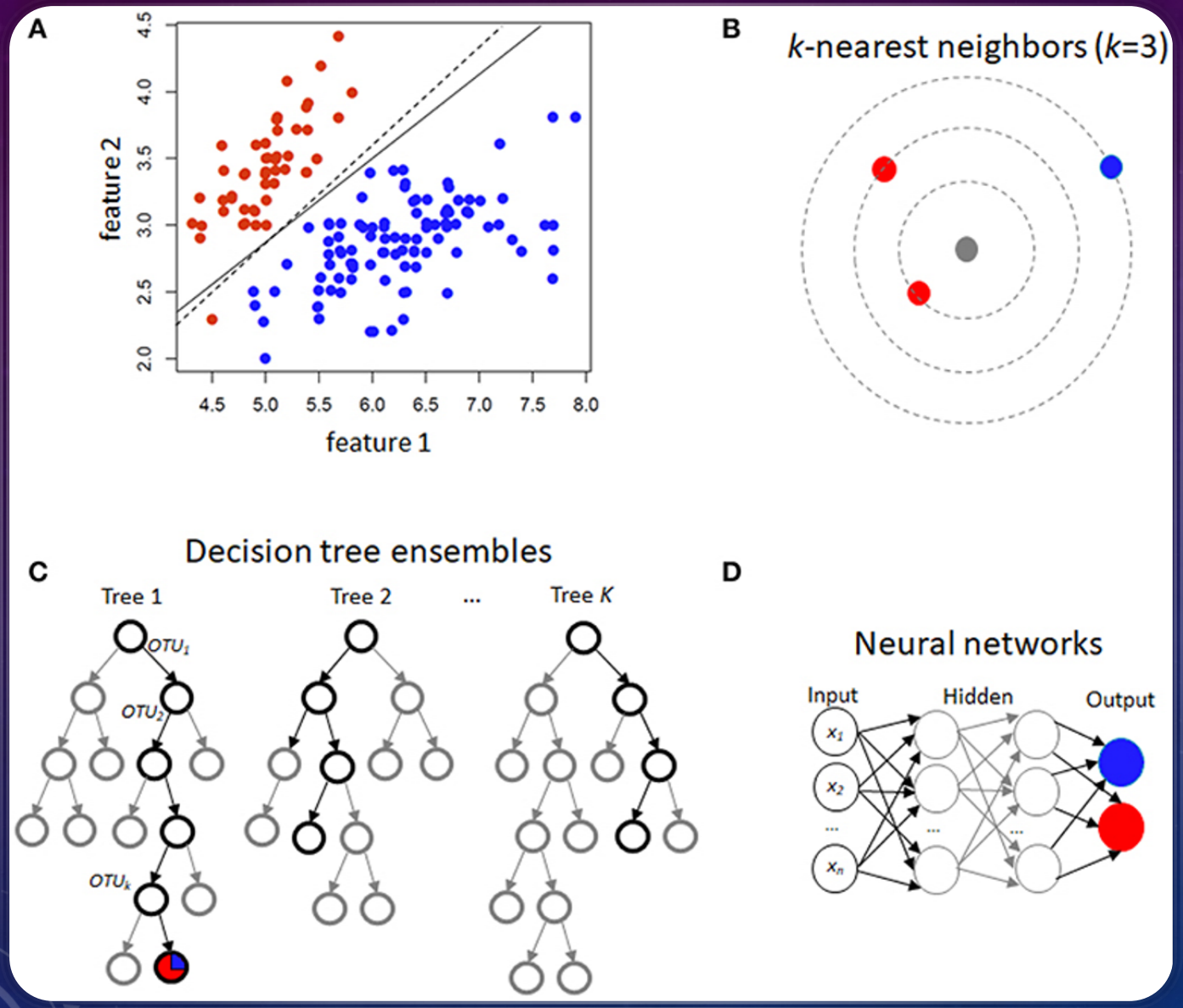
DISEASE PREDICTIONS FROM OPERATIONAL TAXONOMIC UNITS *A MACHINE LEARNING APPROACH*

TEAM 4

AGAZ WANI, ALEX DEAN, CHANG LI, NATHAN VAN BIDDER, PETER RADULOVIC, YIBO DONG, GREG HERBERT, KRISTEN DOMINGUEZ, NGOC TRAN, XIAOMING LIU

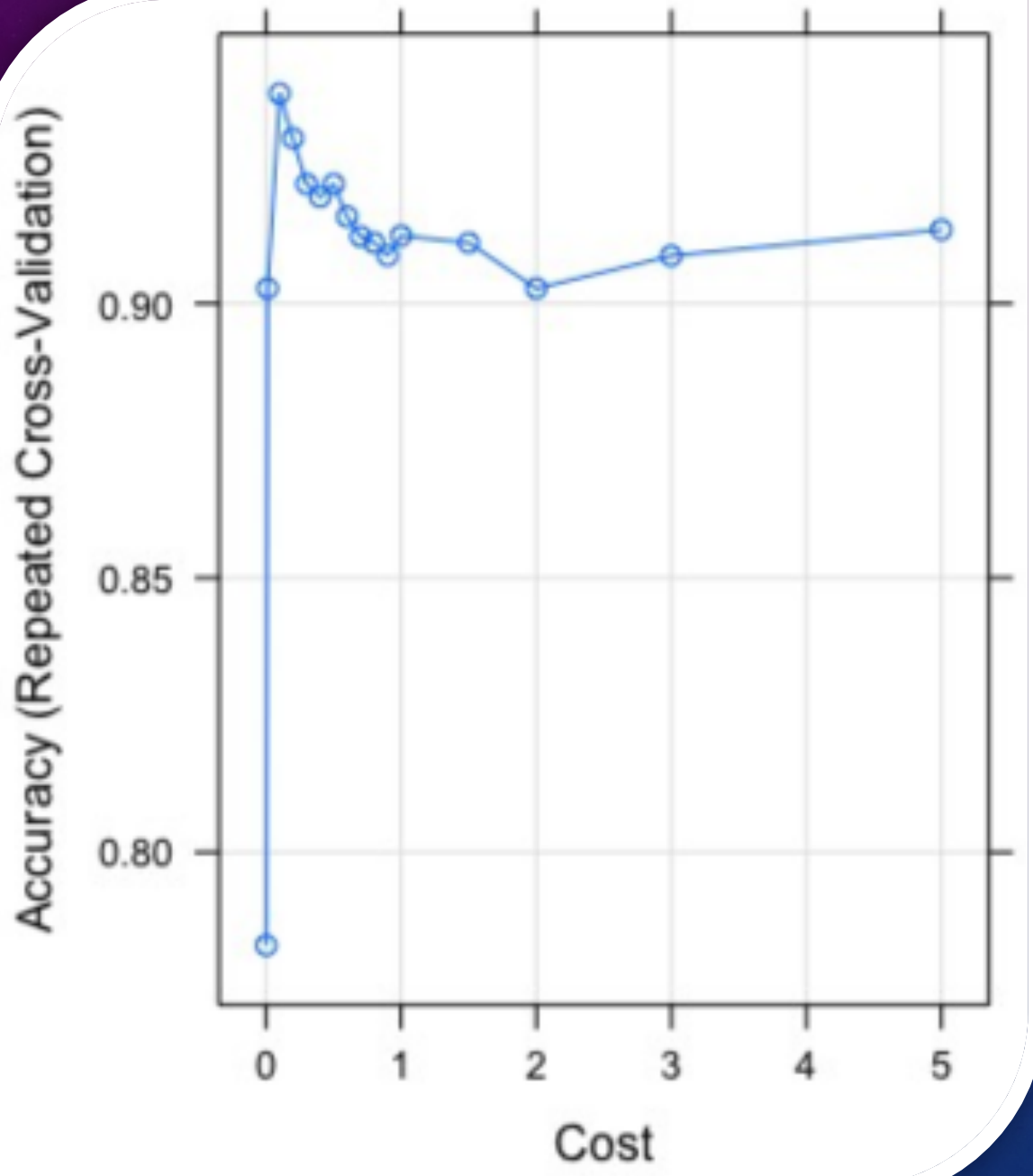


- Figure 1: Schematic illustration of several machine learning prediction methods using case/control (red/blue) status. For two features,
- **(A)** illustrates linear discrimination methods. The solid line shows the linear discriminant line corresponding to equally probable outcomes, while the dashed line shows the midpoint of the maximum-margin support vector machine.
- **(B)** For *k*-nearest neighbors, the gray point is predicted using an average of the neighbors (red, in this instance).
- **(C)** Decision tree ensembles include random forests, which average over bootstrapped trees, and boosted trees, where successive residuals are used for fitting. Trees may not extend to the level of individual observations, and modal or mean values in the terminal nodes are used for prediction.
- **(D)** A neural network with few hidden layers.

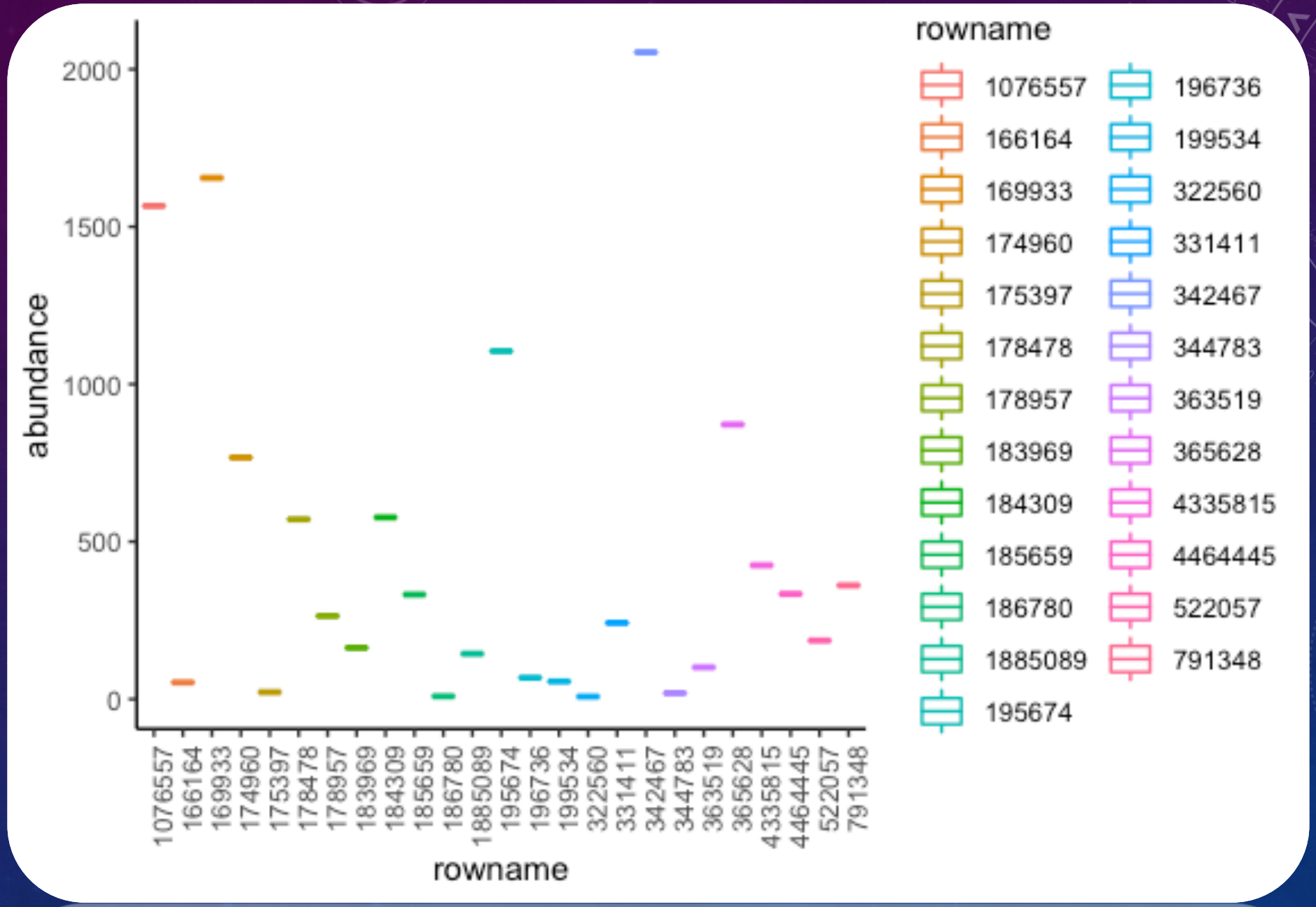
The proper citation: Zhou, Y.H. and Gallins, P., 2019. A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in Genetics*, *10*, p.579.

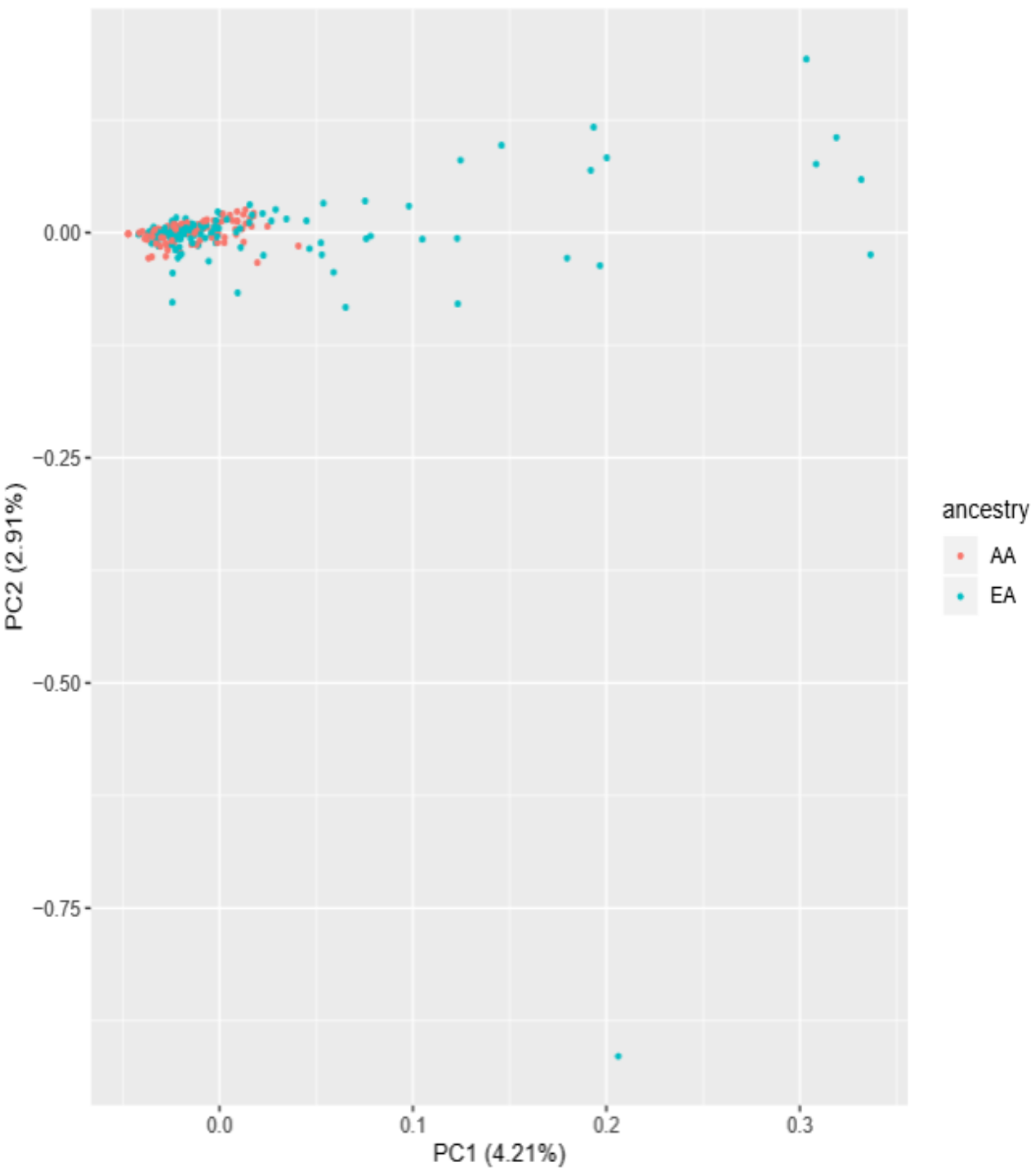
INITIAL MODEL DEVELOPMENT

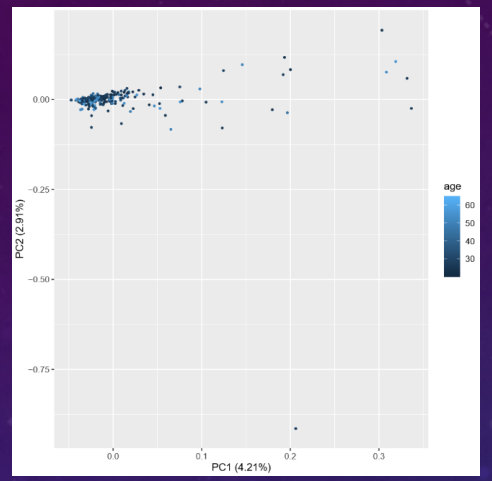
- Add random small number of 0 values in OTU
- Combined features of OTU and age, family, ancestry, zygosity, twin type (No. of features = 5,468)
- Select minimum set of variables, conditioning on which the other variables will be irrelevant (No. of selected features = 26).
- 10-fold cross validation with linear Support Vector Machine to train the model on obese status (Overweightvs. Lean)

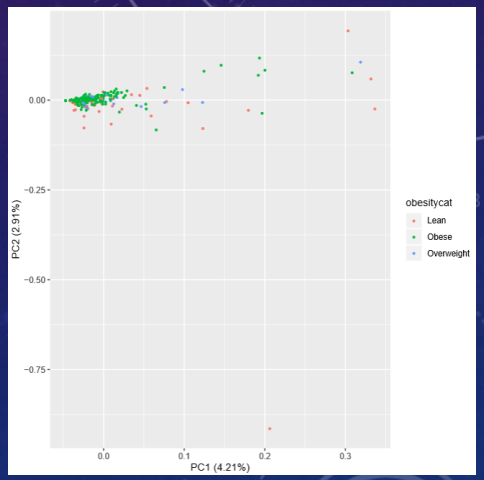


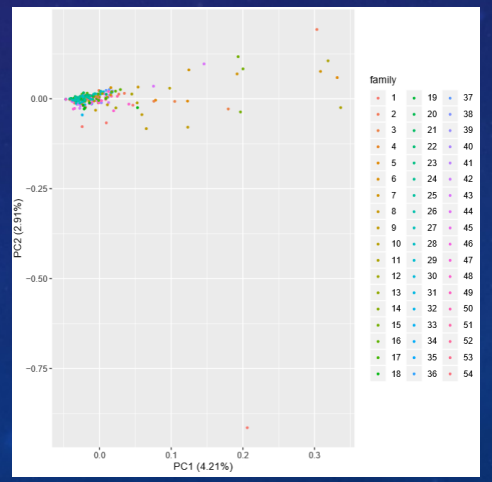

```
Support Vector Machines with Linear Kernel

281 samples
 26 predictor
  2 classes: 'Lean', 'Overweight'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 253, 253, 253, 252, 253, 253, ...
Resampling results across tuning parameters:

  C      Accuracy   Kappa
  0.001  0.7830049  0.0000000
  0.010  0.9027915  0.6579803
  0.100  0.9383005  0.8033662
  0.200  0.9301314  0.7795361
  0.300  0.9218391  0.7538193
  0.400  0.9194581  0.7493853
  0.500  0.9218391  0.7578448
  0.600  0.9158867  0.7433447
  0.700  0.9122742  0.7340943
  0.800  0.9111248  0.7301640
  0.900  0.9087438  0.7229806
  1.000  0.9123974  0.7381868
  1.500  0.9110837  0.7345620
  2.000  0.9027915  0.7108905
  3.000  0.9087028  0.7263280
  5.000  0.9135057  0.7457435
```


abundance
2000
1500
1000
500
0
1076557
166164
169933
174960
175397
178478
178957
183969
184309
185659
186780
1885089
195674
196736
199534
322560
331411
342467
344783
363519
365628
4335815
4464445
522057
791348
rowname


0.00
−0.25
−0.50
−0.75
PC2 (2.91%)
0.0
0.1
0.2
0.3
PC1 (4.21%)
ancestry
AA
EA


0.00
−0.25
−0.50
−0.75
PC2 (2.91%)
0.0
0.1
0.2
0.3
PC1 (4.21%)
age
60
50
40
30


0.00
−0.25
−0.50
−0.75
PC2 (2.91%)
0.0
0.1
0.2
0.3
PC1 (4.21%)
obesitycat
Lean
Obese
Overweight


0.00
−0.25
−0.50
−0.75
PC2 (2.91%)
0.0
0.1
0.2
0.3
PC1 (4.21%)
family