TITLE GOES HERE

Running title: Optimal resolution

Courtney R. Armour, (ande? begum? etc?), 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

- CRC is one of the most common cancers and a leading cause of cancer related death
- There is evidence that the microbiome has a role in CRC development/progression and could be useful for biomarker detection and diagnostics.
- Begum 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).
- A common question when applying ML methods to microbiome data is which method and taxonomic level should be use.
- This analysis utilizes the reproducible framework developed by Begum et al to quantify which ML method and taxonomic level produce the best performing classifier for CRC data.

Results

- Of the five ML methods tested, Random forest was consistently the top performer (supplemental figure of all models?) at most taxonomic levels.
  - RF might be more appropriate anyways since its more suitable for zero inflated data? (need to look into literature)
- Within the RF model, the highest AUCs were observed for family, genus, and otu level data with no significant difference between the three. (Figure 1)

Conclusion

Materials and Methods

- data from prior study {baxter}
- mikropml package
- pvalues as previously described {begum}

Acknowledgements

Figures