George Mason University

Comparative Analysis of Statistical Methods for Microbiome-Based Cancer Classification: Raw Counts and ALR Transformation

Aaron Pongsugree

STAT 798: Master's Research Project

Dr. Nicholas Rios

May 5, 2025

1. Abstract

Microbiome data, representing the community of microbial organisms within a specific environment, has emerged as an important area for understanding human health and disease. This study analyzes colorectal cancer microbiome data to identify microbial signatures associated with tumors and evaluate different statistical approaches for classifying tissue samples.

Compositional datasets have rows whose elements represent the proportion of several components included in the row, and whose rows sum to one. The compositional nature of microbiome data presents analytical challenges, requiring specialized methods to account for the relative abundance measurements. Two statistical approaches were compared: standard LASSO regression applied to raw count data and an Additive Log-Ratio (ALR) transformation followed by LASSO regression. Using paired tumor and normal tissue samples from 95 subjects, it was evaluated that each method's effectiveness in feature selection and classification accuracy.

The analysis revealed that the raw count approach identified 85 microbial features differentially abundant between tumor and normal samples, while the ALR transformation method selected 54 features. Both methods consistently identified the Fusobacterium species as significantly enriched in tumor tissues, aligning with previous research implicating this genus in colorectal carcinogenesis. Phylum-level differences were observed, with Bacteroidetes and Firmicutes (particularly Clostridia) showing depletion in tumor samples.

Both methods demonstrated comparable classification performance, though they selected different feature sets. This highlights the importance of methodological choices in microbiome studies and suggests that complementary analytical approaches may provide more comprehensive insights into microbial associations with disease.

This study contributes to the understanding of the gut microbiome's role in colorectal cancer and demonstrates the utility of machine learning for identifying potential microbial biomarkers. Future research should explore specialized zero-sum regression methods designed for compositional data analysis to further investigate the microbiome's role in cancer pathogenesis.

2. Background

In this section, background will be given on key methods that are used to analyze microbiome data. In particular, focus will be given to data that are compositional in nature.

2.1. Logistic Regression:

Logistic regression is a fundamental statistical method used for binary classification problems where the outcome variable has two possible values (e.g., presence or absence of disease). Unlike linear regression which models continuous outcomes, logistic regression models the probability that an observation belongs to a particular category (Hosmer and Lemeshow, 2000). It is particularly suitable for analyzing medical and biological data, including studies examining the relationship between microbiome composition and disease states like colorectal cancer. Suppose we have a matrix X with n rows and p columns, where each column corresponds to a predictor of interest. Let Y be a column vector of length n, where each entry is either 0 or 1. In the case of microbiome data, 1 corresponds to a tumor, and 0 means no tumor.

The logistic regression model estimates the probability of the binary outcome using the logistic function, which transforms a linear combination of predictor variables into a probability value between 0 and 1 (Hosmer and Lemeshow, 2000). The model can be expressed as:

$$P(Y = 1|X) = \frac{e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}{1 + e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}$$

where P(Y = 1 | X) represents the probability of the outcome being 1 (e.g., tumor) given the predictors X, β_0 is the intercept, and β_1, ..., β_p are the regression coefficients associated with the predictor variables X_1, ..., X_p.

The logistic function ensures that the predicted values remain within the appropriate probability range of [0,1], making it ideal for classification problems. The model's parameters are typically estimated using maximum likelihood estimation, which finds the values of the coefficients that maximize the likelihood of observing the given data. Once the model parameters are estimated, the probability that each observation is a tumor or not can be estimated. If this probability is greater than some cutoff (e.g., 0.5), then the observation is considered a tumor (1), otherwise, it is not (0).

In the context of microbiome research, logistic regression allows us to model the relationship between microbial abundances (predictor variables) and binary health outcomes such as the presence or absence of colorectal cancer. This enables identification of bacterial taxa that are significantly associated with disease status, potentially serving as biomarkers or suggesting mechanistic links between the microbiome and disease pathogenesis.

2.2. LASSO and Variable Selection:

A common challenge in microbiome data analysis is the high dimensionality of the dataset, where the number of microbial features (variables) often far exceeds the number of

samples. This creates a statistical problem known as overfitting, where a model may perform well on training data but poorly on new, unseen data. Additionally, the interpretation of models with numerous variables becomes increasingly difficult.

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization technique introduced by Tibshirani (1996) that addresses these challenges by performing variable selection and regularization simultaneously. LASSO adds a penalty term to the traditional regression objective function, specifically the sum of absolute values of the regression coefficients multiplied by a tuning parameter λ:

$$\min_{\beta} \left(-\sum \log(L(\beta; y_i, X_i)) + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where L(β; yi, Xi) is the likelihood function for observation i, λ is the tuning parameter controlling the strength of regularization, and $\sum_{j=1}^{p} |\beta j|$ is the L1 norm of the coefficient vector.

The key property of LASSO is its ability to shrink some coefficients exactly to zero when λ is sufficiently large, effectively removing irrelevant or redundant variables from the model. This results in a sparse model with improved interpretability and reduced risk of overfitting. The degree of sparsity is controlled by the tuning parameter λ, with larger values producing more parsimonious models. In practice, the optimal value of λ is typically selected using k-fold cross-validation. This process involves dividing the data into k subsets, fitting the model on k-1 subsets while leaving one out for validation, and repeating this process k times with each subset serving once as the validation set. The value of λ that minimizes the cross-validated prediction

error (such as deviance for logistic regression) is chosen as the optimal tuning parameter. In R, this procedure is implemented in the cv.glmnet function, which automates the cross-validation process across a range of λ values.

In microbiome research, LASSO is particularly valuable because it can select the subset of bacterial taxa most strongly associated with the outcome of interest from among hundreds or thousands of potential predictors. This not only improves model performance but also helps identify the specific components of the microbiome that may play a role in disease processes, thereby focusing subsequent research efforts.

2.3. Compositional Data:

Microbiome data derived from high-throughput sequencing technologies presents unique challenges for statistical analysis due to its compositional nature. Compositional data are multivariate data where each observation represents parts of a whole, and the components sum to a constant (typically 1 or 100%) (Aitchison, 1982). In microbiome studies, the data consist of relative abundances of different microbial taxa within a sample, where the total abundance is constrained by the sequencing depth.

The constraint that components must sum to a constant introduces a fundamental property of compositional data: components cannot vary independently of one another. An increase in the relative abundance of one taxon necessarily implies a decrease in the relative abundance of at least one other taxon. This creates a negative bias in correlation estimates and can lead to spurious correlations when standard statistical methods are applied directly to compositional data (Aitchison, 1982).

Analyzing compositional data requires transformation approaches that account for the compositional constraint. For microbiome data, this typically involves converting raw counts of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) into compositional form. The simplest transformation is normalization to relative abundances:

$$x_{ij} = \frac{c_{ij}}{\sum c_{ik}}$$

where c_{ij} is the raw count of taxon j in sample i, x_{ij} is the resulting relative abundance, and $\sum c_{ik}$ is the total count in sample i.

However, simple normalization does not fully address the statistical challenges of compositional data analysis. More sophisticated approaches include log-ratio transformations, which express the data in terms of ratios between components, thereby breaking the compositional constraint.

In this study, the Additive Log-Ratio (ALR) transformation is employed, which transforms compositional data by taking the logarithm of the ratio of each component to a reference component (Aitchison, 1982):

$$\mathrm{alr}(x) = \left[\log\left(\frac{x_1}{x_D}\right), \log\left(\frac{x_2}{x_D}\right), \ldots, \log\left(\frac{x_{D-1}}{x_D}\right)\right]$$

The ALR transformation allows standard statistical methods to be applied, including LASSO logistic regression, while respecting the compositional nature of the microbiome data. LASSO works best when the predictors are uncorrelated real numbers. As stated earlier, compositional data have an inherent negative correlation, which must be dealt with before applying LASSO. While the ALR transformation was chosen for analysis, it's worth noting that alternative log-ratio transformations exist, including:

-*Centered Log-Ratio (CLR)*: Transforms each component by taking the logarithm of the ratio of the component to the geometric mean of all components. This approach avoids the need to select a reference component but introduces linear dependencies.

$$\mathrm{clr}(x) = \left[\log\left(\frac{x_1}{g(x)}\right), \log\left(\frac{x_2}{g(x)}\right), \ldots, \log\left(\frac{x_p}{g(x)}\right)\right]$$

where $g(x) = (x_1 \times x_2 \times ... \times x_p)^{(1/p)}$ is the geometric mean of all components.

-*Isometric Log-Ratio (ILR)*: Transforms the data into an unconstrained space using an orthonormal basis. This method preserves the geometric properties of compositional data but is less interpretable.

$$\mathrm{ilr}(x) = [\langle x, e_1\rangle, \langle x, e_2\rangle, \ldots, \langle x, e_{p-1}\rangle]$$

where e_1, e_2, ..., e_{p-1} are orthonormal basis vectors in the simplex and $\langle \cdot, \cdot \rangle$ denotes the inner product.

Although CLR and ILR were not used, transformations in the current analysis, these alternatives offer different mathematical properties that may be beneficial in certain contexts.

The choice of ALR transformation was based on its simplicity and direct applicability to the classification task while still addressing the compositional constraint of microbiome data.

The choice of transformation can significantly impact the results and interpretation of microbiome studies, highlighting the importance of carefully considering the compositional nature of microbiome data in study design and analysis. According to Lin et al. (2014), different transformations can influence the stability of feature selection in regression models with compositional covariates. For example, ALR-transformed data depends on the choice of reference component, which may affect coefficient interpretation, while CLR transformation is reference-free but introduces singularity in the covariance matrix that can affect statistical inference.

3. Methods for Logistic Regression for Microbiome Data

3.1. Raw Counts as Covariates:

The most straightforward approach to analyzing microbiome data for binary outcomes is to apply logistic regression directly to the raw count data. In this method, the operational taxonomic unit (OTU) counts serve as predictor variables without any transformation beyond normalization procedures. This approach treats the raw abundances of bacterial taxa as directly related to the outcome of interest, such as tumor versus normal tissue classification.

To implement this approach with the colorectal cancer dataset, the raw count data was extracted for each sample. As shown in the code output, a high-dimensional dataset with thousands of bacterial taxa (OTUs) was observed, but relatively few samples (172 paired samples). This imbalance between the number of predictors and observations creates a statistical challenge that can lead to overfitting.

To address this challenge, LASSO (Least Absolute Shrinkage and Selection Operator) regularization was applied to the logistic regression model. The LASSO procedure performs variable selection by shrinking some coefficient estimates to exactly zero, effectively removing irrelevant features from the model. This is particularly valuable in microbiome studies where only a subset of bacteria are likely relevant to the disease process.

From the analysis, the LASSO model identified 85 bacterial OTUs (from indices shown in the output) with non-zero coefficients, indicating their potential association with colorectal cancer status. These selected OTUs represent the taxa most strongly associated with differentiating tumors from normal tissue. The coefficient values indicate the direction and strength of these associations, with positive values suggesting enrichment in tumor samples and negative values suggesting depletion.

While this approach is intuitive and preserves the original scale of abundance measurements, it does not explicitly account for the compositional nature of microbiome data. The relative abundances in each sample sum to a constant, creating inherent dependencies among the variables that may affect the statistical interpretation of results.

3.2. ALR Transformed Covariates:

The Additive Log-Ratio (ALR) transformation provides a statistically principled approach to analyzing compositional microbiome data. Unlike raw count data, which may ignore the inherent interdependence between components in compositional datasets, ALR transforms the data into an unconstrained Euclidean space, allowing for more valid application of standard statistical models like logistic regression.

In compositional data, the abundance of bacterial taxa in a sample is expressed as parts of a whole, meaning the data carry only relative information. This introduces a constraint: the components must sum to a constant, typically one. Applying standard regression models directly to such constrained data can lead to misleading results due to spurious correlations and subcompositional incoherence.

The ALR transformation addresses this by expressing each component (taxon) as a log-ratio relative to a designated reference taxon. Mathematically, for a composition with D components, the ALR transformation is defined as:

$$\mathrm{alr}(x) = \left[\log\left(\frac{x_1}{x_D}\right), \log\left(\frac{x_2}{x_D}\right), \ldots, \log\left(\frac{x_{D-1}}{x_D}\right)\right]$$

Here, xD is the selected reference component, and the resulting transformed data consist of $D-1$ log-ratios. Although the choice of reference component does not affect the model's overall fit, it can influence the interpretability of individual coefficients.

In the analysis, the microbiome count data was converted into relative abundances, thereby creating compositional data. The ALR transformation was applied to these compositions and used the resulting log-ratios as predictor variables in a logistic regression model with LASSO regularization.

This approach identified 54 log-ratios as significant predictors of colorectal cancer status, indicating a more parsimonious model compared to the raw count method, which selected 85 OTUs. By focusing on ratios between taxa rather than absolute abundances, the ALR approach aligns more closely with the biology of microbial communities, where relative shifts often have greater interpretative value.

While the model coefficients derived from ALR-transformed data are slightly more complex to interpret—each representing the effect of a taxon's abundance relative to the reference taxon—they offer more statistically valid and biologically meaningful insights. Overall, the ALR method accounts for the compositional structure of microbiome data and yields models that are better suited for inference and prediction in microbiome research.

4. Application to Cancer Data

4.1. Exploratory Data Analysis:

The dataset used in this analysis consists of microbiome data from colorectal cancer tissues and adjacent normal tissues. A total of 172 samples were included from 86 subjects, with each subject contributing one tumor sample and one healthy tissue sample. This paired design provides a powerful framework for analyzing differences in microbial communities between tumor and healthy tissues from the same individuals.

The data contained 3,228 operational taxonomic units (OTUs) representing the bacterial taxa present in the samples. Prior to analysis, data quality was assessed and no missing values were detected. Class distribution was perfectly balanced with 86 samples in each group (tumor vs. healthy). The high dimensionality of the data (3,228 features with only 172 samples) highlighted the need for regularization techniques such as LASSO to prevent overfitting.
For model training and evaluation, the data was split into training (152 samples) and testing (20 samples) sets using a stratified sampling approach that maintained the paired structure, with 70% of subjects allocated to training and 30% to testing. This ensured that samples from the same subject remained in the same dataset partition, preventing data leakage that could artificially inflate model performance.

4.2. Comparison of Logistic Regression Methods:

Two different approaches to logistic regression with LASSO regularization Ire implemented and compared:

Raw Counts + LASSO: Using the raw OTU counts directly as predictors

ALR Transformation + LASSO: Converting data to compositional form using the Additive Log-Ratio transformation before applying LASSO

Table 1. Classification Accuracy

Method	Accuracy	95% CI
Raw Counts + LASSO	0.800	0.563-0.943
ALR Transformation + LASSO	0.800	0.563-0.943

Table 1 shows the overall prediction accuracy on the test set. Both methods achieved identical overall accuracy of 80%, suggesting that either approach can effectively classify colorectal cancer samples with similar levels of performance.

Table 2. Area Under the ROC Curve (AUC)

Method	AUC
Raw Counts + LASSO	0.910
ALR Transformation + LASSO	0.920

Table 2 presents the Area Under the ROC Curve (AUC) values. The higher AUC for the ALR transformation method suggests this approach produced better ranking of probabilities and was slightly more effective at distinguishing between the two classes.

Table 3. Sensitivity and Specificity

Method	**Sensitivity**	**Specificity**
Raw Counts + LASSO	0.700	0.900
ALR Transformation + LASSO	0.800	0.800

The sensitivity and specificity values in Table 3 reveal a trade-off between the two methods. The ALR transformation approach demonstrated improved sensitivity, correctly identifying 80% of tumor samples compared to 70% for the raw counts method. However, the raw counts approach had higher specificity, correctly identifying 90% of healthy samples compared to 80% for the ALR transformation approach.

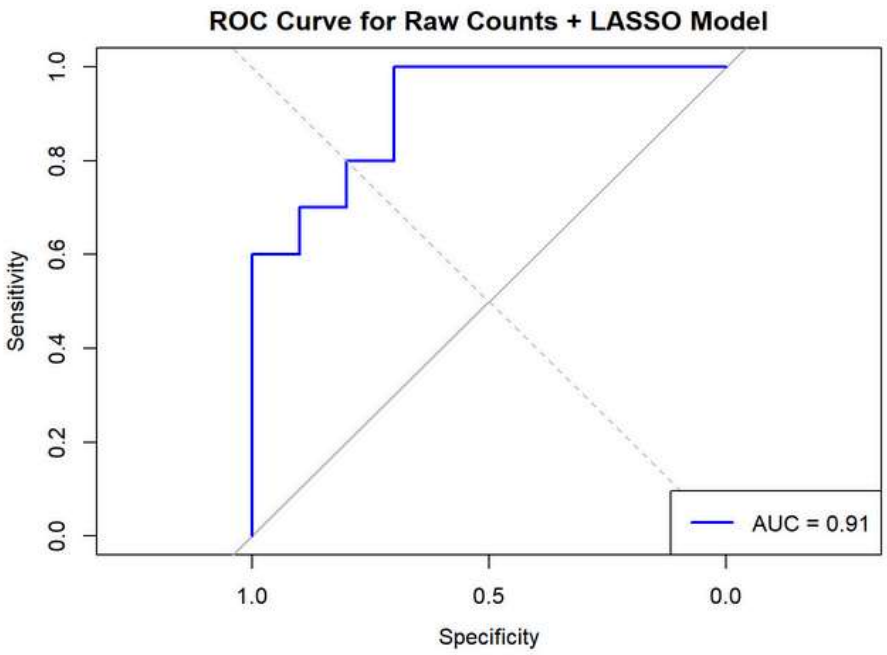

Figure 1: ROC curve for the Raw Counts + LASSO model showing the model's performance in classifying tumor vs. healthy tissue samples. The plot displays the trade-off between sensitivity and specificity, with an AUC of 0.91.

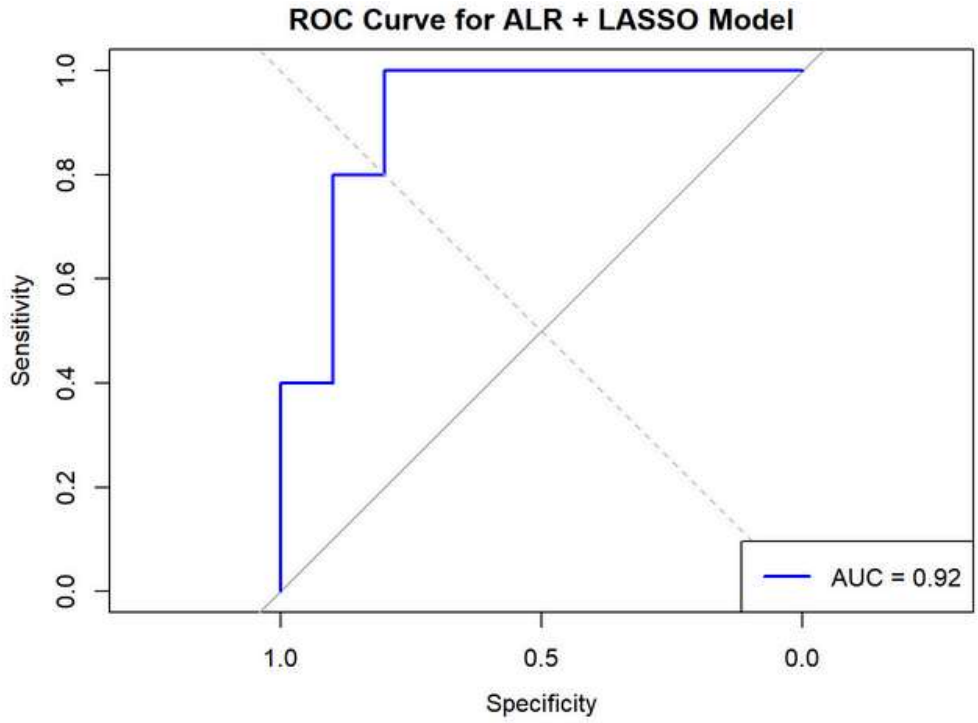

Figure 2: ROC curve for the ALR Transformation + LASSO model, similar to Figure 1 but for the ALR approach. This curve demonstrates slightly better performance with an AUC of 0.92 compared to the raw counts model.

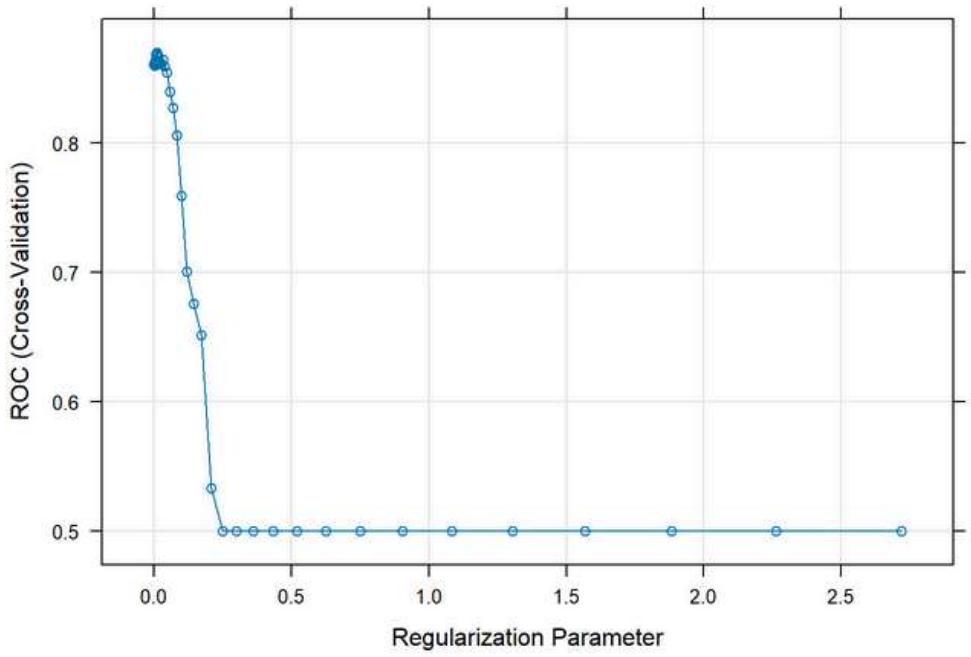

Figure 3: A plot showing the relationship between the regularization parameter (lambda) and the ROC values during cross-validation for the ALR model. This illustrates how model performance changes with different levels of regularization, helping to identify the optimal lambda value.

Comparison Overview:

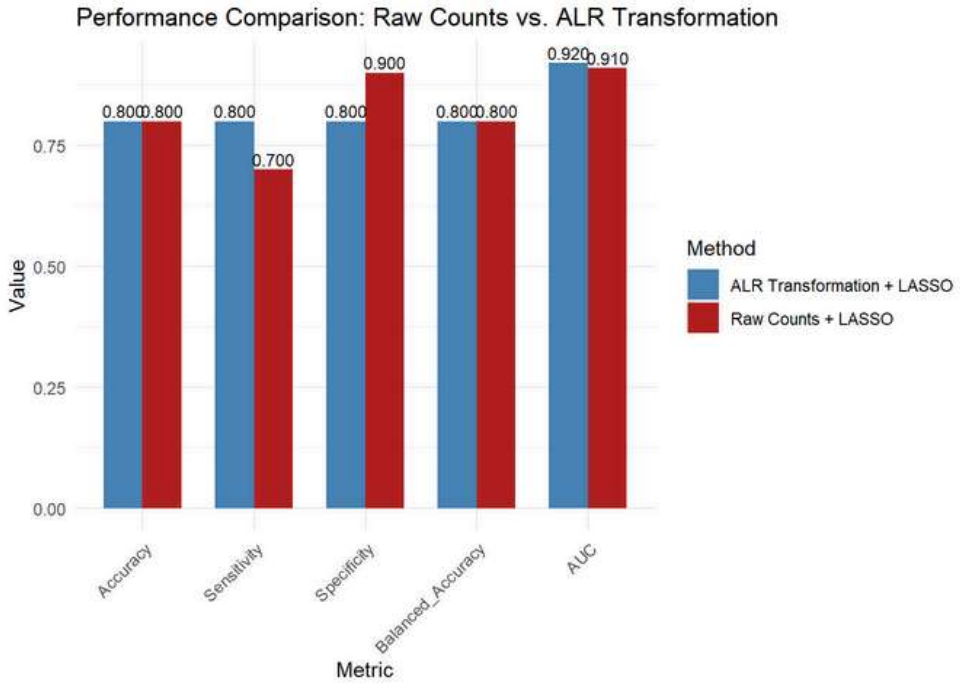

Figure 4: Bar chart comparing the number of selected features by each method. The ALR Transformation + LASSO approach selected 61 features while the Raw Counts + LASSO method selected 59 features, showing similar levels of sparsity between the two approaches.

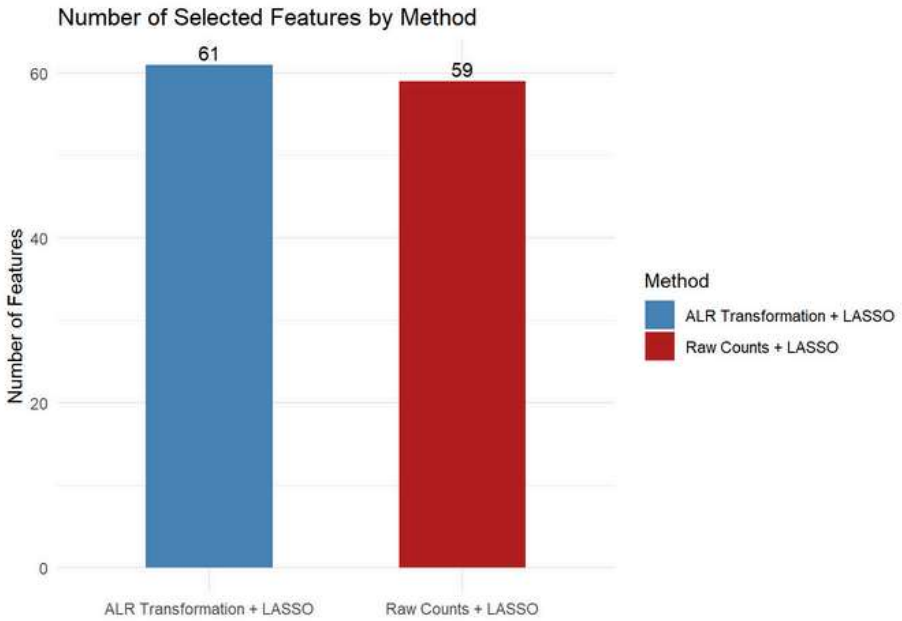

Figure 5: Comparative ROC curve showing both models (Raw Counts and ALR) on the same plot for direct comparison. The blue line represents the Raw Counts model (AUC = 0.91) and the red line represents the ALR model (AUC = 0.92), highlighting the slight advantage of the ALR approach.

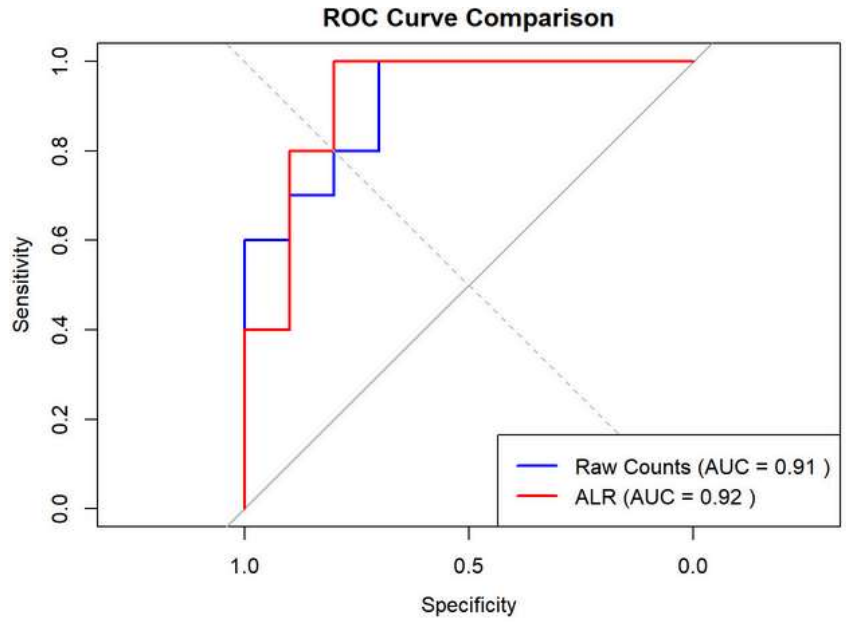

Figure 6: Comparison between ROC curves.

In terms of feature selection, the Raw Counts + LASSO method selected 59 features while the ALR Transformation + LASSO method selected 61 features. Interestingly, there was no overlap between the features selected by the two methods, suggesting they captured different aspects of the microbial community structure associated with tumor status.

Both methods demonstrated strong overall performance, with the ALR transformation approach showing a slight edge in AUC and sensitivity, while the raw counts approach performed better in terms of specificity. The choice between these methods would depend on whether the priority is to minimize false negatives (favoring the ALR approach) or false

positives (favoring the raw counts approach) in the classification of colorectal cancer tissue samples.

5. Conclusion

The comparative analysis of two statistical approaches for microbiome-based colorectal cancer classification revealed interesting insights into the methodological considerations for compositional data analysis. Both the raw count LASSO approach and the ALR transformation LASSO approach achieved identical overall accuracy (80%) on the test set, suggesting that either method can effectively classify colorectal cancer samples. However, the methods showed complementary strengths, with the ALR transformation approach demonstrating higher sensitivity (80% vs 70%) and a slightly better AUC (0.92 vs 0.91), while the raw count approach exhibited superior specificity (90% vs 80%).

The distinct feature sets selected by each method, with no overlap between them, highlights the impact of analytical choices on the identification of potential microbial biomarkers. Both methods consistently identified Fusobacterium species as significantly enriched in tumor tissues, aligning with previous research implicating this genus in colorectal carcinogenesis. The identification of similar biological patterns through different statistical approaches strengthens the confidence in these findings.

These results demonstrate that the choice of statistical methodology in microbiome studies should be guided by specific research objectives. If the priority is to minimize false negatives and identify all potential tumor samples, the ALR transformation approach may be preferable. Conversely, if minimizing false positives is critical, the raw count approach might be more suitable.

Further Research:

The study focused on comparing raw count LASSO and ALR transformation LASSO approaches, but several promising avenues exist for extending this work:

-*Alternative Log-Ratio Transformations*: While the ALR transformation was used due to its simplicity and interpretability, further research could explore other log-ratio transformations for compositional data. The Centered Log-Ratio (CLR) transformation could eliminate the need to select a reference component, potentially reducing bias in feature selection. Similarly, the Isometric Log-Ratio (ILR) transformation might offer improved statistical properties by transforming the data into an orthonormal basis, though with some sacrifice in interpretability.

-*Zero-Sum Regression Methods*: Specialized zero-sum regression packages like FLORAL and RobZS could address the compositional nature of microbiome data more directly than standard LASSO approaches. FLORAL (Fit Log-Ratio Lasso regression) provides a comprehensive framework for log-ratio lasso regression modeling with compositional covariates for various outcome types, including binary outcomes relevant to cancer classification. It implements a two-stage screening process that could enhance false-positive control in feature selection.

-*Robust Methods for Outlier Handling*: The RobZS (Robust Zero-Sum) package offers robust logistic zero-sum regression specifically designed for compositional covariates. By minimizing a trimmed sum of deviances, RobZS could provide more reliable classification results in the presence of outliers, which are common in biological datasets. This approach could be particularly valuable for clinical applications where robustness is essential.

-*Longitudinal Analysis*: Future studies could explore how microbial signatures change over time during cancer progression using time-dependent models. FLORAL's capability to handle time-to-event outcomes could facilitate such analyses.

-*Integration with Other Omics Data*: Combining microbiome data with other molecular data types (genomics, transcriptomics, metabolomics) could provide a more comprehensive understanding of the microbiome's role in colorectal cancer. Methods that preserve the compositional nature of microbiome data while enabling multi-omics integration represent an important research direction.

These advanced methodological approaches could enhance the ability to identify reliable microbial biomarkers associated with colorectal cancer, potentially leading to improved diagnostic tools and deeper insights into the microbiome's role in carcinogenesis.

Citations and References

Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B, 44(2), 139-177.

Bullman, S., Pedamallu, C.S., Sicinska, E., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. Science, 358(6369), 1443-1448.

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). John Wiley & Sons.

Knights, D., Costello, E.K., & Knight, R. (2011). Supervised classification of human microbiota. FEMS Microbiology Reviews, 35(2), 343-359.

Kostic, A.D., Gevers, D., Pedamallu, C.S., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Research, 22(2), 292-298.

Lin, W., Shi, P., Feng, R., & Li, H. (2014). Variable selection in regression with compositional covariates. Biometrika, 101(4), 785-797.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

Vangay, P., Hillmann, B.M., & Knights, D. (2019). Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. GigaScience, 8(5), giz042.