Investigating Differential m6A Regulation: A Comparative Analysis of Long Non-Coding and Coding RNAs

Feiyue Yang[a], **Chase Armer**[a], **Claudia Aiello**[a]
[a]Columbia University, Deep Sequencing Fall 2023

Abstract. This study investigates the differential m6A methylation patterns and their regulatory impacts on long non-coding RNAs (lncRNAs) and coding RNAs in mouse embryonic stem cells. The primary focus is on the role of m6A methyltransferases (writers), particularly METTL3, and m6A-binding proteins (readers), specifically YTHDF proteins, in modulating RNA stability and expression. Our results reveal that knockout of METTL3 leads to a relative downregulation of lncRNAs compared to protein-coding RNAs, with a notable variance in expression based on the number of m6A sites present in the genes. This pattern is consistent across different m6A regulatory proteins, as evidenced by parallel observations in YTHDF1/2/3 triple-knockout conditions. We also explore the m6A reading strategy differences between coding RNAs and lncRNAs, employing integrated analysis of differential expression in METTL3-knockout and YTHDF1/2/3-triple-knockout datasets. Our findings suggest that while there is a similarity in the total m6A reading manners of YTHDF proteins between coding RNAs and lncRNAs, individual YTHDF proteins exhibit distinct non-redundant reading abilities, especially in lncRNAs. The study also delves into the complex relationship between m6A modifications and the transcriptional resilience of genes, with a particular emphasis on housekeeping genes. Our data indicate that m6A modifications impart a nuanced influence on gene expression fidelity, suggesting the involvement of additional regulatory mechanisms beyond m6A writers and readers. This research provides novel insights into the epitranscriptomic regulation of gene expression and highlights the intricate interplay between m6A modifications and RNA stability in both coding and non-coding RNA landscapes.

Keywords: m6A, METLL3, YTHDF, long non-coding RNA, coding RNA.

1 Introduction

N6-methyladenosine (m6A), a methyl modification of adenosine, is the most abundant chemically modified nucleotide. It plays a key role in many physiological and pathological processes, such as acute myeloid leukemia(AML), primarily by regulating RNA stability.[1] m6A modifications are deposited co-transcriptionally onto transcripts by m6A methyltransferases (writers) in the nucleus, some of which can be subsequently removed by the demethylases (erasers). Once exported to the cytoplasm, m6A is recognized by cytoplasmic m6A-binding proteins (readers) including YTHDF1, YTHDF2, and YTHDF3, which all promote mRNA degradation.[2] While extensive discoveries in transcripts of protein coding genes have laid a foundation for understanding the basics of m6A regulation, accumulating researches have also indicated m6A's presence and essential role on non-coding RNAs.[3–7] However, whether m6A regulation patterns differ between coding and non-coding RNAs and the underlying mechanisms remain underexplored.

2 Differential impact of METTL3, an m6A writer, on long non-coding and coding RNA

In the present study, we first sought to determine differences in the regulation patterns of long non-coding RNA (lnc-RNA) and coding RNA by m6A. m6A modifications are regulated by several proteins known as writers, readers, and erasers.[8] m6a writers function to methylate adenosine and include proteins such as Mettl3, Mettl14, WTAP, and more.[8] In this study, we chose to focus specifically on the m6A writer Mettl3 since it's the most crucial part of the m6A writer complex.

From the resultant data, it was determined that knockout of Mettl3, a m6A writer, led to a relative downregulation of lnc-RNA compared to protein coding RNA (Fig 1.a). Given these results, we were prompted to explore how variances in the number of m6A sites in a given gene would influence the expression of the two RNA species of interest.

To explore the question outlined above, genes were annotated according to the presence of 0, 1, 2, 2-4, or 5+ m6A sites. The resultant graphs allowed for the visualization of the impact of m6A variation on

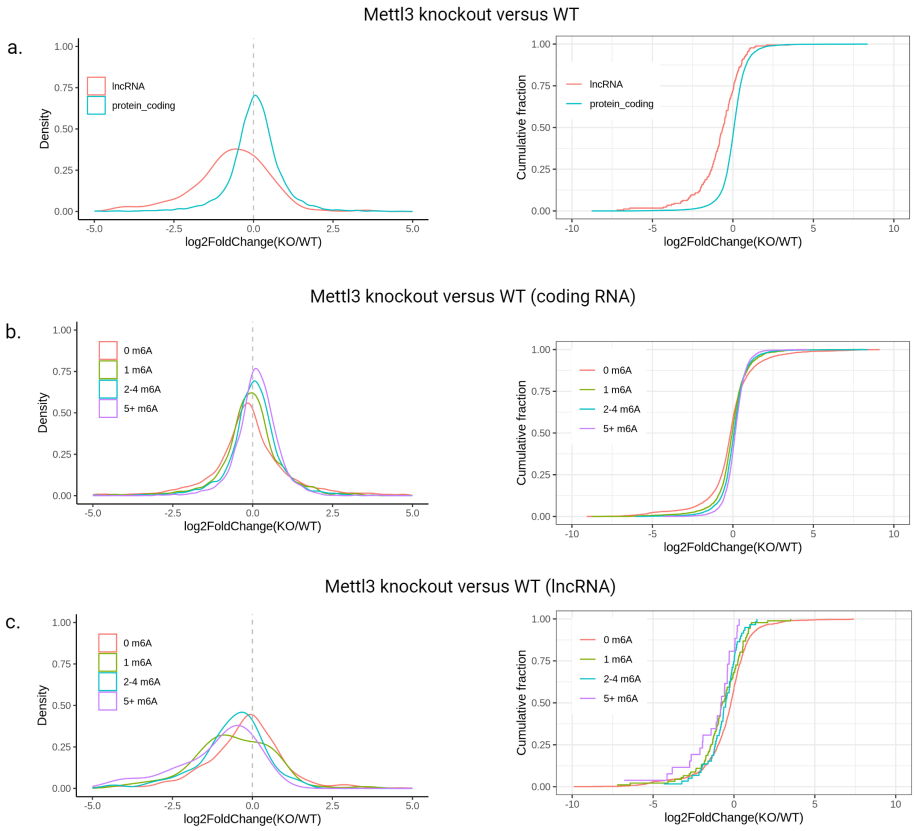

Fig 1 The differential impact of m6A reader knockout on various RNA species. The log fold change in expression of genes of interest were compared between long non-coding and coding RNAs (a) and plotted as a density curve. The log fold change in expression of these genes were then compared for coding (b) long non-coding (b) RNAs separately with varying numbers of m6A sites present (0, 1, 2-4, 5+).

the regulation patterns of lnc-RNA and coding RNA. It was found that among protein coding genes, those with more m6A sites were upregulated relative to genes with less m6A sites upon knockout of Mettl3 (Fig 1.b) Interestingly, the opposite trend was observed when m6A variation was plotted for lnc-RNA. This data demonstrated that lnc genes with fewer m6A sites were upregulated relative to lnc genes with more m6A sites upon Mettl3 knockout (Fig 1.c). These results indicate that m6A regulation differs depending on whether a given RNA is coding or non-coding, and that the differential regulation pattern may be dependent upon the number of m6A sites in a particular gene.

3 Differential impact of YTHDF, an m6A reader, on long non-coding and coding RNA

Given the results obtained from knockout of the m6A writer, Mettl3, we wondered whether the differential regulation patterns observed between lnc-RNA and coding RNA was a phenomenon specific to Mettl3, or represented a trend associated with m6A as a whole. As mentioned previously, m6A modifications are regulated by several proteins, therefore we next chose to query m6A readers to see if data patterns were maintained between writer and reader knockout data. Readers, as their name describes, function to read methylated proteins and are specific to m6A.[8] Several readers exist for m6A, however the readers available in our perturbation dataset were YTHDF1/2/3, and as such were the focus of the subsequent investigation.

The differences in m6A regulation patterns in lncRNA versus coding observed in the Mettl3 knockout were strikingly maintained in the YTHDF1/2/3 dataset. This alignment was in line with our expectations as knockout of either the writer or reader should impair m6A modifications and therefore have the

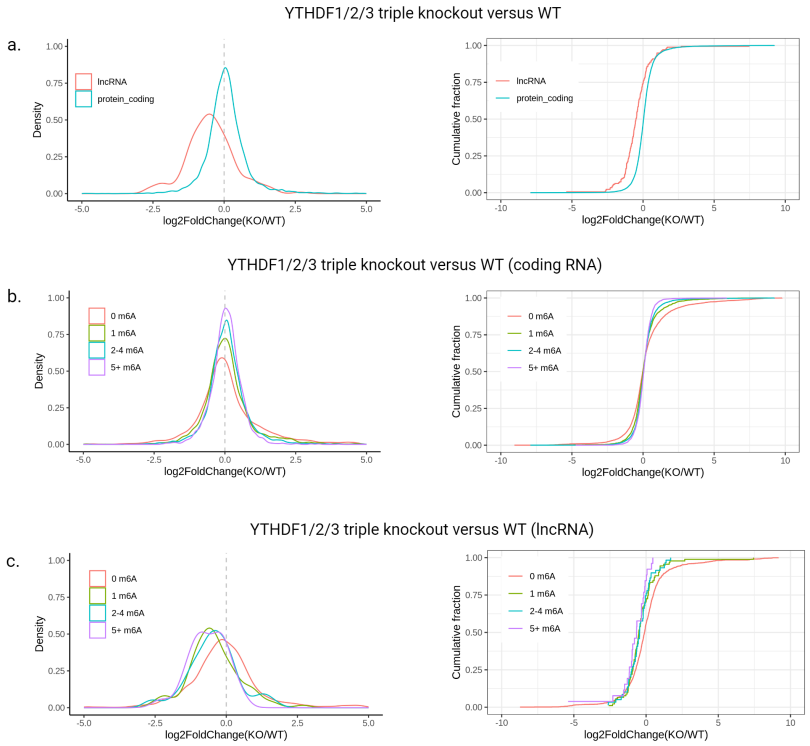

Fig 2 The differential impact of m6A writer (YTHDF) knockout on various RNA species. The log fold change in expression of genes of interest were compared between long non-coding and coding RNAs (a) and plotted as a density curve. The log fold change in expression of these genes were then compared for coding (b) long non-coding (b) RNAs separately with varying numbers of m6A sites present (0, 1, 2-4, 5+).

same impact. It was observed that knockout of YTHDF1/2/3 led to a relative downregulation of lnc-RNA compared to protein coding RNA (Fig 2.a), protein coding genes with more m6A sites were upregulated relative to genes with less m6A sites upon YTHDF knockout (Fig 2.b), and lnc-RNA genes with more m6A sites were downregulated compared to those with less m6A sites upon YTHDF knockout (Fig 2.c). An additional observation made from these results is that protein coding RNAs appear more tightly regulated by m6A than lnc-RNAs due to their lower relative log fold change (Fig 2.a), and that the degree of regulation correlates is positively correlated with the number of m6A sites (Fig 2.b).

These data indicate that the differential regulation of lncRNA and coding RNA by m6A is a direct effect of m6A modifications rather than an alternative effect mediated by knockout of the reader or writer proteins specifically. One of the major roles of the m6A gene is promotion of transcript degradation in the cytoplasm.[2] Based on this knowledge, we would expect genes with more m6A sites to undergo a relative increase in gene expression when regulatory proteins such as Mettl3 and YDHTF are nonfunctional and unable to enact the known function of m6A - transcript degradation. This is the exact trend that was observed in the coding RNA data, but the opposite trend than what was observed in the lnc-RNA data, indicating that m6A may regulate lnc-RNA in a non-canonical fashion with potentially interesting biological implications.

4 Integrated analysis of m6A modifiers

With different m6A regulation patterns observed between coding RNAs and lncRNAs in both Mettl3-knockout and YTHDF1/2/3-triple-knockout conditions, we aimed to explore whether such difference roots in distinct m6A reading strategy.

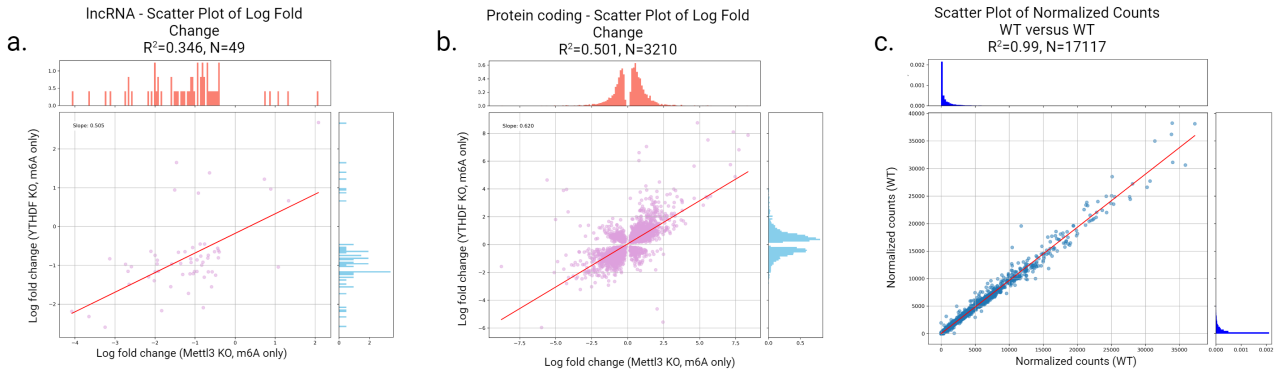

Fig 3 Integrated analysis of m6A writer and readers knockout datasets. The expression fold change between knockout and wild-type in Mettl3-knockout and YTHDF1/2/3-triple-knockout conditions were compared, with (a) lncRNA and (b) coding RNAs separately plotted in scatterplots. (c) The correlation between DESeq2 normalized read counts of two wild-type replicates.

Previous studies on YTHDF proteins' m6A reading functions have focused primarily on protein coding genes. Whether they function the same on lncRNAs remains underexplored. Therefore, we hypothesized that m6As on lncRNAs interact with YTHDF proteins in a different manner compared with coding RNAs. To test this hypothesis, we incorporated differential expression analysis results from Mettl3-knockout and YTHDF1/2/3-triple-knockout datasets.

In this integrated analysis, only genes differentially expressed between both knockout conditions and wild-type with a significance of FDR lower than 0.05 are considered. In addition, genes with no m6A sites were filtered out so that only those directly impacted by m6A were analyzed. The expression fold change between knockout and wild-type in Mettl3-knockout and YTHDF1/2/3-triple-knockout conditions were subsequently plotted and compared in the scatterplots with regression analysis shown below (Fig 3.a, Fig 3.b).

According to our results (Fig 3.a, Fig 3.b), the R^2 between the two conditions in coding RNAs and lncRNAs are similarly low between 0.3-0.5, and the slopes of the regression lines have only minor difference (R^2_{coding} = 0.501, R^2_{lnc} = 0.346, $slope_{coding}$ = 0.620, $slope_{lnc}$ = 505). The fact that the R^2 and regression slopes are far from 1 suggest that there might be other m6A readers or writers playing an essential role in m6A regulation of both coding RNAs and lncRNAs, which complicates the correlation between METTL3-knockout and YTHDF1/2/3-triple-knockout. The decrease in the R^2 in lncRNAs compared to coding RNAs might be explained by the small sample size of lncRNAs after filtering. This could be a consequence of the biased ability of polyA RNA-seq to capture lncRNAs, which might not reflect true biological insights, and the small sample size made it difficult to reach to a conclusion with confidence that the 0.16 difference between the two R^2 indicates different m6A reading manners. Therefore, based on our results, the summarization of the three YTHDF proteins' m6A reading functions in coding RNAs and lncRNAs are similar.

It should be noted that these correlation results could also stem from experimental noise. To address this concern, we conducted regression analysis between replicates using read counts normalized by DESeq2 internally (Fig 3.c). Results show that there is high correlation between two replicates with a R^2 of 0.99, demonstrating high quality of the dataset and ruling out the possibility that experimental noise contributed to the low correlation found in Fig 3.a and Fig 3.b.

Following that, we further explored whether individual YTHDF proteins function distinctively between coding RNAs and lncRNAs by utilizing RNA-seq data of YTHDF1/2/3 individual knockouts in GSE147849 (Table S1). While the results suggest a higher variation of each YTHDF protein's non-redundant reading ability of m6As on lncRNAs compared to coding RNAs, particularly the dominant

role of YTHDF1 on lncRNAs, the limited number of lncRNAs after filtering restricted the confidence in drawing a final conclusion.

Together, our results show no predominant difference in m6A reading strategy between coding RNAs and lncRNAs, suggesting that the differences in m6A regulation patterns between coding RNAs and lncRNAs could arise somewhere else. One possible explanation lies in the distinct m6A distribution on coding RNAs and lncRNAs, with m6As highly enriched in the last exon and particularly near the stop codon in coding RNAs while a uniform m6A distribution is observed on lncRNAs.[9] Another possibility is that m6A regulation patterns are coupled with translation,[7,10,11] which lncRNAs circumvent, and it can be tested by simply conducting similar analysis on RNA-seq dataset in translation-inhibited conditions.

5 Investigating the Tight Regulation of High m6A-Count Protein-Coding Genes

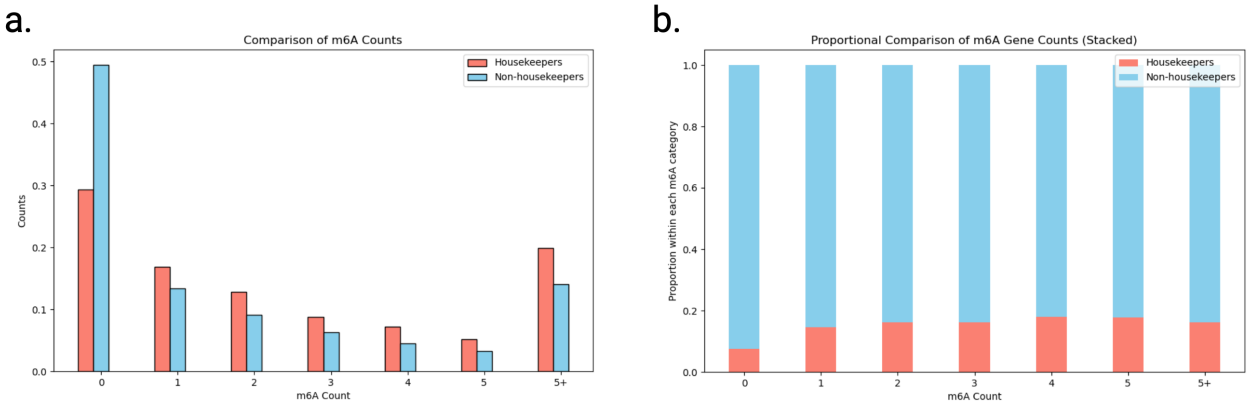

Fig 4 The relationship between housekeeping genes and m6A frequency. Our analytical approach employed two distinct visualizations: the first (a) delineates the independent frequency distributions of housekeeping versus non-housekeeping genes across m6A modification count bins, normalized within each category to facilitate intra-category comparisons. The second (b) represents the proportional distribution of housekeeping and non-housekeeping genes within each m6A count bin, thereby enabling an inter-category comparison.

In the present study, we explored the ramifications of m6A modification machinery disruption on gene expression profiles. Anticipating a pronounced expression alteration in genes with a higher incidence of m6A modifications, our data instead revealed a nuanced landscape: such genes exhibited a marginal increase in expression post-machinery knockout, yet the distribution of their log fold changes was notably tighter than that of m6A-deficient sequences.

This observation prompted us to consider alternative regulatory frameworks that might confer resilience to genes with elevated m6A marks. We postulated a potential interplay between m6A modifications and the stringency of gene regulation. Specifically, we hypothesized that genes with a high frequency of m6A modifications are subject to more stringent regulatory control beyond m6A mechanisms, thereby exhibiting robustness to perturbations such as the loss of m6A-related machinery.

To probe this hypothesis, we conducted an analysis to discern whether housekeeping genes, which are characterized by their consistent expression across various tissues and are often indicative of tight transcriptional control, were overrepresented in the cohort of genes bearing high m6A modification counts. Our comparative analysis aimed to reveal if a relationship between m6A modifications and the housekeeping status of genes might explain their expression stability.

Our findings revealed a modestly higher proportion of housekeeping genes within the 5+ m6A count bin compared to their non-housekeeping counterparts(Fig 4.a). However, this overrepresentation was not

as pronounced when considering the relative distribution within the higher m6A bins(Fig 4.b). Consequently, the data do not support a straightforward correlation between housekeeping gene status and m6A modification frequency.

The implications of these findings are multifold. While they partially align with our initial postulate—suggesting a nuanced correlation between m6A density and transcriptional regulation—they also indicate the presence of additional, yet-to-be-elucidated mechanisms that impart resilience to gene expression changes following m6A reader and writer protein knockouts. This observation paves the way for future inquiries into the complex regulatory networks governing m6A-modified genes and challenges us to delve deeper into the molecular underpinnings of epitranscriptomic influence on gene expression fidelity.

6 Methods

6.1 RNA-sequencing dataset

We used GSE147849[12] as our source data. In this dataset, mouse embryonic stem cells (strain: KH2) were knocked-out of each of the following proteins: METTL3, YTHDF1, YTHDF2, YTHDF3, and YTHDF1/2/3 triple-KO. RNA-seq was generated in each of these KOs and in WT, with two replicates for each condition. RNA-seq libraries were prepared by extracting total RNAs first and then purifying for polyadenylated RNAs. Single-end sequencing data was obtained from NextSeq 550.

6.2 Data pre-processing

Raw fastq files were downloaded from the European Nucleotide Archive(ENA). Quality control was conducted with FastQC and reads were trimmed using cutadapt[13] (version 4.5) (parameters: -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -a AAAAAAAAAA -a TTTTTTTTTT –times 2 -q 20 -m 25). Reads were mapped to genome mm10 using STAR[14] (version 2.7.9) (parameters: –sjdbOverhang 74 –alignEndsType EndToEnd –outFilterMismatchNoverLmax 0.05 –twopassMode Basic). Only uniquely mapped reads were used for downstream analysis. Sample counting was done using StringTie[15] (version 2.2.1), quantifying mm10 Ensembl annotated genes.

6.3 Differential expression analysis and gene annotation

Normalization of the counts and differential expression analysis were performed using DESeq2[16] (version 1.42.0) in R (version 4.3). The number of m6A sites on each gene in mESC was annotated using the comprehensive m6A knowledgebase m6A-Atlas v2.0, which contains 266,643 m6A sites detected in mouse using 7 high-resolution technologies. The biotype of each gene was annotated using the mm10 gene annotation file from GENCODE. The density plots and cumulative distribution function plots were generated with ggplot2 (version 3.4.4) in R.

6.4 Integrated Analysis of Mettl3-knockout and YTHDF1/2/3-triple-knockout Datasets

Filtering m6A-Modified Genes To focus on genes with m6A modifications, we utilized the filter_out_m6a_only_genes.py script. This Python script selectively filters genes based on the presence of m6A modifications (denoted as 'm6A' in the 'm6A_ESC_mm' column) from datasets comparing METTL3 and YTHDF triple-knockout to wild-type conditions in mouse embryonic stem cells. Filtered datasets containing only m6A-modified genes were saved for subsequent analyses.

Replicate Correlation Analysis Using replicate_scatterplot.py, we assessed the consistency between biological replicates. This script plots scatter diagrams of normalized gene expression counts between pairs of replicates, such as WT replicates, and identifies outliers for removal. Linear regression is applied to quantify the correlation between replicates, with results displayed in scatter plots accompanied by histograms representing the distribution of counts in each replicate.

Comparing Writer vs. Reader Knockout Effects The writer_vs_reader_ko_scatterplot.py script was developed to compare the effects of knocking out m6A writer (METTL3) and reader (YTHDF) proteins on gene expression. This analysis was performed on datasets filtered for m6A-modified genes. The script generates scatter plots comparing log2 fold changes in gene expression between METTL3 and YTHDF knockouts, separately for lncRNA and protein-coding genes. Linear regression analysis provides insights into the correlation between the two knockouts' effects, highlighting potential differences in m6A regulatory mechanisms between gene types.

6.5 Analysis of Housekeeping Genes in m6A Modification Patterns

Housekeeping Gene Selection For this study, we utilized a dataset of housekeeping genes, obtained from the Housekeeping and Reference Transcript Atlas (https://housekeeping.unicamp.br).[17] Housekeeping genes, essential for the maintenance of basic cellular functions, are expected to be expressed uniformly across all cell types and developmental stages. In our analysis, a housekeeping gene is defined as one exhibiting at least one protein-coding transcript with a consistent presence in all analyzed cell types or tissues, maintaining a non-zero expression level. The selection criteria for these genes were stringent: they must be expressed at a level above 1 RPKM (reads per kilobase million) in every tissue and cell type studied, exhibit low expression variability (standard deviation of log2 RPKM ¡ 1), and have a maximum fold change (MFC) less than 2. Additionally, only genes without known pseudogenes and having well-supported transcript models were included.

Data Preparation and Analysis We employed our custom Python script, get_housekeeping_genes.py, to segregate housekeeping genes from non-housekeeping genes within our dataset. This script was executed on datasets obtained from mouse embryonic stem cells (strain: KH2) subjected to METTL3 knockout (KO) and wild-type (WT) conditions. The script identifies housekeeping genes based on their presence in the Housekeeping Genes list and partitions the dataset accordingly into housekeepers and non-housekeepers.

m6A Modification Pattern Analysis To visualize the influence of m6A modifications on these gene categories, we developed two Python scripts for data visualization: plot_m6a_histogram_housekeepers.py and plot_m6a_housekeepers_barplot.py. The first script generates histograms comparing the frequency of m6A modifications in housekeeping versus non-housekeeping genes. The second script provides a stacked bar plot to exhibit the proportional distribution of m6A gene counts within each category. These visualizations aid in understanding the relationship between m6A modification frequency and the regulatory stringency of housekeeping genes, as well as their differential expression in response to METTL3 knockout.

7 Code and Data Availabilty

All code, data, and figures can be accessed in our Github repository:
https://github.com/Chasearmer/DeepSequencingFinalProject

8 Supplementary Information

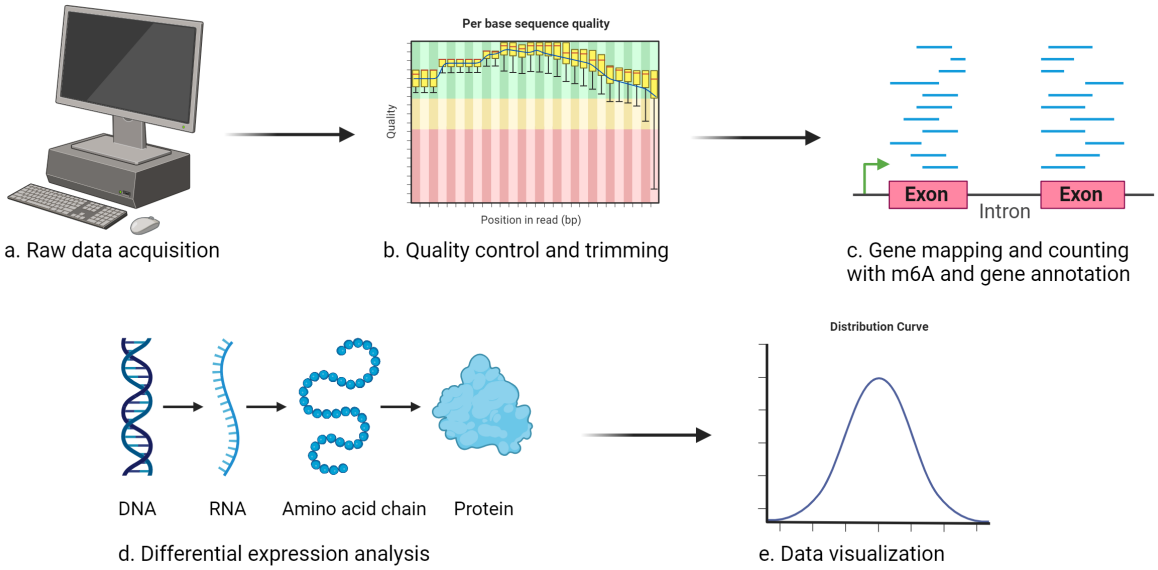

Fig 5 Overview of analysis workflow

Table 1 Statistics of differentially expressed genes in each condition compared to WT

Treatment	Coding RNA	LncRNA	Coding RNA (%)	LncRNA (%)
YTHDF-1 KO	637	19	0.166	0.306
YTHDF-2 KO	450	10	0.117	0.161
YTHDF-3 KO 3	9	0	0.122	0.145
YTHDF triple KO	223	62	1	1

Notes: Genes with FDR<0.05 and at least 1 m6A site are shown. Percentage is calculated as the gene number in individual YTHDF KO condition over YTHDF triple KO condition.

References

1 S. Delaunay, M. Helm, and M. Frye, "RNA modifications in physiology and disease: towards clinical applications," *Nat Rev Genet* , Sept. 2023.

2 S. Murakami and S. R. Jaffrey, "Hidden codes in mRNA: Control of gene expression by m6A," *Mol Cell* **82**, pp. 2236–2251, June 2022.

3 D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, and G. Rechavi, "Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq," *Nature* **485**, pp. 201–206, Apr. 2012.

4 K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons," *Cell* **149**, pp. 1635–1646, June 2012.

5 N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan, "N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions," *Nature* **518**, pp. 560–564, Feb. 2015.

6 D. Yang, J. Qiao, G. Wang, Y. Lan, G. Li, X. Guo, J. Xi, D. Ye, S. Zhu, W. Chen, W. Jia, Y. Leng, X. Wan, and J. Kang, "N6-Methyladenosine modification of lincRNA 1281 is critically required for mESC differentiation potential," *Nucleic Acids Res* **46**, pp. 3906–3920, May 2018.

7 H. Huang, H. Weng, and J. Chen, "m6A Modification in Coding and Non-coding RNAs: Roles and Therapeutic Implications in Cancer," *Cancer Cell* **37**, pp. 270–288, Mar. 2020.

8 X. Jiang, B. Liu, Z. Nie, L. Duan, Q. Xiong, Z. Jin, C. Yang, and Y. Chen, "The role of m6A modification in the biological functions and diseases," *Signal Transduct Target Ther* **6**, p. 74, Feb. 2021.

9 X. Cui, Z. Wei, L. Zhang, H. Liu, L. Sun, S.-W. Zhang, Y. Huang, and J. Meng, "Guitar: An R/Bioconductor Package for Gene Annotation Guided Transcriptomic Analysis of RNA-Related Genomic Features," *Biomed Res Int* **2016**, p. 8367534, 2016.

10 K. D. Meyer, "m6A-mediated translation regulation," *Biochim Biophys Acta Gene Regul Mech* **1862**, pp. 301–309, Mar. 2019.

11 Y. Mao, L. Dong, X.-M. Liu, J. Guo, H. Ma, B. Shen, and S.-B. Qian, "m6A in mRNA coding regions promotes translation via the RNA helicase-containing YTHDC2," *Nat Commun* **10**, p. 5332, Nov. 2019.

12 L. Lasman, V. Krupalnik, S. Viukov, N. Mor, A. Aguilera-Castrejon, D. Schneir, J. Bayerl, O. Mizrahi, S. Peles, S. Tawil, S. Sathe, A. Nachshon, T. Shani, M. Zerbib, I. Kilimnik, S. Aigner, A. Shankar, J. R. Mueller, S. Schwartz, N. Stern-Ginossar, G. W. Yeo, S. Geula, N. Novershtern, and J. H. Hanna, "Context-dependent functional compensation between Ythdf m6A reader proteins," *Genes Dev* **34**, pp. 1373–1391, Oct. 2020.

13 M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal* **17**, pp. 10–12, May 2011. Number: 1.

14 A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics* **29**, pp. 15–21, Jan. 2013.

15 M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nat Biotechnol* **33**, pp. 290–295, Mar. 2015.

16 M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology* **15**, p. 550, Dec. 2014.

17 B. W. Hounkpe, F. Chenou, F. de Lima, and E. V. De Paula, "HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets," *Nucleic Acids Res* **49**, pp. D947–D955, Jan. 2021.