5-methylcytonsine profiles in mouse transcriptomes suggest the randomness of m^5C formation catalyzed by RNA methyltransferase

Junfeng Liu[1,2*]

*Correspondence: liujf@big.ac.cn

[1]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

[2]China National Center for Bioinformation, Beijing 100101, China

Email address: liujunfeng@big.ac.cn for Junfeng Liu

Abstract

5-methylcytosine (m^5C) is a type of chemical modification on the nucleotides and is widespread in both DNA and RNA. Although the DNA m^5C has been extensively studied over the past years, the distribution and biological function of RNA m^5C still remain to be elucidated. Here, I explored the profiles of RNA m^5C in four mouse tissues and found that the methylation rates of cytosine were the same with the averages of methylation level at single-nucleotide level. Furthermore, I gave a mathematical formula to describe the observed relationship and analyzed it deeply. The sufficient necessary condition for the given formula suggests that the methylation levels at most m^5C sites are the same in four mouse tissues. Therefore, I proposed a hypothesis that the m^5C formation catalyzed by RNA

methyltransferase is random and with the same probability at most m^5C sites, which is the methylation rate of cytosine. My hypothesis can be used to explain the observed profiles of RNA m^5C in four mouse tissues and will be benefit to future studies of the distribution and biological function of RNA m^5C in mammals.

Keywords: RNA 5-methylcytosine, methylation rate, methylation level

Background

5-methylcytosine (m^5C) is a type of chemical modification on the nucleotides and is widespread in both DNA and RNA. DNA m^5C is a common epigenetic modification which is crucial for diverse biological processes, including gene silencing, imprinting and X chromosome inactivation [1], and the aberrant m^5C has been associated with a variety of diseases, such as Alzheimer's disease and cancer [2, 3]. Although the DNA m^5C has been extensively studied over the past years, the distribution and biological function of RNA m^5C still remain to be elucidated [4]. A combination of bisulfite treatment of RNA and followed by PCR-based amplification of cDNA and DNA sequencing is an important approach to detect the m^5C sites [5]. After bisulfite treatment, the unmethylated cytosines are converted into uracils through deamination and methylated cytosines are intact. The advantage of BS-RNAseq is that it can analyze the methylated data at single base pair resolution. The high-throughput sequencing of RNA treated with bisulfite (RNA-BisSeq) can be used to profile RNA m^5C at single-nucleotide resolution. Currently, some tools have been developed to analyze RNA-BisSeq, such as meRanTK [6] and Episo [7]. By analyzing RNA-BisSeq data of mouse embryonic stem cells and murine brain, Amort *et al.* [8] observe a pronounced

accumulation of m^5C sites in the vicinity of the translational start codon, depletion in coding sequences, and mixed patterns of enrichment in the 3'UTR. By analyzing human and mouse RNA-BisSeq data, Yang *et al.* [9] reveal that m^5C modification is enriched in CG-rich regions and in regions immediately downstream of translation initiation sites and Liu *et al.,* (2020) find that the RNA m^5C is not evenly distributed among the transcript isoforms at isoform level. However, only partial m^5C sites were analyzed in the above studies. The methylation level of candidate cytosine positions should be no less than 0.2 [8] and 0.1 [7, 9], respectively.

In this study, I mapped m^5C globally in human HeLa cells and multiple mouse tissues using RNA-BidSeq and deeply analyzed the relationship among global methylation rate, methylation level at single-nucleotide resolution and at gene resolution. Collectively, these data suggest that the m^5C formation catalyzed by RNA methyltransferase is random.

Results and discussion

5-methylcytosine profiles in mouse transcriptomes

To explore the profiles of m^5C in mouse transcriptomes, I applied Episo [7] to published RNA-BisSeq data in mouse (liver, kidney, heart and brain). According to the mapping results from Episo, I computed the methylation rate of cytosine and the methylation rate of reads (Table 1). In four mouse tissues (liver, kidney, heart and brain), the methylation rates of cytosine were all 0.001 and the methylation rates of reads were 0.034, 0.029, 0.035 and 0.038 respectively. For computing the methylation level at single-nucleotide level, I analyzed the sites with coverage depth >=30. In four mouse tissues, the averages of methylation level at single-nucleotide level were all 0.001. The averages of methylation level at gene level in four

mouse tissues (liver, kidney, heart and brain were 0.034, 0.030, 0.030, and 0.039 respectively. Intriguingly, the methylation rates of cytosine in four mouse tissues were all the same and were the same with the averages of methylation level at single-nucleotide level. In addition, the methylation rates of reads were also close to the averages of methylation level at gene level in four mouse tissues.

A hypothesis

The profiles of m^{5}C in mouse transcriptomes showed that the methylation rates of cytosine were the same with the averages of methylation level at single-nucleotide level in four mouse tissues. This can be described as the following formula:

$$\left(\frac{a_1}{b_1}+\frac{a_2}{b_2}+\cdots+\frac{a_n}{b_n}\right)\Big/n \approx \frac{a_1+a_2+\cdots+a_n}{b_1+b_2+\cdots+b_n} \quad (1)$$

Where n denotes the number of m^{5}C sites; a_i denotes the number of reads with methylation at the i_{th} m^{5}C site; b_i denotes the number of reads at the i_{th} m^{5}C site. The left of formula (1) means the average of methylation level at single-nucleotide level and the right of formula (1) means the methylation rate of cytosine. In mathematics, the sufficient necessary condition for the formula (1) is $\frac{a_1}{b_1}=\frac{a_2}{b_2}=\cdots=\frac{a_n}{b_n}$ (Additional file 1). Because $\frac{a_i}{b_i}$ is the methylation level at the i_{th} m^{5}C site, the sufficient necessary condition for the formula (1) suggests that the methylation levels at most m^{5}C sites are the same. How to explain the suggestion from the sufficient necessary condition for the formula (1)? If the m^{5}C formation catalyzed by RNA methyltransferase is random and with the same probability at most m^{5}C sites, the methylation level at the i_{th} m^{5}C site is the probability of methylation at the i_{th} m^{5}C site according to law of large numbers when b_i is sufficient large and the suggestion

from the sufficient necessary condition for the formula (1) can be explained. Therefore, I proposed a hypothesis that the m^5C formation catalyzed by RNA methyltransferase is random and with the same probability at most m^5C sites, which is the methylation rate of cytosine. According to my hypothesis, the average of methylation level at gene level should be close to the methylation rate of reads (Table 1). Furthermore, I explored the profile of m^5C in human HeLa cells. The methylation rate of cytosine was still the same with the average of methylation level at single-nucleotide level, and was equal to the methylation rates of cytosine in four mouse tissues.

Discussion

In my hypothesis, the probability at most m^5C sites can be obtained by computing the methylation rate of cytosine. In order to test it, I simulated RNA-BisSeq data, in which the m^5C formation was random and the probability of methylation at each m^5C site was 0.001. Then, I applied Episo to the simulated RNA-BisSeq data and computed the methylation rate of cytosine. The results showed that the methylation rate of cytosine is equal to the probability given by simulating (Additional file 1). In four mouse tissues and human HeLa cells, there are some m^5C sites with significant high methylation level. There are two reasons that can be used to explain this phenomenon. The first reason is that the coverage depth is low. If the coverage depth at one m^5C site is 10, the methylation level is at least 0.1 when the m^5C site was catalyzed by RNA methyltransferase. The second reason may be the high probability of methylation at some m^5C sites because of aberrant RNA methyltransferase.

Conclusions

In this study, I explored the profiles of m^5C in mouse transcriptomes by computing the methylation rate and the methylation level. I found that the methylation rates of cytosine were the same with the averages of methylation level at single-nucleotide level in four mouse tissues. Furtherly. I explored the profile of m^5C in human HeLa cells and observed the same relationship between the methylation rates of cytosine and the average of methylation level at single-nucleotide level. I gave a mathematic formula to describe the above relationship and analyzed it deeply. The sufficient necessary condition for the given formula suggests that the methylation levels at most m^5C sites are the same in four mouse tissues and human HeLa cells. In order to explain the above suggestion, I proposed a hypothesis that the m^5C formation catalyzed by RNA methyltransferase is random and with the same probability at most m^5C sites, which is the methylation rate of cytosine. Finally, I simulated RNA-BisSeq data with the randomness of m^5C formation catalyzed by RNA methyltransferase to test whether the probability at most m^5C sites can be obtained by computing the methylation rate of cytosine. I think my hypothesis will be benefit to future studies of the distribution and biological function of RNA m^5C in mammals.

Methods

Data sources

I downloaded RNA-BisSeq data for human and mouse from the BIG Data Center under accession number PRJCA000315 [10]. Furthermore, I downloaded reference genome and

transcriptome for human (version GRCh37) and mouse (version GRCm38) from the Ensemble database [11].

RNA-BisSeq bioinformatics analysis

The alignment procedure was performed by using Episo [7], which maps RNA-BisSeq reads to the reference genome and reference transcriptome. Episo can convert the m^5C sites in transcriptome and junction sequences to corresponding genome locus. The methylation rate of cytosine is the proportion of converted cytosine in all examined RNA-BisSeq data. The methylation rate of reads is the proportion of the reads with methylation in all examined RNA-BisSeq reads. The methylation level at single-nucleotide level is defined as $i/(i+j)$, where i denotes the number of reads with methylation at the given m^5C site, and j denotes the number of reads lack of methylation at the given m^5C site. The methylation level at gene level is defined as $R_{m,g} \Big/ R_g$, where $R_{m,g}$ denotes the number of reads that carry at least one methylated cytosine site from the given gene, and R_g denotes the number of reads that are from the given gene. I only analyzed the sites with coverage depth >30.

Simulation

I simulated an RNA-Seq experiment using the FluxSimulator with default parameters [12], a freely available software package that models whole-transcriptome sequencing experiments with the Illumina Genome Analyzer. The software works by first randomly assigning expression values to the transcripts provided by user, constructing an amplified, size-selected library, and then sequencing it. Human transcripts assembled by Cufflinks [13]

according to the experimental data were supplied to the FluxSimulator. FluxSimulator then randomly assigned expression levels to 40,205 transcripts and produced paired-end RNA-Seq reads which of length is 101-bp and the numbers are 23 million. Then, I simulated the bisulfite treatment using the Bisulfitefq, which is from the package Episo and converts the non-methylated cytosines into thymines when the probability of methylation at each m^5C site is 0.001.

Declarations

Acknowledgements

Not applicable.

Funding

Not applicable.

Availability of data and materials

The supplementary tables (Additional file 1: Table S1) contain simulation results.

Author's contributions

Not applicable.

Ethic approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

References

1. Law JA, Jacobsen SE: Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics. 2010;11(3):204-20.

2. Urdinguio RG, Sanchez-Mut JV, Esteller M: Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. Lancet Neurology. 2009;8(11):1056-72.

3. Robertson KD: DNA methylation and human disease. Nature Reviews Genetics. 2005;6(8):597-610.

4. Edelheit S, Schwartz S, Mumbach MR, 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 Genetics. 2013;9(6).

5. Schaefer M, Pollex T, Hanna K, Lyko F: RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Research. 2009;37(2).

6. Rieder D, Amort T, Kugler E, Lusser A, Trajanoski Z: meRanTK: methylated RNA analysis ToolKit. Bioinformatics. 2016;32(5):782-5.

7. Liu J, An Z, Luo J, Li J, Li F, Zhang Z: Episo: quantitative estimation of RNA 5-methylcytosine at isoform level by high-throughput sequencing of RNA treated with bisulfite. Bioinformatics. 2020;36(7):2033-9.

8. Amort T, Rieder D, Wille A, Khokhlova-Cubberley D, Riml C, Trixl L, Jia X-Y, Micura R, Lusser A: Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain. Genome Biology. 2017;18.

9. Yang X, Yang Y, Sun B-F, Chen Y-S, Xu J-W, Lai W-Y, Li A, Wang X, Bhattarai DP, Xiao W, et al: 5-methylcytosine promotes mRNA export-NSUN2 as the methyltransferase and ALYREF as an m(5)C reader. Cell Research. 2017;27(5):606-25.

10. Members, B.I.G.D.C. The BIG Data Center: from deposition to integration to translation. Nucleic Acids Research. 2017;45(D1):D18-24.

11. Yates, A. et al: Ensembl 2016. Nucleic Acids Research, 2016;44(D1):D710-16.

12. Montgomery, S.B. et al: Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464(7289):773-U151.

13. Trapnell, C. et al: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010;28(5):511-U174.

Tables

Table 1. 5-methylcytosine profiles in mouse transcriptomes

Tissue	methylation rate_A	methylation level_A	methylation rate_B	methylation level_B
Liver	0.001	0.001	0.034	0.034
Kidney	0.001	0.001	0.029	0.030
Heart	0.001	0.001	0.035	0.030
Brain	0.001	0.001	0.038	0.039

Note. methylation rate_A means the methylation rate of cytosine; methylation level_A means the average of methylation level at single-nucleotide level; methylation rate_B means the methylation rate of reads; methylation level_B means the average of methylation level at gene level. For computing the methylation level at single-nucleotide level, I analyzed the sites with coverage depth >=30.

Additional Files

Additional file 1: Supplementary Material (DOCX).

This additional file includes the more details for the sufficient necessary condition for the formula (1) and the simulation results.