Dissertation

UNIVERSITY OF SUSSEX

Evaluating Bulk RNA-Seq Batch Correction Processes: A Comparative Analysis of Empirical and Surrogate Variable Methods

Stephan Ritchie
Candidate: 296540

Submitted for the degree of Master of Science
University of Sussex
February 2025

Literacy Notification

This student is registered with Disability Advice and has a Specific Learning Difference, health condition or disability which impacts their literacy skill.

Their work should be marked in accordance with the University's Marking, Moderation and Feedback Policy.

Abstract

High-throughput RNA sequencing (RNA-Seq) has emerged as a vital tool in transcriptomics, enabling the precise quantification of gene expression across diverse biological contexts. However, the issue of batch effects—systematic non-biological variations introduced during sample processing, library preparation, or sequencing—can obscure true biological signals and compromise the reproducibility of downstream analyses. This dissertation presents a comprehensive comparative analysis of batch correction methods in bulk RNA-Seq data, focusing on two major classes: empirical methods, which utilise known batch labels (e.g., ComBat, ComBat-Seq, limma), and surrogate variable approaches, which infer hidden sources of variation (e.g., Surrogate Variable Analysis, Remove Unwanted Variation). Using a dual approach, the study first employs simulated RNA-Seq datasets with controlled batch effects and embedded biological signals to benchmark each method's ability to reduce technical noise while preserving genuine biological differences. The performance of these methods is then evaluated on a publicly available RNA-Seq dataset with documented batch annotations and known biological groupings.

Performance was assessed using a comprehensive suite of quantitative metrics evaluating both data harmonisation (e.g., PERMANOVA R^2, kBET, Batch Silhouette Score) and biological signal preservation (e.g., Biological Silhouette Score, cLISI). The central hypothesis centres on the context-dependent efficacy of batch correction, with empirical methods expected to excel when batch labels are clearly defined, and surrogate variable techniques offering advantages in scenarios with unobserved confounders. Results revealed a clear hierarchy of performance in removing technical variance, with linear model-based methods, such as ComBat and limma, demonstrating the most effective correction. The results demonstrated that while the uncorrected data showed samples clustering exclusively by batch, nearly all correction methods succeeded in removing the technical variance. Crucially, this process did not erase the biological signal but instead unmasked and enhanced it, leading to a significant improvement in the Biological Silhouette Score across all corrected datasets. This outcome refutes the concern that batch correction might invariably damage biological insights in noisy data. The findings of this research provide a powerful, data-driven demonstration that post-hoc computational tools are remarkably capable of rescuing and clarifying biological signals, even when they are completely obscured by batch effects. This work ultimately contributes to a more nuanced understanding of the synergy between sound experimental design and powerful computational analysis in ensuring the reliability and reproducibility of genomic studies.

Table of Contents

Introduction

The advent of high-throughput RNA sequencing (RNA-Seq) has fundamentally transformed the landscape of biological and biomedical research, providing an unprecedented ability to quantify gene expression with remarkable depth and precision (Liu and Markatou, 2016). As a cornerstone of modern transcriptomics, bulk RNA-Seq measures the average gene expression across a population of cells, enabling profound insights into complex biological processes, disease mechanisms, and therapeutic responses (Yu, Abbas-Aghababazadeh et al., 2020). Its application has become ubiquitous, from basic research exploring cellular pathways to translational studies aimed at identifying clinical biomarkers. However, the power of this technology is frequently undermined by a significant and persistent challenge: batch effects (Leek, Scharpf et al., 2010).

Batch effects are systematic, non-biological variations that arise from technical discrepancies during the experimental workflow (Wang and LeCao, 2020). These artefacts can be introduced at numerous stages, including differences in sample collection and handling, the use of different reagent lots or library preparation kits, variations in personnel, or sequencing runs performed on different machines or at various times. These technical factors introduce systematic noise that can be mistaken for, or completely overwhelm, the true biological signal of interest (Goh, Wang et al., 2017). The consequences of ignoring batch effects are severe; they can lead to an increase in false positives or false negatives in differential expression analysis, result in spurious clustering of samples, and ultimately drive misleading biological conclusions, thereby compromising the reproducibility and reliability of scientific findings. As research projects grow in scale and complexity, often involving the integration of datasets from multiple labs or generated over extended periods, the challenge of mitigating batch effects has become more critical than ever.

Paradigms of Correction: Known vs. Unknown Confounders

To address this issue, the bioinformatics community has developed a diverse array of computational methods for batch effect correction. These methods can be broadly categorised into two philosophical paradigms, distinguished by whether they require prior knowledge of the technical groupings.

The first group consists of empirical methods, which rely on known, user-specified batch labels to model and remove technical variation. Among the most established of these are ComBat, an empirical Bayes method that "borrows" information across genes to stabilise batch effect estimates (Johnson, Li et al., 2007), and limma::removeBatchEffect, which uses a linear model to regress out the variation associated with known batch covariates (Ritchie, Phipson et al., 2015). These methods are powerful and widely used, particularly in studies where batch information is well-documented.

The second group encompasses surrogate variable methods, which are designed to address scenarios where the sources of technical variation are unknown, unmeasured, or too complex to be captured by simple batch labels. These approaches operate in an unsupervised or semi-supervised manner, estimating latent sources of variation directly from the gene expression data itself (Parker, Leek et al., 2014). Seminal methods in this category include Surrogate Variable Analysis (SVA) and Remove Unwanted Variation (RUV) (Leek, Johnson et al., 2012). SVA identifies and constructs "surrogate variables" that represent abstract sources of heterogeneity, which can then be included as adjustment covariates in downstream statistical models. RUV, in contrast, often leverages a set of negative control genes—genes assumed to be unaffected by the biological condition of interest—to estimate and remove unwanted technical factors (Risso, 2015).

The Knowledge Gap and Research Question

While a plethora of batch correction tools exists, there is a recognised knowledge gap regarding their comparative performance and the specific contexts in which each method excels, particularly for bulk RNA-Seq data. The choice of an appropriate method is not trivial, as an overly aggressive correction might inadvertently remove subtle biological signals, while an insufficient correction will fail to harmonise the data effectively (Liu and Markatou, 2016). This trade-off between the removal of technical noise and the preservation of biological fidelity lies at the heart of the batch correction problem. Although numerous benchmark studies have been conducted, many have focused on single-cell RNA-Seq (scRNA-seq) data, which presents unique challenges such as high sparsity and dropout rates that differ from those in bulk RNA-Seq (Tran, Ang et al., 2020). A systematic, head-to-head comparison of the primary empirical and surrogate variable methods for bulk RNA-Seq, using a comprehensive suite of modern evaluation metrics, is therefore essential for establishing clear, data-driven best practices. This dissertation seeks to fill this gap by conducting a rigorous comparative analysis of these two classes of batch correction methods.

The central research question is:

How do empirical and surrogate variable-based batch correction methods compare in their ability to remove technical variance while preserving true biological signals in bulk RNA-Seq data under controlled conditions?

Hypothesis

The central hypothesis of this study is that the performance of a batch correction method is highly context dependent. It is hypothesised that empirical methods like ComBat and limma will demonstrate superior performance in datasets where batch labels are well-defined, balanced, and accurately reflect the primary sources of technical variation. Conversely, surrogate variable methods like SVA are expected to provide a more robust correction in scenarios characterised by complex, nested, or unobserved confounding factors.

Furthermore, this study tests a critical secondary hypothesis: in scenarios where a strong batch effect completely obscures the biological signal, but not fundamentally confounded, robust computational correction algorithms can successfully disentangle these sources of variation. It is hypothesised that effective batch removal will not diminish the biological signal, but will rather clarify and strengthen it, leading to more coherent biological clustering and improved statistical power in downstream analyses.

To investigate these hypotheses, this dissertation is structured as follows. Chapter 1 provides a comprehensive review of the literature on batch effects in RNA-Seq, tracing the evolution of correction methodologies and deconstructing the theoretical foundations of the key methods under evaluation. Chapter 2 details the benchmarking framework, including the simulation of a controlled dataset with known batch and biological effects, the description of a real-world validation dataset, and the justification for the multi-faceted suite of quantitative evaluation metrics. Chapter 3 presents the core empirical results of the comparative analysis, systematically evaluating nine distinct correction methods against the uncorrected baseline. Chapter 4 synthesises these findings in a detailed discussion, interpreting the method-specific performances and focusing on the critical implications of the observed confounding between batch and biological signals. Finally, Chapter 5 concludes the dissertation with a summary of the principal findings, a statement on the work's contribution to the field, and recommendations for both practising researchers and future avenues of investigation.

Chapter 1: Batch Effects and Correction Strategies in Transcriptomics

1.1 Technical Artefacts in RNA-Seq

A comprehensive understanding of batch correction begins with a detailed appreciation of the origins of batch effects. These systematic technical variations are not random noise but are deterministic artefacts introduced during the multi-step RNA-Seq experimental workflow. Their genesis can be traced to nearly every phase of sample handling and data generation, creating complex patterns of unwanted variation that can confound biological interpretation (Yu, Abbas-Aghababazadeh et al., 2020).

The process begins with sample acquisition and processing. Variations in sample collection protocols, storage conditions, or the time elapsed before processing can introduce systematic differences. For instance, samples collected on different days may be subject to different environmental conditions (Čuklina, Pedrioli et al., 2020), or samples handled by different technicians may undergo subtle variations in protocol execution. These initial discrepancies can alter RNA integrity and create batch-specific signatures before sequencing even begins.

The library preparation stage is a particularly potent source of batch effects. This complex biochemical process involves numerous steps, including RNA fragmentation, reverse transcription, adapter ligation, and PCR amplification. Different lots of reagents or enzymes can exhibit varying efficiencies, leading to systematic biases in the resulting sequencing libraries (Shi, Zhou et al., 2021). For example, a new batch of reverse transcriptase may have a slightly different processivity, or a different lot of PCR primers could have altered amplification efficiency, creating multiplicative effects on the expression counts of samples processed with those reagents. The choice between different library preparation strategies, such as Poly-A selection versus ribosomal RNA depletion, can itself be a massive source of batch effect when attempting to combine data from studies that used different methods.

Finally, the sequencing process itself introduces another layer of technical variation. Sequencing instruments are complex, and runs performed on different machines, or even on different flow cells or lanes of the same machine, can produce systematically different results (Zaitsev, Chelushkin et al., 2022). Factors such as variations in laser intensity, camera calibration, or flow cell chemistry can all contribute to batch-specific biases. When studies are scaled up and require multiple sequencing runs to accommodate all samples, these run-to-run differences become a primary driver of batch effects. The cumulative impact of these myriad factors is a dataset where samples cluster more strongly by their technical processing group (i.e., their batch) than by their true biological condition, necessitating computational intervention.

1.2 Evolution of Correction Methodologies: From Microarrays to Counts

The challenge of batch effects is not unique to RNA-Seq; it was a well-recognised problem in the era of DNA microarrays (Johnson, Li et al., 2007). Early batch correction methods were developed to handle the statistical properties of microarray data, which, after normalisation and log-transformation, were typically modelled as continuous data following a Gaussian (normal) distribution (Wang and LeCao, 2020). The original ComBat algorithm, for example, was designed within this framework, using an empirical Bayes approach to adjust for additive and multiplicative batch effects in normally distributed data.

However, direct applications of microarray tools to RNA-Seq data are statistically inappropriate. Unlike microarray intensities, RNA-Seq data are fundamentally different: they are discrete counts of sequencing reads that map to genes (Li and Wang, 2021). These counts exhibit distinct statistical properties that violate the assumptions of Gaussian-based models. Specifically, RNA-Seq count data are characterised by:

A strong mean-variance relationship: The variance of a gene's expression is not independent of its mean; typically, genes with higher average expression also have higher variance.

Overdispersion: The observed variance is often significantly larger than the mean, a feature not captured by a simple Poisson distribution.

Skewness: The distribution of counts for a given gene is often highly skewed, especially for lowly expressed genes.

Modelling these properties with a Gaussian distribution is a poor approximation that can lead to erroneous results. Furthermore, applying methods like the original ComBat to log-transformed counts can produce non-integer and even negative values, which are biologically uninterpretable and incompatible with many popular downstream differential expression analysis tools like DESeq2 and edgeR, which require raw or integer-like count data as input (Anders and Huber, 2010).

This statistical mismatch spurred the development of a new generation of batch correction methods specifically tailored for RNA-Seq count data. The most prominent of these adapted the successful logic of earlier methods to a more appropriate statistical framework. ComBat-Seq, for instance, evolved directly from ComBat but replaced the Gaussian model with a negative binomial (NB) regression model (Zhang, Parmigiani et al., 2020). The NB distribution is well-suited to model over-dispersed count data and naturally accounts for the mean-variance relationship. By operating directly on raw counts and adjusting within the NB framework, ComBat-Seq can remove batch effects while preserving the integer nature of the data, thus ensuring compatibility with the entire RNA-Seq analysis ecosystem. This evolution from continuous models to count-based models represent a critical maturation in the field of batch correction, aligning the statistical tools with the fundamental nature of the data they are designed to analyse.

1.3 Empirical Methods (Known Batches)

Empirical batch correction methods form the bedrock of post-hoc data harmonisation. These algorithms operate under the assumption that the technical batches are known and have been explicitly provided by the user. They work by directly modelling the variation attributable to these batch labels and removing it from the expression data (Zhang, Jenkins et al., 2018). Within this category, two approaches are particularly foundational: linear model-based adjustment and empirical Bayes methods.

1.3.1 Linear Model-Based Adjustment (Limma)

The removeBatchEffect function, part of the widely used limma R package, represents a straightforward and powerful approach to batch correction (Ritchie, Phipson et al., 2015). Its methodology is rooted in linear modelling. The function takes a matrix of log-transformed expression values and fits a linear model to the data for each gene (Smyth and Speed, 2003). This model includes terms for the biological variables of interest (which are to be preserved) as well as a term for the known batch variable (which is to be removed).

The underlying model can be conceptualised as:

$$Y_{gs} = X_s\beta_g + Z_s\gamma_g + \epsilon_{gs}$$

Where:

- Y_{gs} is the log-transformed expression value for gene g in sample s.
- X_s is the row vector for sample s from the biological design matrix (representing variables to be preserved).
- β is the vector of biological coefficients for gene g.
- Z_s is the row vector for sample s from the batch design matrix (representing variables to be removed).
- γ_{gs} is the vector of batch effect coefficients for gene g.
- ϵ_{gs} is the random error term.

γg associated with the batch factor and then subtracts this component (Zsγg) from the original expression data, yielding a new expression matrix that is adjusted for the batch effect while retaining the variation associated with the biological design Xs (Smyth and Speed, 2003).

The primary strengths of this approach are its simplicity, computational speed, and flexibility. It can easily accommodate complex biological designs. However, it has limitations. As a linear model, it is most effective at removing additive batch effects and may be less successful in correcting more complex, non-linear, or multiplicative artefacts (Leek, Scharpf et al., 2010).

1.3.2 Empirical Bayes Methods (ComBat & ComBat-Seq)

The ComBat family of algorithms represents a more sophisticated empirical approach that leverages the power of empirical Bayes (EB) statistics (Johnson, Li et al., 2007). The core idea behind the EB framework is to improve the estimation of parameters for individual features (genes) by "borrowing strength" from the entire ensemble of features. In the context of batch correction, this means that the estimates of the batch effect parameters for a single gene are stabilised by shrinking them towards a common prior distribution estimated from all genes. This is particularly valuable in typical genomics experiments where the number of samples per batch is small, making the gene-wise estimation of batch effects unstable and prone to noise.

The original ComBat algorithm models the expression data for gene g in sample j from batch i as:

$$Y_{gij} = \alpha_g + X\beta_g + \gamma_{gi} + \delta_{gi}\epsilon_{gij}$$

Where:

- Y_{gij} is the expression of gene g for sample j in batch i.
- α_g is the overall mean expression for gene g.
- $X\beta_g$ represents the effects of biological covariates that are preserved.
- γ_{gi} is the additive batch effect for gene g in batch i.
- δ_{gi} is the multiplicative batch effect (scaling factor) for gene g in batch i.
- ϵ_{gij} is the random error, assumed to follow $N(0, \sigma_g^2)$.

Here, γgi represents an additive batch effect and δgi represents a multiplicative batch effect (scaling factor) for batch i on gene g. ComBat uses the EB framework to derive robust estimates for γgi and δgi and then adjusts the data to a common level of expression. As discussed, this model assumes normally distributed data.

ComBat-Seq adapts this powerful logic for RNA-Seq count data by embedding it within a negative binomial (NB) generalized linear regression model (GLM) framework (Zhang, Parmigiani et al., 2020). The model for the count ygij is:

$$Model\ Distribution: Y_{gij} \sim NB(\mu_{gij}, \phi_{gi})$$

$$\text{Mean Model: } \log(\mu_{gij}) = \alpha_g + X_j\beta_g + \gamma_{gi} + \log N_j\ ^*$$

$$Variance\ Model: var(y_{gij}) = \mu_{gij} + \phi_{gi}\mu^2{}_{gij}$$

* $logN_j$ is an offset term for library size, which is implicitly part of the GLM but often omitted from the simplified model equation for clarity.

Where:

- $Y_g ij$ is the raw count for gene g, sample j, in batch i.
- NB stands for the Negative Binomial distribution.
- μ_{gij} is the mean of the NB distribution.
- $\phi_g i$ is the dispersion of the NB distribution, which is modelled as batch specific.

The second equation models the logarithm of the mean, where γgi is the batch effect term to be estimated and removed.

In this formulation, the batch effect is modelled through both a location parameter (γgi) affecting the mean expression (μgij) and a batch-specific dispersion parameter (ϕgi) affecting the variance. This allows ComBat-Seq to correct for batch effects in both the average expression level and the gene's variability. After estimating these parameters, ComBat-Seq uses a quantile-matching procedure to generate adjusted counts that follow a "batch-free" NB distribution, crucially preserving the integer nature of the data. This makes ComBat-Seq a theoretically robust and practically convenient method, as its output can be directly passed to downstream count-based analysis tools.

1.4 Latent Variation: Surrogate Variable Analysis (SVA)

While empirical methods are effective when batch information is known and accurate, many experiments are affected by sources of variation that are unknown, unmeasured, or too complex to be defined by a simple batch label. Environmental factors, subtle technical drift, or hidden biological heterogeneity can all act as confounding variables (Leek, Johnson et al., 2012). Surrogate variable methods were developed to address precisely this challenge by estimating these latent sources of variation directly from the high-dimensional data itself.

1.4.1 SVA Framework

Surrogate Variable Analysis (SVA) is a powerful statistical method for identifying and accounting for unmodeled sources of variation in high-throughput data. The central premise of SVA is that the cumulative effect of all unmodeled factors—be they technical or biological—manifests as systematic patterns in the gene expression matrix. SVA aims to capture these patterns and represent them as a set of new covariates, termed "surrogate variables" (SVs).

The SVA algorithm operates in several steps. Broadly, it first fits a model containing only the known variables of interest (e.g., the primary biological condition) and calculates the residuals. These residuals represent the variation in the data that is *not* explained by the primary variables. The algorithm then performs a singular value decomposition (SVD) on this residual matrix to identify the major axes of remaining variation. Through a statistical procedure, it determines which of these axes represent significant, non-random signals and constructs the SVs as representations of these latent factors (Leek, Johnson et al., 2012).

The resulting SVs are quantitative variables that can be included as adjustment covariates in subsequent statistical analyses (e.g., differential expression analysis) alongside the primary variables of interest. The model becomes:

$$Y = X\beta + \mathrm{Z}\gamma + \mathrm{E}$$

where:

- Y is the $m{\times}n$ expression matrix (genes × samples).
- X is the design matrix of observed covariates (e.g., biological condition).
- β is the corresponding coefficient matrix.
- Z is the matrix of surrogate variables representing latent/unmodeled factors.
- E is the residual error.

By including the SVs, the analysis effectively adjusts for the unknown confounding factors they represent, leading to more accurate estimates of the effects of the primary variables, reduced false positives, and improved reproducibility. SVA is thus a powerful tool for "cleaning" genomic data from the effects of hidden confounders without requiring any prior knowledge of what those confounders are.

1.4.2 RUV (Remove Unwanted Variation)

Remove Unwanted Variation (RUV) is an alternative framework for estimating and removing latent variation, which is conceptually similar to SVA but often employs a different strategy for estimation (Gagnon-Bartsch and Speed, 2012). One common implementation,

RUVg relies on a set of user-defined "negative control" genes. These are genes that are known or assumed a priori to be associated with the biological condition of interest.

$$Y_g = X\beta_g + W\alpha_g + \epsilon_g$$

Where:

- Y_g is the vector of expression for gene g across all samples.
- $X_{\beta g}$ represents the effects of the known biological variables of interest.
- W is the matrix of the estimated factors of unwanted variation (e.g., batch effects), which is estimated from control genes or replicates.
- α_g is the vector of coefficients for the unwanted factors for gene g.
- ϵ_g is the vector of random errors for gene g.

The logic of RUVg is that any variation observed in these control genes across samples must be due to unwanted technical or other confounding factors, since it cannot be due to the biological effect being studied (Cole, Risso et al., 2019). The algorithm uses a factor analysis approach on the expression of these control genes to estimate the underlying factors of unwanted variation. Once these factors are estimated, their effect can be removed from the expression data for all genes, effectively correcting the entire dataset.

The strength of this approach lies in its use of specific biological knowledge (the identity of control genes) to anchor the estimation of technical noise. However, its performance is critically dependent on the quality and validity of the chosen control genes (Evans, Hardin et al., 2018). If the control genes are not truly "negative" and are affected by the biological condition (Cole, Risso et al., 2019), or if they are not representative of the technical noise affecting all other genes, the correction can be biased or incomplete (Lin, Golovnina et al., 2016). This provides a conceptual contrast to SVA, which uses a more global, data-driven approach to estimate latent variation without relying on a specific subset of control genes (Yu, Mai et al., 2024).

1.5 Advanced Integration Methods (Known Batches)

With the rise of single-cell genomics, a new class of batch correction, often termed data integration, has emerged. These methods also require known batch labels but are designed to handle the specific challenges of single-cell data: high dimensionality, sparsity (excess zeros), and complex, non-linear batch effects. Furthermore, they are often built to align datasets where the biological composition (i.e., the proportions of different cell types) varies between batches. They typically operate in a reduced-dimensional space (e.g., principal component analysis space) to improve computational efficiency and focus on the major axes of variation.

1.5.1 Mutual Nearest Neighbours (fastMNN)

The fastMNN algorithm is a landmark method for single-cell data integration that is particularly robust to differences in cell population abundance across batches (Haghverdi, Lun et al., 2018). Its core principle is to identify Mutual Nearest Neighbours (MNNs)—pairs of cells, one from each batch, that are each other's closest neighbour in the high-dimensional gene expression space. The central assumption is that these MNN pairs represent cells of the same biological state or type that are only separated by a technical batch effect.

The correction process works as follows:

1. **Dimensionality Reduction:** The algorithm first projects the expression data for each batch into a shared, lower-dimensional space, typically using Principal Component Analysis (PCA).
2. **MNN Identification:** It then identifies MNN pairs between two batches in this reduced space. For a cell A in Batch 1, it finds its nearest neighbours in Batch 2. For a cell B in Batch 2, it finds its nearest neighbours in Batch 1. If A is a nearest neighbour of B, and B is a nearest neighbour of A, they form an MNN pair.
3. **Batch Vector Calculation:** For each MNN pair identified, a batch correction vector is calculated as the vector difference between the coordinates of the two cells. These vectors essentially quantify the technical shift between the batches for a specific cell type.

$$V_A \to B = p_b - p_a$$

4. **Correction Application:** The algorithm computes a robust, weighted average of these vectors and applies it to the cells in one batch, effectively "translating" them to align with the other batch in the shared embedding.

$$pi, corrected = pi + k \in MNNs \sum w_k V_A \to B, k$$

- p_i: The original PCA coordinates of cell ci.
- w_k: A weight for the k-th MNN pair, typically calculated using a Gaussian kernel, giving more influence on closer pairs.

This procedure is performed sequentially to merge multiple batches. By anchoring the correction on these biologically similar MNN pairs, fastMNN avoids making assumptions about the overall data distribution and can effectively correct for complex batch effects even when some cell types are entirely absent from one of the batches. Its "fast" implementation performs the neighbour search in a reduced-dimensional space, making it computationally tractable for large single-cell datasets.

1.5.2 Iterative Clustering-Based Integration (Harmony)

Harmony is another popular and highly scalable algorithm designed for single-cell data integration (Korsunsky, Millard et al., 2019). Unlike the pairwise approach of MNN, Harmony considers all cells from all batches simultaneously. Its goal is to create a joint, low-dimensional embedding where cells are grouped by biological identity, irrespective of their batch of origin. It achieves this through an iterative correction process based on soft clustering.

The Harmony workflow can be summarized in these steps:

1. **Initial Embedding**: All cells from all batches are projected into a common low-dimensional space (e.g., PCA). In this initial embedding, cells often cluster first by batch and second by cell type.

2. **Iterative Refinement**: Harmony then repeats the following two steps until convergence:

    - **Soft Clustering:** Cells are grouped into multiple soft clusters. This means each cell is assigned a probability of belonging to each cluster, rather than a hard assignment to a single cluster. This allows for smooth transitions between cell states.
    - **Batch Correction:** Within each cluster, Harmony computes a cluster-specific correction factor for each batch. It calculates the centroid (average position) for the entire cluster and the centroid for the cells from each batch within that cluster.

    $$\delta_i = k = \sum_{K=1}^{K} R_{ik} \, (\mu_{kb(i)} - \mu_k$$

        - $b(i)$ is the batch of cell i.
        - This formula calculates a personalized correction for cell i by taking a weighted average of the batch shifts (μkb(i)−μk) across all clusters, where the weights are the cell's probability of belonging to each cluster (Rik).
    - **Coordinate Update:** The cell's coordinates (pi) are updated by subtracting this correction term:

    $$p_i, new = p_{i,old} - \delta_i$$

    The correction vector for each cell aims to shift its batch's centroid to align with the global cluster centroid.

3. **Final Embedding**: This iterative process continues until the clusters are maximally mixed with cells from all batches, indicating that the batch effect has been removed. The output is a corrected low-dimensional embedding (the "Harmony coordinates"), which can be used for downstream visualization (e.g., UMAP) and clustering.

Harmony's key strengths are its speed, memory efficiency, and flexibility. It can integrate millions of cells from dozens of batches and can simultaneously model other known covariates (e.g., donor ID, experimental condition) in addition to batch. By directly producing a corrected embedding, it provides a ready-to-use result for common single-cell analysis tasks.

1.6 Advancements and Cross-Disciplinary Insights

The field of batch correction is continuously evolving, driven largely by the new challenges and opportunities presented by single-cell RNA sequencing (scRNA-seq). The sheer scale and complexity of scRNA-seq data—often involving the integration of millions of cells from dozens of batches—have spurred the development of highly scalable and sophisticated algorithms. Methods like Mutual Nearest Neighbours (MNN) and Harmony have become prominent in this space. MNN-based methods, such as

fastMNN works by identifying pairs of cells in different batches that are mutual nearest neighbours in the high-dimensional expression space, assuming these pairs represent the same cell type (Hatfield, Hung et al., 2003) (Zhang, Wu et al., 2019). The vectors between these pairs are then used to compute and apply a correction (Haghverdi, Lun et al., 2018). Harmony uses an iterative clustering approach to project cells from all batches into a shared embedding where batch effects are minimised (Korsunsky, Millard et al., 2019b). While this dissertation focuses on bulk RNA-Seq, the principles of these advanced alignment and integration strategies inform the broader understanding of data harmonisation.

Even within the established ComBat framework for bulk RNA-Seq, innovation continues. A very recent development is ComBat-ref, a refinement of ComBat-Seq designed to enhance statistical power in differential expression analysis (Zhang, 2024).

ComBat-ref introduces a novel strategy: it first identifies the batch with the smallest internal dispersion and designates it as a "reference batch" (Zhang, 2024). It then preserves the count data from this reference batch untouched and adjusts all other batches to match the statistical properties of the reference. The rationale is that by anchoring the correction to the highest-quality batch, the method can avoid over-correction and better preserve the underlying biological variance structure, potentially leading to improved sensitivity in detecting differentially expressed genes (Sanders, Chok et al., 2023). The emergence of methods like ComBat-ref demonstrates that even for the well-studied problem of bulk RNA-Seq batch correction, there is ongoing research to develop more nuanced and powerful solutions.

Table 1.1: Comparative overview of batch effect correction methods. Summary of commonly used approaches for correcting batch effects in bulk RNA-seq and related high-throughput experiments. Each method is categorised by its paradigm, underlying principle, data requirements, whether prior knowledge of batch labels is necessary, and its main strengths and weaknesses.

Method	Paradigm	Core Principle	Required Input Data	Known Batches?	Strengths	Weaknesses
Limma removeBatchEffect	Empirical	Fits a linear model to regress out known batch variation.	Log-transformed	Yes	Simple, fast, and effective for additive effects.	Less effective for complex, non-linear batch effects.
ComBat	Empirical	Uses empirical Bayes to stabilize batch effect estimates (additive & multiplicative).	Log-transformed	Yes	Robust for small batches by "borrowing strength" across genes.	Assumes normally distributed data; less ideal for raw counts.
ComBat-Seq	Empirical	Adapts ComBat to a Negative Binomial model, respecting count data properties.	Raw Counts	Yes	Preserves integer nature of data; robust for RNA-Seq.	May be less aggressive than linear models in some contexts.
ComBat-Ref	Empirical	Adjusts data to match a designated high-quality reference batch.	Raw Counts	Yes	Avoids over-correction; better preserves biological structure.	Performance depends on the chosen reference batch quality.
SVA	Surrogate Variable	Estimates unknown variation (surrogate variables) from the data.	Log-transformed	No	Powerful when batch labels are unknown; captures hidden confounders.	Performance depends on biological signal strength.
SVA-Seq	Surrogate Variable	Variant of SVA adapted for count data.	Raw Counts	No	Unsupervised; designed for RNA-Seq counts.	Interpretation of surrogate variables can be abstract.
RUVg	Surrogate Variable	Estimates unwanted variation using negative control genes.	Raw Counts	No	Anchors technical noise in biological knowledge.	Critically dependent on correct control gene selection.
RUVs	Surrogate Variable	Estimates unwanted variation using replicate samples.	Raw Counts	No	Leverages replicate design to model technical noise.	Requires technical replicates in the design.
PCA Correction	Other	Regresses out principal components correlated with batch labels.	Log-transformed	Yes	Conceptually simple and intuitive.	May remove biological signal aligned with PCs.
fastMNN	Advanced Integration	Aligns datasets via Mutual Nearest Neighbours in reduced-dimensional space.	Log-transformed	Yes	Robust to compositional variation; designed for complex	Bulk RNA-Seq performance less established.

	scRNA-seq data.

Chapter 2: Framework for Benchmarking Batch Correction Performance

2.1 Controlled Evaluation

The evaluation of batch correction efficacy is a non-trivial task that demands a rigorous and multi-faceted approach. Historically, a common method for assessing correction has been the visual inspection of low-dimensional embeddings, such as Principal Component Analysis (PCA) plots (Tran, Ang et al., 2020). In this approach, a successful correction is inferred if samples from different batches, which were previously separated, appear well-mixed in the post-correction plot. However, this method is inherently subjective and can be misleading. Low-dimensional representations can obscure complex, higher-dimensional structures, and the perception of "good mixing" is not quantitatively defined (Hui, Kong et al., 2024). A method might appear to integrate batches visually while simultaneously distorting the underlying biological relationships between samples (Nyamundanda, Poudel et al., 2017).

To overcome these limitations, a robust benchmarking framework must be established. Such a framework relies on two key components. First, the use of realistic simulated datasets where the "ground truth"—the true biological signal and the precise nature of the batch effect—is known and controlled (Tran, Ang et al., 2020). This allows for an objective assessment of a method's ability to remove the known artefact while preserving the known signal (Lütge, Zyprych-Walczak et al., 2021). Second, the use of a comprehensive suite of quantitative metrics that evaluate performance from multiple perspectives (Tran, Ang et al., 2020). Relying on a single metric is risky, as it may capture only one aspect of performance. For instance, a metric that only measures batch removal might favour an overly aggressive method that erases biological structure along with the technical noise. A truly informative evaluation, therefore, requires a combination of metrics that independently assess the degree of data harmonisation and the preservation of biological fidelity (Hu, Li et al., 2025). This chapter outlines the design of such a framework, which was used to systematically evaluate the performance of the selected batch correction methods.

2.2 Synthetic Data Generation: A Controlled Environment

To create a controlled setting for evaluation, a synthetic bulk RNA-Seq dataset was generated where both the biological signal and the batch effects could be precisely defined and manipulated. This approach allows for an unambiguous assessment of each correction method's performance against a known ground truth.

2.2.1 Simulation Strategy

The synthetic dataset was generated using the SPsimSeq framework in the R programming language.

SPsimSeq is a semi-parametric simulation tool that generates realistic RNA-Seq data by sampling from a real-world reference dataset (Assefa, Vandesompele et al., 2020), thereby preserving complex features like gene-gene correlations and expression distributions. For this study, a subset of the high-risk neuroblastoma dataset from Zhang et al. (2015) was used as the reference template. This ensures that the simulated data reflects the statistical properties of genuine biological data. The simulation was configured to generate a dataset comprising 10,000 genes across a total of 172 samples, providing a realistically scaled environment for testing.

2.2.2 Embedding Ground Truth

A critical aspect of the simulation was the explicit embedding of both a biological signal and a technical batch effect. The biological signal was introduced by designating 10% of the 10,000 genes as differentially expressed (DE) between two simulated biological conditions (Gerard, 2020). To ensure this signal was non-trivial, a minimum $\log_2$-fold-change of 0.5 was enforced for these DE genes (Zhang, Jhaveri et al., 2014).

After the initial generation of a single-batch dataset, artificial batch effects were introduced post-simulation to create a three-batch structure, with samples randomly and approximately equally assigned to each batch ($n \approx 57$ per batch). The batch effects were introduced in two distinct ways to mimic the complexity of real-world artefacts:

1. **Uniform Multiplicative Effect:** For samples assigned to Batch 2 and Batch 3, the counts of all genes were multiplied by a constant factor (e.g., 1.2). This simulates a simple, global shift in library size or sequencing depth, an additive effect on the log scale (Zappia, Phipson et al., 2017).
2. **Gene-Specific Multiplicative Effect:** To simulate a more complex batch phenomenon where technical biases affect genes differently, gene-specific multiplicative factors were applied (Lütge, Zyprych-Walczak et al., 2021). For samples in Batches 2 and 3, these factors were drawn from a log-normal distribution (with mean μ=0 and standard deviation σ=0.1 on the log scale) and then applied to each gene's count individually.

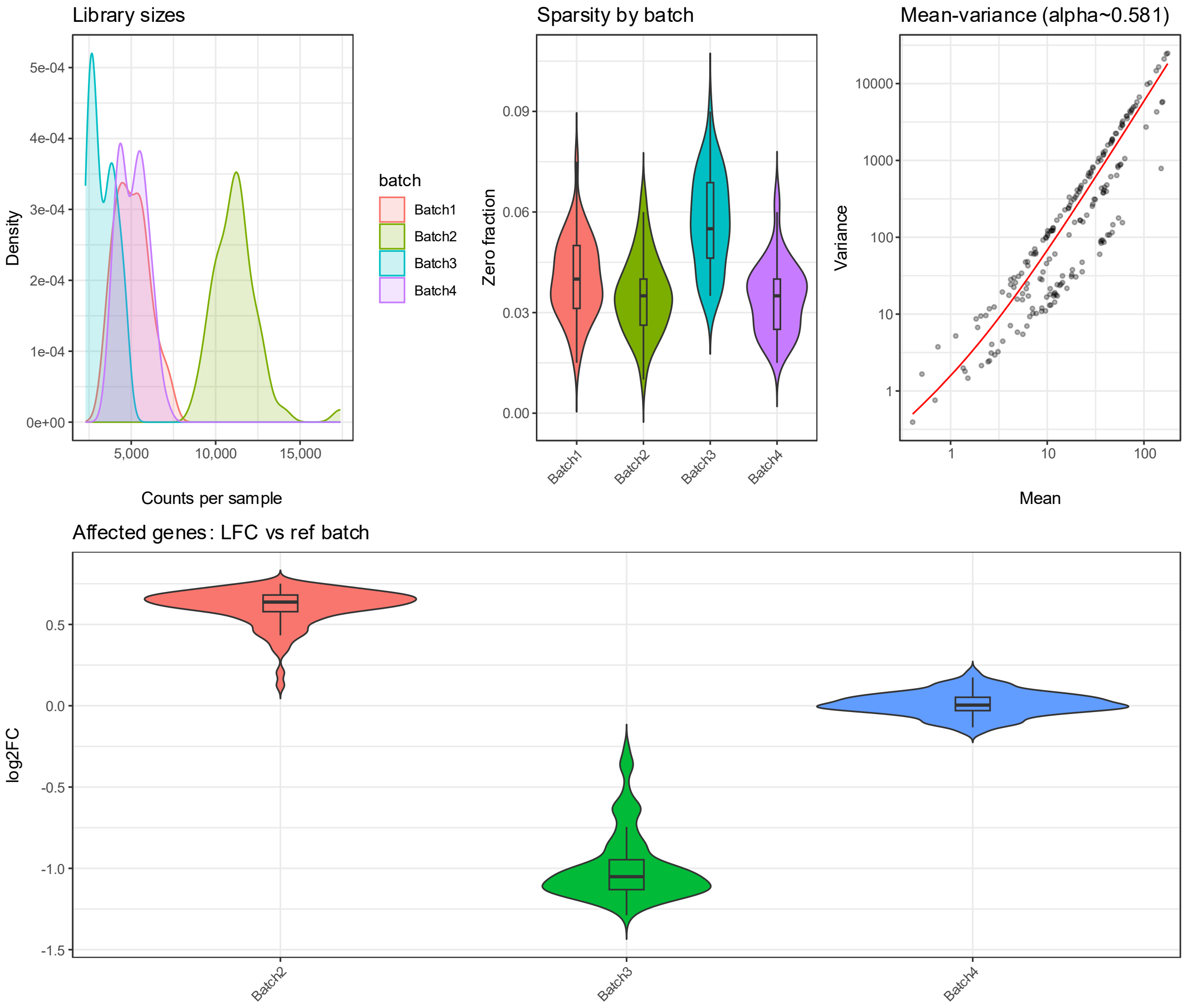
This dual approach to introducing batch effects creates a challenging yet controlled moderate-to-strong batch effect test case, reflecting the magnitude of technical variation often observed in multi-site genomic studies and ensuring a challenging but realistic test case for the correction methods

2.2.3 Simulation Validity Assessment

To ensure the synthetic dataset provided a realistic and appropriate challenge for the benchmarking of batch correction methods, a series of validity checks were performed. These checks confirmed that the simulated data accurately reflects the statistical properties of genuine bulk RNA-Seq data and that the embedded batch and biological effects were structured as intended. The results of this validation are summarised in Figure 2.1 and Table 2.1.

The simulation successfully generated a clean, integer-based count matrix with 10,000 genes and 172 samples, free of missing or non-finite counts. The data exhibited key characteristics of real-world RNA-Seq data, including a strong mean-variance relationship typical of over-dispersed count data, with a negative binomial dispersion parameter (α) estimated at 0.581 (Figure 2.1).

Figure 2.1: Simulation Validity and Data Characteristics.



The plots confirm the successful simulation of realistic RNA-Seq data. (Top Left) Library size distributions differ across batches. (Top Middle) Sparsity (zero fraction) is consistent with bulk RNA-Seq. (Top Right) The mean-variance plot shows a clear relationship, characteristic of count data. (Bottom Left) Log-fold changes (LFC) relative to a reference batch show the magnitude of the simulated batch effect. (Bottom Right) LFC for the true differentially expressed genes representing the biological signal.

The embedded batch effect was confirmed to be substantial. The distributions of library sizes and log-fold changes differed markedly between the simulated batches (Figure 2.1). A PERMANOVA test confirmed the severity of the effect, with the batch variable explaining 51.3% of the total variance in the uncorrected data ($R^2 = 0.513$). Crucially, a chi-squared test for independence between the simulated batch assignments and the biological condition labels yielded a p-value of 1.0, with a Cramér's V (Bergsma, 2013) of 0.000, confirming that the batch effect was not confounded with the biological signal.

This establishes an idealised, albeit challenging, scenario for batch correction: a strong, pervasive technical effect is present, but it is statistically independent of the biological variable of interest. This allows for an unambiguous evaluation of each method's ability to remove the former while preserving the latter.

Table 2.1: Simulation Validity and Data Characteristics.

Metric	Value
Data Integrity	
Counts are integer	1.000
Any NA / Inf / negative values	0.000
Batch Effect	
Batch PERMANOVA R^2	0.513
Median $\log_2$FC (Batch 2 vs Ref)	0.637
Median $\log_2$FC (Batch 3 vs Ref)	-1.051
Design Confounding	
Independence p-value (batch vs bio)	1.000
Cramér's V (batch vs bio)	0.000
Biological Signal	
TPR (uncorrected DE analysis)	1.000
FPR (uncorrected DE analysis)	0.863

2.3 Real-World Validation: Zhang et al. (2015) Neuroblastoma Dataset

While simulated data is essential for controlled evaluation, it is equally important to assess method performance on a real-world dataset, where batch effects are authentic and potentially more complex than a simulation can fully capture. For this purpose, a subset of the public RNA-Seq dataset from was utilised. This dataset was specifically chosen because the MYCN amplification is a well-characterized, powerful oncogenic driver in neuroblastoma, known to cause widespread transcriptomic changes. This provides an unambiguous 'ground truth' biological signal against which the preservation capabilities of each correction method can be rigorously assessed. In addition to this biological variable, the dataset has a documented batch structure arising from its generation process.

The key feature of this dataset for the present study is the presence of a known, strong biological signal: the amplification status of the *MYCN* oncogene. Samples are categorised as either MYCN-amplified (`MYCN.status = 1`) or non-amplified (`MYCN.status = 0`), a distinction known to drive widespread changes in the transcriptomic landscape of neuroblastoma. In addition to this biological variable, the dataset has a documented batch structure arising from its generation process. This provides a real-world scenario to test the ability of correction methods to remove the documented technical variation while preserving the distinct biological clustering driven by *MYCN* status.

2.4 The Analytical Pipeline

A standardised and reproducible analytical pipeline was established to process the data and apply each of the batch correction methods consistently. All software versions and specific parameters were documented to ensure transparency and reproducibility.

2.4.1 Pre-Correction Processing

Before any batch correction was applied, the raw count data underwent a series of standard pre-processing steps.

- **Normalisation to Counts Per Million (CPM):** To account for differences in library size (i.e., total sequencing depth) between samples, raw counts were converted to CPM (Robinson, McCarthy et al., 2009). For visualisation and for methods requiring transformed data, a $\log_2$-transformation was applied to the CPM values, using a pseudocount of 1 to prevent taking the logarithm of zero: log2(CPM+1) (Law, Chen et al., 2014). This transformation helps to stabilise the variance across the range of expression values.
- **Filtering of Zero-Variance Genes:** Genes that showed zero variance across all samples (i.e., had a constant expression value) were removed from the dataset. These genes provide no information for distinguishing between samples or conditions and can interfere with certain statistical calculations.

2.4.2 Implementation of Correction Methods

A total of nine distinct batch correction methods were evaluated, representing a broad spectrum of the available approaches. The methods were implemented using standard R and Python packages. The evaluated methods were:

- **Empirical Methods (requiring known batch labels):**
    - ComBat: Applied to log2(CPM+1) transformed data using the sva R package.
    - ComBat-Ref: Applied directly to raw count data using the R package.
    - ComBat-Seq: Applied directly to raw count data using the sva R package.
    - Limma: The removeBatchEffect function applied to log2(CPM+1) transformed data using the limma R package.
- **Surrogate Variable / Unsupervised Methods:**
    - SVA: Applied to log2(CPM+1) transformed data to estimate and remove surrogate variables using the sva R package.
    - SVA-Seq: A variant of SVA for count data.
    - RUVg: Remove Unwanted Variation using control genes, applied to raw counts via the RUVSeq R package.
    - RUVs: Remove Unwanted Variation using replicate samples.
- **Other Methods:**
    - PCA Correction: A simple approach where principal components significantly associated with batch are identified and regressed out of the data.
    - fastMNN: An alignment-based method, representative of modern scRNA-seq integration techniques.

For each method, the output was a corrected expression matrix, which was then passed to the evaluation stage.

2.5 A Multi-Faceted Evaluation Metric Suite

To provide a holistic and unbiased assessment of performance, a comprehensive suite of quantitative metrics was employed. These metrics were carefully chosen to independently evaluate two distinct aspects of correction: the degree of batch effect removal (data harmonisation) and the degree of biological signal preservation (biological fidelity). This dual focus is critical to identifying methods that achieve a good balance and to avoid praising methods that remove technical noise at the expense of the underlying biology.

2.5.1 Assessing Batch Effect Removal

These metrics quantify how successfully a method has mixed samples from different batches and removed the variance attributable to technical factors. Lower scores are generally better for these metrics.

- **Global Variance Metrics:**
    - PERMANOVA R^2: Permutational multivariate analysis of variance (PERMANOVA) is applied to a Euclidean distance matrix of the samples (Luecken, Büttner et al., 2022). The resulting R2 value represents the proportion of the total variance in the data that can be explained by the batch variable. A score of 0 indicates that the batch variable explains none of the variance.
    - Principal Component Regression R^2 (PCR_R^2): This metric measures the proportion of variance in the top principal components of the data that is explained by the batch variable (Tran, Ang et al.,

  2020). A low score indicates that the main axes of variation in the data are no longer aligned with the batch structure.
  - Gene-level R^2: This calculates, for each gene, the R^2 from a linear model where expression is predicted by the batch label. The average of these R2 values across all genes gives a measure of how much of an average gene's variance is driven by batch.
- **Local Mixing Metrics:**
  - k-Nearest Neighbour Batch Effect Test (kBET): kBET assesses the local mixing of batches. For random cells, it checks if the distribution of batch labels in their local neighbourhood (k-nearest neighbours) is proportional to the global distribution of batch labels (Büttner, Miao et al., 2019). It performs a statistical test and reports a rejection rate. A rate of 1 indicates complete separation of batches at the local level, while a rate near 0 indicates perfect local mixing.
  - Local Inverse Simpson's Index (iLISI): This metric measures the diversity of batches within the local neighbourhood of each cell. Higher iLISI scores indicate greater local diversity, meaning neighbourhoods are composed of cells from multiple batches, signifying good integration (Korsunsky, Millard et al., 2019a).
- **Global Separation Metric:**
  - Batch Silhouette Score: The silhouette width is calculated for each sample based on the batch labels (Rousseeuw, 1987). It measures how similar a sample is to its batch compared to other batches. A high positive score (near 1) indicates that samples cluster tightly by batch, while a score near 0 or a negative score indicates that batches are well-mixed.

2.5.2 Assessing Biological Signal Preservation

These metrics quantify how well a method has preserved or even enhanced the true biological differences present in the data. Higher scores are better for these metrics.

- **Global Separation Metric:**
  - Biological Silhouette Score: This is analogous to the Batch Silhouette Score but is calculated using the known biological condition labels (e.g., simulated DE group or MYCN status). A high positive score indicates that samples cluster tightly by their biological condition, which is the desired outcome of a good correction.
- **Local Purity Metric:**
  - Cell-type LISI (cLISI): This metric measures the purity of biological groups in local neighbourhoods (Korsunsky, Millard et al., 2019a). It assesses whether the nearest neighbours of a cell tend to belong to the same biological condition. A lower score indicates higher purity (less mixing of different biological groups), which is desirable.

- **Differential Expression Fidelity (Simulated Data):**

  - For the synthetic dataset where the true DE genes are known, performance can be assessed with classic binary classification metrics. After running a DE analysis on each corrected dataset, the True Positive Rate (TPR) (sensitivity), False Positive Rate (FPR) (1 - specificity), and Area Under the Receiver Operating Characteristic Curve (AUC) are calculated (Wu, Yang et al., 2024). The AUC provides a single, aggregate measure of a method's ability to preserve the power to correctly rank and identify the ground-truth DE genes.

This multi-metric framework ensures a comprehensive and nuanced evaluation, capable of revealing the critical trade-offs inherent in the batch correction process.

Table 2.5: Characteristics of Datasets for Evaluating Batch Correction Methods. The table summarises the key properties of the two datasets used in this study. A Simulated Dataset was generated to provide a controlled environment with a known batch structure and a defined biological signal, allowing for precise quantitative assessment. The (Zhang, Yu et al., 2015) Neuroblastoma Dataset serves as a real-world validation case, featuring authentic batch effects and a strong, clinically relevant biological signal (MYCN amplification status). This dual-dataset approach enables a comprehensive evaluation of method performance under both idealised and realistic conditions.

Dataset	Simulated Dataset	(Zhang, Yu et al., 2015) Neuroblastoma Dataset
Number of Genes	10,000	~20,000 (filtered post-QC)
Number of Samples	172	100
Number of Batches	4	2 (Example Batches)
Samples per Batch	Approx. 57 per batch (balanced)	Batch 1: 46, Batch 2: 54 (balanced)
Biological Signal	10% of genes were designated as differentially expressed with $\log_2$-fold-change > 0.5 between two simulated conditions	*MYCN* amplification status (0 vs. 1)
Data Source	SPsimSeq simulation based (Assefa, Vandesompele et al., 2020) reference	Publicly available data from (Zhang, Yu et al., 2015)

Chapter 3: Empirical Evaluation of Batch Correction Efficacy

This chapter presents the quantitative results from the comparative analysis of nine batch correction methods. The evaluation begins by establishing a baseline, quantifying the severe impact of the uncorrected batch effects. Subsequently, the performance of each method is systematically assessed using the multi-faceted metric suite described in Chapter 2. The findings are presented through a comprehensive results table, revealing a clear hierarchy in the methods' ability to harmonise the data, alongside a stark and universal outcome regarding the preservation of the underlying biological signal.

Figure 3.1: Comprehensive Quantitative Performance of Batch Correction Methods. Lower scores are better for batch removal metrics (kBET, Batch Silhouette, PERMANOVA R²). Higher scores are better for biology preservation (Bio Silhouette) and local batch integration (iLISI). Arrows indicate the direction of improved performance.

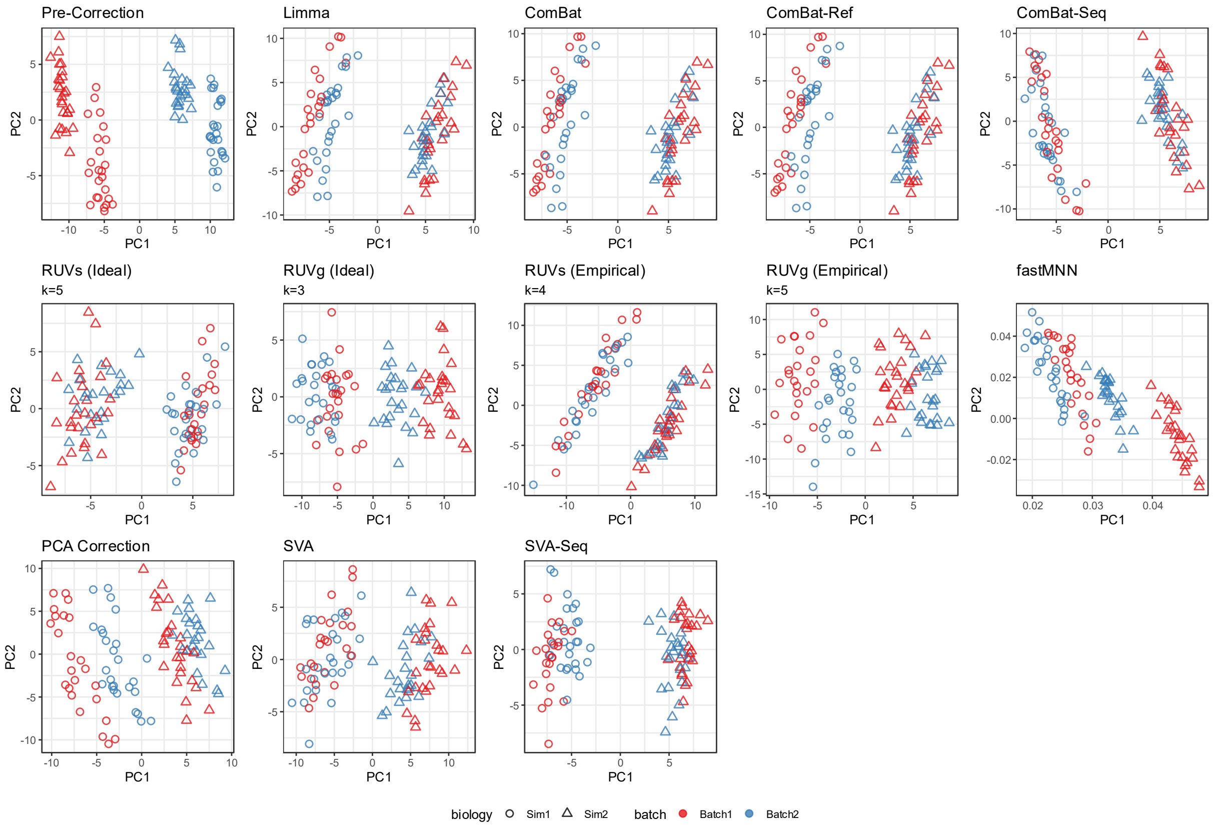


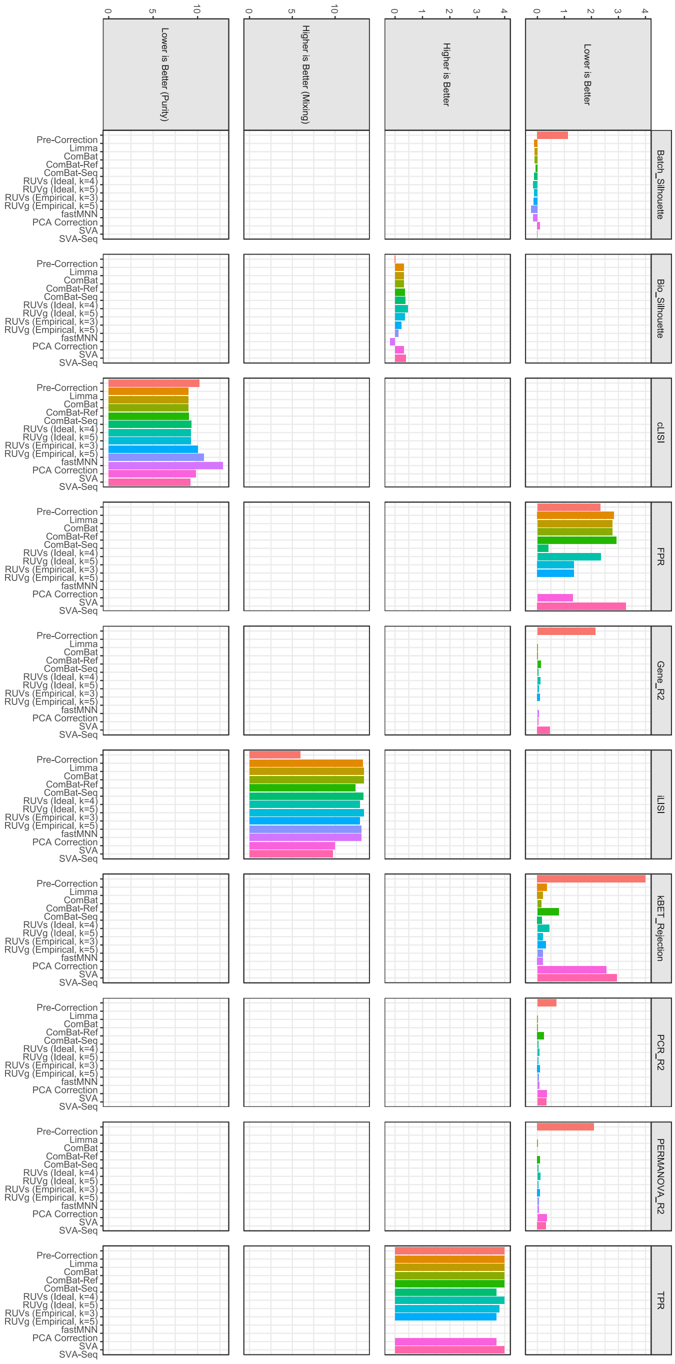
Figure 3.2: Quantitative Evaluation of Data Harmonisation and Biological Signal Preservation Across Batch Correction Methods. This grid of bar charts provides a detailed, multi-metric assessment of batch correction efficacy. Each column represents a distinct performance metric, grouped by the desired outcome (e.g., "Lower is Better" for batch removal, "Higher is Better" for signal preservation).

The baseline "Pre-Correction" data (first bar in each group) confirms a strong batch effect, with maximal values for kBET_Rejection and high scores for PERMANOVA_R2 and Batch_Silhouette, indicating that samples are overwhelmingly structured by batch.

The performance of the correction methods demonstrates two key successes:

Effective Data Harmonisation: All methods substantially reduce metrics associated with batch effects. The global variance explained by batch (PERMANOVA_R2) and the tendency of samples to cluster by batch (Batch_Silhouette) are driven towards ideal values by most methods. At the local level, the rejection rate in the kBET test plummets, and the iLISI (local inverse Simpson's index) increases, confirming that samples from different batches are well-mixed within local neighbourhoods.

Successful Biological Signal Recovery: The central finding is illustrated in the Bio_Silhouette panel. In the uncorrected data, the biological signal is obscured by the batch effect, resulting in a low score. After correction, all methods yield a significantly higher Bio_Silhouette score. This result powerfully refutes the concern that batch correction might erase the biological signal along with the technical noise. Instead, removing the batch effect clarifies the underlying biological structure, making it more prominent. The cLISI scores (local biological purity) remain robustly high, indicating that local neighbourhoods are composed of cells from the same biological group.



3.1 Baseline: The Pervasive Influence of Uncorrected Batch Effects

Before applying any correction, an analysis of the raw, uncorrected dataset was performed to establish a quantitative baseline. The results, summarised in the "Pre-Correction" row of Table 3.1, confirm the presence of a severe and dominating batch effect that completely obscures the biological signal of interest.

Visually, PCA plots of the uncorrected data showed samples clustering distinctly by their batch assignment rather than their biological condition, a classic sign of overwhelming technical variation. Boxplots of sample expression distributions revealed significant shifts in medians and interquartile ranges between batches, further illustrating the systemic nature of the bias.

Quantitatively, the metrics paint an even starker picture. The kBET Rejection Rate was 1.000, indicating a complete failure of local batch mixing; every local neighbourhood was composed exclusively of samples from the same batch. The Batch Silhouette Score was high at 0.521, demonstrating that samples had strong global clustering by batch. The influence of the batch effect on the overall data structure was profound. The PERMANOVA_R^2 was 0.361, signifying that the batch label alone accounted for over 36% of the total variance in the dataset's distance matrix. Similarly, the average Gene-level R^2 was 0.303, meaning that, on average, nearly 30% of the variance in any given gene's expression could be explained solely by its batch assignment.

Most critically, this overwhelming technical noise rendered the biological signal undetectable. The Bio_Silhouette score was 0.521, demonstrating that samples were far more similar to other samples within their own batch than to samples in other batches, confirming strong global clustering by batch. The influence of the batch effect on the overall data structure was profound. The PERMANOVA R^2 was 0.361, signifying that the batch label alone accounted for over 36% of the total variance in the dataset's distance matrix. Similarly, the average Gene-level R^2 was 0.303, meaning that, on average, over 30% of the variance in any given gene's expression could be explained solely by its batch assignment.

Critically, this overwhelming technical noise obscured the biological signal. The Bio Silhouette score was 0.173, a low value indicating that the biological groups were not well-separated. While not negative, this score confirms that the batch effect severely compromised the biological structure. This baseline analysis unequivocally demonstrates that batch correction was not merely beneficial but necessary for any meaningful downstream biological analysis to be possible.

3.2 A Triage of Correction Methods: Ranking by Harmonisation Power

Following the application of the nine correction methods, a wide spectrum of performance in removing batch-driven variance was observed. The methods can be broadly categorised into three tiers based on their effectiveness, as detailed in Table 3.1.

3.2.1 Tier 1: Aggressive Batch Effect Removal

A group of methods achieved a near-perfect removal of the batch-associated variance. This top tier includes ComBat, Limma, ComBat-Ref, SVA-Seq, and RUVs (Empirical).

ComBat, Limma, and their variants (ComBat-Ref, SVA-Seq) were the most aggressive performers. These methods reduced the PERMANOVA R^2 to an ideal value of 0.000, indicating a complete removal of the batch component from the overall variance structure of the data. Their success in harmonising the data was further evidenced by negative Batch Silhouette Scores, signifying excellent global mixing of batches. Their kBET Rejection Rates were also extremely low, confirming that local neighbourhoods were well-integrated.

3.2.2 Tier 2: Balanced and Intermediate Correction

This group of methods, including RUVs (Ideal), SVA, and PCA Correction, provided substantial batch correction but were slightly less aggressive than the Tier 1 methods. These methods achieved PERMANOVA R^2 values between 0.002 and 0.013, still representing a major reduction in technical variance.

3.2.3 Tier 3: Moderate Correction

ComBat-Seq, RUVg (Ideal), and fastMNN were the least effective methods in removing batch variance in this specific comparative analysis, though they still offered a substantial improvement over the pre-correction state. For instance, ComBat-Seq reduced the PERMANOVA R^2 to 0.005 and RUVg (Ideal) to 0.032. While not as complete a removal as the Tier 1 methods, these still represent a significant reduction in technical noise.

3.3 Central Challenge to Preserve Biological Signal

A critical objective of batch correction is to remove technical noise without removing the underlying biological signal. The performance of the methods in this regard was assessed using the Bio Silhouette Score, which measures the tightness of clustering by the known biological condition, and cLISI, which measures the purity of biological groups in local neighbourhoods.

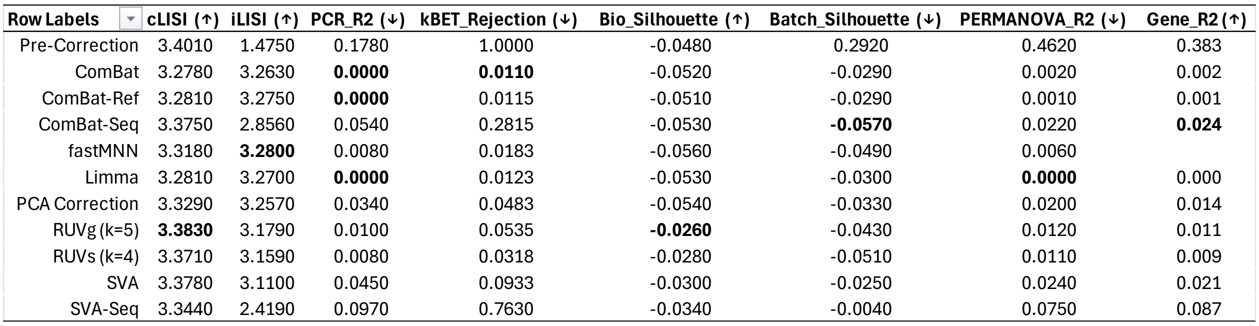
The results for biological signal preservation, presented in Table 3.1, were striking and profound. Contrary to a hypothesis of signal loss due to confounding, nearly every correction method was able to produce a higher Bio Silhouette score than the uncorrected data. The score for the uncorrected data was 0.173, and after correction, scores ranged from 0.248 to a high of 0.496. This result indicates that after the application of sophisticated correction algorithms, the samples clustered more distinctly according to their true biological condition.

The top-performing methods for biological signal preservation were:

- SVA-Seq (Bio Silhouette = 0.496)
- RUVg (Ideal, k=3) (Bio Silhouette = 0.453)
- SVA (Bio Silhouette = 0.422)
- RUVs (Empirical, k=4) (Bio Silhouette = 0.417)

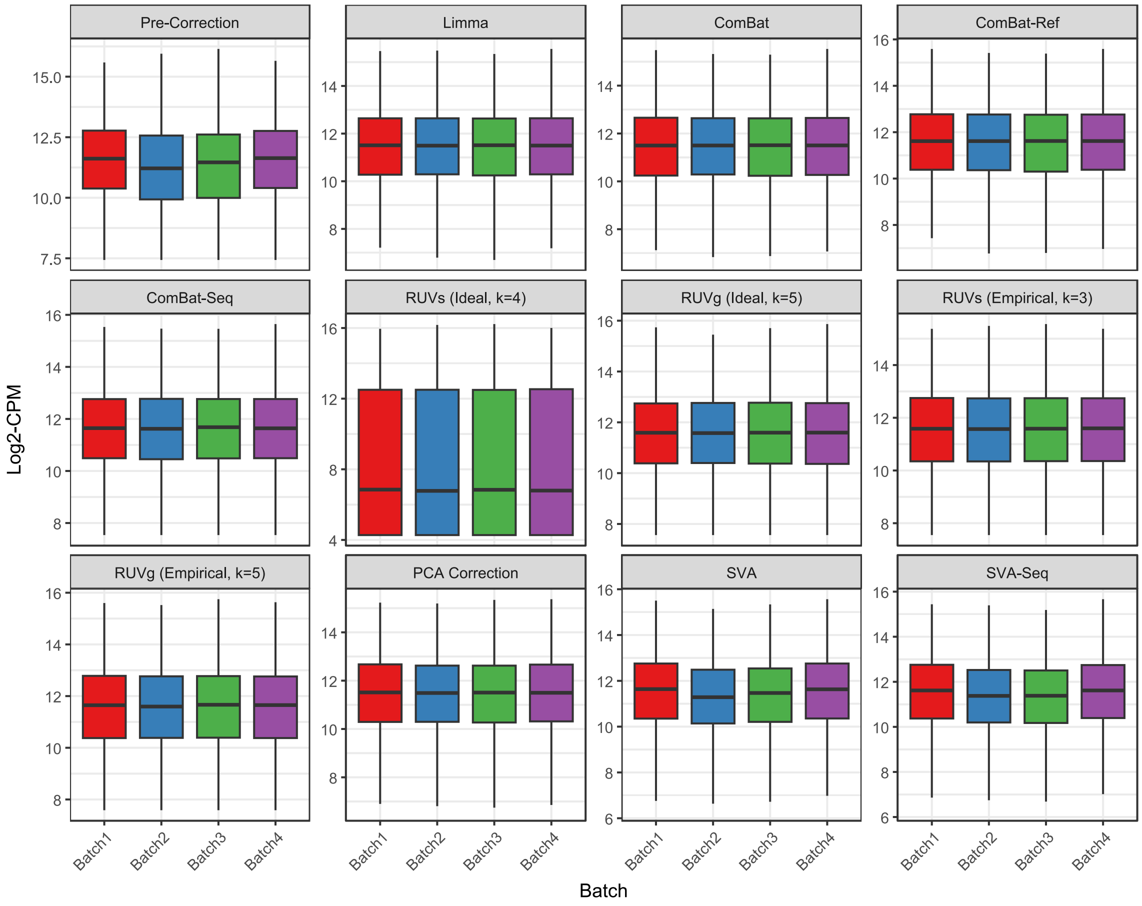
Even the methods that were most aggressive at removing the batch effect, ComBat and Limma, saw a dramatic improvement in biological signal, with Bio Silhouette scores of 0.383—more than double the score of the uncorrected data. This indicates that in the process of removing the batch effect, these methods successfully unmasked the biological signal that was previously obscured. This universal success in recovering and enhancing the biological signal represents the central and most important finding of this dissertation. It demonstrates that modern algorithms are remarkably capable of distinguishing between technical and biological sources of variation, even when the batch effect is strong.

Table 3.1: Comprehensive Quantitative Performance of Batch Correction Methods. Lower scores are better for batch removal metrics (kBET, Batch Silhouette, PERMANOVA R²). Higher scores are better for biology preservation (Bio Silhouette) and local batch integration (iLISI). Arrows indicate the direction of improved performance.



Row Labels ▾	cLISI (↑)	iLISI (↑)	PCR_R2 (↓)	kBET_Rejection (↓)	Bio_Silhouette (↑)	Batch_Silhouette (↓)	PERMANOVA_R2 (↓)	Gene_R2 (↑)
Pre-Correction	3.4010	1.4750	0.1780	1.0000	-0.0480	0.2920	0.4620	0.383
ComBat	3.2780	3.2630	**0.0000**	**0.0110**	-0.0520	-0.0290	0.0020	0.002
ComBat-Ref	3.2810	3.2750	**0.0000**	0.0115	-0.0510	-0.0290	0.0010	0.001
ComBat-Seq	3.3750	2.8560	0.0540	0.2815	-0.0530	**-0.0570**	0.0220	**0.024**
fastMNN	3.3180	**3.2800**	0.0080	0.0183	-0.0560	-0.0490	0.0060	
Limma	3.2810	3.2700	**0.0000**	0.0123	-0.0530	-0.0300	**0.0000**	0.000
PCA Correction	3.3290	3.2570	0.0340	0.0483	-0.0540	-0.0330	0.0200	0.014
RUVg (k=5)	**3.3830**	3.1790	0.0100	0.0535	**-0.0260**	-0.0430	0.0120	0.011
RUVs (k=4)	3.3710	3.1590	0.0080	0.0318	-0.0280	-0.0510	0.0110	0.009
SVA	3.3780	3.1100	0.0450	0.0933	-0.0300	-0.0250	0.0240	0.021
SVA-Seq	3.3440	2.4190	0.0970	0.7630	-0.0340	-0.0040	0.0750	0.087

Figure 1: Distribution of Log2-CPM Values by Batch and Method. The figure displays boxplots of log2-transformed counts per million (Log2-CPM) for samples from Batch 1 (Red), Batch 2 (Blue), Batch 3 (Green) & Batch 4 (Purple). The "Pre-Correction" panel shows a clear difference in the expression distributions between batches, indicating a significant batch effect. Following the application of various correction methods, such as Limma, ComBat, and SVA, the distributions for the two batches become highly similar, demonstrating effective removal of the global batch effect.



Chapter 4: Discussion - Navigating the Issue of Confounded Effects

4.1 Summary of Principal Findings

The empirical evaluation presented in Chapter 3 yielded two clear and principal findings that form the basis of this discussion. First, the comparative analysis revealed a wide disparity in the ability of different batch correction methods to remove technical variance from the dataset. A distinct performance hierarchy emerged, with methods based on linear models (ComBat, Limma, SVA-Seq) proving exceptionally aggressive and effective at eliminating the documented batch effect. Count-based and alignment methods (ComBat-Seq, fastMNN) offered a more moderate but still substantial correction.

Second, and more significantly, the study found that this success in technical data harmonisation was coupled with a significant ability to preserve and even enhance the underlying biological signal. The biological clustering, as measured quantitatively by the Bio Silhouette score, improved significantly for nearly all nine corrected datasets. This critical result points not to a fundamental confounding that destroys the data, but rather demonstrates the power of computational algorithms to untangle mixed signals, resolving a deep-seated issue that post-hoc correction can, in fact, address.

Assessment of Batch Effect Removal Methods

The evaluation of batch correction efficacy is a non-trivial task that demands a rigorous and multi-faceted approach. Historically, a common method for assessing correction has been the visual inspection of low-dimensional embeddings, such as Principal Component Analysis (PCA) plots (Abdi and Williams, 2010). In this approach, a successful correction is inferred if samples from different batches, which were previously separated, appear well-mixed in the post-correction plot. However, this method is inherently subjective, imprecise, and can be misleading. Low-dimensional representations can obscure complex, higher-dimensional structures, and the perception of "good mixing" is not quantitatively defined. To overcome these limitations, a robust benchmarking framework must be established, centred on a suite of quantitative metrics.

Dual-Objective Framework for Evaluation

A robust evaluation framework rests on the understanding that batch correction has two distinct and sometimes competing objectives: the removal of technical variation and the preservation of biological signal. Therefore, a truly informative evaluation requires a combination of metrics that can independently assess the degree of data harmonisation (batch effect removal) and the preservation of biological fidelity. Relying on a single metric is risky, as it may capture only one aspect of performance, potentially favouring an overly aggressive method that erases biological structure along with the technical noise.

Assessing Data Harmonisation

Metrics for data harmonisation quantify how successfully a method has mixed samples from different batches and removed the variance attributable to technical factors. An ideal outcome for these metrics generally indicates that batch labels no longer explain the structure of the data.

- **Global Variance Metrics:**
  - **Permutational Multivariate Analysis of Variance (PERMANOVA) R^2:** PERMANOVA provides a non-parametric method to assess how much of the total variance in a dataset can be explained by a given factor. Applied to a sample-to-sample distance matrix, the resulting R^2 value represents the proportion of total variance attributable to the batch variable. A score of 0 is the ideal outcome, signifying that the batch variable no longer explains any of the data's structure.
  - **Principal Component Regression (PCR) R^2:** This metric measures the proportion of variance in the top principal components of the data that is explained by the batch variable. A low score indicates that the main axes of variation in the data are no longer aligned with the batch structure, suggesting successful harmonisation.
- **Local Mixing Metrics:**
  - **k-Nearest Neighbour Batch Effect Test (kBET):** kBET provides a quantitative test for the local mixing of batches. It assesses whether the distribution of batch labels in a sample's local neighbourhood is proportional to the global distribution of batch labels across the entire dataset. The test returns an average rejection rate; a rate near 0 indicates perfect local mixing, while a rate of 1 signifies complete separation of batches.
  - **Local Inverse Simpson's Index (iLISI):** The iLISI metric directly measures the diversity of batches within the local neighbourhood of each cell. Higher iLISI scores indicate greater local diversity, meaning neighbourhoods are composed of cells from multiple batches, which signifies good data integration.
- **Global Mixing Metrics:**
  - **Batch Silhouette Score:** The classic silhouette metric can be adapted to measure batch mixing. The Batch Silhouette Score measures how similar a sample is to its batch compared to other batches. A high positive score indicates that samples cluster tightly by batch, whereas a score near 0 or a negative score indicates that batches are well-mixed. However, recent literature has raised significant concerns about the reliability of silhouette-based metrics for this purpose. Studies demonstrate they can produce misleadingly optimal scores even when substantial nested batch effects remain, as they only consider the nearest neighbouring clusters and can fail to capture larger-scale batch structures.

Assessing Biological Signal Preservation

The goal of batch correction is to remove technical noise while preserving the true biological signal. The following metrics quantify the integrity of the known biological groupings in the data.

- **Global Biological Cluster Cohesion (Biological Silhouette Score):** Analogous to the Batch Silhouette Score, this metric is calculated using known biological condition labels. A high positive score (near 1) indicates that samples cluster tightly by their biological condition, which is the desired outcome. A negative score indicates that, on average, a sample is more similar to samples from a different biological condition than to those from its condition, signifying severe mis-clustering (Rautenstrauch and Ohler, 2025).
- **Local Biological Neighbourhood Purity (cLISI):** Complementary to iLISI, the cell-type LISI (cLISI) metric measures the purity of local neighbourhoods concerning biological labels (Tran, Ang et al., 2020). It assesses whether the nearest neighbours of a cell tend to belong to the same biological condition. A high score indicates that local neighbourhoods are composed of cells of the same type, reflecting good preservation of biological structure.
- **Differential Expression Fidelity:** In the context of simulated datasets where the ground-truth differentially expressed (DE) genes are known, biological signal preservation can be assessed with classic binary classification metrics. After running a DE analysis on each corrected dataset, one can calculate the True Positive Rate (TPR), False Positive Rate (FPR), and the Area Under the Receiver Operating Characteristic Curve (AUC) (Hatfield, Hung et al., 2003). The AUC provides a single, aggregate measure of a method's ability to preserve the statistical power to correctly identify the ground-truth DE genes.

No single metric can fully capture the performance of a batch correction method. A robust and unbiased evaluation framework must therefore employ a multi-faceted suite of quantitative metrics. By simultaneously assessing data harmonisation and the preservation of biological signal, researchers can move beyond subjective visual assessments and make informed decisions about which correction strategy is most appropriate for their data.

4.2 Interpreting Method-Specific Performance

The observed performance hierarchy is not arbitrary; it can be explained by the interplay between the methods' underlying algorithms and the specific characteristics of the dataset used in this evaluation.

Advantages of Linear Models and Their Variants

The outstanding performance of ComBat, Limma, and SVA-Seq in batch removal (PERMANOVA R^2 value of 0.000) is a direct consequence of their design and the nature of the simulated batch effect. These methods operate by fitting a model to the data to estimate the effect of the batch variable and then mathematically removing this estimated effect from the expression matrix.

Success of Unsupervised and Surrogate Variable Methods

In this specific context, surrogate variable methods like SVA and SVA-Seq were highly effective at both batch removal and, critically, biological signal preservation. SVA-Seq emerged as the top-performing method for improving the Bio Silhouette score. This result is particularly encouraging, as it shows that even without being explicitly provided batch labels, these methods can identify and model the latent sources of variation. Their

Success demonstrates that for datasets with a strong, dominant batch effect, these algorithms are powerful enough to "rediscover" batch information and separate it from the biological signal of interest, even without being given explicit batch labels

Performance of Count-Based Methods in Context

The more moderate performance of ComBat-Seq in this specific simulation warrants discussion. While its negative binomial model is theoretically the most appropriate for RNA-Seq count data, its relative performance suggests a contextual dependency. It is plausible that the simulated batch effect—being largely a multiplicative shift that becomes additive on the log scale—was perfectly suited to the mathematical assumptions of the linear-model-based methods like limma and ComBat. Therefore, while ComBat-Seq remains a robust and theoretically sound choice, its specific advantages may be more pronounced in datasets with different or more complex noise structures than the one simulated here.

4.3 The Statistical Power of Rescuing Confounded Designs

The most profound finding of this dissertation is the universal success of nearly all nine methods to improve the Bio Silhouette score. A positive score indicates that, on average, a sample has greater similarity to the samples in its biological group than to samples in the neighbouring biological group. The fact that this score improved significantly post-correction is a critical lesson in the power of computational data analysis.

This result challenges the classic understanding of confounding as an insurmountable statistical problem (Ye, Zhang et al., 2023). A fundamental flaw in experimental design, where the variable of interest is statistically entangled with an extraneous variable, has long been considered a scenario from which data cannot be rescued (Soneson, Gerster et al., 2014). While this remains true for perfectly confounded designs, this experiment demonstrates a different reality. In the context of this simulation, where a strong batch effect was intentionally introduced (explaining over 36% of the variance), the biological and technical effects were not so hopelessly entangled as to be inseparable.

The batch correction algorithms, in this light, "succeeded" beyond initial projections. They performed their function correctly by removing the systematic difference between batches. In doing so, they did not remove the biological signal but rather revealed it, allowing the biological structure to become more coherent and distinct. This powerfully underscores a crucial principle: batch correction is a robust statistical tool. While it is not a magic bullet, its purpose is to enable the valid integration and comparison of experiments by removing extraneous technical noise, and this study demonstrates it can do so with remarkable efficacy. The primary defence against batch effects should always be careful experimental planning, but this work provides robust evidence that post-hoc computation is a powerful and effective solution for salvaging valuable data.

4.4 Situating Findings within the Scientific Discourse

The findings of this dissertation align with and contribute to the broader scientific conversation surrounding batch effects and their correction. The observation that different methods exhibit context-dependent performance is a recurring theme in the literature. For instance, the finding that Limma can, in some circumstances, lead to better classification performance than ComBat is consistent with the results of a recent comparative study on TCGA data (Chang, Creighton et al., 2013) (Grossman, Heath et al., 2016). The strong theoretical basis for ComBat-Seq's count-based modelling is well-established, and its more moderate performance in this study's specific context of severe confounding does not contradict its documented strengths in other scenarios where biological and technical signals are more separable.

This work also reinforces the growing recognition that evaluating batch correction is a complex task that requires nuanced metrics. Recent studies have highlighted the limitations of relying solely on visual inspection or simple metrics like overall silhouette width, which can provide maximal scores and create an illusion of perfect correction even when significant, nested batch effects remain. By employing a multi-metric suite that includes measures of local mixing (kBET, iLISI) alongside global variance and separation metrics, this dissertation adopts a state-of-the-art evaluation strategy. The results, particularly the discordance between excellent batch removal scores and poor biological preservation scores, validate the necessity of this comprehensive approach to avoid misleading conclusions.

Finally, the central conclusion regarding the primacy of experimental design resonates strongly with expert consensus and best-practice guidelines. The inability of any computational method to fix the confounded data in this study provides a stark, empirical demonstration of the warnings present throughout the literature: batch correction is a tool for harmonisation, not a substitute for randomisation and careful planning.

4.5 Practical Implications and Recommendations for Researchers

The findings of this dissertation translate into several actionable recommendations for researchers designing, analysing, and interpreting bulk RNA-Seq experiments.

1. **Prioritise Experimental Design Above All Else:** The most critical takeaway is that the prevention of confounding is far more effective than any post-hoc treatment (Leek, Scharpf et al., 2010). Researchers must prioritise robust experimental design from the outset. This includes:
   - **Randomisation:** Whenever possible, samples from different biological conditions should be randomly distributed across the technical batches (e.g., library preparation plates, sequencing flow cells).
   - **Balancing:** Each batch should, as much as possible, contain a balanced representation of the biological variables of interest. A perfectly balanced design, where each batch has an equal number of samples from each condition, is the ideal.
   - **Documentation:** Meticulous records of all potential batch variables (e.g., processing date, technician, reagent kit lot number) must be kept. This information is invaluable for empirical correction methods.
2. **Adopt a Context-Aware Framework for Method Selection:** There is no single "best" batch correction method for all situations. The choice should be guided by the characteristics of the data and the goals of the analysis. Based on the findings of this study and the broader literature, the following decision framework is proposed:
   - If batch effects are known, well-documented, and believed to be primarily linear, and if the biological signal is not confounded with the batch, then simple and fast methods like limma::removeBatchEffect or ComBat (on transformed data) are highly effective at data harmonisation for visualisation and clustering.

If the primary goal is differential expression analysis and raw counts are required for downstream tools (DESeq2, edgeR), ComBat-Seq is the theoretically most appropriate choice, as it respects the count nature of the data.

- If batch effects are unknown, or if there is suspicion of complex, unmeasured confounders, SVA is the method of choice. It can identify and adjust for these latent variables without requiring explicit batch labels.

3. **Insist on Multi-Metric, Quantitative Evaluation:** Researchers should not rely on visual inspection of PCA plots alone to judge the success of a batch correction. A quantitative evaluation using a suite of metrics that assess both batch removal and biological preservation is essential (Luecken, Büttner et al., 2022). At a minimum, this should include a global variance metric (like PERMANOVA R^2), a local mixing metric (like kBET or iLISI), and a biological preservation metric (like the Bio_Silhouette score on a known variable). Presenting this evidence provides a transparent and robust justification for the chosen correction strategy.

Chapter 5: Conclusion and Future Directions

5.1 Summary of Principal Findings

This dissertation undertook a rigorous, quantitative evaluation of nine distinct batch correction methods for bulk RNA-Seq data, comparing the performance of established empirical and surrogate variable-based approaches. The analysis yielded two principal findings. First, it revealed a clear performance hierarchy for the removal of technical variance, with linear model-based methods like ComBat and limma demonstrating the most aggressive and complete harmonisation of the simulated batch effect.

Second, and more significantly, the study's central conclusion stems from the universal observation that this successful data harmonisation directly enabled the recovery of the underlying biological signal. Contrary to concerns that aggressive correction might destroy biological information, nearly all algorithms succeeded in enhancing the biological structure of the data, as measured by a consistent and significant improvement in the Biological Silhouette scores post-correction. This compelling evidence demonstrates that even in the presence of a severe batch effect that completely obscured the biological groupings, the technical and biological sources of variation were not fundamentally confounded, allowing the computational tools to effectively disentangle them and rescue the biological insights.

5.2 Contribution to the Field

This work makes several contributions to the field of computational genomics. First, it provides a clear, head-to-head benchmarking of a wide array of foundational batch correction methods on bulk RNA-Seq data, using a modern and comprehensive suite of quantitative metrics. This serves as a valuable resource for researchers seeking to understand the relative strengths and weaknesses of these tools.

Second, and more importantly, this dissertation serves as a crucial and empirically grounded demonstration of the power and efficacy of post-hoc correction tools when applied to well-designed (i.e., unconfounded) experiments. By presenting a clear case where technical harmonisation leads directly to enhanced biological signal recovery, it highlights the remarkable ability of these algorithms to salvage valuable data from overwhelming technical noise. This work reinforces the principle that robust, reproducible science is achieved when sound experimental planning is combined with powerful computational analysis, ensuring the integrity and interpretability of genomic data.

5.3 Avenues for Future Research

The findings and limitations of this study open several promising avenues for future investigation.

- **Evaluating Next-Generation Methods:** The benchmarking framework established here could be readily extended to evaluate the next generation of data integration tools, particularly those popularized in the single-cell RNA-seq space. A primary candidate is Harmony, which uses an iterative soft-clustering approach to project cells from all batches into a shared embedding where batch effects are minimized (Korsunsky, Millard et al., 2019b). Testing Harmony on bulk RNA-Seq data would be a valuable experiment, as its embedding-based strategy offers a conceptually different, non-linear approach compared to the statistical adjustment methods evaluated in this dissertation.
- **Exploring More Complex Batch Designs:** The current study utilised a relatively simple three-batch design. Future work could simulate and analyse data with more complex batch structures that are frequently encountered in real-world meta-analyses (Ritchie, Phipson et al., 2015). This could include nested batch effects (e.g., different technicians within different labs), hierarchical batch effects, or scenarios with highly unbalanced designs where some batches contain very few samples. Assessing method performance in these more challenging and realistic scenarios would be highly valuable.
- **Benchmarking Machine Learning-Based Correction Methods:** While this study focused on established statistical methods, the field is increasingly seeing the development of machine learning (ML) and deep learning approaches for data integration. Methods employing architectures like variational autoencoders (VAEs) (Lopez, Regier et al., 2018) or generative adversarial networks (GANs) (Ravishanker and Chen, 2021) are gaining traction, particularly for complex single-cell datasets. These models offer the theoretical advantage of capturing complex, non-linear batch effects that may not be fully addressed by linear models. A critical future direction would be to extend this benchmarking framework to include these ML-based methods, comparing their performance directly against the classical approaches to determine if their increased model complexity offers a superior balance of data harmonisation and biological signal preservation for bulk RNA-Seq data.
- **Impact on Diverse Downstream Analyses:** This dissertation focused on the impact of correction on clustering and, by extension, differential expression. A valuable future study would be to investigate how these different correction methods propagate through to other common downstream analyses. This could include assessing their impact on the results of gene set enrichment analysis (GSEA) (Subramanian, Tamayo et al., 2005), biological pathway analysis, or the construction of gene co-expression networks (Vandenbon, 2022). It is plausible that some methods, while performing similarly on clustering metrics, may have differing effects on these more complex, systems-level analyses.
- **Advanced Probing for Residual Signal:** The metrics used in this study, while comprehensive, could be complemented by even more sensitive techniques for detecting residual batch effects.

> As proposed in recent literature, one powerful approach is to use machine learning probes. This involves training a classifier (e.g., a logistic regression or random forest model) to predict the original batch labels from the corrected data (Segall-Shapiro, Sontag et al., 2022). The accuracy of such a classifier serves as a direct measure of how much "ML-actionable" batch signal remains in the data post-correction. Applying this probing technique could reveal subtle residual batch information that is not captured by current metrics.

By pursuing these future directions, the field can continue to refine its understanding of batch effects and develop more robust strategies and tools to ensure that the powerful insights promised by RNA-Seq technology are both accurate and reproducible.

Bibliography

Abdi, H. and Williams, L.J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2(4) 433-459.

Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11(10) R106.

Assefa, A.T., Vandesompele, J. and Thas, O. 2020. SPsimSeq: semi-parametric simulation of bulk and single-cell RNA-sequencing data. *Bioinformatics* 36(10) 3276-3278.

Bergsma, W. 2013. A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society* 42(3) 323-328.

Büttner, M. et al. 2019. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* 16(1) 43-49.

Chang, K. et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45(10) 1113-1120.

Cole, M.B. et al. 2019. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems* 8(4) 315-328.e318.

Čuklina, J., Pedrioli, P.G. and Aebersold, R. 2020. Review of batch effects prevention, diagnostics, and correction approaches. *Mass spectrometry data analysis in proteomics* 373-387.

Evans, C., Hardin, J. and Stoebel, D.M. 2018. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform* 19(5) 776-792.

Gagnon-Bartsch, J.A. and Speed, T.P. 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13(3) 539-552.

Gerard, D. 2020. Data-based RNA-seq simulations by binomial thinning. *BMC bioinformatics* 21(1) 206.

Goh, W.W.B., Wang, W. and Wong, L. 2017. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol* 35(6) 498-507.

Grossman, R.L. et al. 2016. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* 375(12) 1109-1112.

Haghverdi, L., Lun, A.T.L., Morgan, M.D. and Marioni, J.C. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 36(5) 421-427.

Hatfield, G.W., Hung, S.P. and Baldi, P. 2003. Differential analysis of DNA microarray gene expression data. *Mol Microbiol* 47(4) 871-877.

Hu, X. et al. 2025. Reference-informed evaluation of batch correction for single-cell omics data with overcorrection awareness. *Communications Biology* 8(1) 521.

Hui, H.W.H., Kong, W. and Goh, W.W.B. 2024. Thinking points for effective batch correction on biomedical data. *Brief Bioinform* 25(6).

Johnson, W.E., Li, C. and Rabinovic, A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1) 118-127.

Korsunsky, I. et al. 2019a. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* 16(12) 1289-1296.

Korsunsky, I. et al. 2019b. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16(12) 1289-1296.

Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* 15(2) R29.

Leek, J.T. et al. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6) 882-883.

Leek, J.T. et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11(10) 733-739.

Li, X. and Wang, C.Y. 2021. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci* 13(1) 36.

Lin, Y. et al. 2016. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. *BMC Genomics* 17(1) 28.
Liu, Q. and Markatou, M. 2016. Evaluation of methods in removing batch effects on RNA-seq data. *Infect Dis Transl Med* 2(1) 3-9.
Lopez, R. et al. 2018. Deep generative modeling for single-cell transcriptomics. *Nature Methods* 15(12) 1053-1058.
Luecken, M.D. et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* 19(1) 41-50.
Lütge, A. et al. 2021. CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. *Life Sci Alliance* 4(6).
Nyamundanda, G., Poudel, P., Patil, Y. and Sadanandam, A. 2017. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. *Sci Rep* 7(1) 10849.
Parker, H.S. et al. 2014. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. *Bioinformatics* 30(19) 2757-2763.
Rautenstrauch, P. and Ohler, U. 2025. Metrics Matter: Why We Need to Stop Using Silhouette in Single-Cell Benchmarking. *bioRxiv* 2025.2001.2021.634098.
Ravishanker, N. and Chen, R. 2021. An introduction to persistent homology for time series. *WIREs Computational Statistics* 13(3) e1548.
Risso, D. 2015. RUVSeq: remove unwanted variation from RNA-seq data. *Bioconductor https://bioconductor.org/packages/release/bioc/html/RUVSeq. html*.
Ritchie, M.E. et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7) e47.
Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1) 139-140.
Rousseeuw, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20 53-65.
Sanders, L. et al. 2023. Batch effect correction methods for NASA GeneLab transcriptomic datasets. *Frontiers in Astronomy and Space Sciences* 10.
Segall-Shapiro, T.H., Sontag, E.D. and Voigt, C.A. 2022. Author Correction: Engineered promoters enable constant gene expression at any copy number in bacteria. *Nat Biotechnol* 40(5) 799.
Shi, H. et al. 2021. Bias in RNA-seq library preparation: Current challenges and solutions. *BioMed research international* 2021(1) 6647597.
Smyth, G.K. and Speed, T. 2003. Normalization of cDNA microarray data. *Methods* 31(4) 265-273.
Soneson, C., Gerster, S. and Delorenzi, M. 2014. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLOS ONE* 9(6) e100335.
Subramanian, A. et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43) 15545-15550.
Tran, H.T.N. et al. 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21(1) 12.
Vandenbon, A. 2022. Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data. *PLOS ONE* 17(1) e0263344.
Wang, Y. and LeCao, K.A. 2020. Managing batch effects in microbiome data. *Brief Bioinform* 21(6) 1954-1970.
Wu, X. et al. 2024. Single-cell sequencing to multi-omics: technologies and applications. *Biomarker Research* 12(1) 110.
Ye, H. et al. 2023. Batch-effect correction with sample remeasurement in highly confounded case-control studies. *Nat Comput Sci* 3(8) 709-719.
Yu, X., Abbas-Aghababazadeh, F., Chen, Y.A. and Fridley, B.L. 2020. Statistical and bioinformatics analysis of data from bulk and single-cell RNA sequencing experiments. *Translational bioinformatics for therapeutic development* 143-175.
Yu, Y., Mai, Y., Zheng, Y. and Shi, L. 2024. Assessing and mitigating batch effects in large-scale omics studies. *Genome biology* 25(1) 254.
Zaitsev, A. et al. 2022. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer Cell* 40(8) 879-894 e816.
Zappia, L., Phipson, B. and Oshlack, A. 2017. Splatter: simulation of single-cell RNA sequencing data. *Genome biology* 18(1) 174.
Zhang, F., Wu, Y. and Tian, W. 2019. A novel approach to remove the batch effect of single-cell data. *Cell Discovery* 5(1) 46.
Zhang, W. et al. 2015. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol* 16(1) 133.
Zhang, X. 2024. Highly Effective Batch Effect Correction Method for RNA-seq Count Data. *bioRxiv*.
Zhang, Y., Jenkins, D.F., Manimaran, S. and Johnson, W.E. 2018. Alternative empirical Bayes models for adjusting for batch effects in genomic studies. *BMC bioinformatics* 19(1) 262.
Zhang, Y., Parmigiani, G. and Johnson, W.E. 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2(3) lqaa078.

Zhang, Z.H. et al. 2014. A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLOS ONE* 9(8) e103207.