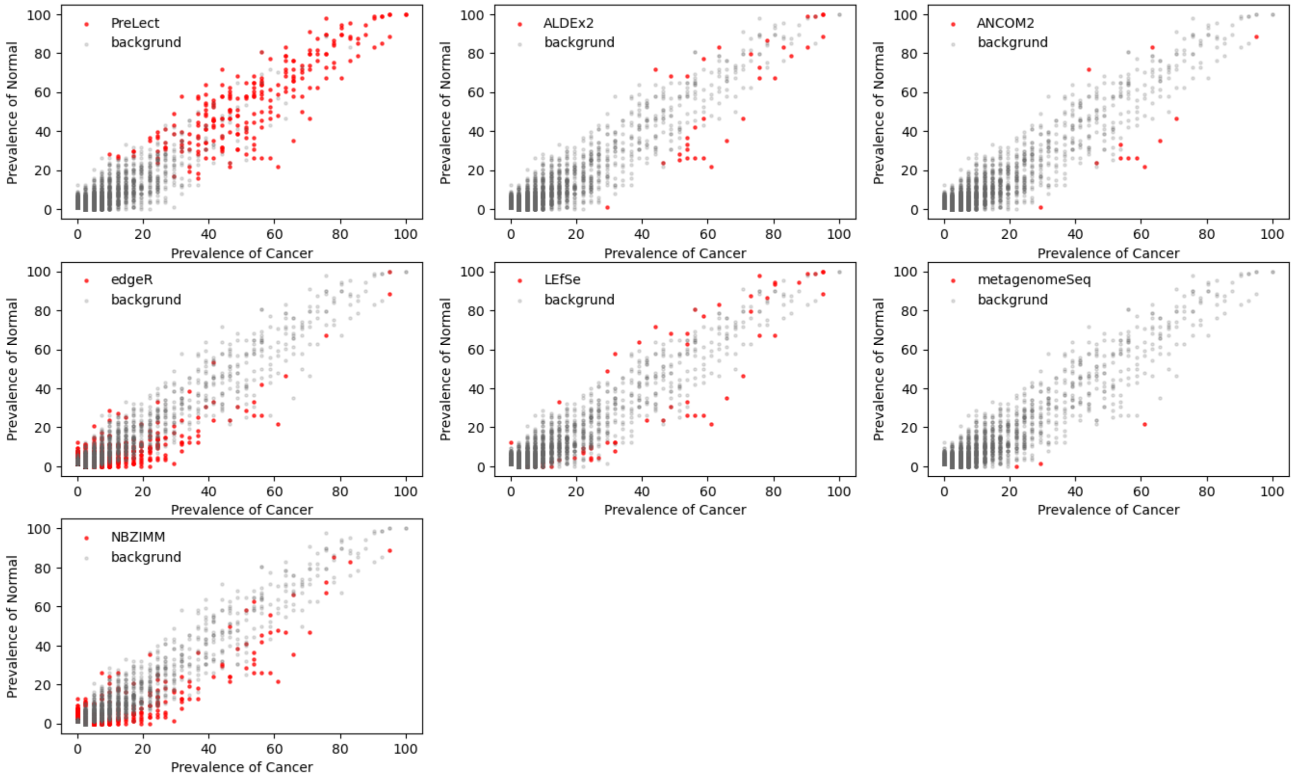
Supporting Information

PreLect: Prevalence Leveraged Consistent Feature Selection Decodes Microbial Signatures across Cohorts

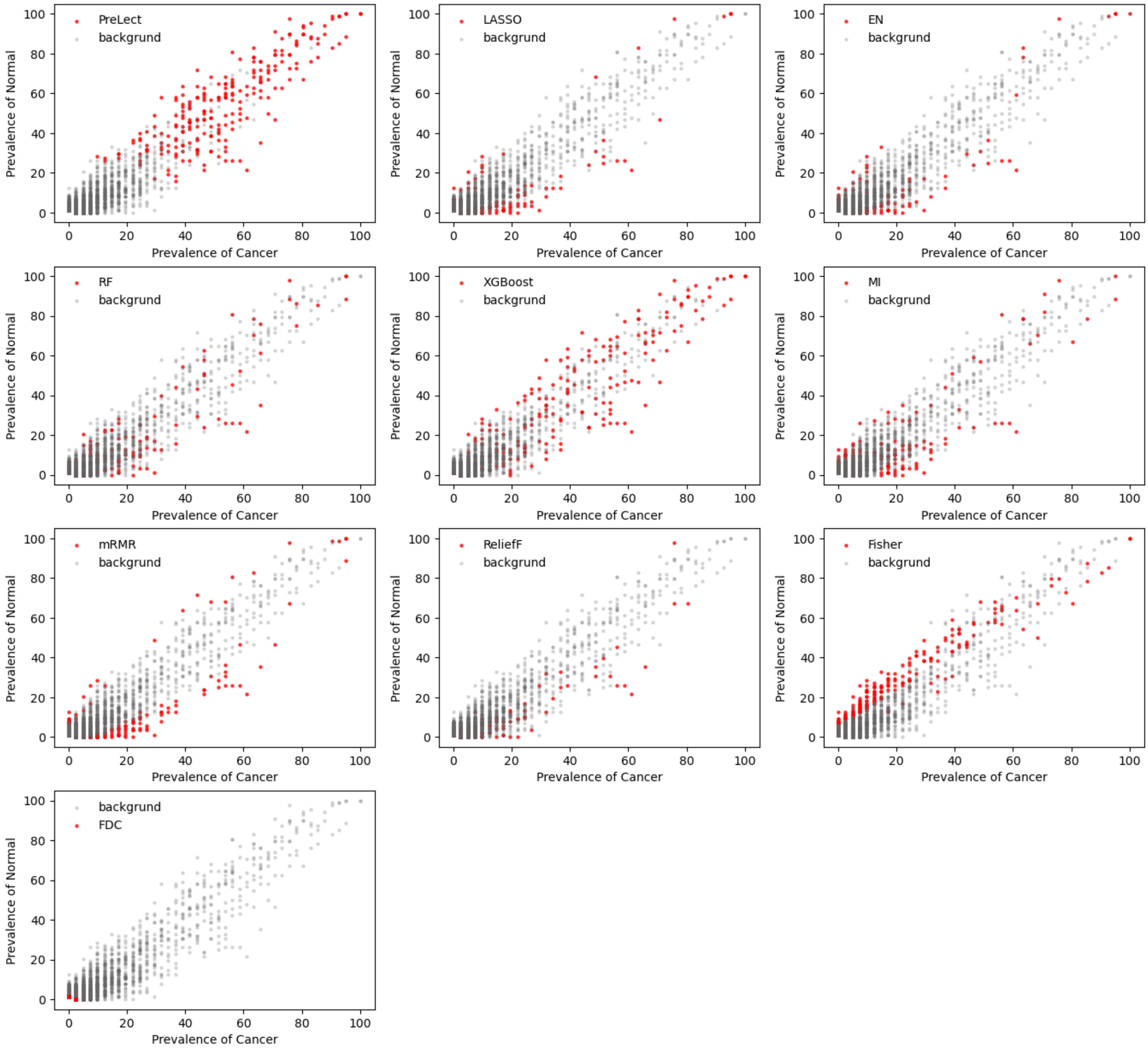
Yin-Cheng Chen, Yin-Yuan Su, Tzu-Yu Chu, Ming-Fong Wu, Chieh-Chun Huang, Chen-Ching Lin

SI Figures



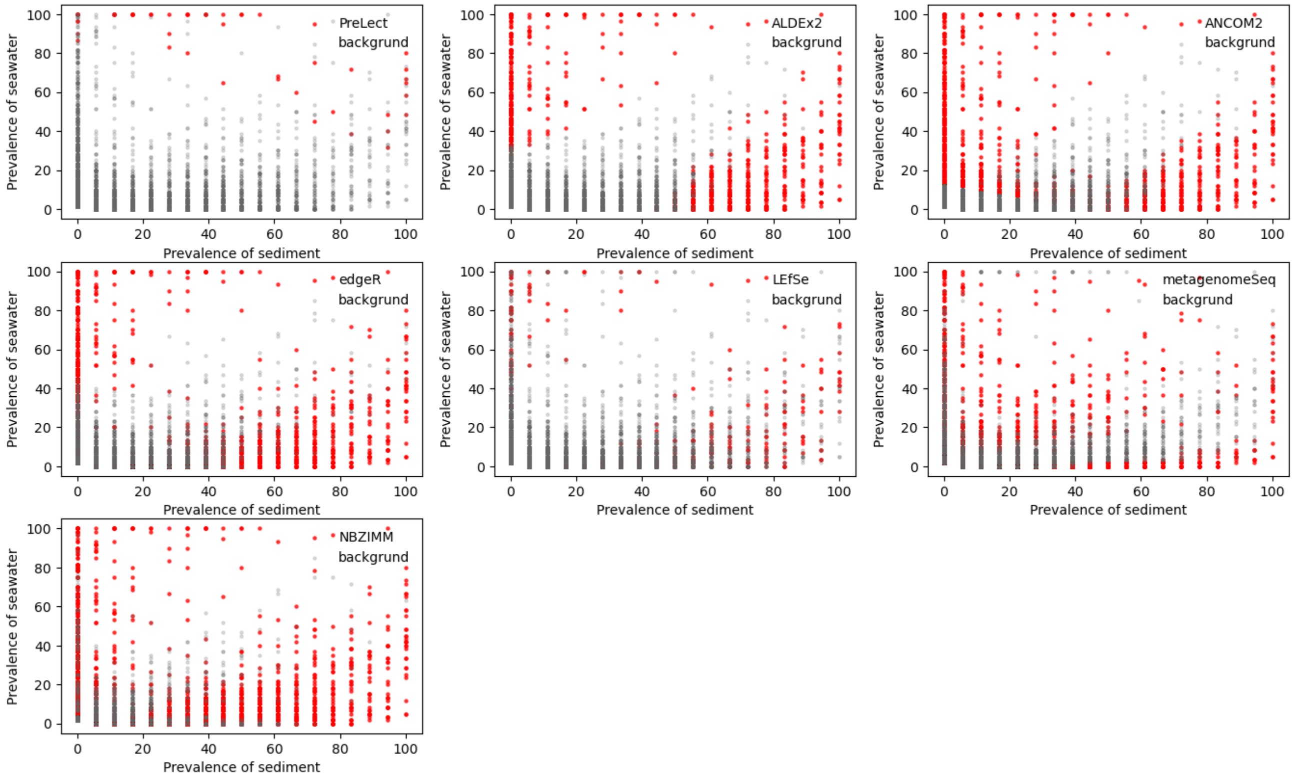
Supplementary Figure 1. The feature prevalence profile among statistics-based methods in the Zeller_CRC dataset.

The selected features with each method were highlighted as red dots. The x-axis represents the feature prevalence in the CRC patients, and the y-axis indicates the feature prevalence in control.



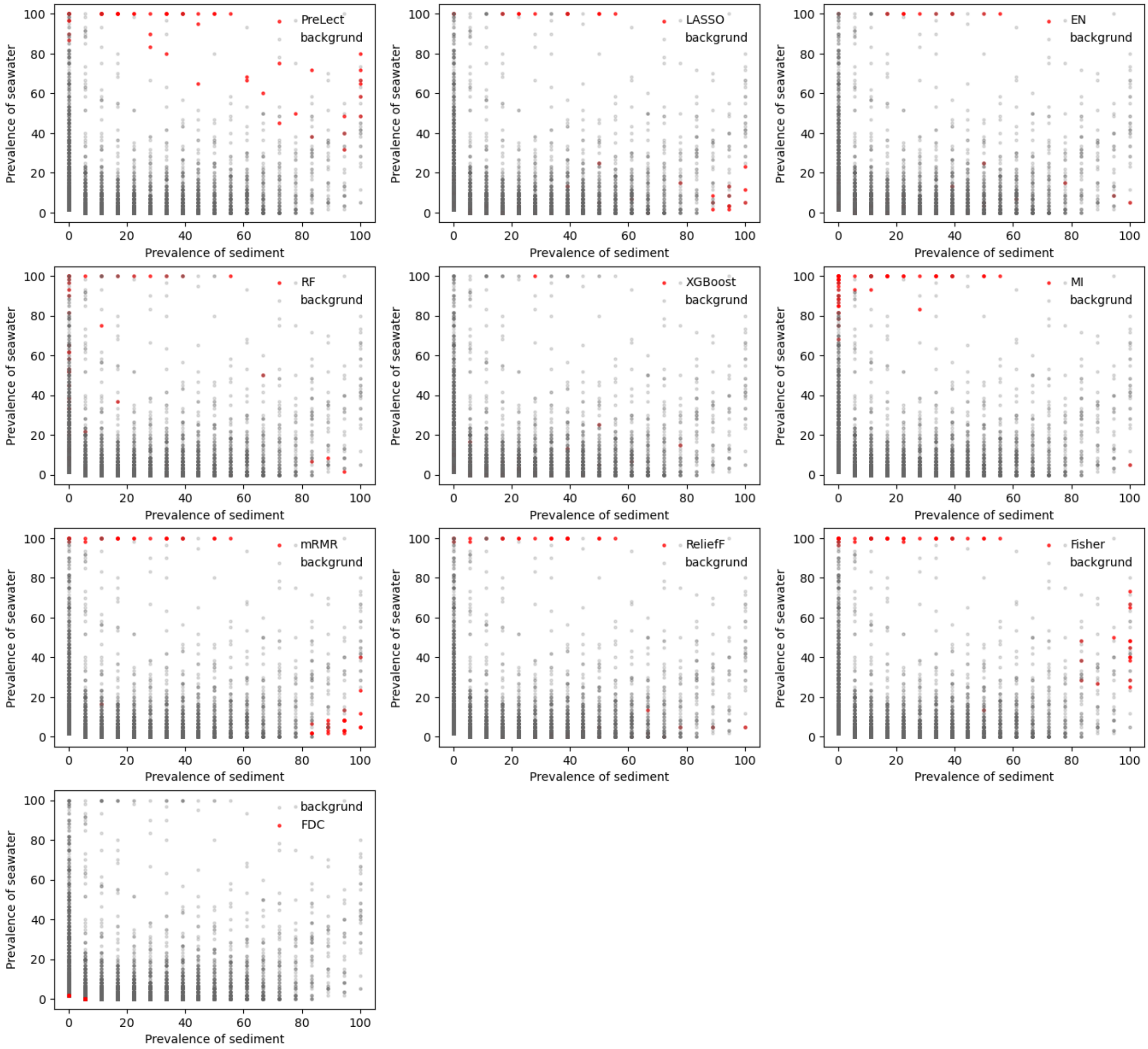
Supplementary Figure 2. The feature prevalence profile among ML-based methods in the Zeller_CRC dataset.

The selected features with each method were highlighted as red dots. The x-axis represents the feature prevalence in CRC patients, and the y-axis indicates the prevalence in control.



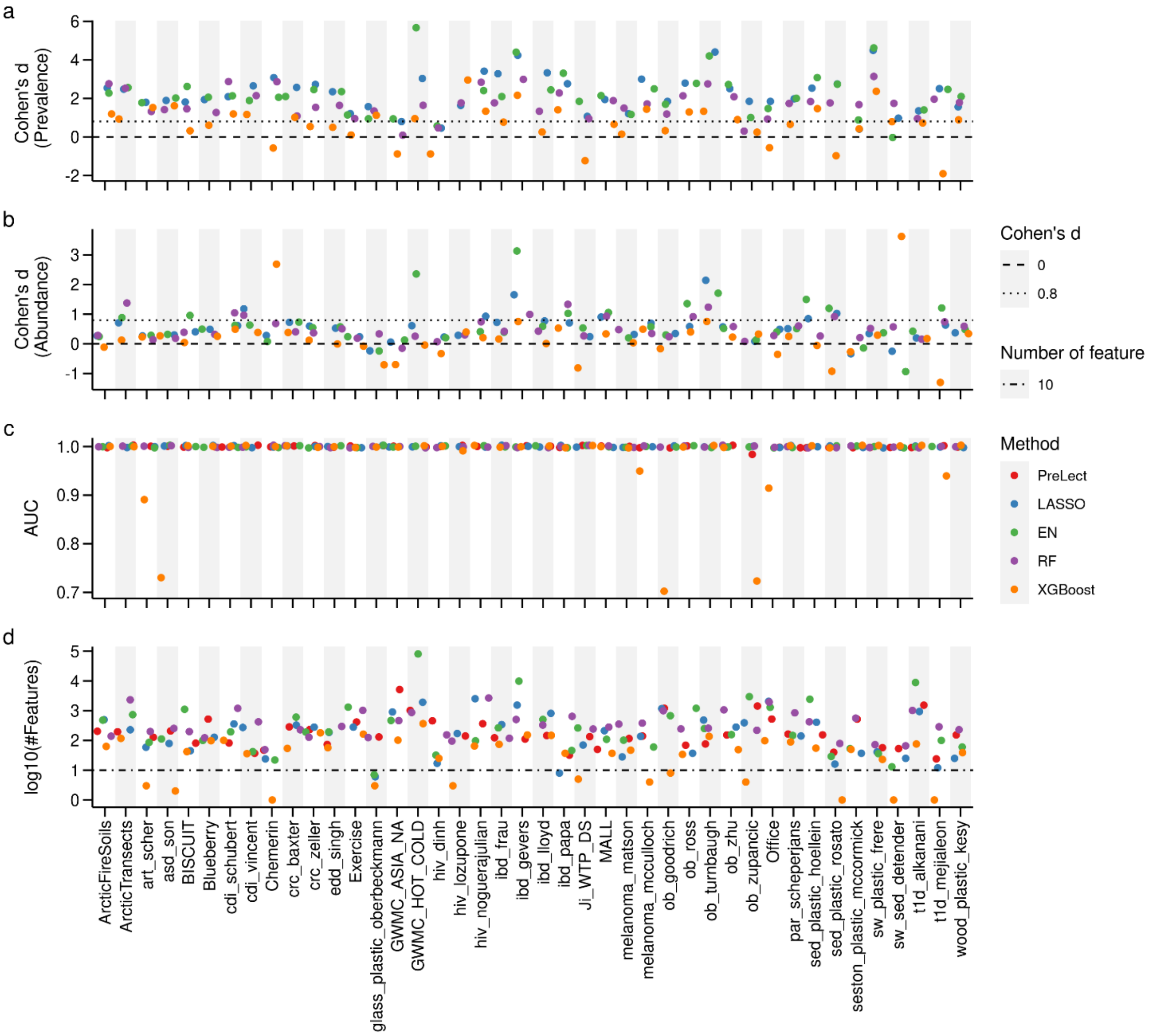
Supplementary Figure 3. Feature prevalence profile in the sw_sed_detender dataset using statistics-based methods.

This figure highlights the selected features as red dots. The y-axis represents feature prevalence in sediment, while the x-axis indicates feature prevalence in seawater. This visualization helps to compare the prevalence of features selected by different methods across the two environmental conditions.

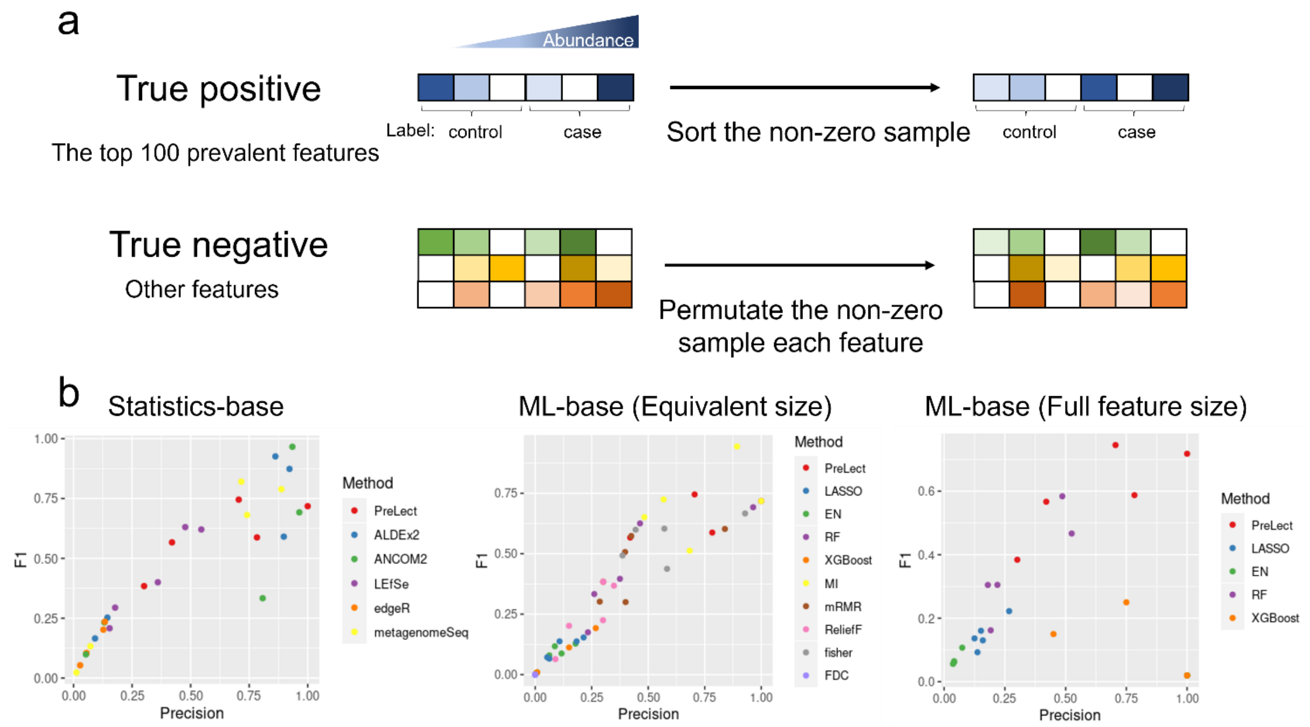


Supplementary Figure 4. The feature prevalence profile among ML-based methods in the sw_sed_detender dataset.

The selected features with each method were highlighted as red dots. The x-axis represents the feature prevalence in the sediment, and the y-axis indicates the feature prevalence in seawater.

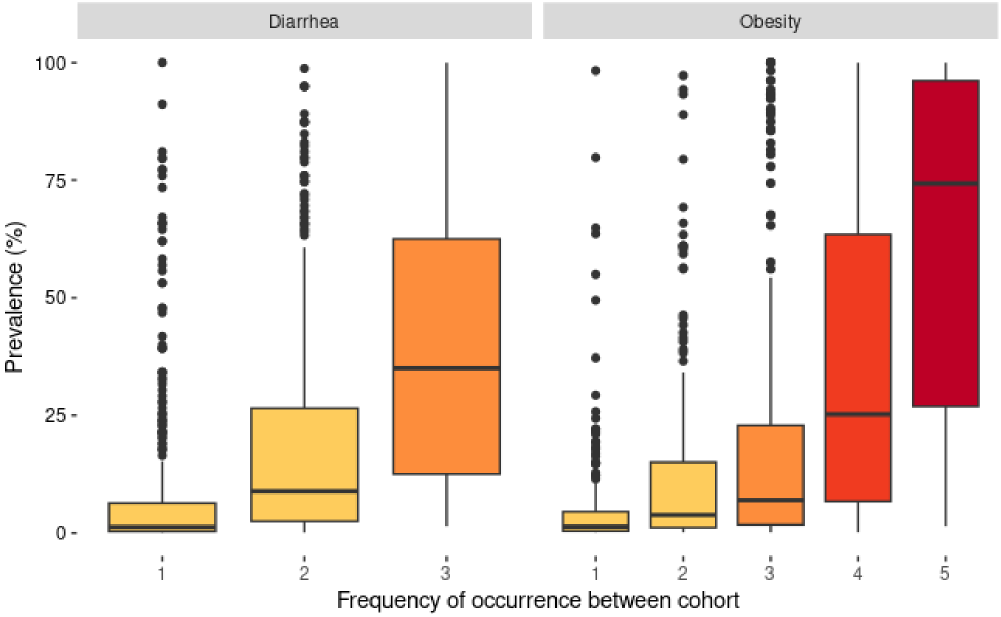


Supplementary Figure 5. Comparison of PreLect with the full feature set of ML-base methods. (A) and (B) Effect size of prevalence and abundance difference. Cohen's D measures the effect size difference between PreLect and other benchmarked machine learning methods. Values above 0.8 (dotted line) indicate a notable higher feature prevalence of PreLect. (C) Classification performance. The area under the receiver operating characteristic curve (AUC) is derived from a naïve logistic regression model to classify case and control samples. Herein, all the ML-based methods use the default number of features, and (D) shows the number of features used in each method.



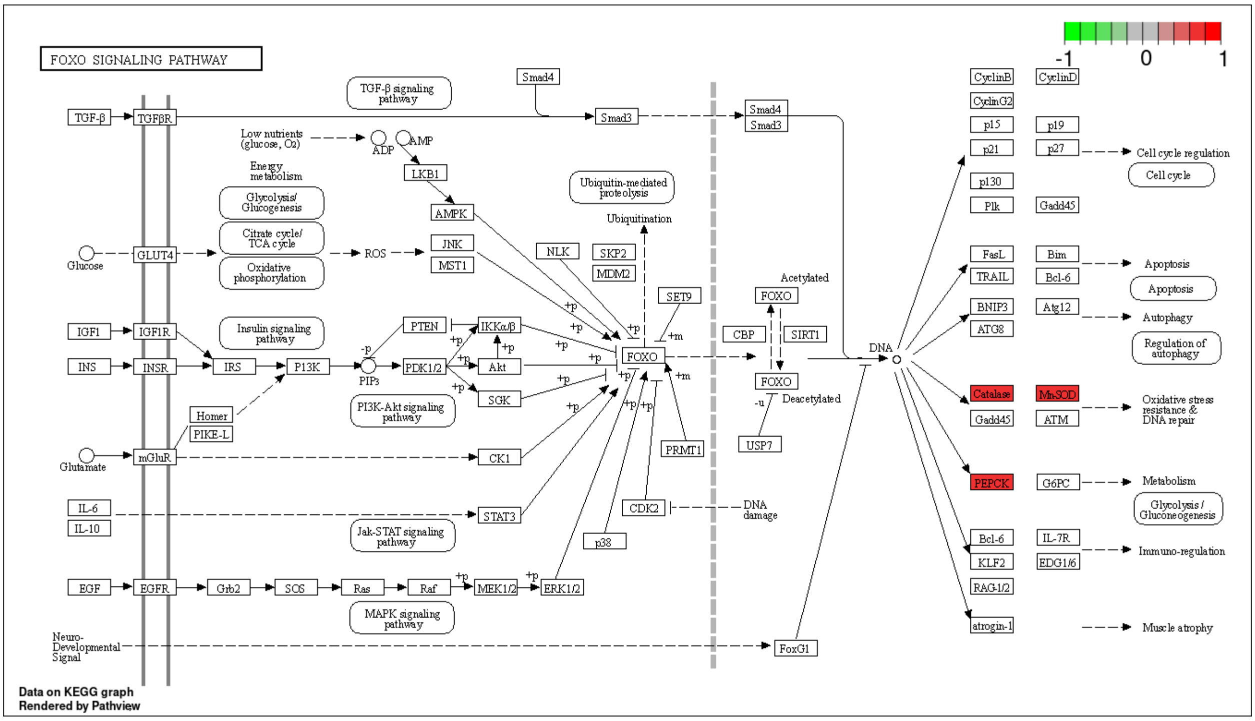
Supplementary Figure 6. Synthetic data strategy and results.

(A) This panel illustrates the synthetic data strategy used to generate true positive and true negative features, ensuring a controlled environment to assess feature selection methods accurately. (B) This panel displays the precision and F1 scores for each benchmarking method, providing a quantitative comparison of their performance in identifying true positive features within the synthetic datasets.



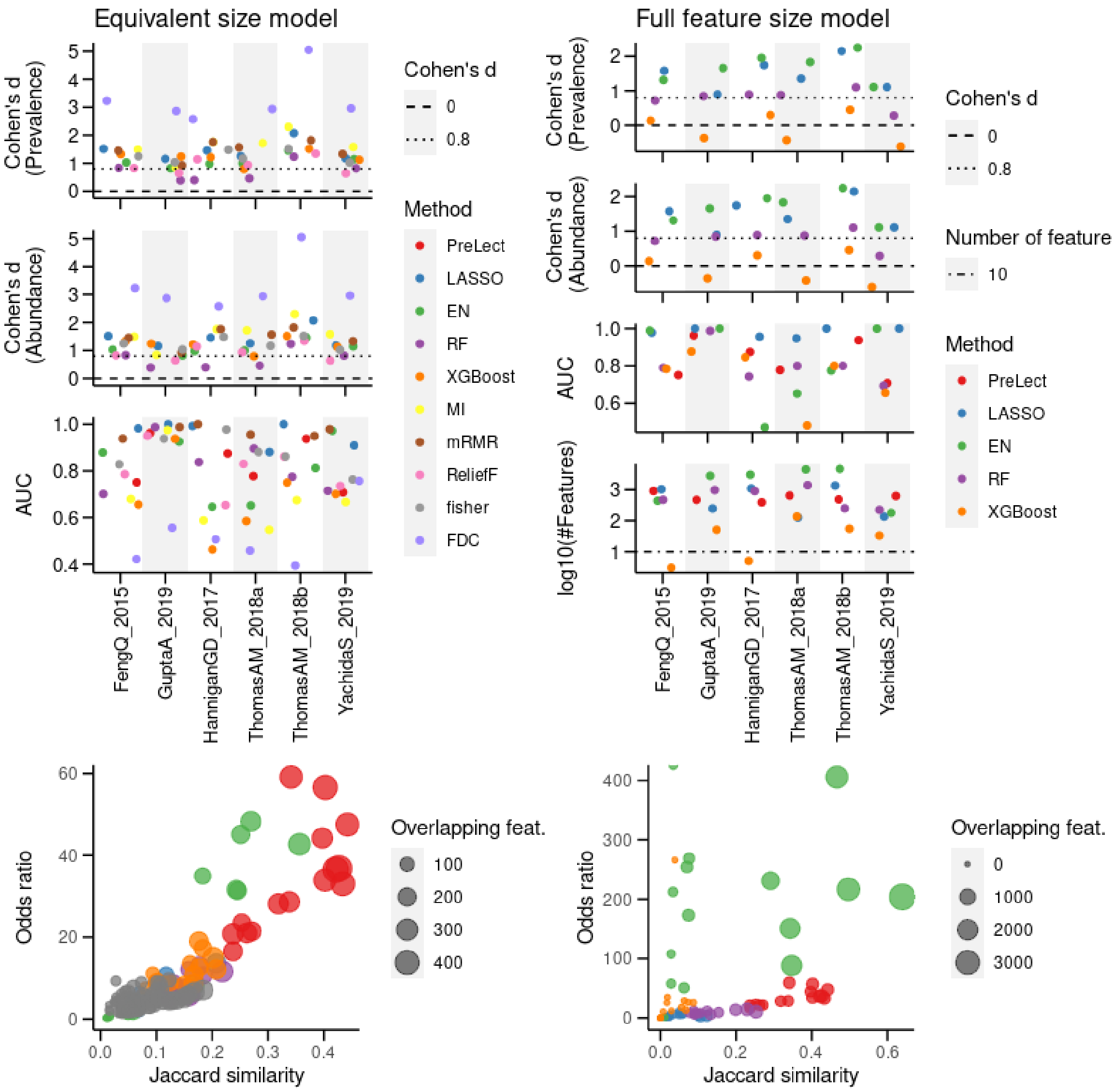
Supplementary Figure 7. Universality of prevalent features across cohorts.

The frequency of features is calculated based on their occurrence across different cohorts and is compared with their prevalence within each individual dataset. This approach highlights the relationship between multi-cohort occurrence and dataset-specific prevalence.



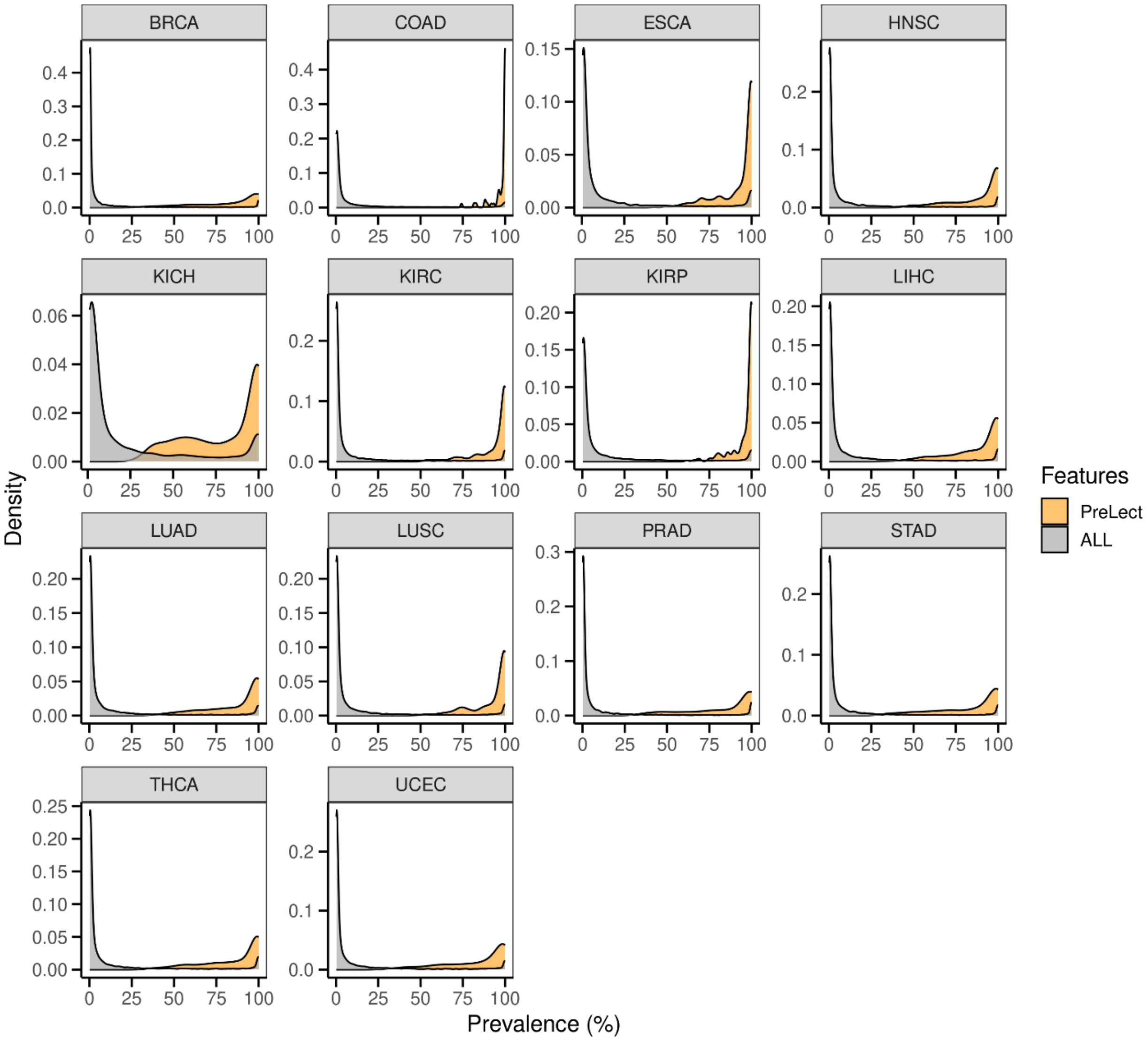
Supplementary Figure 8. Enriched KO in FoxO signaling pathway.

The catalase (K03781) and superoxide dismutase (K04564) were found to be significantly enriched in colorectal cancer (CRC) based on GSEA and are highlighted. A color scheme is used to depict the fold-changes of these KOs: red signifies KOs enriched in cancer patients, and green indicates KOs enriched in normal samples.



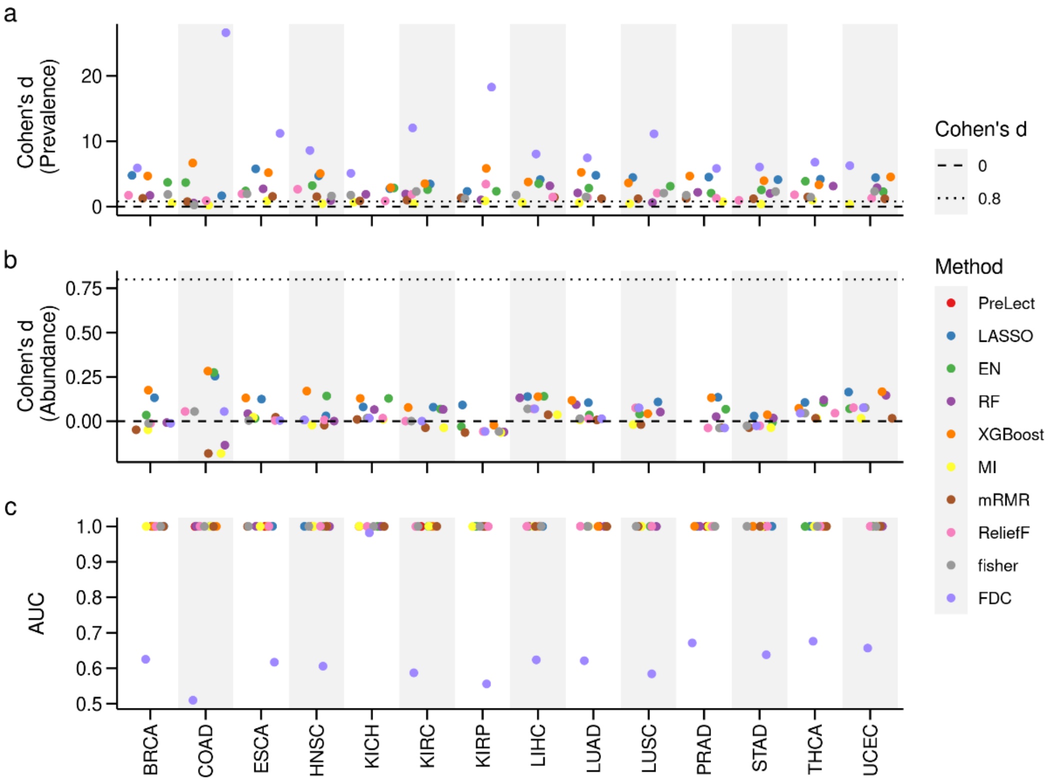
Supplementary Figure 9. Comparison of PreLect with other ML-based methods in shotgun dataset.

The left panel illustrates the selection profile using the Equivalent Size Model, where the number of features is constrained to match those selected by PreLect for nine benchmarking methods. The right panel displays the results of the Full Feature Set Model, showcasing the benchmarked outcomes when all available features are considered.



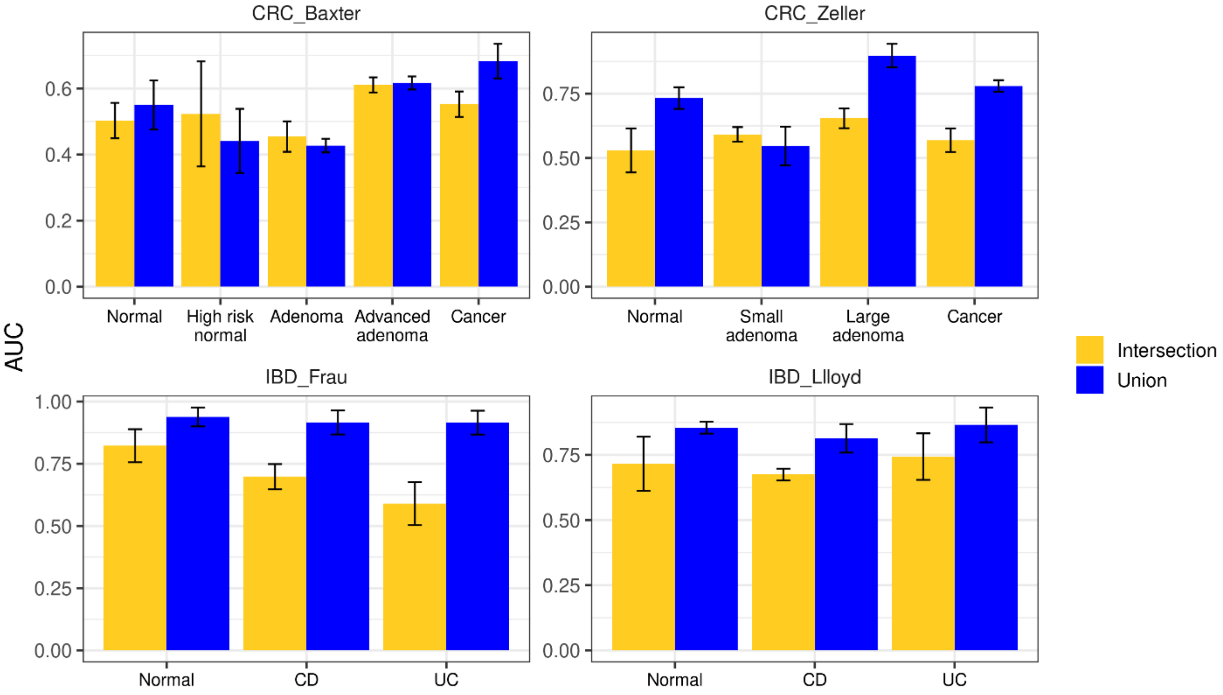
Supplementary Figure 10. The sparsity of miRNA dataset.

To demonstrate the sparsity in the miRNA dataset, we illustrated the whole feature prevalence by density (grey) and the prevalence distribution of features selected by PreLect (orange).

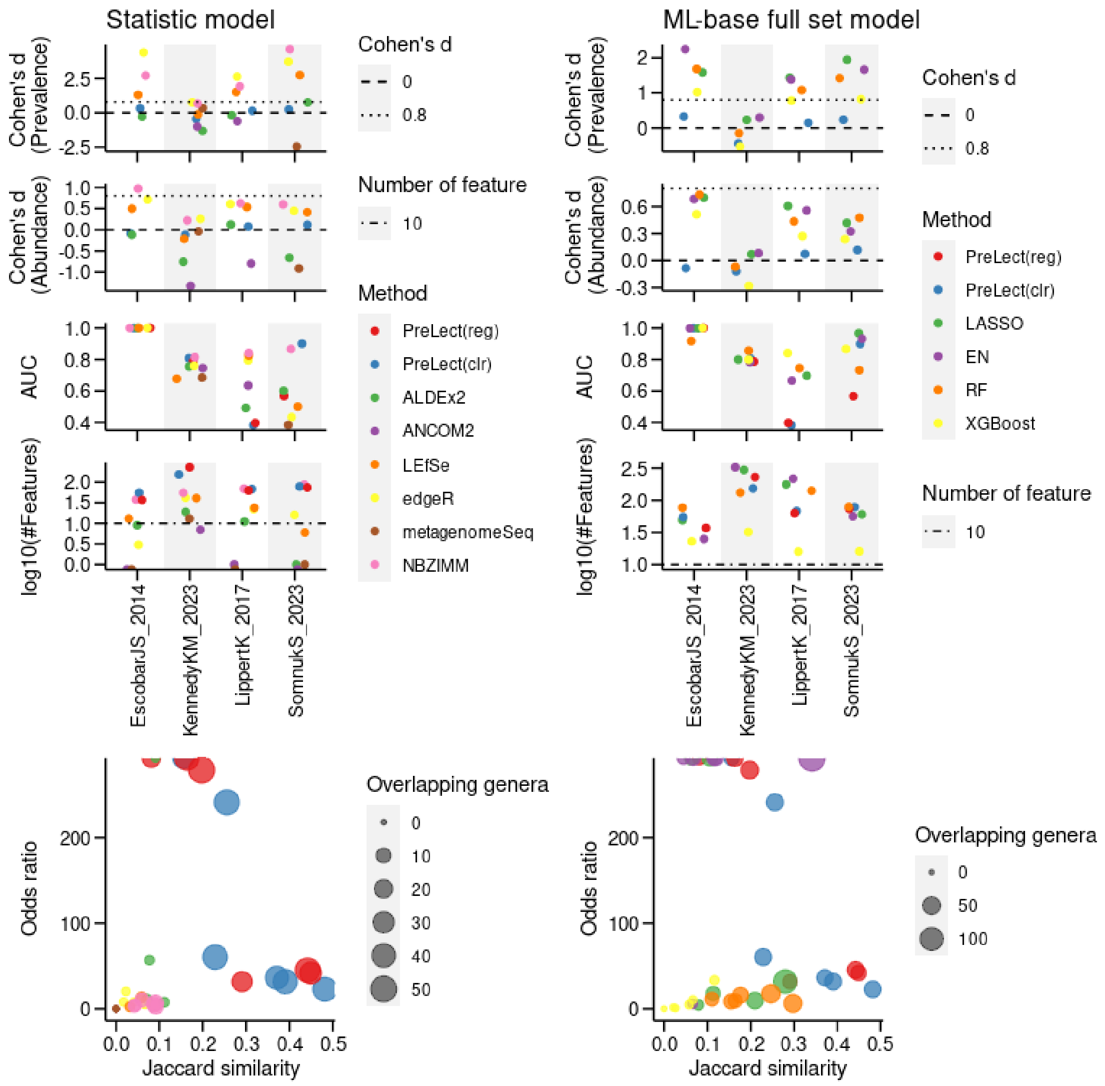


Supplementary Figure 11. Comparison of PreLect with other ML-based methods in microRNA dataset.

(A) and (B) Effect size of prevalence and abundance difference. Cohen's D measures the effect size of the prevalence difference between PreLect and other benchmarked methods. Values above 0.8 (dotted line) indicate a notable higher feature prevalence or abundance of PreLect. (C) Classification performance. The AUC value is derived from a naïve logistic regression model to classify case and control samples. A full feature set was applied to evaluate the classification performance between normal and tumor samples.

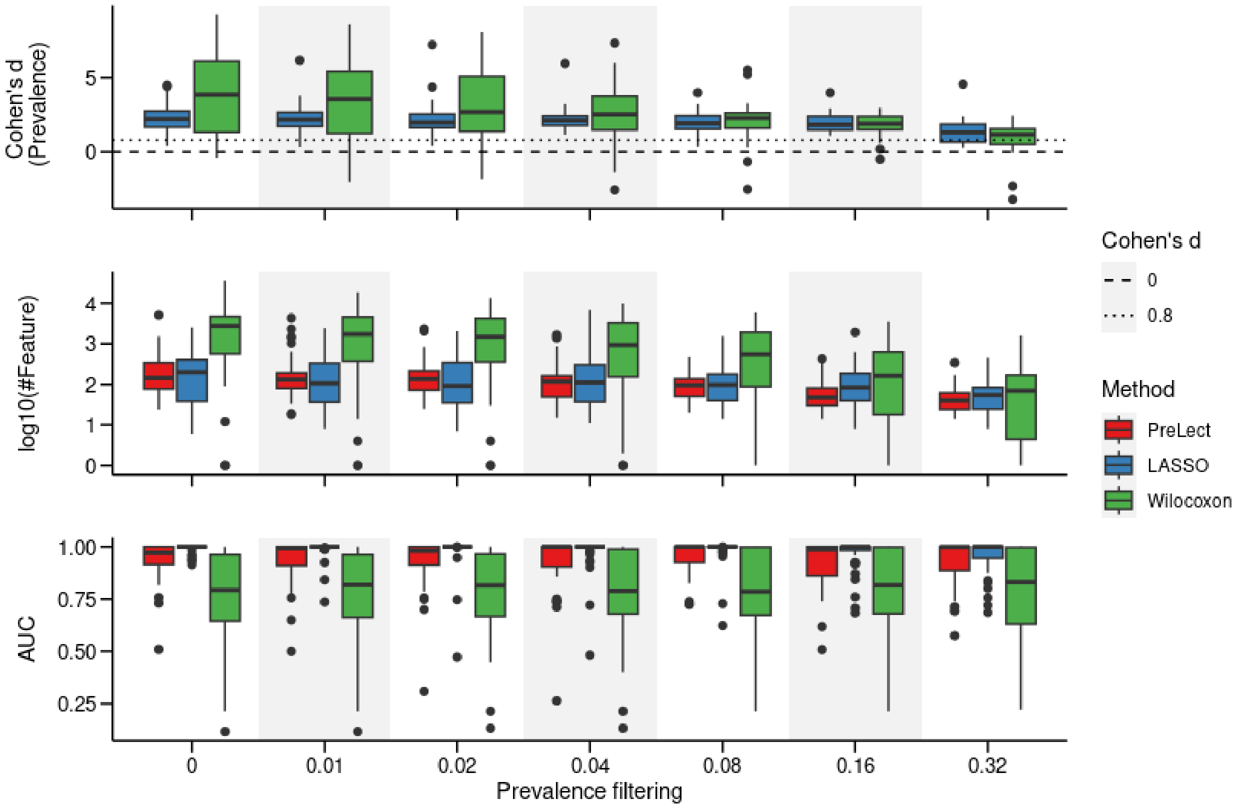


Supplementary Figure 12. Exploring the potential of PreLect for multi-class classification. The feature set selected by PreLect was validated using logistic regression with one-vs.-rest strategy. 3-fold cross-validation was performed separately on each of the four datasets, and the mean and standard deviation are indicated by barplot and error bars.



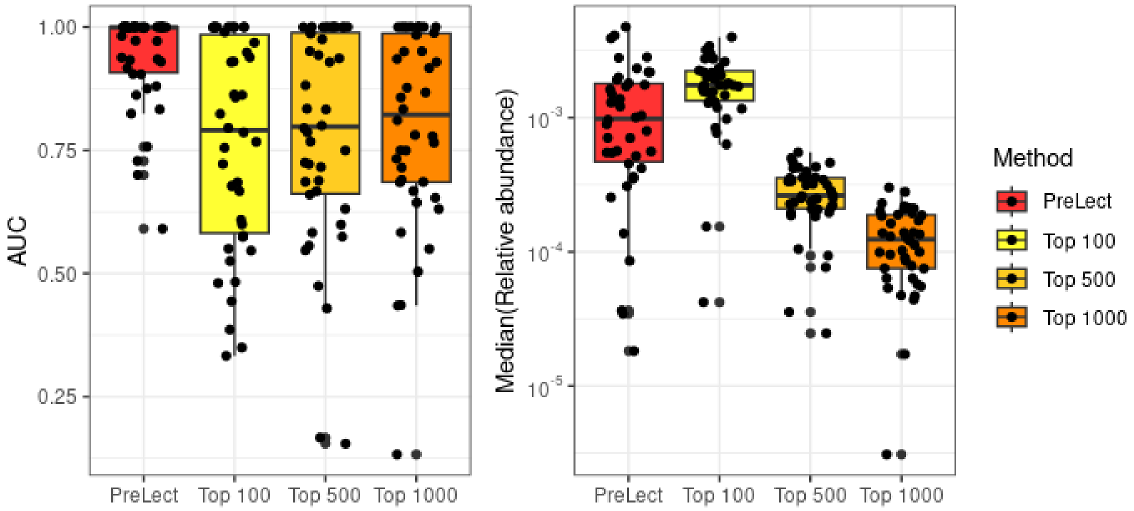
Supplementary Figure 13. PreLect regression applied to obesity 16S amplicon data.

This figure contrasts the feature sets obtained from PreLect regression (PreLect(reg)) with those from PreLect classification (PreLect(clr)), as well as with other benchmarking methods. A Cohen's *d* value exceeding 0.8, indicated by the dotted line, signifies significantly higher feature prevalence or abundance in the PreLect regression compared to other methods.



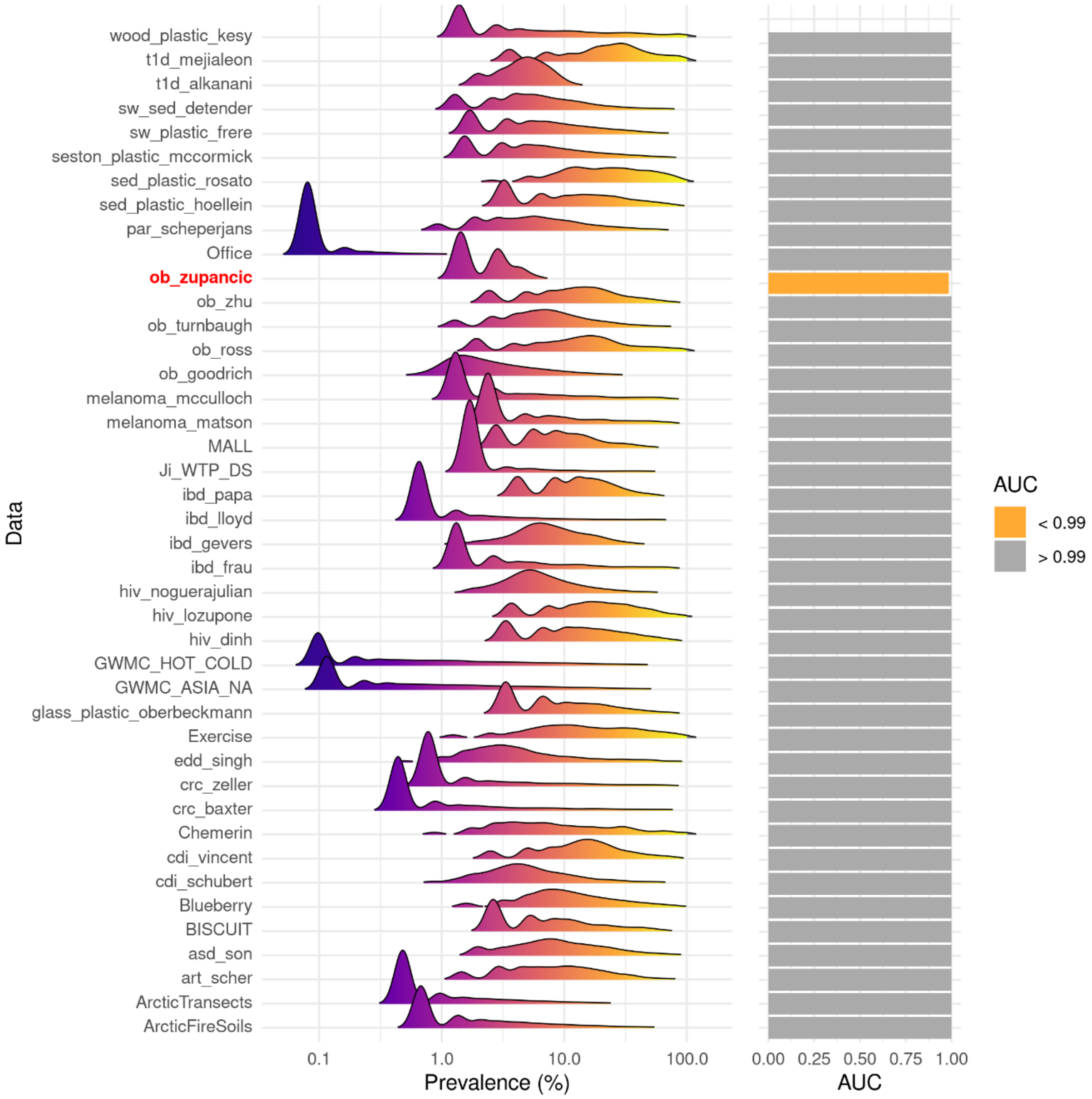
Supplementary Figure 14. Comparative analysis of PreLect and conventional feature selection methods using prevalence filtering strategy across 42 microbiome datasets.

The upper panel presents the effect size of the prevalence difference. A positive Cohen's *d* value indicates that features selected by PreLect exhibit higher prevalence compared to those selected by the two conventional methods. The middle panel displays the number of features each method selected for each dataset. The lower panel shows the classification performance, where the AUC score, derived from a basic logistic regression model, evaluates the ability of the selected features to distinguish between case and control samples.



Supplementary Figure 15. Classification capability of prevalent features.

We selected the top 100, 500, and 1000 prevalent features from 42 benchmark datasets and assessed their classification performance using logistic regression, while also analyzing their abundance. The results suggest that PreLect effectively balances the selection of prevalent and informative features, enhancing overall performance.



147

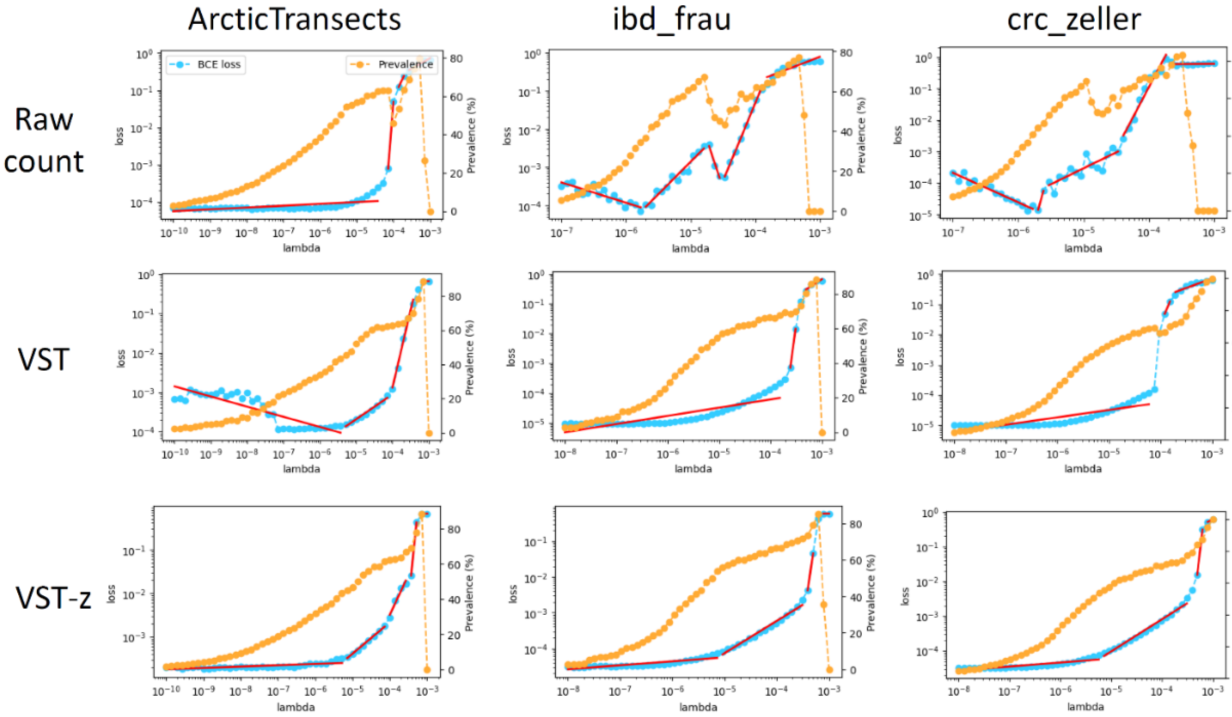
148 **Supplementary Figure 16. The prevalence distribution of the features could influence PreLect's**

149 **performance.**

150 The density plots on the left panel illustrate each dataset's feature prevalence distribution. The

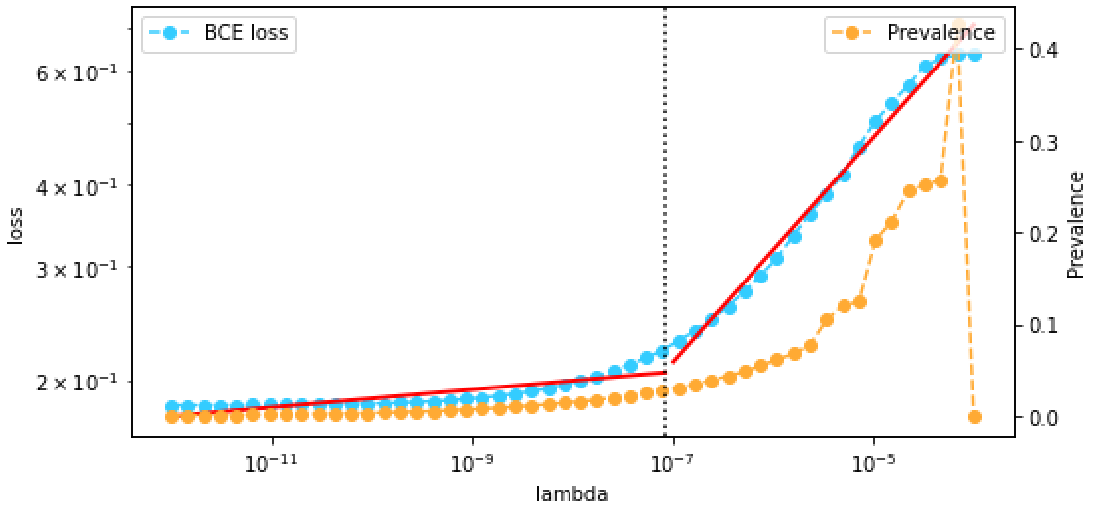
151 performance of the feature set selected by PreLect is shown on the right panel, with the AUC as the

152 metric. Datasets with AUC scores lower than 0.99 are highlighted.



Supplementary Figure 17. Stability of PreLect with VST-transformed data and z-Score standardization.

We conducted lambda scanning of PreLect using three different data processing methods: raw counts, variance stabilizing transformation (VST)-transformed data, and VST-transformed data with z-score standardization. Our analysis shows that using VST with z-score standardization results in the smoothest loss curves, indicating enhanced stability in the model's performance.



Supplementary Figure 18. PreLect application in real-sim dataset.

The lambda selection in real-sim of libsvm is shown, where segmented regression was conducted with k = 2 in loss history. The blue points indicate the mean of loss, and the orange dots represent the mean of prevalence for each lambda within five folds CV.

Supplementary Notes

Supplementary Note 1: Multi-class classification

In order to implement the multi-class task in PreLect with the one-vs-rest strategy, we design a perception $w_{d \times l}$ where $l \in [c] = 1,2, \ldots, c$ and c is the number of categories in labels, each column of perception w representing the different classifier, and the objective function is modified as the following equation.

$$\boldsymbol{min f(w) = \frac{1}{c} \sum_{l}^{c} \left(BCE(y_l, \hat{y_l}) + \lambda \sum_{j}^{d} \frac{|w_{j,l}|}{p_{j,l}} \right)} \qquad (1)$$

One hot encoding is implemented in the label, which is denoted as $y_{i \times l}$ and $y_{i \times l} \in [0,1]$, each column of y is the response variable in each binary classifier. The $p_{j,l}$ is the prevalence of feature i that only considers samples belonging to category l, the loss is defined as the mean of BCE with L_1-regularization in each classifier, and the PGD is also utilized to optimize the perception. Like the single classification of PreLect, we examine the lambda with k-fold CV from 10^{-8} to 10^{-2} and select the suitable lambda at the turning point that loss value from the horizontal line to dramatic rising.

Supplementary Note 2: PreLect regression

We have developed a regression version of PreLect with the following objective function.

$$\boldsymbol{\min f(w) = MSE(y, \hat{y}) + \lambda \sum_{j}^{d} \frac{|w_j|}{p_j}} \qquad (2)$$

Mean squared error (MSE) was used as the loss function, consistent with the classification version. We employed PGD to address the non-differentiability of the L1-norm, and optimized the parameters using RMSprop. The lambda tuning strategy remains consistent with the classification version, employing k-fold CV scanning. The optimal lambda is determined by segmented regression on the MSE loss curve.

Supplementary Note 3: Benchmarked methods

The benchmarked methods used in this study are listed below:

1. ALDEx2[1]: This method estimates abundance from count data using Monte Carlo sampling to generate a Dirichlet distribution with a uniform prior for each sample. It employs the centered log-ratio (CLR) transformation for scale invariance and sub-compositional coherence. Feature significance is evaluated using Wilcoxon tests, with Benjamini-Hochberg (BH) adjusted p-values. Significant features are selected based on a p-value threshold of 0.05.
2. ANCOM2[2]: This framework addresses sparsity in abundance analysis by ignoring zeros. Outlier zeros and structural zeros are identified, and a pseudo count is applied to the dataset for additive log ratios of features. The Wilcoxon rank-sum test examines significance, and p-values are adjusted using the BH method and using 0.05 as threshold.
3. edgeR[3]: This method applies pseudo count addition and relative log expression scaling to the raw count table. The exactTest function is used on negative binomial data for feature identification, with adjusted p-values corrected using the BH method. Features with corrected p-values lower than 0.05 are selected.
4. LEfSe[4]: Raw count datasets are transformed into frequencies by dividing each feature count by the total library size. The effect size is calculated with linear discriminant analysis (LDA), and significance is estimated using the Wilcoxon rank-sum test. The default thresholds for feature selection are LDA > 2.0 and p-value < 0.05.
5. metagenomeSeq[5]: The raw count table was normalized using cumulative-sum scaling (CSS), and a Zero-Inflated Log-Normal mixture model was fitted for each feature. The p-values were adjusted using the BH method. Significant features were selected based on a corrected p-value threshold of 0.05.
6. NBZIMM[6]: NBZIMM comprises two integral components. Firstly, a logistic model predicts excess zeros, while the second component employs a negative binomial distribution to model dispersed counts. In the study, the raw count table is utilized directly for significant estimation. All samples are treated as independent subjects, and features are selected based on BH corrected p-values below 0.05.
7. LASSO[7]: This is conventional L1-regularization, which is based on the absolute size of the regression coefficients for each feature. The regularization term is implemented with logistic

regression, and features with non-zero coefficients are selected after training.

8. Elastic Net (EN)[8]: This hybrid approach combines the penalizations of L1 and L2 regularization from LASSO and ridge methods. Users can assign a ratio for using L1 and L2 regularization. Like LASSO, features with non-zero weights are selected after training.

9. Random Forest (RF)[9]: This ensemble method combines numerous individual binary decision trees by bootstrapping the sample set. Features with zero weight, not included in the model, are dropped.

10. eXtreme Gradient Boosting (XGBoost)[10]: This well-known ensemble method iteratively combines several random forests into a single strong learner. XGBoost with L1 and L2 regularization is conducted using the xgboost package.

11. Mutual Information (MI)[11]: This metric is calculated as the difference between the joint probability distribution's entropy and the sum of the marginal distributions' entropies. It is used to evaluate the relationship strength between the feature and the target variable.

12. mRMR[12]: This method selects features based on maximum relevance to the outcome variable and minimum redundancy with previously selected features. It assigns a weight to each feature to evaluate its relationship with the target variable.

13. Relief-F[13]: This method iteratively samples instances from the dataset and assigns a weight to each feature based on how well it differentiates the sampled instance from other instances in the dataset.

14. Fisher Score[14]: This measure calculates the ratio of between-class variance to within-class variance, providing a measure of each feature's discriminatory power to a particular outcome variable.

15. Feature Dispersion Criterion (FDC)[15]: This unsupervised method estimates each feature's importance by measuring its dispersion. It computes the relevance criterion for a feature by dividing the arithmetic mean (AM) by the geometric mean (GM).

Computational Issues Encountered

During the analysis, we encountered several computational issues. ALDEx2 encountered memory exhaustion on six larger datasets (GWMC_ASIA_NA, GWMC_HOT_COLD, hiv_dinh, ob_zupancic, Office, and t1d_alkanani), despite utilizing a machine with 256GB of RAM. ANCOM2 took an excessively long time to run on three datasets (melanoma_matson, ob_zupancic, and sw_plastic_frere), exceeding five days, which led us to terminate the computations. LEfS failed to run on four larger datasets (ArcticTransects, GWMC_ASIA_NA, GWMC_HOT_COLD, and ob_zupancic). While edgeR, metagenomeSeq, and NBZIMM successfully processed 42 datasets, but metagenomeSeq failed to identify significant features in nine datasets (art_scher, asd_son, BISCUIT, cdi_vincent, melanoma_matson, melanoma_mcculloch, ob_zupancic, par_scheperjans, and t1d_alkanani). All methods successfully processed all datasets. However, Elastic Net (EN) did not select any features in the hiv_lozupone dataset.

References

1 Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 1-13 (2014).

2 Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology* **8**, 2114 (2017).

3 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, 1-9 (2010).

4 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome biology* **12**, 1-18 (2011).

5 Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10**, 1200-1202 (2013).

6 Zhang, X. & Yi, N. NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC bioinformatics* **21**, 1-19 (2020).

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

8 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301-320 (2005).

9 Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).

10 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

11 Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Physical review E* **69**, 066138 (2004).

12 Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-

dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* **27**, 1226-1238 (2005).

13 Kononenko, I., Šimec, E. & Robnik-Šikonja, M. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence* **7**, 39-55 (1997).

14 Gu, Q., Li, Z. & Han, J. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725* (2012).

15 Artur Ferreira, M. a. F. in *proceedings, European Symposium on Artificial Neural Networks* (Bruges, 2011).