

RESEARCH ARTICLE

magpie: A power evaluation method for differential RNA methylation analysis in N6-methyladenosine sequencingZhenxing Guo¹, Daoyu Duan², Wen Tang², Julia Zhu³, William S. Bush², Liangliang Zhang², Xiaofeng Zhu², Fulai Jin⁴, Hao Feng^{2*}

1 School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, Guangdong, China, 2 Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America, 3 Hathaway Brown School, Shaker Heights, Ohio, United States of America, 4 Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America

* These authors contributed equally to this work.

* hxf155@case.edu

OPEN ACCESS

Citation: Guo Z, Duan D, Tang W, Zhu J, Bush WS, Zhang L, et al. (2024) *magpie*: A power evaluation method for differential RNA methylation analysis in N6-methyladenosine sequencing. PLoS Comput Biol 20(2): e1011875. <https://doi.org/10.1371/journal.pcbi.1011875>

Editor: Mark Ziemann, Burnet Institute, AUSTRALIA

Received: September 27, 2023

Accepted: January 30, 2024

Published: February 12, 2024

Copyright: © 2024 Guo et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The proposed power assessment tool *magpie* is publicly available as a R/Bioconductor package at: <https://bioconductor.org/packages/magpie/>. The simulation tutorial with instructions is available at <https://github.com/dxd429/magieSims>.

Funding: This research was partially supported by the intramural funding from Case Western Reserve University awarded to HF, and the intramural funding UDF01002936 awarded to ZG from The Chinese University of Hong Kong, Shenzhen

Abstract

Recently, novel biotechnologies to quantify RNA modifications became an increasingly popular choice for researchers who study epitranscriptome. When studying RNA methylations such as N6-methyladenosine (m^6A), researchers need to make several decisions in its experimental design, especially the sample size and a proper statistical power. Due to the complexity and high-throughput nature of m^6A sequencing measurements, methods for power calculation and study design are still currently unavailable. In this work, we propose a statistical power assessment tool, *magpie*, for power calculation and experimental design for epitranscriptome studies using m^6A sequencing data. Our simulation-based power assessment tool will borrow information from real pilot data, and inspect various influential factors including sample size, sequencing depth, effect size, and basal expression ranges. We integrate two modules in *magpie*: (i) a flexible and realistic simulator module to synthesize m^6A sequencing data based on real data; and (ii) a power assessment module to examine a set of comprehensive evaluation metrics.

Author summary

Sample size and sequencing depth are two essential quantitative factors determined prior to high throughput sequencing experiments, for statistical power maximization with limited budget. Due to the complex structure of data from m^6A RNA methylation sequencing, analytical derivations for both quantities remain challenging in experimental designs. In response to this challenge, we propose a simulation-based statistical framework, together with a user-friendly R/Bioconductor package *magpie*, to comprehensively assess the power of the differential m^6A methylation detection at varied sample sizes, effect sizes, baseline expression levels, and sequencing depths. Using in-silico synthetic data that

(CUHK-Shenzhen). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

mimic real data well, *magpie* provides several major evaluation metrics to assist users in study design and statistical power evaluation.

Introduction

RNA methylation represents another layer of epigenetic regulation in addition to the well-studied DNA methylation and histone modification. Among different types of RNA methylation, N6-methyladenosine, i.e. m⁶A, is the most common form. It has been identified as one of the post-transcriptional regulatory markers on mRNA, rRNA, tRNA, circRNA, miRNA and long-noncoding RNA, and plays important roles in regulating pre-RNA splicing, RNA translation, stability, and degradation [1–3]. The effects of m⁶A suggest its involvement in multiple cellular processes such as cell differentiation and reprogramming [4, 5]. Studies also suggest linkages between the dysregulation of m⁶A and many human diseases such as cancers and neural disorders [2, 6, 7].

MeRIP-seq/m⁶A-seq was developed to characterize transcriptome-wide m⁶A profiles [8, 9]. This technique typically relies on immunoprecipitation of m⁶A-containing RNA fragments (m⁶A-IP), followed by high-throughput next generation sequencing. These samples are generally referred to as the IP (immunoprecipitated) samples. In addition to IP samples, cDNA libraries are also prepared for input control mRNAs to measure the background mRNA abundance. These input controls are essentially the transcriptome from regular RNA-seq. The m⁶A methylation level, for each region, is then quantified by the enrichment of IP over input, roughly the normalized ratio between IP and input control counts. If the m⁶A enrichment is significantly high, then the called peak of that region suggests an underlying m⁶A residue. MeRIP-seq is becoming a popular and indispensable tool to profile transcriptome-wide m⁶A, since the invention of this technique. One feature of MeRIP-seq is that, it immunoprecipitates each IP sample independently, which can potentially induce technical variabilities. Such technical artifacts lead to erroneous peak calling of methylated regions. This problem becomes prominent in studies with small sample sizes [10], which is often the case given the high expenses associated with the current experimental protocols. As an improved alternative, in m⁶A-seq2 [11], a single IP experiment is performed on the pooled RNAs of all samples, where RNAs from different samples are uniquely barcoded and demultiplexed after sequencing. The multiplexed profiling procedure by m⁶A-seq2 is expected to be widely applied to interrogate the distribution and functional consequences of m⁶A.

To study the biological implications of m⁶A, one fundamental task is to identify the Differentially Methylated Regions (DMRs) across different conditions. Although several DMR detection methods have been developed [12–14] and evaluated [15] in either MeRIP-seq or m⁶A-seq2 experiments, the sample size calculation with their associated statistical power remains an open question due to the complexities of sequencing experiments. Further, due to the uniqueness in data structure, power analysis tools developed for other types of analyses such as Differential Expression (DE) gene detection from RNA-seq cannot be applied to MeRIP-seq and m⁶A-seq2 experiments. First, data simulated for power assessment in DE gene detection from RNA-seq are barely equivalent to the input control data alone. No statistical model is available to generate their matched IP counts. Second, the effect size of methylation in m⁶A data analysis is based on the ratios of IP/input, not the input data alone. Therefore, the count coverage of each gene may affect power and other metrics in a way differing from that in DE analyses. Additionally, the impact of baseline expression of each gene and sequencing depth of the whole sample on the power of DMR detection are also unignorable. Therefore, an

appropriate power analysis tool specifically for epitranscriptome studies is needed, especially with its increasing popularity. To our best knowledge, no method is currently available.

Here, we propose a comprehensive power evaluation method named *magpie* (**m**⁶**A** genome-wide differential analysis **p**ower **i**nference). *magpie* first learns characteristics of real data, and then synthesizes data that mimics the real data well. In simulations, *magpie* allows for the adjustment of sample size, sequencing depth and effect size. It can evaluate the epitranscriptome study design using multiple metrics including sensitivity, specificity, precision, false discovery rate, and more. Building upon these functionalities, *magpie* fills in the knowledge gap by providing a comprehensive biostatistical tool for statistical power evaluation, sample size calculation, and data analysis planning, which are almost always required in general experimental designs. This makes it the first available tool to guide the practical experimental design by comprehensively investigating the relationship between statistical metrics and associated factors in m⁶A differential analysis. *magpie* is publicly available as an R/Bioconductor package at <https://bioconductor.org/packages/magpie/>.

Materials and methods

An overview of *magpie*

We assess the effect of experimental design on the power of DMR detection purely based on simulations, where the whole procedure is divided into two components. First, *magpie* preprocesses .bam files from MeRIP-seq sequencing and obtains read counts in candidate regions from all samples (Fig 1), where candidates are identified with conditional binomial tests. With the counts from the identified candidate regions, *magpie* simulates count matrices for both IP and Input samples with a Gamma-Poisson model. Parameters involved are estimated from the candidates to mimic the actual MeRIP-seq data in aspects of marginal distribution read counts, and the distribution of biological dispersion in methylation levels (Fig 1). With data simulated, we then evaluate power and error rates on them (Fig 1). The two components, Gamma-Poisson simulation and power assessment, are independent so that *magpie* allows the assessment on data by different simulation strategies.

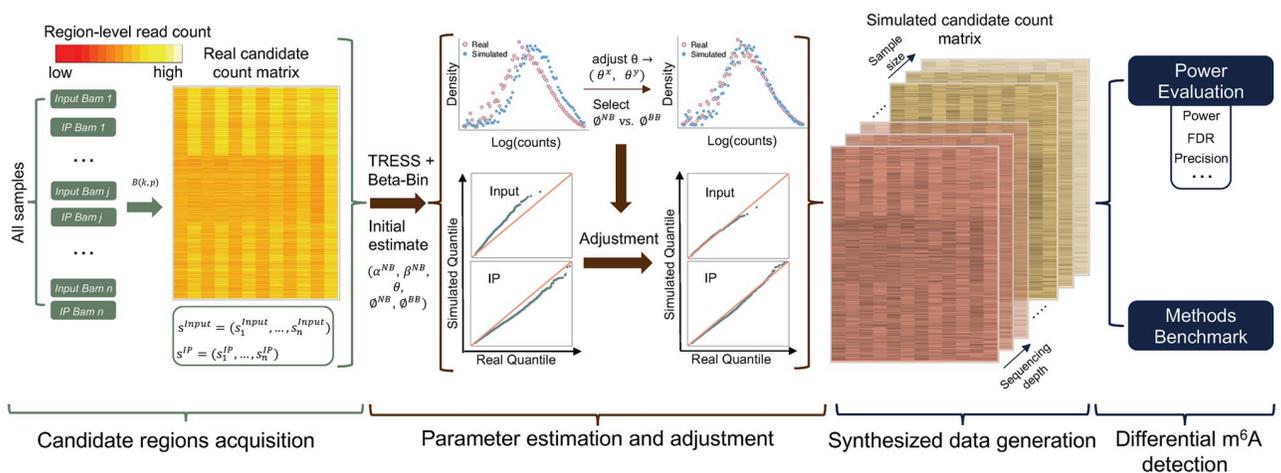


Fig 1. Overview of *magpie*. *magpie* provides power evaluation for differential m⁶A methylation analysis. It takes pilot MeRIP-seq data as the input. Based on the pilot data, it obtains candidate regions, estimates key parameters, and conducts real-data-based simulations for statistical power evaluation.

<https://doi.org/10.1371/journal.pcbi.1011875.g001>

Data generative model

Here we describe how *magpie* simulates the MeRIP-seq count matrices given existing real MeRIP-seq data from different conditions. *magpie* processes *.bam* files by splitting the transcriptome into bins, aggregating read counts, and testing for significance of IP enrichment over Input. Using a bump-finding algorithm, significant bins are combined into candidate regions. *magpie* then focus on these candidates in simulations, as other regions lack IP enrichment and biological relevance. Suppose there are in total N pairs of IP and Input samples from all conditions, and M candidate DMRs generated after preprocessing. Let X_{ij} and Y_{ij} denote input and IP counts in candidate DMR i from sample j . We assume that $X_{ij} \sim \text{Poisson}(s_j^x \lambda_{ij}^x)$ and $\lambda_{ij}^x \sim \text{Gamma}(\alpha_{ij}^x, \theta_i)$. Similarly, $Y_{ij} \sim \text{Poisson}(s_j^y \lambda_{ij}^y)$ and $\lambda_{ij}^y \sim \text{Gamma}(\alpha_{ij}^y, \theta_i)$. Here s_j^x and s_j^y represent the normalizing factors for input and IP samples, such as the library sizes. λ_{ij}^x and λ_{ij}^y are normalized poisson rates. $\alpha_{ij}^x, \alpha_{ij}^y$, and θ_i are the shape and scale parameters of corresponding gamma distributions. Given above assumptions, naturally $\frac{\lambda_{ij}^y}{\lambda_{ij}^x + \lambda_{ij}^y} \sim \text{Beta}(\alpha_{ij}^y, \alpha_{ij}^x)$. Further, denote $\mu_{ij} = \frac{\alpha_{ij}^y}{\alpha_{ij}^x + \alpha_{ij}^y}, \phi_{ij} = \frac{1}{\alpha_{ij}^x + \alpha_{ij}^y + 1}$, then marginally,

$$\begin{aligned} X_{ij} | \mu_{ij}, \phi_{ij} &\sim \text{NB}((1 - \mu_{ij})(\phi_{ij}^{-1} - 1), \frac{s_j^x \theta_i}{1 + s_j^x \theta_i}), \\ Y_{ij} | \mu_{ij}, \phi_{ij} &\sim \text{NB}(\mu_{ij}(\phi_{ij}^{-1} - 1), \frac{s_j^y \theta_i}{1 + s_j^y \theta_i}) \end{aligned} \tag{1}$$

In above equations, μ_{ij} and ϕ_{ij} represent the mean and dispersion of the methylation level for candidate region i in sample j .

We begin by simulating size factors, for which we directly use the values estimated from real data:

$$\begin{aligned} s_j^x &= \sum_b x_{bj} / \text{median}\{\sum_b x_{b1}, \dots, \sum_b x_{bN}, \sum_b y_{b1}, \dots, \sum_b y_{bN}\}, \\ s_j^y &= \sum_b y_{bj} / \text{median}\{\sum_b x_{b1}, \dots, \sum_b x_{bN}, \sum_b y_{b1}, \dots, \sum_b y_{bN}\}, \end{aligned}$$

where x_{bj} and y_{bj} are read counts in bin b from the j th Input and IP samples.

Next, for each candidate region i , *magpie* simulates a baseline methylation level μ_i or equivalently e^{α_i} in the structure of $\log(\mu_{ij}) = \alpha_i + Z_j^T \beta_i$ where Z_j contains the attributes of sample j , and β_i represents corresponding coefficient. To do that, we randomly sample α_i from one parametric distribution, or from its empirical distribution estimated from real data. Distributions of $\hat{\alpha}_i$'s from five MeRIP-seq datasets are presented in Fig D in [S1 Appendix](#).

After simulating the baseline methylation, we simulate β_i 's for all regions. Because we can hardly know the actual number of DMRs and their degree of differential methylation, specific settings are adopted based on both reasonable assumptions and empirical observations. First, *magpie* sets the proportion of DMRs as 10%, assuming that DM is present in only a small subset of regions in most experiments. Then, for non-DMRs, $\beta_i = 0$. For DMR i , $\beta_i = \hat{\beta}_i$ if its estimated effect size $\hat{\beta}_i$ is greater than the 50% quantile of all regions. Otherwise, $\beta_i \sim U(1, 2)$. Here, $\hat{\beta}_i$'s are directly derived from real pilot data, using the DMR detection method TRESS.

The dispersion has been shown to be substantial across several real datasets, which justifies the necessity of its modeling (Fig F in [S1 Appendix](#)). We can simulate it again from a parametric distribution or sample from empirical distributions. To ensure the robustness, the empirical distribution can be estimated by TRESS from raw counts or by Beta-binomial regressions

from normalized counts. Denote \tilde{Y}_{ij} and $T_{ij} = \tilde{X}_{ij} + \tilde{Y}_{ij}$ as the normalized IP and total counts, the Beta-Binomial regressions are established as follows:

$$\begin{aligned}\tilde{Y}_{ij}|T_{ij} &\sim \text{Bin}(T_{ij}, p_{ij}) \\ p_{ij}|\mu_{ij}, \phi_i &\sim \text{Beta}(\mu_{ij}, \phi_i), \\ \text{logit}(\mu_{ij}) &= \tilde{a}_i + z_j^T \tilde{\beta}_i.\end{aligned}\quad (2)$$

where μ_{ij} and ϕ_i represent the mean and dispersion of methylation level. As noted, for the convenience of estimation, above Beta-binomial regressions (as well as TRESS) assume $\phi_{ij} = \phi_i$ for all j .

Empirically for the same real dataset, ϕ_i 's estimated by the Negative-binomial model in TRESS are usually greater than those estimated by Beta-binomial regressions. Without golden truth, our setting for ϕ_i relies on a data driven approach. Specifically, by comparing to the real data, *magpie* will calculate a KL-divergence for each of the synthetic counts by model in (1). Those $\hat{\phi}_i$ resulting in a significantly lower KL divergence between simulated data and real data will be kept for final data generation. If there is no significant difference in KL-divergence, $\hat{\phi}_i$ estimated by NB will be adopted.

Lastly, we simulate the scale parameter θ_i in (1). Again, it can be simulated by parametric distributions, or sampled directly from empirical distributions. For the parametric distribution, we set its mean as a function of ϕ_i observed in the previous peak detection method [14]. No matter the strategy employed, the first-round generated θ_i 's will be further scaled by the fold change between real and first-round simulated counts. Such an adjustment again helps to reduce the disparity between the simulated and real distributions, thereby improving the reliability of the results in follow-up power assessments.

DMR detection

After generating the simulated read counts in candidate DMRs, an existing software developed for MeRIP-seq is applied to detect DMRs. We implement an interface for calling TRESS and exomePeak2. Each method reports test statistics, p -values and FDRs for all candidate regions. These results are then used for the downstream power assessment. Users also have the option to adopt other DMR detection methods for their own evaluation, by following our simulation tutorial with detail instructions at <https://github.com/dxd429/magieSims>. This resource enables users to conduct their own analyses using the synthetic data generated by our simulation and evaluation framework.

Power assessment measures

We adopted several evaluation metrics in the statistical power assessment for differential analysis using MeRIP-seq data. These metrics include classic criteria in hypothesis testings such as the false discovery rate (FDR), power, and precision. We also inspected the false discovery cost (FDC, defined below) and targeted power [16], aiming to provide a comprehensive statistical power evaluation.

Because not all DMRs are of biological interest to us, especially those with low effect sizes, we introduce a cutoff Δ for the effect size β . Only those DMRs with $|\beta| \geq \Delta$ are considered as 'targeted DMRs', which are of biological interest in research. We denote the number of non-DMRs, non-targeted DMRs, and targeted DMRs as R_0 , R_1 , and R_2 , respectively. Suppose T_r represents the testing result of region r , where $T_r = 1$ denotes the discovered DMR, and $T_r = 0$ otherwise. The confusion matrix in the DMR detection is summarized in Table 1.

Table 1. The confusion matrix in m⁶A DMR detection, when taking biological significance into consideration.

Simulated True	Testing Result		Total
	$T_r = 0$	$T_r = 1$	
Non-DMR	N_0	P_0	R_0
Non-targeted DMR ($0 < \beta < \Delta$)	N_1	P_1	R_1
Targeted DMR ($ \beta \geq \Delta$)	N_2	P_2	R_2
Total	N	P	R

<https://doi.org/10.1371/journal.pcbi.1011875.t001>

The false discovery rate (FDR) and precision are statistical metrics that jointly provide insights into the balance between true and false discoveries among the significant features. In this context, FDR and precision are defined as $E\left[\frac{P_0}{P}\right]$ and $E\left[\frac{P_1+P_2}{P}\right]$, respectively. Statistical power is defined, naturally, as $E\left[\frac{P_1+P_2}{R_1+R_2}\right]$. To investigate the power of detecting targeted DMRs that are biologically interesting with $|\beta| \geq \Delta$, targeted power is introduced and defined as $E\left[\frac{P_2}{R_2}\right]$. To better illustrate the trade-off between false positives and true positives, we propose an additional metric, False Discovery Cost (FDC), $E\left[\frac{P_0}{P_2}\right]$, which is defined as the expected number of false positives per targeted true positive. The rationale behind this is straightforward: this cost is the expected number of false discoveries, per true discovery we are interested in.

Finally, our proposed evaluation framework allows for the examination of aforementioned metrics using simulations under various combinations of sample size, sequencing depth, input expression stratum, and FDR threshold. Each user-defined scenario is repeated for 100 times, and these metrics are computed and averaged to generate empirical estimations.

Implementation

Given a MeRIP-seq dataset in *.bam* files, various experimental scenarios (such as sample size, sequencing depth, FDR threshold, etc.), and a chosen differential methylation testing method, *maggie* generates evaluation results for each proposed study design. Functions incorporated in *maggie* allow users to export these results in an *.xlsx* file, and to visualize them through line plots. Users have the option to provide small pilot data, which could include only several chromosomes. We would estimate major parameters from these pilot data, to guide larger-scale simulations for power evaluation for future experimental designs. Alternatively, when pilot MeRIP-seq datasets are unavailable or unattainable, the function *quickPower* can produce power evaluation results within seconds. This is achieved by directly extracting our in-house evaluation results based on three public N6-methyladenosine datasets on GEO as the pilot data [17–19]. Our package also comes with a vignette that provides thorough instructions and examples of its applications in differential analysis experimental design on N6-methyladenosine.

Results

Larger sample size benefits DMR detection

Under simulation settings outlined in Simulation Settings in [S1 Appendix](#), we next examine the relationship between sample size and power in DMR detection, given that determining sample size is a primary objective in our method. We adopt sample sizes of 2, 3, 5, 7, and 10 per group, and nominal FDR values of 0.05, 0.1, 0.15, and 0.2, both of which are common choices in MeRIP-seq experiments. Note that we have validated our synthetic data against the

pilot data, ensuring that our strategy effectively captures the characteristics of real data (Fig G in S1 Appendix). The empirical results for major metrics are shown in Fig 2. Grouped by sample size and nominal FDR level, power, targeted power, FDC, and FDR averaged over 100 simulations are shown in Fig 2A–2D. For a fixed sample size, metrics like power, FDC, and FDR diminish under lower FDR thresholds. This occurs as lower FDR values lead to greater stringency, which in turn reduces false positives. The power will drop, as expected, when using stringent FDRs. As sample size increases, these differences become smaller, particularly for statistical power (Fig 2A). Here, power remains consistently high across all FDR levels with 7 and 10 replicates per group. This highlights the benefit of using a larger sample size that helps detect DMRs with limited effect sizes, where a type II error would often occur when the sample size is small. Such trend is observed consistently when using different pilot data (Fig A in S1 Appendix). At the same time, these results give researchers the knowledge to optimize the sample size based on their budgets. Using Fig 2A as an example, a power around 0.8 is achieved with 7 samples per group, and sample size of 7 is considered large in current MeRIP-seq studies. The benefit of expanding the sample size to 10 is marginal, but the associated costs could be significantly higher.

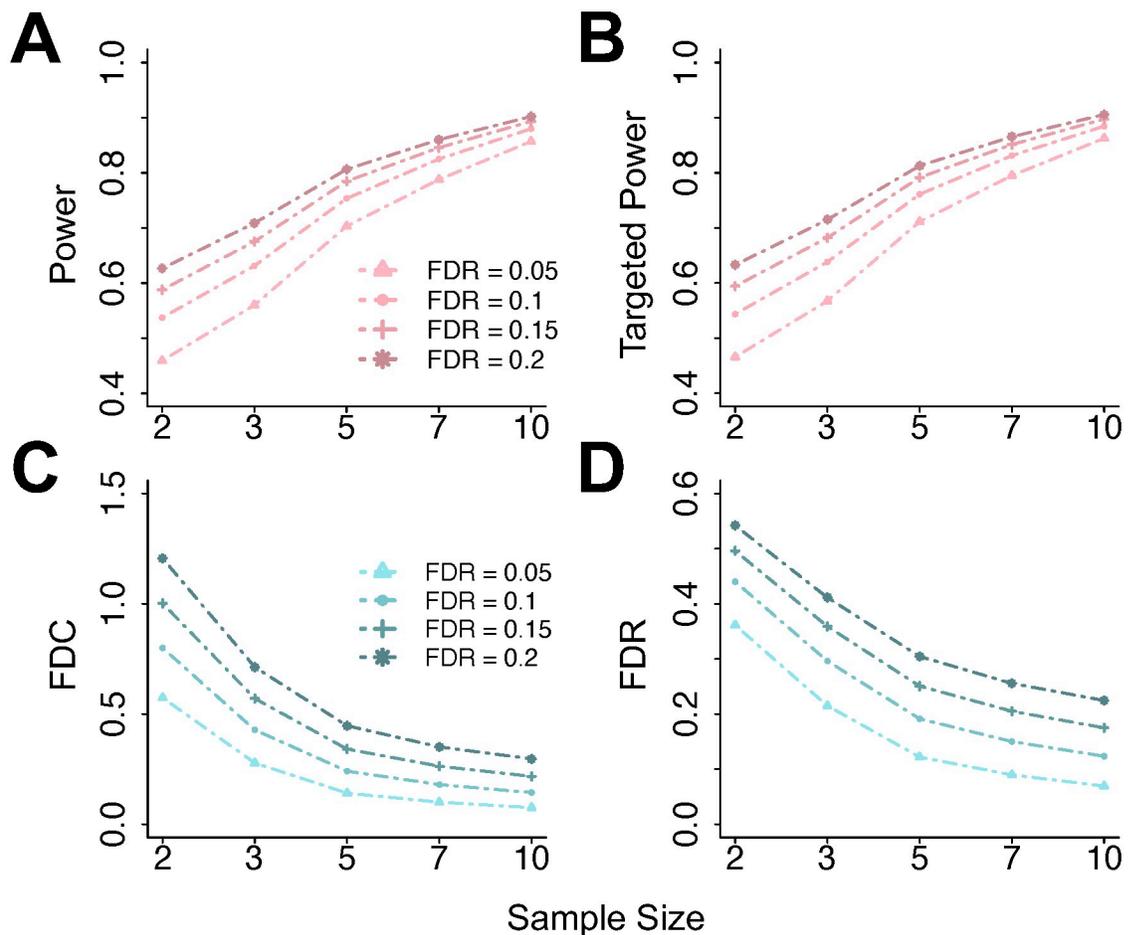


Fig 2. Statistical power evaluation metrics for DMR detection, at various sample sizes and FDR thresholds. (A) Power versus sample size, with each line presenting one FDR cutoff. (B)–(D) Similar to (A) but for other metrics: targeted power, FDC, and FDR. Targeted power and FDC are computed for DMRs with $|\beta| \geq 2$. Each point on the line plots is an averaged value over $N = 100$ simulations based on real MeRIP-seq data.

<https://doi.org/10.1371/journal.pcbi.1011875.g002>

Impact of baseline expression values

It is useful for researchers to understand the effects of heterogeneity in baseline expression levels in DMR detection. In MeRIP-seq data, the basal expression level is represented by input control read counts, thus we stratify power metrics by input control ranges. Six strata are obtained based on following quantiles of mean input counts: stratum 1 (0%-10%), stratum 2 (10%-30%), stratum 3 (30%-50%), stratum 4 (50%-70%), stratum 5 (70%-90%), and stratum 6 (90%-100%). At a nominal FDR of 0.05, the average targeted power and FDC for the six strata are shown in Fig 3A and 3B. Overall, reduced targeted power is observed in the lower strata, a trend that is more evident when sample sizes are small. This is expected, as true differences in low-expressed regions are often obscured by noise, making DMRs harder to detect. A limited sample size further exacerbates this issue. This suggests the potential benefits of increasing the sequencing depth, particularly when biological replicates are limited and more samples are hard to obtain. Here, relatively low strata will enjoy the benefit of more drastic power improvement. Interestingly, higher FDCs are reported in the upper strata, suggesting that more false positives are detected per true discovery in these highly expressed regions. However, this trend diminishes with increasing sample size. Given that these metrics were computed across various simulation scenarios, we further explore the variability of the results, using visualizations within a specific stratum and sample size in Fig 3C–3F. With elevated sample sizes, there is reduced variability in both targeted power and FDC (Fig 3C and 3D). This is not surprising since it is more likely to capture the true dispersion with more replicates, leading to more consistent power estimates. However, this trend is not observed across the strata at a fixed sample

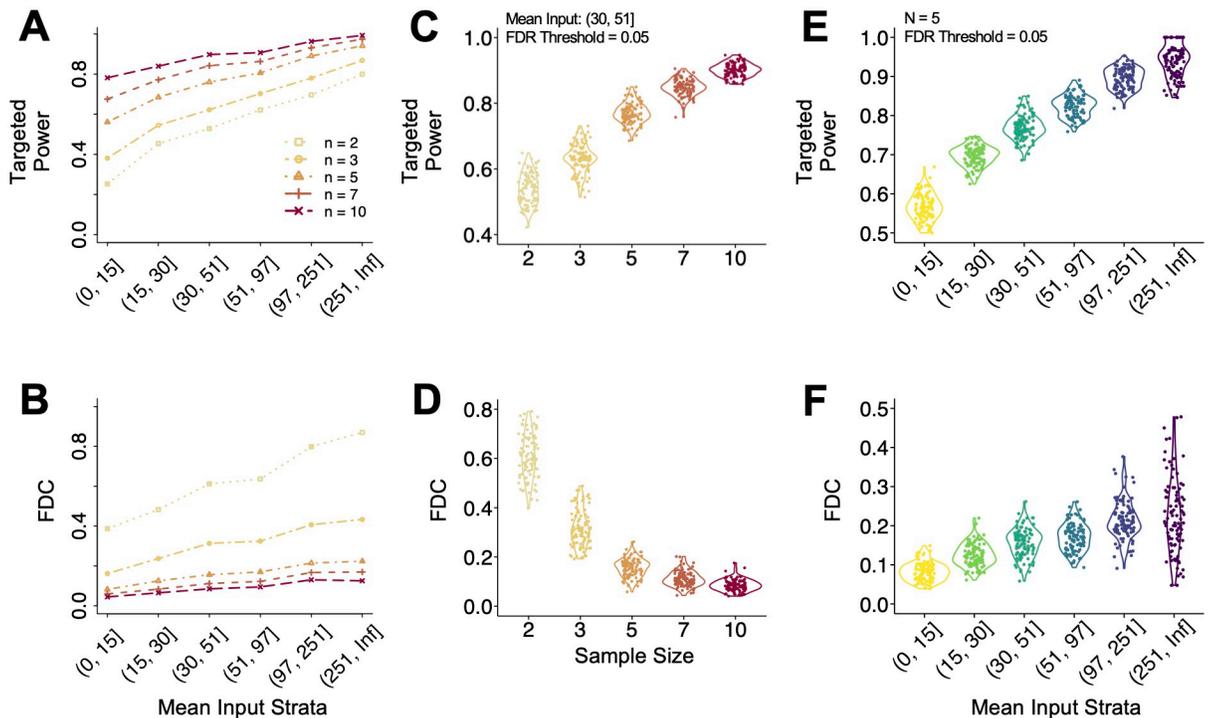


Fig 3. Targeted power and FDC stratified by mean input values for DMRs with $|\beta| \geq 2$. Six strata are defined based on input count data quantiles: stratum 1 (0%, 10%), stratum 2 (10%, 30%), stratum 3 (30%, 50%), stratum 4 (50%, 70%), stratum 5 (70%, 90%), and stratum 6 (90%, 100%). A nominal FDR value of 0.05 is used to define significance. (A), (B) Mean targeted power and FDC along strata. Each line represents one sample size choice. (C), (D) Targeted power and FDC distributions in stratum 3, separated by sample size. (E), (F) Targeted power and FDC distributions with 5 replicates per group, stratified by mean input count values. N = 100 simulations are conducted.

<https://doi.org/10.1371/journal.pcbi.1011875.g003>

size, suggesting the benefit of increasing sample size over sequencing depth for more reliable inferences. Under a fixed sample size (Fig 3E and 3F), an upward trend is observed across the strata for both targeted power and FDC. This trend aligns with the observations in Fig 3A and 3B, though some variability is evident. A heatmap panel is also available to illustrate the stratified results (Fig C in S1 Appendix).

Consistency among major DMR calling methods

It is worth noting that the targeted power and FDC presented in the Results section are computed for DMRs with odds ratios (OR) exceeding $\Delta = 2$, using TRESS. To evaluate the fluctuations of these two metrics across various effect sizes (OR), sample sizes and DMR detection methods, we also consider Δ values of 1.5, 2, 4, 6, 8, and 10, for TRESS, exomePeak2, and RADAR. In Fig 4, the targeted power and FDC are plotted against the sample size and are grouped by odds ratio thresholds. At all sample sizes, there is an increased targeted power (Fig 4A, 4C and 4E) and a higher FDC to identify DMRs with a larger odds ratio. Specifically, for FDC (Fig 4B, 4D and 4F), substantially higher values are observed among DMRs with exceptionally large odds ratios ($\Delta = 8, 10$). This indicates that detecting DMRs with these large ORs might lead to a significant increase in false positives. These patterns hold true for TRESS, exomePeak2, RADAR. While all three methods show improvements in targeted power with added replicates across all odds ratio thresholds, a discrepancy is noted for FDC that it tends to increase with larger sample sizes when using exomePeak2 and RADAR. This discrepancy, however, is not universally observed when applying our proposed framework to different pilot data sets (Fig B in S1 Appendix). Further examinations have been conducted to explore the choices of OR on the balance between sensitivity and specificity, as well as the trade-off between precision and recall (Fig E in S1 Appendix). These findings highlight the importance of utilizing the users' designated DMR detection methods during power calculation to ensure accurate estimations.

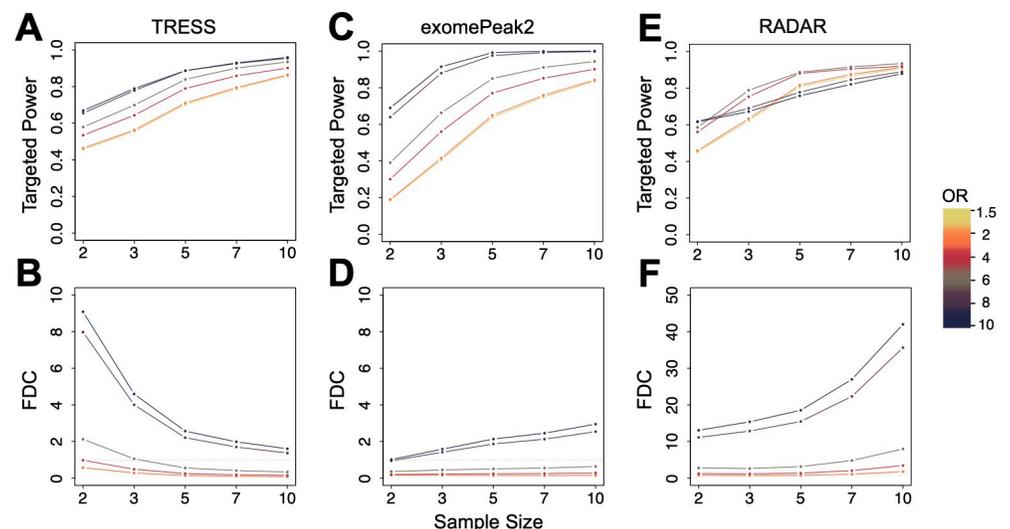


Fig 4. Comparing power evaluation results between major DMR detection methods TRESS (A)-(B) and exomePeak2 (C)-(D). Targeted power and FDC are shown at various Odds Ratios (OR, representing effect size) and sample sizes. A nominal FDR value of 0.05 is used to define significance. Points on the line plots are averaged over $N = 100$ simulations.

<https://doi.org/10.1371/journal.pcbi.1011875.g004>

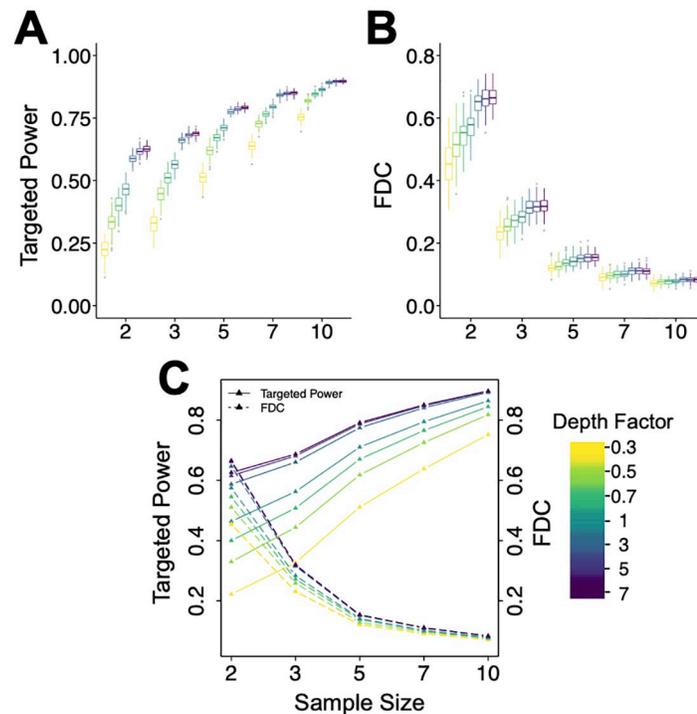


Fig 5. Sequencing depth affects targeted power and FDC for DMRs with $|\beta| \geq 2$. The ‘depth factor’ is the relative ratio of the new dataset’s library size over that from the original dataset. It reflects the impact of enlarging or down-sampling the sequencing depth of pilot data. (A), (B) Targeted power and FDC under different sequencing depths, grouped by sample size. (C) Joint visualization of the mean targeted power and FDC, over various sequencing depths and sample sizes. $N = 100$ simulations are conducted. The average sequencing depths of ‘Input’ and ‘IP’ from the pilot data are 3.51X and 0.54X, respectively.

<https://doi.org/10.1371/journal.pcbi.1011875.g005>

Impact of sequencing depth

As shown previously, sequencing depth is another critical factor in MeRIP-seq study design. Building upon our analysis of sequencing coverage strata, here we examine another aspect of sequencing depth by introducing a “depth factor”. This is a relative ratio to reflect the effect to enlarge or down-sample the sequencing coverage of the pilot data. As illustrated in Fig 5A, the targeted power rises with increased sequencing depth in all sample sizes. The incremental gain from increasing sequencing depth diminishes at high depths or large sample sizes, but benefits the small sample size the most. In Fig 5B for FDC, a similar pattern is observed as in the stratified analysis: FDCs increase with sequencing depth, but stabilize in scenarios with larger sample sizes. We also provide an integrated visualization in Fig 5C, presenting targeted power and FDC in the same panel, aiding users in understanding the tradeoffs between them. Researchers could consult similar figures, generated by *magpie* using their own pilot data, to select a customized increase in sequencing depth to achieve the desired power.

Discussions

Sample size and power evaluation are pivotal and routine tasks in experimental design using sequencing data. Here we present the first tool to address the immediate needs of sample size calculation and power estimation for DMR detection in MeRIP-seq experiments. Traditionally, sample size calculation or power evaluation in hypothesis testings depends on inputs such

as the effect size, variance from pilot studies, and the significance level. In contrast, for MeRIP-seq experiments with transcriptome-wide data, these scalar parameters must be considered as distributions. In addition, the distributions of sequencing depth and input control level can also significantly influence the statistical power, as we have shown in the results. We thus propose a statistically rigorous approach to address all these challenges, and draw information from pilot real data for simulation and empirical power evaluation.

We have a flexible simulation framework that allows switching models to mimic the real data well. In sequencing studies, data from varied tissues or cell types can exhibit unique expression and RNA methylation distributions across features (i.e. genes or regions). To address this, our tool allows users to provide pilot data analogous to their intended studies, serving as the basis for the estimated and adopted parameters in downstream simulations. To ensure that the simulated data accurately reflects actual data characteristics, *magpie* can adopt both negative-binomial and beta-binomial models and choose the one that aligns with the real data distributions best.

Both increased sequencing coverage and a larger sample size can significantly enhance the statistical power, as demonstrated in Results. Given that the total sequencing reads are often predetermined before experiments, researchers can benefit from our tool to optimize the balance between sequencing depth and sample size, to ensure the best possible experimental design in differential RNA methylation studies.

In our stratified analysis, significantly lower power is observed in regions of low input levels. This suggests the potential of refining the filtering strategy. While excluding low-expressed strata certainly means losing some true positives among these regions, it boosts the power to detect DMRs that are highly expressed, which are often of greater biological interest. Our proposed tool *magpie* can offer a foresight into the overall power gain, should the researchers want to weigh the tradeoffs before initiating their data analyses.

Our proposed approach captures real data characteristics, simulates data under various experimental settings, and produces common power evaluation metrics. This statistical framework has been implemented into a user-friendly R/Bioconductor package *magpie*. The package allows users to save power evaluation results as an Excel file and visualize their relationship with aforementioned factors with line plots. Recognizing that users might not have their own pilot MeRIP-seq data, we also develop a “*quickPower*” function. This function can generate comprehensive power evaluation outputs in seconds, by retrieving pre-calculated results from three published studies. *magpie* is available at <https://bioconductor.org/packages/magpie/>.

Supporting information

S1 Appendix. Simulation settings and additional results.
(PDF)

Author Contributions

Conceptualization: Zhenxing Guo, Wen Tang, Hao Feng.

Data curation: Zhenxing Guo, Daoyu Duan, Wen Tang.

Formal analysis: Zhenxing Guo, Daoyu Duan, Julia Zhu.

Funding acquisition: Hao Feng.

Investigation: Daoyu Duan, Julia Zhu, William S. Bush, Liangliang Zhang, Xiaofeng Zhu, Fulai Jin.

Methodology: Zhenxing Guo, Daoyu Duan.

Project administration: William S. Bush, Hao Feng.

Resources: Daoyu Duan, Fulai Jin.

Software: Daoyu Duan.

Supervision: Zhenxing Guo, William S. Bush, Liangliang Zhang, Xiaofeng Zhu, Fulai Jin, Hao Feng.

Validation: Daoyu Duan, Wen Tang.

Visualization: Daoyu Duan, Wen Tang, Julia Zhu.

Writing – original draft: Zhenxing Guo, Daoyu Duan, Hao Feng.

Writing – review & editing: Zhenxing Guo, Daoyu Duan, Hao Feng.

References

1. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, et al. N 6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*. 2014; 505(7481):117–120. <https://doi.org/10.1038/nature12730> PMID: 24284625
2. Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AAF, Kol N, Salmon-Divon M, et al. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science*. 2015; 347(6225):1002–1006. <https://doi.org/10.1126/science.1261417> PMID: 25569111
3. Oerum S, Meynier V, Catala M, Tisne C. A comprehensive review of m6A/m6Am RNA methyltransferase structures. *Nucleic Acids Research*. 2021; 49(13):7239–7255. <https://doi.org/10.1093/nar/gkab378> PMID: 34023900
4. Lasman L, Hanna JH, Novershtern N. Role of m6a in embryonic stem cell differentiation and in gametogenesis. *Epigenomes*. 2020; 4(1):5. <https://doi.org/10.3390/epigenomes4010005> PMID: 34968239
5. Chen T, Hao YJ, Zhang Y, Li MM, Wang M, Han W, et al. M6A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell*. 2015; 16(3):289–301. <https://doi.org/10.1016/j.stem.2015.02.011> PMID: 25683224
6. Chen XY, Zhang J, Zhu JS. The role of m6A RNA methylation in human cancer. *Molecular Cancer*. 2019; 18(1):1–9.
7. Lan Q, Liu PY, Haase J, Bell JL, Huttelmaier S, Liu T. The critical role of RNA M6A methylation in cancer. *Cancer Research*. 2019; 79(7):1285–1292. <https://doi.org/10.1158/0008-5472.CAN-18-2965> PMID: 30894375
8. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*. 2012; 485(7397):201. <https://doi.org/10.1038/nature11112> PMID: 22575960
9. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*. 2012; 149(7):1635–1646. <https://doi.org/10.1016/j.cell.2012.05.003> PMID: 22608085
10. McIntyre ABR, Gokhale NS, Cerchietti L, Jaffrey SR, Horner SM, Mason CE. Limits in the detection of m6A changes using MeRIP/m6A-seq. *Scientific Reports*. 2020; 10(1). <https://doi.org/10.1038/s41598-020-63355-3>
11. Dierks D, Garcia-Campos MA, Uzonyi A, Safra M, Edelheit S, Rossi A, et al. Multiplexed profiling facilitates robust m6A quantification at site, gene and sample resolution. *Nature Methods*. 2021; 18(9):1060–1067. <https://doi.org/10.1038/s41592-021-01242-z> PMID: 34480159
12. Tang Y, Chen K, Song B, Ma J, Wu X, Xu Q, et al. M6A-Atlas: A comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Research*. 2021; 49(D1): D134–D143. <https://doi.org/10.1093/nar/gkaa692> PMID: 32821938
13. Zhang Z, Zhan Q, Eckert M, Zhu A, Chryplewicz A, De Jesus DF, et al. RADAR: Differential analysis of MeRIP-seq data with a random effect model. *Genome Biology*. 2019; 20(1):1–17. <https://doi.org/10.1186/s13059-019-1915-9> PMID: 31870409
14. Guo Z, Shafik AM, Jin P, Wu H. Differential RNA methylation analysis for MeRIP-seq data under general experimental design. *Bioinformatics (Oxford, England)*. 2022; 38(20):4705–4712. <https://doi.org/10.1093/bioinformatics/btac601> PMID: 36063045

15. Duan D, Tang W, Wang R, Guo Z, Feng H. Evaluation of epitranscriptome-wide N6-methyladenosine differential analysis methods. *Briefings in Bioinformatics*. 2023; 24(3):1–11. <https://doi.org/10.1093/bib/bbad139> PMID: 37039682
16. Wu H, Wang C, Wu Z. PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*. 2015; 31(2):233–241. <https://doi.org/10.1093/bioinformatics/btu640> PMID: 25273110
17. Niu Y, Zhao X, Wu YS, Li MM, Wang XJ, Yang YG. N6-methyl-adenosine (m6A) in RNA: An Old Modification with A Novel Epigenetic Function. *Genomics, Proteomics and Bioinformatics*. 2013; 11(1):8–17. <https://doi.org/10.1016/j.gpb.2012.12.002> PMID: 23453015
18. Schwartz S, Mumbach MR, Jovanovic M, Wang T, Maciag K, Bushkin GG, et al. Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Reports*. 2014; 8(1):284–296. <https://doi.org/10.1016/j.celrep.2014.05.048> PMID: 24981863
19. Barbieri I, Tzelepis K, Pandolfini L, Shi J, Millán-Zambrano G, Robson SC, et al. Promoter-bound METTL3 maintains myeloid leukaemia by m6A-dependent translation control. *Nature*. 2017; 552(7683):126–131. <https://doi.org/10.1038/nature24678> PMID: 29186125