Li *et al. BMC Bioinformatics* (2020) 21:31
https://doi.org/10.1186/s12859-020-3380-6

BMC Bioinformatics

SOFTWARE Open Access

m6Acorr: an online tool for the correction and comparison of m^6A methylation profiles

Jianwei Li[1,2†], Yan Huang[1†], Qinghua Cui[2] and Yuan Zhou[2*]

Abstract

Background: The analysis and comparison of RNA m^6A methylation profiles have become increasingly important for understanding the post-transcriptional regulations of gene expression. However, current m^6A profiles in public databases are not readily intercomparable, where heterogeneous profiles from the same experimental report but different cell types showed unwanted high correlations.

Results: Several normalizing or correcting methods were tested to remove such laboratory bias. And m6Acorr, an effective pipeline for correcting m^6A profiles, was presented on the basis of quantile normalization and empirical Bayes batch regression method. m6Acorr could efficiently correct laboratory bias in the simulated dataset and real m^6A profiles in public databases. The preservation of biological signals was examined after correction, and m6Acorr was found to better preserve differential methylation signals, m^6A regulated targets, and m^6A-related biological features than alternative methods. Finally, the m6Acorr server was established. This server could eliminate the potential laboratory bias in m^6A methylation profiles and perform profile–profile comparisons and functional analysis of hyper- (hypo-) methylated genes based on corrected methylation profiles.

Conclusion: m6Acorr was established to correct the existing laboratory bias in RNA m^6A methylation profiles and perform profile comparisons on the corrected datasets. The m6Acorr server is available at http://www.rnanut.net/m6Acorr. A stand-alone version with the correction function is also available in GitHub at https://github.com/emersON106/m6Acorr.

Background

RNA m^6A methylation, one of the most common RNA modifications [1], is crucial for the post-transcriptional regulations of gene expression processes, such as mRNA degradation, translation, and alternative splicing [2, 3]. Recent studies also suggest its critical roles in various biological processes, including stem cell fate decision, oncogenesis, and long-term memory consolidation [4–6].

In line with its emerging functional importance, current specialized databases, such as MeT-DB and RMBase, have accumulated a sizable amount of m^6A methylation profiles [7, 8]. In one methylation profile, the relative methylation level of each gene can be described as an enrichment score comparing the methylated read counts (m^6A-IP library) to the total read counts (input library). Ideally, the hyper- (hypo-) methylated genes of each sample can be easily determined on the basis of the enrichment score if the methylation profiles are readily intercomparable.

However, preliminary analysis of MeT-DB data failed to validate the intercomparability of current methylation profiles. Notably, all methylation profiles from MeT-DB were generated by the same computational pipeline to remove algorithm discrepancies. Therefore, intuitively, in the absence of other prominent biases, the methylation profiles of the same cell or tissue type should show prominent similarity relative to those from the same experimental report. However, the current results on the human dataset indicated the opposite: the methylation profiles from the same experiment but different cell types showed unwanted high correlations, whereas the

* Correspondence: zhouyuanbioinfo@hsc.pku.edu.cn
†Jianwei Li and Yan Huang contributed equally to this work.
[2]Department of Biomedical Informatics, School of Basic Medical Sciences, Center for Noncoding RNA Medicine, Peking University, Beijing, China
Full list of author information is available at the end of the article

correlations among the same cell type profiles across different experiments were only moderate (Fig. 1a). This tendency can also be observed in the mouse dataset (Additional file 1: Figure S1A). These results highlighted the serious but previously ignored laboratory bias in current m^{6}A methylation profiles. Several well-known normalization and batch regression methods [9] for gene expression profile correction were attempted to correct such bias, and a valid correction pipeline was finally established. This pipeline has been made available as a web server named m6Acorr, wherein users could correct their m^{6}A methylation profiles and perform comparative analysis based on the corrected profiles.

Implementation

Real-world m^{6}A profile dataset

After removing methylation profiles with small amounts of genes (e.g., methylation profiles of small RNAs), 36

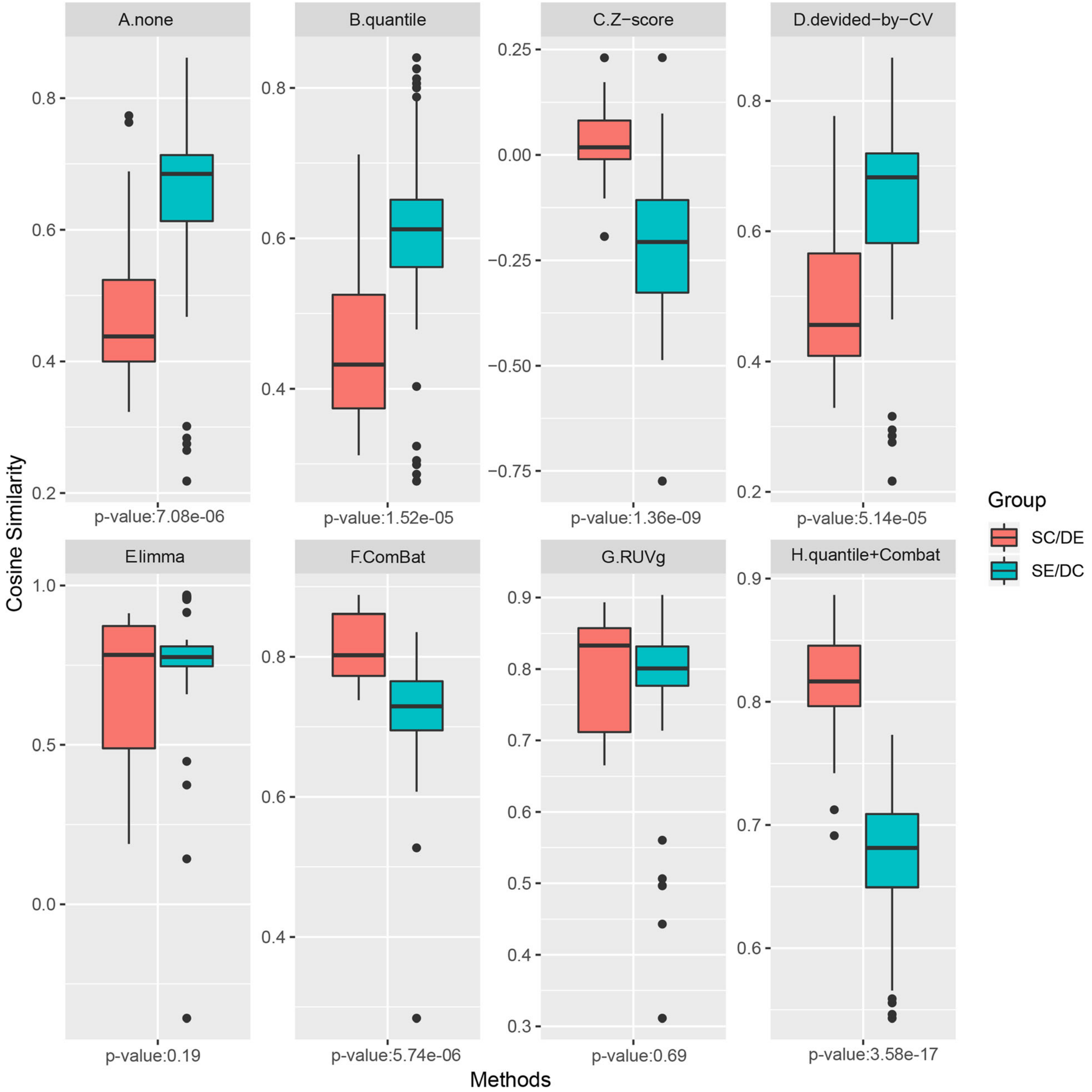

Fig. 1 Comparison of intragroup correlations between the SE/DC and SC/DE groups in the human methylation dataset. SE/DC, same experiment but different cell types; SC/DE, same cell type across different experiments. Intuitively, methylation profiles with low bias should have significantly higher correlation in SC/DE group than that in the SE/DC group, which can be achieved by combining the ComBat method and quantile normalization. *P*-values were obtained by t-test

human samples from seven experiments and 13 cell/tissue types were obtained from MeT-DB v2.0. The samples were then divided into two groups: SE/DC (same experiment but different cell types) and SC/DE groups (same cell type across different experiments). Cosine correlation was used to demonstrate the similarity between the two methylation profiles within the same group to avoid artifacts resulting from zero scores. Finally, intra-group similarities based on the two grouping criteria were compared to check the potential laboratory bias.

Simulated dataset

R (v3.6.1) package Splatter (v1.10.0) was utilized to simulate RNA m^{6}A methylation profiles with laboratory bias. Splatter gathers the parameters that reflect the distribution of real data and generates simulative data with these estimated parameters. *splatEstimate* and *splatSimulate* functions were respectively used to acquire parameters and generate an artificial dataset with laboratory bias. Two parameters, namely, batch.facLoc and batch.facScale, which respectively reflect location and scale for the batch effect factor log-normal distribution, were optimized through grid search to match the distribution of the m6A profiles from the real-world dataset (Additional file 1: Table S1). Finally, batch.facLoc and batch.facScale were set to 0.3 and 0.2 respectively which were closest to the real world. An extreme case that set batch.facLoc and batch.facScale to 0.4 was also tested and consistent results could be achieved (SE/DC under SC/DE after correction, $P = 3.11\text{e-}17$). Finally, an artificial dataset with 20 profiles, including four batches (experiments) and two cell types, was obtained to test the performance of m6Acorr.

Methylation profile correction method

The methylation correction pipeline is closely related to popular gene expression profile correction methods. Seven popular methods for normalizing and/or correcting RNA-seq and microarray data were tested to correct observed laboratory bias in the methylation profiles. (1) Quantile normalization. According to the standard setting [10], each column of the m^{6}A profiles matrix X was first sorted to obtain $X_{ordered}$, and then each row of $X_{ordered}$ was replaced with the average of the row. Finally, the matrix X was replaced by the corresponding average value based on the $X_{ordered}$ and acquired $X_{quantile}$. Q1, Q2, and Q3 quantiles were also attempted to replace the average (i.e., the standard setting); the standard setting performed slightly better than using alternative settings (Additional file 1: Figure S2). (2) Z-score normalization among samples from the same experiment. The Z-score can be defined as follows:

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the enrichment score of the gene; μ represents the average enrichment score of one experiment; and σ is the standard deviation of the sample. (3) Division of the per experiment coefficient of variation (CV). CV can be described by the following equation:

$C_v = \frac{\sigma}{\mu}$ (2)where σ is the standard deviation of the sample space, and μ represents the sample average. (4) The empirical Bayes-based batch regression method from limma package (v3.42.0) [11] was performed with the default parameters. (5) The empirical Bayes-based batch regression method from SVA package (v3.34.0), which is also known as the ComBat method [12], was performed with the default parameters. (6) The RUVg method from RUVseq package (v1.20.0), which removes the batch effect according to the control genes [13], was performed. In this study, the top 10% genes showing the most constant expression in the GTEx transcriptome atlas were selected as the control genes [14]. (7) Considering the possible differences between the profiles, these genes were adjusted to the same distribution with quantile normalization to coordinate with the Combat model for correction. The combined pipeline is finally adopted in m6Acorr, which is closely related to the recommended pipeline for gene expression correction [9].

Analysis of corrected methylation profiles

The m6Acorr server is available at http://www.rnanut.net/m6Acorr. In addition to profile correction, the m6Acorr server also provides the comparative analysis functions of the corrected profiles. After correction, hyper- (hypo-) methylated genes could be obtained on the basis of the Z-score comparison of user-provided samples to the MeT-DB samples. The functional enrichment analysis of the hyper- (hypo-) methylated genes is also enabled by investigating the enrichment of gene sets curated from (1) WikiPathways pathway gene sets [15]; (2) MSigDB hallmark gene sets [16]; (3) RNA-binding protein target genes from ENCODE project [17]; and (4) miRNA target genes from miRTarBase using the hypergeometric test [18]. Finally, a stand-alone version with the correction function is also available in GitHub at https://github.com/emersON106/m6Acorr.

Results

As previously described in the Background Section, the high correlation among the samples from the same experiment but different cell types (i.e. the SE/DC group) suggested serious laboratory bias. This result indicates that heterogeneous profiles show unwanted higher similarity compared with those from the same cell type across different experiments (i.e. the SC/DE group) (SE/

DC over SC/DE, t-test, *P* = 7.08e-06; Fig. 1 a). Several normalization methods have been applied, and only Z-score normalization partially reversed this biased correlation (Fig. 1b–d). Several popular batch-regression methods were further applied because normalization methods were not intended for experimental batch bias correction. Among these methods, the ComBat method successfully reversed the high correlation in the SE/DC group (SE/DC under SC/DE, *P* = 5.74e-06; Fig. 1f). Finally, by combining ComBat and quantile normalization, the unwanted high correlation in the SE/DC group was prominently reversed, implying effective bias elimination (SE/DC under SC/DE, *P* = 3.58e-17; Fig. 1h). The mouse dataset from MeT-DB V2.0 and the simulated data for independent verification of the correction pipeline were further adopted. Similar results were observed for the mouse dataset (Additional file 1: Figure S1), where the unwanted high correlation (SE/DC over SC/DE, *P* = 2.44e-08) was successfully reversed (SE/DC under SC/DE, *P* = 9.52e-14). Such a tendency reversal on the mouse dataset verified the validity of the "quantile + Combat" correction pipeline for the real-world dataset. In addition, considering the limited coverage of current m^{6}A profiles, the correction pipeline was applied to an artificial simulated dataset. Notably, the "quantile + Combat" correction pipeline on this simulated dataset also exhibited competitive efficiency (SE/DC under SC/DE, *P* = 1.99e-17 after correction; Additional file 1: Figure S3). Therefore, the "quantile + Combat" combination pipeline, which was named m6Acorr hereafter, was selected for m^{6}A methylation profile correction.

Although m6Acorr could efficiently reduce laboratory bias, one prominent concern is the elimination of biological signals after correction. The retention of differential methylation signals was first examined to further check if the m6Acorr pipeline would disturb biological signals and thus obstruct downstream applications. To this end, two representative pairs of profiles in MeT-DB V2.0, which investigate the alteration in m^{6}A methylation after the knockdown of methyltransferases *METTL3* and *METTL14* (the enzymes catalyzing m^{6}A methylation), were selected (p007_HeLa1_ctrl with p007_HeLa1_KO_M14, and p007_HeLa2_ctrl with p007_HeLa2_KO_M3). Note also that these profiles were independent from the above assessment of correction pipelines because they were derived from the m6A enzyme mutant cells rather than wild-type cells. The shared differentially methylated genes before and after correction were compared by calculating the Jaccard index between the top 20% differentially methylated genes. As shown in Fig. 2, m6Acorr exhibited better preservation of differentially methylated genes than the alternative methods. In addition, the fraction of shared differentially methylated genes could not be achieved by randomly selected genes. Given that the fraction of preserved differentially methylated genes was only moderate, the methods were examined by checking if the differentially methylated genes identified after correction could show good consistency with the functional m^{6}A target genes [2, 19]. Two typical classes of functional m^{6}A target genes were considered. The first class is the genes whose translational efficiency is intensively regulated by m^{6}A modification. These genes showed remarkable decreases in translational efficiency after *METTL3* or *METTL14* knockdown as recorded in the GEO dataset GSE63591. The second class is the genes whose mRNA stability is intensively regulated by m^{6}A modification. These genes showed significant increases in mRNA stability after *METTL3* or *METTL14* knockdown as recorded in the GEO dataset GSE49339. The comparison results are summarized in Fig. 3. Interestingly, differentially methylated genes identified after correction showed good consistency with either class of functional m^{6}A target genes, even when compared with the differentially methylated genes identified from the uncorrected methylation profiles. These results indicated that methylation profile correction by m6Acorr is also helpful for finding the important functional targets of m^{6}A regulation. Finally, the comparative analysis of the biological features of frequently and occasionally methylated genes was performed using the corrected methylation profiles. Previous comparative analysis using the uncorrected methylation profiles indicated that the overall m^{6}A methylation breadth across samples was correlated with gene importance-related features, including dN/dS ratio, tissue expression specificity, and PPI network degree [20]. Notably, such significant correlations were retained after correction (Additional file 1: Figure S4), indicating the preservation of biological signatures.

Discussion

In the above section, we have demonstrated the effectiveness of the 'normalization + correction' pipeline for m^{6}A methylation profile correction. The pipeline adopts well-known methods for gene expression profile correction [9]. However, the suitability and effectiveness of such methods on m^{6}A methylation profile correction is not naturally guaranteed. In fact, at the start point, to which extent laboratory bias exists in m^{6}A methylation profiles has not been systematically explored. As for the intuitive assumption according to the principle of m^{6}A profiling technique (i.e. MeRIP-seq) [7, 8], since the methylation levels are derived by comparing the methylated read counts against the total read counts from the

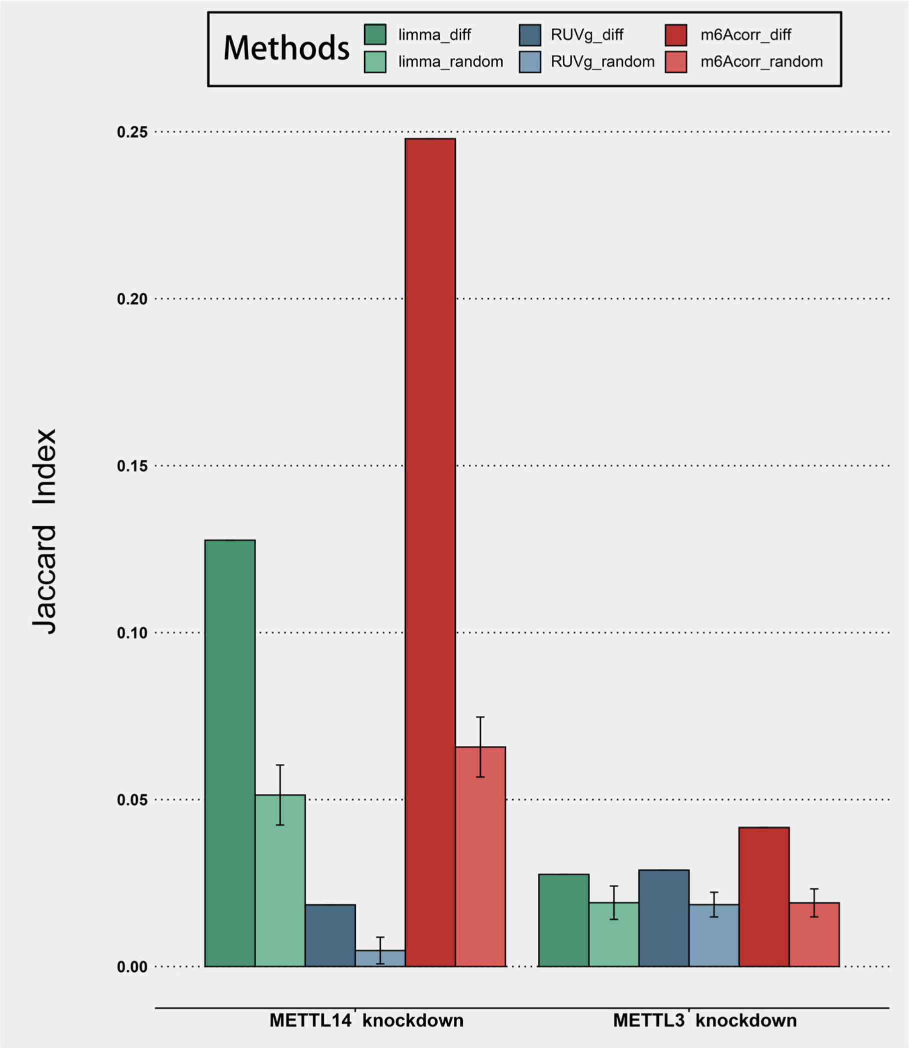

Fig. 2 Jaccard index depicting the shared fraction between the differentially methylated genes identified before and after correction by three methods. Diff: the top 20% differentially methylated genes; random: randomly selected same amount of genes (repeated 100 times, error bar showing the standard error)

same sample, the laboratory bias should be canceled out during such intra-sample comparison. But as what we have shown above, the laboratory bias turns out to be quite serious, and on the other hand, not all of the well-known methods work well for the methylation profile correction. Therefore, the novelty of this study is focused on why and how the correction pipeline should be applied to the m^{6}A methylation profiles.

The pipeline also has been made available as the m6Acorr server, which could perform methylation profile correction based on the user-provided batch (experiment) assignment. If no batch is assigned, then the entire dataset will be treated as one experiment. In addition, users could assign experimental groups of samples (e.g., diseased and healthy). Thus, the hyper- (hypo-) methylated genes for each group can be derived on the basis of the Z-score by comparing the intra-group methylation level with the methylation level in other profiles. Moreover, their enriched functions can be analyzed on the basis of the curated gene set annotations in the m6Acorr server (Fig. 4).

Conclusions

The current work focused on the existing laboratory bias in RNA m^{6}A methylation profiles of public databases and developed m6Acorr, a pipeline for m^{6}A profile correction based on quantile normalization and empirical Bayes batch regression methods. m6Acorr achieved the favorable results in real and artificial datasets. While the bias was eliminated by m6Acorr, the biological signals

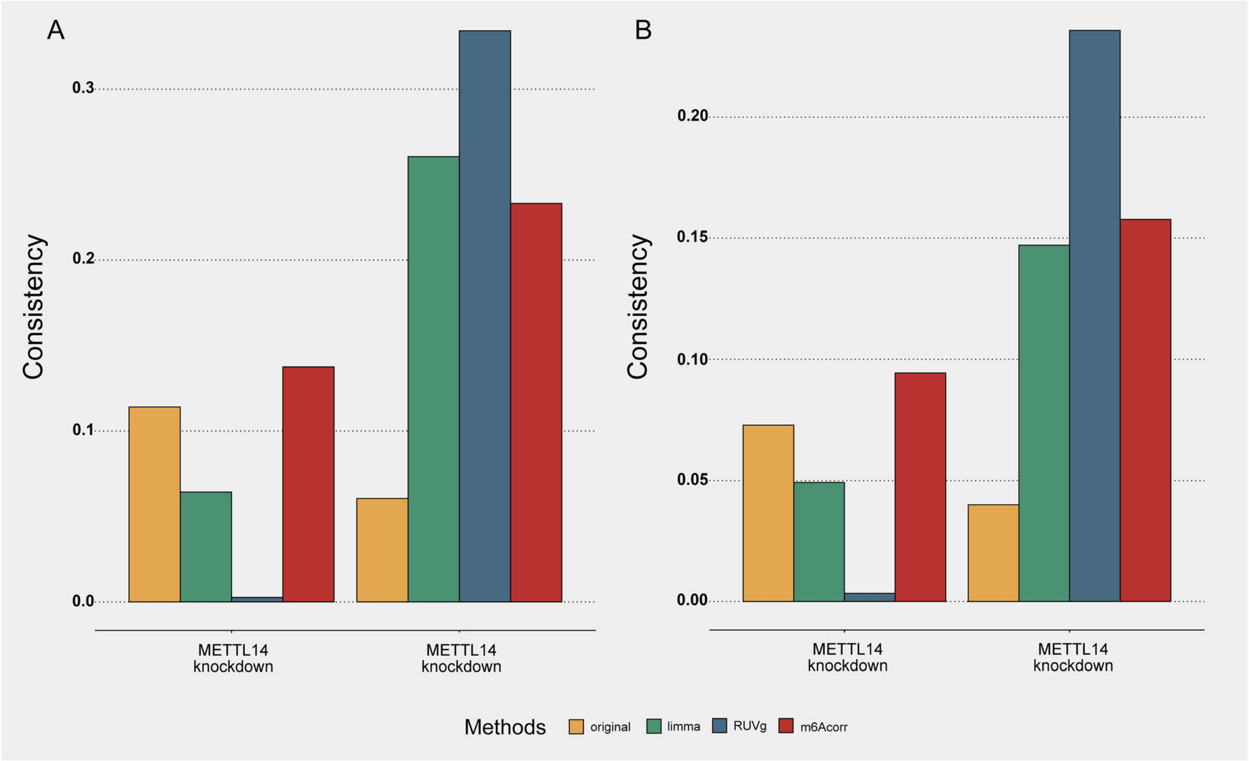

Fig. 3 Consistency of differentially methylated genes (before and after corrections) with functional m^{6}A target genes. **a**. The consistency with the m^{6}A target genes whose translation efficiency is significantly reduced after *METTL3* or *METTL14* knockdown. **b**. The consistency with the m^{6}A target genes whose mRNA stability is significantly increased after *METTL3* or *METTL14* knockdown

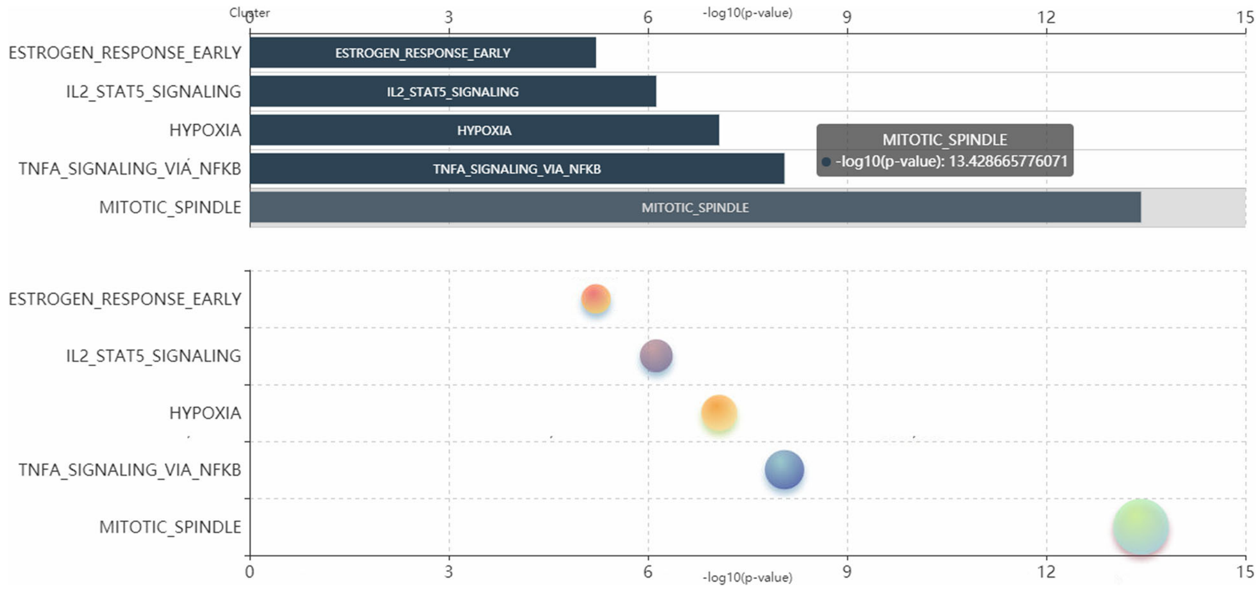

Fig. 4 An example functional enrichment analysis results of hypermethylated genes

were well preserved after correction. The m6Acorr server could also be used to compare m^6A profiles and conduct the functional analysis of hyper- (hypo-) methylated genes based on corrected methylation profiles. Overall, the m6Acorr server could be a useful tool for the correction and comparison of m^6A methylation profiles.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3380-6.

Additional file 1: Figure S1. Comparison between intra-group correlation among SE/DC group and SC/DE group in mouse methylation dataset. **Figure S2.** Comparison between intra-group correlation among SE/DC group and SC/DE group under different quantiles. **Figure S3.** Comparison between intra-group correlation among SE/DC group and SC/DE group in the simulated dataset. **Figure S4.** Correlation curves between the m^6A regulation breadth and various gene importance-related features. **Table S1.** Grid search of the parameters to fit the real world laboratory bias

Abbreviations

CV: Coefficient of variation; IP: Immunoprecipitation; m^6A: N^6-methyladenosine; miRNA: MicroRNA; mRNA: Messenger RNA; PPI: Protein–protein interaction; SC/DE: Same cell but different experiments; SE/DC: Same experiment but different cells

Acknowledgements

We appreciate the researchers who shared their m^6A methylation profiles and analysis work.

Availability and requirements

Project name: m6Acorr.
Project home page: http://www.rnanut.net/m6Acorr (online webserver version) and https://github.com/emersON106/m6Acorr (standalone version).
Operating system(s): Platform independent.
Programming language: Python, R, JavaScript, PHP.
Other requirements: Online webserver requires Chrome, Firefox or Microsoft Edge browser.
License: FreeBSD.
Any restrictions to use by non-academics: None.

Authors' contributions

YH collected the data and designed and implemented the experiments, and wrote the paper. YZ conceived and led the project, designed the experiments and analyzed the results and wrote the paper. JL and Q.C. evaluated the methods, suggested improvements and analyzed the results. All authors have read and approved the final manuscript.

Funding

This study was supported by National Natural Science Foundation of China [81672113 to J.L., 31801099 to Y.Z.], Natural Science Foundation of Hebei Province [C2018202083 to J.L.] and the Fundamental Research Funds for the Central Universities [BMU2017YJ004 to Y.Z.]. The funding body has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the MeT-DB v2.0 repository, http://www.xjtlu.edu.cn/metdb2.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute of Computational Medicine, School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. [2]Department of Biomedical Informatics, School of Basic Medical Sciences, Center for Noncoding RNA Medicine, Peking University, Beijing, China.

Received: 2 October 2019 Accepted: 21 January 2020
Published online: 29 January 2020

References

1. Lewis CJ, Pan T, Kalsotra A. RNA modifications and structures cooperate to guide RNA-protein interactions. Nat Rev Mol Cell Biol. 2017;18(3):202–10.
2. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature. 2014;505(7481):117–20.
3. Xiao W, Adhikari S, Dahal U, Chen YS, Hao YJ, Sun BF, Sun HY, Li A, Ping XL, Lai WY, et al. Nuclear m(6) a reader YTHDC1 regulates mRNA splicing. Mol Cell. 2016;61(4):507–19.
4. Batista PJ, Molinie B, Wang J, Qu K, Zhang J, Li L, Bouley DM, Lujan E, Haddad B, Daneshvar K, et al. M(6) a RNA modification controls cell fate transition in mammalian embryonic stem cells. Cell Stem Cell. 2014;15(6):707–19.
5. Choe J, Lin S, Zhang W, Liu Q, Wang L, Ramirez-Moya J, Du P, Kim W, Tang S, Sliz P, et al. mRNA circularization by METTL3-eIF3h enhances translation and promotes oncogenesis. Nature. 2018;561(7724):556–60.
6. Zhang Z, Wang M, Xie D, Huang Z, Zhang L, Yang Y, Ma D, Li W, Zhou Q, Yang YG, et al. METTL3-mediated N(6)-methyladenosine mRNA modification enhances long-term memory consolidation. Cell Res. 2018;28(11):1050–61.
7. Liu H, Wang H, Wei Z, Zhang S, Hua G, Zhang SW, Zhang L, Gao SJ, Meng J, Chen X, et al. MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. Nucleic Acids Res. 2018;46(D1):D281–7.
8. Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, Qu LH, Yang JH. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res. 2018;46(D1):D327–34.
9. Muller C, Schillert A, Rothemeier C, Tregouet DA, Proust C, Binder H, Pfeiffer N, Beutel M, Lackner KJ, Schnabel RB, et al. Removing batch effects from longitudinal gene expression - Quantile normalization plus ComBat as best approach for microarray Transcriptome data. PLoS One. 2016;11(6):e0156594.
10. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
11. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
12. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.
13. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):896–902.
14. Consortium GT. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
15. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Melius J, Cirillo E, Coort SL, Digles D, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46(D1):D661–7.
16. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
17. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46(D1):D794–801.
18. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, et al. MiRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018;46(D1):D296–302.

19. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. N(6)-methyladenosine modulates messenger RNA translation efficiency. Cell. 2015;161(6):1388–99.
20. Zhou Y, Cui Q. Comparative analysis of human genes frequently and occasionally regulated by m(6) a modification. Genomics Proteomics Bioinformatics. 2018;16(2):127–35.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.