Bioinformatics, 36(7), 2020, 2033–2039
doi: 10.1093/bioinformatics/btz900
Advance Access Publication Date: 3 December 2019
Original Paper

Genome analysis

Episo: quantitative estimation of RNA 5-methylcytosine at isoform level by high-throughput sequencing of RNA treated with bisulfite

Junfeng Liu[1,†], Ziyang An[1,2,†], Jianjun Luo[3], Jing Li[1], Feifei Li[1,*] and Zhihua Zhang[1,2,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [2]School of Life Science, University of Chinese Academy of Sciences, Beijing 100049, China and [3]Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jan Gorodkin

Received on June 4, 2019; revised on September 7, 2019; editorial decision on November 26, 2019; accepted on November 29, 2019

Abstract

Motivation: RNA 5-methylcytosine (m^5C) is a type of post-transcriptional modification that may be involved in numerous biological processes and tumorigenesis. RNA m^5C can be profiled at single-nucleotide resolution by high-throughput sequencing of RNA treated with bisulfite (RNA-BisSeq). However, the exploration of transcriptome-wide profile and potential function of m^5C in splicing remains to be elucidated due to lack of isoform level m^5C quantification tool.

Results: We developed a computational package to quantify Epitranscriptomal RNA m^5C at the transcript isoform level (named Episo). Episo consists of three tools: *mapper*, *quant* and *Bisulfitefq*, for mapping, quantifying and simulating RNA-BisSeq data, respectively. The high accuracy of Episo was validated using an improved m^5C-specific methylated RNA immunoprecipitation (meRIP) protocol, as well as a set of *in silico* experiments. By applying Episo to public human and mouse RNA-BisSeq data, we found that the RNA m^5C is not evenly distributed among the transcript isoforms, implying the m^5C may subject to be regulated at isoform level.

Availability and implementation: Episo is released under the GNU GPLv3+ license. The resource code Episo is freely accessible from https://github.com/liujunfengtop/Episo (with Tophat/cufflink) and https://github.com/liujunfengtop/Episo/tree/master/Episo_Kallisto (with Kallisto).

Contact: liff@big.ac.cn or zhangzhihua@big.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Post-transcriptional modifications in RNAs have drawn much attention in recent literature (Chi, 2017), as rapidly growing evidence has suggested that reversible RNA modifications may be a new layer of epigenetic regulation in gene expression (Zhao *et al.*, 2017), and its disruption may lead to life-threatening disease like cancer (Frye *et al.*, 2016; Popis *et al.*, 2016). 5-Methylcytosine is a type of chemical modification on the nucleotides that can be found in both DNA and RNA. To identify transcriptome-wide m^5C, several advanced techniques have been employed, including high-throughput sequencing of RNA treated with bisulfite (RNA-BisSeq) (Schaefer *et al.*, 2010; Squires *et al.*, 2012), RNA m^5C-RNA immunoprecipitation (RIP) (Jayaseelan *et al.*, 2011), 5-azacytidine-mediated RNA immunoprecipitation (Aza-IP) (Khoddami and Cairns, 2013) and methylation individual-nucleotide-resolution crosslinking and immunoprecipitation (miCLIP) (Hussain *et al.*, 2013a, b). Among these, RNA-BisSeq is considered the gold standard for RNA m^5C (Amort *et al.*, 2017; Hussain *et al.*, 2013a,b; Squires *et al.*, 2012; Yang *et al.*, 2017). Hereinafter in this article, m^5C only refers to RNA m^5C, unless otherwise indicated.

The transcriptome-wide function of m^5C is just starting to be investigated (Edelheit *et al.*, 2013; Hussain *et al.*, 2013a, b; Shelton *et al.*, 2016). It is clear that m^5C is distributed widely over protein coding and non-coding RNAs in human (Squires *et al.*, 2012) and mouse (Amort *et al.*, 2017), as well as enriched in CG-rich regions and the initiation site of coding regions (Yang *et al.*, 2017). RNA m^5C is conserved from bacteria to mammals and plants, and its functions have been suggested to include structural and metabolic stabilization and translational regulation (Blanco *et al.*, 2014; Burgess *et al.*, 2015; Chen *et al.*, 2019; Gabriel Torres *et al.*, 2014; Popis *et al.*, 2016; Schaefer *et al.*, 2010; Schwartz *et al.*, 2013;

Squires *et al.*, 2012; Yang *et al.*, 2019). At the transcription level, alternative splicing is a predominant contributor to the diversity of protein types in higher eukaryotic cells (Black, 2003). Any changes in this regulated process could result in severe phenotypes (Pan *et al.*, 2008). However, both the distribution of m^{5}C among the transcript isoforms and its functional relevance to the splicing programme have been barely investigated.

In order to study the relationship between m^{5}C and splicing, a tool that is able to quantify m^{5}C at isoform level is necessary. However, to the best of our knowledge, such method has yet been found in literature. There are several tools for RNA-BisSeq or Aza-IP data analysis (Bormann *et al.*, 2018; Legrand *et al.*, 2017; Liang *et al.*, 2016; Rieder *et al.*, 2016). meRanTK is a toolkit composed of tools for RNA-BisSeq read mapping, methylation calling and differentially methylation identification (Rieder *et al.*, 2016). BS-RNA maps and annotates RNA-BisSeq data with more attention on the 'dovetailing' reads (Liang *et al.*, 2016). BisRNA considers the possible artifact that may be stochastically introduced by the experiment (Legrand *et al.*, 2017). BisAMP is more specific to targeted RNA m^{5}C analysis (Bormann *et al.*, 2018). However, none of above tools supports the analysis of RNA-BisSeq data at isoform level.

To address this issue, we developed a probabilistic model to quantify Epitranscriptomal RNA m^{5}C at the transcript isoform level (named Episo), which utilizes single-nucleotide resolution m^{5}C data from RNA-BisSeq. Episo consists of three tools: *mapper*, *quant* and *Bisulfitefq*, for mapping, quantifying and simulating RNA-BisSeq data, respectively. Both *in silico* and wet experiments showed that the prediction of Episo is highly accurate. By applying Episo to recently published m^{5}C data (Yang *et al.*, 2017), we generated the first transcript isoform level m^{5}C profiles for HeLa cells and four mouse tissues. We found that the distribution of m^{5}C in the transcript isoforms was remarkably dissimilar to random shuffled data, suggesting that there may exist a latent regulatory layer for m^{5}C at isoform level.

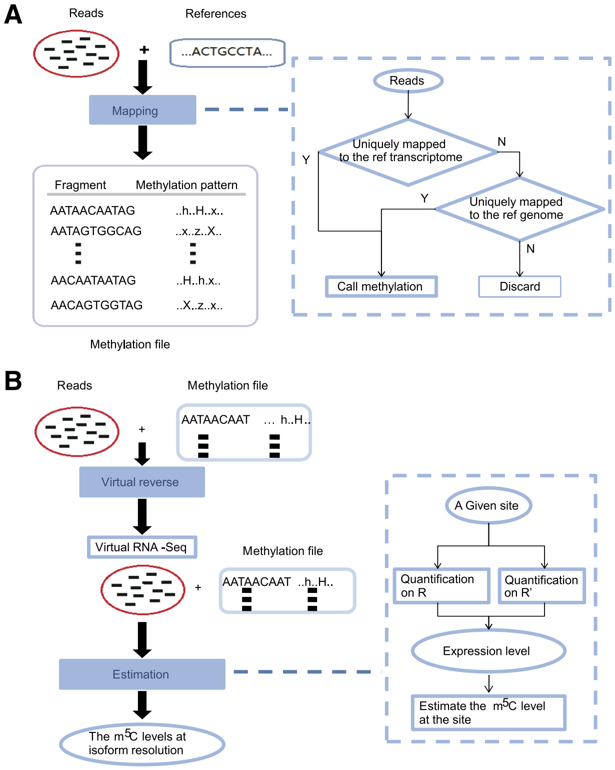

Fig. 1. The Episo pipeline. (**A**) The mapping procedure. Incoming RNA-BisSeq reads are mapped to reference genome and transcriptome. The output methylation file contains two columns that represent mapped fragments and methylation pattern. The symbols Z, X and H represent cytosines in CpG, CHG and CHH, respectively, whereas H can be A, C or T. The upper- and lowercase letters represent methylated and unmethylated cytosines, respectively. (**B**) The quantification procedure. For any given cytosine site, the total reads that cover the site and the reads that carry methylated cytosine at the given cytosine site are denoted as R and R′, respectively

2 Materials and methods

Reference genome and transcriptome for human and mouse, version GRCh37 and GRCm38, respectively, were downloaded from the Ensembl database (Yates *et al.*, 2016). The RNA-BisSeq data were downloaded from the BIG Data Center under accession number PRJCA000315 (Members, 2017).

Episo consists of three tools: *mapper*, *quant* and *Bisulfitefq*, for mapping, quantifying and simulating RNA-BisSeq data, respectively (Fig. 1). Before reads mapping, Episo has all low-quality bases and adaptor sequences removed by Cutadapt (Martin, 2011).

The *mapper* maps RNA-BisSeq reads to the reference genome and reference transcriptome. We adopted the mapping strategy used in Bismark to map RNA-BisSeq reads with modifications (Krueger and Andrews, 2011). First, all RNA-BisSeq reads were C-to-T and G-to-A transformed, and the resultant data were denoted as BSC-T and BSG-A, respectively. Second, the reference transcriptrome was also C-to-T and G-to-A transformed, and the transformed references were denoted as RefC-T and RefG-A, respectively. Last, the four types of mapping (BSC-T versus RefC-T, BSC-T versus RefG-A, BSG-A versus RefC-T and BSG-A versus RefG-A) were performed by Bowtie (version 1.1.2, see Fig. 1). The uniquely mapped reads, i.e. those that were uniquely mapped to a genome locus in at least one of four above mappings, but not necessarily mapped to a unique transcript, were used in subsequent processes.

The *quant* quantifies m^{5}C level at transcription isoform level from RNA-BisSeq data. The *quant* consists of two steps. The first step estimates transcription level from RNA-BisSeq data. To accomplish this, *quant* constructs a virtual RNA-seq dataset, i.e. for all RNA-BisSeq reads that contain unmethylated cytosines, *quant* transforms them back to their native cytosine states. With such virtual RNA-seq data, *quant* estimates gene transcription level using third party tools, which has two choice in current implement, i.e. Tophat (version 2.1.0)/Cufflink(version 2.2.1) (Trapnell *et al.*, 2009, 2010) and Kallisto (version 0.44.0) (Bray *et al.*, 2016). The users can also replace them with any favorite tools easily by making input data format acceptable for Episo. For example, Episo takes fragments per kilobase of transcript per million mapped reads (FPKM) as the default input format, while the output of Kalliato is estimated transcripts per million (TPM). So, we need to convert TPM to FPKM using the following formula.

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j}\right) \times 10^6 \quad (1)$$

We present the performance and data analysis using the results from Tophat/Cufflink in this article; however, the results are essentially similar when using Kalliato.

The second step of *quant* estimates the RNA m^{5}C level at each putative methylation site in the isoforms. We define the methylation rate at global, isoform and single-nucleotide levels as follows. The global methylation rate is the proportion of cytosine sites that have been methylated in all examined RNAs. This rate is estimated by directly counting the unconverted cytosines in RNA-BisSeq data. The methylation rate at isoform level is defined as $R_{m,iso}/R_{iso}$, $R_{m,iso} \subseteq R_{iso}$, where $R_{m,iso}$ denotes the RNAs that carry at least one methylated cytosine site, and R_{iso} denotes all RNAs of the given isoform *iso*. The methylation rate of a single-nucleotide at the level of a given set of isoform(s) is defined as $R_{m,c}/R_c$, $R_{m,c} \subseteq R_c$, where $R_{m,c}$ denotes the RNAs of the given isoform(s) from the methylated cytosine sites, and R_c denotes all RNAs of the given isoform(s) that carry this cytosine site.

To estimate the RNA m^{5}C rate at isoform level, one needs to estimate the probability that a read $r \in R$ was generated from a given isoform $t \in T$, where R denote a set of RNA-BisSeq reads and T denote a set of isoform(s). Let's denote this probability as $P(r, t)$. One way to calculate this probability was showed by Trapnell *et al.* (2009),

$$P(r,\ t) = \frac{\rho_t \tilde{l}(t)}{\sum_{u \in T} \rho_u \tilde{l}(u)} \left(\frac{F(l_t(r))}{l(t) - l_t(r) + 1} \right) \tag{2}$$

where ρ_t denotes the proportion of reads that were generated from isoform t, $l(t)$ denotes the length of an isoform, $\tilde{l}(u)$ denotes the effective length of an isoform, $l_t(r)$ denotes the implied length of r, assuming it originated from isoform t and F denotes the distribution function of read length. The first term in the above formula (2) is the probability that a read selected at random originates from isoform t (denoted as *Pr* (*trans*=*t*)), and the second term is the conditional probability that read r was observed when it originated from isoform t. The effective length of an isoform is defined as

$$\tilde{l}(t) = \sum_{i=1}^{l(t)} F(i)(l(t) - i + 1) \tag{3}$$

Therefore, the likelihood function, the maximum likelihood estimate and the 95% confidence interval of ρ_t can be easily derived from formula (2). It is the ρ_t and the 95% confidence interval needed for Episo; however, users can take these two values from any third party tools. In the current release of Episo, the users can either choose Kallisto (version 0.44.0) (Bray *et al.*, 2016) or Tophat (version 2.1.0)/Cufflink(version 2.2.1) (Trapnell *et al.*, 2010) for these two inputs.

Now, we estimate the RNA m^{5}C level for a given isoform t (denoted as M_t). Let R' denotes the reads that carry at least one methylated cytosine, i.e. unconverted cytosines after bisulfite treatment. We define the relative transcription level of methylated isoform, i.e. the proportion of reads from isoform *t in* R' is denoted as $\bar{\rho}_t$. Following logic identical to that in formula (2), we can calculate the maximum likelihood estimation and 95% confidence interval of $\bar{\rho}'_t$. To simplify the derivation, we assumed the independence between transcription and post-transcriptional modification, i.e. $\bar{\rho}_t \perp \bar{\rho}'_t$. M_t can be estimated according to the delta method, as

$$M_{t=} \left(\frac{m}{n} \right) \left(\frac{\bar{\rho}'_t}{\bar{\rho}_t} + \frac{\bar{\rho}'_t \sigma_t^2}{(\bar{\rho}_t)^3} \right) \tag{4}$$

where m and n denote the number of reads in R' and R, respectively, $-\rho_t$ and $-\rho'_t$ denote the estimated ρ_t and $\rho^{'}_t$, respectively, and σ_t^2 is the variance of $-\rho_t$. Moreover, *quant* can estimate the m^{5}C level for a subset of isoforms of a given gene. Let B be a subset containing k isoforms from gene A; then, the RNA m^{5}C level in B can be estimated, as

$$\left(\frac{m}{n} \right) \left(\frac{\sum^k{}_{t=1} \bar{\rho}'_t}{\sum^k{}_{t=1} \bar{\rho}_t} + \frac{\left(\sum^k{}_{t=1} \bar{\rho}'_t \right) \left(\sum^k{}_{t=1} \sigma_t^2 \right)}{\left(\sum^k{}_{t=1} \bar{\rho}_t \right)^3} \right) \tag{5}$$

Finally, we estimate the RNA m^{5}C level for a given cytosine site. Let $\bar{R}$ denote the reads that cover the given cytosine site, and let $\bar{R}'$ denote the reads that carry methylated cytosine at the given site. Following logic identical to that in formulas (2)–(5), the RNA m^{5}C of a given cytosine site on an isoform or a subset of isoforms can be estimated.

The *Bisulfitefq* simulates bisulfite treatments, and the sequencing process were simulated by the FluxSimulator with default parameters (Montgomery *et al.*, 2010). *Bisulfitefq* consists of *Bisulfitefq*-reads, *Bisulfitefq*-fragment and *Bisulfitefq*-fragment-multirate. The *Bisulfitefq*-reads component generates RNA-BisSeq data with a given global m^{5}C level, i.e. it randomly transforms a given proportion of cytosines to thymines in the input reads. The *Bisulfitefq*-fragment component transforms all cytosines in a given proportion of randomly selected reads into thymines. The *Bisulfitefq*-fragment–multirate component can then simulate an epitranscriptome with multiple m^{5}C levels between isoforms. It can take up to three m^{5}C levels as input parameters. For each m^{5}C level, it transforms all cytosines in the given proportion of randomly selected reads from one assigned isoform into thymines. For example, let a gene A have three isoforms, namely A1, A2 and A3, while P1, P2 and P3 are three m^{5}C levels we want to simulate for the isoforms, respectively. The *Bisulfitefq*-fragment–multirate transforms all cytosines with 1-P1, 1-P2 and 1-P3 of randomly selected reads from isoform A1, A2 and A3, respectively, into thymines.

Gene ontology (GO) analysis of RNA m^{5}C-containing genes was performed using DAVID (Dennis *et al.*, 2003), and the sequence motifs of the RNA m^{5}C sites were discovered using MEME (Bailey *et al.*, 2009).

The human cervical carcinoma cell line HeLa are cultured in DMEM high glucose medium (Hyclone, SH30243.01) with 10% FBS (Biowest, S1580-500) and 1% Penicillin–Streptomycin (Corning, 30002283). Cell identity was verified by STR analysis (DNA fingerprinting), and mycoplasma contamination was regularly tested for cell cultures.

The regions spanning nt 914–1465 of *E.coli* 16S rRNA sequence was amplified by PCR with a forward primer harboring a T7 promoter sequence . 0.5 µg production was used as template for in vitro transcription with MEGAScript Kit (Promega, P1440) according to the protocol. The transcribed RNA was treated 15 min at 37°C with 1 µl RQ1 RNase-free DNase and purified using RNA clean kit (ZYMO REASEARCH R1015) following the protocol. The spike-in RNA was subpackaged and stored in RNase-free water at -80°C.

Following the manufacturer's recommendations, about 3 × 10^7 HeLa cells were treated with 6 ml TRIzol Reagent (Invitrogen, 15596026) for total RNA isolation. Isolated RNA was then subjected to two rounds of poly(A) RNA enrichment using fresh Dynabeads (Ambion, 61006) and treated with 1U of DNase I (Thermo Scientific, 00383793) for 15 min at 37°C. The mRNA was cleaned using RNA clean kit (ZYMO REASEARCH R1015) and concentration was determined in nanodrop (Thermo Scientific, Nanodrop 1000) by measuring absorbance at 260 and 280 nm.

Thirty microgram purified mRNA was mixed with 50 ng spike-in RNA. The mixture was divided equally into three portions. Four microgram anti-m^{5}C antibody (1.36 µg/µl, Diagenode, C15200081), 2 µg random 25nt oligonucleotides and 50 µl Dynabeads Protein G (Novex, 10007D) was incubated in 300 µl IP buffer [10 mM Tris–HCl pH 7.5, 150 mM NaCl, 0.05%Triton-X(v/v)] at 4°C for 2 h on a rotating wheel. The same procedure was performed using Mouse IgG (1 µg/µl, Diagenode, C15400001) as control. We used 250 µl IP buffer to wash bead–antibody complexes three times, then added RNA mixture and finally brought to 250 µl with IP buffer. The mixtures were incubated at 4°C overnight (>12 h) with 1 µl RNasin (Invitrogen, N8080119) on a rotating wheel. The RNA–antibody–beads complexes were gentle washed three times using IP buffer and incubated in 300 µl elution buffer (5 mM Tris–HCl pH 7.5, 1 mM EDTA, 0.05%SDS, 80 µg Proteinase K) for 1 h at 50°C. After removing beads complexes, the eluted production was cleaned with RNA clean kit (ZYMO REASEARCH R1015). We repeated the beads–antibody incubation, RNA-complexes incubation and RNA–antibody–beads complexes elution processes for five times. The supernatant was collected in a new tube and cleaned with RNA clean kit (ZYMO REASEARCH R1015). The supernatant concentration was determined in nanodrop by measuring absorbance at 260 and 280 nm.

Five tubes of cleaned eluted MeRIP production and 300 ng supernatant RNA were reverse transcribed using SuperScript III reverse transcriptase (Invitrogen 18080-093). Each tube contains 11 µl RNA solution, 1 µl dNTPs and 1 µl random 6 bp Hexamers. The tubes were incubated at 65°C for 5 min and 0°C at least 1 min. Four microlitre 5× First strand buffer, 1 µl 0.1 M DTT, 1 µl RNasin and 1 µl SuperScript III reverse transcriptase were mixed and added the mixture into each tube. The tubes were incubated at 55°C for 60 min followed by 70°C for 15 min. The reverse transcribed products were stored at -20°C.

The spike-in RNA was used as internal unspecific binding control to normalize binding of RNA between IgG control and m^{5}C sample. Five tubes of reverse transcribed MeRIP products were mixed and measured for the enrichment using $2^{-\Delta\Delta Ct}$ method by comparing the anti-m^{5}C sample with the anti-IgG sample. Data were expressed as the expression of target genes relative to the spike-in control in the anti-m^{5}C sample compared with the anti-IgG sample. PCR primers can be found in the Supplementary Table S1.

If our model is correct, we should expect

$$\text{Predicted MR rate } \tilde{}\text{Observed MR rate}$$

where

$$\text{Predicted MR rate} = \frac{\text{m5c level from isoform A}}{\text{m5c level from isoform B}}$$

$$\text{Observed MR rate} = \frac{\text{observed m5c level in isoform A}}{\text{observed m5c level in isoform B}}$$

$$= \frac{\dfrac{\text{m5C counts from isoform A}}{\text{IgG counts from isoform A}}}{\dfrac{\text{m5C counts from isoform B}}{\text{IgG counts from isoform B}}}$$

$$= \frac{\text{FCmeRIP(isoform A)}}{\text{FCmeRIP(isoform B)}}$$

3 Results

We present an RNA-BisSeq data analysis package, named Episo, to quantify m^{5}C at the transcript isoform level. Episo consists of three tools: named *mapper*, *quant* and *Bisulfitefq*, for mapping, quantifying and simulating RNA-BisSeq data, respectively (Fig. 1). The detailed algorithm can be found in the Section 2.

3.1 *In silico* assessment of Episo

To *in silico* assess the performance of Episo, we simulated paired-end RNA-BisSeq data with three global methylation rates, 0.1, 1 and 10% (about 23 million 101-bp length paired-end reads for each methylation rate) using *Bisulfitefq*. Comparing to meRanTK, the mapping rates of *mapper* were consistently higher at all three methylation rates tested (86.6% versus 80.72%). Next, we assessed the accuracy of *quant* at global and single-nucleotide level. The *quant* accurately estimated the relative global RNA m^{5}C rates at all nucleotide contexts, i.e. CpG, CHG and CHH, and at all the three methylation rates (Supplementary Table S2). Because Episo, to the best of our knowledge, is the first computational tool that enables isoform level quantification of m^{5}C, there are no existing methods to compare with. Thus, to further assess the accuracy of *quant* at isoform level, we simulated RNA-BisSeq data with 10, 5 and 1% m^{5}C isoform level methylation (Fig. 2A). These three methylation levels were examined because most methylated transcript isoforms were estimated carrying 1–10 percent m^{5}C in real samples (Fig. 2B). Because the m^{5}C level is so low, when gene expression level is also low, the data would be too noisy to made meaningful predictions, e.g. the average differences are 0.2076, 0.4211 and 0.4765 compared with simulated levels of 10, 5, 1%, respectively (Supplementary Fig. S1). We only considered the transcripts that have sufficient expression level, i.e. fragments per kilobase per million reads (FPKM) >2 in our tests. We found that the average differences between estimated and simulated m^{5}C levels were minor (the average differences are 0.0070, 0.0013 and 0.0006 compared with simulated levels of 10, 5, 1%, respectively, Fig. 2A). To assess the accuracy of *quant* at single-nucleotide level, we simulated RNA-BisSeq data with 60, 40 and 10% level single-nucleotide m^{5}C methylation. These three methylation levels were tested because most methylated cytosines have an m^{5}C level between 10 and 60% in real samples (Fig. 2C). In this dataset, we also found that the average differences between the estimated and simulated RNA m^{5}C levels were nearly zero (the average differences are 0.0267, 0.0113 and 0.0013, when the simulated levels are 60, 40 and 10%, respectively, Fig. 2A). Last, Episo was found to predict isoform m^{5}C levels specifically and sensitively in tested methylation level (Supplementary Fig. S2).

3.2 Experimental assessment of Episo

To experimentally assess the accuracy of Episo's prediction by MeRIP followed by qPCR (Fig. 2D and Supplementary Fig. S3), we applied Episo to RNA-BisSeq data of HeLa cells and predicted m^{5}C level of transcripts at isoform level. To distinguish the methylation levels between isoforms easily, we looked for genes satisfied all the following conditions for experimental validation. First, this gene should have and only have one Guanine site that its corresponding cytosine site in the RNA production be methylated. Second, the gene generates at least two isoforms that carrying the methylated site. Third, the predicted methylation level at the site is no less than 0.1 in tested isoforms, because the accuracy of current RNA m^{5}C examination technology, i.e. MeRIP were limited when the methylation level is low. Fourth, there is at least one unique exon to enable proper primers design distinguishing the two tested isoforms (Fig. 2D). We identified seven sites after filtering (Supplementary Table S3). Finally, we picked three genes (PIK3R2, TUBGCP2 and PRKCA). The rest four genes are excluded from experimental validation because of unsuccessful primer design (STK32C and COPS7A) or having too distinguished expression levels between the isoforms (fold change > 30, RALY and GPAA1).

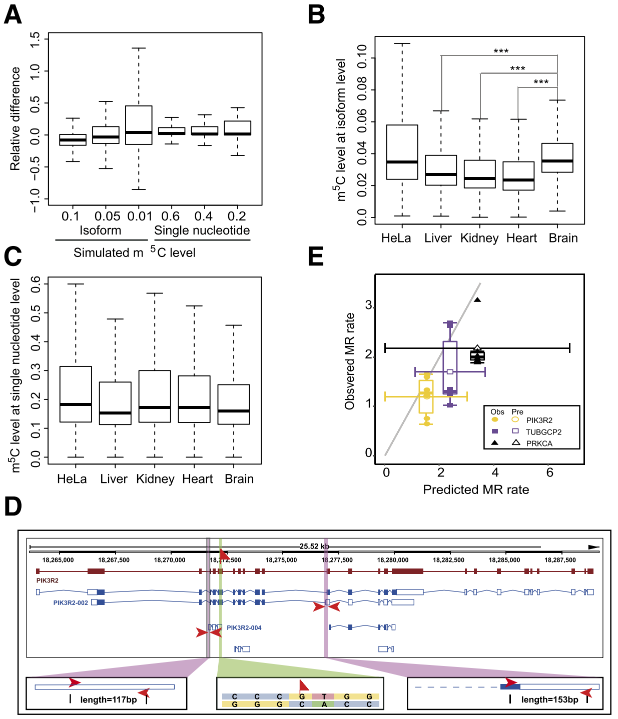

Fig. 2. The accuracy of Episo and the distribution of estimated m^{5}C levels in human cells and mouse tissues. (**A**) The distribution of estimation errors of Episo. The difference of m^{5}C levels between the simulated and Episo estimated data is shown at the resolution of RNA isoforms and single cytosine. At both resolutions, the comparisons were made at three m^{5}C levels that covered the main range of estimations in real data from human cells and mouse tissues as shown in (**B**) and (**C**). (B) The distribution of estimated m^{5}C levels at the resolution of RNA isoforms. In the mouse tissues tested, the m^{5}C level is significantly higher in brain than that in the other three tissues. The '***' indicates significant difference of P-value < 0.001. (C) The distribution of estimated m^{5}C levels at the resolution of single cytosine. For both metrics, we found no significant different among mouse tissues tested. (**D**) Experiment design for PIK3R2. The unique exons for isoform PIK3R2-002 and PIK3R2-004 are marked in pink shadow and the primers design are indicated as red arrows. The unique methylated cytosine is marked with red flag. See Supplementary Figure S3 for PRKCA and TUBGCP2. (**E**) The comparison of the predicted MR ratio with the observed MR ratio. The hollow points and horizontal whiskers represent the expectation and 95% confident interval of predicted MR rate, respectively. The Y coordinates for the hollow points are the mean of experimental data. The solid points and vertical boxplots represents the experimental measured MR rates and their distributions of six replicates (three independent experiments with two replicates each), respectively. The diagonal line is $Y = X$. (Color version of this figure is available at *Bioinformatics* online.)

The experimental observed m^{5}C levels have consistent trends with the predictions (Fig. 2E). We measured the methylation ratio (MR) between the isoforms with three independent experiments (each independent experiment had two replications) using a modified methyl-RNA immunoprecipitation (meRIP) assay (Section 2). The order of average MR ratio in the experiments are PIK3R2

(1.195) < TUBGCP2 (1.701) < PRKCA (2.18), which was rather close to Episo's predictions of PIK3R2 (1.507) < TUBGCP2 (2.3865) < PRKCA (3.3903) (Fig. 2E). However, the 95% confident interval of the predictions are large, particular for the PRKCA, which need to be improved in the further. Nevertheless, our experimental data suggested that Episo can finely predict m^{5}C level at isoform level from BisSeq-seq data.

Taken together, the assessment using both simulated and wet experimental data suggested that Episo is a fine RNA-BisSeq data analysis tool and it can estimate m^{5}C level at global, isoform, exons and single-nucleotide level.

3.3 Uneven m^{5}C distribution among tissues at isoform level

To explore the distribution of m^{5}C level in real samples, we applied *quant* to recently published RNA-BisSeq data in human (HeLa) and mouse (liver, kidney heart and brain). At the isoform level, although the overall m^{5}C is low in all the samples tested, it is indeed higher in HeLa cells and the mouse brains than other tissues and the difference is moderate (The median m^{5}C level in isoforms of HeLa cells, mouse liver, kidney, heat and brain were estimated as 0.035, 0.027, 0.025, 0.023 and 0.035, respectively) (Fig. 2B). Intriguingly, the distribution of m^{5}C at single-nucleotide level is undistinguishable in all mouse tissues and HeLa cells (Fig. 2C), suggesting the presence of more methylated cytosine sites in HeLa cells and mouse brain compared to other mouse tissues.

It is well known that the transcriptome pattern of brain is distinct from other tissues in mouse (Zheng-Bradley *et al.*, 2010). Therefore, we asked if the distinct m^{5}C pattern in mouse brain stemmed from gene expression level per se. We found a weak negative correlation between RNA m^{5}C at isoform level and its expression (Spearman's R = -0.31, $P < 2.2e\text{-}16$) in mouse brain data. However, this weak correlation may only hold for transcripts with relatively low m^{5}C level because when we limit the analysis to the region of m^{5}C larger than 0.3, the correlation disappeared (Spearman's $R = 0.15$, $P = 0.26$). A similar pattern can be found in other mouse tissues and HeLa cells (Supplementary Fig. S4). Therefore, the distinct m^{5}C pattern in mouse brain and HeLa cells may be largely independent of gene transcription activity.

3.4 A fraction of cytosines are methylated specifically towards certain isoforms and CG dinucleotide

Next, we asked whether m^{5}C was evenly distributed in genome. If m^{5}C, in general, is subjected to a certain regulation, we reasoned that the distribution of m^{5}C over isoforms, or between genomic contexts, should deviate from random expectation. Thus, we compared the diversities of m^{5}C as estimated by Episo in real data to the random controls, which were simulated with comparable total expression and methylation levels (about 0.1%) using *Bisulfitefq*.

First, we found that the RNA m^{5}C in real data is less evenly distributed over isoforms than random expectation (Fig. 3A). We employed the coefficient of variation (CV) to index the diversity. The CV is a statistic measures the diversity of a distribution, and is defined as the ratio of the standard deviation σ to the mean μ of the distribution and a distribution with a larger CV indicates that it is more diverse than distributions with smaller CVs. We compared the distribution diversity of m^{5}C over isoforms between real data and randomly simulations. The percentages of mRNAs that carrying m^{5}C are comparable between real and simulated control (29, 27, 34, 31 and 40% of mRNAs in the HeLa, liver, kidney, heart and brain, respectively, compared to about 40% in the control, Fig. 3B). However, for those methylated mRNA in real data, their diversity is much higher than that in control (Fig. 3B and C). The average CV of m^{5}C in HeLa, liver, kidney, heart and brain was 2.28, 2.06, 2.04, 2.08 and 1.94, respectively, higher than the expected CV in simulation data (1.34, KS test's *P*-value <2.2e-16 for all the samples tested). The largest CV in simulation was less than three, while the distribution in real tissues was heavily tailed towards larger CVs. For example, 19, 9, 8, 9 and 7% genes for HeLa, liver, kidney, heart and brain, respectively, had CVs larger than 3 (Fig. 3D). This results suggest, at least, these fraction of cytosines are methylated specifically towards certain isoforms.

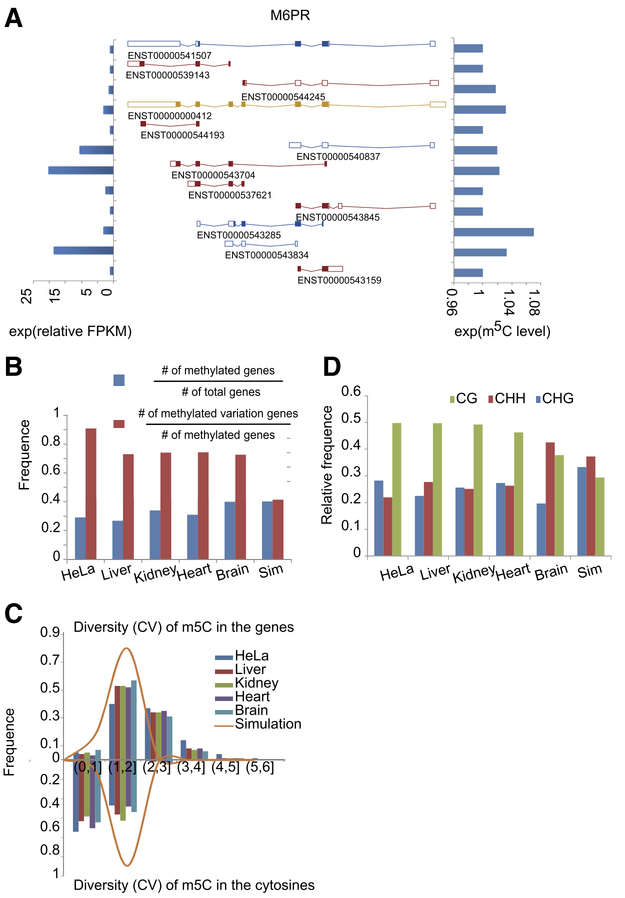

Fig. 3. Diversity of m^{5}C levels. (**A**) The distribution of expression and m^{5}C levels between RNA isoforms of the M6PR gene in HeLa cells. The relative FPKMs were calculated as the absolute FPKMs of isoforms divided by the average FPKM of the isoforms of M6PR. (**B**) The proportion of m^{5}C-containing and m^{5}C-variable genes. The blue bars represent the proportion of genes that have at least one m^{5}C- containing mRNA over all protein coding genes in the human HeLa cells, four mouse tissues and simulated data (sim). The red bars represent the proportion of genes with diversity of m^{5}C CV > 1 over that of all genes represented by blue bars. (**C**) The proportion of diversity singleton sites (CV > 1) in the genomic context of CG, CHG and CHH. (**D**) The distribution of diversity of m^{5}C at isoform and single-cytosine resolution. The orange curve represents the expected distribution of CV as obtained by random shuffling of real data. (Color version of this figure is available at *Bioinformatics* online.)

Next, we wonder whether this larger than expected CVs were also true at single-nucleotide level. Because the CV will be zero if a m^{5}C site was methylated in one isoform, we divided m^{5}C sites into two classes. The two classes, named singletons and multitons, contain the m^{5}C sites that be methylated in only one isoform and multiple isoforms, respectively. We noticed that the CVs in multitons are much smaller than control (Fig. 3C). The average CV of multitons in HeLa, liver, kidney, heart, brain and control was 0.87, 0.97, 0.99, 0.98, 1.02 and 1.42 (KS test, $P < 2.2e\text{-}16$ for all tissues), respectively. This lower than expected CV implies that, if a cytosine is methylated in multiple isoforms, the methylation tends to be even.

There are 991, 870, 872, 1261 and 2959 singletons in HeLa, mouse liver, kidney, heart and brain, respectively. We asked if they are biased to certain nucleotide type. By comparing the relative frequencies of m^{5}C at three di-/tri-nucleotide (CG, CHH and CHG) contexts between these singletons and control, we detected a strong bias towards CG in all tissue types we tested (Fig. 3C). And m^{5}C is also enriched at CHH in brain. GO analysis of genes with CV > 0 showed that they were enriched for several post-translational modifications, while motif analysis on singletons showed that they were enriched in the simple repeats region (Supplementary Figs S5 and S6). Together, at the whole gene level, the above analysis implies that the distribution of m^5C may be isoform-specific, while at single-nucleotide level, it may be more CG specific.

4 Discussion

In this work, we described a computational tool, Episo, to quantify the RNA m^5C at isoform level and single cytosine sites. The ability to distinguish m^5C level between isoforms of the same gene distinguishes Episo from existing tools, e.g. meRanTK. To the best of our knowledge, Episo is currently the only method with this feature. The accuracy of Episo mainly depends on the expression level of the host gene and the methylation level per se (Fig. 2). Although the expected estimation error is almost zero, variation cannot be completely ignored. Given the low m^5C level in most real samples, ultra-high sequencing depth would be recommended. A recent work by Zhang and colleagues showed that there might be technical biases in RNA-BisSeq data (Huang *et al.*, 2019). Thus, further development is needed, computationally and experimentally, to reduce the variations and the costs in the RNA-BisSeq analysis, respectively.

Episo utilizes RNA-BisSeq data for m^5C estimation, even though the most abundant modification type is N6-methyladenosine (m^6A) (Wang *et al.*, 2016), not m^5C, in mammals. Episo may not the best to handle m^6A data, as current assays for m^6A detection are mostly antibody-based and seldomly reach single-nucleotide resolution (Zhou *et al.*, 2016). However, we foresee the use of Episo for m^6A in the future, most likely by taking advantage of rapidly evolving third-generation sequencing technologies. With the help of the third-generation sequencing technology, chemical modifications on DNA could now potentially be directly identified (Shendure *et al.*, 2017). However, the two most popular third-generation sequencing technologies, Pacbio (Eid *et al.*, 2009) and Nanopore (Branton *et al.*, 2008), suffer from severe sequencing errors. In real practice, data correction using second-generation data is a widely used strategy for both technologies (Mahmoud *et al.*, 2017). Thus, when the third-generation sequencing technology has evolved to the point where it can directly identify chemical modifications on RNAs with short reads, Episo might be a powerful tool for quantifying m^6A and all detectable modifications at isoform level.

We applied Episo to RNA-BisSeq data from human HeLa cells and mouse liver, kidney, heart and brain samples. We showed that RNA m^5C at isoform level tends to be tissue-specific, particularly in brain sample and HeLa cells. Evidence suggested that not all m^5Cs are randomly methylated, implying a potential layer of regulation in the m^5C program. The biological significance of low-level chemical modifications detected in transcripts is always a concern (Agris, 2015; Song and Yi, 2017). In general, the present analysis showed that some m^5C sites do not mimic the distribution from pure stochastic sampling. It is this portion of m^5C sites that will draw much more attention in future functional investigation.

The current version of Episo only takes into account the static m^5C level in a given sample. If the m^5C program is indeed involved in important biological processes, dynamic changes of m^5C level between samples should be expected (Supplementary Text, Fig. S7and Table S4). Thus, a quantitative model for differential methylations is needed in the future.

Acknowledgements

David Martin performed English language editorial services.

Author contributions

Z.Z. conceived this project. J.F.L. developed Episo, J.F.L. and Z.Z. analyzed data, Z.A., J.L. and F.L. performed the experiments, and Z.Z., F.L. and J.F.L. prepared the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Nature Science Foundation of China [91940304, 91540114, 31671342, 31871331, 31271398].

Conflict of Interest: none declared.

References

Agris,P.F. (2015) The importance of being modified: an unrealized code to RNA structure and function. *RNA*, **21**, 552–554.

Amort,T. *et al.* (2017) Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain. *Genome Biol.*, **18**, 1.

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.

Blanco,S. *et al.* (2014) Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *EMBO J.*, **33**, 2020–2039.

Bormann,F. *et al.* (2018) BisAMP: a web-based pipeline for targeted RNA cytosine-5 methylation analysis. *Methods.*, **156**, 121–127.

Branton,D. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, **26**, 1146–1153.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

Burgess,A.L. *et al.* (2015) Conservation of tRNA and rRNA 5-methylcytosine in the kingdom Plantae. *BMC Plant Biol.*, **15**, 199.

Chen,X. *et al.* (2019) 5-Methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nat. Cell Biol.*, **21**, 978–990.

Chi,K.R. (2017) The RNA code comes into focus. *Nature*, **542**, 503–506.

Dennis,G.,Jr. *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Edelheit,S. *et al.* (2013) Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m^5C within archaeal mRNAs. *PLoS Genet.*, **9**, e1003602.

Eid,J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Frye,M. *et al.* (2016) RNA modifications: what have we learned and where are we headed? *Nat. Rev. Genet.*, **17**, 365–372.

Gabriel Torres,A. *et al.* (2014) Role of tRNA modifications in human diseases. *Trends Mol. Med.*, **20**, 306–314.

Huang,T. *et al.* (2019) Genome-wide identification of mRNA 5-methylcytosine in mammals. *Nat. Struct. Mol. Biol.*, **26**, 380–388.

Hussain,S. *et al.* (2013a) Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol.*, **14**, 215.

Hussain,S. *et al.* (2013b) NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep.*, **4**, 255–261.

Jayaseelan,S. *et al.* (2011) RIP: an mRNA localization technique. *Methods Mol. Biol.*, **714**, 407–422.

Khoddami,V. and Cairns,B.R. (2013) Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.*, **31**, 458–464.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

Legrand,C. *et al.* (2017) Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. *Genome Res.*, **27**, 1589–1596.

Liang,F. *et al.* (2016) BS-RNA: an efficient mapping and annotation tool for RNA bisulfite sequencing data. *Comput. Biol. Chem.*, **65**, 173–177.

Mahmoud,M. *et al.* (2017) Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. *Genomics.*, **111**, 43–49.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10–12.

Members,B.I.G.D.C. (2017) The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res.*, **45**, D18–D24.

Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**.

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

Popis,M.C. *et al.* (2016) Posttranscriptional methylation of transfer and ribosomal RNA in stress response pathways, cell differentiation, and cancer. *Curr. Opin. Oncol.*, **28**, 65–71.

Rieder,D. *et al*. (2016) meRanTK: methylated RNA analysis ToolKit. *Bioinformatics*, **32**, 782–785.

Schaefer,M. *et al*. (2010) RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev*., **24**, 1590–1595.

Schwartz,S. *et al*. (2013) High-resolution mapping reveals a conserved, widespread, dynamic mrna methylation program in yeast meiosis. *Cell*, **155**, 1409–1421.

Shelton,S.B. *et al*. (2016) Who watches the watchmen: roles of RNA modifications in the RNA interference pathway. *PLoS Genet*., **12**, e1006139.

Shendure,J. *et al*. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345–353.

Song,J. and Yi,C. (2017) Chemical modifications to RNA: a new layer of gene expression regulation. *ACS Chem. Biol*., **12**, 316–325.

Squires,J.E. *et al*. (2012) Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res*., **40**, 5023–5033.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–U174.

Wang,X. *et al*. (2016) Structural basis of N-6-adenosine methylation by the METTL3-METTL14 complex. *Nature*, **534**, 575.

Yang,X. *et al*. (2017) 5-Methylcytosine promotes mRNA export-NSUN2 as the methyltransferase and ALYREF as an m^{5}C reader. *Cell Res*., **27**, 606.

Yang,Y. *et al*. (2019) RNA 5-methylcytosine facilitates the maternal-to-zygotic transition by preventing maternal mrna decay. *Mol. Cell*., **75**, 1188–1202.e1111.

Yates,A. *et al*. (2016) Ensembl 2016. *Nucleic Acids Res*., **44**, D710–D716.

Zhao,B.S. *et al*. (2017) Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol*., **18**, 31–42.

Zheng-Bradley,X. *et al*. (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*., **11**, R124.

Zhou,Y. *et al*. (2016) SRAMP: prediction of mammalian N-6-methyladenosine (m(6)A) sites based on sequence-derived features. *Nucleic Acids Res*., **44**, e91.