Estimation of N^6-*Methyladenosine* (m^6A) Specific Binding mRNA Fragment Proportion

Jingxuan Bao

Xi'an Jiaotong-Liverpool University

Superviser: Jionglong Su
Co-supervisor: Jia Meng

May 7, 2019

Overview

Introduction

1) Biological Background
2) Motivation and Objectives

Methodology

1) ISOpureR Method
2) CLARKE Method

Results and Discussions

1) Simulated Data
2) Real Data

Conclusions

1) Conclusions
2) Contributions

Reference

Q & A

Introduction

Biological Background

What Is m^6A

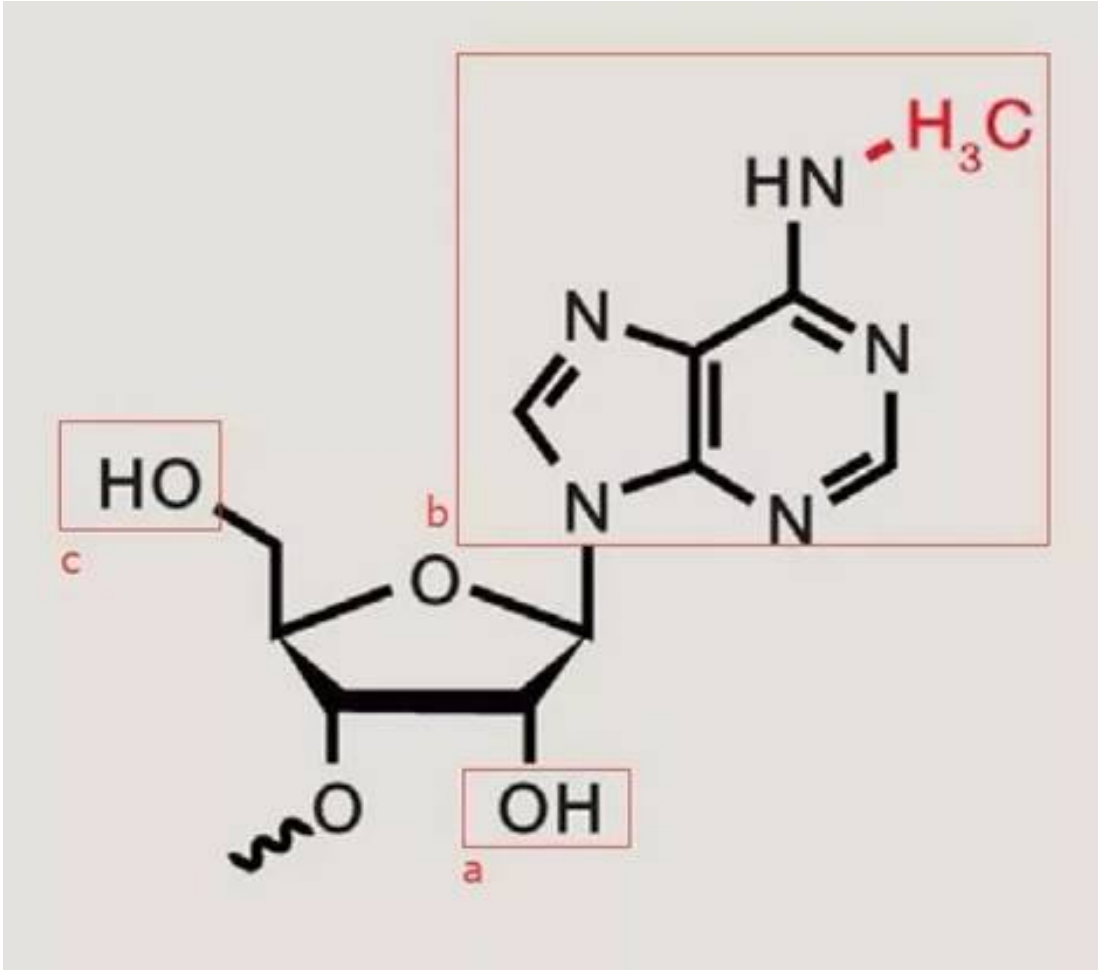
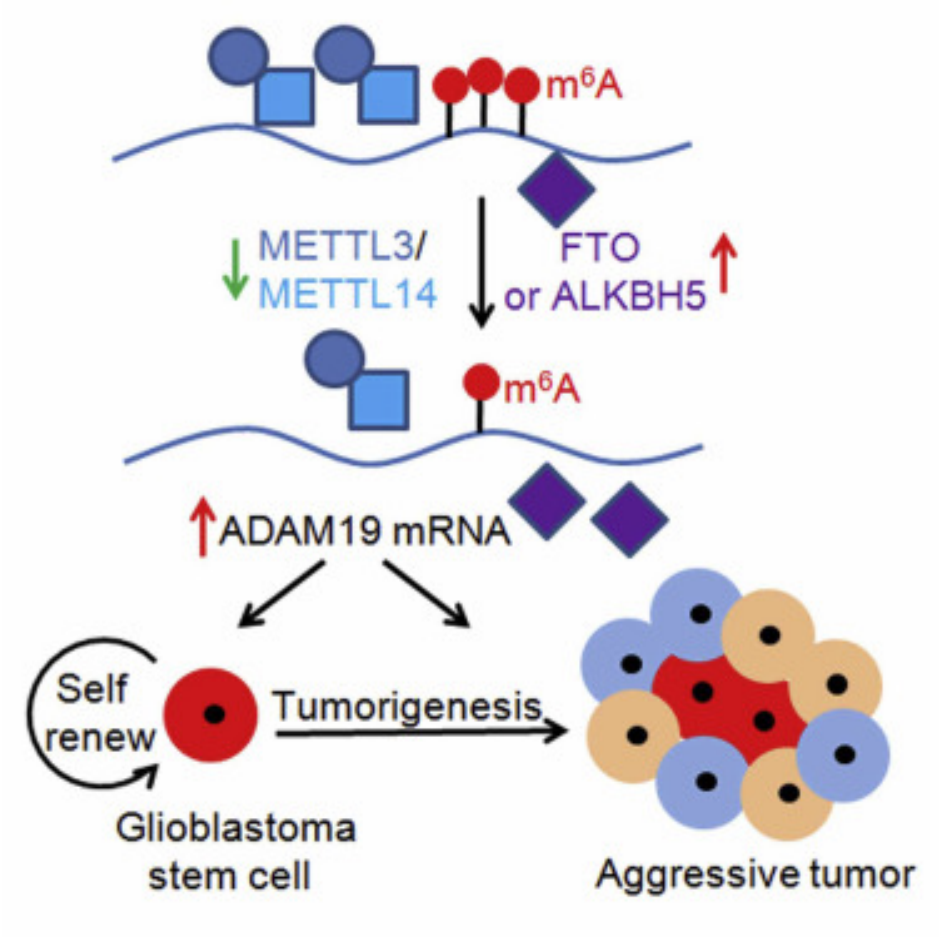


Figure: N^6-*Methyladenosine*(Roundtree et al., 2017)

Why Is It Important

RNA methylation patterns play an important biological function in the regulation of different cellular processes, such as metabolism, embryonic development, and stem cell self-renewal.

Why Is It Important



There are links between alteration in m^6A levels and abnormal cellular differentiation states present in cancer.

Figure: Self-renewal of glioblastoma stem cells is regulated by m^6A RNA methylation (Cui et al., 2017)

Motivation and Objectives

Motivation

Problems of RNA Methylation Experiment

- High Expense;
- Low output;
- ...

Alternative Way

Estimate the level of RNA methylation with application of gene expression data.

Alternative Way

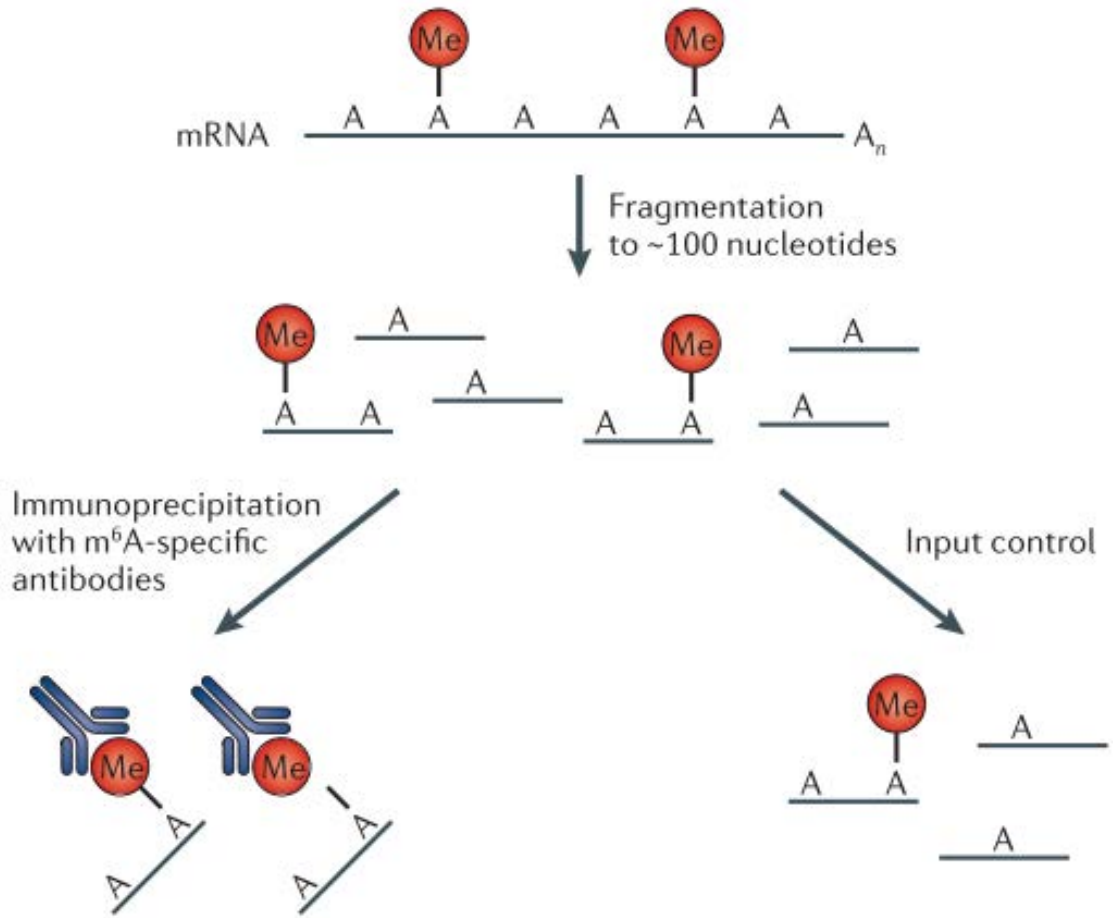


Figure: Introduction to The Gene Expression Data (Dominissini et al., 2013)

Aims and Objectives

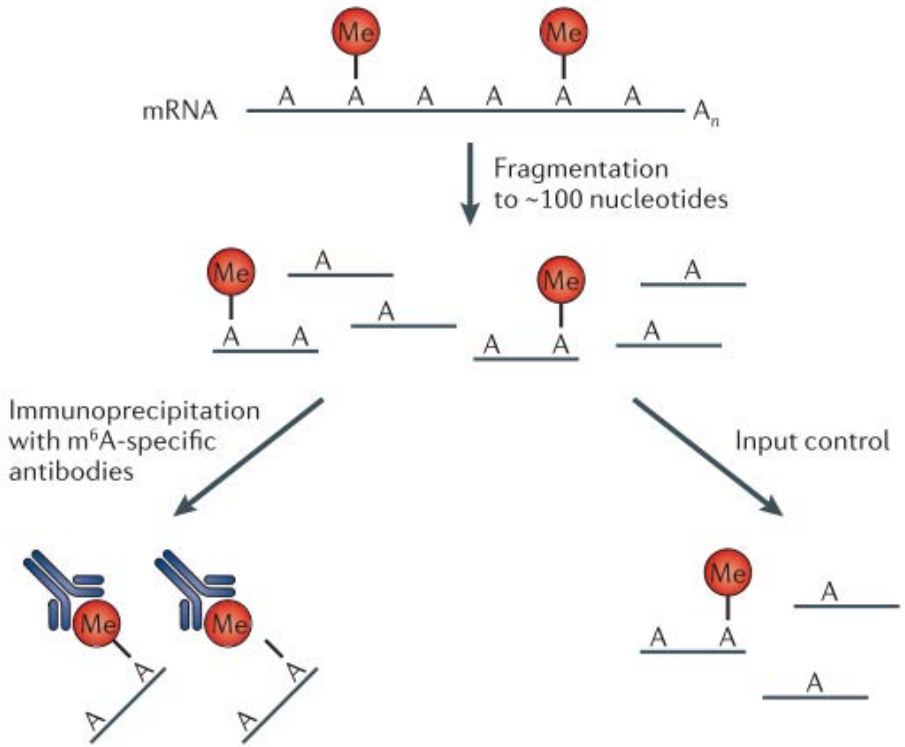


Figure: Introduction to The Gene Expression Data (Dominissini et al., 2013)

Estimation of m^6A-specific binding mRNA fragment proportion from IP sample with reference of input sample

Methodology

Estimation by Comparison

Two Problems

The problem of **deconvolution of m^6A specific binding mRNA in the IP sample** with the reference of input sample.

The problem of **deconvolution of pure cancer cell in the tumour sample** with the reference of normal sample.

Reasons

Both data are biological data, which implies we have to handle with the biological noise.

Previous Work - Tumour Purity

- ISOpureR Method
- CLARKE Method

ISOpureR Method

ISOpureR Method

Method Description

Quon et al. (2013) describe the problem as follows (all the "**bold**" symbols refer to a vector or a matrix):

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

- $\boldsymbol{t_n}$: vector of tumour sample;
- α_n: the proportion of cancer cells in the tumour sample;
- $\boldsymbol{h_n}$: vector of normal sample;
- $\boldsymbol{e_n}$: error.

ISOpureR Method

Method Description

Quon et al. (2013) describe the problem as follows (all the "**bold**" symbols refer to a vector or a matrix):

$$\boldsymbol{t_n} = \alpha_n \boldsymbol{c_n} + \sum_{r=1}^{R} \theta_{n,r} \boldsymbol{b_r} + \boldsymbol{e_n}$$

- $\boldsymbol{t_n}$: Vector of tumour sample;
- α_n: The proportion of cancer cells in the tumour cells;
- $\boldsymbol{c_n}$: Vector of pure cancer sample;
- $\boldsymbol{b_r}$: Vector of healthy profile;
- $\boldsymbol{e_n}$: Error.

ISOpureR Method

Data Transformation

Discertisation

Round each element of $\boldsymbol{t_n}$ to the nearest non-negative integer to obtain our transformed tumour profiles $\boldsymbol{x_n}$.

Reasons:

- Rescale the tumour profiles;
- Balance the influence of shared parameters.

Rescaling

Divide each normal profile $\boldsymbol{b_r}$ by the sum of its elements.

Reasons:

Allow $\boldsymbol{b_r}$ to be interpreted as a discrete probability distribution over transcripts.

Model Formulation

Method Description

Quon et al. (2013) describe the problem as follows (all the "**bold**" symbols refer to a vector or a matrix):

$$\boldsymbol{x_n} = \alpha_n \boldsymbol{c_n} + \sum_{r=1}^{R} \theta_{n,r} \boldsymbol{b_r}$$

- $\boldsymbol{x_n}$: Discretised tumour profiles;
- α_n: The proportion of cancer cells in the tumour cells;
- $\boldsymbol{c_n}$: Vector of pure cancer sample;
- $\boldsymbol{b_r}$: Vector of healthy profile.

ISOpureR Method

Model Formulation

Maximise the complete likelihood function to find the most proper estimator:

$$\mathbb{L} = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n}|\boldsymbol{v}) p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n})$$

where

- $p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) = \mathit{Dirichlet}(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega})$;
- $\boldsymbol{m}$ = Reference cancer profile estimated from the tumour profile data;
- k' = The strength parameter of the Dirichlet distribution over $\boldsymbol{m}$;
- $\boldsymbol{B} = [\boldsymbol{b_1}\ \boldsymbol{b_2}\ ...\ \boldsymbol{b_R}]$, with $\boldsymbol{b_r}$ being the vector of healthy profile;
- $\boldsymbol{\omega}$ = The weights on the normal profiles $\boldsymbol{b_r}$.

ISOpureR Method

Model Formulation

Maximise the complete likelihood function to find the most proper estimator:

$$\mathbb{L} = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n}|\boldsymbol{v}) p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n})$$

where

- $p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) = Dirichlet(\boldsymbol{c_n}|k_n\boldsymbol{m})$;
- $\boldsymbol{c_n}$ = Vector of pure cancer sample;
- k_n = Strength parameter of the Dirichlet distribution over $\boldsymbol{c_n}$ given $\boldsymbol{m}$;
- $\boldsymbol{m}$ = Reference cancer profile estimated from the tumour profile data.

ISOpureR Method

Model Formulation

Maximise the complete likelihood function to find the most proper estimator:

$$\mathbb{L} = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n}|\boldsymbol{v}) p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n})$$

where

- $p(\boldsymbol{\theta_n}|\boldsymbol{v}) = Dirichlet(\boldsymbol{\theta_n}|\boldsymbol{v})$;
- $\boldsymbol{\theta_n} = [\theta_{n,1}\ \theta_{n,2}\ ...\ \theta_{n,R}\ \alpha_n]$;
- $\boldsymbol{v}$ = Both the mean and strength of a Dirichlet distribution over $\boldsymbol{\theta_n}$.

ISOpureR Method

Model Formulation

Maximise the complete likelihood function to find the most proper estimator:

$$\mathbb{L} = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n}|\boldsymbol{v}) p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n})$$

where

- $p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) = \mathit{Multinomial}(\boldsymbol{x_n}|\hat{\boldsymbol{x_n}})$;
- $\boldsymbol{x_n}$ = Discretised tumour profiles;
- $\boldsymbol{B} = [\boldsymbol{b_1}\ \boldsymbol{b_2}\ ...\ \boldsymbol{b_R}]$, with $\boldsymbol{b_r}$ being the vector of healthy profile;
- $\boldsymbol{\theta_n} = [\theta_{n,1}\ \theta_{n,2}\ ...\ \theta_{n,R}\ \alpha_n]$;
- $\boldsymbol{c_n}$ = the cancer profiles.

ISOpureR Method

Model Formulation

To **maximise the complete likelihood function**

$$\mathbb{L} = p(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c_n}|k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n}|\boldsymbol{v}) p(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n})$$

is to **minimise the nimus complete log-likelihood function**, $-log\mathbb{L}$, which is

$$-logp(\boldsymbol{m}|k', \boldsymbol{B}, \boldsymbol{\omega}) - \sum_{n=1}^{N} \Big[logp(\boldsymbol{c_n}|k_n, \boldsymbol{m}) + logp(\boldsymbol{\theta_n}|\boldsymbol{v}) + log(\boldsymbol{x_n}|\boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) \Big]$$

by applying the flexible preconditioned conjugate gradient method.

CLARKE Method

CLARKE Method

Method Description

For the pure sample A and B, with mixed sample AB, Clarke et al. (2010) describe the problem as follows (all the "**bold**" symbols refer to a vector or a matrix):

$$E_i(AB) = p_A E_i(A) + (1 - p_A) E_i(B) + \epsilon.$$

where

- $\boldsymbol{E}(A) = (E_1(A), \cdots, E_m(A))$, gene expression of pure sample A;
- $\boldsymbol{E}(B) = (E_1(B), \cdots, E_m(B))$, gene expression of pure sample B;
- $\boldsymbol{E}(AB) = (E_1(AB), \cdots, E_m(AB))$, gene expression of mixed sample;
- $p_A =$ The true proportion of profile A in profile AB;
- ϵ: error.

CLARKE Method

Method Description

According to Gosink et al. (2008), ratio of gene expression is given by

$$R_i = \frac{E_i(AB)}{E_i(A)}$$

In the noiseless case, we have

$$R_i = p_A \frac{E_i(A)}{E_i(A)} + (1 - p_A)\frac{E_i(B)}{E_i(A)}$$

Under the assumption that $E_i(B) \to 0$, we have $min_i R_i = p_A$, that is,

$$\lim_{E_i(B)\to 0} R_i = p_A + (1 - p_A)\frac{E_i(B)}{E_i(A)} = p_A.$$

CLARKE Method

Data Transformation

Reason: With the influence of noise, the minimum ratio is likely to be **underestimated** compared with the true proportion; performance can be improved by **increasing the small ratio values** while **decreasing the large ratio values** (Gosink et al., 2007).

Transformation: According to Clarke et al. (2010), we transforming both $\boldsymbol{E}(A)$ and $\boldsymbol{E}(AB)$ into the form

$$tE_i(AB) = log(1 + \alpha E_i(AB))$$

$$tE_i(A) = log(1 + \alpha E_i(A))$$

for some $\alpha > 0$ for all i.

CLARKE Method

Model Formulation

By Clarke et al. (2010), the mean of tR_i as a function of α is defined as

$$\overline{tR_i(\alpha)} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{log(1+\alpha E_i(AB))}{log(1+\alpha E_i(A))}\right]$$

CLARKE Method

Principal Components Analysis

Clarke et al. (2010) plot the scree plot of CLARKE model,

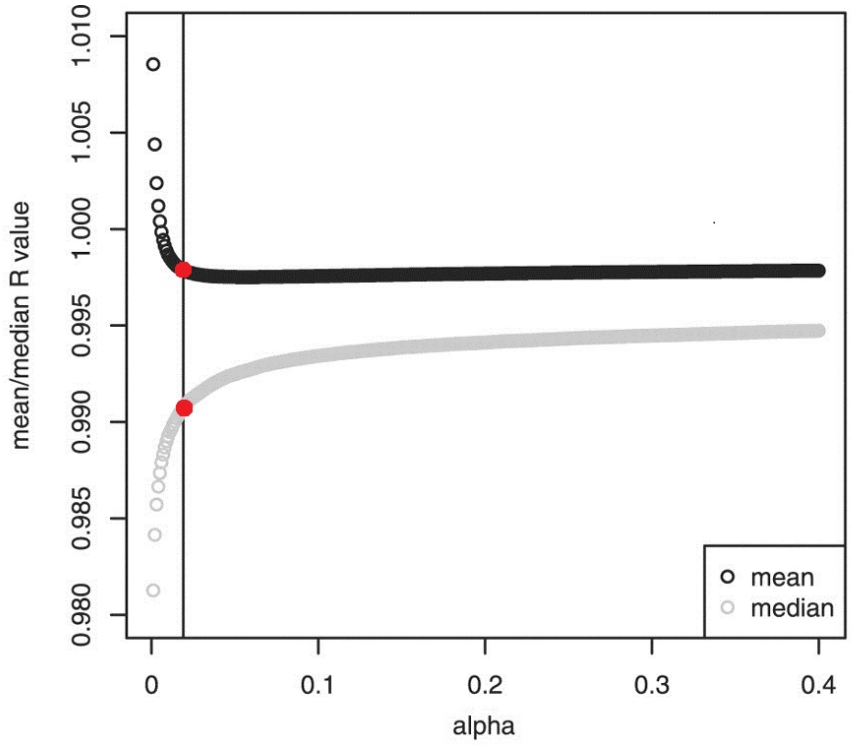


Figure: Scree Plot of the Dataset 1

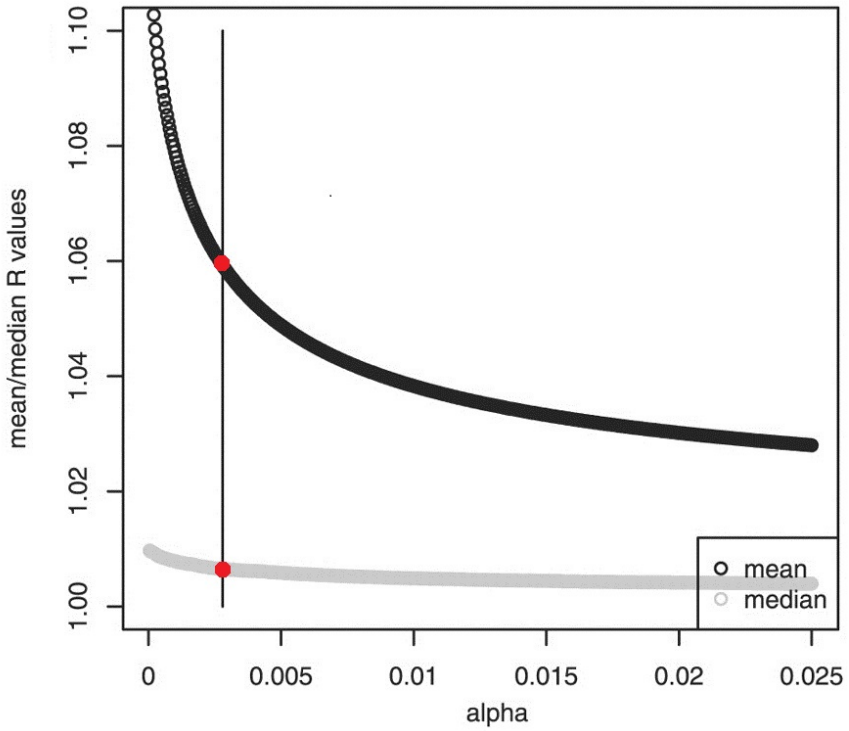


Figure: Scree Plot of the Dataset 2

CLARKE Method

Principal Components Analysis

Clarke et al. (2010) plot the scree plot of CLARKE model,

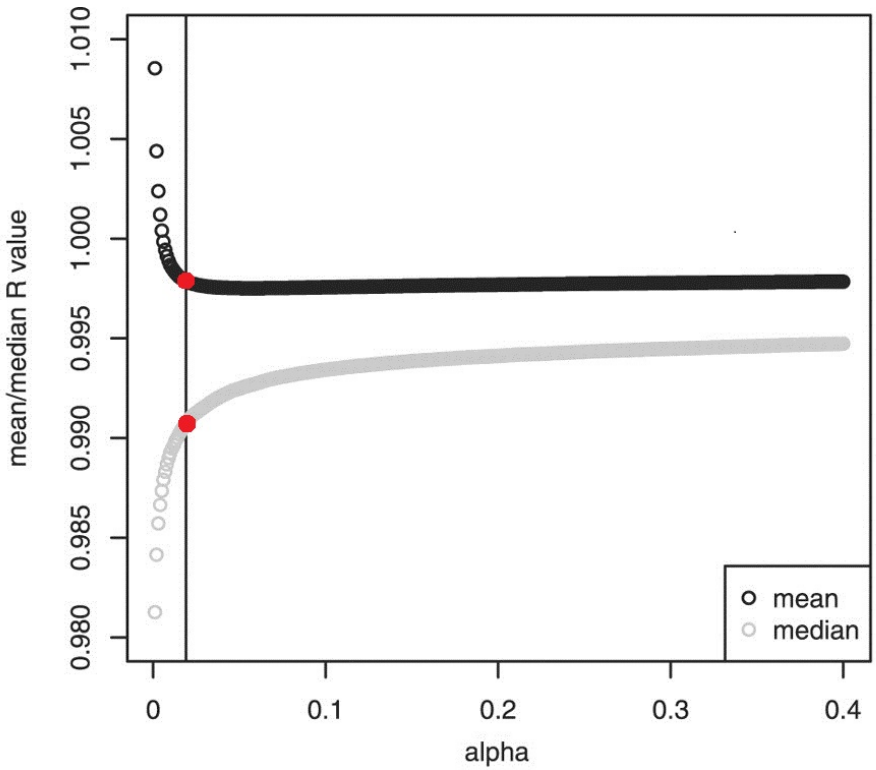


Figure: Scree Plot of the Dataset 1

The point in red gives the minimum ratio, or say, the most accurate estimated proportion; and according to the knowledge of principal components analysis, this point is located at the 'knee' or 'elbow' of the curve.

CLARKE Method

Algorithm

Original Model

Parametrise the curve ⇒ Calculate the first derivative ⇒ Calculate the arc-length ⇒ **Reparametrise the curve into unit-speed curve** ⇒ Calculate the second derivative ⇒ Calculate the curvature ⇒ The radius of curvature is the reciprocal of obtained curvature

Improved Model

Parametrise the curve ⇒ Calculate the first derivative ⇒ Calculate the second derivative ⇒ **Calculate the cross product of first and second derivatives** ⇒ Calculate the curvature ⇒ The radius of curvature is the reciprocal of obtained curvature

Results and Discussions

Simulated Data

Simple Extreme Data

Simulated Data - Simple Extreme Data

Gene Sites	**Input Sample**	**Pure m^6A Sample**
Gene site 1	1000	0
Gene site 2	1000	0
...	...	...
...	...	...
Gene site 250	1000	0
Gene site 251	0	1000
Gene site 252	0	1000
...	...	...
...	...	...
Gene site 499	0	1000
Gene site 500	0	1000

Table: Generated Input Sample and Pure m^6A Sample

Simulated Data - Simple Extreme Data

Generation of IP Sample

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

Simulated Data - Simple Extreme Data

Results of Simple Extreme Data

Purity / Methods	Purity 10%			Purity 20%		
	Set1	Set2	Set3	Set1	Set2	Set3
ISOpureR	0.000	0.000	0.000	0.000	0.000	0.000
CLARKE	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
	Purity 30%			Purity 40%		
ISOpureR	0.000	0.000	0.000	0.000	0.000	0.000
CLARKE	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
	Purity 50%			Purity 60%		
ISOpureR	0.002	0.003	0.000	0.286	0.286	0.286
CLARKE	0.127	0.127	0.127	0.246	0.244	0.244

Table: Results of Simple Extreme Datasets

Simulated Data - Simple Extreme Data

Results of Simple Extreme Data

Purity / Methods	Purity 60%			Purity 70%		
	Set1	Set2	Set3	Set1	Set2	Set3
ISOpureR	0.286	0.286	0.286	0.575	0.570	0.571
CLARKE	0.246	0.244	0.244	0.175	0.180	0.178
	Purity 80%			Purity 90%		
ISOpureR	1.000	1.000	1.000	1.000	1.000	1.000
CLARKE	0.121	0.118	0.119	0.056	0.058	0.056

Table: Results of Simple Extreme Datasets (Continued)

Simulated Data - Simple Extreme Data

Gene Sites	**Input Sample**	**Pure m^6A Sample**
Gene site 1	1000	0
Gene site 2	1000	0
...	...	...
...	...	...
Gene site 250	1000	0
Gene site 251	0	1000
Gene site 252	0	1000
...	...	...
...	...	...
Gene site 499	0	1000
Gene site 500	0	1000

Table: Generated Input Sample and Pure m^6A Sample

Simulated Data

Simple Real Data - Generation

Simulated Data - Simple Real Data

Generate simulated data using experiment 'human-A549-C'

- Generate the pure m^6A sample by **permutating** the input sample;
- Randomly **choose the same 500 gene sites** from input sample and pure m^6A sample to form our simulated input sample and simulated pure m^6A sample;
- Apply **binomial distribution** to generate the IP samples with different mixing proportions (10%, 20%, $\cdots$, 90%) with respect to the **input profiles**; and for each proportion, we generate **30** different samples.

Simulated Data

Simple Real Data - General Results

Simulated Data - Simple Real Data

General Impression

Figure: Box Plot of ISOpureR Method

Simulated Data - Simple Real Data

General Impression

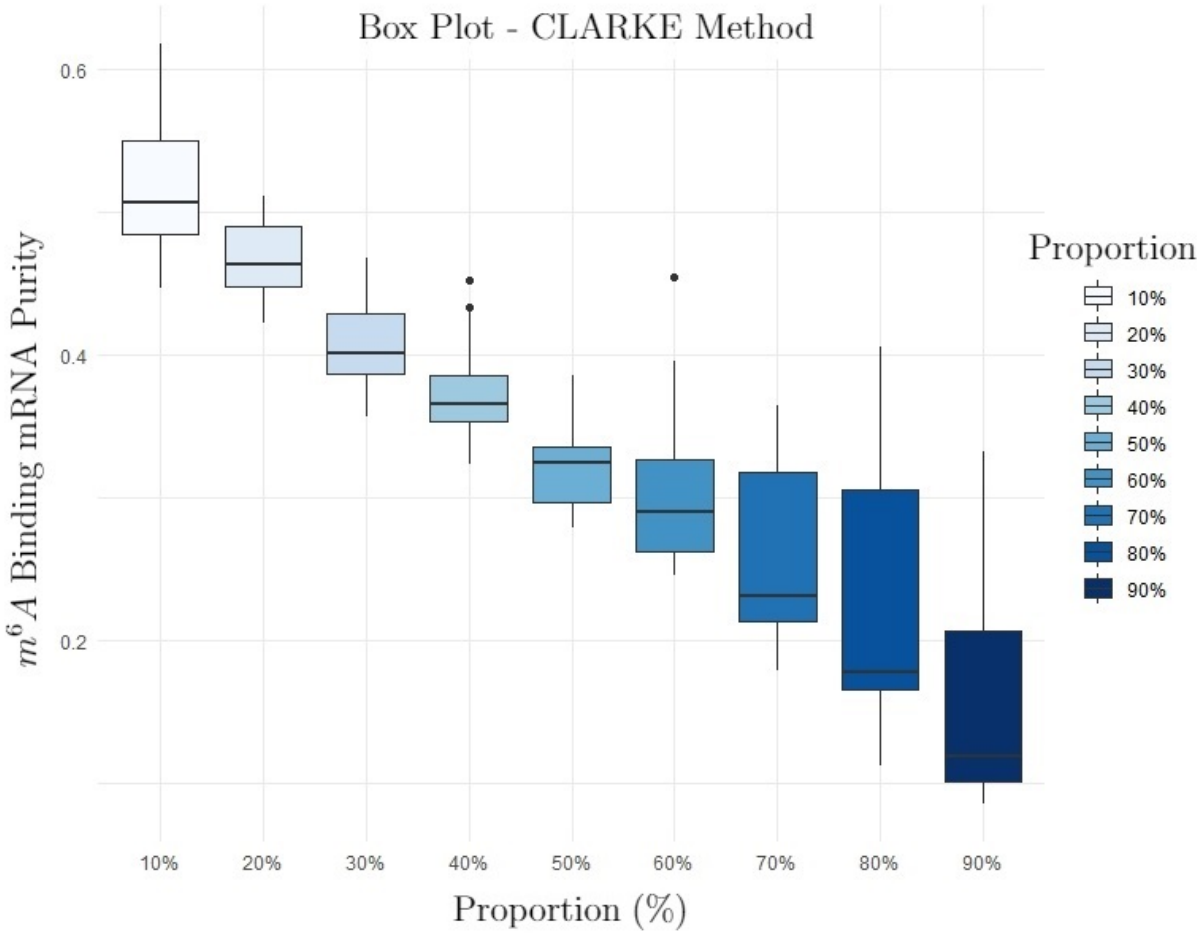


Figure: Box Plot of CLARKE Method

Simulated Data

Simple Real Data - Data Analysis of ISOpureR Method

Simulated Data - Simple Real Data

Linear Regression

Assume the results of ISOpureR method satisfies the linear regression model, which is

$$y = \beta_0 + \beta_1 x + \epsilon$$

with assumption (Prabhakaran, 2016),

- The y-values (or the errors) are independent.
- The y-values can be expressed as a linear function of the x variable.
- Variation of observations around the regression line (the residual SE) is constant (homoscedasticity).
- **The residuals of y value (or the error) are normally distributed.**

Simulated Data - Simple Real Data

Linear Regression

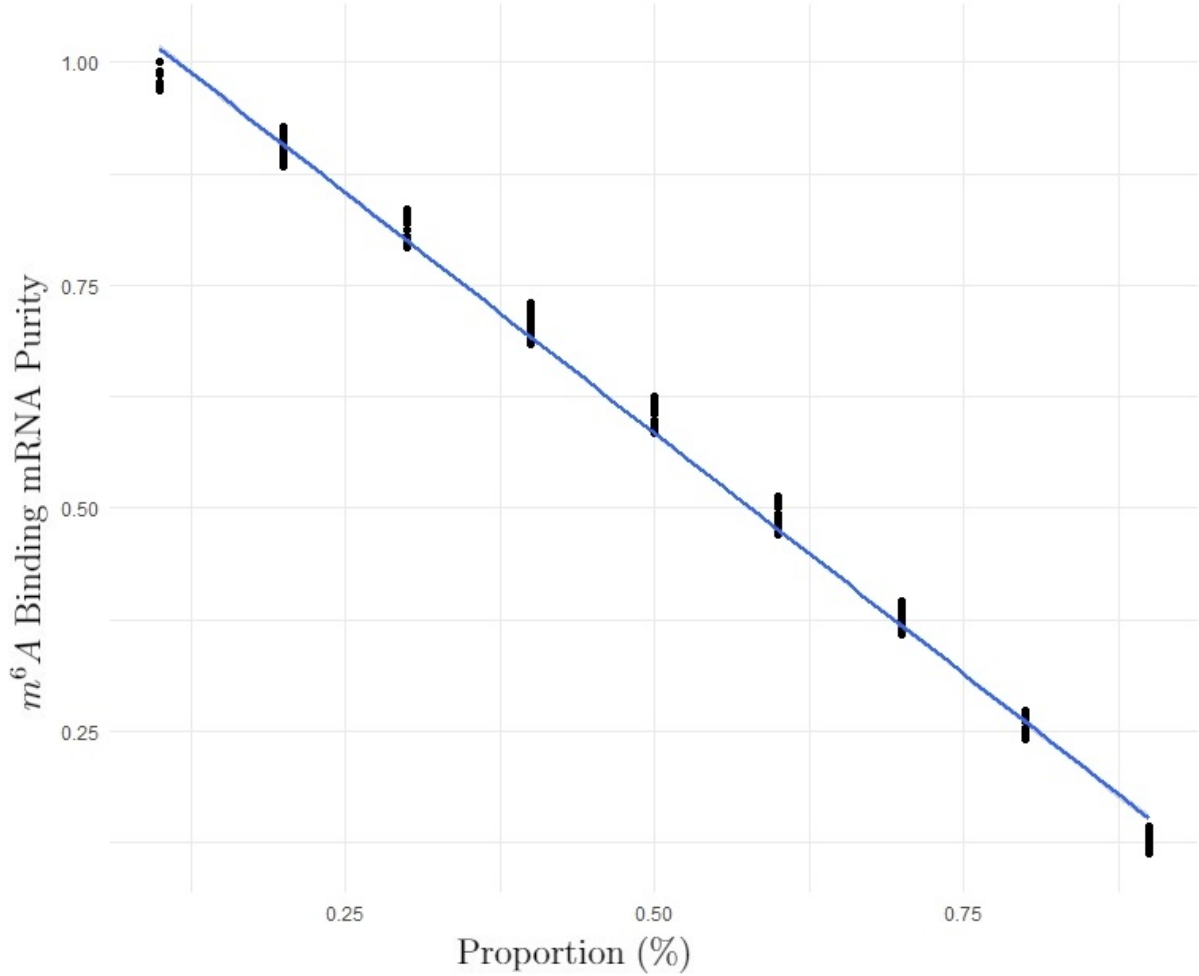


Figure: Linear Regression of ISOpureR Method

Simulated Data - Simple Real Data

Test of Significance

- H_0 : *the coefficients of the linear regression model is equal to 0.*
- H_1 : *the coefficients of the linear regression model is not equal to 0.*

Results: *p-value* $< 2 \times 10^{-16}$, $R^2 = 0.9952$

Lack-of-Fit Test

- H_0: *linear model adequately fits data*.
- H_1: *linear model does not adequately fit data*.

Results: *p-value* $< 2.2 \times 10^{-16}$

Simulated Data - Simple Real Data

Quadratic Polynomial Regression

Assume the results of ISOpureR method satisfies the quadratic polynomial regression model, which is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

with assumption (Prabhakaran, 2016),

- The y-values (or the errors) are independent.
- The y-values can be expressed as a quadratic polynomial function of the x variable.
- Variation of observations around the regression line (the residual SE) is constant (homoscedasticity).
- **The residuals of y value (or the error) are normally distributed.**

Simulated Data - Simple Real Data

Quadratic Polynomial Regression

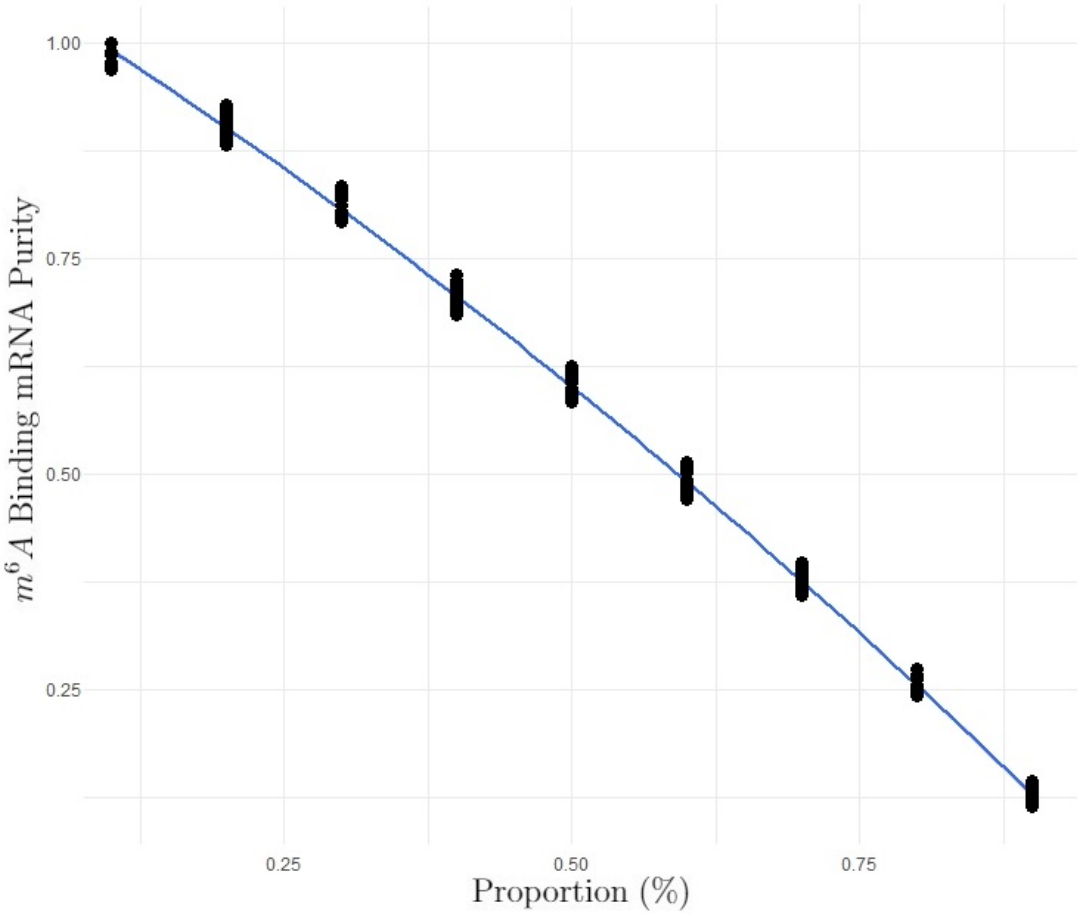


Figure: Quadratic Polynomial Regression of ISOpureR Method

Simulated Data - Simple Real Data

Test of Significance

- H_0 : *the coefficients of the quadratic polynomial regression model is equal to 0.*
- H_1 : *the coefficients of the quadratic polynomial regression model is not equal to 0.*

Results: *p-value* $< 2 \times 10^{-16}$, $R^2 = 0.9981$

Lack-of-Fit Test

- H_0: *quadratic polynomial model adequately fits data.*
- H_1: *quadratic polynomial model does not adequately fit data.*

Results: *p-value* $= 0.9577$

Simulated Data - Simple Real Data

Diagnostic Test

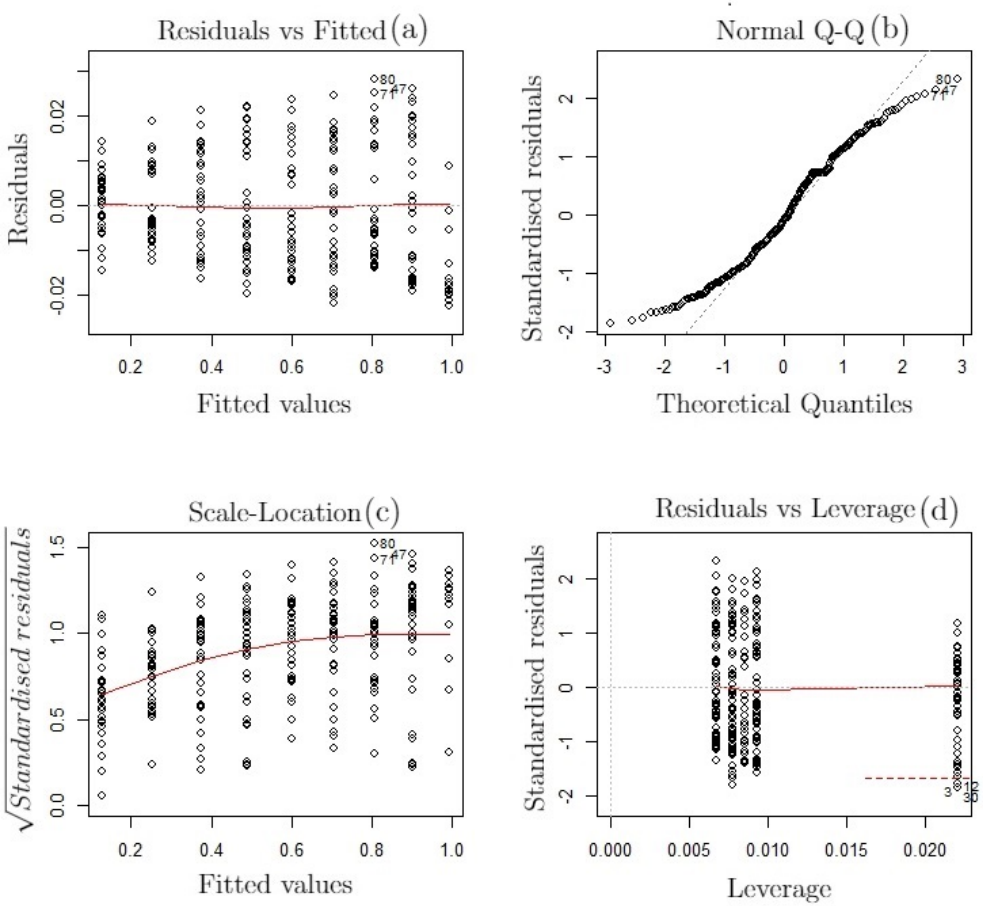


Figure: Diagnostic Test for Quadratic Polynomial Regression of ISOpureR Method

Simulated Data - Simple Real Data

Shapiro-Wild Test

- H_0: *the sample came from a normally distributed population*.
- H_1: *the sample did not come from a normally distributed population*.

Results: *p-value* $= 1.365 \times 10^{-5}$

Simulated Data - Simple Real Data

Second Order Median Regression

Assume the results of ISOpureR method satisfies the median regression model, which is

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

where α_0, α_1 and α_2 are the coefficients of the median regression model; F represents the common distribution function of the errors **without any distributional assumptions**.

Lack-of-Fit Test

- H_0: *second order median regression model adequately fits data*.
- H_1: *second order median regression model does not adequately fit data*.

Results: *p-value* $= 0.21$

Simulated Data - Simple Real Data

Second Order Median Regression

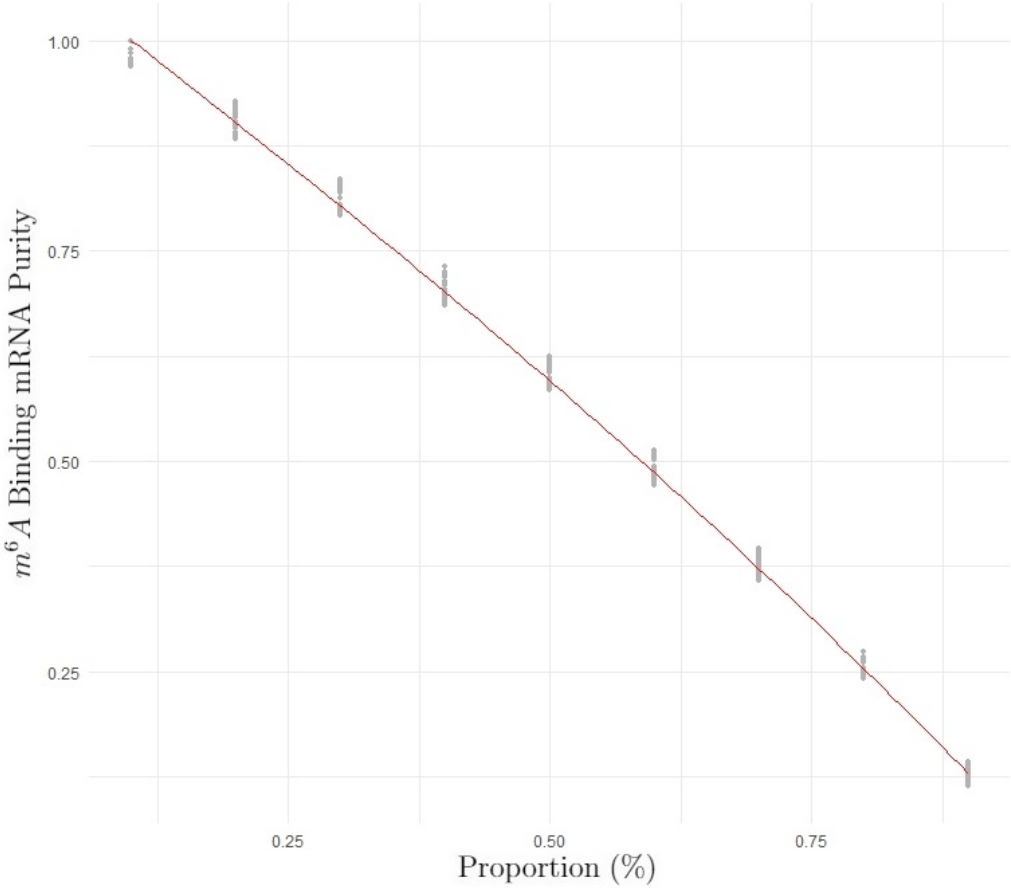


Figure: Second Order Median Regression of ISOpureR Method

Simulated Data

Simple Real Data - Data Analysis of CLARKE Method

Simulated Data - Simple Real Data

Linear Regression

Assume the results of CLARKE method satisfies the linear regression model, which is

$$y = \beta_0 + \beta_1 x + \epsilon$$

with assumption (Prabhakaran, 2016),

- The y-values (or the errors) are independent.
- The y-values can be expressed as a linear function of the x variable.
- Variation of observations around the regression line (the residual SE) is constant (homoscedasticity).
- **The residuals of y value (or the error) are normally distributed.**

Simulated Data - Simple Real Data

Linear Regression

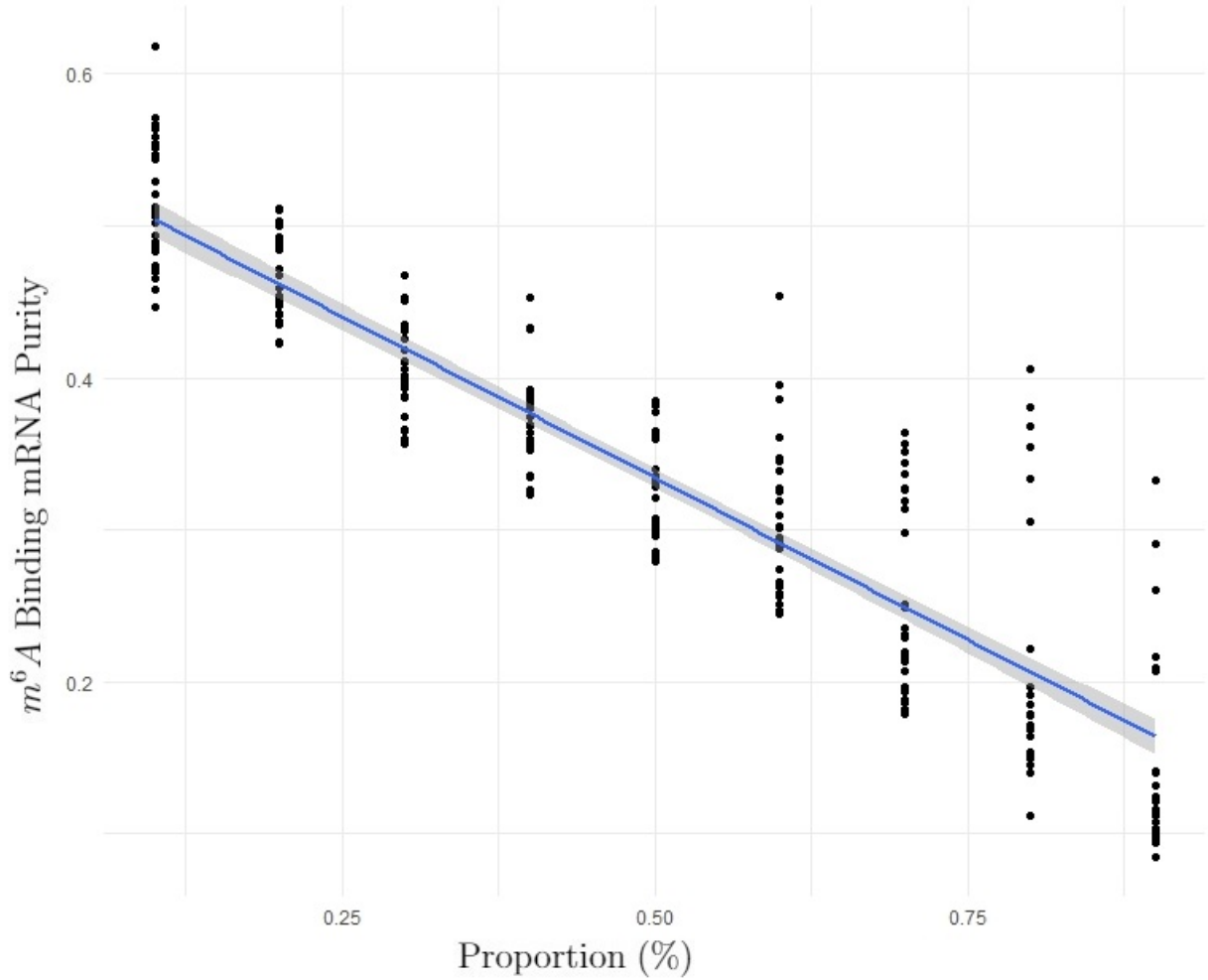


Figure: Regression of CLARKE Methods

Simulated Data - Simple Real Data

Test of Significance

- H_0 : *the coefficients of the linear regression model is equal to 0.*
- H_1 : *the coefficients of the linear regression model is not equal to 0.*

Results: *p-value* $< 2 \times 10^{-16}$, $R^2 = 0.8154$

Lack-of-Fit Test

- H_0: *linear model adequately fits data.*
- H_1: *linear model does not adequately fit data.*

Results: *p-value* $= 0.1153$

Simulated Data - Simple Real Data

Diagnostic Test

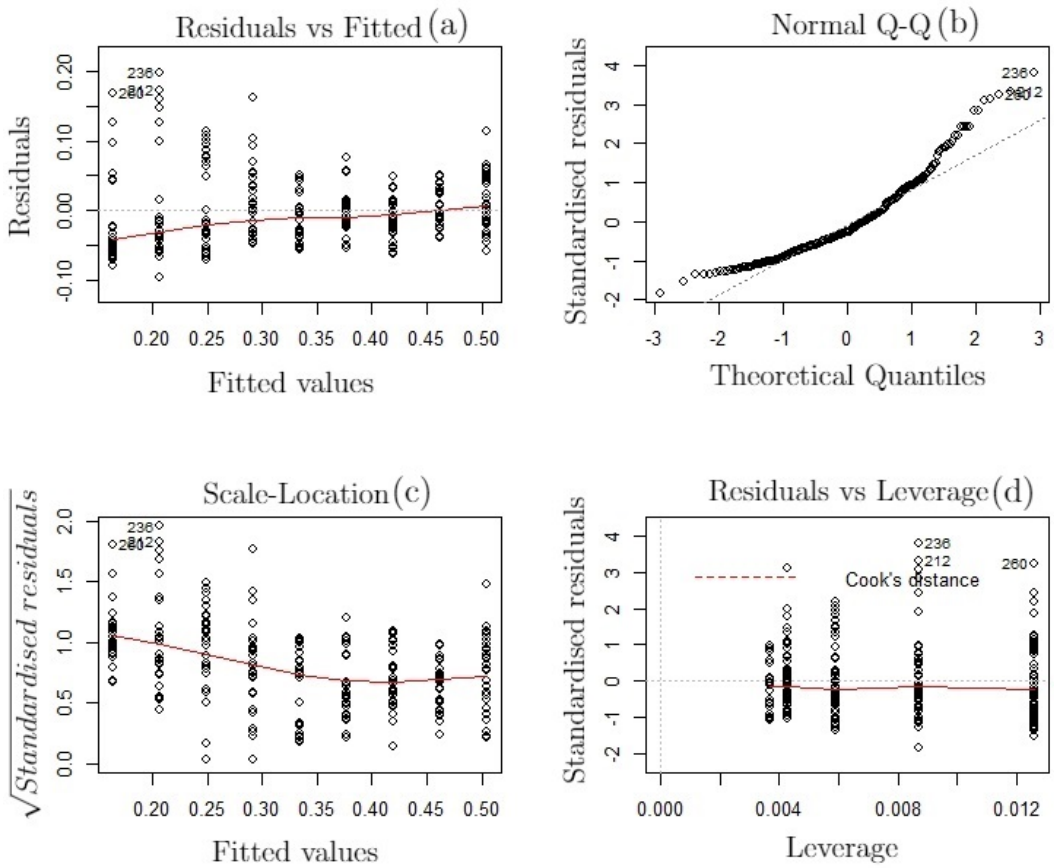


Figure: Diagnostic Test of CLARKE Regression Model

Simulated Data - Simple Real Data

Shapiro-Wild Test

- H_0: *the sample came from a normally distributed population*.
- H_1: *the sample did not come from a normally distributed population*.

Results: *p-value* $= 1.136 \times 10^{-11}$

Simulated Data - Simple Real Data

Median Regression

Assume the results of CLARKE method satisfies the median regression model, which is

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

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**.

Lack-of-Fit Test

- H_0: *median regression model adequately fits data*.
- H_1: *median regression model does not adequately fit data*.

Results: *p-value* $= 0.09$

Simulated Data - Simple Real Data

Median Regression

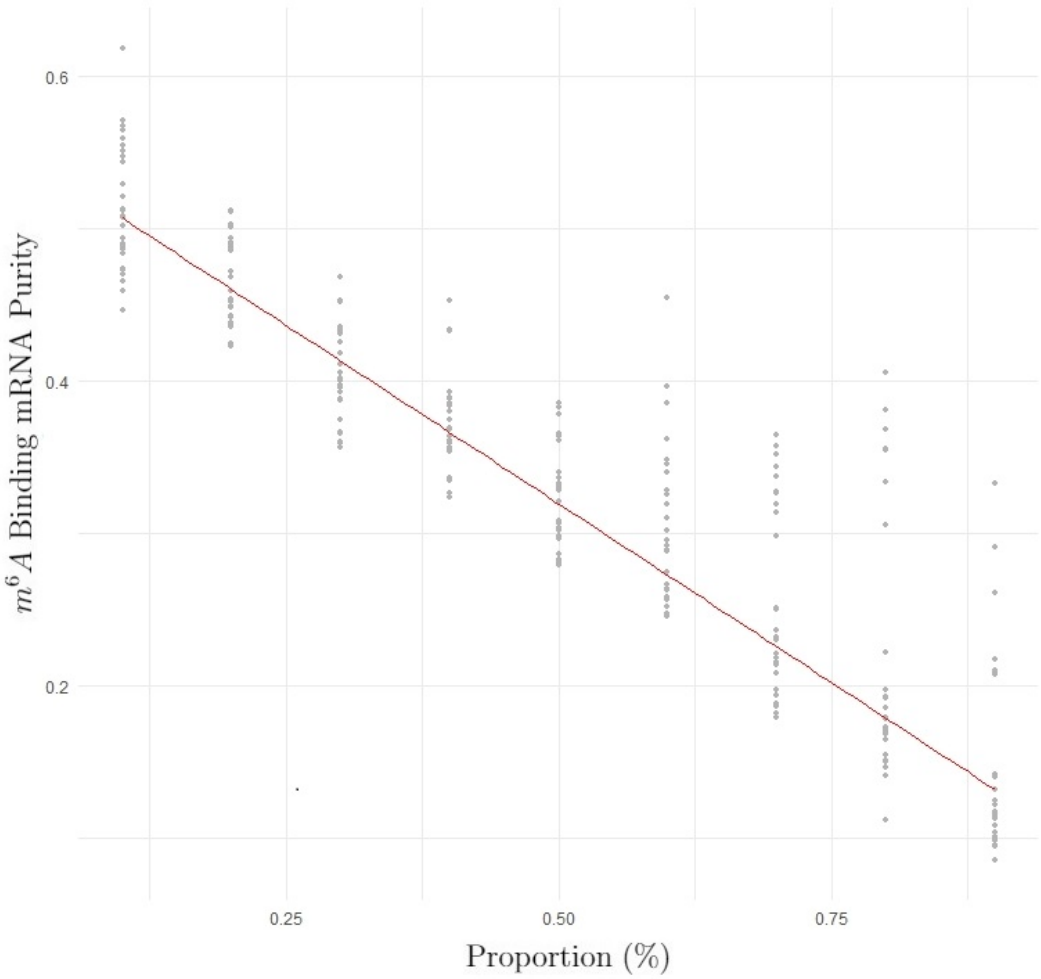


Figure: Median Regression of CLARKE Method

Real Data

Real Data

Example of real data

Gene Sites	SRR1182615	···	SRR1182630
Site 1	32	···	0
Site 2	98	···	0
···	···	···	···
···	···	···	···
Site 33448	97	···	17
Site 33449	106	···	19
Site 33450	90	···	18
···	···	···	···
···	···	···	···
Site 69445	15	···	9
Site 69446	3	···	8

Table: Real Data - 'human-A549-METTL3-'

Real Data

Results of real data

Method	human-A549-C			
	SRR1182619	SRR1182621	SRR1182623	
ISOpureR	1.000	1.000	1.000	
CLARKE	0.397	0.452	0.450	
	human-A549-METTL3-			
	SRR1182615	SRR1182617	SRR1182629	
ISOpureR	1.000	1.000	1.000	
CLARKE	0.464	0.540	0.375	
	human-A549-METTL14-			
	SRR1182607	SRR1182609	SRR1182611	SRR1182613
ISOpureR	1.000	1.000	1.000	1.000
CLARKE	0.445	0.461	0.430	0.386

Table: Results of Real Datasets

Real Data

Results of real data

Method	human-A549-METTL14-			
	SRR1182607	SRR1182609	SRR1182611	SRR1182613
ISOpureR	1.000	1.000	1.000	1.000
CLARKE	0.445	0.461	0.430	0.386
	human-H1ESC-T48			
	SRR1035218	SRR1035220		
ISOpureR	1.000	1.000		
CLARKE	0.686	0.698		
	human-hESC-C			
	SRR1035222	SRR1035224		
ISOpureR	1.000	1.000		
CLARKE	0.665	0.619		

Table: Results of Real Datasets (Continued)

Conclusions

Conclusions and Contributions

Conclusions

ISOpureR

- Small number of gene sites $\Rightarrow$ More accurate;
- Large number of gene sites $\Rightarrow$ Converge to 1;
- Simple extreme dataset $\Rightarrow$ Inaccurate.

CLARKE

- Small number of gene sites $\Rightarrow$ Underestimated;
- Large number of gene sites $\Rightarrow$ More accurate;
- Simple extreme dataset $\Rightarrow$ Chaos.

Original Work

- Estimation by comparison;
- Approximation of first derivative;
- Calculation of curvature;
- Results analysis.

Reference

References I

Bertrand Clarke, Jennifer Clarke, and Pearl Seo. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, 26(8): 1043–1049, 03 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq097. URL https://doi.org/10.1093/bioinformatics/btq097.

Qi Cui, Hailing Shi, Peng Ye, Li Li, Qiuhao Qu, Guoqiang Sun, Guihua Sun, Zhike Lu, Yue Huang, Cai-Guang Yang, Arthur D. Riggs, Chuan He, and Yanhong Shi. m 6 a rna methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. *Cell Reports*, 18: 2622–2634, 03 2017. doi: 10.1016/j.celrep.2017.02.059.

Dan Dominissini, Sharon Moshitch-Moshkovitz, Mali Salmon-Divon, Ninette Amariglio, and Gideon Rechavi. Transcriptome-wide mapping of n6-methyladenosine by m 6a-seq based on immunocapturing and massively parallel sequencing. *Nature protocols*, 8:176–89, 01 2013. doi: 10.1038/nprot.2012.148.

References II

Mark Gosink, Howard Petrie, and Nicholas Tsinoremas. Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics (Oxford, England)*, 23:3328–34, 01 2008. doi: 10.1093/bioinformatics/btm508.

Mark M. Gosink, Howard T. Petrie, and Nicholas F. Tsinoremas. Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics*, 23(24):3328–3334, 10 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm508. URL `https://doi.org/10.1093/bioinformatics/btm508`.

Selva Prabhakaran. *Assumptions of Linear Regression*. r-statistics.co, 2016. doi: http://r-statistics.co/Assumptions-of-Linear-Regression.html.

References III

Gerald Quon, Syed Haider, Amit Deshwar, Ang Cui, Paul C Boutros, and Quaid Morris. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome medicine*, 5:29, 03 2013. doi: 10.1186/gm433.

Ian Roundtree, Molly Evans, Tao Pan, and Chuan He. Dynamic rna modifications in gene expression regulation. *Cell*, 169:1187–1200, 06 2017. doi: 10.1016/j.cell.2017.05.045.

Q & A