A search for a proper taxonomic resolution in order to accurately classify colorectal cancer

Anonymous Student Contributor

4/22/2022

Contents

Introduction

Colorectal cancer is one of the more common cancers in the world. If caught early, colorectal cancer is highly preventable. However, the detection of this cancer requires invasive, time consuming and expensive procedures and because of this, many cancer diagnosis are made too late and result in deaths that early detection could have prevented. Thus, a simpler method to detect colorectal cancer, resulting in earlier detection, would lead to greater survivability in patients diagnosed with it.

Background

Machine learning classification models have been used on gut microbiome data at operational taxonomic unit (OTU) levels in order to predict the presence of screen relevant colonic neoplasia. Recently, some research has suggested that lowering the taxonomic resolution to the level of amplicon sequence variants (ASV) may lead to better classification models. In this work, the authors attempt to answer whether or not ASV level resolution leads to better results.

Biological hypothesis

Does performance of classification models increase as the taxonomic resolution decreases.

Significance

Obtaining the best possible models would maximize the number of true-positive results generated by these models and would then lead to a larger number of correct diagnosis.

Materials and Methods

The Mothur software was used for generating taxonomic abundance data. The Mikropml R package was used to create the following models: Random forest, L2-Regularized logistic regression, Decision trees, XGBoost, Support vector machine

and also for the following data processing: Normalization, 0 and near 0 variance removal and collapsing colinear features

The DADA2 R package was used for generating an alternative set of ASVs in order to validate the correctness of ASVs generated by Mothur.

Samples

Raw 16S rRNA gene amplicon sequences from 490 human stool samples, split into 261 normal and 229 screen relevant neoplasia samples. The normal samples served as the controls.

Experimental procedure

The samples used in this work come from publicly available data and downloaded from NCBI. Each sample was then categorized into "normal" or "screen relevant neoplasia" categories. The taxonomic abundance tables were generated using Mothur. Then the machine learning models were run using the mikropml R package to learn a model classifying the data into one of the diagnosis categories. The performance of each model was then evaluated using the area under the receiver operator curve metric.

Results

The performance of each of the models increased when moving in resolution from the Phylum, Class and Order levels down to the the Family, Genus and OTU levels, which saw similar performances, and dropped off at the ASV level.

```
### FUNCTIONS ########
read_combined <- function(tax_level, model){
  read_csv(paste0(path, "/combined-", tax_level, "-", model, ".csv"),
           col_types = c(method = col_character(),.default=col_double()),
           na = c("","NA","NaN")) %>% #some of the Pos/Neg Pred Values were NaN
    mutate(level=tax_level)
}

### VARIABLES ########
models <- c("rf","glmnet","xgbTree","svmRadial","rpart2")
levels <- c("phylum","class","order","family","genus","otu","asv")
levels_names <- c("Phylum","Class","Order","Family","Genus","OTU","ASV")
pal <- park_palette("Everglades",length(models))
names(pal) <- models
pal2 <- park_palette("Everglades",length(models))
names(pal2) <- c("Random Forest","Logistic Regression","XGBoost","SVM Radial","Decision Tree")

### MAIN ########
```

```r
# Generate dataframe of all results and tables of the median cv and test AUCs
full_df <- NULL
auc_df <- NULL

for( m in models ){
  model_df <- map_dfr(levels, read_combined, model = m)

  full_df <- bind_rows(full_df,model_df)
}

# adjust table
df_all <- full_df %>%
  select(cv_metric_AUC,AUC,method,level) %>%
  rename(test_AUC=AUC,cv_AUC=cv_metric_AUC) %>%
  pivot_longer(cols=c("cv_AUC","test_AUC"),names_to="auc_type",values_to="AUC") %>%
  mutate(level=factor(level,levels=c("phylum","class","order","family","genus","otu","asv"),
                      labels=c("Phylum","Class","Order","Family","Genus","OTU","ASV")),
         method=factor(method,levels=c("rf","glmnet","xgbTree","svmRadial","rpart2"),
                       labels=c("Random Forest","Logistic Regression","XGBoost","SVM Radial","Decision

#plot results
df_all %>%
  filter(auc_type == "test_AUC") %>%
  ggplot(aes(x=level,y=AUC,fill=method)) +
  geom_hline(yintercept = 0.5,color="grey",lty="dashed") +
  geom_boxplot(alpha = 0.9) +
  theme_bw() + xlab("") + ylab("AUROC") +
  theme(legend.position ="top",
        axis.text.x = element_text(angle = 0),
        axis.text = element_text(size = 14)) +
  scale_fill_manual(values=pal2,name="Model")
```

It appears that the lower resolution at the ASV level of data does not provide for better model results.

To further reinforce the point, a closer look at the results from one of the models is helpful and since the random forest model outperformed the other models, it can be examined. Figure 2 shows further the difference in the results and its mean when the model is trained on Family, Genus or OTU taxonomic levels and when trained on ASV level.

```r
pal1 <- park_palette("Everglades")

df_all %>%
  filter(auc_type == "test_AUC" & method == 'Random Forest') %>%
  ggplot(aes(x=level,y=AUC,fill = method)) +
  geom_violin(alpha = 0.9) +
  theme_bw() + xlab("") + ylab("AUROC") +
  theme(legend.position ="top",
        axis.text.x = element_text(angle = 0),
        axis.text = element_text(size = 14)) +
  scale_fill_manual(values = pal1, name = "Model") +
  geom_jitter() +
  stat_summary(fun = "median", geom = "crossbar", aes(color = "Median")) +
  stat_summary(fun = "mean", geom = "crossbar", aes(color = "Mean")) +
  scale_colour_manual(values = c("red", 'green'), name = '')
```

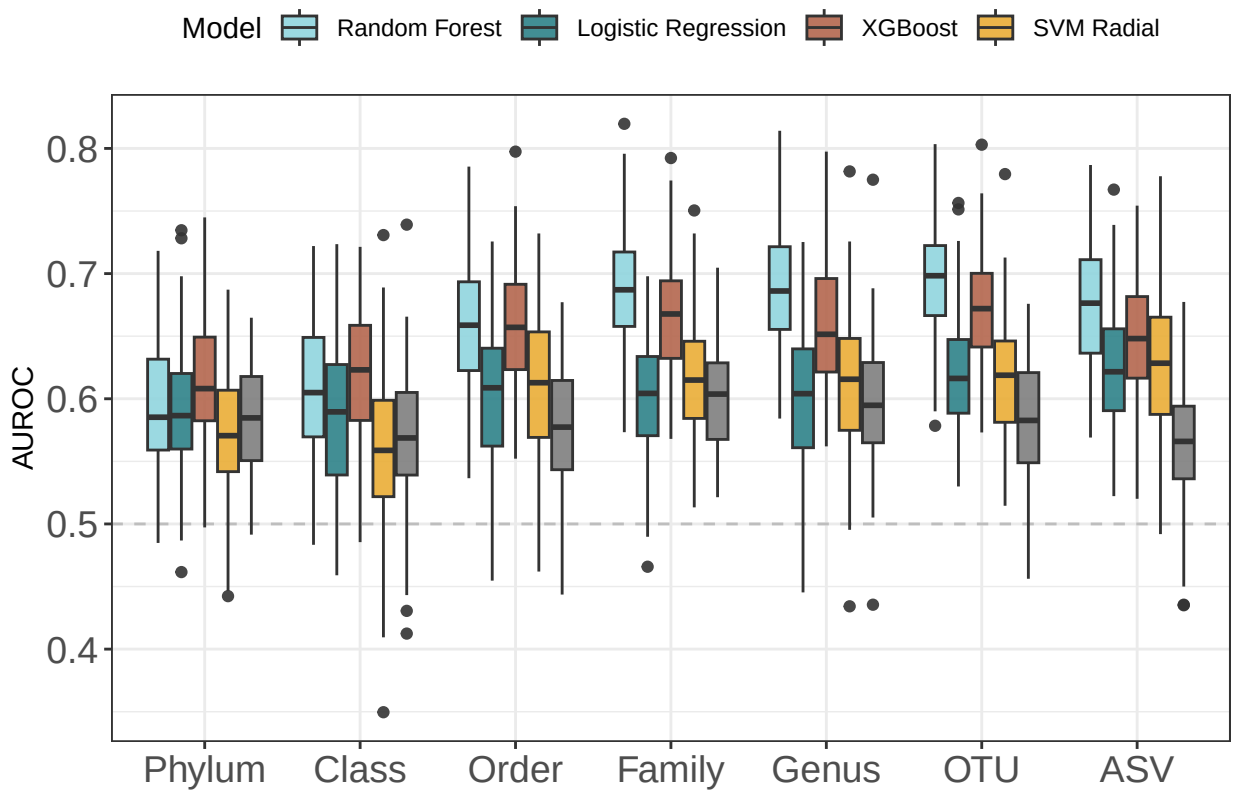

Figure 1: Model performance across taxonomy. Each boxplot shows AUROC values of each model per each taxonomy

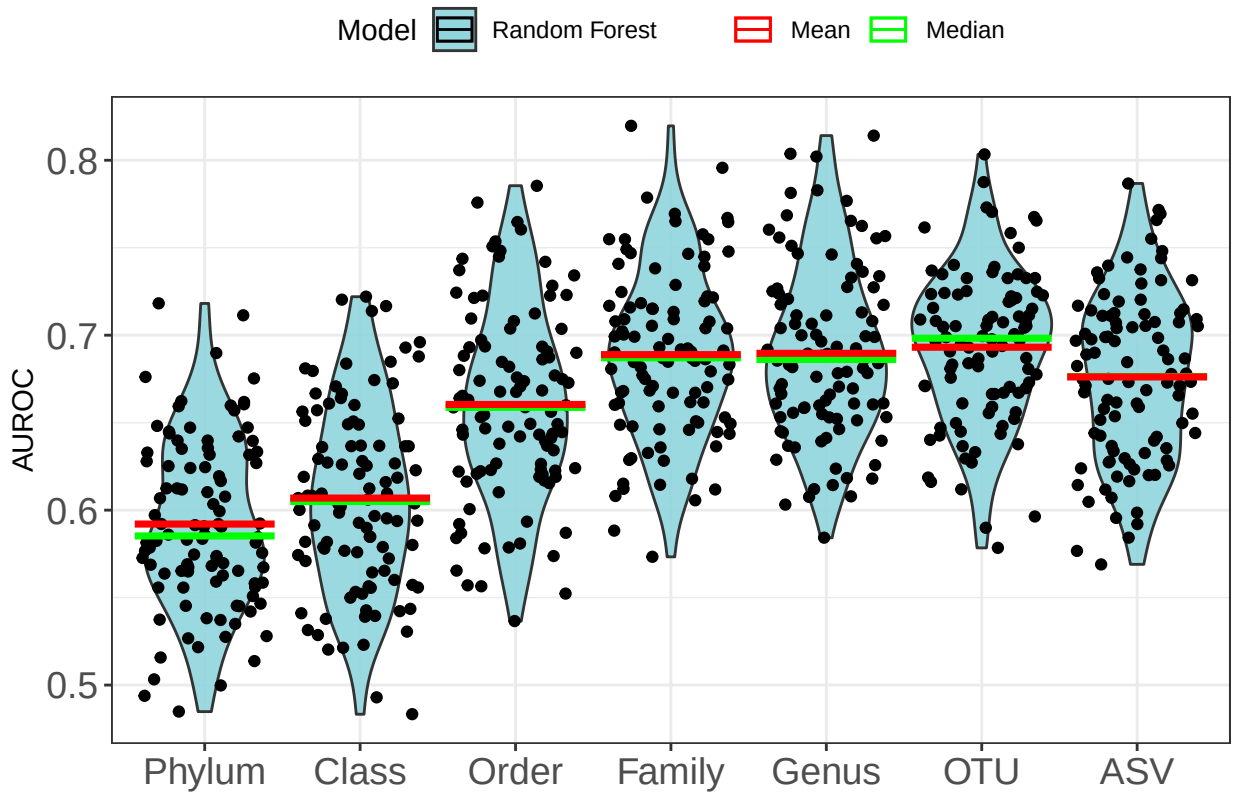

Figure 2: Random forest performance across taxas. The bars are the mean and the median and the black circles are each a result of 1 of 100 training runs.

Conclusions

Given the drop in model performance using the ASV taxonomic level, classifications screening for screen relevant neoplasia should continue to be done at the OTU levels instead. At least when using the five models tested in this work.

Biological conclusions

While the work here is focused on methods, there is an interesting question that can be raised based on the results. Since the ASV level represents a lower taxonony, does the interaction in the gut microbiome resulting in biomarkers for colorectal cancer happen at a taxonomic level higher than what ASVs represent? That is, are the markers for colorectal cancer a result of interactions at individual, or higher, taxonomic levels and not at a more molecular level?

Future work

This work tested five classification models only. There are other models that may yet yield better results using ASVs. Also, the ASV resolution may have lead to overfitting and more work can be done to reduce the effects of such.

References

A Goldilocks Principle for the Gut Microbiome: Taxonomic Resolution Matters for Microbiome-Based Classification of Colorectal Cancer Armour, C. R., Topçuoğlu, B. D., Garretto, A., & Schloss, P. D. (2022). Mbio, 13(1), e03161-21; DOI: 10.1128/mbio.03161-21

Code repository

https://github.com/SchlossLab/Armour_Resolution_mBio_2021

Data file

Data: https://github.com/SchlossLab/Armour_Resolution_mBio_2021/tree/master/data/process