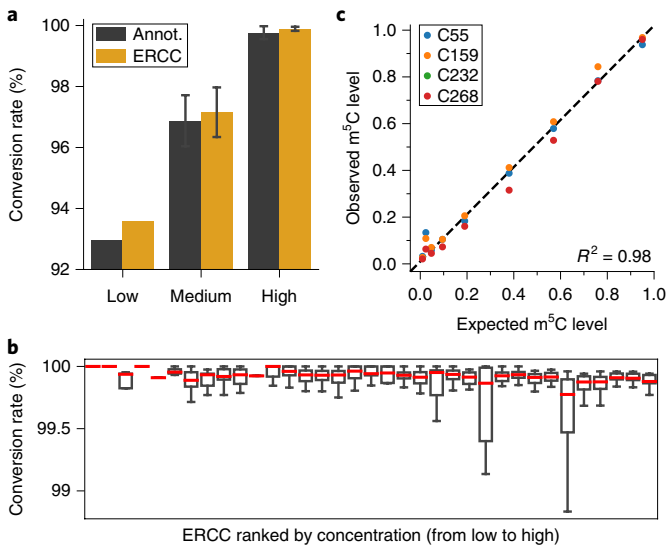

Fig. 1 | The evaluation of BS-seq library construction protocols with different reaction conditions. a, Comparison of conversion rates estimated by ERCC mixes or all annotated genes between libraries constructed using different conversion conditions. We integrated all libraries that were constructed in this study for analysis. Samples constructed using low-stringency conditions, HEK293T cells; medium-stringency conditions, HEK293T cells and multiple mouse tissues ($n=14$); high-stringency conditions, HEK293T cells, multiple mouse and human tissues ($n=17$). Center line represents the mean and error bars indicate standard deviation. Annot., all genes from Ensembl annotation. **b**, Boxplot for conversion rates of ERCC mixes in samples treated using high-stringency condition. ERCC mixes on the *x* axis are sorted by concentration. Only ERCC transcripts with C-position coverage of ≥1,000 were used for the analysis. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ±1.5× interquartile range (IQR). **c**, In vitro-transcribed transcripts with either m^{5}Cs or Cs were mixed at different frequencies (1%, 2%, 5%, 10%, 20%, 40%, 60%, 80% and 100%). Colors indicate different m^{5}C sites. R^2, squared Pearson correlation coefficient between the expected values and the observed values.

overall conversion rates with different conditions ranged from 93% to 99.9% (Fig. 1a and Supplementary Table 2) and the high-stringency condition reached an average conversion rate of 99.8%. Second, the coverages of ERCC mixes were correlated with their concentrations (Supplementary Fig. 1b). No obvious coverage difference was observed between different conversion conditions for most transcripts, indicating that the additional heat treatment in high-stringency condition did not have a notable effect on the integrity of sequenced transcripts. Third, in the high-stringency condition, the conversion rates were equal for ERCC mixes with different concentrations (Fig. 1b); thus, this bisulfite treatment condition could fully convert the unmethylated cytosines (Cs) in the transcripts with dynamic expression ranges.

To confirm that the quantification of m^{5}C level under the high-stringency bisulfite treatment condition is accurate, we generated in vitro–transcribed m^{5}C and non-m^{5}C transcripts, spiked into the poly(A)-selected RNA samples and measured ratios in prepared mixtures using RNA BS-seq. We found that the observed m^{5}C frequency is highly consistent with the expected frequency (Fig. 1c and Supplementary Fig. 1c), particularly when the m^{5}C level is >5%.

The source of noise in BS-seq. We next sought to identify sources of noise in BS-seq data by characterizing the conversion rate, m^{5}C site distribution and raw reads of each gene using BS-seq data generated by others and us (Supplementary Fig. 2a,b and Supplementary Note 2). We found that the conversion rates of individual genes varied and some were much lower than the overall conversion rate (Fig. 2a and Supplementary Fig. 2c). So, the statistical test with a single representation of conversion rate, which was used in previous BS-seq analysis, was inadequate to remove false positives. Furthermore, we observed that 52–87% of the putative m^{5}C sites reported from previous studies[20,29] were clustered in specific genes (Supplementary Fig. 2d) and were called from reads with multiple non-converted Cs (C-reads) (Supplementary Note 3). In addition, the clustered sites and C-reads varied among studies (Fig. 2b and Supplementary Note 3). These results indicate that reads with multiple non-converted Cs were probably generated due to the conversion failure, thus they need to be removed using a carefully designed computational filter.

Since there is no knowledge about whether and to what extent the real m^{5}C sites are clustered, the parameter of C-reads filter could only be experimentally estimated. Here, we used m^{5}C RIP–sequencing (RIP-seq) with in vitro–transcribed m^{5}C transcripts to determine this parameter. We first confirmed the feasibility and the specificity of m^{5}C pulldown using dot blot and m^{5}C RIP real-time PCR (rtPCR) (Supplementary Note 4 and Supplementary Fig. 3a–e). Next, we performed m^{5}C RIP-seq using fragmented HEK293T and HeLa mRNAs (Supplementary Table 3). The in vitro–transcribed m^{5}C and non-m^{5}C transcripts were spiked into HeLa mRNA for m^{5}C RIP-seq (Methods). Reads were mapped, each gene was split into the 100-nt sliding windows and an enrichment fold was calculated for each window, as previously described[31]. We found that over the 784,372 and 697,964 mRNA windows examined in HEK293T and HeLa cells, fewer than ten windows had an enrichment fold greater than that of the in vitro–transcribed transcripts with 5 m^{5}C sites (Fig. 2c,d and Supplementary Fig. 3f). Additionally, although the low affinity of anti-m^{5}C antibody prevents it from calling individual m^{5}C peaks directly, the windows with non-clustered m^{5}C sites tended to have higher enrichment folds compared with the non-m^{5}C control windows (Fig. 2e and Supplementary Fig. 3g,h). In contrast, no such difference was found for the windows with clustered sites (Fig. 2e and Supplementary Fig. 3g,h). These results together indicate that the sites called from C-reads (for example, 150 base pair (bp) reads with multiple Cs) were false-positive sites and were derived from the bisulfite treatment–resistant regions, which laid the foundation for the C-cutoff filter of our pipeline below.

The computational pipeline for m^{5}C discovery in mRNAs. Next, we developed a computational pipeline to accurately identify mRNA m^{5}C sites (Fig. 2f and Methods). The characteristic features that distinguish our approach are (1) the careful design of read preprocessing and mapping steps; (2) the incorporation of the Gini coefficient to determine the C-cutoff to remove reads with Cs (Supplementary Fig. 4a and Supplementary Note 5); (3) the signal ratio filter to evaluate the conversion status of a region and further remove false-positive sites in bisulfite treatment–resistant regions (Supplementary Fig. 4b–d and Supplementary Note 6); (4) the identification of statistically significantly methylated sites with binomial *P* values on the basis of the gene-specific conversion rate; (5) the exclusion of genes with low conversion rates; and (6) the choice of Stouffer's method to calculate the combined *P* value for biological replicates, which gives high specificity. These analysis steps were designed to remove false discoveries caused by errors introduced during the construction and sequencing of BS-seq libraries, stochastic process in bisulfite conversion, incorrect mapping of short reads and conversion failure in bisulfite treatment–resistant regions.

Pipeline verification and the identification of sequence and structural features of mRNA m^{5}C. To verify our computational pipeline, we first applied it to the BS-seq data from HeLa NSUN2-knockdown and control samples[20], as NSUN2 significantly affects the mRNA

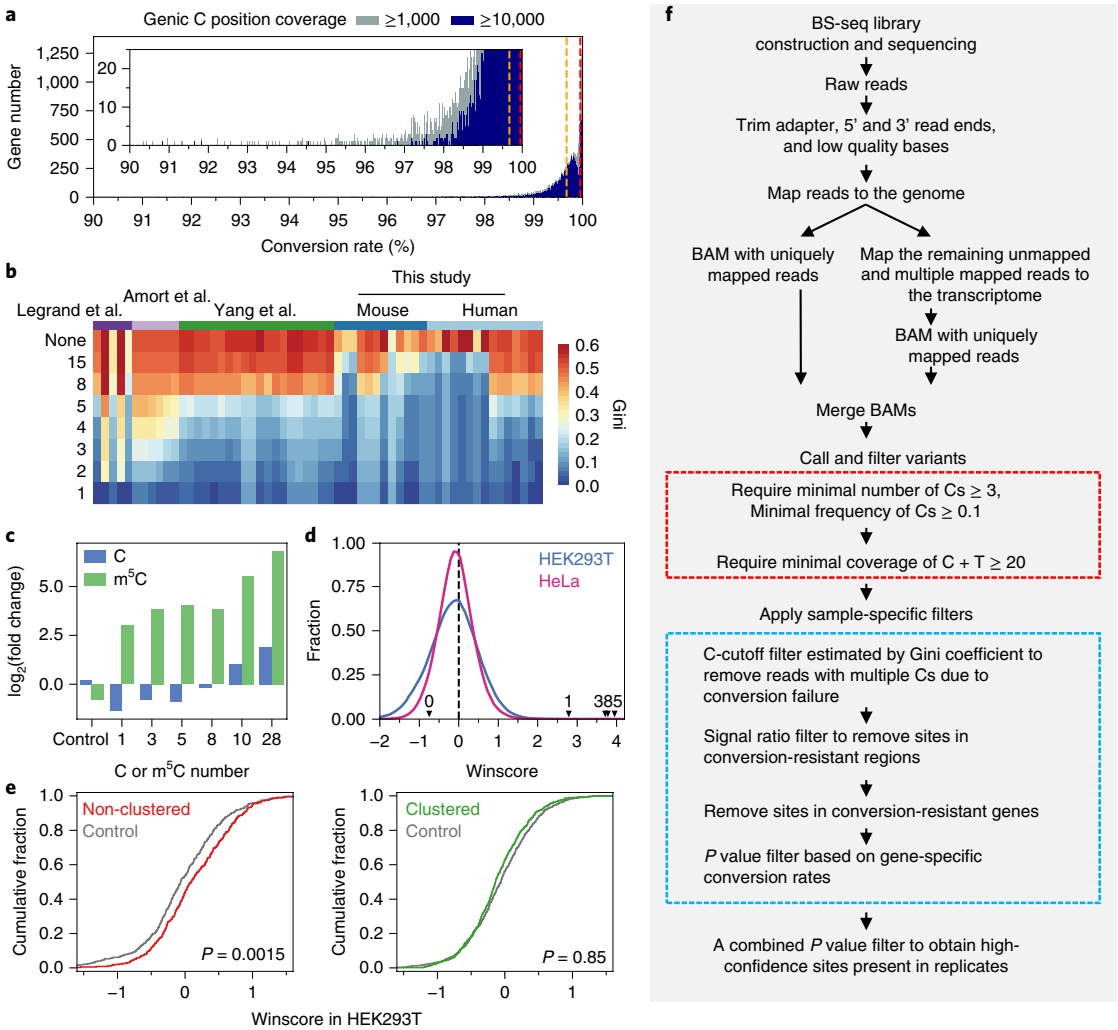

Fig. 2 | Development of a computational framework to identify high-confidence m^{5}C sites. a, The distribution of gene-specific conversion rates in a HEK293T sample using high-stringency conditions. Genes with two coverage cutoffs are shown: >1,000 (gray) and >10,000 (blue). Overall conversion rate (orange) and conversion rate estimated by ERCC mixes (red) are indicated with dashed lines. A zoomed inset is also shown. **b**, Gini coefficients among different samples. All samples from different studies were analyzed using our computational pipeline, and the Gini coefficients were calculated on the basis of the sites called using the same mapping procedure and filters listed in **f** with different C-cutoffs. This study, 15 human samples and 12 mouse samples; Yang et al.[20], 16 mouse samples and four HeLa cell samples; Amort et al.[29], three mouse brain and three mouse endothelial stem cells samples; and Legrand et al.[28], five mouse samples. **c**, Fold change of spike-in transcripts between immunoprecipitation and input. Seven 100-nt oligos containing different numbers of Cs were transcribed with cytidine triphosphate (CTP) or 5-methyl-cytidine triphosphate (m5CTP). The control oligo containing 23 Cs was transcribed only with CTP. These in vitro-transcribed transcripts were spiked into the fragmented mRNA for RIP-seq. **d**, Winscore distribution of mRNAs in HEK293T and HeLa cells. Each gene was split into the 100-nt sliding windows for enrichment score calculation. The enrichment fold of in vitro-transcribed transcripts with different numbers of m^{5}C modifications is marked by arrows. **e**, Cumulative distributions of winscores of windows with either non-clustered or clustered m^{5}C sites in HEK293T cells. Non-clustered, windows containing the high-confidence sites; clustered, windows containing the sites called without sample-specific filters minus windows containing the high-confidence sites. Control windows were selected from upstream and downstream regions relative to windows containing the m^{5}C site (Methods and Supplementary Fig. 3g). Windows with reads per kilobase per million mapped reads (RPKM) values of ≥1 in the input were used for analysis. The *P* value was determined using a one-sided Kolmogorov–Smirnov test. **f**, Schematic diagram of the RNA bisulfite sequencing analysis pipeline. C + T, cytosine and thymine.

m^{5}C levels in HeLa cells, based on mass spectrometry results[20]. Of the sites called using our method, 61.8% had significantly decreased m^{5}C levels in the knockdown sample (Fig. 3a), suggesting that these were real m^{5}C sites and NSUN2 dependent. The proportions of the sites with significantly decreased m^{5}C levels after NSUN2 knockdown were greatly increased using our filters (Supplementary Fig. 4e). As expected, no m^{5}C level change was observed for the clustered sites (Fig. 3b), and the proportion of the sites with significantly decreased m^{5}C levels after NSUN2 knockdown was negatively correlated with the cluster status of a site (Supplementary Fig. 4f).

Remarkably, a strong 3′ G-rich triplet (3′ NGGG) motif was found in NSUN2-dependent sites (Fig. 3c), indicating a previously unknown sequence preference of NSUN2. Further, a 3′ TCCA motif was observed in NSUN2-independent sites (Fig. 3c). Among the known RNA methyltransferases expressed in HeLa cells, only NSUN2 seemed to regulate mRNA m^{5}C (Supplementary Fig. 5 and Supplementary Note 7), indicating the presence of one or more unknown mRNA methyltransferases with a 3′ TCCA motif preference. Next, we analyzed the local structure of the putative m^{5}C sites. We found that the clustered sites tended to be located in the

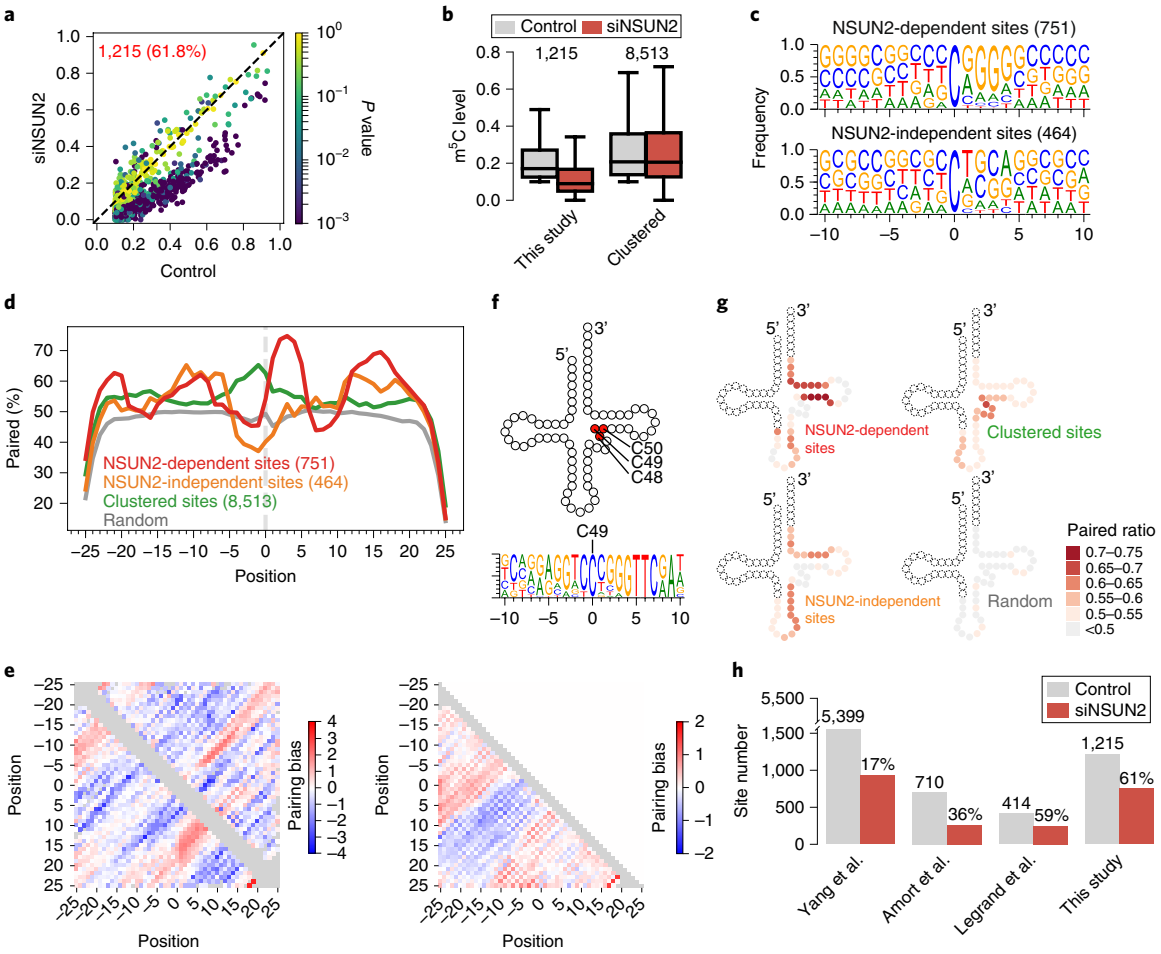

Fig. 3 | Performance of our computational pipeline and sequence and structural features of mRNA m^{5}C. a, Comparison of m^{5}C levels of individual sites between NSUN2-knockdown and control cells. The significant difference in m^{5}C levels was determined by a two-sided Fisher's exact test. The number of exonic sites called in the control sample is indicated in red. The percentage of sites with significantly decreased ($P < 0.05$) m^{5}C levels in the knockdown sample is given in parenthesis. **b**, Boxplot showing m^{5}C level change between NSUN2 knockdown and control samples. High-confidence sites, sites called using our pipeline ($n = 1,215$); clustered sites, sites called without sample-specific filters minus high-confidence sites ($n = 8,513$). Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ±1.5× IQR. **c**, The sequence context of C sites with (top) or without (bottom) significantly decreased m^{5}C levels after NSUN2 knockdown. **d**, Metaprofiles of the secondary structure of m^{5}C sites and the flanking regions. Sites identified in HeLa cells were used for analysis. Position 0 represents the m^{5}C sites. Each negative or positive value indicates the distance between an upstream or downstream position and the m^{5}C site. Percentage of paired bases at each position was calculated, and 100,000 randomly selected Cs from the transcribed regions of the human genome were used as the control. The numbers of the different sites analyzed are given in parenthesis. **e**, Visualizing the metaprofiles of the RNA secondary structure. Left: NSUN2-dependent sites (below diagonal) and NSUN2-independent sites (above diagonal). Right: clustered sites. Pairing bias was defined as $\log_2(\text{Ratio}_{\text{Observed}}(i,j)/(\text{Ratio}_{\text{Random}}(i,j))$, where Ratio(i,j) is the percentage of paired bases between position i and position j. **f**, Top: three NSUN2-specific substrates on tRNA. Bottom: sequence context of methylated C49 obtained from Aza-IP VarScan results[22]. **g**, The predicted secondary structures in **d** were centrally aligned to position C49 on tRNA, and the paired ratio of 20-nt flanking regions are shown. **h**, Comparison of our method with recent efforts by other groups. The number of total sites analyzed and the percentage of sites with significantly decreased m^{5}C levels in the knockdown sample is given above the barplot. To determine the significant difference, only sites that are present in control samples and with coverage ≥10 in knockdown samples were used.

stem region (Fig. 3d,e), which is probably resistant to bisulfite treatment. Both NSUN2-dependent sites and NSUN2-independent sites showed distinct structural preferences; they tended to be located at the 5′ end and the loop region of a stem–loop structure, respectively (Fig. 3d,e). It is known that NSUN2 methylates specific positions (C48, C49 and C50) in the vast majority of the tRNAs in humans and mice[32]. We examined the sequence and structure of these NSUN2-specific m^{5}C sites in tRNAs and found that they were also located in the 5′ end of a stem region and had a 3′ G-rich triplet motif (Fig. 3f). Furthermore, when we projected the predicted secondary structure of the mRNA m^{5}C flanking region onto a tRNA model (Methods), we found that the structure of NSUN2-dependent sites mirrors the structure of NSUN2-specific m^{5}C flanking region in tRNAs (Fig. 3g). To rule out the possibility that the m^{5}C sites with the 3′ G-rich triplet motif were derived from contaminating tRNAs that had been incorrectly mapped to the mRNA sequences, we extracted all reads that cover the mRNA m^{5}C sites and mapped them to the tRNA sequences. We found that very few reads could be mapped to the tRNA sequences (Supplementary Fig. 6). These data suggest that the primary nucleotide sequence and the secondary structure together shape the landscape of mRNA m^{5}C targets.

To evaluate the performance of our method and to carry out a fair comparison with other methods, we applied methods from others and us to the BS-seq data of HeLa NSUN2-knockdown and control samples. Compared with Amort et al.'s method[29] and Yang et al.'s method[20], our method had a much higher accuracy, reflected by the

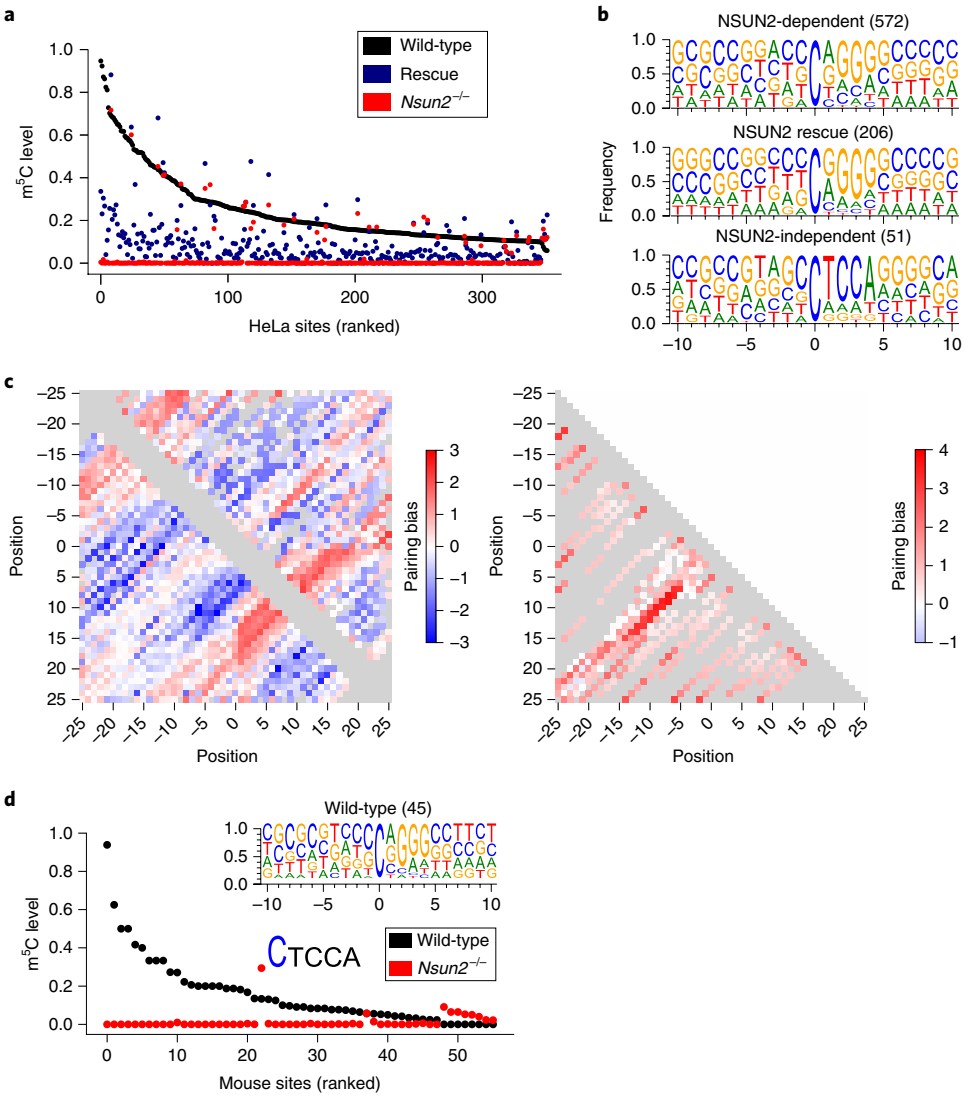

Fig. 4 | The validation of mRNA m^{5}C sites using human and mouse NSUN2-knockout models. a, m^{5}C methylation levels of sites measured from wild-type, NSUN2-knockout and NSUN2-rescue HeLa cells. A total of 352 sites that were covered by at least 20 reads in all samples and with a methylation level of ≥0.1 in either wild-type cells or NSUN2-knockout cells are shown. **b**, The sequence context of different groups of m^{5}C sites. High-confidence m^{5}C sites called from wild-type HeLa cells were used for analysis. NSUN2-dependent, sites that were not methylated (<5% methylation) in NSUN2-knockout cells; NSUN2-independent, sites that were methylated (≥5% methylation) in NSUN2-knockout cells; NSUN2 rescue, sites that were methylated (≥5% methylation) in rescue cells. **c**, Visualization of metaprofiles of the RNA secondary structure at m^{5}C sites. Three groups of sites defined in **b** were analyzed. Left: NSUN2-dependent sites (bottom diagonal) and sites rescued by transient expression of the wild-type NSUN2 construct in the knockout cells (top diagonal). Right: NSUN2-independent sites. **d**, m^{5}C methylation levels of 56 sites measured from the skin samples of wild-type and *Nsun2*$^{-/-}$ mice. Sites that were covered by at least five reads and with a methylation level of >3% in either wild-type mice or *Nsun2*$^{-/-}$ mice are shown. The 3′ flanking sequence of the site with high-level methylation in *Nsun2*$^{-/-}$ mice is shown, with the m^{5}C site highlighted in blue. The sequence context of the m^{5}C sites that were not methylated in *Nsun2*$^{-/-}$ mice is shown in the inset.

higher percentage of downregulated sites after NSUN2 knockdown (Fig. 3h). Compared with Legrand et al.'s method[28], ours identified twice as many sites with comparable accuracy (Fig. 3h). Both our method and Legrand et al.'s method[28] used a C-cutoff filter to remove non-converted C-reads, which are the major noise in BS-seq data. Legrand et al. used an arbitrary C-read cutoff (C-cutoff <3), regardless of the read length and BS-seq data quality[28]. Instead, we used the Gini coefficient to determine C-cutoff, which takes the variability of BS-seq library-construction protocol into account. Also, we provided experimental support for use of the C-cutoff filter.

To further validate that the mRNA m^{5}C sites called using our pipeline were real, we generated an NSUN2-knockout HeLa cell line via CRISPR-Cas9-induced mutagenesis. This mutant cell line has a frameshift deletion in the *Nsun2* coding sequence (CDS) region that results in an enzymatically dead truncated protein (Supplementary Fig. 7a–c). We found that 92% of the m^{5}C sites identified in wild-type HeLa cells were not methylated (m^{5}C level < 5%) in the knockout cells (Fig. 4a and Supplementary Table 4), confirming these sites are indeed NSUN2 dependent. We further performed a rescue experiment by transiently expressing wild-type NSUN2 in the knockout cells, which rescued the methylation of some sites (Fig. 4a, Supplementary Fig. 7d and Supplementary Table 4). Similar to the knockdown data above, the sites that were not methylated in the mutant or rescued by the exogenously expressed NSUN2 had the 3′ G-rich triplet motif and tended to be located at the 5′ end of a stem–loop structure (Fig. 4b,c), resembling the feature of NSUN2 tRNA

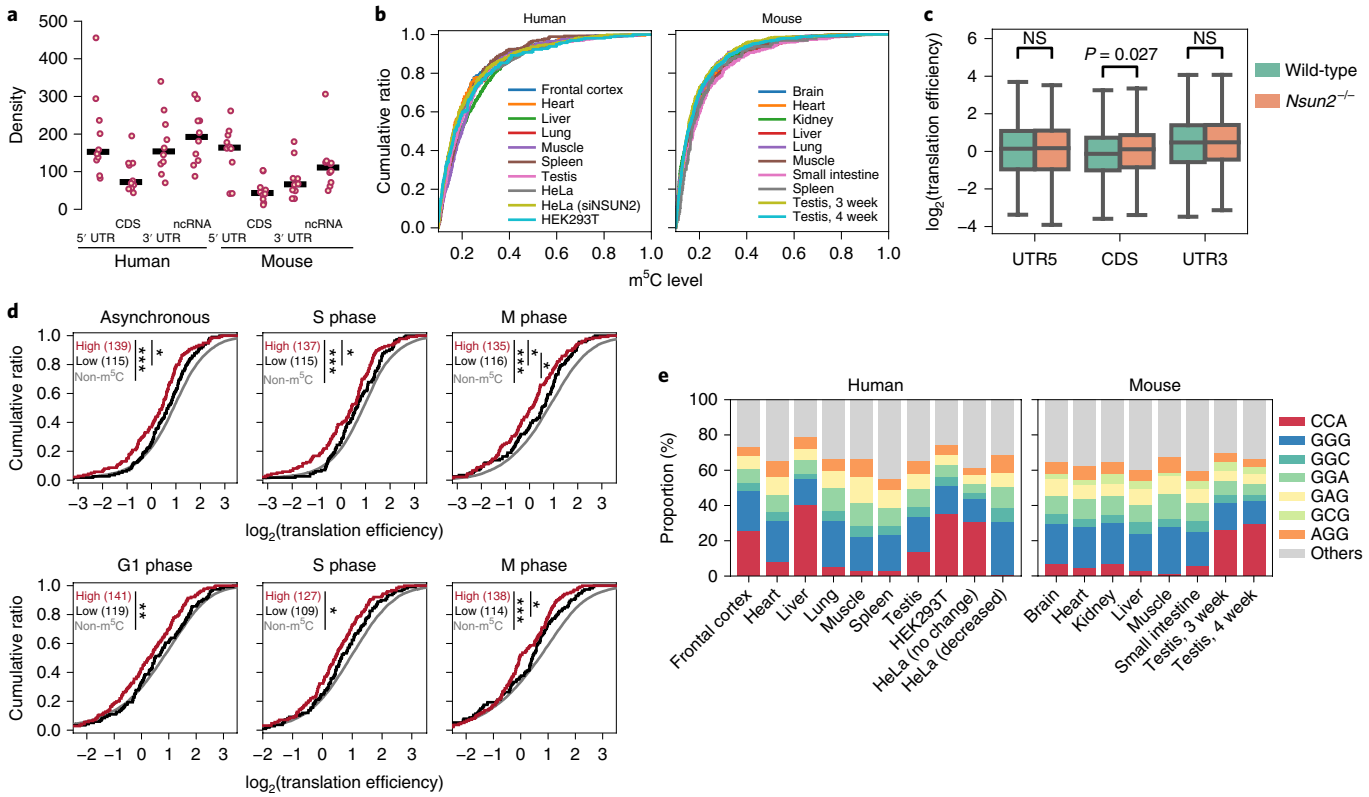

Fig. 5 | Global characterization of mRNA m^{5}C and the effect of CDS m^{5}C sites on translation. a, Densities of m^{5}C sites for each indicated genic locations. The numbers of sites per 1 Mb of sequence covered by ≥20 reads are shown. Center line represents the median. Human samples, $n=10$; mouse samples, $n=10$. **b**, The distribution of m^{5}C levels in human and mouse tissues and cells. **c**, Comparison of translation efficiency of m^{5}C-containing genes between wild-type and *Nsun2*$^{-/-}$ mice. m^{5}C-containing genes were grouped by the genic locations of the m^{5}C sites. Genes with m^{5}C sites in multiple genic locations were excluded. All mouse m^{5}C sites identified in this study were used for analysis. *P* values were calculated using two-sided Kolmogorov–Smirnov test. NS, not significant; $^*P<0.05$. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ±1.5×IQR. The number of genes analyzed (from left to right): 216, 169, 627, 511, 483 and 169. **d**, Comparison of translation efficiency between m^{5}C and non-m^{5}C genes in HeLa cells, using data from Park et al.[34] (top) or Stumpf et al.[35] (bottom). CDS sites were used for analysis. m^{5}C genes with low-level (bottom third) or high-level (top third) methylation were analyzed separately. The number of m^{5}C genes and non-m^{5}C genes are indicated. *P* values were calculated using two-sided Kolmogorov–Smirnov test. $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$. **e**, The downstream sequence preference of high-confidence m^{5}C sites in human and mouse. The downstream triplets (+2 to +4) of m^{5}C sites among human and mouse samples are shown. Only triplets that had a percentage over 5% in at least one out of ten human samples or eight mouse samples are highlighted. The tissues with >200 sites were used for analysis. The number of sites used for each tissue is shown in Supplementary Fig. 9a.

substrate. In contrast, a strong 3′ TCCA motif and loop-region preference were found in the sites that were methylated in the NSUN2-knockout cells (Fig. 4b,c). We also validated that the mouse mRNA m^{5}C sites we called were real by examining the BS-seq data from the skin samples of wild-type and NSUN2-knockout (*Nsun2*$^{-/-}$) mice[13]. Because these BS-seq libraries were not constructed using poly(A)-selected RNAs, only a small fraction of the reads was mapped to the mRNAs. Of the 56 mRNA sites that were covered and methylated in either wild-type mice or *Nsun2*$^{-/-}$ mice, most sites were methylated in wild-type mice but not in *Nsun2*$^{-/-}$ mice (Fig. 4d). Consistent with that, the NSUN2-dependent sites had a 3′ G-rich triplet motif (Fig. 4d). In contrast, a site with high-level methylation (NSUN2 independent) in *Nsun2*$^{-/-}$ mice contained a 3′ TCCA motif (Fig. 4d). These results highlight the accuracy of our approach and the motif conservation of NSUN2 mRNA targets between humans and mice.

Global characterization of mRNA m^{5}C and the effect of CDS m^{5}C sites on translation. Using our method, we next examined the landscape of RNA m^{5}C sites in mammals. We profiled BS-seq libraries constructed from seven types of human tissue and 11 types of mouse tissue. For each tissue type, at least two biological replicates were used for high-confidence site calling. The high-confidence sites shared between replicates typically had a higher methylation level, while the sites unique to one replicate had a much lower methylation level and their numbers varied among samples (Supplementary Fig. 8). In total, we compiled comprehensive lists of 3,212 (5′ untranslated region (UTR), 461 sites; CDS, 1,353 sites; 3′ UTR, 1,000 sites; and non-coding RNA (ncRNA), 398 sites) and 2,498 (5′ UTR, 411 sites; CDS, 1,080 sites; 3′ UTR, 762 sites; ncRNA, 245 sites) high-confidence exonic sites in human and mouse tissues (Supplementary Table 4). About 100–1,300 sites were identified in each tissue, partly dependent on sequencing depth (Supplementary Fig. 9a). The m^{5}C sites were distributed throughout the gene body (Supplementary Fig. 9b–c). The 5′ UTR regions showed high m^{5}C site density in both human tissues and mouse tissues, with ~200 sites per megabase pair (Mb) of covered sequence length (covered by at least 20 reads). CDS regions had the lowest density (Fig. 5a). In addition, they were not clustered together (Supplementary Fig. 9d). Overall, the m^{5}C sites spanned a wide range of methylation levels in different tissues (Fig. 5b). The median methylation level of mRNA m^{5}C sites was about 15–18%. In any given tissue or cell type, about 62–70% of the sites were lowly methylated (<20% level) and 8–10% of the sites were moderately or highly methylated (>40% level). Altogether, the maximum methylation level of 1,293 and 1,669 sites was <20%, and that of 227 and 211 sites was >40%, in human tissue and mouse tissue, respectively.

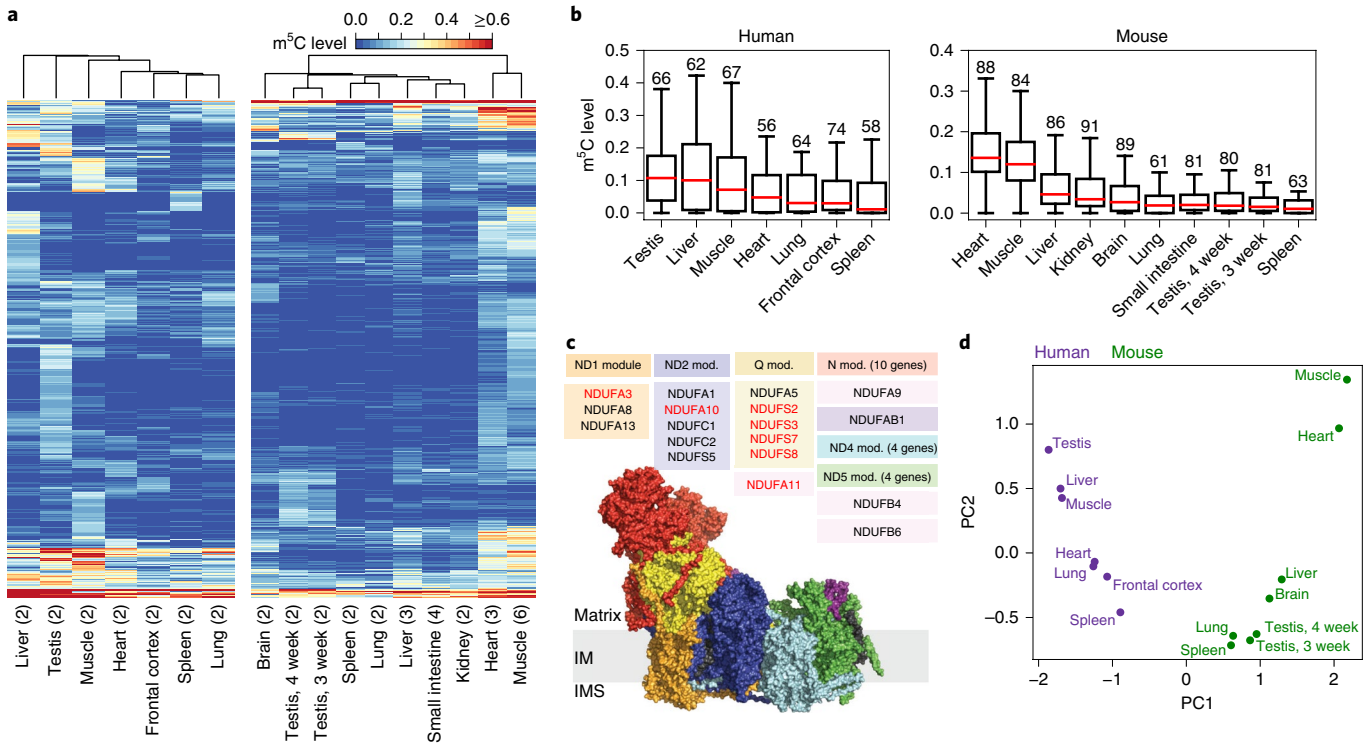

Fig. 6 | The profiles of mRNA m^{5}C in human and mouse. a, Heatmap and dendrogram of m^{5}C levels in human samples (left) and mouse samples (right). For each species, only sites with methylation levels of ≥10% in at least one sample were selected for analysis. The m^{5}C levels were determined by combining biological replicates of the same tissue type. The number of replicates that were combined is given in parenthesis. **b**, Boxplot showing the methylation levels of sites within genes that were annotated as mitochondrion-related genes by PANTHER functional classification in various human and mouse tissues. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ±1.5× IQR. The number of genes in each tissue is indicated. **c**, Subunits of mitochondrial complex I. The modular composition of mitochondrial complex I is shown. The image of the different modules and their individual subunit composition is adapted from Stroud et al.[37]. Subunits that were not within a module have been removed for clarity. IM, inner membrane; IMS, intermembrane space. Genes encoding m^{5}C-containing transcripts are highlighted in red. **d**, Principal component analysis of m^{5}C levels of various human and mouse tissues. Tissue types that were profiled in both humans and mice were selected. A total of 525 sites that were conserved between human tissues and mouse tissues were used for analysis. A conserved site was defined as a site that is a C in both the human genome and mouse genome and has methylation level of ≥10% in at least one tissue.

Despite the fact that artificially introduced m^{5}C in the bacterial mRNA CDS region can lead to a decreased yield of protein product in an in vitro–translation assay[33], its effect has not been characterized in a mammalian system in vivo. To understand the effect of m^{5}C methylation on translation in vivo, we analyzed ribosome profiling data from the skin samples of wild-type and *Nsun2*$^{-/-}$ mice[13]. We noted that genes containing m^{5}C sites in CDS regions had increased translation efficiency in *Nsun2*$^{-/-}$ mice (Fig. 5c). In contrast, genes containing m^{5}C sites in the 5′ UTR or 3′ UTR did not have such an effect (Fig. 5c). A similar observation was made using ribosome profiling data[34,35] from HeLa cells (Fig. 5d). These results together indicate that m^{5}C sites in CDS regions could negatively regulate the translation in vivo.

We also examined the sequence context of the m^{5}C flanking regions. The context of our sites was distinct from that in previous studies (Supplementary Note 8). The 3′ G-rich triplet motif was consistently observed in multiple human and mouse tissues (Fig. 5e and Supplementary Fig. 6), indicating that NSUN2 is a major mRNA methyltransferase in multiple tissues. Notably, the 3′ CCA motif was also present in multiple human and mouse tissues (Fig. 5e).

Dynamic landscape and properties of m^{5}C sites. Quantitatively, the methylation levels of m^{5}C sites vary across tissues in both human and mouse. Sites that are constitutively methylated in multiple tissues were identified (Fig. 6a and Supplementary Fig. 10a,b); sites that were methylated exclusively or preferentially in only one tissue type were also present (Fig. 6a and Supplementary Fig. 10a,b). mRNA m^{5}C methylation occurred more frequently in mouse muscle and heart than in other tissues (Fig. 6a). Genes encoding molecules with mitochondrial and transport functions were enriched with m^{5}C-containing genes in mouse muscle and heart (Supplementary Fig. 10c), as well as with a high methylation level (Fig. 6b).

For example, complex I (NADH:ubiquinone oxidoreductase) is the first enzyme of the mitochondrial respiratory chain and is composed of 45 subunits in mammals[36]. Complex I is required for the generation of a transmembrane proton gradient used for ATP synthesis. Complex I is also a major source of damaging reactive oxygen species, and its mis-regulation is associated with mitochondrial disease, Parkinson's disease and aging[36]. We found that seven (NDUFS2, NDUFS3, NDUFS7, NDUFS8, NDUFA3, NDUFA10 and NDUFA11) of the 45 complex I subunits[37] in mouse contained m^{5}C sites (Fig. 6c). In particular, four of the five members of Q-module that bridged the matrix and membrane arms were involved in transfer of electrons along Fe-S clusters to ubiquinone containing m^{5}C sites (Fig. 6c). In humans, five (NDUFB7, NDUFV3, NDUFS7, NDUFS8 and NDUFA11) of the 45 complex I subunits also contained m^{5}C sites. Another interesting gene is the gene encoding voltage-dependent anion channel 1 (VDAC1). VDAC1 is the most abundant protein on the outer membrane of mitochondria. VDAC1 is the gatekeeper for the passages of metabolites, nucleotides and ions. VDAC1 plays a crucial role in regulating apoptosis and is important not only for

metabolic functions of mitochondria but also for cell survival[38]. We found that mouse VDAC1 contains three sites that were highly methylated in muscle and heart. In summary, these data indicate that mRNA m^{5}C may be important for mitochondrial function, particularly in organs that require the most energy, such as muscles and the heart.

***Cis*-directed regulation of RNA m^{5}C across species.** To understand the evolutionary landscape of RNA m^{5}C genome wide, we compared RNA m^{5}C in human samples with that in mouse samples, focusing on sites that were conserved between the two species. Seven common tissue types profiled in both human samples and mouse samples were selected, and a total of 525 conserved sites present in these tissues were used for principal component analysis (Methods). We found that the samples were grouped by species rather than by tissue type (Fig. 6d), suggesting that *cis*-acting elements exert a greater effect on RNA m^{5}C than do *trans*-acting factors. This result parallels recent findings that RNA editing or RNA splicing is primarily *cis*-directed[39–41] and is in sharp contrast to gene expression programs, which exhibit tissue-specific signatures[39,40].

Discussion

Although, in principle, m^{5}C identification via RNA BS-seq is straightforward, in practice, the tools developed were insufficient to accurately identify m^{5}C sites, due to the uneven conversion rates across the structural mRNAs and the computational complexity of analyzing reads with reduced sequence complexity and removing false positives. Here, we propose a discovery pipeline for m^{5}C site identification. Our approach achieved high specificity by applying a high-stringency bisulfite conversion condition and developing a computational pipeline that implemented meticulous mapping and filtering steps to remove false positives. With our approach, we estimated that the average density of mRNA m^{5}C sites was about 100 site per Mb in a given tissue or cell type in mammals. Among the mRNA m^{5}C sites, more than half were methylated at a level of less than 20%, and about 10% were moderately or highly methylated at a level of more than 40%. Targeted bisulfite sequencing using a microfluidics-based multiplex PCR and sequencing, as previously described for RNA-editing quantification[42], may further allow large-scale quantification of m^{5}C modification in a cost-effective and efficient manner.

Despite recent efforts, no consensus methylation motif or structural feature was identified for mRNA m^{5}C substrates, raising doubts about the validity and biological importance of most previously reported putative mRNA m^{5}C sites. If mRNA m^{5}C does represent an important regulatory mechanism, it is probably tightly regulated and has a specific sequence feature. Using a set of high-confidence sites, we revealed that the primary sequence context and the secondary structure together determined the landscape of mRNA m^{5}C substrates. We found that NSUN2, as a major mRNA methyltransferase, targeted Cs at the 5′ end of a stem region that contained a 3′ G-rich triplet motif. The sequence and structural features of mRNA substrates were very similar to the sequence context of NSUN2-specific tRNA methylation sites. In addition, we observed another group of mRNA m^{5}C substrates that were located in the loop region of a stem–loop structure and harbored a 3′ TCCA motif. These m^{5}C sites were not the substrates of known methyltransferases, raising the possibility of the presence of one or more unknown mRNA methyltransferases that need to be further investigated.

In summary, as the impact and functional importance of individual mRNA m^{5}C sites are recently emerging, our framework to reliably identify mRNA m^{5}C sites paves the way for deeper understanding of this post-transcriptional process. In addition, the m^{5}C maps we generated in human and mouse samples provide a valuable resource to help unravel the regulation and function of mRNA m^{5}C.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41594-019-0218-x.

Received: 23 October 2018; Accepted: 27 March 2019;
Published online: 6 May 2019

References

1. Li, S. & Mason, C. E. The pivotal regulatory landscape of RNA modifications. *Annu. Rev. Genomics Hum. Genet.* **15**, 127–150 (2014).
2. Gilbert, W. V., Bell, T. A. & Schaening, C. Messenger RNA modifications: form, distribution, and function. *Science* **352**, 1408–1412 (2016).
3. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **169**, 1187–1200 (2017).
4. Machnicka, M. A. et al. MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic Acids Res.* **41**, D262–D267 (2013).
5. Grozhik, A. V. & Jaffrey, S. R. Distinguishing RNA modifications from noise in epitranscriptome maps. *Nat. Chem. Biol.* **14**, 215–225 (2018).
6. Ramaswami, G. et al. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat. Methods* **9**, 579–581 (2012).
7. Bass, B. et al. The difficult calls in RNA editing. Interviewed by H Craig Mak. *Nat. Biotechnol.* **30**, 1207–1209 (2012).
8. Bahn, J. H. et al. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* **22**, 142–150 (2012).
9. Peng, Z. et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol.* **30**, 253–260 (2012).
10. Ramaswami, G. et al. Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods* **10**, 128–132 (2013).
11. Blanco, S. & Frye, M. Role of RNA methyltransferases in tissue renewal and pathology. *Curr. Opin. Cell Biol.* **31**, 1–7 (2014).
12. Schaefer, M. et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev.* **24**, 1590–1595 (2010).
13. Blanco, S. et al. Stem cell function and stress response are controlled by protein synthesis. *Nature* **534**, 335–340 (2016).
14. Sharma, S., Yang, J., Watzinger, P., Kotter, P. & Entian, K. D. Yeast Nop2 and Rcm1 methylate C2870 and C2278 of the 25S rRNA, respectively. *Nucleic Acids Res.* **41**, 9062–9076 (2013).
15. Schosserer, M. et al. Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. *Nat. Commun.* **6**, 6158 (2015).
16. Luo, Y., Feng, J., Xu, Q., Wang, W. & Wang, X. NSun2 deficiency protects endothelium from inflammation via mRNA methylation of ICAM-1. *Circ. Res.* **118**, 944–956 (2016).
17. Li, Q. et al. NSUN2-mediated m5C methylation and METTL3/METTL14-mediated m6A methylation cooperatively enhance p21 translation. *J. Cell Biochem.* **118**, 2587–2598 (2017).
18. Shen, Q. et al. Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation. *Nature* **554**, 123–127 (2018).
19. Guallar, D. et al. RNA-dependent chromatin targeting of TET2 for endogenous retrovirus control in pluripotent stem cells. *Nat. Genet.* **50**, 443–451 (2018).
20. Yang, X. et al. 5-methylcytosine promotes mRNA export—NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Res.* **27**, 606–625 (2017).
21. Cheng, J. X. et al. RNA cytosine methylation and methyltransferases mediate chromatin organization and 5-azacytidine response and resistance in leukaemia. *Nat. Commun.* **9**, 1163 (2018).
22. Khoddami, V. & Cairns, B. R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.* **31**, 458–464 (2013).
23. Hussain, S. et al. NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.* **4**, 255–261 (2013).
24. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
25. Hussain, S., Aleksic, J., Blanco, S., Dietmann, S. & Frye, M. Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol.* **14**, 215 (2013).
26. Squires, J. E. et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **40**, 5023–5033 (2012).
27. Edelheit, S., Schwartz, S., Mumbach, M. R., Wurtzel, O. & Sorek, R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m^{5}C within archaeal mRNAs. *PLoS Genet.* **9**, e1003602 (2013).

28. Legrand, C. et al. Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. *Genome Res.* **27**, 1589–1596 (2017).
29. Amort, T. et al. Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain. *Genome Biol.* **18**, 1 (2017).
30. David, R. et al. Transcriptome-wide mapping of RNA 5-methylcytosine in *Arabidopsis* mRNAs and noncoding RNAs. *Plant Cell* **29**, 445–460 (2017).
31. Batista, PedroJ. et al. m6A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**, 707–719 (2014).
32. Blanco, S. et al. Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *EMBO J.* **33**, 2020–2039 (2014).
33. Hoernes, ThomasP. et al. Nucleotide modifications within bacterial messenger RNAs regulate their translation and are able to rewire the genetic code. *Nucleic Acids Res.* **44**, 852–862 (2016).
34. Park, J. E., Yi, H., Kim, Y., Chang, H. & Kim, V. N. Regulation of poly(A) tail and translation during the somatic cell cycle. *Mol. Cell* **62**, 462–471 (2016).
35. Stumpf, C. R., Moreno, M. V., Olshen, A. B., Taylor, B. S. & Ruggero, D. The translational landscape of the mammalian cell cycle. *Mol. Cell* **52**, 574–582 (2013).
36. Sazanov, L. A. A giant molecular proton pump: structure and mechanism of respiratory complex I. *Nat. Rev. Mol. Cell Biol.* **16**, 375 (2015).
37. Stroud, D. A. et al. Accessory subunits are integral for assembly and function of human mitochondrial complex I. *Nature* **538**, 123 (2016).
38. Colombini, M. VDAC: the channel at the interface between mitochondria and the cytosol. *Mol. Cell. Biochem.* **256**, 107–115 (2004).
39. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
40. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).
41. Tan, M. H. et al. Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249–254 (2017).
42. Zhang, R. et al. Quantifying RNA allelic ratios by microfluidic multiplex PCR and sequencing. *Nat. Methods* **11**, 51–54 (2014).

Acknowledgements

We thank J.B. Li and members of R.Z.'s laboratory for critical discussion of the project and L. Wu for manuscript editing. We thank SYSU Ecology and Evolutionary Biology Sequencing Core Facility for the sequencing service. This study was supported by grants from the National Key R&D Program of China (no. 2018YFC1003100), Guangdong Major Science and Technology Projects (no. 2017B020226002 to R.Z.), Guangdong Innovative and Entrepreneurial Research Team Program (no. 2016ZT06S638 to R.Z.) and National Natural Science Foundation of China (nos. 91631108 and 31571341 to R.Z.).

Author contributions

R.Z. conceived the project. T.H. and W.C. conducted the experiments. J.L performed the bioinformatics analysis. J.L, T.H., W.C., N.G. and R.Z. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-019-0218-x.

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

Correspondence and requests for materials should be addressed to R.Z.

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

Methods

Sample collection. Seven types of human tissue from several donors were purchased from the Chinese Brain Bank Center, including frontal cortex, heart, liver, lung, muscle, spleen and testis. These tissues were collected post-mortem from individuals with no known medical history. The informed consent of human tissue samples was obtained from the patients' families. We have complied with all relevant ethical regulation and the study protocols were approved by the Ethical Committee of Sun Yat-Sen University (SYSU). Samples were lysed and homogenized in TRIzol Reagent (Thermo) using Precellys evolution tissue homogenizer (Bertin). Total RNA was extracted using chloroform and isopropanol following the manufacturer's protocol. The quality of the total RNA was determined by agarose gel electrophoresis. For each tissue type, two biological replicates with the best RNA quality were selected for RNA BS-seq.

Male and female C57BL/6J mice 8–10 weeks of age were purchased from the Laboratory Animal Center of SYSU. All experiments were performed at the Animal Center of SYSU, in accordance with the Guide for the Care and Use of Laboratory Animals. The study protocols were approved by the Ethical Committee of SYSU.

Cell culture. The HEK293T and HeLa cell lines were obtained from Cell Bank of Type Culture Collection of Chinese Academy of Sciences (CBTCCCAS). Both cell lines have been identity verified using short tandem repeat analysis and determined to be free from mycoplasma contamination by CBTCCCAS. Cell lines were maintained in DMEM (Gibco) supplemented with 10% FBS (HyClone).

Construction of NSUN2-knockout and rescue cells. NSUN2-knockout HeLa cells were generated via CRISPR-Cas9-induced mutagenesis. In brief, a single guide RNA (sgRNA) sequence (GAGCTCAAGATCGTGCCCGA) was designed using CRISPR-ERA (http://CRISPR-ERA.stanford.edu). The sgRNA template oligonucleotide was synthesized and cloned into lentiCRISPR v.2 plasmid (Addgene no. 52961). The plasmid was transfected into HeLa cells using Lipofectamine 3000 (Thermo) following the manufacturer's instructions. Transfected cells were selected using puromycin (1 μg ml^{-1}, Sigma). A mutant clone that produced a 1-nt frameshift at position 177 of the coding sequence, leading to a premature stop codon at amino acid position 65, was selected for the experiments. The loss of NSUN2 protein expression was verified with NSUN2 Antibody (Proteintech, 20854) by western blot.

To construct NSUN2 expression plasmid, complementary DNA from HeLa cells was reverse transcribed using HiScript II Q Select RT SuperMix for quantitative PCR (qPCR) (Vazyme) and full-length NSUN2 CDS fragment was amplified with the following primers: 5′-TGTCCTAGGATGGGGCGGCGGTCGC-3′ (forward) and 5′-GATACCGGTTCACCGGGGTGGATGGACC-3′ (reverse). The NSUN2 fragment was inserted into the AvrII and AgeI sites of the pCDH-3xFLAG vector to generate pCDH-3xFLAG-NSUN2 plasmid.

For the rescue experiment, NSUN2-knockout cells were plated in a six-well plate. Then, 3 μg of pCDH-3xFLAG-NSUN2 plasmid was transfected using Lipofectamine 3000 (Thermo) following the manufacturer's instructions. After 24 h, a second transfection was performed in the same way. Then, 48 h after the second transfection, proteins of the cells were collected to confirm the expression of NSUN2 with mouse anti-FLAG antibody (Sigma, F1804) by western blot. Total RNA of the cells was purified with TRIzol Reagent (Thermo) for RNA BS-seq.

mRNA preparation. Total RNA was isolated from tissues or cultured cells using chloroform and isopropanol following the manufacturer's protocol. Polyadenylated RNA was separated from total RNA using either GenElute mRNA miniprepKit (Sigma-Aldrich) or Oligo dT Magnetic Beads (Vazyme).

RNA spike-in controls. ERCC RNA mixes (Thermo) were used as the external RNA control for RNA BS-seq. For each sample, 5 μl 1:50 diluted ERCC RNA mixes were added before bisulfite treatment.

To quantify m^{5}C level on in vitro–transcribed transcripts, we generated m^{5}C and non-m^{5}C transcripts. An artificial dsDNA template (Oligo-5C/m^{5}C) with 5Cs was synthesized by the Synbio Tech company. In vitro transcription with cytosine or 5-methylcytosine was performed following the manufacturer's protocol (Thermo, TranscriptAid T7 High Yield Transcription Kit). m5CTP was purchased from TriLink BioTechnologies, LLC. DNase I (Thermo) treatment was performed after transcription. RNA was recovered by the RNA Clean and Concentrator kit (Zymo Research). m^{5}C and non-m^{5}C RNAs were mixed at different ratios, then 100 pg in vitro–transcribed transcripts were spiked into 1 μg mRNA for BS-seq. The sense strand sequence of the double-stranded DNA (dsDNA) template is listed in Supplementary Table 5.

For the dot blot, m^{5}C RIP rtPCR, m^{6}A RIP rtPCR and m^{5}C RIP-seq experiments, the transcripts containing unmodified nucleotide, 6-methyladenine or 5-methylcytosine were transcribed in vitro from the dsDNA templates using TranscriptAid T7 High Yield Transcription Kit (Thermo). m6ATP and m5CTP were purchased from TriLink BioTechnologies, LLC. All sense strand sequences of the dsDNA templates are listed in Supplementary Table 5.

Dot blot. The in vitro–transcribed RNAs containing either m^{5}C or C were denatured at 65 °C for 5 min and then spotted onto a nylon membrane (GE Healthcare). RNA was fixed onto the membrane by cross-linking in a UV Stratalinker at 200 mJ. After blocking with 5% BSA in Tris-buffered saline and Tween 20 buffer, the membrane was incubated with mouse anti-m^{5}C monoclonal antibody (Diagenode, no. 15200003, 1:1,000) overnight at 4 °C. The membrane was then washed with 1× Tris-buffered saline and Tween 20, followed by incubation with HRP-conjugated goat anti-mouse monoclonal antibody (CST) at 1:10,000 for 1 h at room temperature. The membrane was again washed and developed with ECL (Millipore). Loading was assessed by methylene blue (Sigma, M9140) staining of the membrane.

RIP rtPCR and RIP-seq. m^{5}C RIP was performed as previously described[43] with some modifications. In brief, polyadenylated mRNA was first fragmented with Magnesium RNA Fragmentation Module (NEB, E6150S) at 94 °C for 5 min and then cleaned up using ethanol precipitation. Then, 100 ng RNA was saved to serve as the input control and the rest was incubated with 5 μg anti-m^{5}C antibody in immunoprecipitation buffer (150 mM NaCl, 0.1% Igepal CA-630, 10 mM Tris-HCl, pH 7.4) overnight at 4 °C. 60 μl protein G magnetic beads (Thermo, 1004D) was pre-blocked with 0.5 mg ml^{-1} BSA at 4 °C for 2 h and then added for immunoprecipitation to m^{5}C-Ab mixture at 4 °C for another 2 h. After washing with immunoprecipitation buffer three times, RNA was competitively eluted from the beads with 20 mM 5-methylcytosine hydrochloride (Sigma-Aldrich, M6751), followed by ethanol precipitation. RNA was resuspended in 8 μl water and used for m^{5}C RIP rtPCR or m^{5}C RIP-seq.

For m^{5}C RIP rtPCR, the in vitro–transcribed m^{5}C and non-m^{5}C transcripts (Supplementary Table 5) were spiked into the fragmented RNA before immunoprecipitation. Immunoprecipitation and input RNAs were reverse-transcribed using HiScript II Q Select RT SuperMix for qPCR (Vazyme) with gene-specific primers, followed by rtPCR using ChamQ SYBR qPCR Master Mix (Vazyme). The primers used for reverse transcription and rtPCR were listed in Supplementary Table 5. The $2^{-\triangle\triangle Ct}$ method was used to determine the enrichment of in vitro–transcribed m^{5}C transcripts relative to the non-m^{5}C transcripts[44].

m^{6}A RIP rtPCR was performed the same as for m^{5}C, except rabbit anti-m6A polyclonal antibody (SYSY, 202003) was used for immunoprecipitation and N6-methyladenosine (Sigma-Aldrich, M2780) was used for competitive elution. All sequences of dsDNA template used for synthesizing in vitro–transcribed transcripts are listed in Supplementary Table 5.

For m^{5}C RIP-seq, the VAHTS Stranded mRNA-sequencing (mRNA-seq) library prep kit (Vazyme) was used for library construction of both immunoprecipitated and input RNAs.

m^{5}C RIP-seq peak calling. Adapters were first trimmed by cutadapt[45] (-e 0.1 -m 30 -q 20). Ribosomal RNA reads were then removed with SortMeRNA-2.1 (ref. [46]). Clean reads were mapped to the reference genome (GRCh37) with Tophat2 (ref. [47]) (–library-type fr-firststrand).

m^{5}C peak analysis was conducted as previously described[31]. In brief, each gene was split into the 100-nt sliding windows, and an enrichment fold (winscore) was calculated for each window.

$$\text{winscore} = \log_2\left(\frac{\text{MeanWinIP}/\text{MedianGeneIP}}{\text{MeanWinControl}/\text{MedianGeneControl}}\right)$$

MeanWinIP and MeanWinControl are the mean coverage for each window for immunoprecipitation and input control, respectively. MedianGeneIP and MedianGeneControl are gene median coverages for immunoprecipitation and input control, respectively.

RNA BS-seq library construction. Bisulfite treatment was performed using the EZ RNA methylation kit (Zymo Research) with some modifications. In brief, 1 μg of polyadenylated RNA was converted using three different conversion conditions (Supplementary Table 1). The quantity of the converted RNA was determined by Qubit. Libraries were then constructed using VAHTS Stranded mRNA-seq Library Prep Kit (Vazyme). In short, 20 ng converted RNA was fragmented into ~150- to 200-nt fragments by incubation at 94 °C for 3 min in 2× frag/prime buffer. The fragmented RNA was then used for library construction, following the manufacturer's protocol. Libraries were sequenced in GENEWIZ. In brief, libraries first underwent quality control assessment using Agilent TapeStation. The libraries that passed quality control were sequenced on HiSeq X (Illumina) to produce paired-end 150-bp reads. All libraries are summarized in Supplementary Table 2.

ERCC coverage calculation. To calculate the normalized coverage of each ERCC mix, the following formula was used:

$$\text{Normalized coverage} = \frac{N_{\text{ERCC coverage}}}{N_{\text{Observed}} \times N_{\text{Total ERCC coverage}}} \times 10^6$$

$N_{\text{ERCC coverage}}$ denotes the sum of the coverages of each base in a given ERCC transcript. N_{Observed} denotes the detected length of ERCC transcript. $N_{\text{Total ERCC coverage}}$ denotes the library size.

Performance test for mappers. Simulated reads were generated by the R package polyester[48] based on the GRCm38.87 transcriptome with the following settings: 250-bp fragment size (s.d. of 25 bp), error rate 0.001 and 50× coverage. About 5 million stranded 50-bp paired-end reads and 5 million stranded 100-bp paired-end reads were generated for analysis. Different mappers were used to map the simulated reads into both the C-to-T and G-to-A converted reference genomes and/or transcriptome. Alignment results were examined with the original genomic coordinates of the simulated reads.

Mapping of RNA BS-seq reads. For stranded paired-end reads (BS-seq data generated in this study for m^{5}C site calling), we trimmed adapters, the first 10 bp of the reads, the last 6 bp of the reads and the low-quality bases using cutadapt[45] (-e 0.25 -q 25 -trim-n) and Trimmomatic[49]. Clean read pairs were mapped to both the C-to-T and G-to-A converted reference genomes (GRCh37 or GRCm38) by HISAT2 (ref. [50]) (-k 10,–fr,–rna-strandness FR,–no-mixed). Uniquely mapped reads were extracted and stored as a BAM file (named BAM.genome) for further analysis. Unmapped and multiple mapped reads were further mapped to a C-to-T converted transcriptome (in combination with spike-in sequences) by Bowtie2 (ref. [51]) (-end-to-end,–fr,–gbar 5,–mp 5, -k 10, -R 2, -D 5). Only read pairs mapped to a single gene were further considered. If a read was mapped to multiple isoforms of a gene, the one mapped to the longest isoform was selected. For the read pairs that had only one read mapped to a transcript, if the forward read was uniquely mapped to the 3′ end of a transcript and the reverse read had >80% adenines, the forward read was retained. Reads mapped to the transcriptome were stored as a BAM file (named BAM.transcriptome) and the transcript coordinates were then liftovered to the genomic coordinates according to Ensembl GTF annotation. Last, BAM.genome and BAM.transcriptome were merged for further analysis.

For stranded single-end reads generated by Amort et al.[29], reads were preprocessed and mapped by HISAT2 (ref. [50]) (-k 10,–rna-strandness F) and Bowtie2 (ref. [51]) (-end-to-end,–norc,–gbar 5,–mp 5, -k 10, -R 2, -D 5) with parameters specified for single-end reads.

m^{5}C site calling and false-positive filtering. After the merging of BAM files, the reads were used to detect mismatches that may be putative m^{5}C sites. We inferred the mismatch type of each site on the basis of the strand of overlapping annotated genes. We inspected all positions with C-to-T mismatches and only took variant positions into consideration if they conformed to our requirements for number, frequency and quality of bases that vary from the converted reference sequences. We specifically required that each variant was supported by three or more variant nucleotides having a base quality score of ≥30, a mismatch frequency of ≥0.1 and coverage of C + T ≥ 20. Furthermore, we required that (1) the variant still satisfied this criteria after the removal of the overlapped C-reads on the basis of the Gini coefficient determined C-cutoff filter, (2) the signal ratio of the variant was ≥0.9, (3) the variant was not located at conversion-resistant genes and (4) the *P* value calculated using a one-sided binomial test on the basis of the gene-specific conversion rate was <0.001. Finally, to determine the set of high-confidence sites in a specific tissue or cell type, we required the presence of a site in the biological replicates. A combined *P* value was calculated by Stouffer's *Z*-score method. Sites with a combined *P* value of <0.001 were considered to be high-confidence m^{5}C sites. If a site was present in only one of the replicates but not others due to the coverage issue, we further required that at least five variant nucleotides in that sample achieve high specificity. To identify high-confidence sites in human, 14 libraries from this study and 4 HeLa cell libraries from Yang et al.[20] were used. To identify high-confidence sites in mouse, 12 libraries from this study and 16 libraries from Yang et al.[20] were combined for analysis.

The mismatch frequency is defined as the total number of reads with C as compared to all reads with C or T. C-content is defined as the number of Cs in a given read. A C-cutoff means the use of C-content cutoff to remove reads with multiple Cs due to conversion failure. The signal ratio is defined as the signal (the number of reads with C-content ≤ C-cutoff) divided by the total number of reads covered. The gene-specific conversion rate is defined as the coverage of Ts divided by the sum of coverages of Cs and Ts in all C positions in a gene. Only genes whose coverages in reference C positions was ≥1,000 and conversion rates >0.95 were used for analysis. The remaining genes were considered to be conversion-resistant genes, and the sites in these genes were removed.

Cluster status evaluation. The transcriptomic coordinate of each m^{5}C site was determined on the basis of GTF annotation. Intergenic and intronic sites were excluded from further analysis. For sites on the same transcript, a hierarchy linkage clustering on site coordinates was done with scipy ('single' method, known as the 'nearest point algorithm'). In brief, hierarchical clustering was performed on the condensed m^{5}C distance matrix 'X' first (*Z* = scipy.cluster.hierarchy.linkage(X)). Then, flat clusters were formed from the hierarchical clustering defined by the linkage matrix X (cluster = scipy.cluster.hierarchy.fcluster(*Z*, 50, criterion = 'distance')). The threshold applied to form flat clusters was set to 50. Cluster degree was defined as the number of putative m^{5}C sites in a given cluster. To reveal the pattern of m^{5}C cluster, we binned the putative m^{5}C sites by cluster degree (1, 2–4, 5–9, 10–20 and >20).

m^{5}C site calling using other methods. For Yang et al.'s method[20], sites called from the original study were used for analysis directly.

For Amort et al.'s method[29], sites in control samples were called with the same parameters used in the original studies. In brief, meRanTK-1.20 was used in alignment and site calling. During alignment, default parameters of meRanGs were used. Overlaps of paired-end reads were clipped with bamUtil to avoid duplications. Sites in controls were called by meRanCall ($Q \geq 30$, coverage ≥ 10, m^{5}C level ≥ 0.2 and false-discovery rate < 0.01). The conversion rates were estimated by meRanCall. Only sites annotated 'T-to-C mutated' ('refBase' is C and 'CalledBase' is C, T and CT) were processed to structure folding filter: the longest isoform in the Ensembl gene model was folded with RNAfold in VennaRNA-2.2.8 (–MEA 0.1,–maxBPspan 150, -T 70); the structure of the flanking 300 nt region was used if maximum expected accuracy (MEA) folding failed. Only sites presented in both replicates were used.

For Legrand et al.'s method[28], sites in control samples were called with the same parameters used in the original studies. In brief, clean reads were mapped to the Ensembl transcriptome (mRNA and ncRNA) as well as tRNAs from GtRNAdb[52] and rRNAs from SILVA[53]. Reads were piled up and a C-cutoff of 2 was applied to remove false positive reads. Pileups were processed to R package 'BisRNA'. For each sample, the lambda parameter was estimated using the RNAmeth.poisson.test. *P* values were adjusted for multiple tests using the Benjamini–Hochberg method. Only sites with coverage of ≥20, an m^{5}C level of ≥0.1 in both replicates and a combined *P* value of <0.05 were used.

Motif discovery. For m^{5}C sites identified by BS-seq, the flanking sequence of each site was extracted from the reference genome. Motif logos were plotted with WebLogo v.3.5 (ref. [54]).

RNA secondary structure prediction. The upstream and downstream 25 bp sequences of the m^{5}C sites were extracted from the genome and folded with the RNAfold tool in the ViennaRNA Package[55]. The Python API of VennaRNA-2.4.2 with default parameters was used: RNA.fold function for RNA structure prediction and RNA.ptable function for base-pairing event parsing.

Cross-species position conversion. We converted the coordinates of sites between human and mouse using the LiftOver tool (http://genome.ucsc.edu). For positions that were successfully lifted over we determined the nucleotide using the pairwise alignments in axt format from the University of California, Santa Cruz Genome Browser.

m^{5}C level call from BS-seq data of wild-type and *Nsun2*$^{-/-}$ mice. BS-seq libraries from Blanco et al. were used for analysis[13]. Because these BS-seq libraries were not constructed using poly(A)-selected RNAs, no sufficient reads can be used to call m^{5}C sites in mRNAs de novo. Therefore, we applied the following pipeline to obtain m^{5}C levels for known mouse sites identified using the mouse tissue samples we studied. In brief, we first preprocessed the FASTQ files before mapping by trimming the adapters and the first 5 bp of the reads, filtering low-quality bases, discarding reads <35 bp. Clean reads were mapped to the genome and subsequently mapped to the transcriptome as we described above. All replicates were merged for m^{5}C level call. A C-cutoff of four was used to obtain m^{5}C levels of known mouse sites.

Translation efficiency calculation. Translation efficiency of each gene was calculated as previously described[34], with some modifications. In brief, we first preprocessed the FASTQ files before mapping by trimming the adapters, filtering low-quality bases, discarding reads <20 nt and shortening reads to 26 nt (cutadapt -m 20 -l 26 -q 25). Clean reads were then mapped to rRNA sequences by Bowtie2 (ref. [51]) (-N 1 -L 20–norc) to eliminate rRNA contamination. rRNA sequences were downloaded from SILVA database (https://www.arb-silva.de/). The remaining reads were then mapped to the reference genome and transcriptome (GRCh37 or GRCm38) by Tophat v.2.1.1 (ref. [47]) (-M -N 1). To avoid ribosome stalling around start codon, the first 16 codons were discarded and only open reading frames with ≥40 codons were used[56]. Only uniquely mapped reads were selected and processed to HTseq-count for feature counting. All replicates were merged for analysis.

To quantify mRNA or ribosome-protected mRNA fragment abundance, genes with ≥60 mapped reads were selected and normalized using the trimmed mean of *M* values method implemented in the edgeR Bioconductor package[57]. Translation efficiency was calculated by dividing the trimmed mean of *M* value–normalized ribosome-protected mRNA fragment value with that of gene expression value. Only genes with ≥60 mapped reads in the mRNA-seq data were used, and averaged translation efficiencies were obtained.

Small interfering RNA (siRNA) knockdown, rtPCR and western blot. siRNAs were transfected using Lipofectamine RNAiMAX (Thermo) following the manufacturer's instructiona. Cells were collected 48 h after transfection. For rtPCR, total RNA was reverse transcribed with Oligo dT primer, followed by rtPCR with primers listed in Supplementary Table 5. For western blot, antibodies to NSUN2 (Proteintech, 20854), NSUN4 (Abcam, ab101625), NSUN6 (Proteintech, 17240) and β-actin (Zsbio, TA-09) were used.

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

Data availability

The sequence data have been deposited in the NCBI GEO database under the accession code GSE122260. All other data are available from the corresponding author upon reasonable request.

Code availability

All relevant code and data processing pipelines have been deposited in GitHub (https://github.com/SYSU-zhanglab/RNA-m5C).

References

43. Dominissini, D. et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201–206 (2012).
44. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative CT method. *Nat. Protocols* **3**, 1101 (2008).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
46. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
47. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
48. Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).
49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
53. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
54. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
55. Lorenz, R. et al. ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
56. Artieri, C. G. & Fraser, H. B. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.* **24**, 2011–2021 (2014).
57. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

nature research

Corresponding author(s): Rui Zhang

Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All computer code and softwares used to collect the data were described in details in the Methods section.
Data analysis	All computer code and softwares used to analyze the data were described in details in the Methods section.

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

Data

Policy information about availability of data

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

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

The sequence data have been deposited in the NCBI GEO database under the accession code GSE122260.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sample size was chosen based on material available ensuring that it will be appropriate for statistical analysis.
Data exclusions	no data exclusion
Replication	For each tissue type, at least two biological replicates were used for high-confidence m5C site calling.
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	The antibodies used are anti-m5C antibodies (Diagenode, C15200003; Zymo Research, A3001; Abcam, ab10805), anti-m6A polyclonal antibody (SYSY, 202003), anti-FLAG antibody (Sigma, F1804) and NSUN2 Antibody (Proteintech, 20854).
Validation	The antibodies have been validated by the respective vendors.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The HEK293T and HeLa cell lines were obtained from Cell Bank of Type Culture Collection of Chinese Academy of Sciences (CBTCCCAS).
Authentication	Both cell lines have been identity verified using STR analysis by CBTCCCAS.
Mycoplasma contamination	Both cell lines have been checked for mycoplasma contamination by CBTCCCAS.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	8- to 10-weeks-old C57BL/6J mice
Wild animals	n/a
Field-collected samples	n/a

Human research participants

Policy information about studies involving human research participants

Population characteristics	Adult Chinese males (age from 39 to 57) with no known medical history.
Recruitment	n/a