Supplementary B

The Statistical Method Details

Supplementary B – Statistical Analysis

The details for statistical analysis are provided in this Supplementary Material.

Supplementary B.1 – Variable Selection with SIAMCAT

Statistical analysis of microbiome composition data was performed using R (v4.3.1, www.r-project.org) and RStudio (v2023.06.2, www.posit.co) with the SIAMCAT library. [1] Relative abundance data were analyzed separately at five taxonomic levels: phylum, class, order, family, and genus. The primary outcome was the classification of samples into CRC and AP groups. Five unsupervised feature selection methods were applied at each taxonomic level with a cutoff value of 0.001: abundance (AB), cumulative abundance (CA), prevalence (PR), variance (VA), and pass-through filtering (PA, retaining all features). Filtered features were then normalized using two approaches: (1) the rank-unit method (RU), which ranks each feature before scaling, and (2) logarithmic transformation (LU) followed by normalization. These preprocessing steps ensured standardized input for downstream statistical and machine learning analyses.

The dataset was randomly split into training and testing sets, and five-fold cross-validation was performed with five resampling iterations. Three machine learning methods were applied: logistic regression with LASSO and RIDGE penalties, and Random Forest (RF). Each model was trained on the training data, and five-fold cross-validated predictions were generated for the testing sets. Model performance was evaluated using receiver operating characteristic (ROC) curves, and the area under the ROC curve (AUC) was calculated and compared across models. For interpretation and visualization, several plots were generated: a heatmap of metagenomic features standardized as Z-scores, bar plots representing feature weights and robustness, and boxplots depicting the minimum inclusion frequency of each feature across models (0.001 for LASSO and RIDGE, 0.01 for RF). These visualizations facilitated identification of robust and predictive microbial features for colorectal cancer classification.

Supplementary B.2 – Spike and Slab Logistic Regression

Bayesian variable selection using spike-and-slab logistic regression has been widely applied in medical research, including functional magnetic resonance imaging (fMRI) [2] and microbiome data analysis [3-5].

In this study, we consider a binary response variable $E_i \in \{0,1\}(1 \le i \le n)$, representing colorectal cancer (CRC) versus adenomatous polyps (AP). For each patient, covariates corresponding to a specific sampling site and taxonomic level are denoted by $X_i \in \mathbb{R}^p$. In the following standard logistic regression model with coefficient $\beta \in \mathbb{R}^p$, the goal is to find the non-zero β:

$$P(E_i = 1 \,|X_i, \beta) = \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} \quad , \qquad i = 1,2, \dots, n \qquad (1)$$

In this regard, the following spike and slab priors for β are used:

$$\beta_j | \lambda_j \sim \left(1 - \lambda_j\right)\alpha_0 + \lambda_j N(0, \tau^2) \qquad 1 \le j \le p \qquad (2)$$

$$P(\lambda_j = 1) = 1 - P(\lambda_j = 0) = q_j \qquad (3)$$

The λ_j $(1 \leq j \leq p)$ is the latent binary variable for the β_j with $\lambda_j = 1$ for active or signal β_j (slab part) and $\lambda_j = 0$ otherwise or noise (spike at zero) assuming independent Bernoulli priors. The hyperparameters are $\tau > 0$ to control the variance of slab part and the inclusion probability $0 < q_j < 1$ controls the signals. The α_0 is the point mass at zero. The combination of three Metropolis-Hastings updates were used as MCMC sampling methods in this model. The number of iterations is 50,000 with burnout samples equals 5,000 updates and summary statistics including mean, standard deviation, and inclusion probability across all iterations are calculated. The only covariates with an inclusion probability greater than 0.05 are selected. Five-fold CV with five repeated samples was used and the AUC was reported based on the full models. All calculations are performed using the *BoomSpikeSlab* package in R [6, 7]

Supplementary B.3 – Concordance Plot

For each algorithm, the features corresponding to the highest AUC across five filtration methods and two normalization approaches were identified and visualized using a concordance plot. The plot consists of four overlapping ellipses arranged in a star-like configuration, with each ellipse representing one algorithm. Feature names are displayed within the corresponding ellipse, and features shared across multiple algorithms appear in the overlapping regions. Features selected by all four algorithms are located at the center of the plot, indicating high consistency, robustness, and potential biological relevance. This approach extends prior work in microbiome data analysis, where three-circle concordance plots were previously used [8]. All visualizations were generated using the ggVennDiagram package in R.

Supplementary B.4 – Bayesian Model Averaging

Bayesian Model Averaging (BMA) combines the predictive distributions of multiple candidate models into a weighted average, accounting for model uncertainty in parameter estimation and future predictions. In this study, BMA was applied to logistic regression models using Bayesian Adaptive Sampling (BAS), which efficiently explores the model space through stochastic or deterministic sampling without replacement from the posterior distribution. Analyses were performed in R using the BAS and BMA packages. The logistic regression framework was formulated as a generalized linear model (GLM) with a binary outcome representing CRC versus AP. BMA was employed to identify robust predictive features by averaging across models with different combinations of covariates, thereby stabilizing estimates and improving out-of-sample predictive performance. The method allows the evaluation of posterior probabilities for individual features and models, facilitating rigorous feature selection and interpretation.[9-13]:

$$p(Y_i) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a(\phi_0)} + c(Y_i, \phi_0)\right\}, i = 1, \dots, n \qquad (4)$$

Where $a(.), b(.)$ and $c(.,.)$ are functions for distribution. The observations $Y_1, \dots, Y_n$ are independent but they are not identically distributed and the parameters θ_i have the following relationship $\theta_i = \theta(\eta_{M,i})$ and $\eta_M = 1_n\alpha + X_M\beta_M$ existed under model M. The design matrix X under the full model has full column rank p and the column space C(**X**) does not contain $\mathbf{1_n}$.
The true model, M_T , assumed included in the 2^p model under consideration with true values of intercept and coefficients $\alpha^*_{M_T}$ and $\beta^*_{M_T}$.
The posterior probabilities with g-priors for a model M is defined as:

$$p(M|Y,g) = \frac{p(Y|M,g)\,\pi(M)}{\sum_{M'} p(Y|M')\,\pi(M')} \tag{5}$$

$$p(Y|M,g) = \int\int p(Y|\alpha,\beta_M,M)p(\alpha)p(\beta_M|M,g)d\alpha d\beta_M \tag{6}$$

The $\pi(M)$ and $p(Y|M,g)$ are the prior probability and marginal likelihood of model M.
For any model M, with uniform prior $p(\alpha) \propto 1$ and $\hat{\eta}_{M,i} = \hat{\alpha}_M + x^T_{M,i}\,\hat{\beta}_M$, we have:

$$J_n(\hat{\eta}_M) = diag(d_i) \quad where \quad d_i = Y_i\theta''(\hat{\eta}_{M,i}) + (b \circ \theta)''(\hat{\eta}_{M,i})$$

$$J_n(\hat{\alpha}_M) = 1_n^T J_n(\hat{\eta}_M) 1_n$$

$$J_n(\hat{\beta}_M) = X_M^T(I_n - P_{1_n})^T J_n(\hat{\eta}_M)(I_n - P_{1_n})X_M$$

$$P_{1_n} = 1_n(1_n^T J_n(\hat{\eta}_M)1_n)^{-1}1_n^T J_n(\hat{\eta}_M) \tag{7}$$

The posterior distribution conditional on M with Laplace approximations are:

$$\beta_M|Y,M,g \to (D)\; N\left(\frac{g}{1+g}\hat{\beta}_M, \frac{g}{1+g}J_n(\hat{\beta}_M)^{-1}\right) \tag{8}$$

$$\alpha|Y,M,g \to (D)\; N(\hat{\alpha}_M, J_n(\hat{\alpha}_M)^{-1}) \tag{9}$$

We have in GLMs with uniform prior on α and g-prior for any model M , the approximate marginal likelihood for M and g have closed form

$$p(Y|M,g) = \int p(Y|\beta_M,M)p(\beta_M|M,g)d\beta_M$$

$$\propto p(Y|\hat{\alpha}_M,\hat{\beta}_M,M)J_n(\hat{\alpha}_M)^{-\frac{1}{2}}(1+g)^{-\frac{p_M}{2}}\exp\left\{-\frac{Q_M}{2(1+g)}\right\} \tag{10}$$

$$Q_M = \hat{\beta}_M^T J_n(\hat{\beta}_M)\hat{\beta}_M \tag{11}$$

Now we called $u = \frac{1}{1+g}$ complementary shrinkage factor:

$$p(Y|M,u) \propto u^{\frac{p_M}{2}}\exp\left(-\frac{uQ_M}{2}\right) \tag{12}$$

Then $u \sim tCCH(t,q,r,s,v,\kappa)$ from truncated Compound Confluent Hypergeometric Distribution:

$$p(u|t,q,r,s,v,\kappa) = \frac{v^t\exp\left(\frac{s}{v}\right)}{B(t,q)\Phi_1\left(q,r,t+q,\frac{s}{v},1-\kappa\right)}\frac{u^{t-1}(1-vu)^{q-1}e^{-s}u}{[\kappa+(1-\kappa)vu]^r}1_{\{0<u<\frac{1}{v}\}} \tag{13}$$

where
$t > 0, q > 0\,, r \in \mathbb{R}, s \in \mathbb{R}, v \geq 1\; and\;, \kappa > 0$
$B(t,q)$: $Beta\; Function$

$$\Phi_1(\alpha,\beta,\gamma,x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}(\alpha)_{m+n}\;(\beta)_n\,x^m y^n/[(\gamma)_{m+n}m!\,n!] \tag{14}$$

The Confluent hypergeometric function of two variables or Humbert series is defined as:

$$u \sim tCCH\left(\frac{a}{2},\frac{b}{2},r,\frac{s}{2},v,\kappa\right) \tag{15}$$

where

$a, b, \kappa > 0$, $r, s \in \mathbb{R}, v \geq 1$
Then, we have:

$$p(Y|M) \propto p\left(Y\middle|\hat{\alpha}_M, \hat{\beta}_M, M\right) J_n(\hat{\alpha}_M)^{-\frac{1}{2}} v^{-\frac{p_M}{2}} \exp\left\{-\frac{Q_M}{2v}\right\} . \frac{B\left(\frac{a+p_M}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b+p_M}{2}, \frac{s+Q_M}{2v}, 1-\kappa\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)} \quad (16)$$

The posterior distribution of u under model M is tCCH distribution asymptotically

$$u|\, Y, M \to (D)\; tCCH\left(\frac{a+p_M}{2}, \frac{b}{2}, r, \frac{s+Q_M}{2}, v\,, \kappa\right) \quad (17)$$

And the Bayes factor defined as

$$BF_{M:M_\phi} = \left[\frac{J_n(\hat{\alpha}_{M_\phi})}{J_n(\hat{\alpha}_M)}\right]^{\frac{1}{2}} v^{-\frac{p_M}{2}} \exp\left\{\frac{z_M}{2} - \frac{Q_M}{2v}\right\} \frac{B\left(\frac{a+p_M}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b+p_M}{2}, \frac{s+Q_M}{2v}, 1-\kappa\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)} \quad (18)$$

Model selection criterion is Confluent Hypergeometric Information Criterion or CHIC. CHIC g-prior has the following priors with changing hyperparameters: **1 to 4- CH:** Four priors including:

- **A)** $CH(a = \frac{1}{2}, b = n)$ [**CCH(a=1/2,b=n,s=0)**],
- **B)** $CH(a = a, b = n)$ [**CCH(a=1,b=n,s=0)**],
- **C)** $CH(a = \frac{1}{2}, b = \frac{n}{2})$ [**CCH(a=1/2,b=n/2,s=0)**],
- **D)** $CH(a = 1, b = \frac{n}{2})$ [**CCH(a=1,b=n/2,s=0)**] are special case of the CHIC family and is a generalized Beta Distribution: $p(u|t,q,s) = \frac{u^{t-1}(1-u)^{q-1}\exp(-su)}{B(t,q)\,{}_1F_1(t,t+q,-s)} 1_{\{0<u<1\}}$ with $u \sim CH(\frac{a}{2}, \frac{b}{2}, \frac{s}{2})$, $u|Y, M \to (d)\; CH(\frac{a+p_M}{2}, \frac{b}{2}, \frac{s+Q_M}{2})$ and $p(Y|M) \propto p\left(Y\middle|\hat{\alpha}_M, \hat{\beta}_M, M\right) J_n(\hat{\alpha}_M)^{-\frac{1}{2}} . \frac{B\left(\frac{a+p_M}{2},\right)\,{}_1F_1\left(\frac{a+p_M}{2}, \frac{a+b+p_M}{2}, -\frac{s+Q_M}{2}\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right)\,{}_1F_1\left(\frac{a}{2}, \frac{a+b}{2}, -\frac{s}{2}\right)}$.
- **Beta-prime** [**beta.prime(n=n)**]: $u \sim Beta\left(\frac{1}{4}, \frac{n-p_M-1.5}{2}\right)$.
- **ZS adapted** [**CCH(a=1,b=2,s = n + 3)**].
- **Benchmark** [**CCH(a=0.02,b=0.02*max(n, p^2),s=0)**]: $\boldsymbol{u \sim Beta(c, c.max(n, p^2))}$.
- **Robust** [**robust(n=n)**] : $\mathrm{p_r(u)} = \mathrm{a_r}[\rho_r(\mathrm{b_r} + \mathrm{n})]^{\mathrm{a_r}} \frac{\mathrm{u}^{\mathrm{a_r}-1}}{[1+(\mathrm{b_r}-1)\mathrm{u}]^{\mathrm{a_r}+1}} 1_{\{0<u<\frac{1}{\rho r(\mathrm{b_r}+n)+(1-\mathrm{b_r})}\}}$.
- **Intrinsic** [**intrinsic(n=n)**] : $g = \frac{n}{p_m+1} . \frac{1}{w}$, $w \sim Beta(\frac{1}{2}, \frac{1}{2})$.
- **Hyper g/n** [**hyper.g.n(alpha=3,n=n)**] : $p(g) = \frac{\mathrm{a_h}-2}{2\mathrm{n}} \left(\frac{1}{1+\frac{g}{n}}\right)^{\frac{\mathrm{a_h}}{2}}$ where $2 < \mathrm{a_h} \leq 4$.
- **DBF g=n** [**g.prior(n)**] : Data-Based Bayes Factor (DBF) with Q_M replaced by z_M in TBF.
- **TBF g=n** [**testBF.prior(n)**] : Test-Based Bayes factors (TBF) with $TBF_{M:M_\phi} = \frac{G(z_M; \frac{p_M}{2}, \frac{1}{2(1+g)})}{G(z_M; \frac{p_M}{2}, \frac{1}{2})} = (1+g)^{-\frac{p_M}{2}} \exp\left\{\frac{g z_M}{2(1+g)}\right\}$ and with asymptotic distribution of z_M under M and M_ϕ where $G(z_M; a, b)$ is Gamma distribution with mean $\frac{a}{b}$.

- **Jeffreys [Jeffreys()]**: $u \sim Beta\left(\frac{a_h}{2} - 1\ , 1\right) where\ 2 < a_h \leq 4\ with\ a_h = 2.$
- **Hyper-g [CCH(a=1,b=2,s=0)]**: $u \sim Beta\left(\frac{a_h}{2} - 1\ , 1\right) where\ 2 < a_h \leq 4\ \ with\ a_h = 3.$
- **Uniform [CCH(a=2, b=2,s=0)]**: $u \sim Beta\left(\frac{a_h}{2} - 1\ , 1\right) where\ 2 < a_h \leq 4\ with\ a_h = 2.$
- **Local EB (EB.local()):** $\hat{g}_M^{LEB} = \arg\max_{g \geq 0} p(Y|M, g)\ \ and$ Under g-prior $\hat{g}_M^{LEB} = \max(\frac{Q_M}{p_M} - 1\ , 0)$ (fixed value of g).
- **AIC [aic.prior()].**
- **BIC [bic.prior(n)].**

The predictive performance of out-of-sample models was assessed using at least 250 bootstrap cross-validation iterations. Four metrics were calculated: the area under the ROC curve (AUC), the calibration slope (CS), the logarithmic score (LS), and the Brier score (the average squared difference between the predicted probability $\hat{\mu}$ and the observed outcome Y). Higher AUC and CS values (closer to 1) indicate better model discrimination and calibration. Lower LS and Brier scores reflect improved predictive accuracy and reliability.

Supplementary B.5 – Interpretable Machine Learning (IML)

To obtain interpretable machine learning results for Random Forest (RF), the dataset was randomly partitioned into 50% training and 50% testing sets. RF models were trained using the randomForestSRC package in R [14]. Variable importance was assessed using the Breiman–Cutler permutation method with 1,000 trees, and 1,000 subsampling iterations were performed to estimate confidence intervals and standard errors. Marginal (unadjusted) and partial (adjusted) dependence plots were generated to evaluate the effect of each covariate on the predicted class probability, with partial plots displaying ± two standard errors.

Hyperparameters were tuned using the train() function from the caret package [15] with grid search and five-fold cross-validation. For example, the number of trees was tested at {500, 1000}, and node sizes at {1,5}. The optimal model was subsequently used to compute Local Interpretable Model-Agnostic Explanations (LIME) for individual predictions using the lime package [16]. Finally, SHapley Additive exPlanation (SHAP) values [17] were calculated and visualized with the DALEX package [18], providing a detailed assessment of the contribution of each covariate to model predictions on a per-sample basis.

Supplementary B.6 – Statistical Significance with GAMLSS-BEZI

Generalized Additive Models for Location, Scale, and Shape (GAMLSS) with a zero-inflated beta (BEZI) family were fitted using the metamicrobiomeR package in R [19]. This approach models the relative abundance of microbiome features ,y, while accounting for the excess zeros commonly observed in microbiome data, allowing robust estimation of feature-specific effects and statistical significance:

$$Beta(\mu, \phi) \rightarrow f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} y^{\mu\phi - 1}(1-y)^{(1-\mu)\phi - 1}\ , y\ \in (0,1) \qquad (19)$$

$$0 \leq\ \mu \leq 1\ , \phi > 0\ \ and\ \Gamma(.)\ is\ gamma\ function$$

$$f(y;\alpha,\mu,\phi) = \begin{cases} \alpha, & if\ y = 0, \\ (1-\alpha)f(y;\mu,\phi), & if\ y \in (0,1) \end{cases}$$

The coefficients, p-values, and false discovery rate (FDR)-adjusted p-values were calculated for each taxa feature, taxonomic level, and sampling site separately. Model coefficients represent group differences between CRC and AP and were adjusted for gender. Additionally, GAMLSS-BEZI models with random effects were fitted, treating sampling sites as a random effect to account for site-specific variability. Features were filtered based on prevalence (threshold ≥ 0.05) and relative abundance (threshold ≥ 0.00005) to ensure robust estimation.

Zero-Inflated Generalized Dirichlet Multinomial (ZIGDM) tests were performed for each sampling site and taxonomic level, with a false discovery rate (FDR) threshold set at 0.05. Two sets of tests were conducted: (1) Non-Zero-Inflated ZIGDM, corresponding to the standard GDM model, and (2) Zero-Inflated ZIGDM, applied to taxa with a minimum of 10 zero counts. For each set, three types of outcomes were assessed: (i) differential mean, (ii) differential dispersion, and (iii) differential presence–absence frequency. All analyses were performed using the miLineage R package [20]

References

1. Wirbel, J., et al., *Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox.* Genome biology, 2021. **22**(1): p. 93.
2. Cao, X., K. Lee, and Q. Huang, *Bayesian variable selection in logistic regression with application to whole-brain functional connectivity analysis for Parkinson's disease.* Statistical Methods in Medical Research, 2021. **30**(3): p. 826-842.
3. Shankar, J., et al., *Using Bayesian modelling to investigate factors governing antibiotic-induced Candida albicans colonization of the GI tract.* Scientific reports, 2015. **5**(1): p. 8131.
4. Shankar, J., et al., *Looking beyond respiratory cultures: microbiome-cytokine signatures of bacterial pneumonia and tracheobronchitis in lung transplant recipients.* American Journal of Transplantation, 2016. **16**(6): p. 1766-1778.
5. Shankar, J., et al., *A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses.* BMC bioinformatics, 2015. **16**: p. 1-18.
6. Scott, S.L., *BoomSpikeSlab: MCMC for spike and slab regression*. 2023.
7. Tüchler, R., *Bayesian variable selection for logistic models using auxiliary mixture sampling.* Journal of Computational and Graphical Statistics, 2008. **17**(1): p. 76-94.
8. Susin, A., et al., *Variable selection in microbiome compositional data analysis.* NAR Genomics and Bioinformatics, 2020. **2**(2): p. lqaa029.
9. Li, Y. and M.A. Clyde, *Mixtures of g-priors in generalized linear models.* Journal of the American Statistical Association, 2018. **113**(524): p. 1828-1845.
10. Liang, F., et al., *Mixtures of g priors for Bayesian variable selection.* Journal of the American Statistical Association, 2008. **103**(481): p. 410-423.
11. Porwal, A. and A.E. Raftery, *Comparing methods for statistical inference with model uncertainty.* Proceedings of the National Academy of Sciences, 2022. **119**(16): p. e2120737119.

12. Sampling, B.B.V.S.a.M.A.u.B.A., *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. 2023.
13. Averaging, B.B.M., *BMA: Bayesian Model Averaging*. 2022.
14. Hemant Ishwaran, U.B.K., *randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. 2024.
15. Kuhn, M., *caret: Classification and Regression Training*. 2023.
16. Hvitfeldt, E., *lime: Local Interpretable Model-Agnostic Explanations*. 2022.
17. Biecek, P. and T. Burzykowski, *Explanatory model analysis: explore, explain, and examine predictive models*. 2021: Chapman and Hall/CRC.
18. Biecek, P., *DALEX: Explainers for complex predictive models in R.* Journal of Machine Learning Research, 2018. **19**(84): p. 1-5.
19. Ho, N.T., et al., *metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models.* BMC bioinformatics, 2019. **20**: p. 1-15.
20. Tang, Z.-Z. and G. Chen, *Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis.* Biostatistics, 2019. **20**(4): p. 698-713.