# Supplemental Material: Evaluating trait-based sets for taxonomic enrichment analysis applied to human microbiome data sets

Quang P. Nguyen[1,2,*], Anne G. Hoen[1,2,+.*], H. Robert Frost[2,+,*]

**1 Department of Epidemiology, Geisel School of Medicine at Dartmouth College**
**2 Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth College**

**+ These authors jointly supervised this research**
*** Co-corresponding authors: quangpmnguyen@gmail.com, hildreth.r.frost@dartmouth.edu, anne.g.hoen@dartmouth.edu**

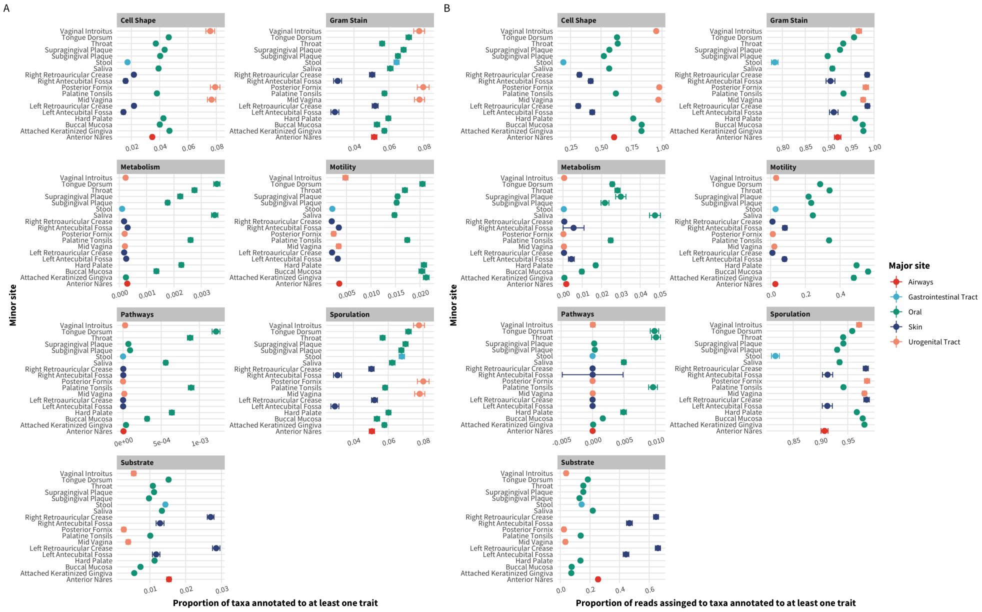

**Figure 1.** Trait coverage statistics for samples profiled with 16S rRNA gene metabarcoding. Panel **(A)** illustrates the proportion of present taxa per sample annotated to at least one trait. Panel **(B)** illustrates the proportion of reads assigned to taxa annotated to at least one trait which accounts for taxa relative abundances. Each plot facet represents different trait categories that were evaluated. Error bar represents the standard error of the evaluation statistic of interest across the the total number of samples evaluated per body site.

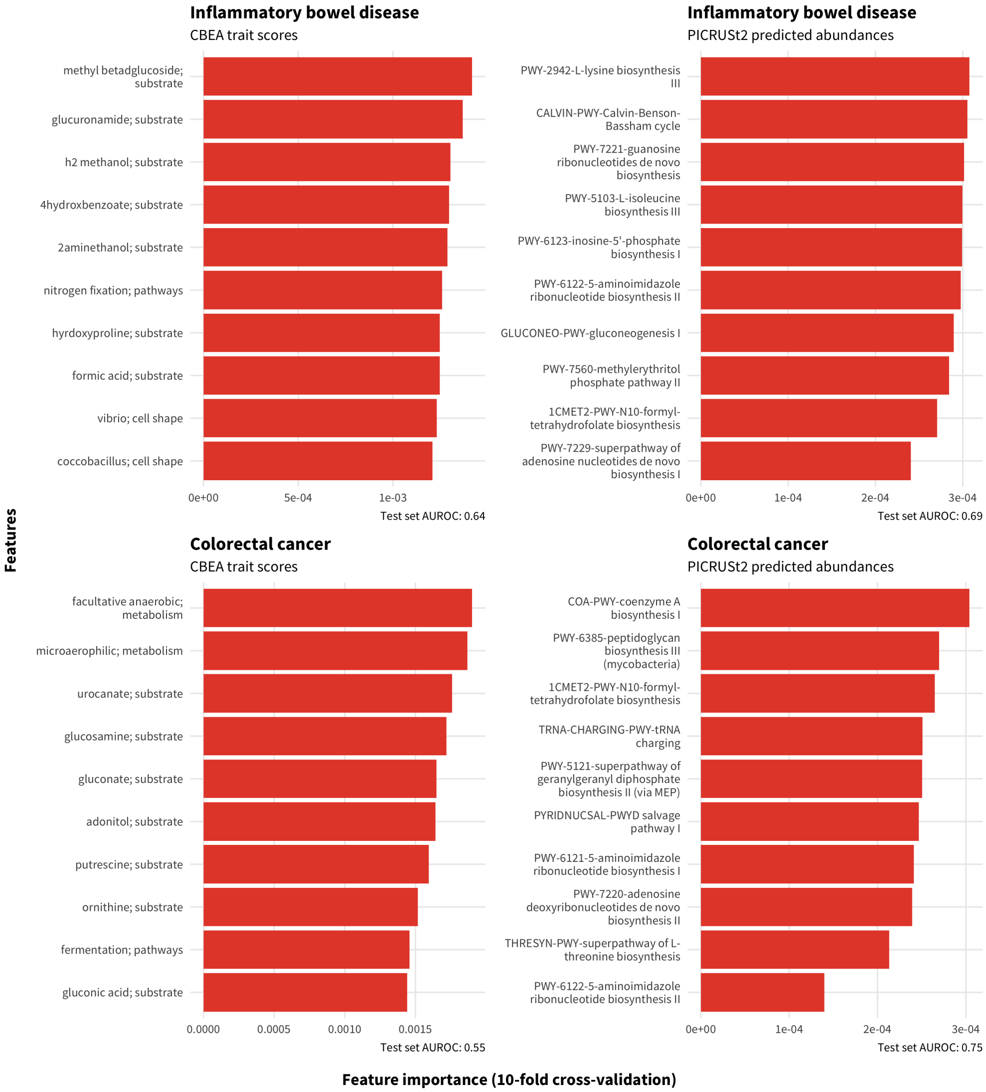

**Figure 2.** Top 10 important features based on random forest model fitted on different inputs from data sets profiled with 16S rRNA gene sequencing. Features were selected from mean decrease in Gini impurity averaged across 500 decision trees and 10-fold cross-validation (nested with the training set) as implemented in `scikit-learn`. AUROC scored on a held-out test set is also presented for each input type and disease condition.