Taxonomic Resolution Matters for Microbiome-Based Classification of Colorectal Cancer

Courtney R. Armour, Begüm D.Topçuoğlu, Andrea Garretto, Patrick D. Schloss †

† To whom correspondence should be addressed:

pschloss@umich.edu

Department of Microbiology

University of Michigan

Ann Arbor, MI 48109

observation format - max 1200 words, 2 figures, 25 ref

10 **Abstract (max 250 words)**

11 **Importance (max 150 words)**

Introduction (~250 words)

Colorectal cancer is one of the most common cancers in men and women and a leading cause of cancer related deaths in the United States (1). Early detection and treatment are essential to increase survival rates, but for a variety of reasons including the invasiveness and high cost of screening (i.e. colonoscopy), many people do not comply with recommended screening guidelines (2) prompting a need for low cost, non-invasive detection methods. A growing body of research points to the gut microbiome as a promising target for non-invasive screening to detect screen relevant neoplasias (neoplasms?) (SRNs) consisting of pre-cancerous polyps (i.e. advanced adenomas) and carcinomas. Efforts to realize the diagnostic potential of the gut microbiome in detecting SRNs have focused on machine learning (ML) methods using abundances from operational taxonomic unit (OTU) classifications based on amplicon sequencing of the 16S rRNA gene {}. However, whether this is the optimal taxonomic resolution for classifying SRNs from microbiome data is unknown. Additionally, recent work has pushed for the use of amplicon sequence variants (ASVs) to replace OTUs for marker-gene analysis because of the improved resolution with ASVs (3). However, whether the additional resolution provided by ASVs is useful for ML classification is unclear {pat preprint?}. Since ML classification relies on consistent differences between groups, its possible that the resolution at the ASV level is too individualized to accurately differentiate groups. Topçuoğlu *et al* (4) recently demonstrated effective application of machine learning (ML) to microbiome based classification problems and developed a framework for applying ML practices in a more reproducible way (mikropml). This analysis utilizes the reproducible framework developed by Topçuoğlu *et al* to quantify which ML method and taxonomic level produce the best performing classifier for detecting SRNs from microbiome data.

Results (~700 words)

Utilizing publicly available 16S rRNA sequence data from stool of patients with SRNs and healthy controls, we generated abundance tables with Mothur {} annotated to Phylum, Class, Order, Family, Genus, OTU and ASV levels. Using the taxonomic abundance data, we quantified how accurately samples could be classified as “normal” or “SRN” (i.e. advanced adenoma or carcinoma) using five machine learning methods with the Mikropml R package (methods). Across the five machine learning methods tested, model performance tended to increase with taxonomic level usually peaking around genus/OTU level before dropping off slightly with ASVs (Supplemental Figure 1). Random forest (RF) was consistently the top performer at most taxonomic levels. Within the RF model, the highest AUROCs were observed for family (median AUROC: 0.687), genus (median AUROC: 0.686), and OTU (median AUROC: 0.698) level data with no significant

difference between the three (Figure 1). Performance with ASVs (median AUROC: 0.676) was significantly lower than OTU ($p < 0.01$), but borderline equivalent to family ($p = 0.06$) and genus ($p = 0.05$) levels. These results were consistent for ASVs generated with DADA2 (5) (median AUROC: 0.66). These results suggest that finer taxonomic resolution is not necessarily better for differentiating a persons cancer status from the microbiome.

While AUROC values between models is a useful way to compare the overall model performance across all thresholds, they can be misleading (6). Depending on the intended implementation of the model, one may want to optimize the true-positive rate(or sensitivity) over the false-postivite rate (or specificity). To further compare the model performance across taxonomic levels we compared the sensitivity of the models

- on average, how many positives are missed with ASV?

One hypothesis for the observation that model performance increases from phylum to OTU level then drops slightly at ASV level is that at higher taxonomic levels (e.g. phylum) there are too few taxa and too much overlap to reliably differentiate between cases and controls. At the level of genus/OTU data there is enough data and variation but at the ASV level, the data is too specific to individuals and doesn't overlap enough. Examination of the prevalence of taxa in samples at each level supports this idea. A majority of taxa are present in greater than 75% of samples at the phylum (67% of taxa) and class (63% of taxa) levels. The opposite is observed at the OTU and ASV level where 60% and 53% of taxa respectively are only present in less than 25% of samples (Supplemental Figure 3). While the resolution provided by ASVs is useful in

TODO: what are important taxa, patterns along levels (e.g fuso always in top)

Of note, the ML pipeline includes a pre-processing step that occurs prior to training and classifying the ML models (methods). As part of this step, features are removed that will not provide useful information to build the model. For example, strongly correlated features provide the same information to build the model and thus can be collapsed. Additionally, features with zero or near-zero variance will not help the model differentiate groups and thus can be removed. Interestingly, despite starting with 104106 features at the ASV level, only 478 (0.5%) remained after pre-processing. At the OTU level, 20079 of the 705 features (3.5%) remained after preprocessing (Table 1).

TODO: check why ASVs were removed in preprocessing (nzv?)

(depending on answer can discuss why features were removed - e.g. only in one or a few samples)

Mikropml{}

Discussion/Conclusions (~250 words)

These results demonstrate that consideration of the appropriate taxonomic resolution for utilizing the microbiome as a predictive tool. In general, we found that finer resolution (e.g. ASVs) does not add additional sensitivity to model prediction and at the ASV level actually impedes model performance due to the sparsity of shared taxa.

- Utilizing genus or family level data could allow for merging data generated from different 16S regions or sequencing platforms. (is this true?)

While predictive models utilizing microbiome data alone are not that great, microbiome in combination with FIT results are better than either alone {}. This analysis wanted to quantify difference between

Materials and Methods

Dataset. Raw 16S rRNA gene amplicon sequence data isolated from human gut samples {Baxter} was downloaded from NCBI SRA (accession #). This dataset contains stool samples from 490 subjects. Based on the available metadata, samples categorized as normal, high risk normal, or adenoma were labeled "normal" for this analysis and samples categorized as advanced adenoma or carcinoma were labeled as "screen relevant neoplasia" (SRN). This resulted in a total of 261 "normal" samples and 229 "SRN" samples.

Data processing. Sequence data was processed with Mothur (1.44.3) using the SILVA reference database (v132) to produce count tables for phylum, class, order, family, genus, OTU, and ASV following the Schloss Lab MiSeq SOP described on the Mothur website {}. ASV level data was also produced using DADA2 to ensure consistent results with a different pipeline. Data was processed following the DADA2 pipeline, but setting pool=TRUE to infer ASVs from the whole dataset rather than per sample. The resulting ASV table was subsampled for consistency with the Mothur data.

Machine Learning. Machine learning models were run with the R package mikropml (v0.0.2) {} to predict the diagnosis category (normal vs SRN) of each sample. Data was preprocessed to normalize values (scale/center), remove values with zero or near-zero variance, and collapse collinear features using default parameters. Initially the models were run with default hyperparameters, but were expanded if the peak performance was at the edge of the hyperparameter range. Each taxonomic model taxonomic level combination (e.g. Random Forest on genus) was run with 100 different seeds (1-100). Each seed split the data

into a training (80%) and testing (20%) set, and output performance of the training and testing as area under the receiver operating curve (AUROC).

To compare performance between taxonomic levels and models, P values were calculated as previously described {begum}. Another output from the mikropml package is the permuted feature importance which is quantified by iteratively permuting each feature in the model and assessing the change in model performance. Features are presumed to be important if the performance of the model, measured by the AUROC, decreases when that feature is permuted. Ranking of feature importance was determined by ordering the features based on the change in AUROC where features with a larger decrease in AUROC are ranked higher in importance.

To quantify prevalence of the features, the number of samples with non-zero abundance was divided by the total number of samples resulting in values ranging from 0 to 1 where 0 indicates the feature is not found in any samples, 0.5 indicates the feature is found in half of the samples, an 1 indicating the feature is found in all of the samples.

All code is available at: **TODO: link to code**

Acknowledgements

Figures

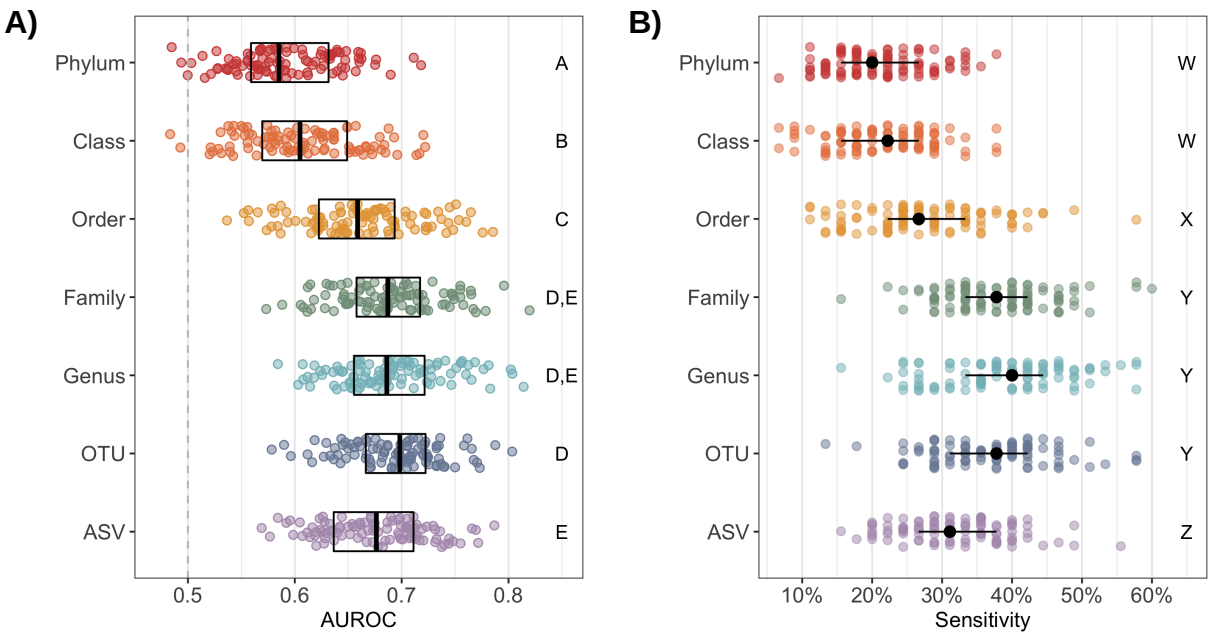


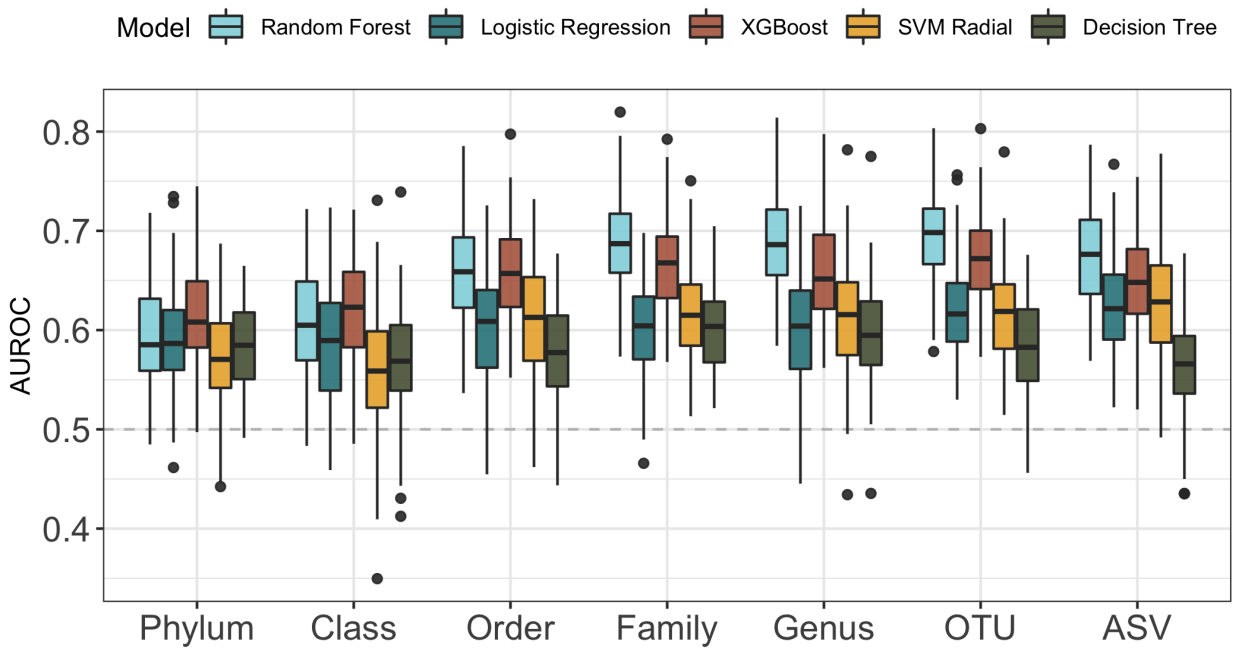
Figure 1: Random Forest Model Performance. A) Boxplots with points of area under the receiver operating characteristic curve (AUROC) values on the test dataset for 100 seeds predicting SRNs using a Random Forest model. Dashed line denotes AUROC of 0.5 which is equivalent to random classification. Significance between taxonomic levels was quantified by comparing the difference in mean AUROC and is denoted by letters A through E on the right side of the plot; taxonomic levels with the same letter are in the same significance group and are not significantly different from one another. **B)** Strip plot of the sensitivity at a specificity of 90% across the 100 model iterations. Black points denote the median and the lines denote the IQR. The letters W through Z denote the significance groups.

Tables

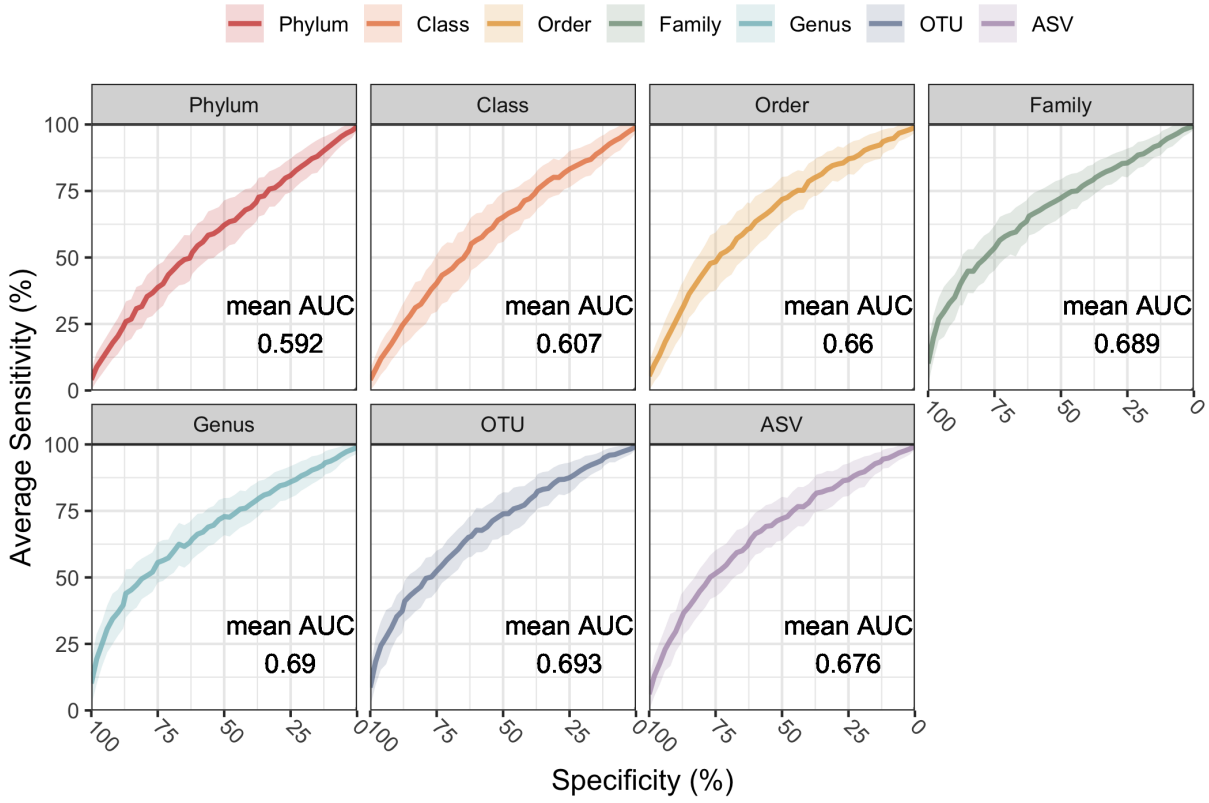
Taxonomic Level	Number of Features	Number of Features After Preprocessing	Percent of Features Kept After Preprocessing
Phylum	19	9	47.4 %
Class	36	19	52.8 %
Order	65	28	43.1 %
Family	124	54	43.5 %
Genus	316	115	36.4 %
OTU	20079	705	3.5 %
ASV	104106	478	0.5 %

Table 1: Summary of Features. Overview of the number of features at each taxonomic level before and after preprocessing as described in the methods.

Supplemental Figures



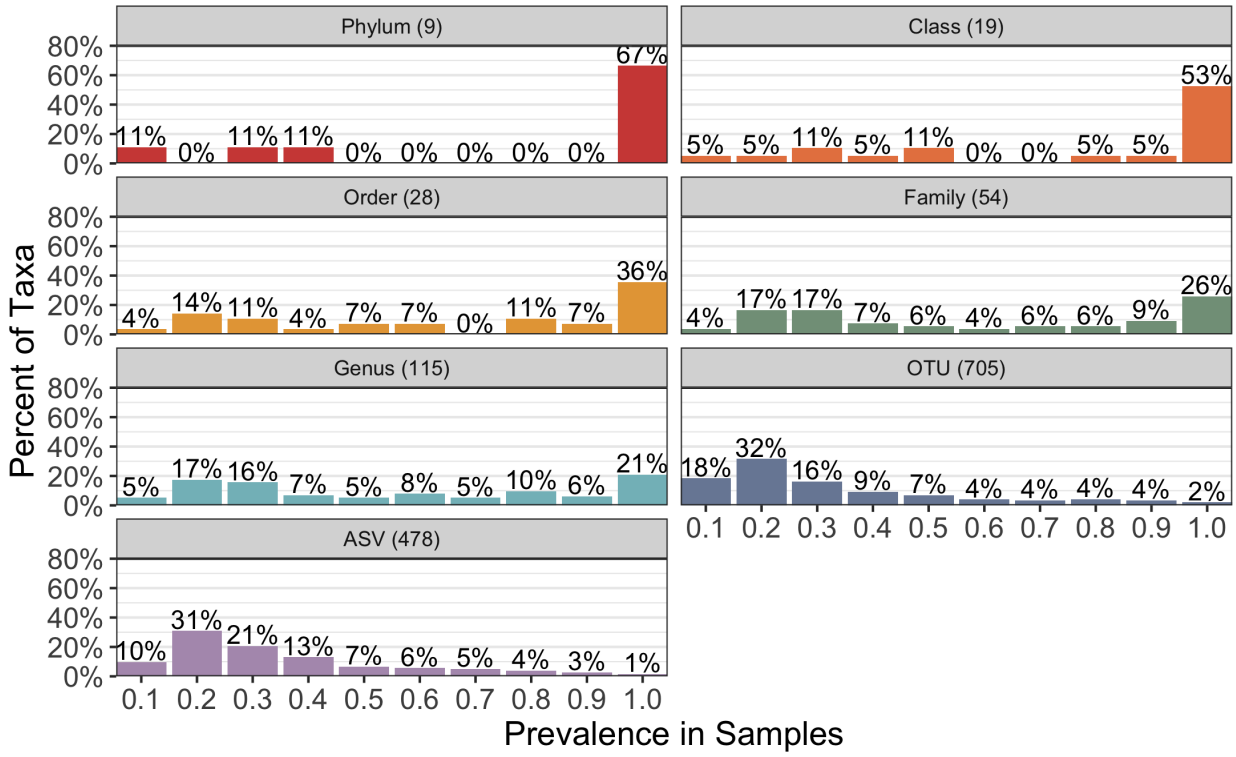
Supplemental Figure 1: Model Performance across Taxonomy. Boxplot of AUROC values on the test dataset for 100 seeds for each model type across all taxonomic levels.



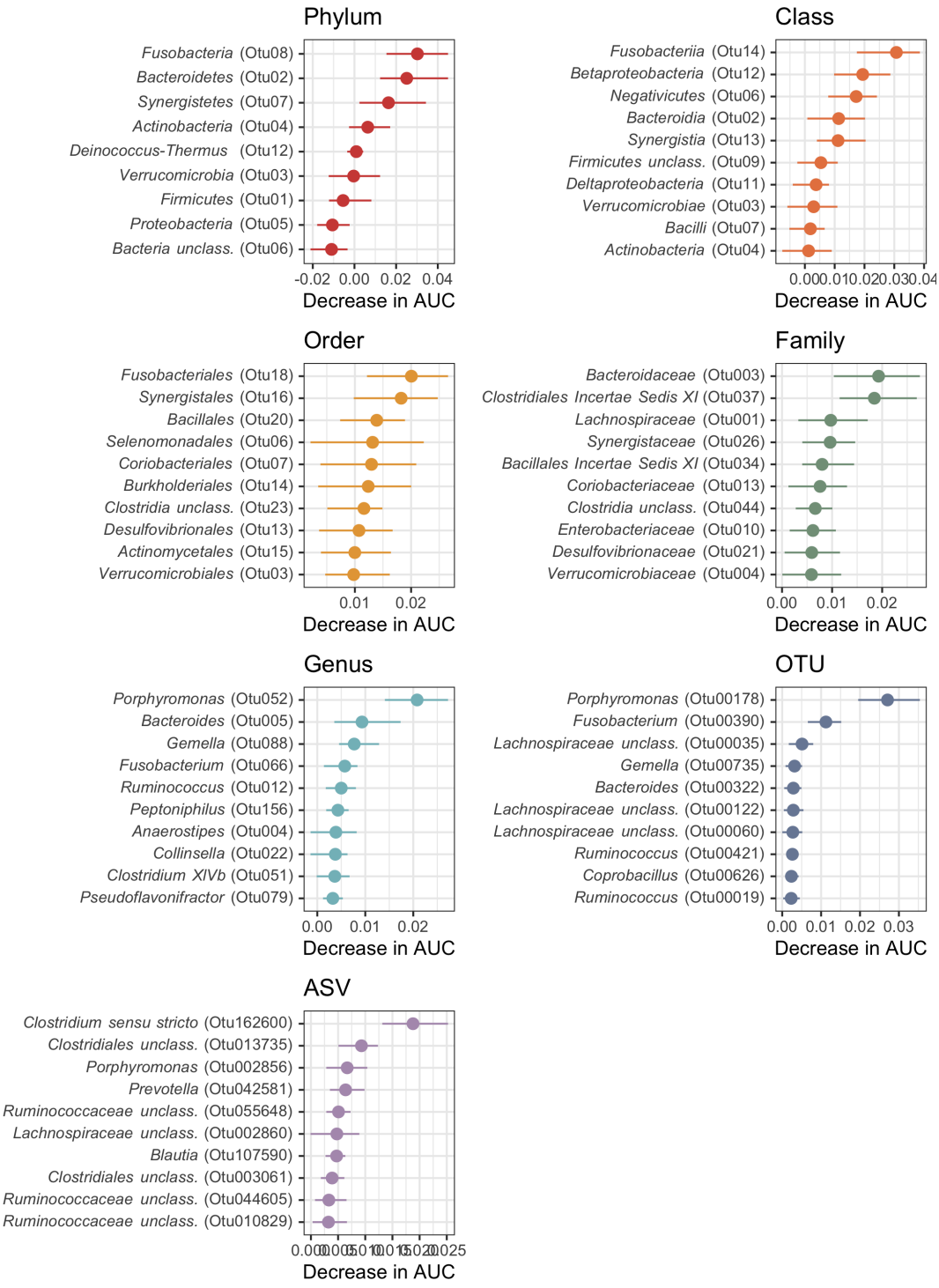
130

131 **Supplemental Figure 2: Averaged ROC curves** ROC curves averaged across the 100 iterations of the

132 model. The shaded region represents the standard deviation form the mean.



Supplemental Figure 3: Prevalence of Taxa in Samples. Distribution of taxa prevalence in samples at each taxonomic level.



Supplemental Figure 4: Top 10 important taxa at each taxonomic level. Summary of the 10 most important taxa for the random forest models at each taxonomic level based on the average decrease in AUC when the feature is permuted.

References

1. **Siegel RL**, **Miller KD**, **Sauer AG**, **Fedewa SA**, **Butterly LF**, **Anderson JC**, **Cercek A**, **Smith RA**, **Jemal A**. 2020. Colorectal cancer statistics, 2020. CA: A Cancer Journal for Clinicians **70**:145–164. doi:https://doi.org/10.3322/caac.21601.

2. **García G**, **Z A**. 2011. Factors Influencing Colorectal Cancer Screening Participation. Gastroenterology Research and Practice **2012**:e483417. doi:10.1155/2012/483417.

3. **Callahan BJ**, **McMurdie PJ**, **Holmes SP**. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal **11**:2639–2643. doi:10.1038/ismej.2017.119.

4. **Topçuoğlu BD**, **Lesniak NA**, **Ruffin MT**, **Wiens J**, **Schloss PD**. 2020. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. mBio **11**. doi:10.1128/mBio.00434-20.

5. **Callahan BJ**, **McMurdie PJ**, **Rosen MJ**, **Han AW**, **Johnson AJA**, **Holmes SP**. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods **13**:581–583. doi:10.1038/nmeth.3869.

6. **Lobo JM**, **Jiménez-Valverde A**, **Real R**. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography **17**:145–151. doi:10.1111/j.1466-8238.2007.00358.x.