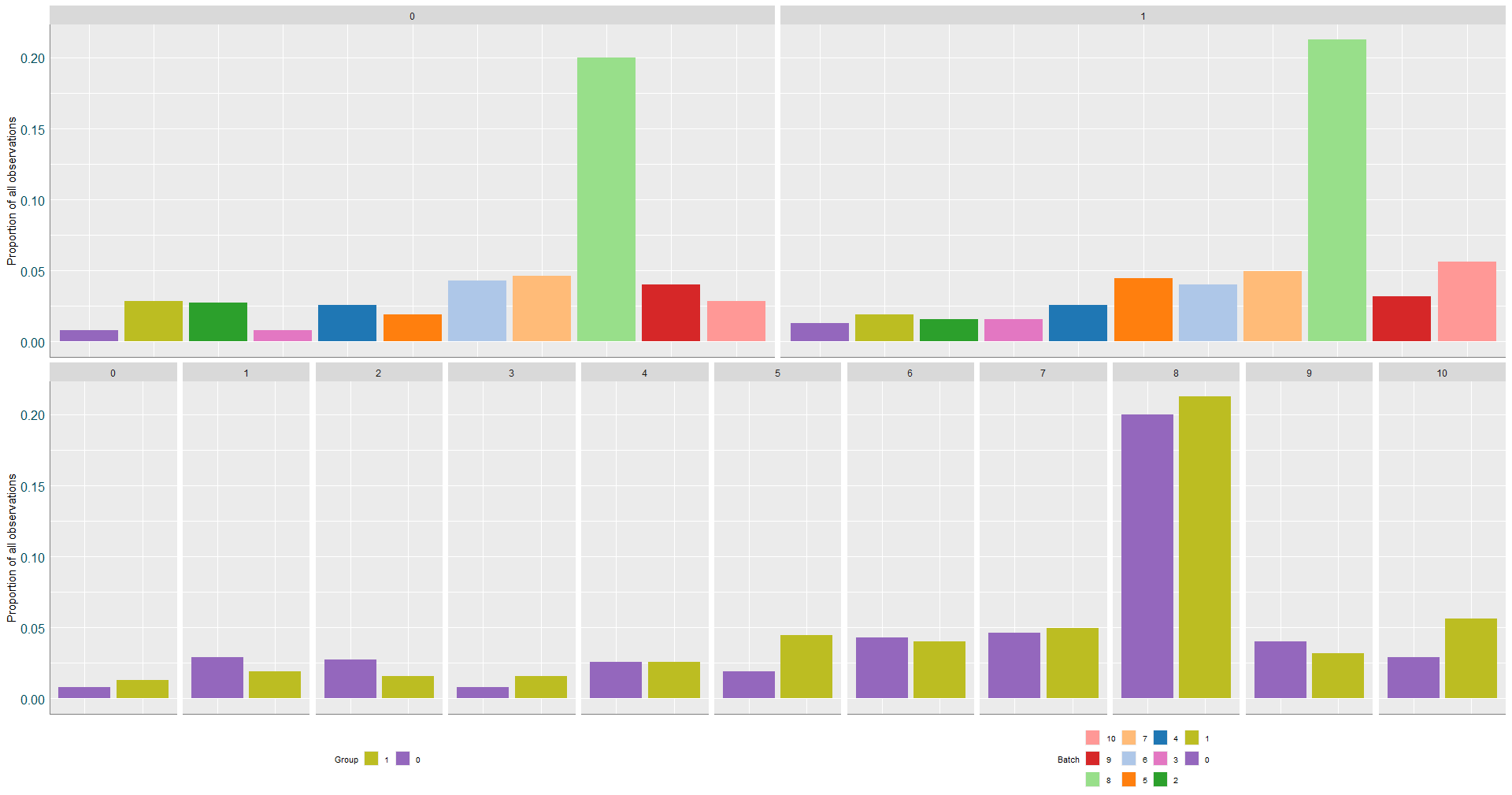
Pre-correction

1. Mosaic plot



The mosaic plot presented in this figure illustrates the distribution of Batch and Outcome (phenotype) before batch correction is applied. The plot is divided into two facets:

Facet 1 shows the distribution of samples from different batches across the two phenotype classes (Positive/Negative).

Facet 2 illustrates the distribution of the two phenotypes across the different batches.

The metric displayed in this plot is the proportion of observations within each combination of Batch and Phenotype. This allows us to assess how the data is distributed across batches and phenotypes before batch correction. The height of each bar represents the proportion of samples belonging to a particular batch and phenotype class.

To interpret this plot, look for patterns where certain batches dominate one phenotype or where the proportions of phenotypes within batches are disproportionate. If one batch contains mostly one phenotype, or if the phenotypic classes are unevenly distributed across batches, it suggests the presence of batch effects that could confound downstream analyses.

To determine whether batch correction is necessary, we use both statistical testing and visual inspection. The primary statistical test used is the Chi-Square test, which assesses whether the distribution of phenotypes is significantly different across batches. First, we create a contingency table that shows the counts of each phenotype within each batch. We then perform a Chi-Square test on this table to check for significant differences. If the p-value is less than 0.05, it indicates a significant imbalance in phenotype distribution across batches, suggesting that batch effects are present. In such cases, batch correction is recommended.

In addition to the statistical test, we also visually inspect the mosaic plot. This plot displays the proportions of each phenotype across batches, with imbalances in proportions indicating potential batch effects. If the plot shows that one batch is consistently associated with a specific phenotype or that batch distribution is uneven across phenotypes, this is a clear sign that batch effects may be influencing the data.

By combining the statistical p-value and visual inspection, we can confidently decide whether batch correction is necessary.

1. PCA, PVCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that transforms the data into a set of orthogonal components, which are linear combinations of the original features, ordered by the amount of variance they explain. When applied to the dataset, PCA identifies the principal components (PCs) that account for the most variability across the samples. If the samples from different batches cluster separately along the principal components, it indicates the presence of batch effects. In this case, PCA serves as a diagnostic tool to visually assess whether batch effects are confounding the results.

in a Principal Variance Component Analysis (PVCA), we might show how much variance is explained by biological factors (e.g., treatment), technical factors (e.g., batch), and residuals (unexplained variance). A well-controlled dataset will show that biological factors explain a larger proportion of variance compared to technical or batch effects. Visualizing variance components is particularly useful in determining if batch effects are confounding the results. If a significant proportion of the variance is attributed to batch effects, it indicates the need for batch correction to ensure that biological signals are not overshadowed by technical noise.

PCA: input CLR.

1. Correlation Analysis

Correlation Analysis is a statistical technique used to measure and describe the strength and direction of the relationship between two or more variables. In the context of microbiome or high-dimensional data, correlation analysis helps identify how different features (e.g., taxa or gene expressions) relate to each other across samples. High correlations between features can suggest that they share similar patterns of variation, which may be biologically meaningful. In the case of batch effects, correlation analysis can highlight discrepancies in the relationships between features across different batches. If the correlation patterns differ significantly between batches, it indicates the presence of batch effects.

1. Permutation Tests

Permutation Tests can be used to evaluate the necessity of batch effect correction by comparing the observed differences between groups or conditions to those obtained from randomly permuted data. In the context of batch effects, the permutation test helps determine whether the differences in the dataset are due to genuine biological variation or are artifacts caused by batch effects. By randomly shuffling the sample labels within batches and recalculating the test statistic for each permutation, we generate a distribution of test statistics under the null hypothesis (i.e., no batch effect). If the observed test statistic falls outside the range of the permuted statistics, it indicates that batch effects are likely confounding the results. In this case, applying batch correction methods can help to reduce these spurious effects and restore biological signal, which can be validated through repeated permutation testing to confirm that the corrected data reflects meaningful biological variation rather than batch-driven noise.

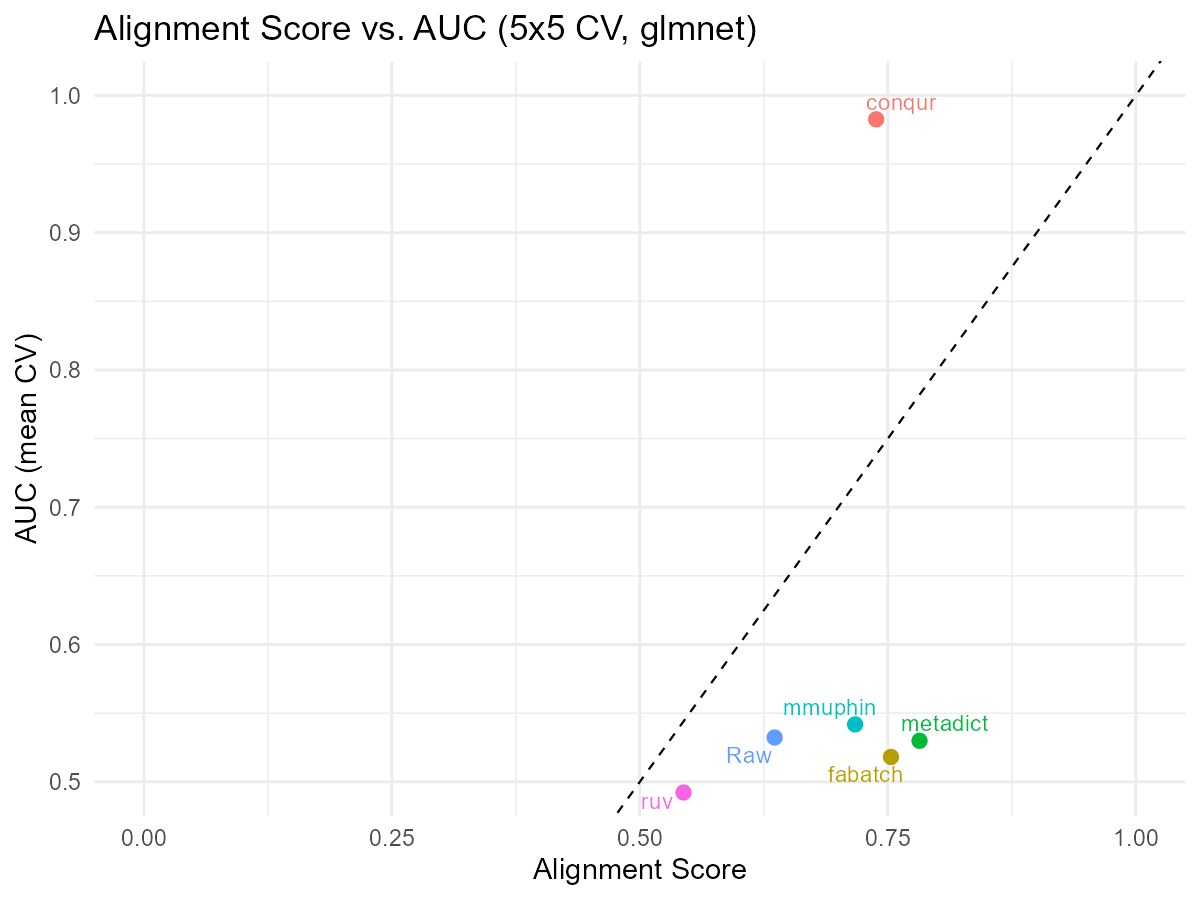
1. Batch-Specific Differential Expression (DE) Analysis

Batch-Specific Differential Expression (DE) Analysis is a statistical approach used to identify genes or features that exhibit differential expression across conditions within individual batches. The primary goal of this analysis is to separate genuine biological variation from confounding batch effects. In cases where batch effects are present, the apparent differential expression may be driven by batch-related factors rather than biological conditions. To address this, batch-specific DE analysis examines the differential expression for each batch separately, controlling for batch effects by including them as covariates in the analysis.

By performing DE analysis within each batch, researchers can identify features that are consistently differentially expressed across batches, while accounting for the batch-related noise. This method helps to detect whether any batch-specific factors are contributing to the observed differences, or if the observed DE signals are primarily biological. If batch effects are significant, performing batch correction before DE analysis is crucial to ensure that the results reflect true biological differences rather than artifacts introduced by batch variations. Batch-Specific DE Analysis can thus inform whether batch effects need to be addressed and guide the decision to apply appropriate correction methods before conducting further downstream analyses.

Post-Correction

1. Alignment Score vs. AUC



his figure shows the relationship between the Alignment Score (AS) and the Area Under the Curve (AUC) from a 5-fold, 5-replication cross-validation, using logistic regression (glmnet) for classification. The Alignment Score (AS) quantifies how well the data from different batches align, with a lower value indicating better alignment and reduced batch effect. The AUC measures the classification performance, where higher values (closer to 1) indicate better discriminatory power between the classes.

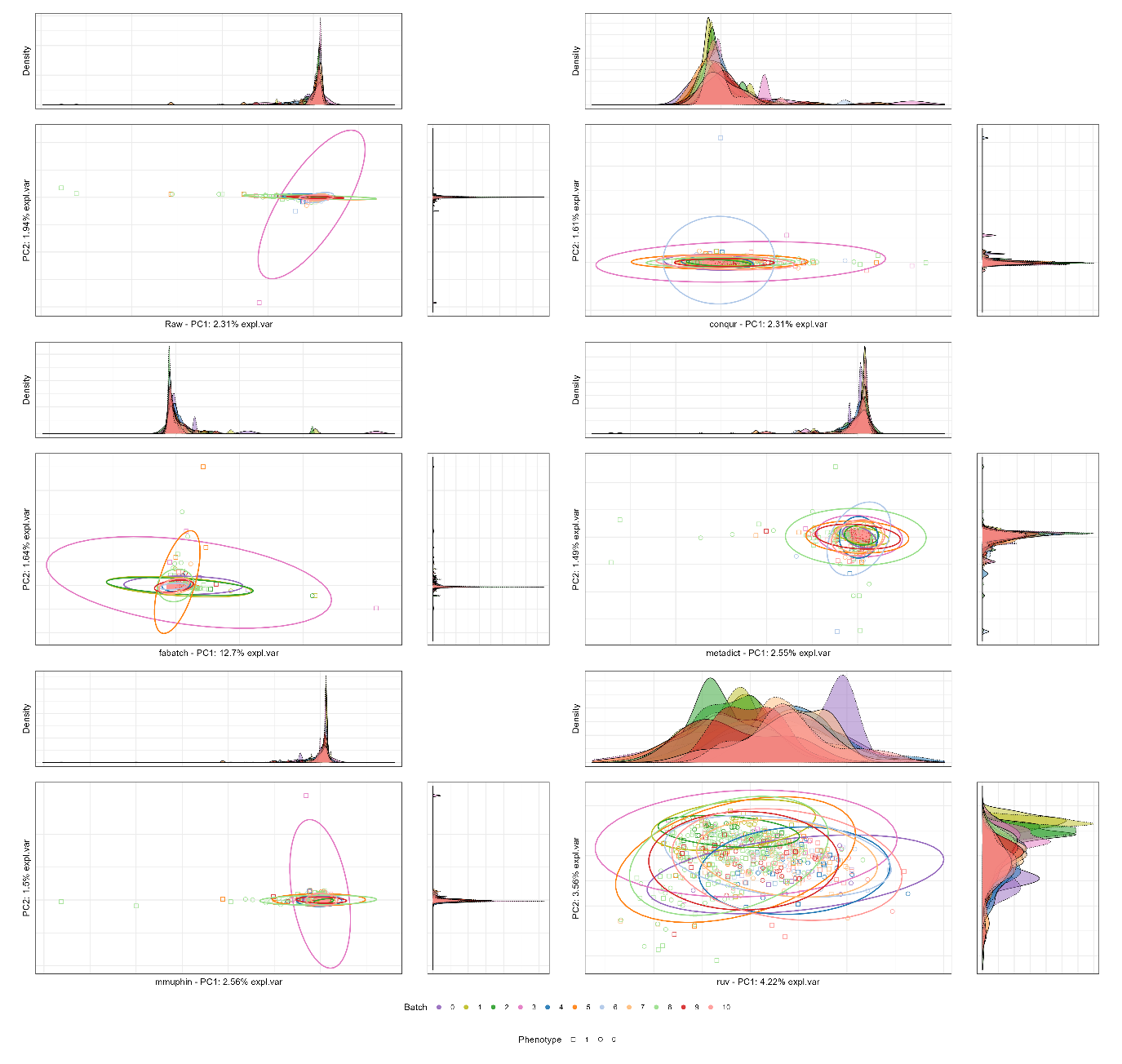
The plot was created by computing the AS and AUC for each method. Each point represents a method, plotted based on its AS and AUC values. A higher AUC and lower AS indicate better performance, with the ideal method showing the best possible AUC (close to 1) and the lowest AS (close to 0).

We plot these two metrics together because they provide complementary insights: AS reflects how well batch effects are controlled, and AUC reflects how well the model can distinguish between the classes. A method that performs well in both metrics is ideal, as it handles batch effects effectively while maintaining high classification performance.

Outliers in this plot may arise due to methods that exhibit either extreme alignment values (indicating poor batch effect handling) or poor classification performance (AUC significantly lower than others). These outliers could result from issues such as overfitting (where the model performs well on the training data but poorly on unseen data) or data leakage (where information from the test set leaks into the training process, artificially inflating performance).

To quantitatively rank the methods, both AUC and AS were normalized to the range [0, 1]. A combined score was computed as a weighted sum of the normalized AUC and 1 - AS, where higher AUC and lower AS are better. Methods with the highest combined score, which balances high classification performance (AUC) and low batch effect (AS), are ranked highest.

0-1

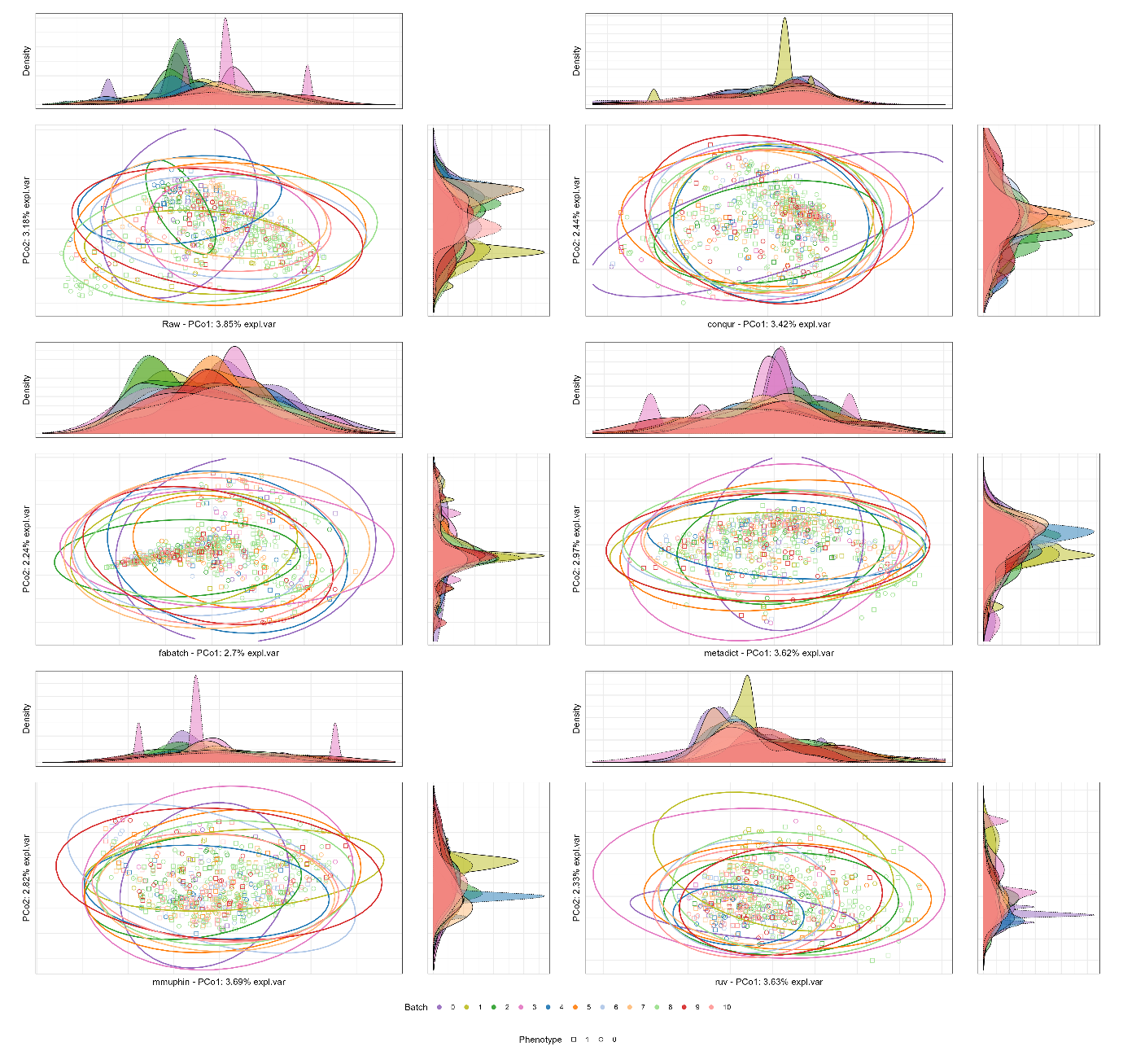
1. PCA

This figure displays the Principal Component Analysis (PCA) of microbiome data, where the first two principal components (PC1 and PC2) are plotted to visualize the separation of samples based on batch and phenotype characteristics. The PCA method was used to reduce the dimensionality of the data, retaining the most informative components that capture the variance across the features. Each point in the plot represents a sample, with colors indicating batch information and shapes representing phenotype groups. The % variance explained by each principal component is shown on the axis labels, giving insight into how much each component contributes to the total variance.

The plot was constructed by performing PCA on the normalized data, followed by plotting the first two principal components. The PCA plot helps identify patterns of separation, with clear clusters indicating strong grouping by batch or phenotype. Ideally, samples from different batches should overlap or show minimal separation after batch effect correction, suggesting that batch effects have been successfully removed and the biological variability is now the dominant factor.

To rank the methods based on PCA, we compute the distance between the centroids of each batch in the PCA space (using the first two principal components). Methods are ranked by the average Euclidean distance between batch centroids, with smaller distances indicating better correction. Methods with smaller centroid distances (indicating that batches are more similar after correction) are ranked higher, suggesting that the batch effects have been effectively minimized. Larger centroid distances imply greater batch separation, indicating poor batch effect correction, and these methods rank lower.

PCOA

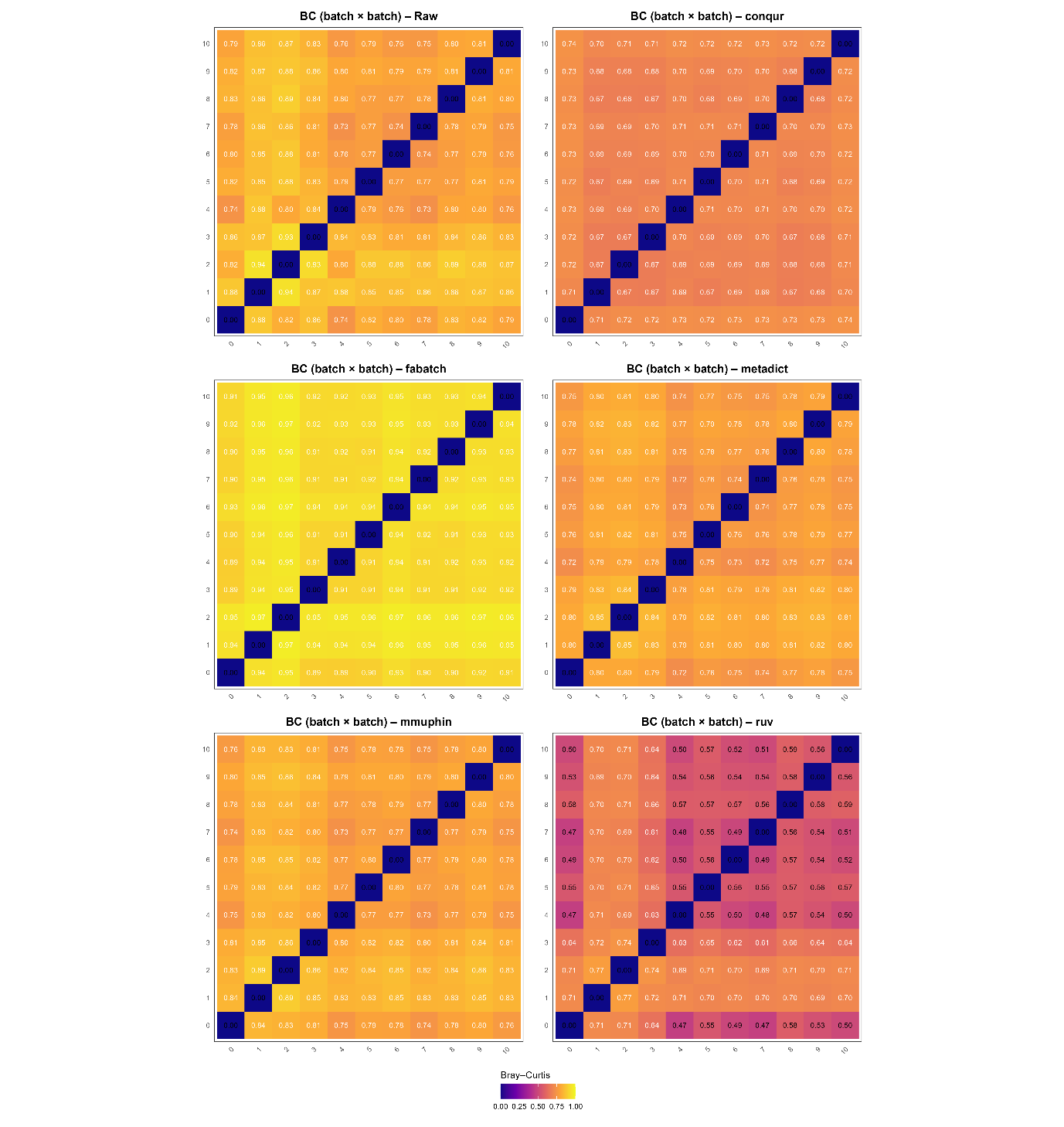


This figure displays the Principal Coordinates Analysis (PCoA) of microbiome data, where the first two principal coordinates (PCo1 and PCo2) are plotted to visualize the separation of samples based on batch and phenotype characteristics. The PCoA method is used to reduce the dimensionality of the data while retaining the most informative variance. Each point in the plot represents a sample, with colors indicating batch information and shapes (if applicable) representing phenotype groups. The % variance explained by each coordinate is shown on the axis labels, reflecting how much each coordinate captures of the total variance.

The plot was constructed by performing PCoA on the normalized data using a distance matrix calculated by Bray-Curtis, Aitchison (CLR), or Euclidean methods, depending on the data characteristics. PCoA helps identify patterns of separation, with clear clusters indicating strong grouping by batch or phenotype. Ideally, samples from different batches should overlap or show minimal separation after batch effect correction, suggesting that batch effects have been successfully removed and the biological variability is now the dominant factor.

To rank the methods based on PCoA, we calculate the distance between the centroids of each batch in the PCoA space (using the first two principal coordinates). Methods are ranked by the average Euclidean distance between these centroids. Methods with smaller centroid distances (indicating that batches are more similar after correction) rank higher, suggesting that batch effects have been effectively minimized. Larger centroid distances imply greater batch separation, indicating poor batch effect correction, and these methods rank lower.

1. Bray–Curtis Heatmap

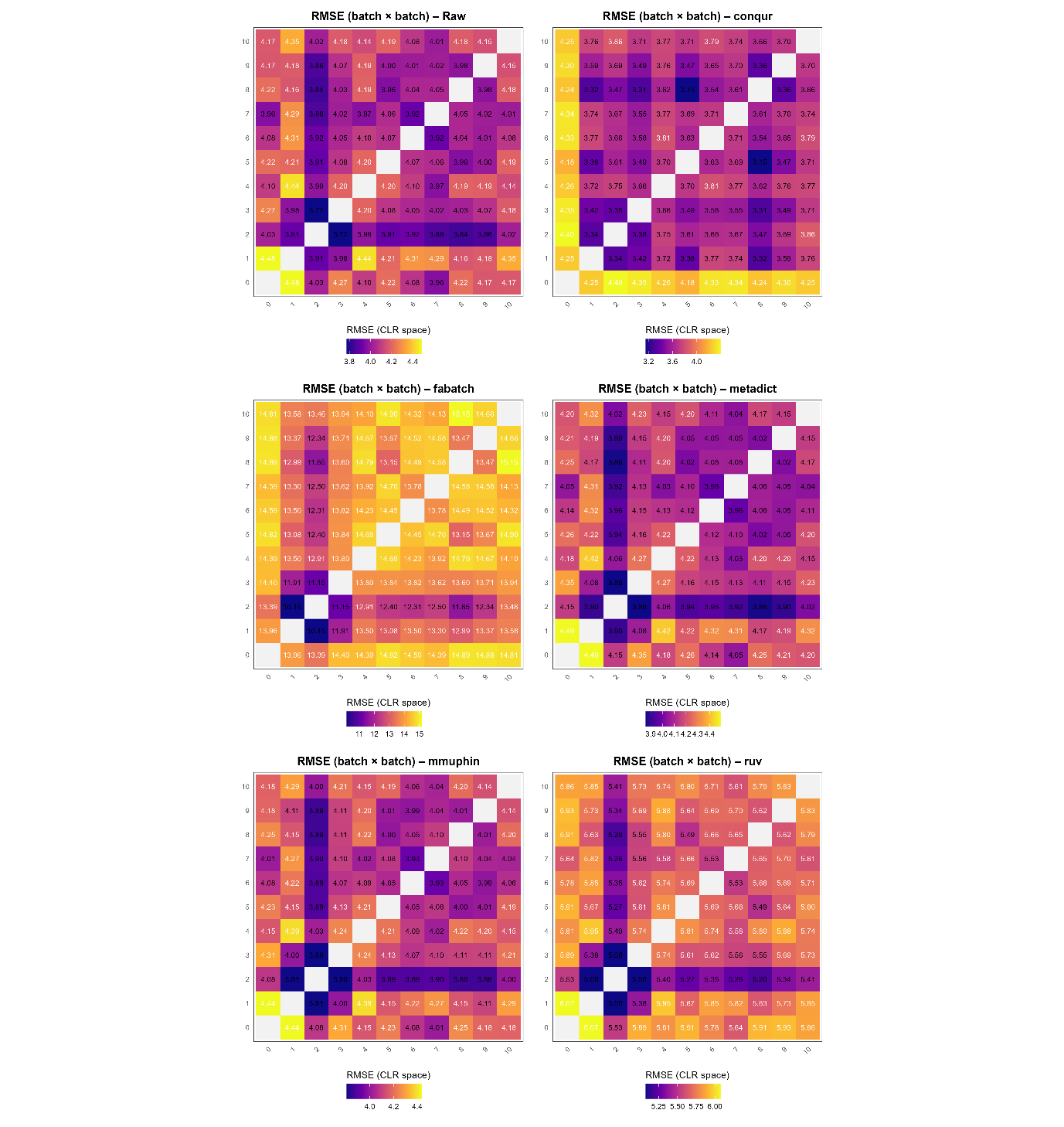


This figure shows the batch × batch Bray-Curtis dissimilarity metric, which quantifies the compositional differences between microbiome samples across batches. The plot was constructed by calculating pairwise Bray-Curtis dissimilarities for each batch, followed by averaging the dissimilarities within each batch combination, resulting in a heatmap. The colors in the plot represent the dissimilarity values, where darker shades indicate higher dissimilarity and lighter shades indicate lower dissimilarity between batches. A value close to 0 represents high batch cohesion (i.e., low batch effect), while a value approaching 1 signifies greater dissimilarity, suggesting stronger batch effects.

An ideal metric would show low dissimilarity values (closer to 0), indicating better batch handling, while higher dissimilarity values are undesirable as they suggest more significant batch effects. Outliers in this plot could arise from samples that exhibit significant differences in composition relative to other samples in the same batch, possibly due to technical errors, sample contamination, or biological variability.

To quantitatively rank the methods, we calculate the average Bray-Curtis dissimilarity for each method across the batch × batch distance matrix. Methods with lower average dissimilarity scores rank higher, indicating better batch correction and less batch effect.

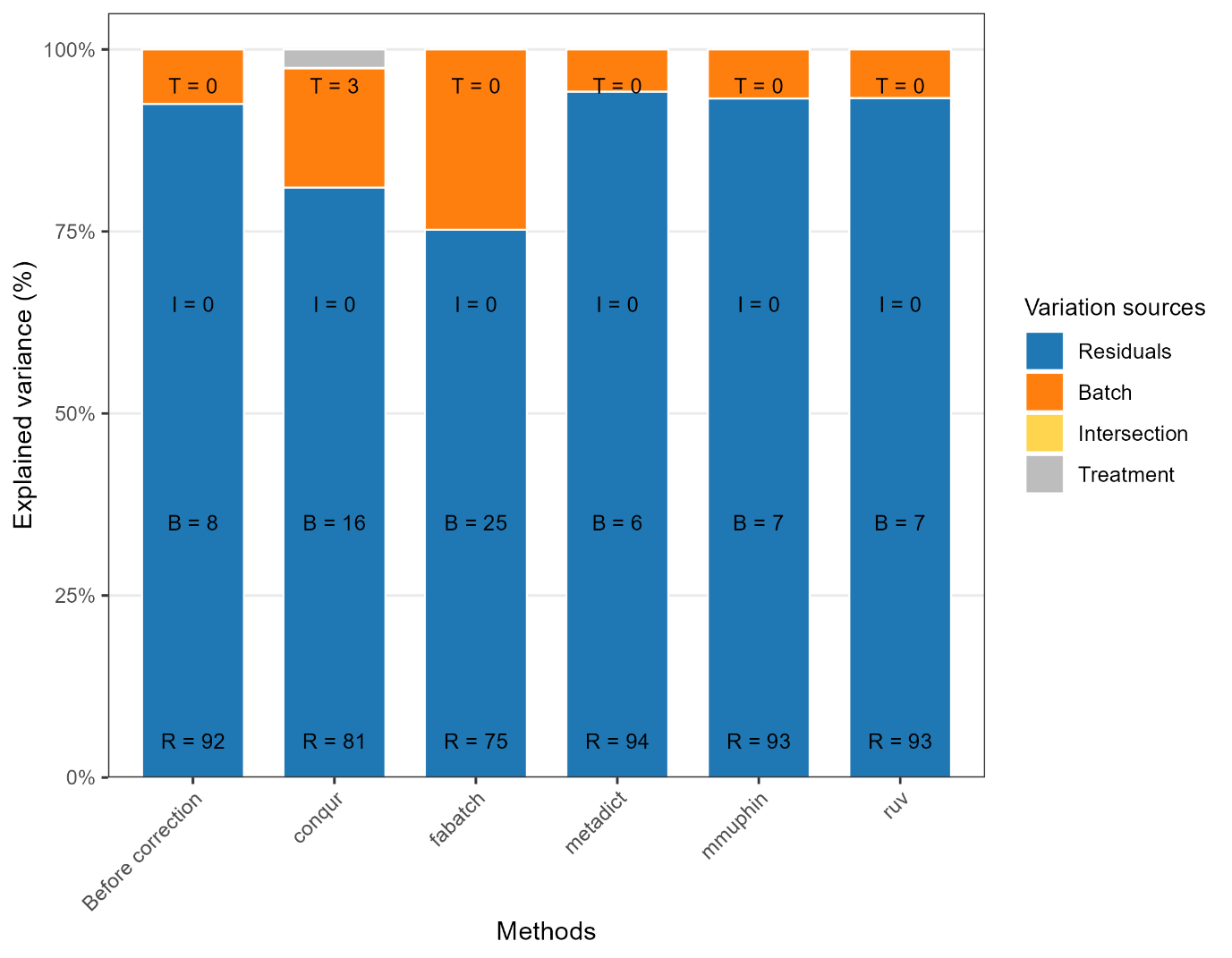
1. RMSE Heatmap



This figure displays the Root Mean Square Error (RMSE) between batches in the CLR-transformed space for multiple methods, where the heatmap represents the RMSE values between all pairwise batch comparisons. The RMSE metric quantifies the difference between batches. The RMSE values are derived from Euclidean distances between CLR-transformed samples, then normalized by the number of features to account for dimensionality.

The plot was created by computing the RMSE matrix for each method, where the matrix entries represent the average pairwise RMSE between samples in different batches. A lower RMSE between batches indicates that the samples are more similar across batches, suggesting better correction of batch effects. A higher RMSE suggests greater separation between batches, indicating that batch effects have not been fully corrected. The heatmap cells are annotated with the RMSE values, which are colored using a viridis scale to represent the magnitude of the RMSE, with darker colors indicating higher RMSE values and lighter colors indicating lower values.

To rank the methods, the mean RMSE values across all batch × batch comparisons were calculated for each method.

1. pRDA (intersection)

This figure shows the results of pRDA variance partitioning for multiple methods, where each stacked bar represents the proportion of explained variance due to Treatment, Batch, Intersection (shared variance), and Residuals. The plot was constructed by calculating the proportion of variance explained by each component in the CLR-transformed space using Principal Coordinate Analysis (pRDA).

Each method’s variance partition is divided into four parts:

Treatment represents the variation explained by the primary experimental factor.

Batch reflects the variation attributed to batch effects.

Intersection quantifies the shared variance between Treatment and Batch.

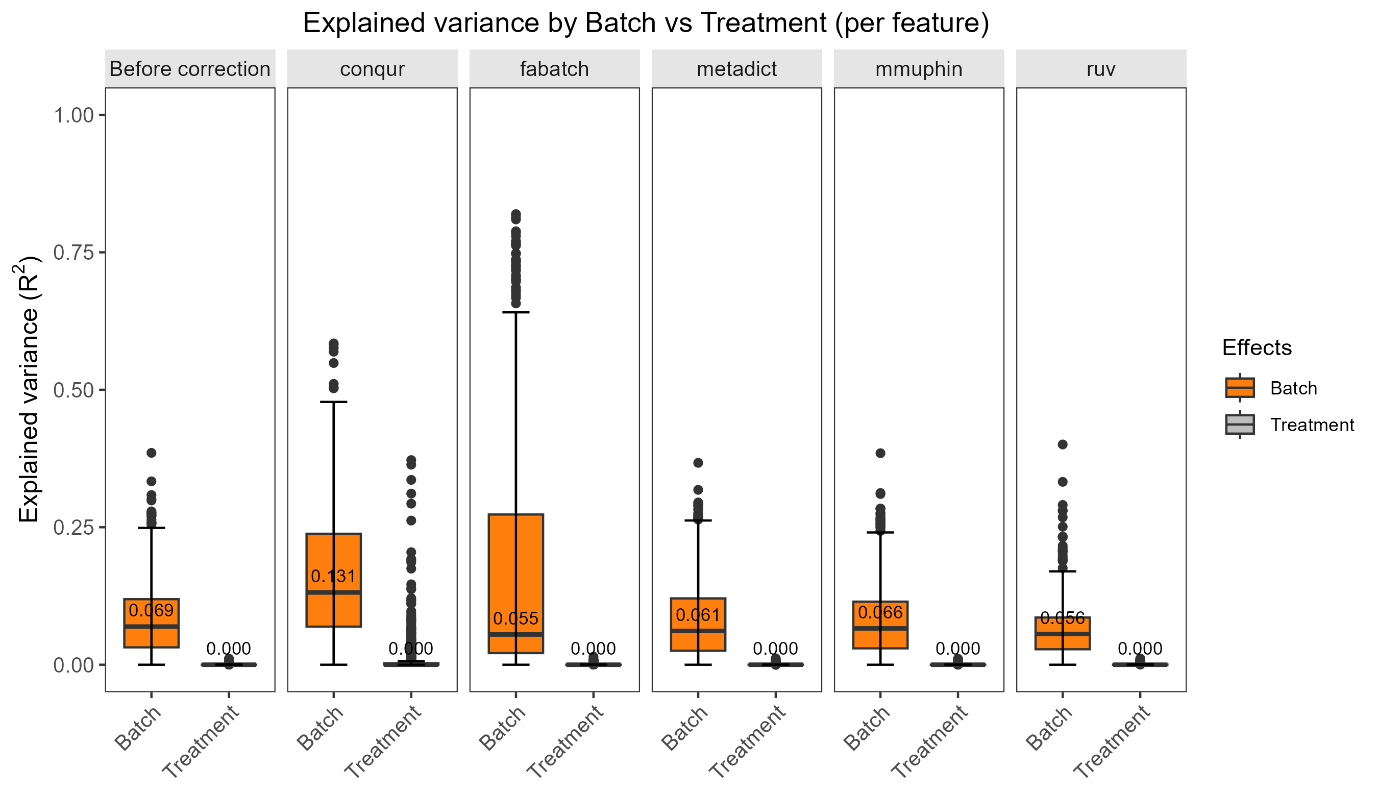
Residuals capture unexplained variation.

The stacked bars display these components, with labels fixed at specific positions on the bars for clarity. The Treatment and Batch components are visually highlighted, with their respective proportions clearly labeled. The plot also includes a legend that categorizes the sources of variance and their respective colors.

To rank the methods, we compute the proportion of variance explained by Treatment and Batch. Methods are ranked based on their Treatment fraction (higher is better) and Batch fraction (lower is better). Methods with larger Treatment fractions and smaller Batch fractions are considered to perform better, as they indicate that the treatment effect is stronger and batch effects are minimized.

In this ranking, methods that explain more variance through Treatment while minimizing the Batch fraction are ranked higher, suggesting better batch effect correction and a more pronounced treatment effect.

1. R2

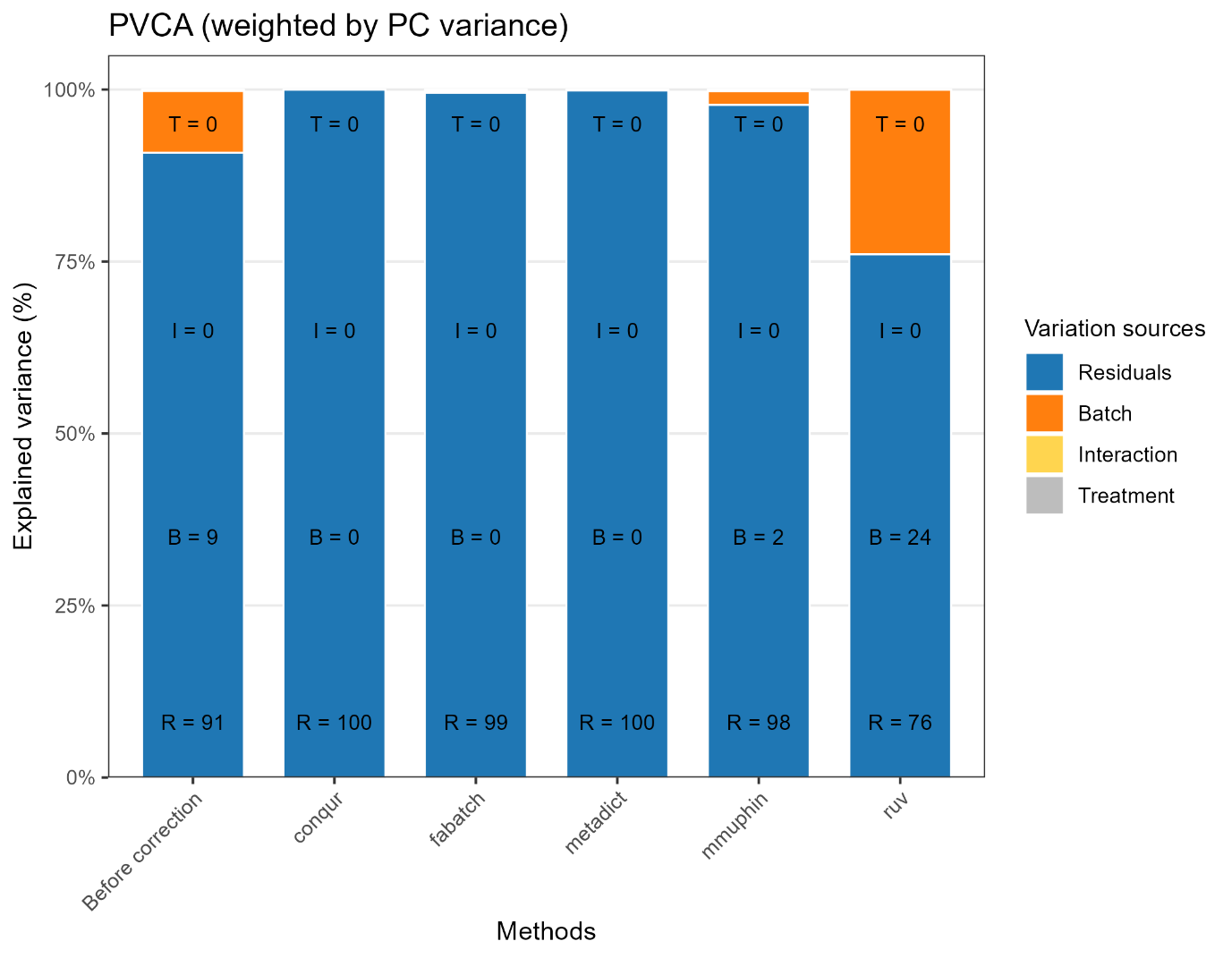


This figure presents the per-feature R² boxplots for the Batch and Treatment effects in the microbiome data, where each boxplot represents the variance explained by these two components for each method. The plot was generated using pRDA (Principal Coordinates Analysis) to partition the variance explained by Treatment and Batch for each feature. The R² values show the proportion of variance explained by the factors of interest, with higher values indicating stronger effects.

Each method is represented by a separate boxplot, and the results are broken down by the two components: Treatment and Batch. The boxplots display the median and interquartile range (IQR) of R² values for each component across features, with labels indicating the median R² for each effect. The color scheme highlights the Batch effect in orange and the Treatment effect in gray, allowing for easy comparison.

To rank the methods, the median R² values for both Treatment and Batch were computed. Methods are ranked based on their Treatment R² (higher is better) and Batch R² (lower is better). Methods with a larger Treatment R² and a smaller Batch R² are considered to perform better, as they suggest that the treatment effect is stronger and batch effects are minimized. The methods are ranked by maximizing the Treatment effect and minimizing the Batch effect, with higher ranks indicating better performance.

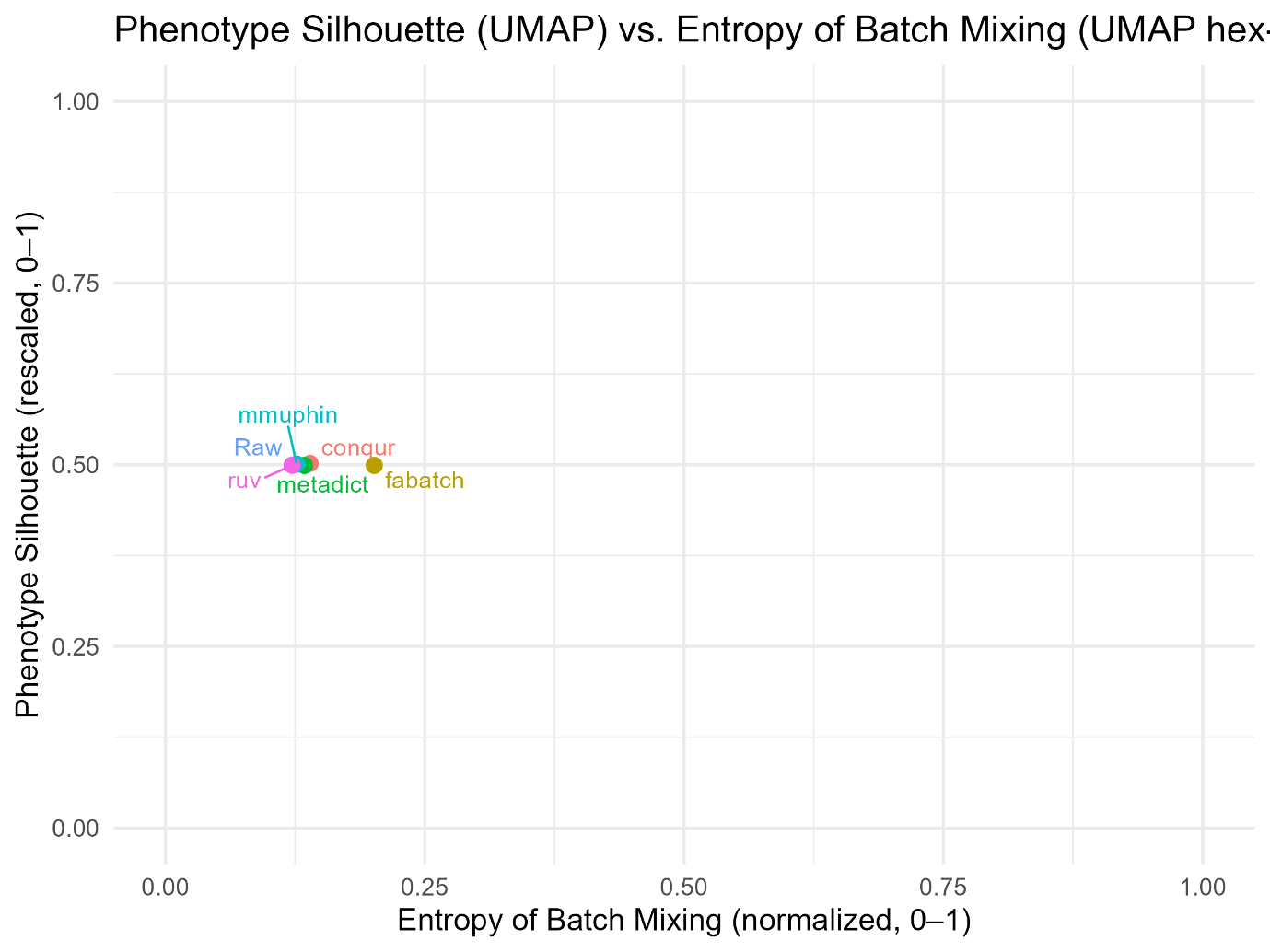
1. PVCA



This figure shows the Principal Variance Component Analysis (PVCA) for multiple methods, partitioning the variance explained by Treatment, Interaction, Batch, and Residuals. The plot visualizes how each method explains variation in the data across these four components, with the stacked bars representing the proportion of variance explained by each component. The Treatment component indicates the variance explained by the experimental factor of interest, while the Batch component reflects variation due to batch effects. Interaction captures the shared variance between Treatment and Batch, and Residuals represent unexplained variance.

To rank the methods, we compute the median fraction of variance explained by Treatment (higher is better) and Batch (lower is better). Methods that explain more variance through Treatment while minimizing Batch variance are ranked higher, as they suggest better separation between groups of interest and more effective control of batch effects.

1. Silhouette vs entropy\_of\_batch\_mixing



This figure presents a comparison of methods based on two key metrics: Silhouette (phenotype separation) and Entropy (batch mixing), both derived from UMAP (Uniform Manifold Approximation and Projection) embeddings. UMAP is a dimensionality reduction technique used for visualizing high-dimensional data by mapping it to lower dimensions (typically 2D or 3D). It preserves both the local and global structure of the data, making it particularly useful for clustering and identifying patterns in complex datasets. We plot these two metrics together because they provide complementary insights: Silhouette quantifies the clarity of the biological signal (phenotype), while Entropy reflects how well batch effects are controlled. Methods that achieve both high phenotype separation and low batch mixing are considered to be the most effective.

The Silhouette score measures the quality of clustering by quantifying the degree of separation between different phenotypes, with higher values indicating better separation. The Entropy score measures the degree of batch mixing, where lower values indicate better separation between batches and reduced batch effects.

The methods are ranked based on their performance in these two metrics. Methods with higher Silhouette scores (closer to 1) and lower Entropy scores (closer to 0) are ranked higher, as they indicate better phenotype separation and minimal batch mixing. This ranking is based on the premise that a good method will produce distinct phenotype clusters while minimizing the interference of batch effects.