

1 5-methylcytosine profiles in mouse 2 transcriptomes suggest the randomness 3 of m⁵C formation catalyzed by RNA 4 methyltransferase

5 Junfeng Liu^{1,2*}

6 *Correspondence: liujf@big.ac.cn

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

9 ²China National Center for Bioinformation, Beijing 100101, China

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

11 Abstract

12 5-methylcytosine (m⁵C) is a type of chemical modification on the nucleotides and is
13 widespread in both DNA and RNA. Although the DNA m⁵C has been extensively studied over
14 the past years, the distribution and biological function of RNA m⁵C still remain to be
15 elucidated. Here, I explored the profiles of RNA m⁵C in four mouse tissues and found that the
16 methylation rates of cytosine were the same with the averages of methylation level at
17 single-nucleotide level. Furthermore, I gave a mathematical formula to describe the
18 observed relationship and analyzed it deeply. The sufficient necessary condition for the given
19 formula suggests that the methylation levels at most m⁵C sites are the same in four mouse
20 tissues. Therefore, I proposed a hypothesis that the m⁵C formation catalyzed by RNA

21 methyltransferase is random and with the same probability at most m⁵C sites, which is the
22 methylation rate of cytosine. My hypothesis can be used to explain the observed profiles of
23 RNA m⁵C in four mouse tissues and will be benefit to future studies of the distribution and
24 biological function of RNA m⁵C in mammals.

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

26 Background

27 5-methylcytosine (m⁵C) is a type of chemical modification on the nucleotides and is
28 widespread in both DNA and RNA. DNA m⁵C is a common epigenetic modification which is
29 crucial for diverse biological processes, including gene silencing, imprinting and X
30 chromosome inactivation [1], and the aberrant m⁵C has been associated with a variety of
31 diseases, such as Alzheimer's disease and cancer [2, 3]. Although the DNA m⁵C has been
32 extensively studied over the past years, the distribution and biological function of RNA m⁵C
33 still remain to be elucidated [4]. A combination of bisulfite treatment of RNA and followed by
34 PCR-based amplification of cDNA and DNA sequencing is an important approach to detect
35 the m⁵C sites [5]. After bisulfite treatment, the unmethylated cytosines are converted into
36 uracils through deamination and methylated cytosines are intact. The advantage of
37 BS-RNAseq is that it can analyze the methylated data at single base pair resolution. The
38 high-throughput sequencing of RNA treated with bisulfite (RNA-BisSeq) can be used to
39 profile RNA m⁵C at single-nucleotide resolution. Currently, some tools have been developed
40 to analyze RNA-BisSeq, such as meRanTK [6] and Episo [7]. By analyzing RNA-BisSeq data of
41 mouse embryonic stem cells and murine brain, Amort *et al.* [8] observe a pronounced

42 accumulation of m⁵C sites in the vicinity of the translational start codon, depletion in coding
43 sequences, and mixed patterns of enrichment in the 3'UTR. By analyzing human and mouse
44 RNA-BisSeq data, Yang *et al.* [9] reveal that m⁵C modification is enriched in CG-rich regions
45 and in regions immediately downstream of translation initiation sites and Liu *et al.*, (2020)
46 find that the RNA m⁵C is not evenly distributed among the transcript isoforms at isoform
47 level. However, only partial m⁵C sites were analyzed in the above studies. The methylation
48 level of candidate cytosine positions should be no less than 0.2 [8] and 0.1 [7, 9], respectively.

49 In this study, I mapped m⁵C globally in human HeLa cells and multiple mouse tissues
50 using RNA-BidSeq and deeply analyzed the relationship among global methylation rate,
51 methylation level at single-nucleotide resolution and at gene resolution. Collectively, these
52 data suggest that the m⁵C formation catalyzed by RNA methyltransferase is random.

53 Results and discussion

54 5-methylcytosine profiles in mouse transcriptomes

55 To explore the profiles of m⁵C in mouse transcriptomes, I applied Episo [7] to published
56 RNA-BisSeq data in mouse (liver, kidney, heart and brain). According to the mapping results
57 from Episo, I computed the methylation rate of cytosine and the methylation rate of reads
58 (Table 1). In four mouse tissues (liver, kidney, heart and brain), the methylation rates of
59 cytosine were all 0.001 and the methylation rates of reads were 0.034, 0.029, 0.035 and 0.038
60 respectively. For computing the methylation level at single-nucleotide level, I analyzed the
61 sites with coverage depth ≥ 30 . In four mouse tissues, the averages of methylation level at
62 single-nucleotide level were all 0.001. The averages of methylation level at gene level in four

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

68 A hypothesis

69 The profiles of m⁵C in mouse transcriptomes showed that the methylation rates of cytosine
70 were the same with the averages of methylation level at single-nucleotide level in four
71 mouse tissues. This can be described as the following formula:

$$72 \left(\frac{a_1}{b_1} + \frac{a_2}{b_2} + \dots + \frac{a_n}{b_n} \right) / n \approx \frac{a_1 + a_2 + \dots + a_n}{b_1 + b_2 + \dots + b_n} \quad (1)$$

73 Where n denotes the number of m⁵C sites; a_i denotes the number of reads with
74 methylation at the i_{th} m⁵C site; b_i denotes the number of reads at the i_{th} m⁵C site. The
75 left of formula (1) means the average of methylation level at single-nucleotide level and the
76 right of formula (1) means the methylation rate of cytosine. In mathematics, the sufficient
77 necessary condition for the formula (1) is $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \dots = \frac{a_n}{b_n}$ (Additional file 1). Because $\frac{a_i}{b_i}$ is
78 the methylation level at the i_{th} m⁵C site, the sufficient necessary condition for the formula (1)
79 suggests that the methylation levels at most m⁵C sites are the same. How to explain the
80 suggestion from the sufficient necessary condition for the formula (1)? If the m⁵C formation
81 catalyzed by RNA methyltransferase is random and with the same probability at most m⁵C
82 sites, the methylation level at the i_{th} m⁵C site is the probability of methylation at the i_{th}
83 m⁵C site according to law of large numbers when b_i is sufficient large and the suggestion

84 from the sufficient necessary condition for the formula (1) can be explained. Therefore, I
85 proposed a hypothesis that the m⁵C formation catalyzed by RNA methyltransferase is
86 random and with the same probability at most m⁵C sites, which is the methylation rate of
87 cytosine. According to my hypothesis, the average of methylation level at gene level should
88 be close to the methylation rate of reads (Table 1). Furthermore, I explored the profile of m⁵C
89 in human HeLa cells. The methylation rate of cytosine was still the same with the average of
90 methylation level at single-nucleotide level, and was equal to the methylation rates of
91 cytosine in four mouse tissues.

92 Discussion

93 In my hypothesis, the probability at most m⁵C sites can be obtained by computing the
94 methylation rate of cytosine. In order to test it, I simulated RNA-BisSeq data, in which the
95 m⁵C formation was random and the probability of methylation at each m⁵C site was 0.001.
96 Then, I applied Episo to the simulated RNA-BisSeq data and computed the methylation rate
97 of cytosine. The results showed that the methylation rate of cytosine is equal to the
98 probability given by simulating (Additional file 1). In four mouse tissues and human HeLa
99 cells, there are some m⁵C sites with significant high methylation level. There are two reasons
100 that can be used to explain this phenomenon. The first reason is that the coverage depth is
101 low. If the coverage depth at one m⁵C site is 10, the methylation level is at least 0.1 when the
102 m⁵C site was catalyzed by RNA methyltransferase. The second reason may be the high
103 probability of methylation at some m⁵C sites because of aberrant RNA methyltransferase.

104 Conclusions

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

120 Methods

121 Data sources

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

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

126 RNA-BisSeq bioinformatics analysis

127 The alignment procedure was performed by using Episo [7], which maps RNA-BisSeq reads
128 to the reference genome and reference transcriptome. Episo can convert the m⁵C sites in
129 transcriptome and junction sequences to corresponding genome locus. The methylation rate
130 of cytosine is the proportion of converted cytosine in all examined RNA-BisSeq data. The
131 methylation rate of reads is the proportion of the reads with methylation in all examined
132 RNA-BisSeq reads. The methylation level at single-nucleotide level is defined as $i/(i+j)$, where
133 i denotes the number of reads with methylation at the given m⁵C site, and j denotes the
134 number of reads lack of methylation at the given m⁵C site. The methylation level at gene
135 level is defined as $R_{m,g}/R_g$, where $R_{m,g}$ denotes the number of reads that carry at
136 least one methylated cytosine site from the given gene, and R_g denotes the number of
137 reads that are from the given gene. I only analyzed the sites with coverage depth >30.

138 Simulation

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

144 according to the experimental data were supplied to the FluxSimulator. FluxSimulator then
145 randomly assigned expression levels to 40,205 transcripts and produced paired-end
146 RNA-Seq reads which of length is 101-bp and the numbers are 23 million. Then, I simulated
147 the bisulfite treatment using the Bisulfitefq, which is from the package Episo and converts the
148 non-methylated cytosines into thymines when the probability of methylation at each m⁵C
149 site is 0.001.

150 **Declarations**

151 **Acknowledgements**

152 Not applicable.

153 **Funding**

154 Not applicable.

155 **Availability of data and materials**

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

157 **Author's contributions**

158 Not applicable.

159 **Ethic approval and consent to participate**

160 Not applicable.

161 Consent for publication

162 Not applicable.

163 Competing interests

164 Not applicable.

165 References

166 1. Law JA, Jacobsen SE: Establishing, maintaining and modifying DNA methylation patterns

167 in plants and animals. *Nature Reviews Genetics*. 2010;11(3):204-20.

168 2. Urduingio RG, Sanchez-Mut JV, Esteller M: Epigenetic mechanisms in neurological

169 diseases: genes, syndromes, and therapies. *Lancet Neurology*. 2009;8(11):1056-72.

170 3. Robertson KD: DNA methylation and human disease. *Nature Reviews Genetics*.

171 2005;6(8):597-610.

172 4. Edelheit S, Schwartz S, Mumbach MR, Wurtzel O, Sorek R: Transcriptome-Wide

173 Mapping of 5-methylcytidine RNA Modifications in Bacteria, Archaea, and Yeast Reveals

174 m(5)C within Archaeal mRNAs. *Plos Genetics*. 2013;9(6).

175 5. Schaefer M, Pollex T, Hanna K, Lyko F: RNA cytosine methylation analysis by bisulfite

176 sequencing. *Nucleic Acids Research*. 2009;37(2).

177 6. Rieder D, Amort T, Kugler E, Lusser A, Trajanoski Z: meRanTK: methylated RNA analysis

178 ToolKit. *Bioinformatics*. 2016;32(5):782-5.

179 7. Liu J, An Z, Luo J, Li J, Li F, Zhang Z: Episo: quantitative estimation of RNA

180 5-methylcytosine at isoform level by high-throughput sequencing of RNA treated with

181 bisulfite. *Bioinformatics*. 2020;36(7):2033-9.

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

185 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,
186 et al: 5-methylcytosine promotes mRNA export-NSUN2 as the methyltransferase and
187 ALYREF as an m(5)C reader. *Cell Research*. 2017;27(5):606-25.

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

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

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

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

196 **Tables**

197 **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 |

198 Note. methylation rate_A means the methylation rate of cytosine; methylation level_A means
199 the average of methylation level at single-nucleotide level; methylation rate_B means the
200 methylation rate of reads; methylation level_B means the average of methylation level at
201 gene level. For computing the methylation level at single-nucleotide level, I analyzed the
202 sites with coverage depth ≥ 30 .

203 Additional Files

204 **Additional file 1: Supplementary Material (DOCX).**

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