Computational and Structural Biotechnology Journal 27 (2025) 58–64

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Highly effective batch effect correction method for RNA-seq count data

Xiaoyu Zhang

Department of Computer Science and Information Science, California State University San Marcos, 333 S. Twin Oaks Valley Rd, San Marcos, CA 92096, USA

ARTICLE INFO

Keywords:
Batch effect correction
RNA-seq data
Negative binomial distribution
Minimum dispersion
Generalized linear model

ABSTRACT

RNA sequencing (RNA-seq) has become a cornerstone of transcriptomics, providing detailed insights into gene expression across diverse biological conditions and sample types. However, RNA-seq data are often confounded by batch effects, systematic non-biological variations that compromise data reliability and obscure true biological differences. To address these challenges, we introduce ComBat-ref, a refined batch effect correction method designed to enhance the statistical power and reliability of differential expression analysis in RNA-seq data. Building on the principles of ComBat-seq, ComBat-ref employs a negative binomial model for count data adjustment but innovates by selecting a reference batch with the smallest dispersion, preserving count data for the reference batch, and adjusting other batches towards the reference batch. Our method demonstrated superior performance in both simulated environments and real-world datasets, including the growth factor receptor network (GFRN) data and NASA GeneLab transcriptomic datasets, significantly improving sensitivity and specificity compared to existing methods. By effectively mitigating batch effects while maintaining high detection power, ComBat-ref provides a robust solution for improving the accuracy and interpretability of RNA-seq data analyses.

1. Introduction

RNA sequencing (RNA-seq) has emerged as a cornerstone technology in transcriptomics, providing unparallel insights into gene expression profiles across various biological conditions and sample types. However, the reliability of RNA-seq data is often undermined by batch effects—systematic non-biological differences that arise during sample processing and sequencing across different batches. These batch effects can be on a similar scale or even larger than biological differences of interest, significantly reducing the statistical power to detect differentially expressed (DE) genes.

The presence of batch effects in RNA-seq data is a well-recognized challenge, prompting the development of various strategies to mitigate their impact. One widely used method is ComBat [1], which employs an empirical Bayes framework to correct for both additive and multiplicative batch effects. Methods such as SVASeq [2] and RUVSeq [3] model batch effects from unknown sources, while popular DE analysis packages such as edgeR [4] and DESeq2 [5] allow the inclusion of batch as a covariate in linear models to account for these effects. ComBat-seq [6], which extends ComBat by using a generalized linear model (GLM) with a negative binomial distribution, retains the integer count data and has demonstrated better statistical power than its predecessors. More recently, machine learning methods [7–9] have been proposed to address batch effects by modeling discrepancies among batches. NPMatch [10], for instance, uses a nearest-neighbor matching-based method to adjust for batch effects.

Among these methods, ComBat-seq offers the advantage of preserving the integer count matrix in adjusted data, making it particularly suitable for downstream DE analysis using tools like edgeR and DESeq2. Additionally, it achieves higher statistical power in detecting DE genes compared to other methods, especially when batches with different dispersion parameters are pooled. This improved performance is largely attributed to its accurate modeling of count data using negative binomial (gamma-Poisson) distributions. Despite these advantages, ComBat-seq still exhibits significantly lower power in DE analysis compared to batch-free data, particularly when using the false discovery rate (FDR; adjusted p-value) for statistical testing, as recommended by edgeR and DESeq2.

In this paper, we introduce ComBat-ref, a refined batch effect adjustment method that builds on ComBat-seq while incorporating key improvements. ComBat-ref models RNA-seq count data using a negative binomial distribution but innovates by estimating a pooled (shrunk) dispersion parameter for each batch and selecting the batch with the lowest dispersion as the reference. The count data of all other batches are then adjusted to align with this reference batch. We demonstrate that ComBat-ref retains exceptionally high statistical power—comparable to

E-mail address: xiaoyu@csusm.edu.

https://doi.org/10.1016/j.csbj.2024.12.010
Received 11 October 2024; Received in revised form 13 December 2024; Accepted 14 December 2024
Available online 16 December 2024

data without batch effects—even when there is significant variance in batch dispersions. Furthermore, ComBat-ref outperforms other methods when FDR is used for DE analysis, making it a robust tool for addressing batch effects in RNA-seq data.

2. Materials and methods

Similar to ComBat-seq [6], we model RNA-seq count data using a negative binomial distribution, with each batch potentially having different dispersions. Consider a gene g in batch i and sample j. Let c_j represent the biological condition of sample j and n_{ij}^g denote the measured count. The count n_{ij}^g is modeled as follows:

$$n_{ij}^g \sim \mathrm{NB}\left(\mu_{ij}^g, \lambda_i^g\right)$$

where μ_{ij}^g is the expected expression level of gene g in sample j and batch i, and λ_i^g is the dispersion parameter for batch i. ComBat-seq estimates λ_i^g for each gene and batch and computes an average dispersion per gene for data adjustment:

$$\overline{\lambda}^g = \frac{1}{N_{\mathrm{batch}}} \sum_i \lambda_i^g$$

However, since the number of samples per batch is typically small, the estimation of λ_i^g can be imprecise, resulting in high variance for $\overline{\lambda}^g$ and reduced statistical power for adjusted data. In the new ComBat-ref method, we pool the gene count data within each batch and estimate a batch specific dispersion λ_i. The batch with the smallest dispersion is selected as the reference batch. Without loss of generality, we assume batch 1 is the reference batch. Unlike [11], which discussed an alternative approach of ComBat [1] using a reference batch, ComBat-ref specifically employs a negative binomial model and selects the reference based on dispersion.

To model the expected gene expression level μ_{ij}^g, we apply a generalized linear model (GLM):

$$\log\left(\mu_{ij}^g\right) = \alpha^g + \gamma_i^g + \beta_{c_j}^g + \log\left(N_j\right)$$

Here, c_j is the biological condition, N_j is the library size of sample j, α^g represents the global "background" expression of gene g, γ_i^g represents the effect of batch i, and $\beta_{c_j}^g$ denotes the effects of the biological condition c_j on the logarithm of gene g's expression level. These model parameters can be estimated using the GLM fit method implemented in edgeR [4], [12], or more computationally intensive MCMC based methods [13]. Since the reference batch has the smallest dispersion, retaining its count data for downstream DE analysis improves statistical power. Thus, in the new ComBat-ref method, RNA-seq count data from other batches are adjusted to align with the reference batch.

2.1. ComBat-ref adjustment

Assuming that reference batch 1 has the smallest dispersion λ_1, the adjusted gene expression level $\widetilde{\mu}_{ij}^g$ for batch $i \neq 1$ and sample j is computed as

$$\log\widetilde{\mu}_{ij}^g = \log\mu_{ij}^g + \gamma_1^g - \gamma_i^g$$

The adjusted dispersion is then set as $\widetilde{\lambda}_i = \lambda_1$. Following the approach of ComBat-seq, the adjusted count $\widetilde{n}_{ij}^g$ is calculated by matching the cumulative distribution function (CDF) of $\mathrm{NB}\left(\mu_{ij}^g, \lambda_i\right)$ at n_{ij}^g and CDF of adjusted distribution $\mathrm{NB}\left(\widetilde{\mu}_{ij}^g, \widetilde{\lambda}_i\right)$ at $\widetilde{n}_{ij}^g$. Care is taken to ensure that the adjusted count does not become infinity when the CDF equals 1, and zero counts are always mapped to zeros.

Setting the adjusted dispersion to λ_1 enhances statistical power in subsequent analyses of the adjusted data, albeit with a potential increase in false positives. This trade-off is often acceptable when pooling samples from multiple batches for DE analysis. In both simulated and real datasets, ComBat-ref demonstrated high sensitivity and controlled FPR, particularly when adjusted p-values (FDR) were used with edgeR or DESeq2.

2.2. Simulations

To evaluate the performance of the new ComBat-ref method and compare it to other batch correction (BC) methods, we followed a procedure similar to that described in [6] to generate realistic simulations of RNA-seq count data. The count data were modeled using a negative binomial (gamma Poisson) distribution, assuming that batch effects could influence both the mean expression of genes and the dispersion of the count distributions. The simulation included two biological conditions and two batches, with three samples for each combination of condition and batch, resulting in a total of 12 samples per experiment.

The count data comprised 500 genes, with 50 up-regulated and 50 down-regulated exhibiting a mean fold change of 2.4. Batch effects were simulated to alter gene expression levels in one random batch by a mean factor (mean_FC), and to increase the dispersion in batch 2 relative to batch 1 by a dispersion factor (disp_FC). Larger values of mean_FC and disp_FC represented stronger batch effects, whereas mean_FC = 1 and disp_FC = 1 indicated no batch effect.

We simulated 16 experiments with varying batch effects, using four levels of mean_FC (1, 1.5, 2, 2.4) and disp_FC (1, 2, 3, 4). The count matrices were generated with the polyester R package[14]. Each experiment was repeated ten times to calculate the average statistics for each BC method. Fig. 1 presents the experimental results, where batch effects increase progressively from left to right and top to bottom in the grid. The ComBat-ref method introduced in this paper outperformed other methods, even in the most challenging scenario represented in the lower-right corner.

3. Results

3.1. Simulations

Simulation data were used to compare the performance of the ComBat-ref method with other popular batch correction (BC) methods for detecting differentially expressed (DE) genes. For comparison, we chose the ComBat-seq method, other methods discussed in [6], and the more recent NPMatch method [10].

First, we analyzed the results using the same threshold of an unadjusted p-value = 0.05 used in the original ComBat-seq paper [6]. Supplemental Figure 1 presents the true positive rates (TPR) and false positive rates (FPR) of DE analysis for each method across all experiments, using the edgeR package[4]. The simulation results for previous methods were consistent with those reported in [6], with the addition of new results for ComBat-ref and NPMatch. We observed that other DE analysis packages, such as DESeq2 [5] produced very similar results; therefore, we only present results from edgeR in this paper.

When there was no change in batch distribution dispersions (disp_FC = 1), i.e., in the first column of Supplemental Figure 1, ComBat-seq and other previous BC methods performed well, achieving high TPR. However, ComBat-ref exhibited a slightly lower FPR compared to ComBat-seq. The NPMatch method also achieved good TPR but consistently had an FPR > 20 % in all experiments, even in cases where no batch effect was present (upper-left corner). This high FPR might be attributed to deficiencies in the NPMatch algorithm, which relies on nearest-neighbor samples for adjustments. As batch dispersion (disp_FC) increased, the new ComBat-ref demonstrated significantly higher sensitivity than all other methods, including ComBat-seq and NPMatch. In the more challenging scenarios of higher disp_FC and mean_FC,

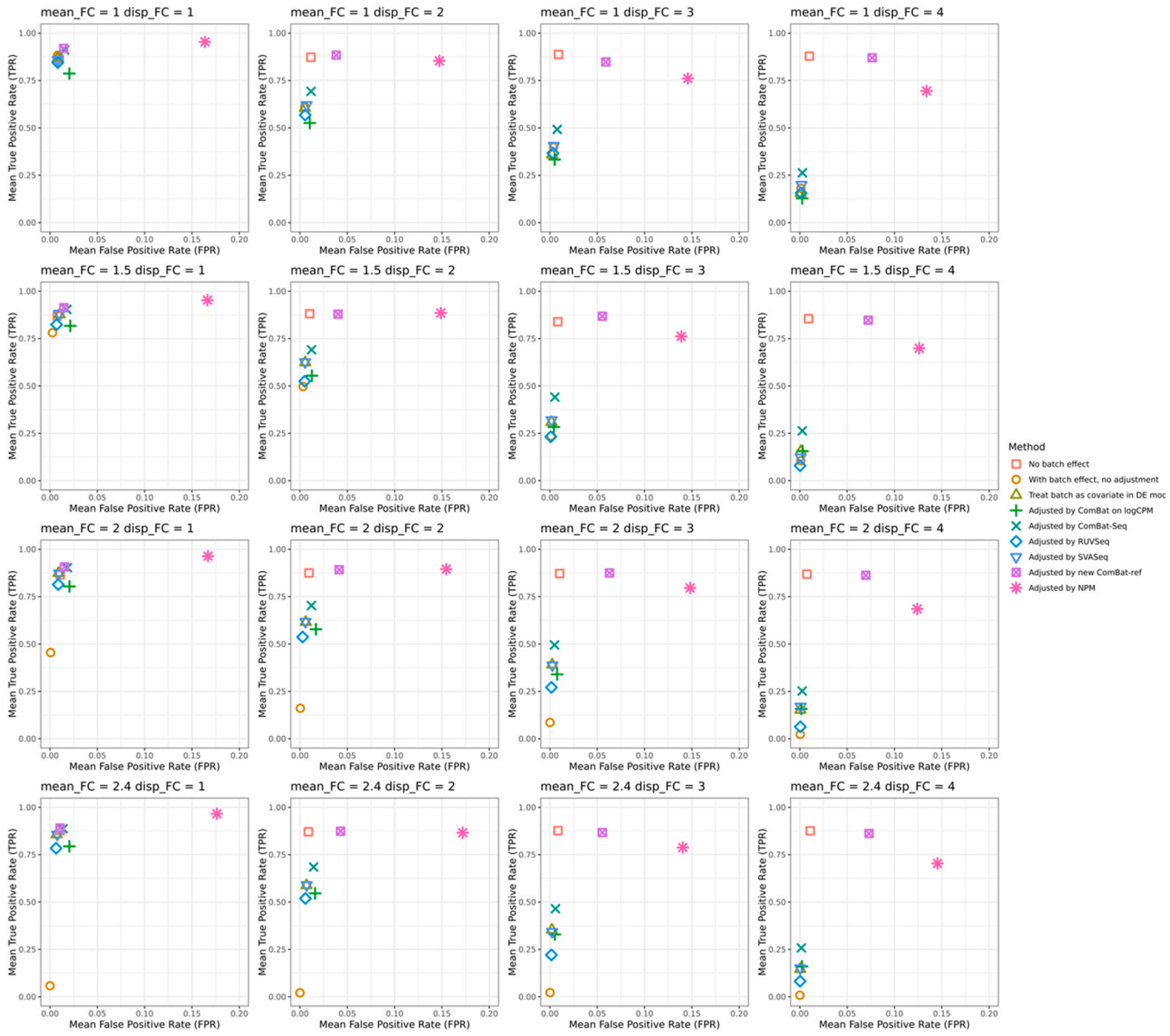

Fig. 1. Simulation results for different levels of mean expression fold change (mean_FC = 1, 1.5, 2, 2.4) and dispersion fold change(disp_FC = 1, 2, 3,4) between batches. The DE genes for each method were identified using edgeR with an FDR ≤ 0.1. As mean_FC and disp_FC increase, all other methods except for ComBat-ref lost power significantly. ComBat-ref consistently maintained high power for DE analysis, comparable to data without batch effects, while controlling the false positive rate (FPR), even in the most challenging scenario of mean_FC = 2.4 and disp_FC = 4.

ComBat-ref maintained a TPR comparable to that seen in cases without batch effects, creating a larger performance gap relative to other methods. Although the improved power of ComBat-ref came with a slightly higher FPR, it was still much lower than that of NPMatch. In DE analysis of samples affected by batch effects, the trade-off of higher power with a potentially higher FPR is often preferred.

When DE analysis was performed using FDR and adjusted p-values, the advantage of ComBat-ref over other methods became even more evident. The original ComBat-seq paper [6] did not include this analysis, likely due to the low power of previous methods in challenging scenarios. Fig. 1 illustrates the results for adjusted p-values and FDR ≤ 0.1, using the edgeR package. Previous methods, including ComBat-seq, experienced a dramatic loss in power as disp_FC and mean_FC increased, with discovery rates approaching zero in the most challenging scenarios. In contrast, ComBat-ref maintained TPR levels comparable to cases with no batch effects, even under the most challenging conditions (mean_FC = 2.4 and disp_FC = 4). Not only did ComBat-ref preserve its near-ideal sensitivity, but it also exhibited much lower FPR when using adjusted p-values. While NPMatch showed higher TPR than previous methods, its performance was inferior to ComBat-ref and exhibited a much higher FPR.

Based on these simulation results, we recommend using ComBat-ref for adjusting batch effects in RNA-seq count data and performing DE analysis with adjusted p-values.

3.2. *Applications to real data*

We first tested the ComBat-ref method on the growth factor receptor network (GFRN) dataset [15], previously analyzed in the original ComBat-seq paper [6]. This dataset contains three batches, each introducing a specific GFRN oncogene to activate downstream pathway signals, with green fluorescent protein (GFP) controls present in all batches. Batch 1 includes 17 samples (five overexpressing HER2 and 12 GFP controls), batch 2 consists of 12 samples (six overexpressing EGFR

and six controls), and batch 3 comprises 18 samples (nine over-expressing KRAS and nine controls). The unadjusted data, shown in the first row of Fig. 2, reveal strong batch effects, with control samples and treated cells grouped by batch. A successful batch adjustment should align control samples across batches while separating treated cells from controls and from each other.

We applied the new ComBat-ref method to the GFRN data and compared its results with those from RUV-Seq, ComBat-seq, and NPMatch. Both ComBat-seq and ComBat-ref effectively adjusted for batch effects, as evidenced by control groups clustering together and treatment groups separating in the PCA plot of Fig. 2. A boxplot displaying the proportion of variation explained by condition and batch further demonstrated the efficacy of the ComBat-ref method. Compared to ComBat-seq, ComBat-ref achieved slightly higher variation explained by condition, while maintaining a similar reduction in variation explained by batch.

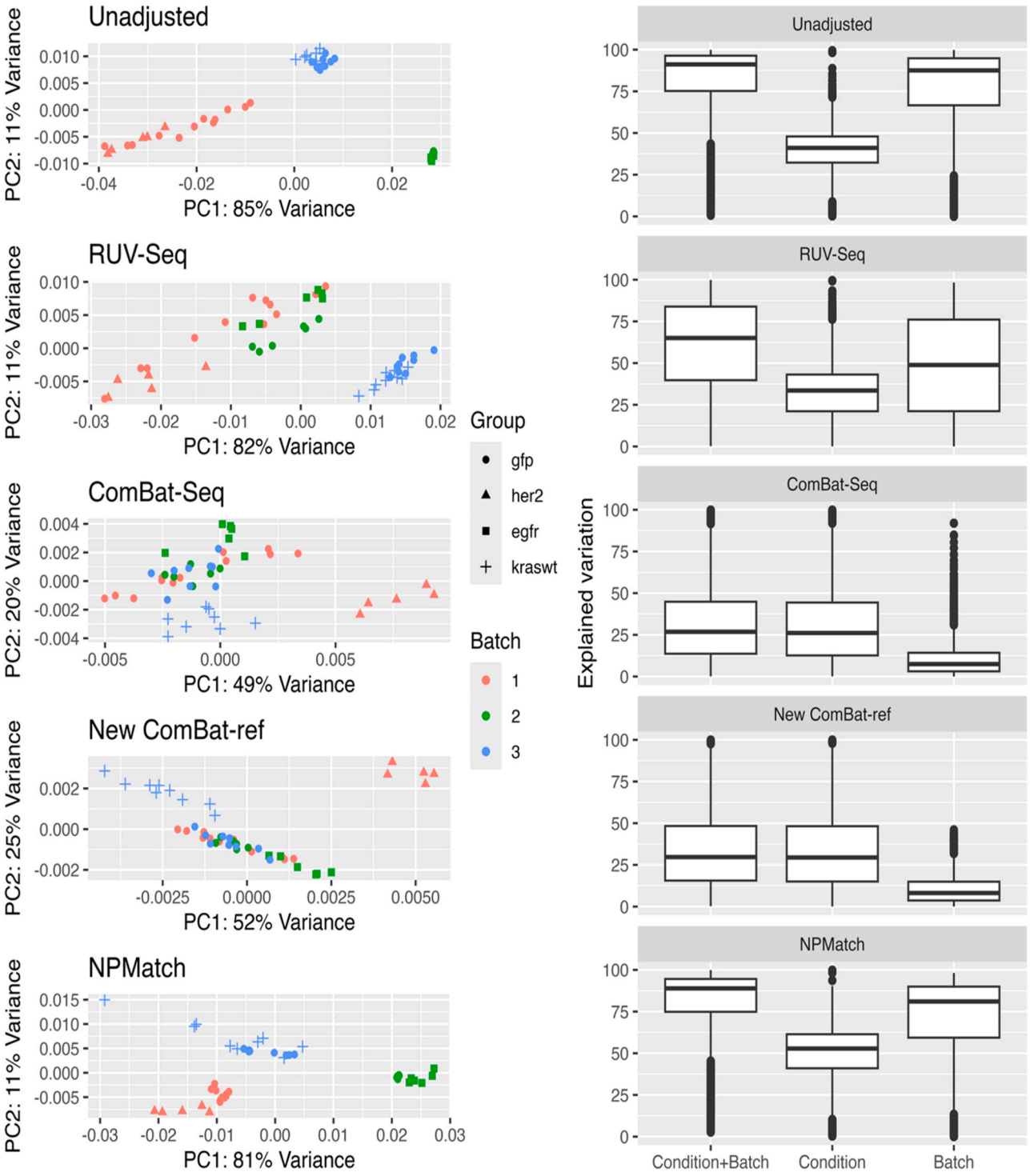

Fig. 2. PCA plots of the GFRN dataset for unadjusted data and data adjusted using RUV-Seq, ComBat-seq, ComBat-ref, and NPMatch methods. The unadjusted data exhibits strong batch effects with all three batches well separated. Both ComBat-seq and ComBat-ref effectively aligned the control (gfp) while preserving biological signals from the treated samples. ComBat-ref outperformed ComBat-seq, as verified by the quantitative measures detailed in the text, and illustrated in the plots of explained variation by condition and batch.

To further evaluate clustering by group in the PCA plot, we computed quantitative measures using the "cluster.stats" function from the fpc package [16]. The unadjusted data showed poor clustering scores by biological condition, with a Gamma score of 0.19 (correlation between distances and a 0–1 vector indicating different clusters) and a Dunn index of 0.78 (ratio oy the minimum inter-cluster dissimilarity to the maximum intra-cluster dissimilarity). After adjustment with ComBat-seq, the clustering scores improved significantly (Gamma = 0.53, Dunn = 1.34). The new ComBat-ref method achieved even higher scores (Gamma = 0.66, Dunn = 1.60), indicating superior batch correction compared to ComBat-seq. In contrast, NPMatch performed poorly for correcting batch effects in this example.

We then performed differential expression (DE) analysis to identify DE genes between treated samples and GFP controls pooled across the three batches, using ComBat-ref, and other approaches discussed in [6]. ComBat-ref demonstrated greater statistical power in DE analysis, identifying more DE genes than ComBat-seq across all three conditions with an FDR ≤ 0.05 (Supplemental Table 1).

Furthermore, ComBat-ref extracted more biologically meaningful DE genes through batch correction. HER2, KRAS and EGFR were expected to be over-expressed in their respective samples. As with other methods, ComBat-ref successfully ranked HER2 and KRAS at the top of the DE gene list. However, detecting differential expression of EGFR was more challenging due to batch effects. EGFR was not identified as DE in the unadjusted dataset and several other adjusted datasets with an FDR ≤ 0.05. ComBat-seq, as demonstrated in [6], successfully detected EGFR as DE (FDR = 0.0016), and ComBat also succeeded in detecting EGFR (FDR = 0.0003). The new ComBat-ref method further improved upon these results, identifying EGFR as DE with an FDR ofy0.0001. Supplemental Figure 2 illustrates how ComBat-ref adjustment led to a more distinct separation of EGFR expression levels in treated and control samples, explaining the statistically significant DE results.

We also evaluated DE genes in the RAS signaling pathway for the KRAS samples, as performed in [6]. Effective batch adjustment is expected to prioritize RAS pathway genes among the top DE genes. Among the top 1000 DE genes, ComBat-ref identified 30 genes in the RAS signaling pathway (Fisher's exact test for enrichment, P = 0.0004), outperforming ComBat (19 genes, P = 0.2) and ComBat-seq (24 genes, P = 0.02) (Supplemental Table 2). Not only did ComBat-ref identify the highest number of DE genes in the RAS pathway, but gene set enrichment analysis (GSEA)[17] demonstrated that ComBat-ref achieved the best enrichment result, as shown in Supplemental Figure 3 plotted using the fgsea package [18], where ComBat-seq was a close second.

Additionally, we applied ComBat-ref to various real-world datasets, including example human RNA-seq datasets from the NPMatch paper [10] and NASA GeneLab transcriptomic datasets [19].

The datasets in [10] exhibited varying levels of batch effects. Below, we discuss results for two representative datasets: GSE182440 (with strong batch effects) and GSE173078 (with minimal batch effects).

The GSE182440 [20] dataset contained 12 control samples and 12 samples from individuals with alcohol use disorder (AUD), distributed across two batches. In the unadjusted dataset, the two batches were distinctly separated (Supplemental Figure 4). ComBat-ref effectively removed batch effects, blending samples from both batches well and significantly reducing batch-associated variation. ComBat-seq also performed well on this dataset. However, NPMatch over-corrected the data, likely introducing more false positives.

The GSE173078 [21] dataset included 36 samples from gingival tissue biopsies (12 healthy control, 12 gingivitis, 12 periodontitis), split across 2 batches. The batches were already well-mixed in the unadjusted dataset (Supplemental Figure 5). Both ComBat-seq and ComBat-ref applied minimal adjustments, further reducing batch variation while preserving the dataset's overall structure. Supplemental Figure 5 shows that ComBat-ref slightly outperformed ComBat-seq in reducing batch effects while maintaining data integrity. Again, NPMatch over-adjusted the dataset, failing to account for the low levels of batch effects.

We also analyzed NASA GeneLab transcriptomic datasets, which included mouse liver RNA-seq data from various space missions and library preparation technologies. These datasets contain multiple batch covariates, such as "mission", "library preparation" and "sequencing facility", with researchers primarily interested in differential expression of mouse liver genes between flight samples and controls. The original study [19] found that batch correction by "library preparation" using ComBat was the most effective approach, followed by ComBat-seq.

For this analysis, we combined those covariates into batch factors and tested ComBat, ComBat-seq and ComBat-ref on a subset of NASA GeneLab datasets comprising three distinct batches: GLDS_137, GLDS_242 and GLDS_48 (Supplemental Table 3), with 12, 9 and 14 samples, respectively. The dataset included 18 flight samples and 17 ground controls. The unadjusted data showed clear batch separation in the PCA plot (Supplemental Figure 6). All three methods effectively removed the majority of batch effects. ComBat-ref identified GLDS_137 as the reference batch and left its count data unadjusted, maintaining a PCA plot shape similar to the unadjusted data. Data from the other two batches were adjusted to remove batch effects, while retaining intra-batch grouping. This adjustment enhanced the power of differential expression analysis. These findings further demonstrate that ComBat-ref is highly effective for correcting batch effects in RNA-seq count data from diverse experiments and sources.

4. Discussion

In this study, we introduced ComBat-ref, a refined batch correction method designed to improve the reliability and interpretability of RNA-seq count data. Our simulations demonstrated that ComBat-ref consistently outperforms existing methods, including ComBat-seq and NPMatch, in mitigating batch effects while preserving biological signals. By employing a negative binomial model with pooled dispersion estimates and aligning other batches to a reference batch, ComBat-ref achieves near-optimal sensitivity and specificity, particularly in challenging datasets with high variability in batch dispersions. Moreover, ComBat-ref retains integer count data, ensuring compatibility with widely used differential expression (DE) analysis tools such as edgeR and DESeq2.

4.1. Contributions and improvements

One of the key innovations of ComBat-ref is its ability to handle datasets with heterogeneous batch effects effectively, even when the number of samples per batch is small. By identifying a reference batch with minimal dispersion and aligning other batches accordingly, ComBat-ref minimizes over-correction while preserving intra-batch structure. This approach addresses limitations observed in methods like NPMatch, which often introduce false positives due to excessive adjustments, and ComBat-seq, which struggles with datasets characterized by small batch sizes and significant batch effects. ComBat-ref consistently improves sensitivity in DE analysis, even in scenarios where batch dispersions and biological conditions introduce substantial complexity, as demonstrated in our simulations.

When applied to real-world datasets, such as the GFRN signal dataset and NASA GeneLab transcriptomic data, ComBat-ref outperformed other methods in recovering biologically meaningful signals. For example, ComBat-ref demonstrated its ability to enhance the detection of critical DE genes, such as EGFR in the GFRN dataset, and improved clustering metrics in PCA analyses. These results underscore its robustness across diverse experimental designs and datasets.

4.2. Potential applications to single-cell RNA-Seq data

Although ComBat-ref was developed for bulk RNA-seq data, its principles and methodology hold considerable potential for application to single-cell RNA sequencing (scRNA-seq) data. Batch effects in scRNA-

seq are particularly pronounced due to technical variations during cell isolation, library preparation, and sequencing. These challenges are compounded by the sparsity of scRNA-seq data, where a large proportion of genes are not detected in individual cells, and the variability of expression levels across cells within the same biological condition.

Adapting ComBat-ref for scRNA-seq could involve the following considerations:

1. **Sparsity Management**: Unlike bulk RNA-seq, scRNA-seq datasets often include a high proportion of zero counts due to dropout effects. Incorporating methods to address sparsity, such as zero-inflated negative binomial (ZINB) models [22], could extend the applicability of ComBat-ref to scRNA-seq data.
2. **Cell Clustering Preservation**: Preserving the integrity of cell clusters is critical in scRNA-seq data. ComBat-ref's approach of aligning batches while retaining the reference batch structure is particularly advantageous, as it minimizes distortions to biological signals during correction.

Applying ComBat-ref to scRNA-seq could unlock its potential for addressing technical variability while retaining cell-type-specific signals. This would improve downstream analyses, including trajectory inference, cell type annotation, and differential gene expression analysis across conditions.

4.3. Limitations and future directions

While ComBat-ref shows strong performance, several limitations and areas for future improvement remain:

1. **Elevated False Positive Rate (FPR)**: Although ComBat-ref achieves high sensitivity, this comes at the cost of a slightly elevated FPR compared to ComBat-seq. Future work could explore integrating more precise dispersion estimation models to better balance sensitivity and specificity.
2. **Applicability to Complex Experimental Designs**: The current framework assumes a single dispersion parameter per batch, which may not fully capture the complexity of datasets with hierarchical or nested batch structures. Extending the model to accommodate such scenarios could enhance its versatility.
3. **Evaluation on scRNA-seq Data**: Adapting ComBat-ref for scRNA-seq datasets will require incorporating sparsity-aware models and validating its performance on large-scale scRNA-seq benchmarks. This step is critical for enabling its adoption in the rapidly growing field of single-cell transcriptomics.

5. Conclusion

ComBat-ref represents a significant advance in batch effect correction for RNA-seq data. By leveraging an innovative reference batch approach and retaining integer count data, it achieves superior sensitivity, specificity, and compatibility with DE analysis tools. Its successful application to datasets like GFRN and NASA GeneLab highlights its utility across diverse experimental contexts. Extending its application to scRNA-seq could further enhance its impact, addressing batch effects in one of the most challenging yet promising areas of transcriptomics. Future efforts to refine dispersion modeling and integrate ComBat-ref with single-cell workflows will undoubtedly expand its utility and performance, paving the way for more robust and reproducible genomic research.

Reproducibility

Code to reproduce the results in this paper are available at https://github.com/xiaoyu12/Combat-ref

Funding

This material is based upon work supported by the National Institutes of Health Grant (SC3GM122659).

CRediT authorship contribution statement

Xiaoyu Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Investigation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGpt in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Conflict of Interest

We have no conflicts of interest to disclose.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.12.010.

References

[1] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics Jan. 2007;8(1):118–27. https://doi.org/10.1093/BIOSTATISTICS/KXJ037.
[2] Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. e161–e161 Nucleic Acids Res Dec. 2014;42(21). https://doi.org/10.1093/NAR/GKU864.
[3] 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 Aug. 2014;32(9):896–902. https://doi.org/10.1038/nbt.2931.
[4] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics Jan. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
[5] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
[6] Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform Sep. 2020;2(3). https://doi.org/10.1093/NARGAB/LQAA078.
[7] Sprang M, Andrade-Navarro MA, Fontaine JF. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. BMC Bioinforma Jul. 2022;23(6):1–15. https://doi.org/10.1186/S12859-022-04775-Y/TABLES/2.
[8] Kotlov N, et al. Procrustes is a machine-learning approach that removes cross-platform batch effects from clinical RNA sequencing data. Commun Biol 2024 7:1 Mar. 2024;7(1):1–14. https://doi.org/10.1038/s42003-024-06020-z.
[9] Shaham U, et al. Removal of batch effects using distribution-matching residual networks. Bioinformatics Aug. 2017;33(16):2539–46. https://doi.org/10.1093/BIOINFORMATICS/BTX196.
[10] Zito A, Martinelli A, Masiero M, Akhmedov M, Kwee I. NPmatch: latent batch effects correction of omics data by nearest-pair matching. 04.29.591524 bioRxiv May 2024;2024. https://doi.org/10.1101/2024.04.29.591524.
[11] Zhang Y, Jenkins DF, Manimaran S, Johnson WE. Alternative empirical Bayes models for adjusting for batch effects in genomic studies. BMC Bioinforma Jul. 2018;19(1):1–15. https://doi.org/10.1186/S12859-018-2263-6/TABLES/2.
[12] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res May 2012;40(10):4288–97. https://doi.org/10.1093/NAR/GKS042.
[13] Stan Development Team, RStan: the R interface to Stan, 2024, ⟨https://mc-stan.org/⟩: R package version 2.32.6.
[14] Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics Feb. 2015;31(17):2778–84. https://doi.org/10.1093/BIOINFORMATICS/BTV272.
[15] McQuerry JA, et al. Pathway activity profiling of growth factor receptor network and stemness pathways differentiates metaplastic breast cancer histological subtypes. BMC Cancer Sep. 2019;19(1):1–14. https://doi.org/10.1186/S12885-019-6052-Z/FIGURES/4.
[16] Hennig Christianfpc: flexible procedures for clustering, 2023, R Package: 2023.

[17] Subramanian A, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci Oct. 2005;102 (43):15545–50. https://doi.org/10.1073/pnas.0506580102.
[18] Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv Jan. 2021:060012. https://doi.org/10.1101/060012.
[19] Sanders LM, et al. Batch effect correction methods for NASA GeneLab transcriptomic datasets. Front Astron Space Sci Jun. 2023;10:1200132. https://doi.org/10.3389/FSPAS.2023.1200132/BIBTEX.
[20] Lim Y, et al. Exploration of alcohol use disorder-associated brain miRNA–mRNA regulatory networks. Transl Psychiatry Dec. 2021;11(1). https://doi.org/10.1038/s41398-021-01635-w.
[21] Kim H, Momen-Heravi F, Chen S, Hoffmann P, Kebschull M, Papapanou PN. Differential DNA methylation and mRNA transcription in gingival tissues in periodontal health and disease. J Clin Periodo Sep. 2021;48(9):1152–64. https://doi.org/10.1111/jcpe.13504.
[22] Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. BioMed Cent Ltd Dec. 01, 2022. https://doi.org/10.1186/s13059-022-02601-5.