Updated Models for CRC Diagnosis Using Microbiome Data

Niel

November 7, 2014

I am trying to develop logit models for distinguishing healthy patients from those with colorectal cancer based on the abundances of bacterial populations in their gut microbiome. To do this I would like to test all possible models containg 1-10 OTUs selected from the 309 most abundant OTUs. Unfortunately that isn't possible, due to the inordinantly large number of possible combinations.

The number of combinations can be calculated using the formula n!/((n-r)!r!) where n is the number of OTUs to choose from (309) and r is the number of OTUs chosen for the model (1 to 10). It can also be calculated with the choose() function in R. I'll calculate how many models need to be tested for each number of OTUs (r=1 to r=10)

```
for(r in 1:10){
  cat(r,' OTUs: ',choose(309,r), '\n')
  }
```

```
## 1  OTUs:  309
## 2  OTUs:  47586
## 3  OTUs:  4869634
## 4  OTUs:  372527001
## 5  OTUs:  2.272e+10
## 6  OTUs:  1.151e+12
## 7  OTUs:  4.984e+13
## 8  OTUs:  1.881e+15
## 9  OTUs:  6.292e+16
## 10  OTUs:  1.888e+18
```

That is way too many models, so I need a way to run fewer of them while still finding the best (or nearly the best) models (i.e. a heuristic.

This is the heuristic I decided to use. First I calculated all possible 3-OTU models (4.8696 × 106 combinations). Then, rather than test all possible 4-OTU models (3.7253 × 108), I took the top 100 of those models and sequentially added each of the 309 OTUs to them. This results approximately 3.09 × 104 models and saves lots of time. I then took the 100 best of thoe 4-OTU models and again added each of the 309 OTUs to them. I repeated this process up to 10-OTU models.

Here's an example for the 4 OTU models

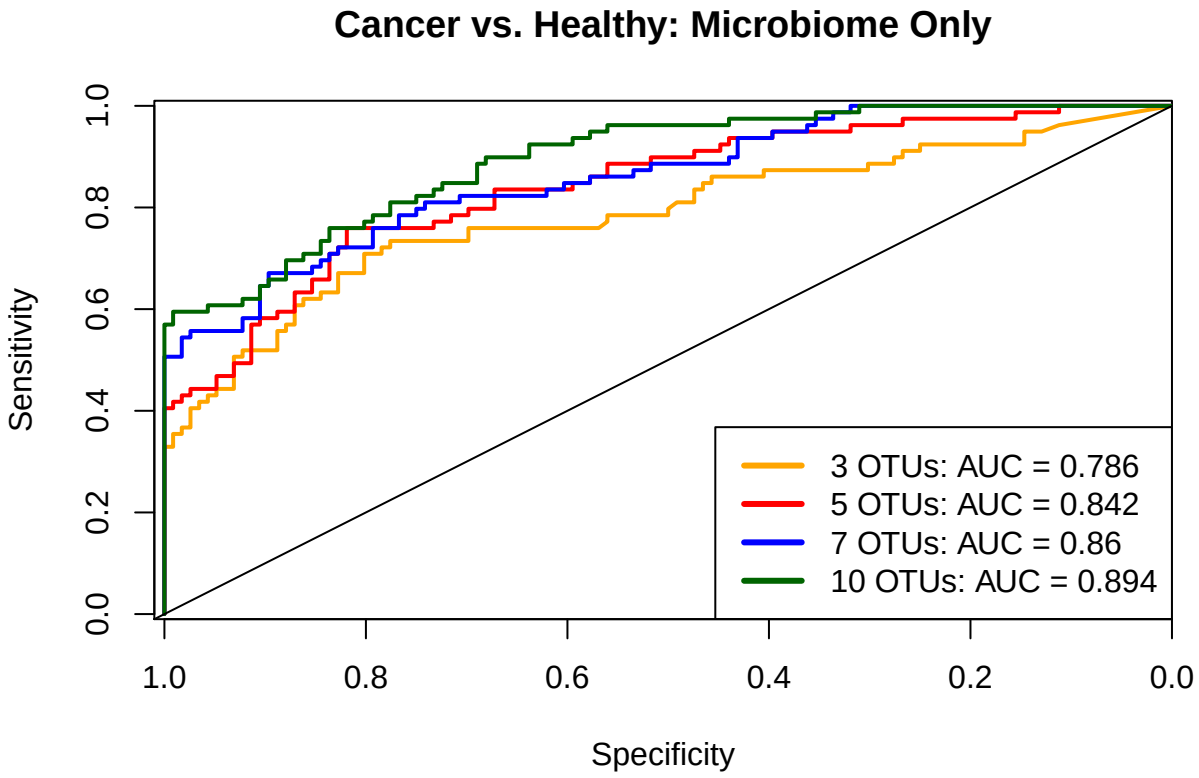
```
library(gtools)
library(AICcmodavg)
setwd('~/Desktop/glne007/')
meta <- read.delim('training.meta.txt', header=T, sep='\t')
shared <- read.delim('training.an.0.03.0.03.subsample.0.03.filter.shared', header=T, sep='\t')
shared <- shared[,-ncol(shared)] #removes rareOtus column from the filtered shared file
shared <- shared[meta$dx!='adenoma',]
meta <- meta[meta$dx!='adenoma',]

meta$dx <- as.character(meta$dx)
```

```
meta$dx[meta$dx=='normal'] <- 0
meta$dx[meta$dx=='cancer'] <- 1
meta$dx <- factor(meta$dx)
mydata <- data.frame(cbind(meta,shared[4:ncol(shared)]))
otus <- colnames(shared[4:ncol(shared)])
goodOTUs <- read.table('3otu.cancer.out') #Reads in a list of all 3-OTU models sorted by AIC
goodOTUs <- goodOTUs[1:10,1:3] # For the sake of time i'm only taking the top 10 in this example

combos <- c()
for(x in otus){ # adds each of the 309 OTUs to the 100 top models
  com <- cbind(goodOTUs, x)
  combos <- rbind(combos, com)
  }

getDuplicates <- function(x){ # removes models with duplicate OTUs
  return(length(x) - length(unique(unlist(x))) )
}
dups <- apply(combos, MARGIN=1, FUN=getDuplicates)
combos <- combos[!dups,]

fun <- function(r){
  return(c(r[1],r[2],r[3],r[4],AICc(suppressWarnings(glm(dx ~ mydata[,r[1]] + mydata[,r[2]] + mydata[,r
}
results <- apply(X=combos, MARGIN=1, FUN=fun)
results <- t(results)
results <- results[order(results[,5], decreasing=F),]
colnames(results) <- c('OTU1','OTU2','OTU3','OTU4','AIC')
head(results, n=10)
```

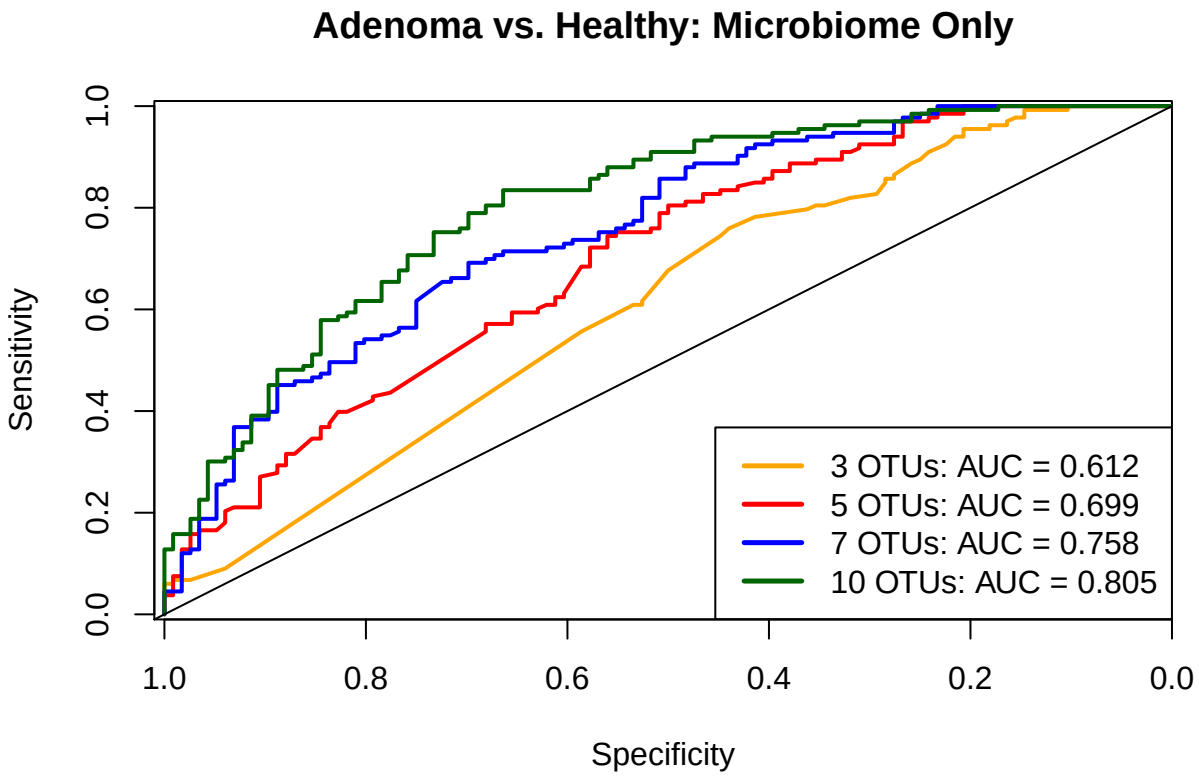
```
##      OTU1        OTU2        OTU3        OTU4        AIC
## 13   "Otu000035" "Otu000048" "Otu002051" "Otu000002" "196.118498930719"
## 1120 "Otu000002" "Otu000035" "Otu000048" "Otu002051" "196.118498930719"
## 553  "Otu000035" "Otu000048" "Otu002051" "Otu000063" "196.769064708944"
## 1114 "Otu000035" "Otu000048" "Otu000063" "Otu002051" "196.769064708944"
## 134  "Otu000035" "Otu000048" "Otu000063" "Otu000014" "197.14647774977"
## 551  "Otu000014" "Otu000035" "Otu000048" "Otu000063" "197.14647774977"
## 133  "Otu000035" "Otu000048" "Otu002051" "Otu000014" "198.783998823214"
## 1111 "Otu000014" "Otu000035" "Otu000048" "Otu002051" "198.783998823214"
## 83   "Otu000035" "Otu000048" "Otu002051" "Otu000009" "198.892468636466"
## 2174 "Otu000035" "Otu000048" "Otu000063" "Otu002974" "198.954604800836"
```

Using the best models with 3-10 OTUs, I regenerated the ROC curves for models with only microbiome data.

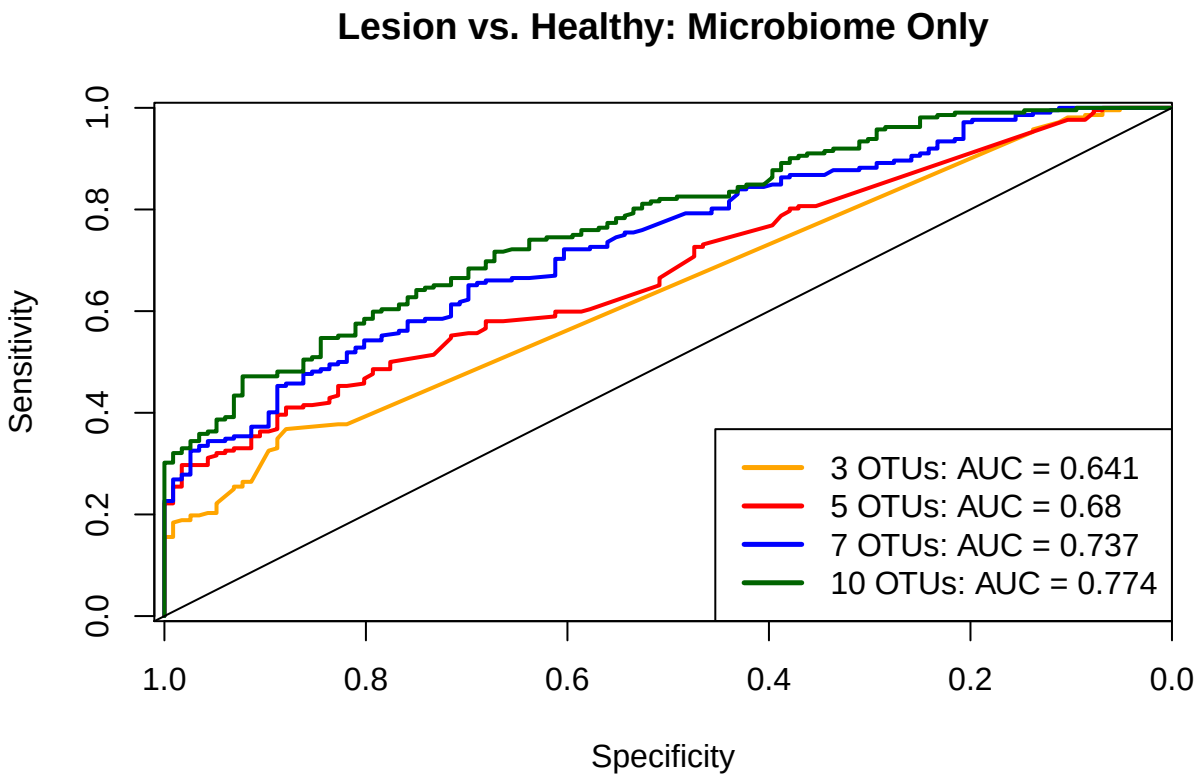
Cancer vs Healthy: 3,5,7,10 OTUs



Adenoma vs Healthy: 3,5,7,10 OTUs



Lesion vs Healthy: 3,5,7,10 OTUs




```
plot(1:10,1:10)
```