Optimal Taxonomic Resolution for Microbiome-Based Classification of Colorectal Cancer

Other title options:

ASV level resolution is unnecessary for prediction of SRNs from microbiome data

OTUs are the Optimal Taxonomic Level 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

Abstract (max 250 words)

Importance (max 150 words)

Introduction

- Colorectal cancer (CRC) is one of the most common cancers in men and women and a leading cause of cancer related death in the United States {}.
- 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 {}.
- Less invasive and more affordable screening methods (e.g. fecal immunochemical test) are available, however these tests are less sensitive than colonoscopies, especially for detecting early stage adenomas.
- A growing body of research points to the gut microbiome as having a role in tumor development and progression {}. For example, studies find that Fusobacterium nucleatum and enterotoxigenic Bacteriodes fragilis tend to be enriched in the gut microbiome of subjects with CRC relative to healthy controls {}, while potentially protective bacteria such as members of Lachnospiraceae tend to be depleted {}.
- The gut microbiome shows promise as diagnostic {}
- Efforts to realize the diagnostic potential of the gut microbiome in detecting screen relevant neoplasias (SRNs) have focused on machine learning (ML) methods using abundances from operational taxonomic unit (OTU) classifications based on the amplicon sequencing of the v4 region 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. However, whether the additional resolution provided by ASVs is useful for ML classification is unclear.
    * 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* (mBio 2020) 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

- Across the five ML methods tested, model performance tended to increase with taxonomic level usually peaking around genus/OTU level before dropping off slightly with ASVs. (Figure)
- Random forest was consistently the top performer at most taxonomic levels.
  - RF might be more appropriate for microbiome analysis since its more suitable for zero inflated data
  - TODO: dig into literature on this
- Within the RF model, the highest AUCs were observed for family (median AUC: 0.687), genus(median AUC: 0.686), and OTU (median AUC: 0.698) level data with no significant difference between the three. (Figure)
  - ASV (median AUC: 0.676) performed significantly lower than OTU and genus ($p < 0.05$) but not family level $P = 0.06$).
  - TODO: compare dada2 performance
- 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.
  - As you reach genus/OTU level 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. (Figure)
- 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 wont 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)
  - TODO: check why ASVs were removed in preprocessing (nzv?)
  - (depending on above can discuss why features were removed - e.g. only in one or a few samples)

- OTU level data provides an idea balance of overlap and distinction to classify samples.
- Overall, the fine resolution of ASVs is unnecessary for ML classification with microbiome data since the vast majority of the ASVs are removed during preprocessing and the model performance is diminshed compared to other taxonomic levels.

other items that could be addressed:

- what are important taxa, patterns along levels (e.g fuso always in top)
- abundance of important taxa: many of the top important features are low abund (Figure)

Conclusions

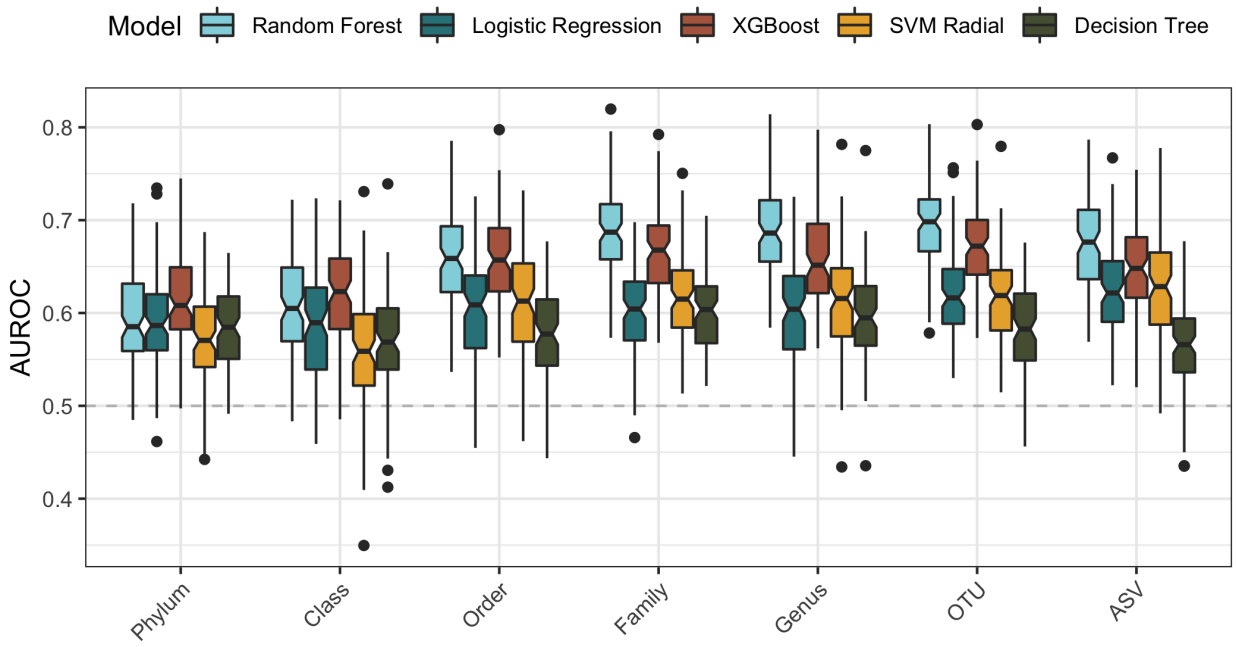
Materials and Methods

- 16S rRNA data from 490 subjects {baxter}
  - 261 controls
  - 229 SRNs (describe how SRN defined)
- processed with mothur v1.44.3
  - SILVA v132 reference
- ML with mikropml package
  - preprocessing -normalize values
    - what is removed and why
      - near zero variance & zero variance removed
      - correlated collapsed
  - hp tuning
  - n seeds
  - for more detail see mikropml package{}
- pvalues calculated as previously described {begum}
- prevalence = number of samples with non-zero abundance / total number of samples

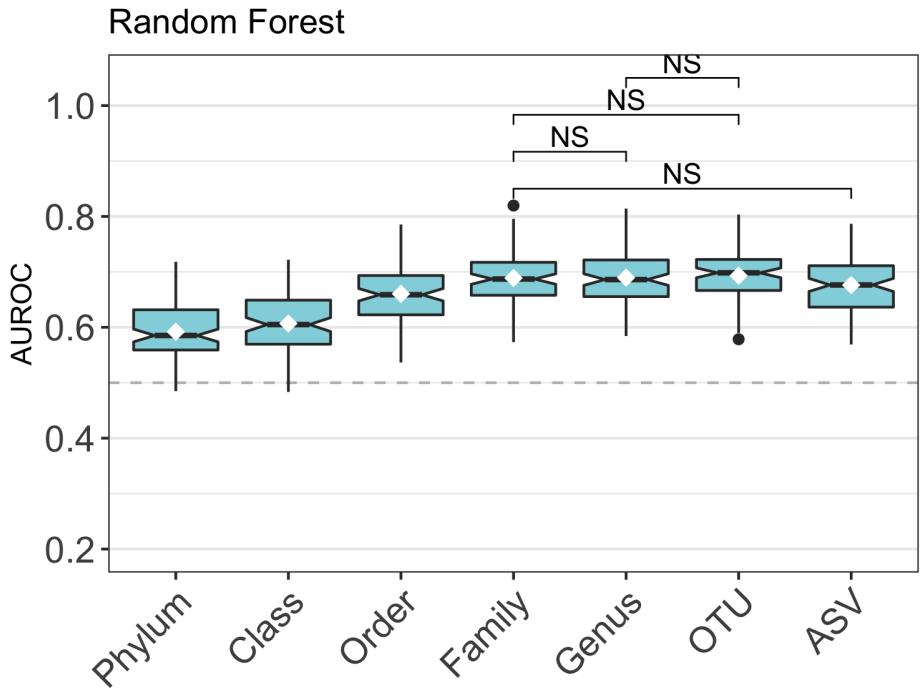
Acknowledgements

Figures

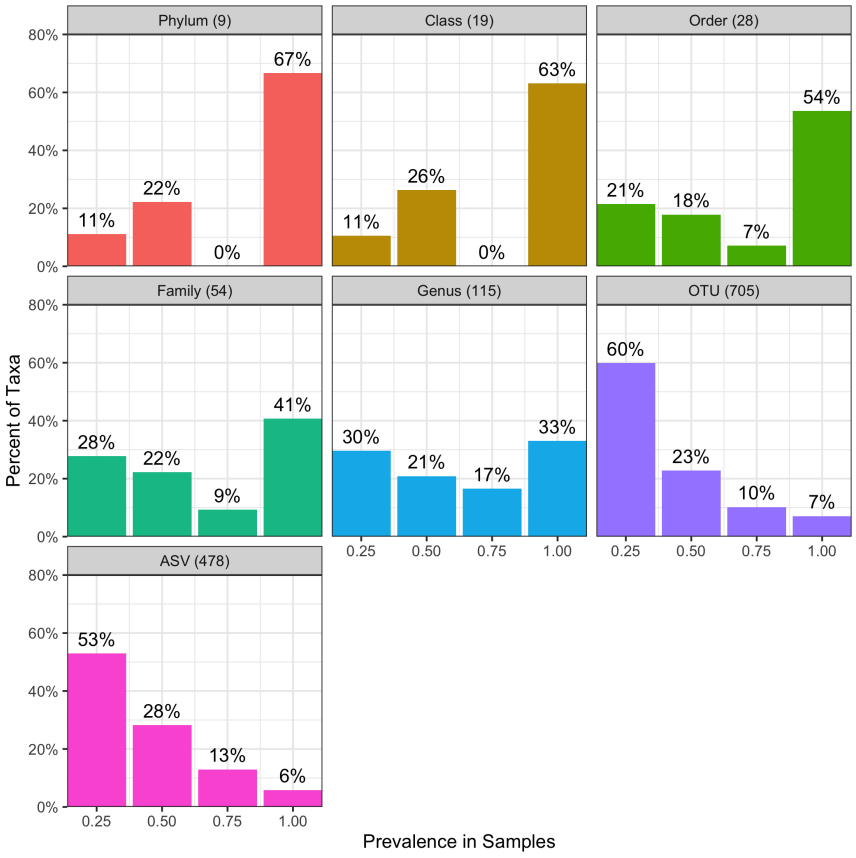
Model Performance across Taxonomy



Random Forest Model Performance with Significance



Prevalence of Taxa in Samples



Summary of Features

level	n_samples	n_features	n_features_preproc	pct_kept
phylum	490	19	9	47.4
class	490	36	19	52.8
order	490	65	28	43.1
family	490	124	54	43.5
genus	490	316	115	36.4
otu	490	20079	705	3.5
asv	490	104106	478	0.5

Relative Abundance of Important Features