Detecting m^{6}A methylation regions from Methylated RNA Immunoprecipitation Sequencing Supplementary materials

Zhenxing Guo[1], Andrew M. Shafik[2], Peng Jin[2], Zhijin Wu[3], Hao Wu[1,*]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA, [2]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA and [3]Department of Biostatistics, Brown University, Providence, RI 02806, USA
[*]Correspondence: hao.wu@emory.edu.

1 Parameter estimation

1.1 Methylation level μ_i

Based on the model specification in the paper, marginally we have negative binomial distributions for observed read counts in both IP and input samples.

$$y_{ij}|\phi_i \sim NB(\mu_i(\phi_i^{-1}-1), \frac{\theta_i s_j^y}{1+\theta_i s_j^y}) \quad x_{ij}|\phi_i \sim NB((1-\mu_i)(\phi_i^{-1}-1), \frac{\theta_i s_j^x}{1+\theta_i s_j^x})$$

Then $E\{y_{ij}|\phi_i\} = \mu_i(\phi_i^{-1}-1)\theta_i s_j^y$, $E\{x_{ij}|\phi_i\} = (1-\mu_i)(\phi_i^{-1}-1)\theta_i s_j^x$. Correspondingly,

$$\frac{E(\frac{y_{ij}}{s_j^y})}{E(\frac{x_{ij}}{s_j^x})} = \frac{\mu_i}{1-\mu_i}$$

Therefore the moment estimate of μ_i is as follows,

$$\hat{\mu}_i = \frac{\sum_j (y_{ij}/s_j^y)}{\sum_j (y_{ij}/s_j^y + x_{ij}/s_j^x)}$$

1.2 Variance of the estimated methylation level $\hat{\mu}_i$

Based on above formula of $\hat{\mu}_i$, we have

$$\begin{aligned}
\text{var}(\hat{\mu}_i|\phi_i) &\approx \frac{\sum_j \text{var}(y_{ij}/s_j^y)}{(E\{\sum_j(y_{ij}/s_j^y + x_{ij}/s_j^x)\})^2} \\
&= \frac{\sum_j \frac{1}{s_j^y}\mu_i(\phi_i^{-1}-1)\theta_i(1+\theta_i s_j^y)}{n^2(\phi_i^{-1}-1)^2\theta_i^2} \\
&= \frac{\mu_i}{n^2\theta_i}\{\sum_j \frac{1+s_j^y\theta_i}{s_j^y}\}\frac{\phi_i}{1-\phi_i} \\
\text{var}(\hat{\mu}_i) &= E[\text{var}(\hat{\mu}_i|\phi_i)] + \text{var}[E(\hat{\mu}_i|\phi_i)] \\
&\approx \frac{\mu_i}{n^2\theta_i}\{\sum_j \frac{1+s_j^y\theta_i}{s_j^y}\}E\{\frac{\phi_i}{1-\phi_i}\} + \text{var}(\mu_i) \\
&= \frac{\mu_i}{n^2\theta_i}\{\sum_j \frac{1+s_j^y\theta_i}{s_j^y}\}E\{\frac{\phi_i}{1-\phi_i}\}
\end{aligned}$$

1.3 Dispersion parameter ϕ_i

The formula to calculate $\text{var}(\hat{\mu}_i)$ in Section 1.2 involves ϕ_i and θ_i. We adapt the maximum likelihood method to get the posterior modes of these two parameters. The joint posterior of ϕ_i and θ_i is proportional to the data likelihood. Then:

$$\begin{aligned}
\log f(\phi_i,\theta_i|Y_{i.},X_{i.},\mu_i) &\propto \log(\prod_j P(Y_{ij}|\phi_i)P(X_{ij}|\phi_i))f(\phi_i) \\
&= \sum_j\{\log\Gamma(\mu_i(\phi_i^{-1}-1)+Y_{ij}) - \mu_i(\phi_i^{-1}-1)\log(1+s_j^y\theta_i) + Y_{ij}\log(\frac{s_j^y\theta_i}{1+s_j^y\theta_i}) \\
&\quad + \log\Gamma((1-\mu_i)(\phi_i^{-1}-1)+X_{ij}) - (1-\mu_i)(\phi_i^{-1}-1)\log(1+s_j^x\theta_i) \\
&\quad + X_{ij}\log(\frac{s_j^x\theta_i}{1+s_j^x\theta_i})\} \\
&\quad - n\log\Gamma(\mu_i(\phi_i^{-1}-1)) - n\log\Gamma((1-\mu_i)(\phi_i^{-1}-1)) \\
&\quad - \log(\phi_i) - \frac{(\log\phi_i - m_\phi)^2}{2\sigma_\phi^2}
\end{aligned}$$

Maximizing the above likelihood provides estimates of ϕ_i and θ_i, denoted as $\tilde{\phi}_i$ and $\tilde{\theta}_i$ hereafter.

1.4 Hyperparamters

The likelihood in Section 1.3 contains m_ϕ and σ_ϕ^2, which should be estimated prior to the estimation of ϕ_i and θ_i. We adapt a three-step procedure to obtain the plug-in estimate of the mean (m_ϕ) and variance (σ_ϕ^2) of dispersion parameter ϕ_i by method of moment, based on our specified hierarchical models:

$$\begin{aligned} X_{ij}|\lambda_{ij}^x &\sim Poisson(s_j^x \lambda_{ij}^x), \\ Y_{ij}|\lambda_{ij}^y &\sim Poisson(s_j^y \lambda_{ij}^y), \\ \lambda_{ij}^x|\phi_i &\sim \Gamma((1-\mu_i)(\phi_i^{-1}-1), \theta_i), \\ \lambda_{ij}^y|\phi_i &\sim \Gamma(\mu_i(\phi_i^{-1}-1), \theta_i), \\ \phi_i &\sim \log N(m_\phi, \sigma_\phi^2). \end{aligned} \tag{1}$$

According to the log-normal prior, we first have a plug-in estimate $\hat{m}_\phi = \log(\bar{\hat{\phi}}) - (1/2) * \log(s_\phi^2/\bar{\hat{\phi}}^2 + 1)$, $\hat{\sigma}_\phi^2 = \log(s_\phi^2/\bar{\hat{\phi}}^2 + 1)$, where $s_\phi^2 = \frac{1}{N-1}\sum_i(\hat{\phi}_i - \bar{\hat{\phi}})^2$ with $\bar{\hat{\phi}} = \frac{1}{N}\sum_i^N \hat{\phi}_i$ and N being the total number of regions. Beause $\frac{\lambda_{ij}^y}{\lambda_{ij}^x + \lambda_{ij}^y} \sim Beta(\mu_i, \phi_i)$, then $\hat{\phi}_i = \frac{\frac{1}{n}\sum_j(\frac{\hat{\lambda}_{ij}^y}{\hat{\lambda}_{ij}^x + \hat{\lambda}_{ij}^y} - \tilde{\mu}_i)^2}{\tilde{\mu}_i(1-\tilde{\mu}_i)}$ with $\tilde{\mu}_i = \frac{1}{n}\sum_j \frac{\hat{\lambda}_{ij}^y}{\hat{\lambda}_{ij}^x + \hat{\lambda}_{ij}^y}$, both of which are plug-in estimates with λs replaced by their moment estimates based on the Poisson distribution of x_{ij} and y_{ij}.

2 Simulation

2.1 Region-level simulation setup

We conduct a number of simulations to evaluate the performance of TRES. The simulations are constructed based on a mouse brain dataset (more details in Section 3.1), in order to mimic the real data characteristics. In each simulation, we assume that there are 5000 candidate regions, with 80% of them being positive (with m^{6}A methylation). Let μ_i^- and μ_i^+ be m^{6}A levels for background and methylated regions respectively. We assume $\mu_i^- \sim N(m_{\mu^-}, 0.05^2)$ for negative regions, and $\mu_i^+ \sim N(m_{\mu^+}, 0.05^2)$ for positive regions. We set $m_{\mu^-} = 0.55$ and $m_{\mu^+} = 0.7$. Here, the background signal m_{μ^-} appears to be high. This is because we have a pre-filtering step, so then even the leftover negative sites can have high counts in the input samples.

We sample the dispersion parameter ϕ_i from a log-normal distribution $\log N(m_\phi, \sigma_\phi^2)$, with m_ϕ = -5 and σ_ϕ = 1.48. These numbers match the estimates of dispersion for candidate regions from a set of mouse brain data.

The scale parameter θ_i in gamma distribution is simulated based on its relationship with ϕ_i. To be specific, we observe strong correlations between the estimates of $\hat{\phi}_i$ and $\hat{\theta}_i$ among candidate regions in the real data (Figure S1). To mimic that in the simulation, we estimate $\hat{\phi}_i$ and $\hat{\theta}_i$ from the mouse brain data, and then fit a simple linear regression of $\log(\hat{\theta}_i)$ against $\log(\hat{\phi}_i)$. The OLS estimates are $a = 4.414$ for the intercept and $b = 0.865$ for the slope. Then we randomly sample the scale parameters from $\theta_i \sim \log N(a + b * \phi_i, 0.5^2)$.

To examine the effect of sample size on dispersion estimation and inference of TRES, we vary the number of replicates to be 2 and 5. Given μ_i, ϕ_i and θ_i, we randomly sample λ_{ij}^x and λ_{ij}^y from gamma distributions specified in Equation (1). Read counts of the input (x_{ij}) and the IP (y_{ij}) for each region are then sampled from the corresponding Poisson distributions given λ_{ij}^x and λ_{ij}^y, with the size factors s_j^x and s_j^y randomly sampled from a uniform distribution U(0.5, 1).

2.2 Bin-level simulation setup

First, we divide the whole transcriptome into 50bp bins and keep bins overlapped with randomly selected 18000 genes. Then, for each bin b in replicate j, the input counts X_{bj} are directly obtained from the same real dataset used to call peaks prior to this simulation. The IP counts Y_{bj} are sampled from Poisson distributions. We assume $Y_{bj}|\lambda_{bj}^y \sim Pois(s_j^y \lambda_{bj}^y)$ and $\lambda_{bj}^y = \lambda_{bj}^x * \frac{p_{bj}}{1-p_{bj}}$, where $\lambda_{bj}^x = X_{bj}/s_j^x$. Here, $p_{bj} = \frac{\lambda_{bj}^y}{\lambda_{bj}^x + \lambda_{bj}^y}$ represents the methylation level of bin b in replicate j, which is simulated using a Hidden Markov Model (HMM). In particular, let z_b denote the methylation status of bin b. $z_b = 1$ if it's overlapped with any called peak from the real data, and $z_b = 0$ otherwise. For bins that are overlapped with peak k, we assume

$$p_{bj}^k \sim Beta(\mu_b^k, \phi_b)$$
$$\mu_b^k \sim N(\mu^k, 0.02^2),$$

where μ^k is the methylation level of pre-called peak k.
For background bins with $z_b = 0$,

$$p_{bj} \sim Beta(\mu_b, \phi_b)$$
$$\mu_b \sim N(\mu_b^{(0)}, 0.01^2)$$

where $\mu_b^{(0)}$ is the bin methylation level estimated using real data.
Dispersion ϕ_b are simulated using log-normal distribution based on our observations from real data (Figure S4B).

It is important to note that, our *bin-level simulation* is based on real data. We have carefully compared the characteristics of simulated vs. real data, including the marginal distribution of bin-level read counts (Figure S4C) , the signal to noise ratios (mu in our model) (Figure S4A), and the dispersion among replicates (phi in our model) (Figure S4B). Overall, the simulated data are very similar to the real data.

2.3 Normality of Wald-test statistics and p-value distribution in region-level simulations.

We propose to use the normal distribution to calculate p-values based on the Wald test statistics. To ensure the validity of normal p-values, we examine the normality of test statistics from our *region-level simulations*.

The histograms and normal quantile–quantile (QQ) plots of Wald test statistics under different settings (Figure S2) show that the test statistics follow a normal distribution very well in the middle of the distribution, while the heavier tail to the right corresponding to the methylation regions. We further show the roughly uniformly distributed p-values under the null (Figure S3) and calculate the type I error rate using different p-value thresholds, FDR using different FDR thresholds under different simulation settings (Table S1 and S2).

2.4 Accuracy of peak calling in bin-level simulations

In addition to compuate the proportion of peaks overlapping with true peaks, we further assess the precision in peak position for each method. In particular, we calculate the percentage of base pairs in a peak that are also covered by true peaks. As shown in Figure S5, TRES still performs best in all scenarios compared to MeTPeak and exomePeak. When sample size increases, the gain of TRES becomes significantly better. Again this is due to its proper modeling of biological variance. In addition, the high precision of TRES under conditions of large dispersion (from top to bottom rows) suggests the benefits of our shrinkage procedure for the dispersion of methylation levels. When the dispersion becomes large, more extreme values will appear, which could cause unstable inferences in peak calling. A shrinkage procedure of the dispersion helps to stabilize the dispersion estimate and generate robust inference.

3 Real data analysis

3.1 Datasets

All data are obtained from the Gene Expression Omnibus (GEO) database. The first dataset (GEO accession number GSE113781) contains samples from mouse adult cortex under two conditions: treated with 15 minute acute restraint stress (*stress*), and left in homecage and sacrificed 4 hours after (*basal*). The goal of the study was to investigate the role of N6-Methyladenosine (m^6A) and N6, 2'-O-dimethyladenosine (m^6Am) in the context of brain's stress response [1]. The IP samples were processed using combined m^6A and m^6Am antibodies to immunoprecipitate the mRNA. Thus, this dataset measures the joint m^6A and m^6Am modifications. There are seven and six biological replicates in basal and stress mouse cortex sample respectively. This dataset is referred to as *Stress mouse data* hereafter. The second dataset (GSE144032) contains six-week old mouse brain samples from four brain regions: cerebellum, cortex, hippocampus and hypothalamus. The goal of this study was to investigate m^6A in different mouse brain regions. Each sample contains two replicates. This dataset is referred to as *Young mouse data* hereafter. The third dataset (GSE46705) contains four samples from human HeLa cell line: one control sample and three treated samples. The treatments correspond to the knock-out of complex METTL3, METLL14 and WTAP respectively. The goal of this study was to investigate the effect of METTL3, METTL14 and WTAP on the dynamics of m^6A on mammalian nuclear RNA [2]. Each sample contains two replicates. This dataset is referred to as *HeLa data*.

3.2 Length of top peaks

Length of top 5000 peaks, motif contents among top 5000 peaks and distance of motif DRACH to the summit of top 5000 peaks called by each method on different datasets are shown in Figure S6, Figure S7(A) and Figure S7(B) respectively.

3.3 De novo motif search

We conduct de novo motif search for top peaks in all samples called by TRES, MeTPeak and MACS2. It turns out that motif DRACH is the top one enriched motif in peaks of all samples called by TRES. As an example, we show the sequence logo of DRACH in basal and stress samples from *Stress mouse data*, cerebellum and hippocampus samples from *Young mouse data*, control and METTL3-knockout samples from *HeLa data* (Figure S8).

In addition to our peaks, we find that DRACH motif also occur in peaks called by the other methods. However, its rank and enrichment score in peaks varies by different methods. For example, among the 10 lists of peaks called by each method for all three datasets, motif DRACH is the top one enriched motif in 10, 6, 9 and 7 lists of peaks called by TRES, MeTPeak, exomePeak and MACS2 respectively. The enrichment score (defined as -log10(p-value) with the p-value reported by HOMER) of DRACH motif is mostly the highest in peaks called by TRES (Figure S9) compared to exomePeak, MeTPeak and MACS2. These results demonstrate that peaks called by TRES are more accurate and better ranked compared to the other three methods.

3.4 Consistency between peaks by TRES and other methods

To show the consistency between TRES and the other methods, we examine the overlapping pattern among top 5000 peaks by each method. All peaks are adjusted to 400 base pair long centered around their summit to avoid potential bias from peak length. As shown in Figure S10, the Venn diagrams suggest that peaks called from different methods have moderate overlaps, while each method has non-trivial number of unique peaks. To further explore the overlaps, we calculate the proportion of peaks by each method that are also reported by at least two other methods. For this, we first obtain a union of peaks by all methods. Then for each method, we generate an indicator vector of the same length with the union. The value of the indicator vector depends on whether the peaks from the union are reported by that method. Given indicator vectors for all methods, we calculate the proportion of 1s in each method that are also 1s in at least another two methods. This proportion indicates the proportion of peaks by each method that are also reported by at least two other methods. As shown in Table S3, TRES has the largest proportion in most samples compared to the other methods. These results indicate that, compared to exomePeak, MeTPeak and MACS2, peaks by TRES are more consistent with peaks by the other methods.

3.5 Effect of sequencing depth

Here, we investigate how the performance of TRES depends on sequencing depth and compare it to exomePeak, MeTPeak and MACS2. To create data with lower depth, we downsample BAM files at different rates, ranging from 0.3 to 0.7. As a criterion of comparison, for each method and each sample, we calculate the percentage of peaks called with raw data that are recaptured at different downsample rates. As shown in Figure S11, TRES reports the highest percentage compared to MeTPeak, exomePeak and MACS2. Although there is an increasing trend in the percentage for all methods as sequencing depth increases, the increasing curve of TRES is more flat, meaning that TRES is more robust to lower sequencing depth than the other methods.

3.6 Location of the peaks over the transcriptome

Transcriptome-wide distributions of top 5000 peaks called by TRES are shown in Figure S12.

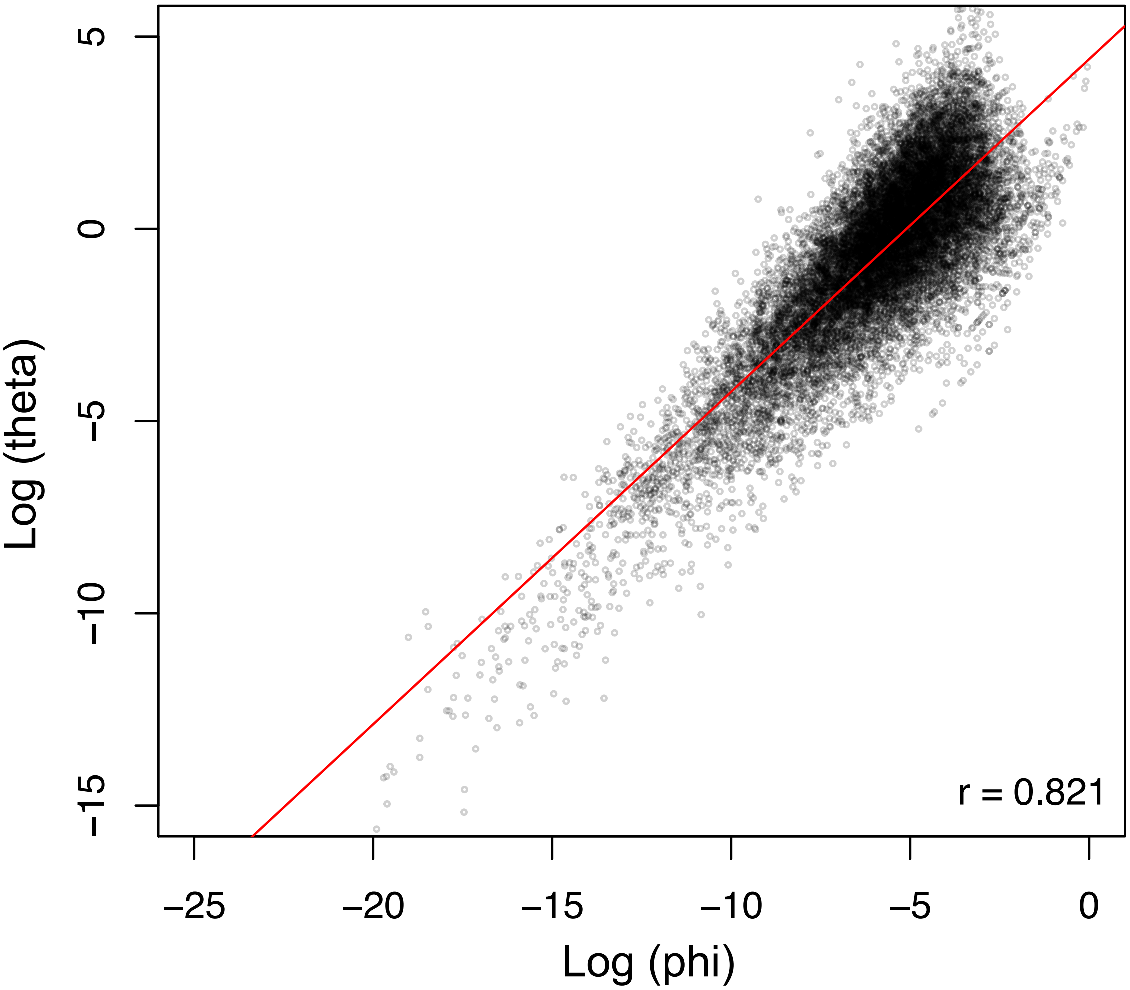

Figure S1: Plots of plug-in moment estimates for ϕ_i and θ_i using *Young mouse data*. The red line is fitted by linear regression between $\log(\hat{\theta}_i)$ and $\log(\hat{\phi}_i)$, with intercept $\alpha = 4.414$ and $\beta = 0.865$. The number to the right bottom is the Pearson's correlation between $\log(\hat{\phi}_i)$ and $\log(\hat{\theta}_i)$.

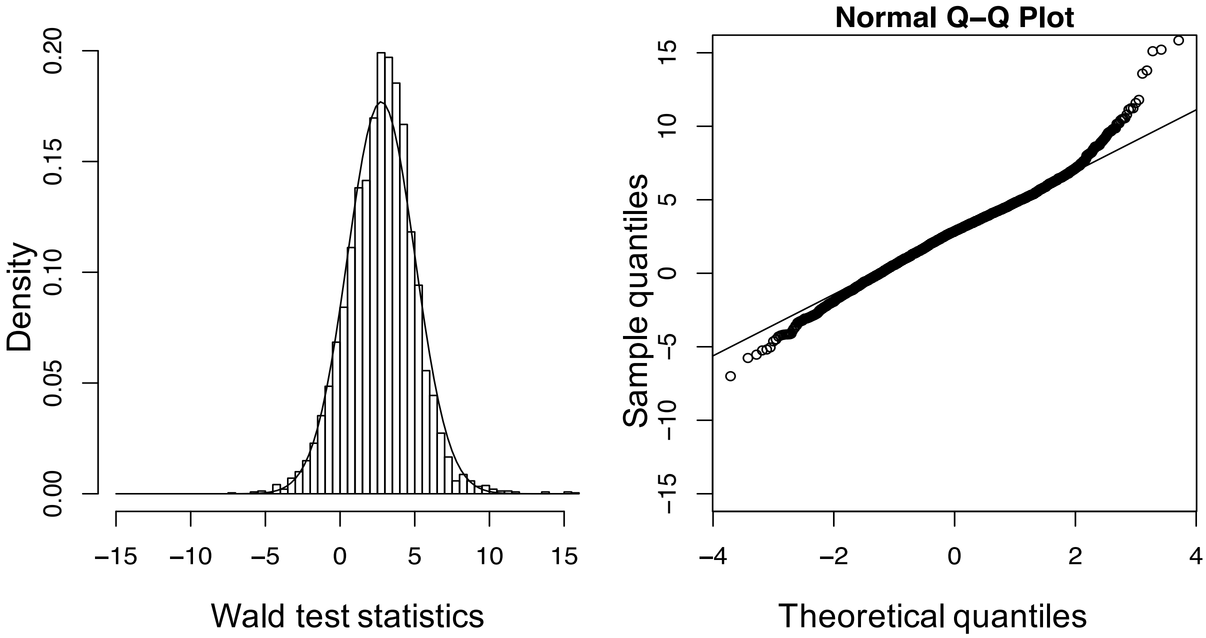

Figure S2: Histogram and QQ-plot of Wald test statistics.

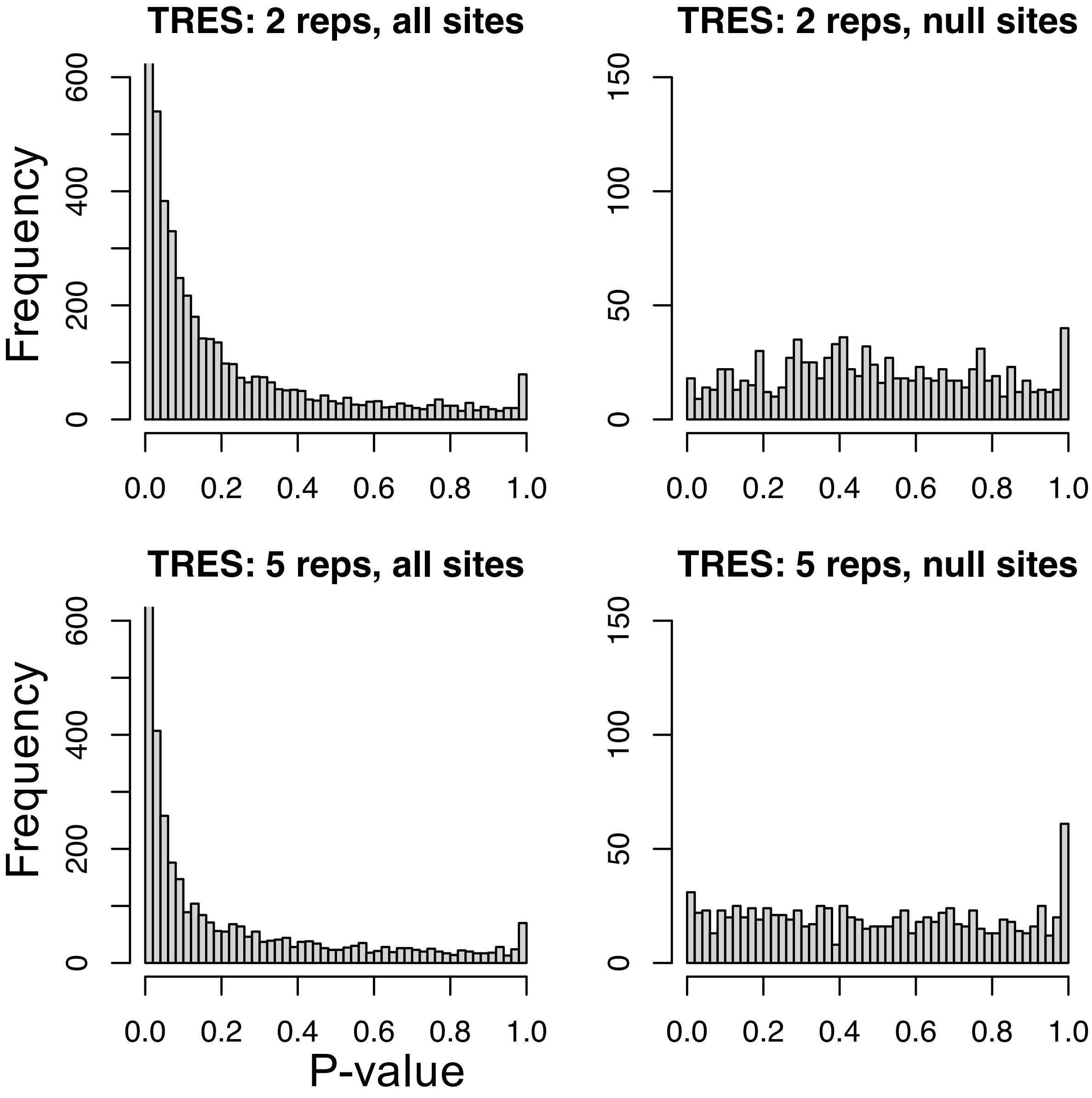

Figure S3: Histogram of p-values for all (left column) and background (right column) regions, when there are two (top row) and five replicates (bottom row).

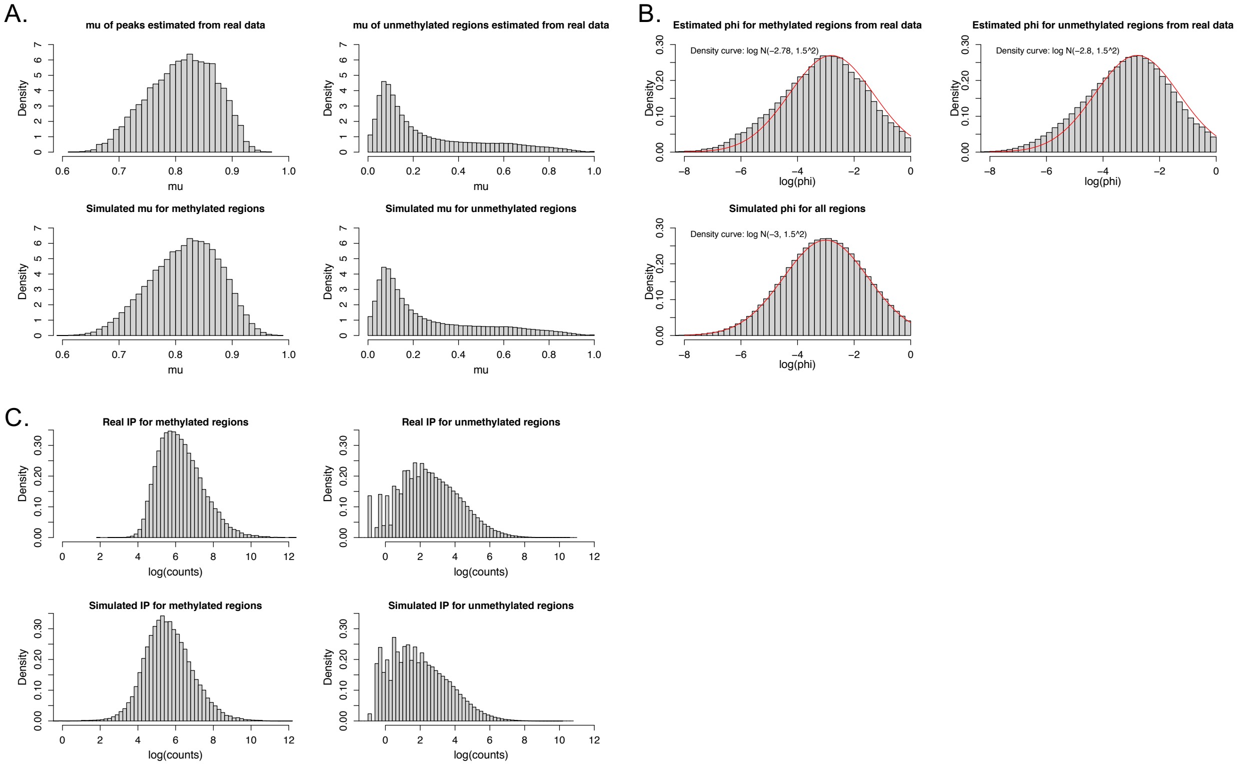

Figure S4: Comparison of real and simulated methylation level (A), dispersion of methylation level (B) and read counts in IP samples (C). (A). Methylation level of peak regions and background bins from real data (left column) and simulated methylation levels by methylation status (right column). (B). Dispersion of bins from real data (left column) and from simulated data (right column). (C). Bin-level read counts in real IP samples (top row) and in simulated IP samples (bottom row).

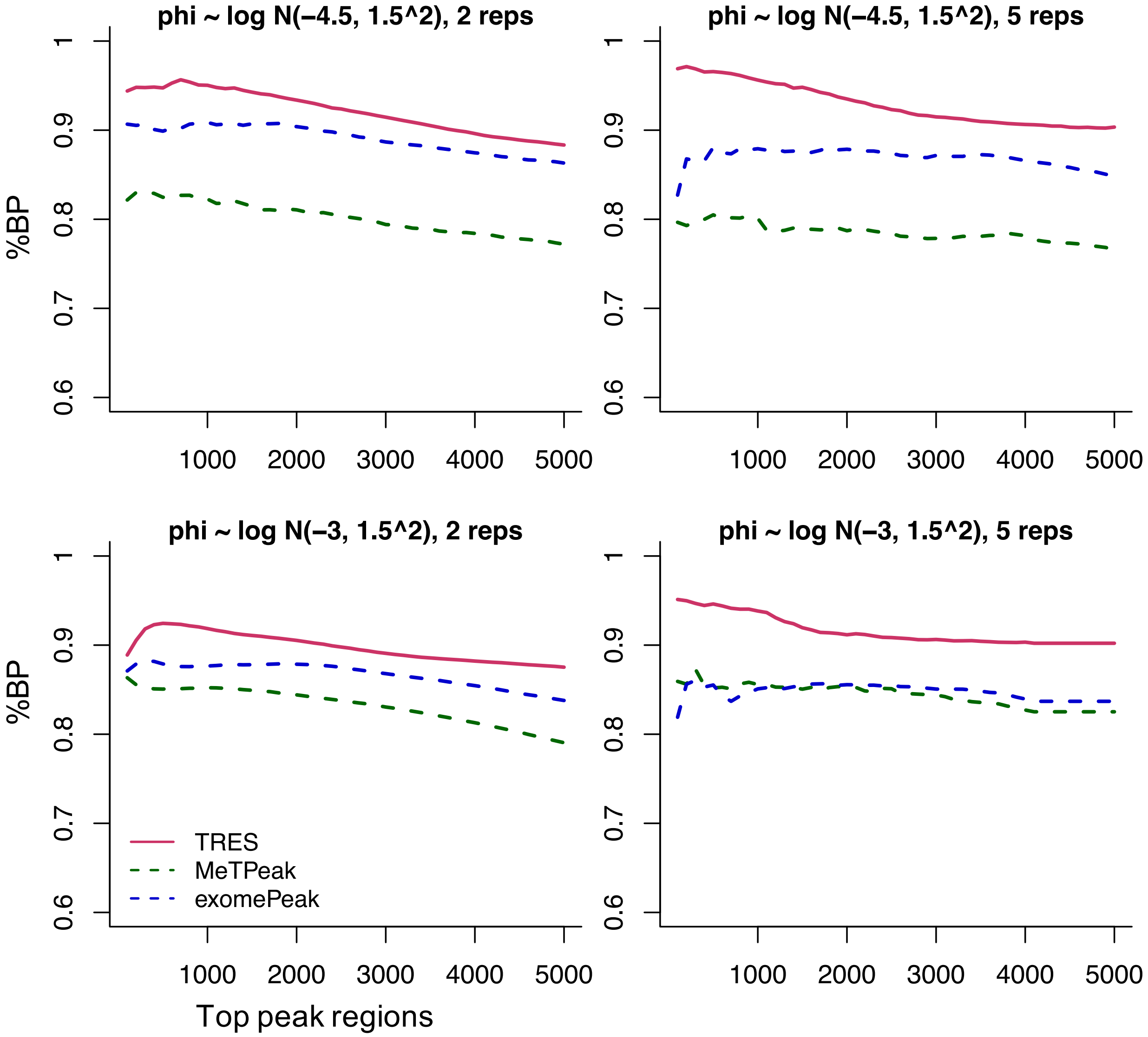

Figure S5: Comparison of percentage of base pairs from called peaks that are also covered by true peaks, among top 5000 peaks called by different methods. One panel presents results under one specific scenario. Panels from left to right in each row contain results with number of replicate 2 and 5. Panels from top to bottom in each column represent results with the mean of log-dispersion -4.5 and -3.

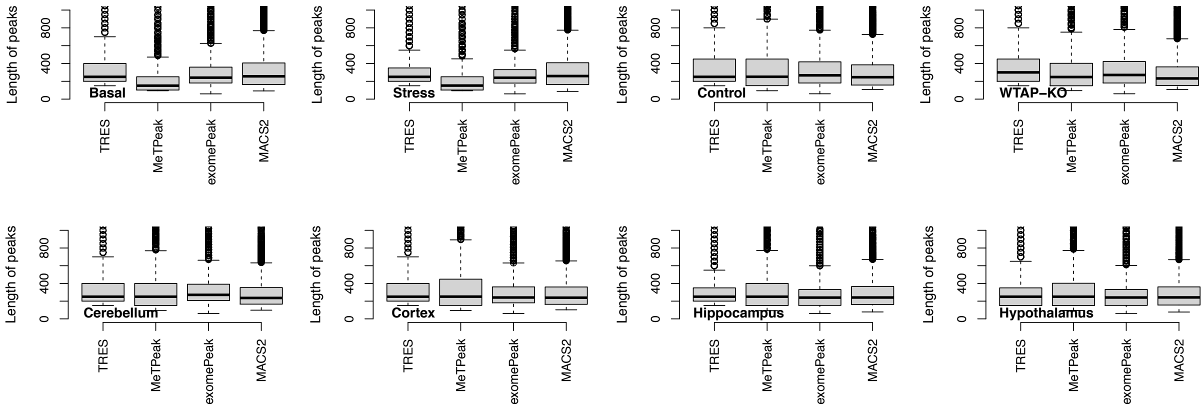

Figure S6: Boxplots for the length of top peaks called by different methods in all samples from *Stress mouse data*, Control and METTL3-knockout samples from *HeLa data* and all samples from *Young mouse data*.

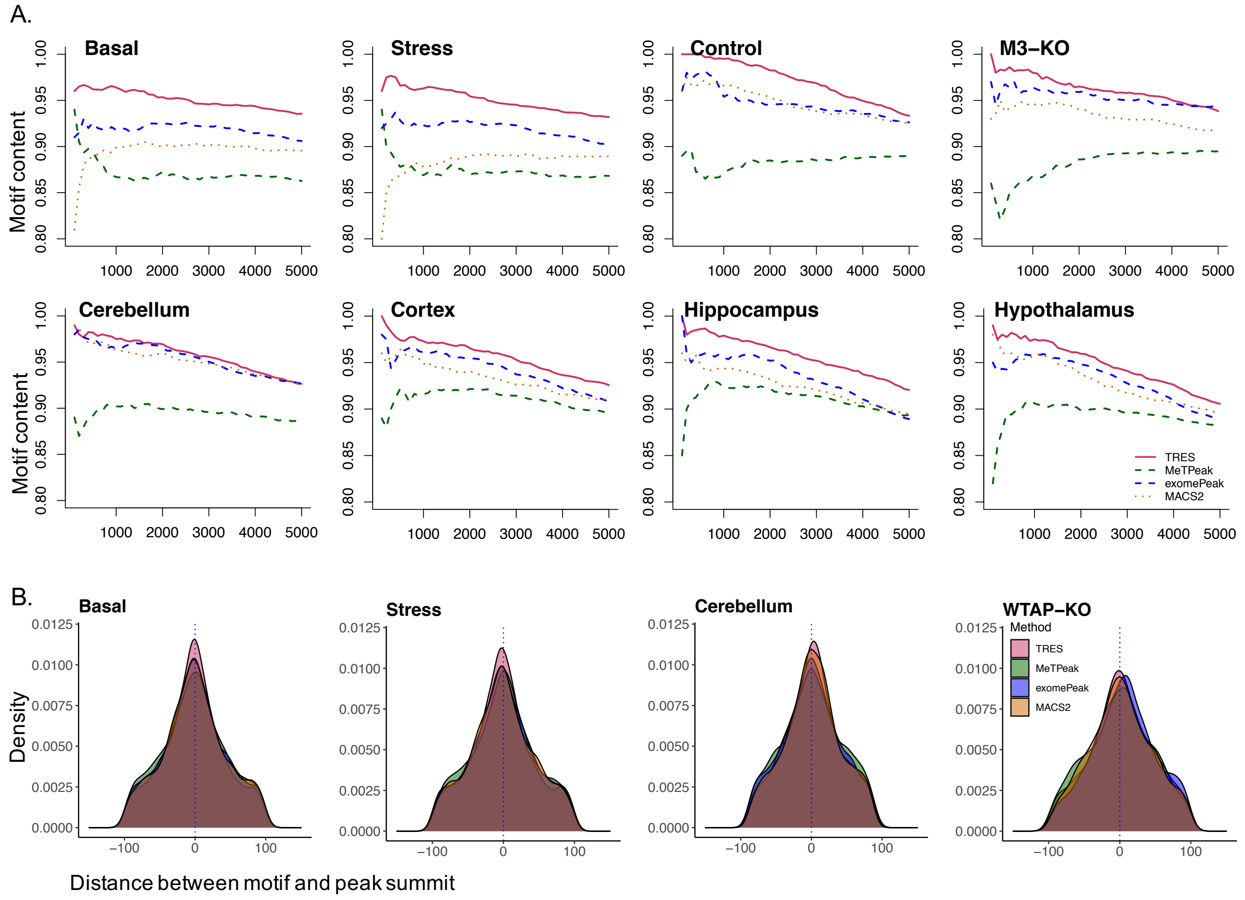

Figure S7: (**A**) Comparison of DRACH motif content among the top 5000 peak regions called by TRES, MeTPeak, exomePeak and MACS2 in basal and stress samples from *Stress mouse data*, in control and M3-knockout samples from *HeLa data*, and all samples from *Young mouse data*. (**B**) Comparison of distances between DRACH motif to peak summits called by TRES, MeTPeak, exomePeak and MACS2, in the basal and stress samples from *Stress mouse data*, the control and WTAP-knockout samples from *HeLa data*.

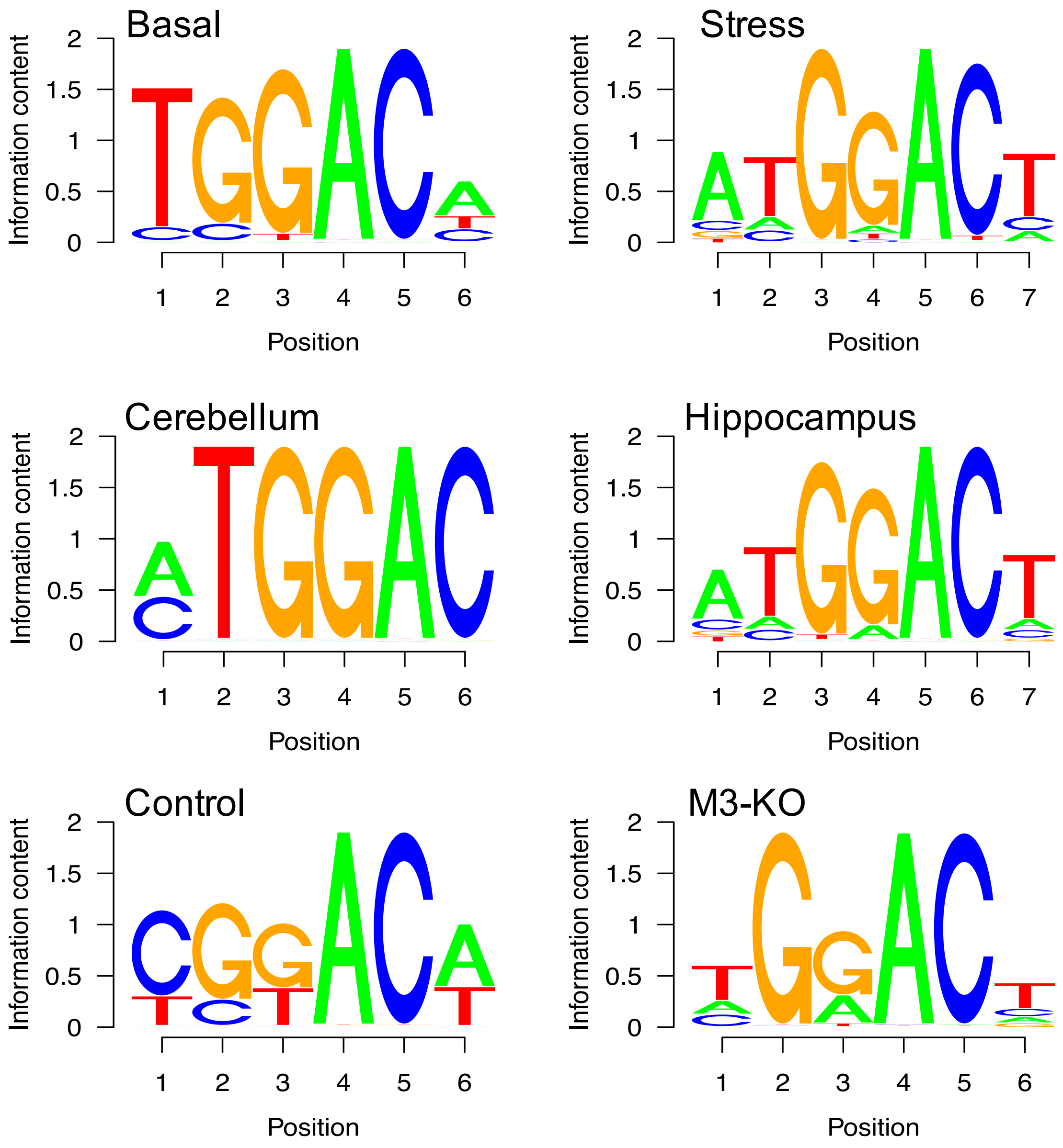

Figure S8: De novo motif search results in basal and stress samples from *Stress mouse data* (top row), and two samples from *Young mouse data* (second rwo), and two samples from *HeLa data* (third row).

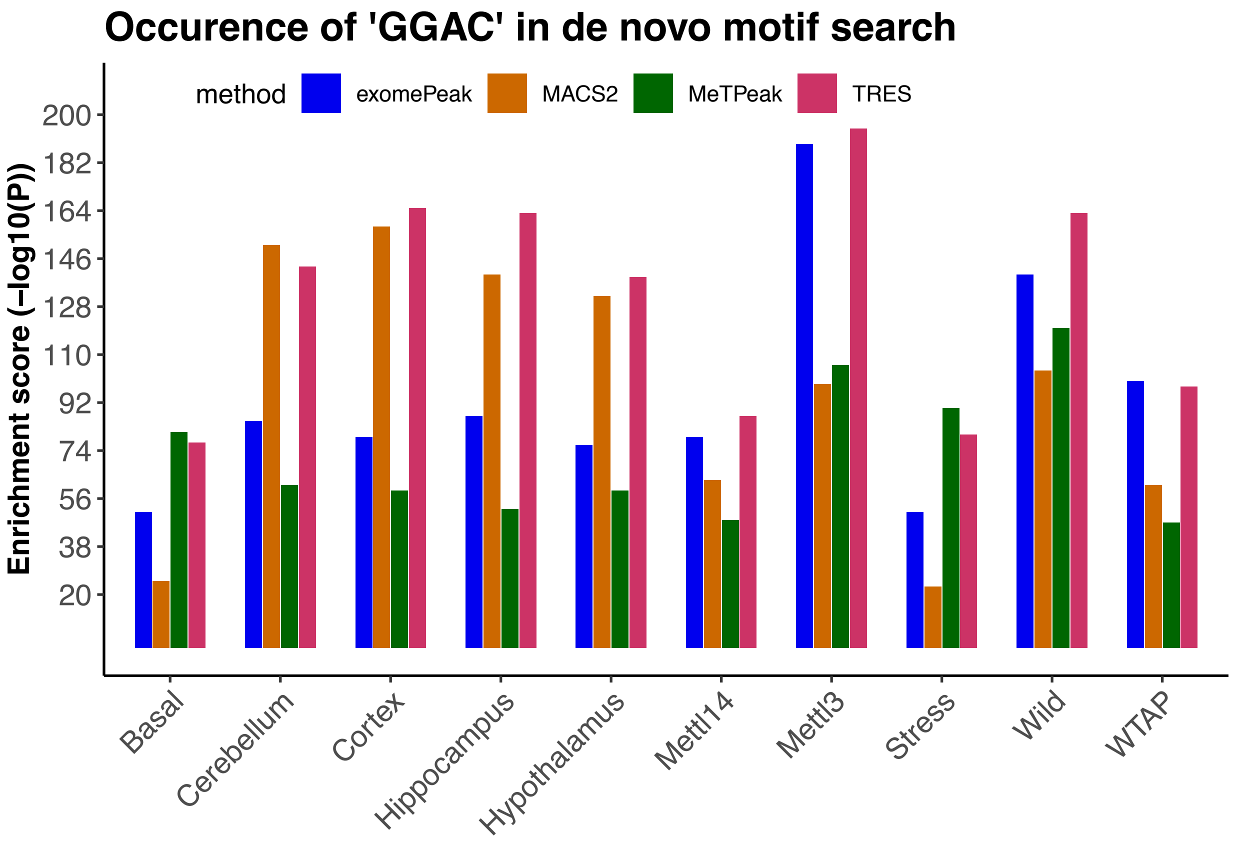

Figure S9: Comparison of motif DRACH enrichment in peaks called by different methods, The enrichment score is -log10(p-value) where all p-values are reported by HOMER in the results of *de novo* motif search.

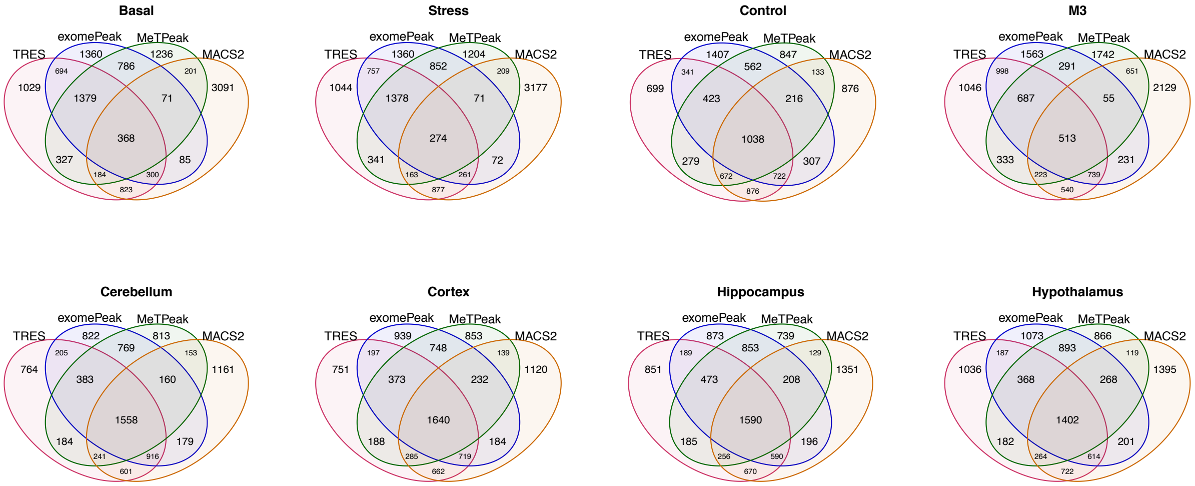

Figure S10: Overlapping patterns among top 5000 significant peaks identified by each method. Data used are all samples from Young mouse data, all samples from *Stress mouse data*, control and METTL3-knockout samples from *HeLa data*.

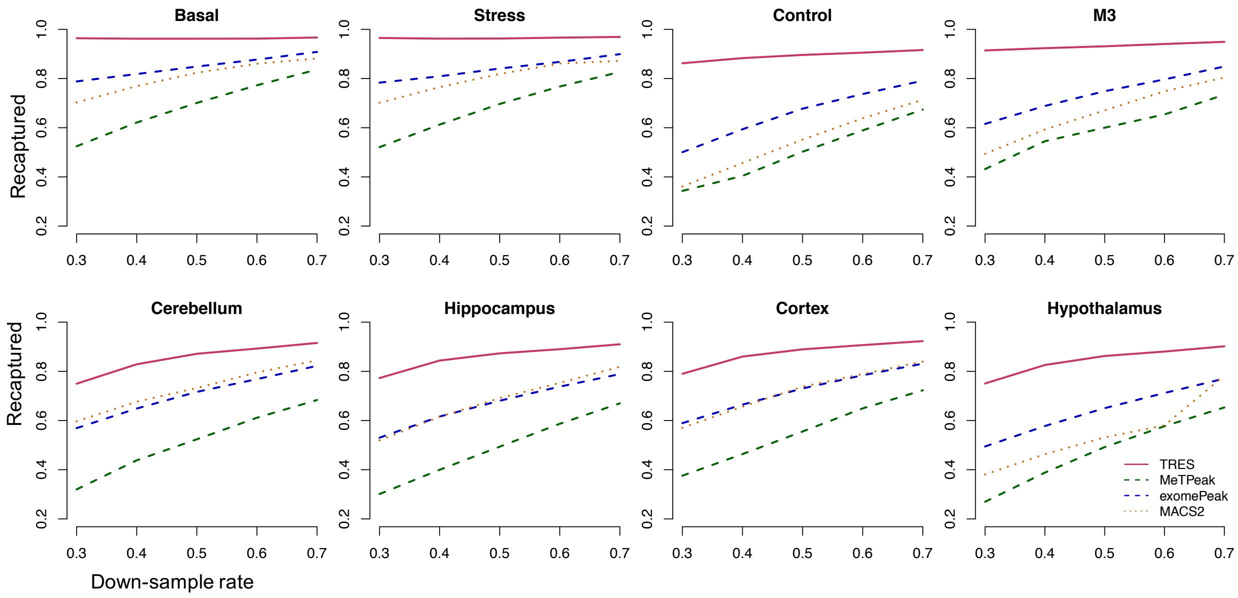

Figure S11: Percentage of peaks called with raw data that are recaptured at different sequencing depths. Data used are all samples from Young mouse data, all samples from *Stress mouse data*, control and METTL3-knockout samples from *HeLa data*. In the calculation of percentage, each method is compared to itself.

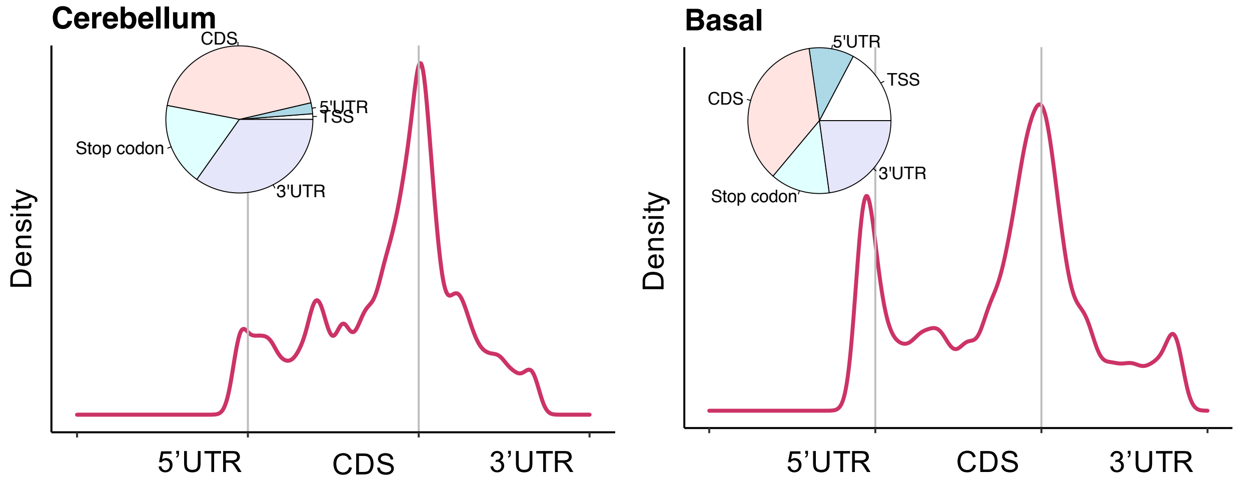

Figure S12: Pie chart and histogram to show the transcriptome-wide level distribution of top 5000 m^6A peaks in the cerebellum sample from *Young mouse data* (left panel), and top 5000 m^6A and m^6Am peaks in the basal sample from *Stress mouse data* (right panel).

Table S1: Type I error rates of TRES under different p-value thresholds.

Replicates	0.05	0.01	0.005	0.001	5e-04	1e-04	1e-05
Two reps	0.0320	0.0080	0.0050	0.0010	0.0010	0	0
Five reps	0.0650	0.0170	0.0120	0.0050	0.0030	0.0010	0

Table S2: FDRs of TRES under different FDR thresholds.

Replicates	0.05	0.01	0.005	0.001	5e-04	1e-04	1e-05
Two reps	0.0102	0	0	0	0	0	0
Five reps	0.0161	0.007	0.0056	0.0018	0	0	0

Table S3: Proportion of peaks by each method that are also reported by at least two other methods.

	TRES	exomePeak	MeTPeak	MACS2
Basal	0.19	0.18	0.17	0.08
Stress	0.17	0.16	0.16	0.06
Control	0.30	0.26	0.25	0.28
M3	0.18	0.17	0.13	0.13
Cerebellum	0.35	0.34	0.26	0.32
Cortex	0.33	0.33	0.28	0.32
Hippocampus	0.32	0.31	0.28	0.29
Hypothalamus	0.28	0.28	0.24	0.27

References

[1] Mareen Engel, Carola Eggert, Paul M Kaplick, Matthias Eder, Simone Röh, Lisa Tietze, Christian Namendorf, Janine Arloth, Peter Weber, Monika Rex-Haffner, et al. The role of m6a/m-rna methylation in stress response regulation. *Neuron*, 99(2):389–403, 2018.

[2] Jianzhao Liu, Yanan Yue, Dali Han, Xiao Wang, Ye Fu, Liang Zhang, Guifang Jia, Miao Yu, Zhike Lu, Xin Deng, et al. A mettl3–mettl14 complex mediates mammalian nuclear rna n 6-adenosine methylation. *Nature chemical biology*, 10(2):93, 2014.