Estimate the Proportion of Non-Specific Binding in MeRIP-Seq Data

Jingxuan Bao[1], Zhen Wei[2], Jia Meng[2], Jionglong Su[1, *]

[1]Department of Mathematical Sciences, [2]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; *To whom correspondence should be addressed: jionglong.su@xjtlu.edu.cn (JS)

*Email: jionglong.su@xjtlu.edu.cn

Abstract

The ISOpureR method uses the gradient decent method to maximize the complete likelihood function of estimating the proper estimators. In this research, we estimate the absolute proportion of m^{6}A specific binding mRNA in the m^{6}A-containing sample (IP sample) with the application of ISOpureR method. We apply the ISOpureR method to simulated data and real data. Our results show that the method performs well under the model of second order median regression in terms of the simulated data with application of regression analysis; and it needs to be further developed when applied to the real data.

1 Introduction

RNA sequencing (RNA-Seq), also known as whole transcriptome shotgun sequencing, is a common method to detect differences in the gene expression in different samples, treatments, as well as different cell populations and experimental conditions. With the sequencing data, bio-statisticians may carry out more accurate predictions of the RNA methylation sites and differential RNA methylation analysis.

Usually, the prediction of RNA methylation sites is carried out using peak calling, a computational method applied to identify regions in a genome with aligned reads enrichment after performing a ChIP-sequencing or MeDIP-seq experiment. With the improvement of high throughput sequencing data, differential RNA methylation analysis is now available with the help of RNA methylation experiments. Nowadays, N^6-methyladenosine (m^{6}A), an abundant modification in mRNA found in some viruses [1, 2] as well as most eukaryotes [2, 3], becomes increasingly important in various biological processes, such as RNA degradation and RNA-protein interaction [4].

In this paper, in order to be integrated with studies on RNA methylation site detection and differential methylation analysis, ISOpureR, a method of estimating the proportion of m^{6}A specific binding mRNA fragments in the m^{6}A-containing sample (IP sample) is introduced and improved upon using statistical approach to calculate the absolute proportion of m^{6}A specific binding mRNA.

2 Methodology

In this section, we discuss the statistical model ISOpureR. The full ISOpureR model is described below with the parameters [5], noting that all the symbols in bold font refer to a vector or a matrix

$$\boldsymbol{t_n} = \alpha_n \boldsymbol{c_n} + (1-\alpha_n)\boldsymbol{h_n} + \boldsymbol{e_n} \tag{1}$$

$$= \alpha_n \boldsymbol{c_n} + \sum\nolimits_{r=1}^{R} \theta_{n,r}\, \boldsymbol{b_r} + \boldsymbol{e_n} \tag{2}$$

where $\boldsymbol{t_n}$ represents the reads count of IP samples where $n = 1,2,\dots,N$; $\boldsymbol{c_n}$ represents the reads count of pure m^6A binding mRNA samples where $n = 1,2,\dots,N$; $\boldsymbol{h_n}$ represents the reads count of normal samples where $n = 1,2,\dots,N$; $\boldsymbol{b_r}$ represents the reads count of input samples which contain very little m^{6}A binding mRNA fragments where $r = 1,2,\dots,R$; α_n represents the absolute proportion of m^{6}A binding mRNA fragments in the IP samples where $n = 1,2,\dots,N$; $\boldsymbol{e_n}$ represents the error following the normal distribution [6] where $n = 1,2,\dots,N$. Here, we assume that $\boldsymbol{h_n}$ can be represented by the combination of input samples, $\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_R}$, provided to the algorithm.

According to Quon *et al.*, to rescale the IP samples so that the total number of observations (the sum of the elements) is approximately the same across all IP samples, and to balance the influence that each IP sample has on the parameters, we discretize each IP samples, $\boldsymbol{t_n}$, to round every element of $\boldsymbol{t_n}$ to the nearest non-negative integer to obtain transformed IP samples $\boldsymbol{x_n}$. Next, we define the $\boldsymbol{\hat{x}_n}$ to be the normalized reconstruction of the IP profiles. The probability of the discretized IP profiles, which is a normalised reconstruction of the IP profile $\boldsymbol{x_n}$ based on the model parameters, follows the multinomial distribution [5]. After discretization, the equation (2) becomes

$$\boldsymbol{\hat{x}_n} = \alpha_n \boldsymbol{c_n} + \sum\nolimits_{r=1}^{R} \theta_{n,r}\, \boldsymbol{b_r} + \boldsymbol{e_n} \tag{3}$$

Now, our task is to estimate α_n, the unknown proportion of pure m^{6}A binding mRNA in the IP samples, by maximizing the complete likelihood function using the gradient decent method.

The full ISOpureR method is defined by Quon *et al.* [5] as follows:

$$\boldsymbol{B} = [\boldsymbol{b_1}\ \boldsymbol{b_2} \dots \boldsymbol{b_r}]$$

$$\boldsymbol{\theta_n} = [\theta_{n,1}\ \theta_{n,2} \dots \theta_{n,r}\ \alpha_n]$$

$$\boldsymbol{x_n} = [\boldsymbol{B}\ \boldsymbol{c_n}]\ \boldsymbol{\theta_n}$$

$$p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) = Dirichlet(\boldsymbol{m}|k'\boldsymbol{B\omega})$$

$$p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) = Dirichlet(\boldsymbol{c_n}|k_n\boldsymbol{m})$$

$$p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) = Multinomial(\boldsymbol{x_n}|\boldsymbol{\hat{x}_n})$$

$$p(\boldsymbol{\theta_n}|\boldsymbol{v}) = Dirichlet(\boldsymbol{\theta_n}|\boldsymbol{v})$$

where the probability mass functions of the Dirichlet distribution and the Multinomial distribution are in the form below

$$Multinomial(\boldsymbol{\gamma}|\boldsymbol{\pi}) = \frac{(\sum_{k=1}^{K}\gamma_k)!}{\prod_{k=1}^{K}\gamma_k!}\prod_{k=1}^{K}\pi_k^{\gamma_k} \tag{4}$$

$$Dirichlet(\boldsymbol{x}|\boldsymbol{a}) = \frac{\Gamma(\sum_{k=1}^{K}a_k)}{\prod_{k=1}^{K}\Gamma\,(a_k)}\prod_{k=1}^{K}x_k^{a_k-1} \tag{5}$$

To estimate the parameters, we define our complete likelihood function as follows

$$L = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega})\prod_{n=1}^{N}\emptyset_{\boldsymbol{n}} \tag{6}$$

where

$$\emptyset_n = p(\boldsymbol{c_n}|k_n, \boldsymbol{m})p(\boldsymbol{\theta_n}|\boldsymbol{v})p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) \tag{7}$$

When maximizing the complete likelihood function, Quon *et al.* first transform the problem into one minimizing the minus log-likelihood function [5]. After transformation, they use the Polack-Ribiere flavor of conjugate gradients method [7] to search directions, and a line search with application of quadratic and cubic polynomial approximations. The Wolfe-Powell stopping criteria [8] is used together with the slope ratio method for guessing initial step sizes.

3 Results and Discussions

In this section, we apply ISOpureR method to estimate the proportion of two datasets, the simulated dataset and the real dataset respectively. After estimation, we analyze the results we obtained.

3.1 Simulated Data

3.1.1 Simple Extreme Data

We have generated a simply extreme simulated input profile and a simple extreme simulated pure m^6A specific binding profile. There are a total of 500 gene sites for each profile, and we set the reads count of the first 250 gene sites of input profile to be 1000 and the rest 250 sites to be 0; for pure m^6A specific binding profile, we set the reads count of the first 250 gene sites to be 0 and the other 250 sites to be 1000.

Next, we apply binomial distribution to generate the IP samples with different mixing proportions with respect to the input profiles, 10%, 20%, …, 90% respectively; and for each proportion, we generate three different samples.

The results of ISOpureR method estimated from the simulated data is shown in the **Table 1**, where the column names, 10%, 20%, …, 90% refer to the true proportions of pure m^6A specific binding sample in the simulated IP dataset, and the numbers refer to the estimated proportion of pure m^6A specific binding sample in the IP sample using ISOpureR method. From **Table 1**, we can say that the ISOpureR method is likely to converge to 0 and 1. To examine whether the convergence is caused by the extreme case of the dataset, we shall further examine the simulated real data.

3.1.2 Simple Real Data

In this part of the section, we generate a simple but more realistic dataset, simulated real dataset. First of all, we use the dataset of the experiment “human-A549-C” containing three IP samples and three input samples. Next, to keep the simulated dataset statistically meaningful, we generate the pure m^6A specific binding profiles by randomly permutating the input sample of experiment “human-A549-C”. We then randomly choose 500 gene sites to form our new dataset. Note that the selected gene sites of both input sample and pure m^6A specific binding sample should be exactly the same. Finally, we apply binomial distribution to generate the IP samples with different mixing proportions with respect to the input profiles, from 10%, 20%, to 90% respectively; and for each proportion, we generate 30 different IP samples.

The box plot of estimated proportion of simple real simulated data using ISOpureR method is shown in **Figure 1**,

Table 1 Results of Simple Extreme Datasets

Method	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
ISOpureR	0.000	0.000	0.000	0.000	0.002	0.286	0.573	1.000	1.000

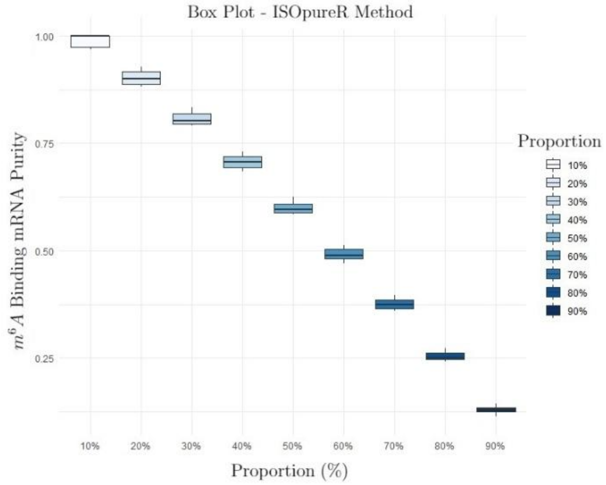

Figure 1 Box Plot of ISOpureR Method

where the *x*-axis refers to the true proportion of input sample in the IP sample, and the *y*-axis refers to the estimated proportion m^{6}A specific binding mRNA in the IP samples. From the graph, the box of each proportion is relatively small. Moreover, the estimated proportions of m^{6}A specific binding mRNA is decreasing while the real mixing proportion of input sample is increasing. Consequently, ISOpureR method is likely to be stable and accurate in terms of the simulated simple real data. To further examine the performance of ISOpureR method, we apply linear regression model to fit our results.

3.1.2.1 Linear Regression Model and Discussions

Now, we assume the results gathered from ISOpureR method satisfy the linear regression model, which is

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (8)$$

with assumptions that residuals of *y* value (or the error) are normally distributed [6]. After fitting the results of ISOpureR method, the fitted model of the ISOpureR method has the coefficients that β_0 is equal to 1.123112 and β_1 is equal to -1.078470. The fitted graph of ISOpureR method is shown in **Figure 2**.

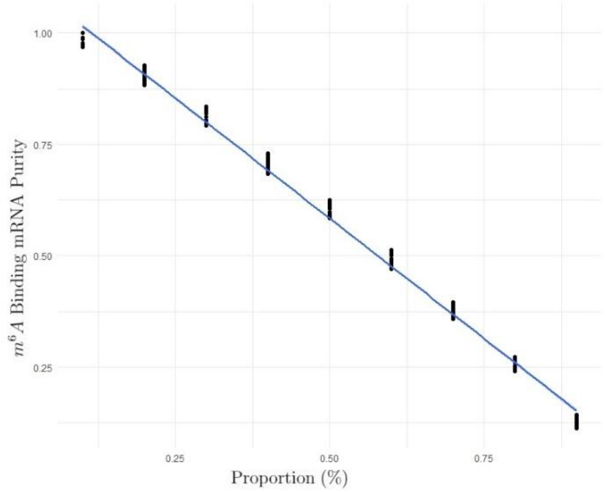

Figure 2 Linear Regression of ISOpureR Method

We next do the diagnostic test to our linear regression model. The normal quantile-quantile (QQ) plot is shown as **Figure 3**.

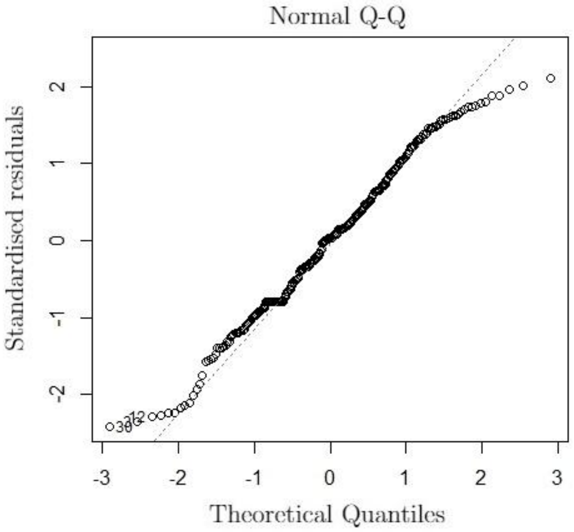

Figure 3 Normal QQ Plot of ISOpureR Method

The deviation from straight line in the normal QQ plot in Figure 3 suggests that the residuals of the linear regression model may not assume normality. After applying of the Shapiro-Wild Test [9] with null hypothesis that the sample came from a normally distributed population and alternative hypothesis that the sample did not come from a normally distributed population. The p-value of Shapiro-Wild Test statistic is equal to 1.365×10^{-5}, suggesting very strong evidence against that the null hypothesis that residuals of the linear regression model are normally distributed. Therefore, the assumption of linear regression model fails.

3.1.2.2 Median Regression Model

In the previous part of the section, we apply the linear regression model to our results. However, the normal QQ plot with Shapiro-Wild Test implies that the residuals do not follow normal distribution which undermines our assumption of linear regression model. Here, we further use the median regression to model our results since according to Furno and Vistocco [10], the median regression does not need the assumption of normality of residuals.

We assume the results of ISOpureR method satisfy the median regression model, which is

$$y = \alpha_0 + \alpha_1 x + F^{-1}(0.5) \qquad (9)$$

where α_0 and α_1 are the coefficients of the median regression model; *F* represents the common distribution function of the errors without any distributional assumptions [10]; and 0.5 refers to the median value. After calculation using the method of least absolute deviance (LAD), we obtain the coefficients α_0 equals 1.12364 and $\alpha_1$equals -1.07808.

Furthermore, we carry out the lack-of-fit test using the method introduced by He and Zhu [11] to our second order median regression model, with null hypothesis that higher order terms are not required to adequately fit our data, and alternative hypothesis that higher order terms are required to adequately fit our data. The p-value of the test is less than 2.2×10^{-16}, suggesting we have very strong evidence against the null hypothesis that higher order terms are not required to adequately fit our data.

3.1.2.3 Second Order Median Regression Model

In the previous part of the section, we examined that the median regression cannot adequately fits our data. We next use the second order median regression to model our results. We assume the results of ISOpureR method satisfy the second order median regression model, which is

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + F^{-1}(0.5) \qquad (10)$$

where α_0, α_1, and α_2 are the coefficients of the second order median regression model; F represents the common distribution function of the errors without any distributional assumptions [10]. 0.5 represents the median value. After calculation with application of LAD method, we obtain the coefficients α_0 equals 1.09092, $\alpha_1$equals -0.88933, and α_2 equals -0.19887. The plot of second order median regression model is shown in the **Figure 4**.

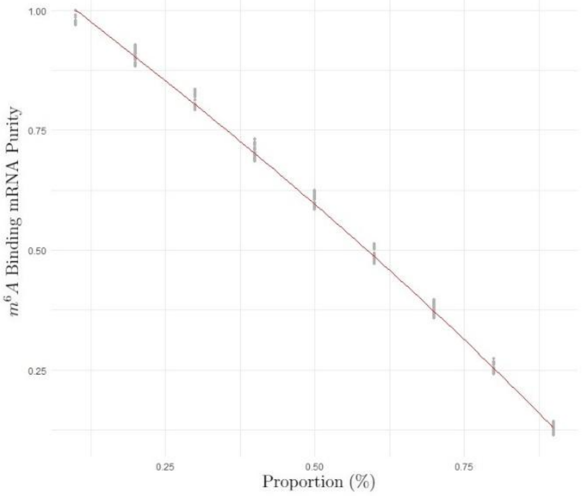

Figure 4 Second Order Median Regression of ISOpureR Method

We further carry out the lack-of-fit test using the method introduced by He and Zhu [11] to our second order median regression model, and the p-value of the test is equal to 0.21 suggesting we have no evidence against the null hypothesis that higher order terms are not required for the second order median regression model to adequately fit our data.

Next, we use the Akaike information criterion (AIC) [12], an estimator of the relative quality of statistical models for a given set of data, to test the goodness of our regression model. The AIC of median regression of is -1324.81, and the AIC of second order median regression model is -1555.036, which is less than the AIC of median regression model, suggesting that the second order median regression is able to describe our data better than the first order model without problem of overfitting.

Therefore, the estimated proportions of ISOpureR method from simulated real dataset show stable and accuracy under the second order median regression model.

3.2 Real Data

After analyzing the simulated data, we apply the ISOpureR method to our real data to estimate the true proportion of m^{6}A specific binding mRNA in the IP sample.

The raw sequencing data comes from an independent study guided by Schwartz *et al.* [13]. The raw data was downloaded from GEO with GEO accession number “human-A549-C”, aligned with Hisat2 [14] to the human reference genome assembly hg19. The aligned reads are counted by the SummarizeOverlap function in Bioconductor package GenomicAlignment. The annotation used for reads count is the single base resolution m^{6}A sites collected from m^{6}A-miCLIP [15] and m^{6}A-CLIP [16] datasets. Only the m^{6}A sites mapped to DRACH motifs are kept, and the sites are extended into 101 bp windows centered by m^{6}A sites for reads counting.

The data we used has in total 69946 gene sites. When we calculating the results of real data, we first randomly select 500 gene sites as a sample and apply ISOpureR method to estimate the true proportion of m^{6}A specific binding mRNA in selected samples. After completing 100 samples of estimation, we calculate the mean and median of the 100 results as final output. The output of estimated proportion of real data by applying the ISOpureR method is shown in the **Table 2**.

From **Table 2**, we observe that almost all the estimated proportions converge to 1. If the result is accurate, as examined when applying the method to the simulated simple real datasets, it implies that almost all the experiments has pure m^{6}A specific binding mRNA sample. Otherwise, when we estimating the proportion, some features that are influential to the results may not be included, for example, GC content bias, implying that we need to further develop the ISOpureR method.

4 Conclusions

In this article, we are estimating the proportion of m^{6}A specific binding mRNA in the IP samples by comparing the method of estimating the proportion of pure cancer cells in the tumour samples. In order to complete the task of estimating m^{6}A specific binding mRNA, we examined the ISOpureR methods. The main idea of this method is to estimate by maximising the complete likelihood function using conjugate gradient method. After introducing the basic idea of ISOpureR method, we apply the method to both simulated data and real data. First, ISOpureR method shows unstable and inaccurate to the simulated simple extreme datasets. However, when we apply the method to the simulated real datasets, the results are suitable and adequately described by the second order median regression

Table 2: Results of Real Datasets

Experiment	**Estimated Proportion**		
human-A549-C	**SRR1182619**	**SRR1182621**	**SRR1182623**
Mean	0.9999694	0.9999694	0.9999694
Median	0.9999906	0.9999906	0.9999906

model. Then, we further apply the method to the real datasets. However, the estimated proportions are likely to converge to 1. If this result is accurate, it implies that the IP samples of the experiment contain only pure m^6A specific binding mRNA samples; otherwise, when we estimating the proportion of pure m^6A binding mRNA fragments, some features, for example, GC content bias, may not be considered. In the future, we may improve the ISOpureR method by adjusting the parameters such that it will include the GC content bias; or to find another method that is more effective to the RNA methylation data.

5 Contributions

In this research, our work is designed to examine the method of deconvolution of tumor samples data, which is ISOpureR method, for different use, estimating the proportion of non-specific binding in MeRIP-seq data. In detail, we introduce the method of regression analysis, a statistical approach, to verify the availability and stability of the method in terms of three different kinds of data, including simulated simple extreme data, simulated real data, and real data.

Funding

This work has been supported by National Natural Science Foundation of China [31671373]; XJTLU Key Programme Special Fund [KSF-T-01]; Jiangsu Six Talent Peak Program [XYDXX-118].

References

[1] K. Beemon and J. Keith, "Localization of N6-methyladenosine in the Rous sarcoma virus genome," *Journal of Molecular Biology*, vol. 113, no. 1, pp. 165-179, 1977/06/15/ 1977.

[2] R. P. Perry, D. E. Kelley, K. Friderici, and F. Rottman, "The methylated constituents of L cell messenger RNA: Evidence for an unusual cluster at the 5′ terminus," *Cell*, vol. 4, no. 4, pp. 387-394, 1975/04/01/ 1975.

[3] R. Desrosiers, K. Friderici, and F. Rottman, "Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells," (in eng), *Proceedings of the National Academy of Sciences of the United States of America*, vol. 71, no. 10, pp. 3971-3975, 1974.

[4] K. D. Meyer and S. R. Jaffrey, "The dynamic epitranscriptome: N6-methyladenosine and gene expression control," *Nature Reviews Molecular Cell Biology*, Review Article vol. 15, p. 313, 04/09/online 2014.

[5] G. Quon, S. Haider, A. G. Deshwar, A. Cui, P. C. Boutros, and Q. Morris, "Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction," *Genome Medicine*, vol. 5, no. 3, p. 29, 2013/03/28 2013.

[6] S. Prabhakaran, "Assumptions of Linear Regression," *R-Statistics*, 2016.

[7] E. Bertolazzi, "Conjugate Direction minimization," PHD Numerical optimization, DIMS, Universita di Trento, 2011.

[8] M. Hintermuller, "Nonlinear Optimization," Basic Course, Berlin School of Mathematics, Humboldt-University of Berlin, 2013.

[9] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)†," *Biometrika*, vol. 52, no. 3-4, pp. 591-611, 1965.

[10] C. Davino, M. Furno, and D. Vistocco, *Quantile Regression: Theory and Applications*. 2013.

[11] X. He and L.-X. Zhu, "A Lack-of-Fit Test for Quantile Regression," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 1013-1022, 2003/12/01 2003.

[12] Y. Sakamoto and G. Kitagawa, *Akaike information criterion statistics*. 1986.

[13] S. Schwartz *et al.*, "Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites," *Cell Reports*, vol. 8, no. 1, pp. 284-296, 2014.

[14] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nat Methods*, vol. 12, 2015.

[15] B. Linder, A. V. Grozhik, A. O. Olareringeorge, C. Meydan, C. E. Mason, and S. R. Jaffrey, "Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome," *Nature Methods*, vol. 12, no. 8, pp. 767-772, 2015.

[16] S. Ke *et al.*, "m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover," *Genes & development*, vol. 31, no. 10, pp. 990-1006, 2017.