ComBat-met: Adjusting Batch Effects in DNA Methylation Data

Junmin Wang[1,*]

[1]Data Sciences and Quantitative Biology, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Waltham, Massachusetts

*To whom correspondence should be addressed: junmin.wang@astrazeneca.com.

Abstract

Integration of genomics data is routinely hindered by unwanted technical variations known as batch effects. Despite wide availability, existing batch correction methods often fall short in capturing the unique characteristics of DNA methylation data. We present ComBat-met, a beta regression framework to adjust batch effects in DNA methylation studies. Our method fits beta regression models to the data, calculates batch-free distributions, and maps the quantiles of the estimated distributions to their batch-free counterparts. Compared to traditional methods, ComBat-met followed by differential methylation analysis shows improved statistical power without compromising false positive rates based on simulated data. Additionally, we demonstrate the ability of ComBat-met to remove cross-batch variations and recover biological signals using data from The Cancer Genome Atlas.

1 Introduction

High-throughput genomics technologies have revolutionized our ability to investigate complex biological systems. Among these, DNA methylation profiling has emerged as a powerful tool for studying epigenetic regulation, providing insights into gene expression, organoid development, disease mechanisms, and environmental influences.

A key step in DNA methylation profiling, whether through bisulfite sequencing or methylation microarrays, is the conversion of unmethylated cytosines to thymines via bisulfite treatment (Frommer et al. 1992). This process is crucial for distinguishing between methylated and unmethylated cytosines, as only unmethylated cytosines are converted, leaving methylated cytosines unchanged. However, variations in the efficiency of this cytosine to thymine conversion across experimental batches can introduce systematic biases in methylation data (Ross et al. 2022; Price and Robinson 2018). Factors such as differences in bisulfite treatment conditions, variations in DNA quality or quantity, and technical inconsistencies in the bisulfite conversion reaction itself can all lead to batch effects (Olova et al. 2018). Failure to address batch effects can obscure true biological signals, impeding the accuracy and reproducibility of downstream analyses (Wang and Novick 2024; Akulenko, Merl, and Helms 2016).

Various methods have been developed to correct for batch effects in genomics data. Among them, ComBat, which uses an empirical Bayes framework by borrowing information across genes, has gained widespread adoption in the field of microarray data analysis (Johnson, Li, and Rabinovic 2007). ComBat-seq, which employs a negative binomial regression model to retain the integer nature of sequence count data, is widely used in RNA-seq studies (Zhang, Parmigiani, and Johnson 2020). For variation from unknown sources, SVA is commonly used to generate surrogate variables, which reduce dependence and stabilize error rate estimates in downstream differential expression analysis (Leek et al. 2012).

Despite the success of ComBat, ComBat-seq, and SVA in their respective domains, the direct application of the aforementioned tools to DNA methylation data remains challenging. This is because DNA methylation data consist of methylation percentages, also known as β-values representing the proportion of methylated alleles at specific genomic loci. β-values are constrained within the range of 0 to 1, and the underlying distribution of β-values often deviates from a Gaussian distribution, exhibiting skewness and over-dispersion. To address this issue, β-values can be converted to M-values through a logit transformation prior to batch correction (Tian et al. 2017). However, the effectiveness of this approach has not been systematically investigated, raising questions of whether it is the best method for preserving true biological signals while correcting for batch effects.

Here we introduce ComBat-met, a method tailored specifically for adjusting batch effects in DNA methylation data. Building upon the principles of ComBat and ComBat-seq, our method employs a beta regression framework to account for the specific characteristics of β-values. ComBat-met estimates the parameters of a beta regression model, calculates a batch-free distribution, i.e., the expected distribution as if there were no batch effects in the data, based on the estimated model parameters, and adjusts the data by mapping the quantiles of the estimated distribution to its batch-free counterpart (Fig. 1). Besides the common cross-batch average, we also allow adjustment of β-values to a reference batch, where all batches are adjusted to the mean and dispersion of the reference (Zhang et al. 2018).

In the rest of this paper, we describe the methodology of ComBat-met and demonstrate its performance through comprehensive benchmarking analyses and real-world applications. By utilizing simulated data, we demonstrated that ComBat-met followed by differential methylation analysis achieved superior statistical power compared to traditional approaches while correctly controlling the Type I error rate in nearly all cases. Furthermore, to illustrate its ability to recover biologically meaningful results, we applied ComBat-met to batch-correct the methylation data in breast cancer patients pulled from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas 2012). Our results suggest that ComBat-met provides an effective solution for mitigating batch effects in methylation data, enabling more accurate and reliable downstream analyses.

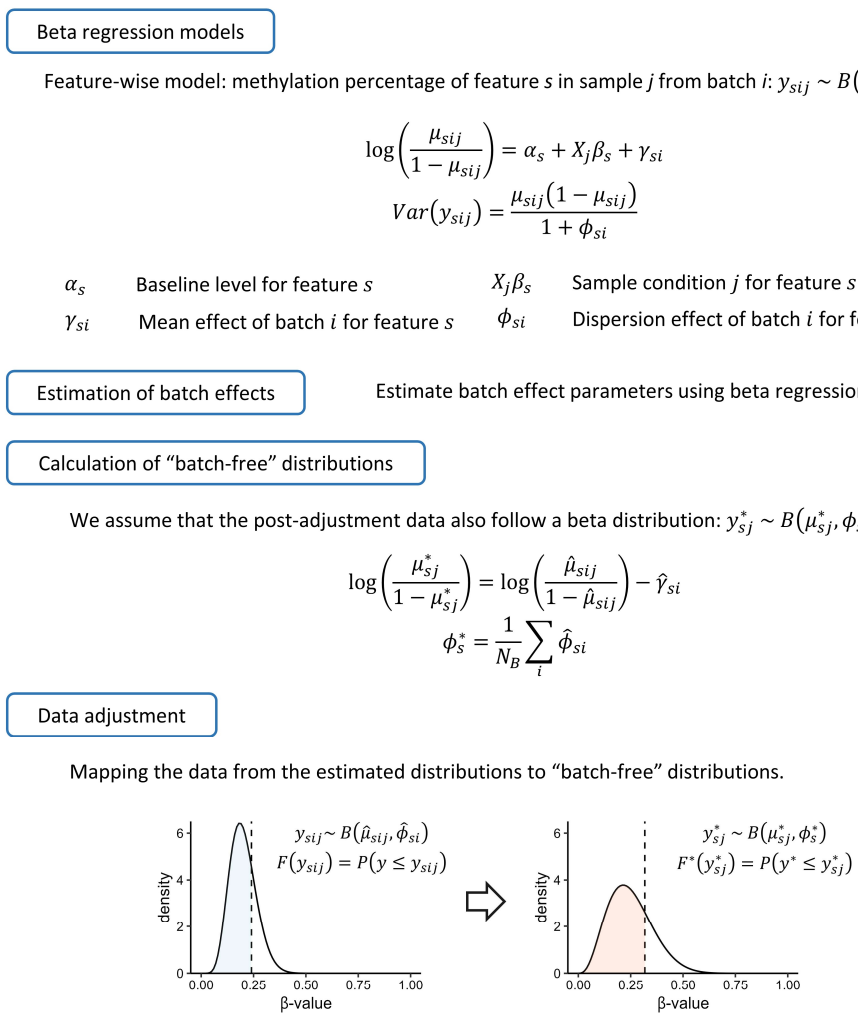

Fig. 1. Diagram of the ComBat-met workflow. Model parameters are estimated using beta regression followed by the calculation of a batch-free distribution. A quantile-matching procedure is subsequently applied to map data from the estimated distribution to its batch-free counterpart.

2 Methods

2.1 Beta regression model

Let y_{sij} denote the β-value of feature s in sample j from batch i. A feature could be a site, a gene, or any molecular entity for which the methylation percentage could be defined. y_{sij} is assumed to follow a beta distribution, where μ_{sij} and ϕ_{si} denote the mean and dispersion parameter of the distribution. The beta regression model is defined as:

$$\log\left(\frac{\mu_{sij}}{1-\mu_{sij}}\right) = \alpha_s + X_j\beta_s + \gamma_{si} \quad (1)$$

$$var(y_{sij}) = \frac{\mu_{sij}(1-\mu_{sij})}{1+\phi_{si}}, \quad (2)$$

where α_s, X_j, and β_s denote the common cross-batch average M-value (i.e., logit transformation of β-values), the covariate matrix, and the corresponding regression coefficients, respectively. γ_{si} represents the batch-associated additive effect on the common cross-batch average M-value. Parameters are estimated via maximum likelihood estimation using the betareg() function in the betareg R package (Cribari-Neto and Zeileis 2010).

In the context of reference-based adjustment, α_s and γ_{si} in Eq. (1) represent the average M-value in the chosen reference batch and the additive difference between the reference batch and batch i, respectively. The rest of the notations remain unchanged.

2.2 Adjustment

To adjust the data, we take a quantile-matching approach similar to that described in ComBat-seq (Fig. 1) (Zhang, Parmigiani, and Johnson 2020). We assume that the adjusted data follow a beta distribution, i.e., $B(\mu^*_{sj}, \phi^*_s)$. Once the regression model is fitted to the data, we calculate the parameters for the batch-free distributions, μ^*_{sj} and ϕ^*_s, as follows:

$$\log\left(\frac{\mu_{sj}^*}{1-\mu_{sj}^*}\right) = \log\left(\frac{\hat{\mu}_{sij}}{1-\hat{\mu}_{sij}}\right) - \hat{\gamma}_{si} \quad (3)$$

$$\phi_s^* = \frac{1}{N_B}\sum_{i=1}^{N_B} \hat{\phi}_{si}\,, \quad (4)$$

where N_B is the number of batches, and $\hat{\mu}_{sij}$, $\hat{\gamma}_{si}$, and $\hat{\phi}_{si}$ represent the maximum likelihood estimates of μ_{sij}, γ_{si}, and ϕ_{si} in Eq. (1) and Eq. (2), respectively. The adjusted data y_{sj}^* are then calculated by matching the quantile of the original data y_{sij} on the estimated distribution $B(\hat{\mu}_{sij},\ \hat{\phi}_{si})$ to its quantile counterpart on the batch-free distribution, i.e., $B(\mu_{sj}^*, \phi_s^*)$. Specifically, each y_{sij} is mapped to its counterpart y_{sj}^* such that the cumulative distribution of y_{sj}^*, i.e., $F^*(y_{sj}^*)$ is closest to that of y_{sij}, i.e., $F(y_{sij})$ (Fig. 1). For reference-based adjustment, μ_{sj}^* and ϕ_s^* are the maximum likelihood estimates of the mean and dispersion parameters in the reference batch, respectively.

2.3 Simulation

The simulation was set up as follows. A total of 1000 features were generated with a balanced design involving two biological conditions and two batches across 20 samples (N = 20). Without loss of generality, we assumed that 100 out of 1000 features were truly differentially methylated, and the methylation percentage of these features was higher under condition 2 than under condition 1 by 10%. Data simulation was implemented using a customized script adapted from the dataSim() function in the methylKit R package, yielding 1000×20 matrices of coverages (i.e., total numbers of bases) and numbers of methylated bases (Akalin et al. 2012). β-values were calculated as the ratios of numbers of methylated bases to coverages for each feature.

All features were affected by batch. Both the mean and dispersion parameter of batch effects were allowed to vary. The methylation percentage in one batch was assumed to be either higher or lower than that in the other batch by 0%, 2%, 5%, or 10%. The dispersion parameter in one batch was set to a baseline value equal to 10, and the dispersion parameter in the other batch was set to be 1-, 2-, 5-, or 10-fold of the baseline value. The remaining simulation parameters were resolved to their default values in dataSim(), yielding data bearing characteristics typical of real-world methylation data (Akalin et al. 2012).

The simulation was repeated 1000 times. The proportion of significant features among those that were and were not truly differentially methylated were counted towards true positive rates (TPRs) and false positive rates (FPRs), respectively. Median TPRs and FPRs over the 1000 repeated simulations were calculated under each parameter setting.

2.4 Other approaches

To validate ComBat-met, we compared its performance with several additional batch correction workflows: the "one-step" approach, M-value ComBat, the SVA approach, and ComBat-biseq. The former three approaches required that β-values be logit-transformed to M-values, which were implicitly assumed to be normally distributed, prior to batch correction. The "one-step" approach included the batch variable directly as a covariate in the differential linear model. The M-value ComBat approach adjusted M-values for batch effects through ComBat (Johnson, Li, and Rabinovic 2007). The SVA approach computed a surrogate variable to be included as a covariate in the downstream differential analysis without reference to the given batch information (Leek et al. 2012).

To accommodate bisulfite sequencing data, we also developed ComBat-biseq, a method which retains the integer nature of count data. Unlike ComBat-met, ComBat-biseq initiates batch correction by fitting beta-binomial regression models to methylated cytosine counts. The remaining steps, however, are akin to ComBat-met (see Supplementary Fig. 1 for details). Beta-binomial regression, as well as the distribution and quantile functions of the beta-binomial distribution were implemented using the aod and updog R packages (Lesnoff and Lancelot 2012; Gerard et al. 2018).

For all M-value-based methods, differential methylation analysis was implemented using the lmFit() function in the limma R package (Ritchie et al. 2015). For data adjusted by ComBat-biseq, a likelihood ratio test was adopted using the diff_methylsig() function in the methylSig R package (Park et al. 2014). To enable a fair comparison, we skipped the empirical Bayes step in limma, as the likelihood ratio test in methylSig employs a frequentist framework (Ritchie et al. 2015; Park et al. 2014).

2.5 Application to real-world experimental data

Level 3 Illumina 450K methylation data (calculated β-values mapped to the genome) for tumor and adjacent normal tissues in breast cancer (BRCA) patients were downloaded from TCGA using the TCGAbiolinks R package (Colaprico et al. 2016; Mounir et al. 2019). To evaluate dispersion correction, we retained only batches containing at least two samples for batch adjustment. Furthermore, we focused exclusively on tumor samples in the luminal B subtype cohort for demonstration and visualization purposes. Our analysis included a total of 11 batches spanning 96 adjacent normal tissue samples and 19 batches spanning 135 tumor tissue samples. Similar to a previous study, tumor and adjacent normal tissue samples were analyzed separately to account for the larger inter-patient heterogeneity in tumor tissues (Supplementary Figs. 2 and 3) (Akulenko, Merl, and Helms 2016).

We conducted data preprocessing following the workflow described in a previous study (Akulenko, Merl, and Helms 2016). First, we removed all sites/probes with missing β-values, as principal component analysis (PCA) requires complete datasets. Next, the remaining probes were mapped to genes and filtered based on their distance to the respective transcription start sites (TSS) (Robinson, McCarthy, and Smyth 2010; Hansen and Aryee 2012). Depending on the strand direction, only probes within 2000 base pairs up- or downstream the respective TSS were retained for further analysis. Subsequently, the β-values of the probes were averaged by the nearest gene, resulting in a total of 4895 and 4639 gene-β-value pairs in normal and tumor tissues, respectively. The resulting β-value matrices were then separately adjusted to the reference batch "A12R", which was present in both tumor and normal samples, using the ComBat-met and M-value ComBat workflows. Differential methylation analysis was performed on the adjusted data using the limma R package (Ritchie et al. 2015). P-values were adjusted via the Benjamini-Hochberg method, and a stringent p-value threshold of 5×10^{-8} was applied due to the large sample size. KEGG enrichment analysis was implemented using the kegga() function in the limma R package (Ritchie et al. 2015).

3 Results

3.1 Using appropriate model assumptions for β-values

To demonstrate the importance of applying appropriate statistical models, we simulated an example of a β-value matrix with a balanced design, consisting of 20 samples (N = 20) pulled from two batches and two conditions

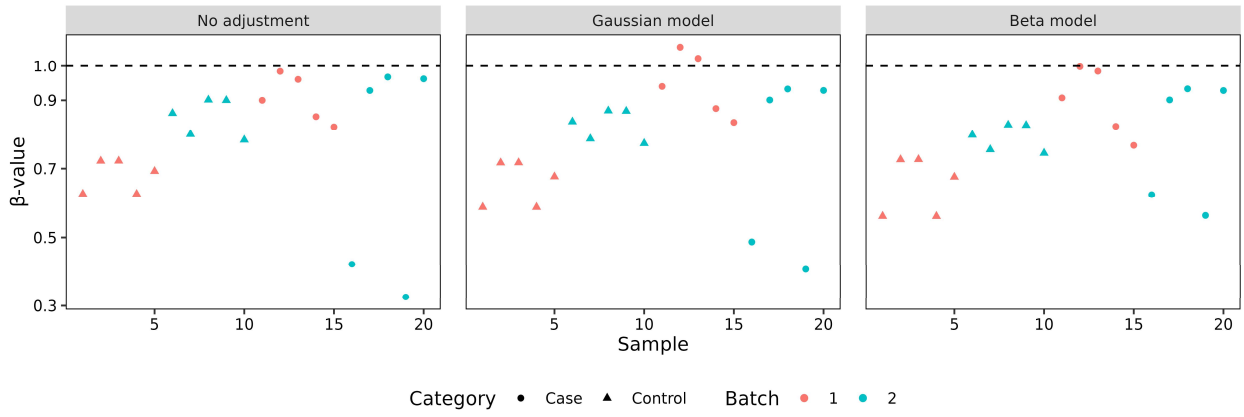

Fig. 2. Example of problems caused by Gaussian model-based batch correction in DNA methylation data. A β-value matrix of 20 samples was simulated with a balanced design involving two batches and two categories. The left, middle, and right panels represent the unadjusted, Gaussian model-adjusted, and beta model-adjusted β-values across 20 samples, respectively. Batches are shown by color, and categories are shown by shape. Dashed lines represent the maximum plausible β-value equal to one.

(Fig. 2). While applying ComBat directly to the β-value matrix reduces cross-batch variations, doing so also causes adjusted β-values to exceed the maximum plausible limit of one in samples #12 and #13, posing challenges for data interpretation and downstream analysis (Fig. 2). This problem arises because ComBat operates based on a Gaussian model (Johnson, Li, and Rabinovic 2007). Similar to count data, β-values exhibit skewness and over-dispersion that cannot be captured by a Gaussian distribution.

Unlike ComBat, ComBat-met assumes that β-values follow a beta distribution, a statistical model much more suitable for the behavior of proportions and percentages. As such, it not only imposes a naturally confined range of (0, 1) on the adjusted β-values but also renders parameter estimation more robust to outliers, yielding the smallest cross-batch variations post-adjustment (Fig. 2).

3.2 ComBat-met demonstrates balanced control of power and false positive rates

We evaluated the FPR of ComBat-met using simulated DNA methylation data. In our simulations, the cross-batch differences in the mean and dispersion parameters of the methylation percentage were allowed to vary. The resulting data were batch-corrected either using ComBat-met, M-value ComBat, ComBat-biseq, or SVA prior to differential methylation analysis, or by directly including the batch variable in the analysis (i.e., "one-step" approach). Details of each method are provided in Section 2.4.

The comparison of different batch correction methods is provided in Fig. 3. ComBat-met, M-value ComBat, and the "one-step" approach demonstrated close control of the FPR around 0.05 in most cases (Fig. 3). Although ComBat-met and M-value ComBat exceeded the threshold of 0.05 for small cross-batch differences in mean and dispersion parameters (e.g., a 2% change in the mean batch percentage and a 2-fold change in the dispersion) (Fig. 3), a slightly inflated FPR could generally be tolerated for discovery studies.

In contrast to the aforementioned methods, ComBat-biseq consistently exhibited severe inflations of FPRs (Fig. 3). This behavior was most likely attributed to the small sample size, leading to breakdowns in the asymptotic argument of the likelihood ratio test downstream ComBat-biseq. Additional simulations supported this hypothesis, as the same workflow maintained much more precise control of the FPR for a larger sample size (N = 100) (Supplementary Fig. 4). On the other hand, SVA showed unstable performance across different parameter settings (Fig. 3). This variability was expected, considering that SVA, unlike ComBat-derived methods, is not optimized for scenarios where technical variations originate from identifiable sources.

The TPRs of these methods were evaluated based on the 10% features that were truly differentially methylated. All batch correction methods yielded higher TPRs than no adjustment, especially notable for a cross-batch difference of 10% in the mean methylation percentage (Fig. 3). This improvement stems from the general belief that failing to account for batch effects can artificially inflate variance estimates within each biological condition, thereby reducing sensitivity.

Among different methods, ComBat-met and M-value ComBat achieved similar TPRs under circumstances with no cross-batch difference or a 2-fold difference in the dispersion parameter (Fig. 3). However, for scenarios involving a 5- or 10-fold change in the dispersion parameter, ComBat-met demonstrated a markedly increased power to detect methylation changes compared to M-value ComBat, without compromising the FPRs (Fig. 3). This highlights the superior performance of ComBat-met, as large dispersion differences typically occur when combining data from heterogeneous batches, such as those from different profiling platforms or studies. While ComBat-biseq achieved even higher TPRs than ComBat-met, its limited ability to control FPRs significantly restricts its suitability for small sample sizes (Fig. 3).

3.3 ComBat-met shows effective adjustment of batch effects

To demonstrate the ability of ComBat-met to recover biological insights, we applied our method to batch-correct the TCGA dataset described in Subsection 2.5. In adjacent normal tissues, samples from the same batches were clustered together in the unadjusted data (Fig. 4). Both M-value ComBat and ComBat-met effectively mitigated batch effects, as evidenced by the absence of batch-related sample separation in the adjusted data (Fig. 4). Analysis of explained variation further suggested that batch-related variance was markedly reduced in the data adjusted by ComBat-met compared to M-value ComBat (Fig. 4). While the tumor samples were

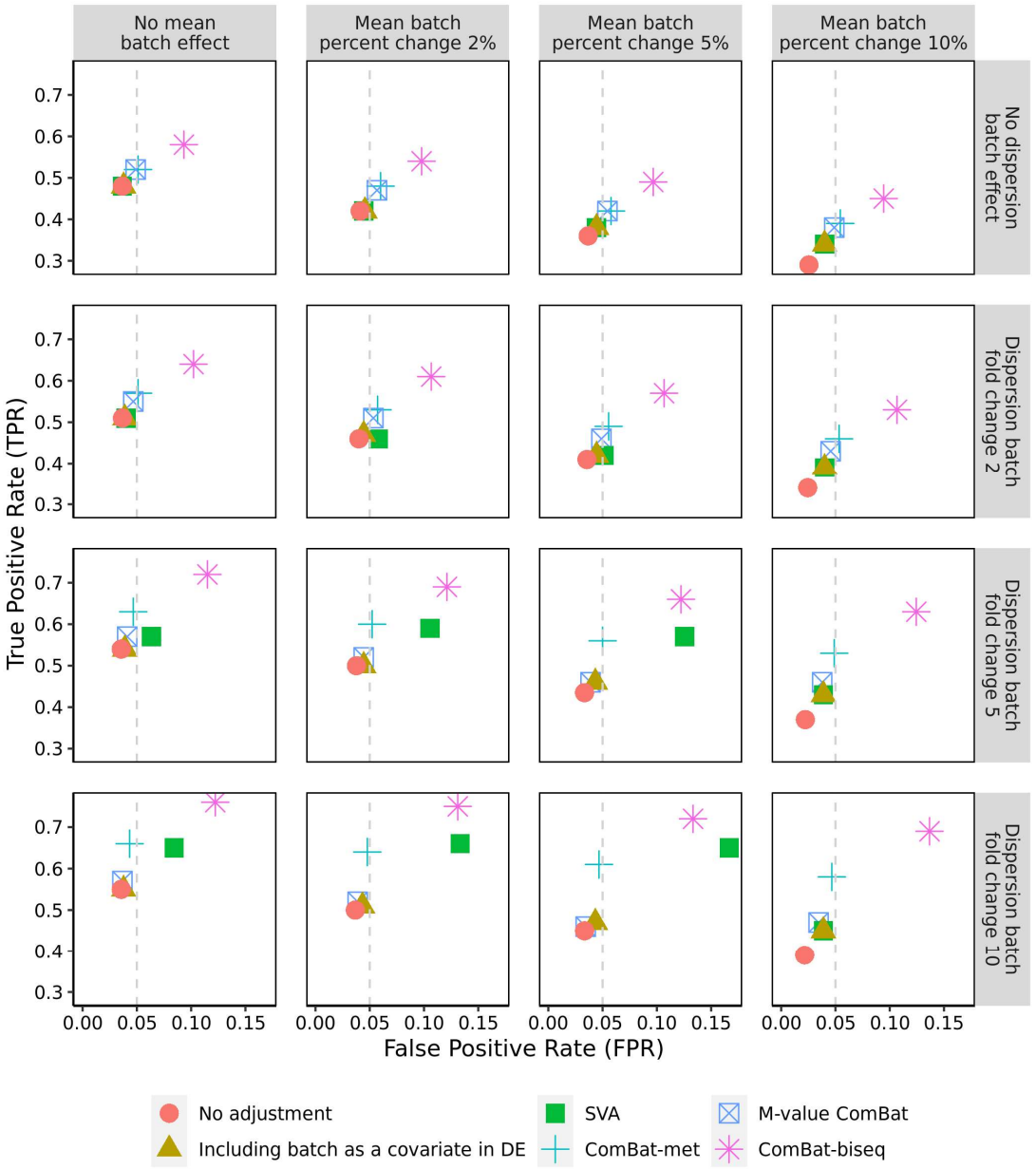

Fig. 3. Median true positive rates and false positive rates calculated based on simulation data. The mean cross-batch difference in methylation percentage was set to 0%, 2%, 5%, or 10%. The dispersion parameters of the batch effect were set to have a 1-, 2-, 5-, or 10-fold change. The simulation was repeated 1000 times. Methods are shown by color. Detailed information about each method is provided in Section 2.4.

seemingly less batch-affected due to larger inter-patient heterogeneity (Supplementary Figs. 2 and 3), quantitative analysis also revealed the greatest reduction of batch-explained variance in the data adjusted by ComBat-met, similar to adjacent normal samples (Supplementary Fig. 5, Fig. 4). Interestingly, applying ComBat-met with parameter shrinkage led to incomplete reduction of batch effects (Supplementary Fig. 6; see Supplementary Methods for details of parameter shrinkage). To corroborate our findings, we repeated the same analysis on probe-level β-values. Probe-level analysis revealed similar patterns, lending additional credibility to the efficacy of ComBat-met (Supplementary Fig. 7).

To identify methylation signatures specific to the luminal B cancer cohort, we compared the methylomes of tumor and adjacent normal tissues. Analysis of the unadjusted data, data adjusted by M-value ComBat, and data adjusted by ComBat-met revealed 2261, 2638, and 2847 differentially methylated genes, respectively. KEGG enrichment analysis further suggested that the list of significant genes only identified in the ComBat-met-adjusted data were overrepresented in glycosaminoglycan biosynthesis (KEGG ID: hsa00533), transcriptional misregulation in cancer (KEGG ID: hsa05202), and small cell lung cancer (KEGG ID: hsa05222). Interestingly, glycosaminoglycans have been increasingly appreciated for their

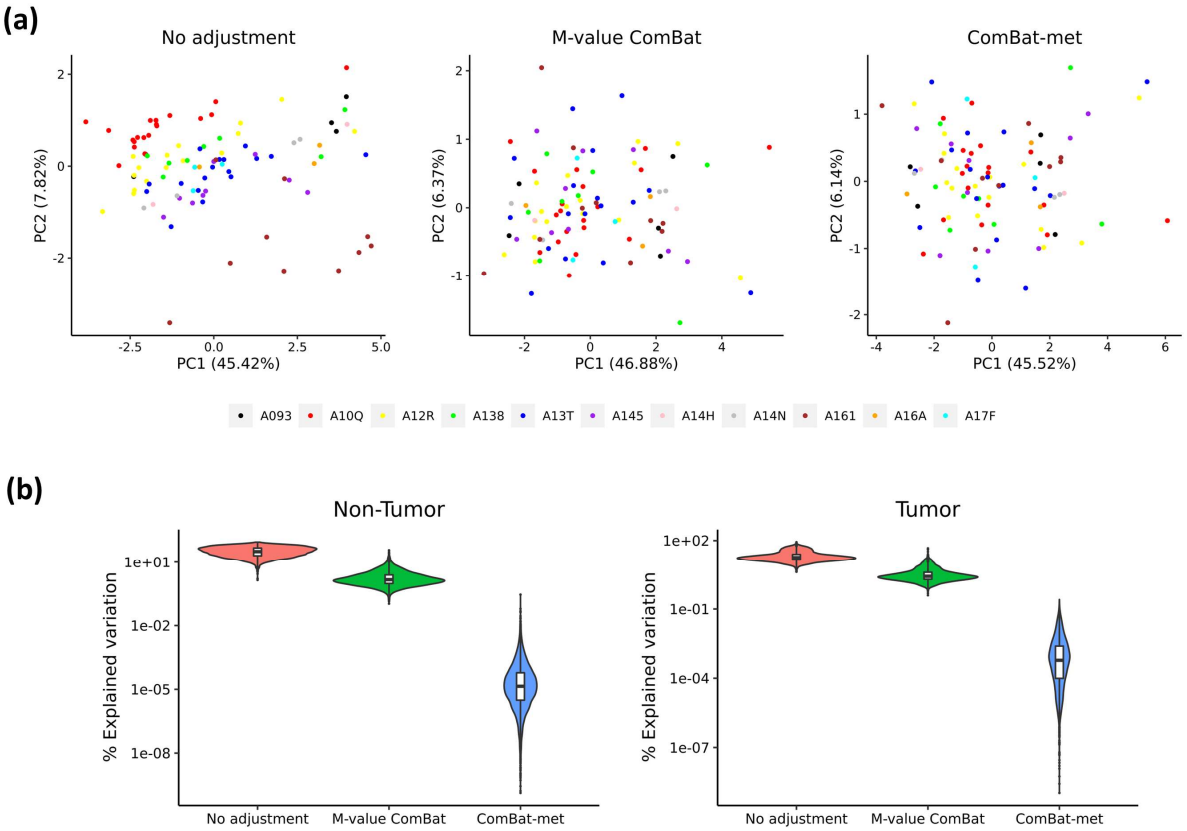

Fig. 4. Application of ComBat-met to remove batch effects in the TCGA data. (a) PCA plots illustrating the separation of adjacent normal tissue samples in the unadjusted gene-level data, gene-level data adjusted by M-value ComBat, and gene-level data adjusted by ComBat-met. Batches are shown by color. (b) Percent of variation explained by batch in the unadjusted gene-level data, gene-level data adjusted by M-value ComBat, and gene-level data adjusted by ComBat-met in tumor and adjacent normal samples.

roles in the proliferation and metastasis of cancer cells (Wei et al. 2020). These results demonstrated the ability of ComBat-met to enhance statistical powers through batch correction, thereby increasing the amount of biological knowledge that could be obtained from the data.

4 Discussion

In this work, we have developed ComBat-met, which adjusts batch effects in DNA methylation data by employing a beta regression framework. Our work not only underscores the importance of batch correction in methylation studies but also highlights the superior performance of ComBat-met through simulations and real-world applications. Compared to the widely adopted M-value ComBat approach, ComBat-met demonstrates more effective removal of batch effects and markedly improved sensitivity to detect methylation changes for scenarios involving large cross-batch dispersion differences, without sacrificing specificity.

ComBat-met is an extension of ComBat and ComBat-seq (Johnson, Li, and Rabinovic 2007; Zhang, Parmigiani, and Johnson 2020). Despite the differences in their probabilistic assumptions, all these methods adopt a regression framework to model the means and dispersion parameters of batch effects followed by an adjustment step, which is in essence, a quantile-mapping procedure. Similar to ComBat, ComBat-met also provides users the option to adjust β-values to not only the common cross-batch average but also any existing batch as the reference (Zhang et al. 2018). In large consortium projects, datasets are typically generated sequentially (Subramanian et al. 2017; Wang and Novick 2023), so batch correction would always need to be reapplied in the event of incoming new datasets if the correction method depends on all datasets in hand. Our reference-based strategy allows new data to be adjusted without impacting the previous results, facilitating data integration across sequential batches.

While ComBat-met brings major improvement to batch correction for methylation studies, we acknowledge several limitations of our approach. First, ComBat-met runs more slowly than M-value ComBat. Unlike linear regression, beta regression lacks a closed-form solution and therefore requires iterative methods for maximum likelihood estimation. Parallel computing may be employed to expedite the regression step in ComBat-met. Second, ComBat-met is designed specifically to address batch effects stemming from identifiable sources. In the presence of unknown sources of variation, SVA or Removal of Unwanted Variation (RUV) should be applied instead (Leek et al. 2012; Gagnon-Bartsch and Speed 2012).

Additionally, ComBat employs an empirical Bayes framework to shrink batch effect parameters (Johnson, Li, and Rabinovic 2007). This approach borrows information across features, making the adjustment robust for data with small sample sizes or outliers. Inspired by the non-parametric empirical Bayes method in ComBat (Johnson, Li, and Rabinovic 2007), we provide a similar option in ComBat-met, enabling users the

freedom to shrink parameters at will. However, in the context of beta regression, parameter shrinkage has been shown to cause under-correction of batch effects (Supplementary Fig. 6). Our observations are consistent with ComBat-seq, which also under-corrects when coupled to parameter shrinkage. Similar to the negative binomial distribution, beta distributions can accommodate a wide range of skewness, conferring more robustness to outliers and making parameter shrinkage unnecessary (Zhang, Parmigiani, and Johnson 2020). Furthermore, it is noteworthy that non-parametric empirical Bayes, which requires extensive Monte-Carlo sampling, becomes extremely time-consuming for large DNA methylomes even if only a subset of features are sampled. While we discourage the use of parameter shrinkage in ComBat-met, it will be interesting to further explore the impact of outlying β-values and refine our method to better accommodate outliers in DNA methylation data.

In conclusion, ComBat-met emerges as a powerful batch correction method for capturing the unique characteristics of DNA methylation data. Our work demonstrates the great potential of this approach in facilitating data integration and accelerating biological discoveries.

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Code availability

ComBat-met, M-value ComBat, and ComBat-biseq, along with the scripts used to generate the simulated data and to analyse the simulated data and the data from TCGA, are available at: https://github.com/JmWangBio/ComBatMet.

Conflict of Interest: none declared.

References

Akalin, A., M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, and C. E. Mason. 2012. 'methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles', *Genome Biol*, 13: R87.

Akulenko, R., M. Merl, and V. Helms. 2016. 'BEclear: Batch Effect Detection and Adjustment in DNA Methylation Data', *PLoS One*, 11: e0159921.

Cancer Genome Atlas, Network. 2012. 'Comprehensive molecular portraits of human breast tumours', *Nature*, 490: 61-70.

Colaprico, A., T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr. 2016. 'TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data', *Nucleic Acids Res*, 44: e71.

Cribari-Neto, F., and A. Zeileis. 2010. 'Beta regression in R', *J Stat Softw*, 34: 1-24.

Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. 1992. 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands', *Proc Natl Acad Sci U S A*, 89: 1827-31.

Gagnon-Bartsch, J. A., and T. P. Speed. 2012. 'Using control genes to correct for unwanted variation in microarray data', *Biostatistics*, 13: 539-52.

Gerard, D., L. F. V. Ferrao, A. A. F. Garcia, and M. Stephens. 2018. 'Genotyping Polyploids from Messy Sequencing Data', *Genetics*, 210: 789-807.

Hansen, K. D., and M. Aryee. 2012. "IlluminaHumanMethylation450kmanifest: Annotation for Illumina's 450k methylation arrays." In, R package.

Johnson, W. E., C. Li, and A. Rabinovic. 2007. 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8: 118-27.

Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. 2012. 'The sva package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics*, 28: 882-3.

Lesnoff, M., and R. Lancelot. 2012. "aod: Analysis of Overdispersed Data." In, R package.

Mounir, M., M. Lucchetta, T. C. Silva, C. Olsen, G. Bontempi, X. Chen, H. Noushmehr, A. Colaprico, and E. Papaleo. 2019. 'New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx', *PLoS Comput Biol*, 15: e1006701.

Olova, N., F. Krueger, S. Andrews, D. Oxley, R. V. Berrens, M. R. Branco, and W. Reik. 2018. 'Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data', *Genome Biol*, 19: 33.

Park, Y., M. E. Figueroa, L. S. Rozek, and M. A. Sartor. 2014. 'MethylSig: a whole genome DNA methylation analysis pipeline', *Bioinformatics*, 30: 2414-22.

Price, E. M., and W. P. Robinson. 2018. 'Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned', *Front Genet*, 9: 83.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. 2015. 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Res*, 43: e47.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Ross, J. P., S. van Dijk, M. Phang, M. R. Skilton, P. L. Molloy, and Y. Oytam. 2022. 'Batch-effect detection, correction and characterisation in Illumina HumanMethylation450 and MethylationEPIC BeadChip array data', *Clin Epigenetics*, 14: 58.

Subramanian, A., R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W. N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, and T. R. Golub. 2017. 'A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles', *Cell*, 171: 1437-52 e17.

Tian, Y., T. J. Morris, A. P. Webster, Z. Yang, S. Beck, A. Feber, and A. E. Teschendorff. 2017. 'ChAMP: updated methylation analysis pipeline for Illumina BeadChips', *Bioinformatics*, 33: 3982-84.

Wang, J., and S. Novick. 2023. 'DOSE-L1000: unveiling the intricate landscape of compound-induced transcriptional changes', *Bioinformatics*, 39.

———. 2024. 'Peptide set test: a peptide-centric strategy to infer differentially expressed proteins', *Bioinformatics*, 40.

Wei, J., M. Hu, K. Huang, S. Lin, and H. Du. 2020. 'Roles of Proteoglycans and Glycosaminoglycans in Cancer Development and Progression', *Int J Mol Sci*, 21.

Zhang, Y., D. F. Jenkins, S. Manimaran, and W. E. Johnson. 2018. 'Alternative empirical Bayes models for adjusting for batch effects in genomic studies', *BMC Bioinformatics*, 19: 262.

Zhang, Y., G. Parmigiani, and W. E. Johnson. 2020. 'ComBat-seq: batch effect adjustment for RNA-seq count data', *NAR Genom Bioinform*, 2: lqaa078.