A8 - Metabarcoding

Roshael Chellappah (20103016)

16/03/2022

GitHub Repository: https://github.com/RoshaelC/metabarcoding (https://github.com/RoshaelC/metabarcoding)

The purpose of this assignment is to examine how communities of plants differ across sample locations. Using information gathered by a former Queen's MSc student, we are hoping to answer the following biological questions:

1. What effect (if any) does garlic mustard have on the plant community?
2. What has a stronger effect on plant communities: the presence/absence of garlic mustard (in/out) or sampling population?

Load Data

```
library(ggplot2)
library(ape)
library(ggtree)
```

```
## ggtree v3.2.1  For help: https://yulab-smu.top/treedata-book/
##
## If you use ggtree in published research, please cite the most appropriate paper(s):
##
## 1. Guangchuang Yu. Using ggtree to visualize data on tree-like structures. Current Protocols in Bioinformatics. 2020, 69:e96. doi:10.1002/cpbi.96
## 2. Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for mapping and visualizing associated data on phylogeny using ggtree. Molecular Biology and Evolution. 2018, 35(12):3041-3043. doi:10.1093/molbev/msy194
## 3. Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017, 8(1):28-36. doi:10.1111/2041-210X.12628
```

```
##
## Attaching package: 'ggtree'
```

```
## The following object is masked from 'package:ape':
##
##     rotate
```

```
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 4.1.3
```

```
## Loading required package: permute
```

```
## Warning: package 'permute' was built under R version 4.1.3
```

```
## Loading required package: lattice
```

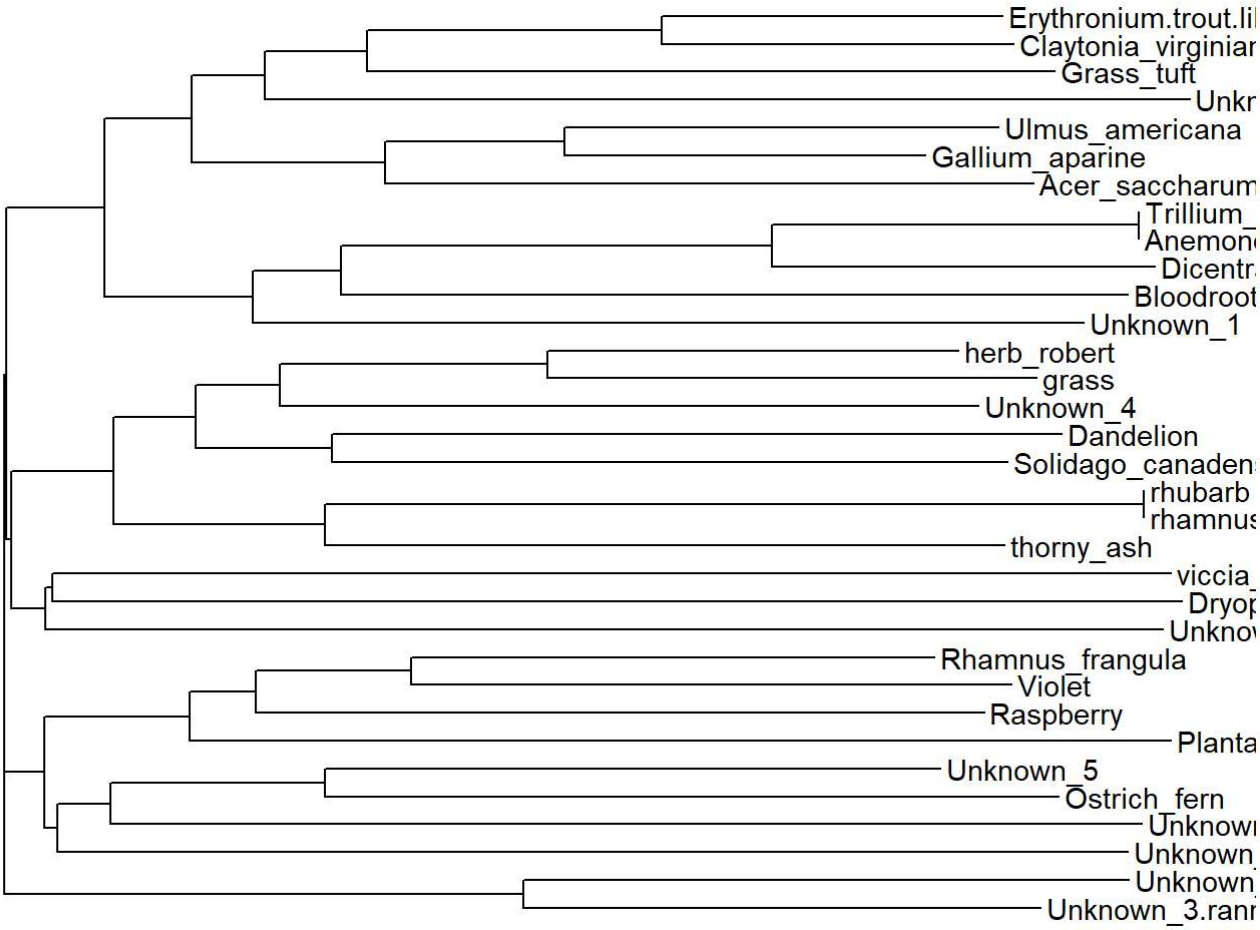
```
## This is vegan 2.5-7
```

```
Data <- read.csv("./FloristicSurvey.csv", fileEncoding="UTF-8-BOM", header = T)
```

Question 1

To find an answer to the first biological question, I will compare the relationship between the presence/absence of garlic mustard to the composition of the various plant species within the different communities.

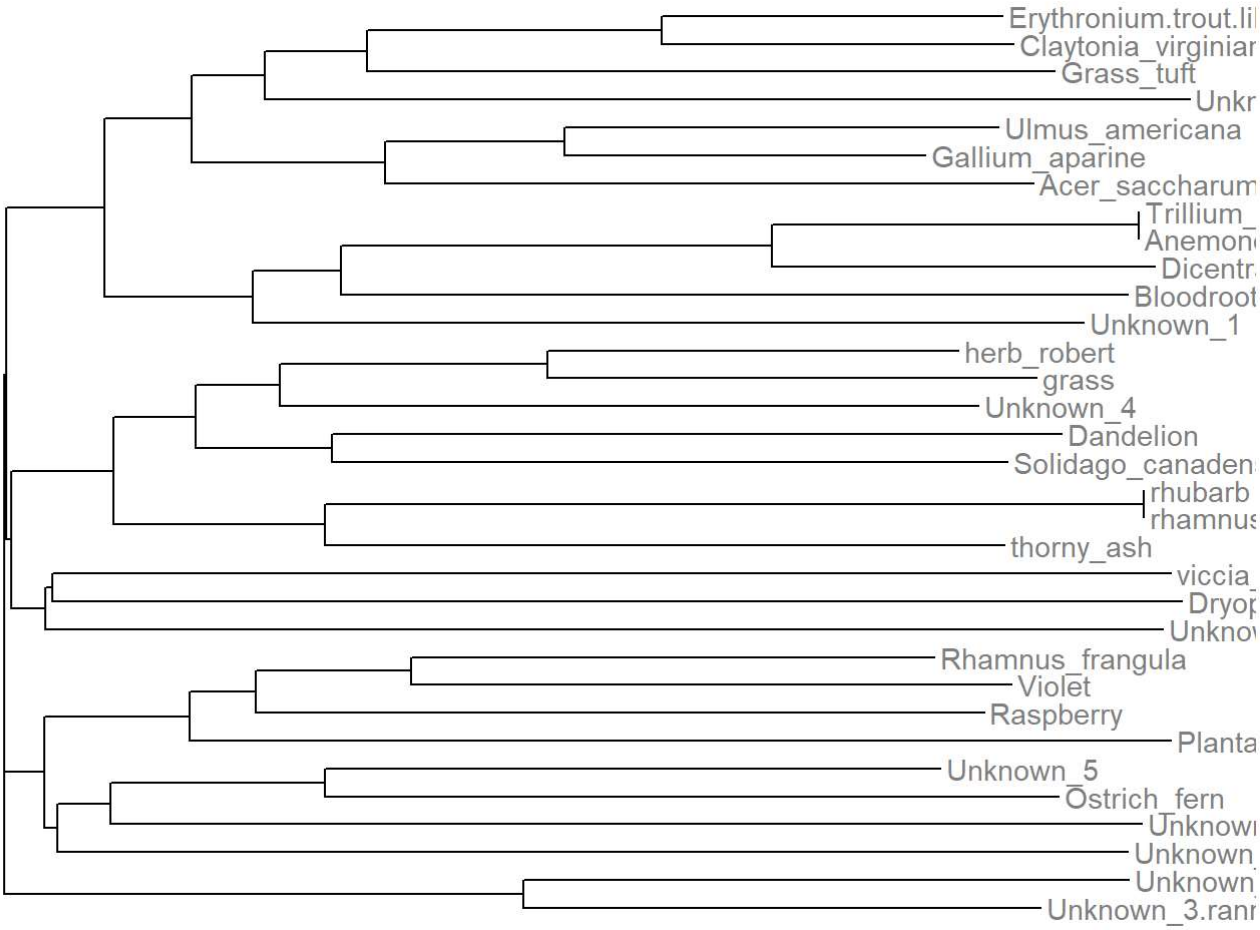
```
Data2 <- Data[,c(1, 11:43)] # isolate the quadrate and plant species columns into new data frame

# make quadrate column into row names
Data3 <- Data2[,-1]
rownames(Data3) <- Data2[,1]

# transpose the data frame
MyData <- as.data.frame(t(Data3))

# Distance Matrix
MyData_bin <- MyData

MyData_bin[MyData_bin > 0] <- 1
MyData_dist <- dist(MyData_bin, method = 'binary')

MyDataDistMat <- as.matrix(MyData_dist)

MyData_tree <- nj(MyDataDistMat)
ggtree(MyData_tree, layout = "rectangular") + geom_tiplab()
```



Although I've been able to visualize the distance matrix, I am unable to show which plants are clustered with which sample locations. This is likely because there is an overlap between the plants found within each location, though I am unsure how to troubleshoot this problem. Ultimately, I would want to show that garlic mustard plants impact the composition of plant species within locations that it inhabits, but I am facing many challenges in understanding and implementing code that could help me demonstrate this.

I have tried creating a secondary dataset, similar to the SamplesInfo.csv file that was used in tutorial, to potentially help in clustering the various nodes from the distance visualization together, and this is shown below. However, it seems that this does not work either.

```
# Binary Distance Matrix
SampleLocations <- read.csv("./A8_Roshael_Chellappah - Locations.csv", fileEncoding="UTF-8-BOM")
# secondary dataset grouping the various quadrates within locations (i.e., all 7o and 7i quadran
ts are from location/population 7)

#NJ
MyData_tree <- nj(MyDataDistMat)
ggtree(MyData_tree, layout = "rectangular") %<+% SampleLocations +
  geom_tiplab(aes(colour = Location)) +
  theme(legend.position = "right")
```




```
MyData_dist <- vegdist(MyData, method = "bray", binary= F)
MyDtaa_tree <- nj(MyData_dist)


#NMDS
set.seed(13)
NMDSdat <- metaMDS(MyData_dist, k = 2) #k=2 dimensions
```

```
## Run 0 stress 0.1191109
## Run 1 stress 0.1246194
## Run 2 stress 0.1329265
## Run 3 stress 0.123143
## Run 4 stress 0.1309924
## Run 5 stress 0.1175122
## ... New best solution
## ... Procrustes: rmse 0.09842243  max resid 0.2341933
## Run 6 stress 0.1236185
## Run 7 stress 0.1263787
## Run 8 stress 0.1219151
## Run 9 stress 0.1252229
## Run 10 stress 0.1209713
## Run 11 stress 0.1269069
## Run 12 stress 0.1162273
## ... New best solution
## ... Procrustes: rmse 0.1111747  max resid 0.2639842
## Run 13 stress 0.1235557
## Run 14 stress 0.1350614
## Run 15 stress 0.1252238
## Run 16 stress 0.1171416
## Run 17 stress 0.1308718
## Run 18 stress 0.1203351
## Run 19 stress 0.12168
## Run 20 stress 0.1223508
## *** No convergence -- monoMDS stopping criteria:
##     10: no. of iterations >= maxit
##     10: stress ratio > sratmax
```

```
NMDSdat <- metaMDS(MyData_dist, k = 2, trymax = 100)
```

```
## Run 0 stress 0.1191109
## Run 1 stress 0.1227708
## Run 2 stress 0.1282102
## Run 3 stress 0.1282339
## Run 4 stress 0.120857
## Run 5 stress 0.12706
## Run 6 stress 0.1327504
## Run 7 stress 0.1241437
## Run 8 stress 0.125709
## Run 9 stress 0.1309525
## Run 10 stress 0.1254027
## Run 11 stress 0.1288614
## Run 12 stress 0.1206262
## Run 13 stress 0.1194447
## ... Procrustes: rmse 0.1042143  max resid 0.2785791
## Run 14 stress 0.1225729
## Run 15 stress 0.1315795
## Run 16 stress 0.1225564
## Run 17 stress 0.1267205
## Run 18 stress 0.1210807
## Run 19 stress 0.1214761
## Run 20 stress 0.118389
## ... New best solution
## ... Procrustes: rmse 0.1186903  max resid 0.2825207
## Run 21 stress 0.1255777
## Run 22 stress 0.1192322
## Run 23 stress 0.1175377
## ... New best solution
## ... Procrustes: rmse 0.1095475  max resid 0.4530673
## Run 24 stress 0.1225513
## Run 25 stress 0.1389358
## Run 26 stress 0.1214224
## Run 27 stress 0.1294765
## Run 28 stress 0.1193614
## Run 29 stress 0.1306239
## Run 30 stress 0.1297883
## Run 31 stress 0.1231773
## Run 32 stress 0.1297406
## Run 33 stress 0.1265199
## Run 34 stress 0.1315788
## Run 35 stress 0.1142128
## ... New best solution
## ... Procrustes: rmse 0.06439251  max resid 0.3030038
## Run 36 stress 0.1236557
## Run 37 stress 0.1297414
## Run 38 stress 0.1287827
## Run 39 stress 0.1231361
## Run 40 stress 0.1218416
## Run 41 stress 0.1313198
## Run 42 stress 0.1184714
## Run 43 stress 0.1253023
## Run 44 stress 0.1197222
```

```
## Run 45 stress 0.1321904
## Run 46 stress 0.1277601
## Run 47 stress 0.1263267
## Run 48 stress 0.1252724
## Run 49 stress 0.1172018
## Run 50 stress 0.1273952
## Run 51 stress 0.1254608
## Run 52 stress 0.1239325
## Run 53 stress 0.1249269
## Run 54 stress 0.1209384
## Run 55 stress 0.1271508
## Run 56 stress 0.1206754
## Run 57 stress 0.1336234
## Run 58 stress 0.1269188
## Run 59 stress 0.1190878
## Run 60 stress 0.1234333
## Run 61 stress 0.1265477
## Run 62 stress 0.116701
## Run 63 stress 0.1303923
## Run 64 stress 0.1324483
## Run 65 stress 0.1205039
## Run 66 stress 0.1240213
## Run 67 stress 0.1266569
## Run 68 stress 0.1279217
## Run 69 stress 0.1239419
## Run 70 stress 0.1178647
## Run 71 stress 0.119668
## Run 72 stress 0.1318111
## Run 73 stress 0.1176273
## Run 74 stress 0.1311976
## Run 75 stress 0.118239
## Run 76 stress 0.1203097
## Run 77 stress 0.1252531
## Run 78 stress 0.1209721
## Run 79 stress 0.1217759
## Run 80 stress 0.1308941
## Run 81 stress 0.1281409
## Run 82 stress 0.1273385
## Run 83 stress 0.1198971
## Run 84 stress 0.1247223
## Run 85 stress 0.1293842
## Run 86 stress 0.1252318
## Run 87 stress 0.1241195
## Run 88 stress 0.124362
## Run 89 stress 0.1271773
## Run 90 stress 0.1264727
## Run 91 stress 0.1274226
## Run 92 stress 0.128066
## Run 93 stress 0.1239196
## Run 94 stress 0.1235229
## Run 95 stress 0.1321858
## Run 96 stress 0.1261425
```

```
## Run 97 stress 0.1228538
## Run 98 stress 0.1202425
## Run 99 stress 0.1275647
## Run 100 stress 0.1242134
## *** No convergence -- monoMDS stopping criteria:
##     57: no. of iterations >= maxit
##     43: stress ratio > sratmax
```

```
## Create data for plotting
PDAT <- data.frame(NMDS1 = NMDSdat$points[,1],
                   NMDS2 = NMDSdat$points[,2],
                   SampleID = row.names(MyData))



# PDat <- merge(PDAT, SampleLocations, by = "Quadrate", all.x = T, all.y = F)
#### maybe need to create a separate data file with the various plant species instead of the sam
ple locations???


# qplot(x = NMDS1, y = NMDS2, colour = Locations, alpha = I(0.6), data = PDat)
#### code does not work with sample locations document
```

Question 2

To answer this second question, I need to understand the information found in the analysis for question 1 (understanding the correlation between the presence/absence of garlic mustard on the plant community composition) as well as the correlation between the sampling population and the plant community.

After trying to work through this assignment, I continue to get stuck with visualizing my data and showing whether there is a correlation between the presence/absence of garlic mustard plants and the composition of plant species within the sample location. I believe that I understand the concept of what information is needed to answer the research questions but I am facing a lot of trouble in getting the code to work. I will need to take more time to fully understand the code and how to use it in settings outside of the tutorials. Unfortunately, I feel that this is all I am able to complete for now, and would prefer to submit this assignment as is than nothing at all.