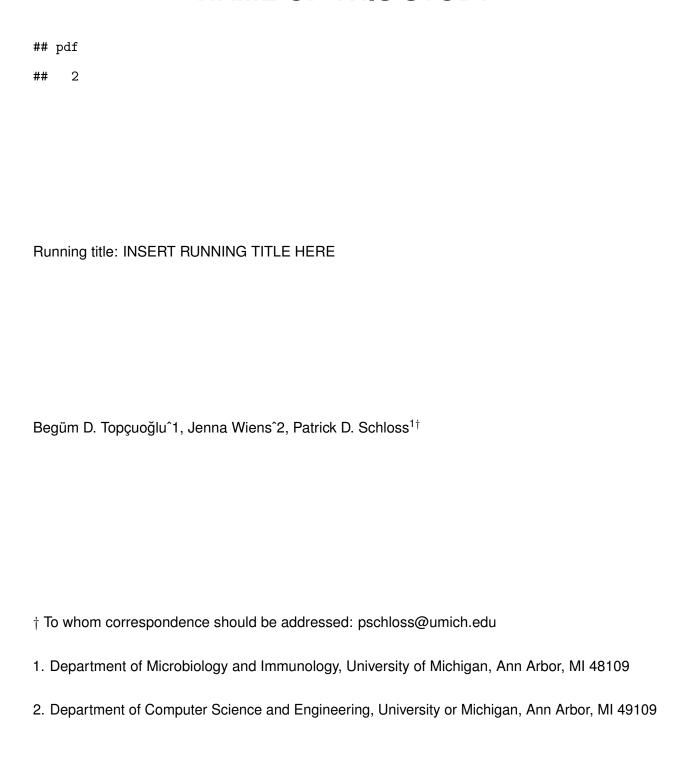
NAME OF THIS STUDY



1 Abstract

Introduction

- 3 As gut microbiome field continues to grow, there will be an ever-increasing demand for reproducible
- 4 machine learning methods to analyze microbiome sequence read count data and to determine
- association with a continuous or categorical phenotype of interest.
- 6 Colorectal cancer is one of the leading cause of death among cancers in the United States. Early
- diagnosis increases the chance of survival. However the current diagnostic methods are expensive
- and invasive. As a less invasive tool, numerous studies use relative abundances of the gut bacteria
- 9 populations to predict disease progression. Most microbial communities are pretty patchy and the
- likelihood of a single feature that explains the differences in health is pretty small. It is likely that
- many biomarkers are needed to account for the patchiness as well as the context dependency of
- 12 the features.
- 13 ML use in microbiome literature is a bit like the wild west with lack of clarity over methods,
- testing, validation, etc. There is a need for guidance on how to properly implement these different
- methods. We need to emphasize good machine learning practices and pipelines and discuss the
- reproducibility, robustness and actionability of models.
- We established a non-leaky pipeline. We performed L1 and L2-regularized logistic regression,
- Linear SVM, Non-Linear SVM, Decision tree, Random forest, XGBoost and Feed Forward Neural
- 19 Net classification models. We evaluated the classification performance of different machine learning
- 20 methods. We also want to discuss the reproducibility, robustness, actionability, interpretibility and
- susceptibility to overfitting of each method.
- 22 Generalisation Perfomance of each model. Is there a maximum threshold of prediction with all
- these methods? Does an increase in model complexity improve predictibility? Synthesis statement
- 24 regarding modeling 16S microbiome data

- 25 Results and Discussion
- 26 Conclusions
- 27 Materials and Methods

- Insert figure legends with the first sentence in bold, for example:
- 29 Figure 1. Number of OTUs sampled among bacterial and archaeal 16S rRNA gene
- sequences for different OTU definitions and level of sequencing effort. Rarefaction curves
- or different OTU definitions of Bacteria (A) and Archaea (B). Rarefaction curves for the coarse
- environments in Table 1 for Bacteria (C) and Archaea (D).

33 References