R Mini-Project Report

Mike Nsubuga

6/4/2021

MSB7102 Mini-project, Semester 1, 2021

Background

This is an assessed exercise that will contribute to coursework for this module. It is aimed at providing an experience with all that has been covered throughout the semester. The tasks are based on two Bioconductor packages; phyloseq and DEseq2. Endeavor to look at the documentation and the links indicated below may be useful. The data files are located in the shared google drive folder.

Data sources and description

The data used in this exercise is derived and was generated by Kolistic et al in their study "The dynamics of the human infant gut microbiome in development and progression toward Type 1 Diabetes". See full publication at https://doi.org/10.1016/j.chom.2015.01.001. Briefly, this was a prospective analysis of developing gut microbiome in infants en route to type 1 diabetes. Infants from Finland and Estonia were recruited at birth based on predisposition to autoimmunity determined by human leukocyte antigen (HLA) genotyping. The cohort consists of 33 infants, 11 of which seroconverted to serum autoantibody positivity and of those, four developed T1D within the three-year time-frame of this study.

Tasks

1. Import the data described above into R, provide descriptive summaries of the subject data (using appropriate graphics and statistical summary measures) given in the diabimmune_16s_t1d_metadata.csv file. In addition, use appropriate test(s) to check for association/independency between disease status and other variables (delivery mode, gender and age). Note that age is given in days.
2. Using phyloseq, create a phyloseq object. This will comprise the OTU abundance, taxonomy (provided in the .txt file) and sample data (provided in the .csv file).
3. Generate Alpha diversity plots and ordination plots. Examine any observed patterns by delivery mode, gender and disease status.
4. Perform differential abundance using DEseq2 ### Useful links

- Importing data: https://joey711.github.io/phyloseq/import-data.html
- Ordination: https://joey711.github.io/phyloseq/plot_ordination-examples.html
- Alpha diversity: https://joey711.github.io/phyloseq/plot_richness-examples.html
- Differential abundance: http://joey711.github.io/phyloseq-extensions/DESeq2.html

Question 1

1.1.1. Descriptive summaries of the subject data using appropriate graphics and statistical summary measures. Using the summary function to get a summary of descriptive statistics which include mean, median, 25th and 75th quartiles, min, max

```
data <- read.csv("diabimmune_16s_t1d_metadata.csv");
summary(data)
```

```
##   Sample_ID          Subject_ID        Case_Control          Gender
##  Length:777         Length:777         Length:777         Length:777
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Delivery_Route     Age_at_Collection
##  Length:777         Min.   :   6.0
##  Class :character   1st Qu.: 229.0
##  Mode  :character   Median : 452.0
##                     Mean   : 482.9
##                     3rd Qu.: 702.0
##                     Max.   :1233.0
```

```
dim(data)
```

```
## [1] 777   6
```

```
class(data)
```

```
## [1] "data.frame"
```

1.1.2. Graphical summary representation of the data. Bar charts representing the number of occurences in each category

```
library(ggplot2)
table(data$Case_Control)
```

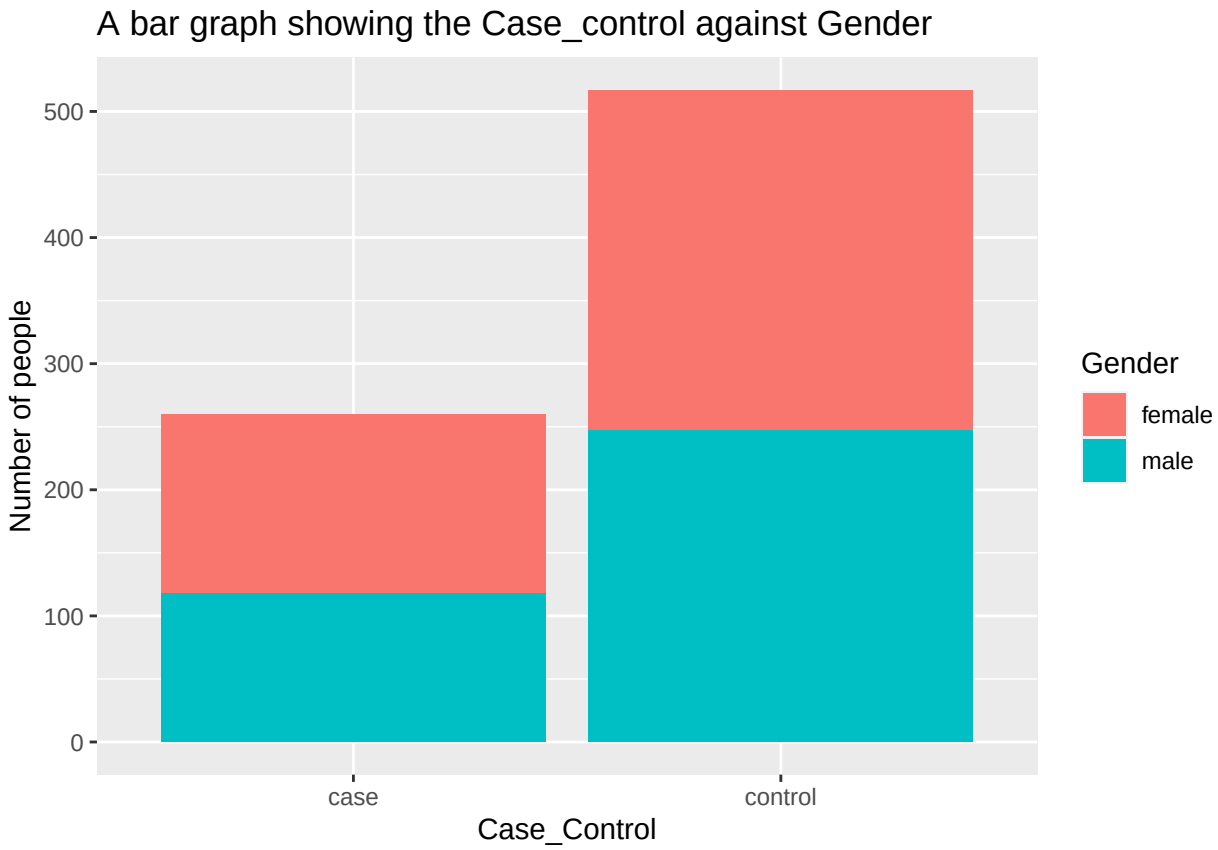
1.1.3 Number of Cases and Controls against the Genders.

```
##
##    case control
##     260     517
```

```
table(data$Gender)
```

```
##
## female   male
##    412    365
```

```
qplot(data$Case_Control, fill = data$Gender) + geom_bar() + labs(title = "A bar graph showing the Case_
```




```
library(ggplot2)
table(data$Case_Control)
```

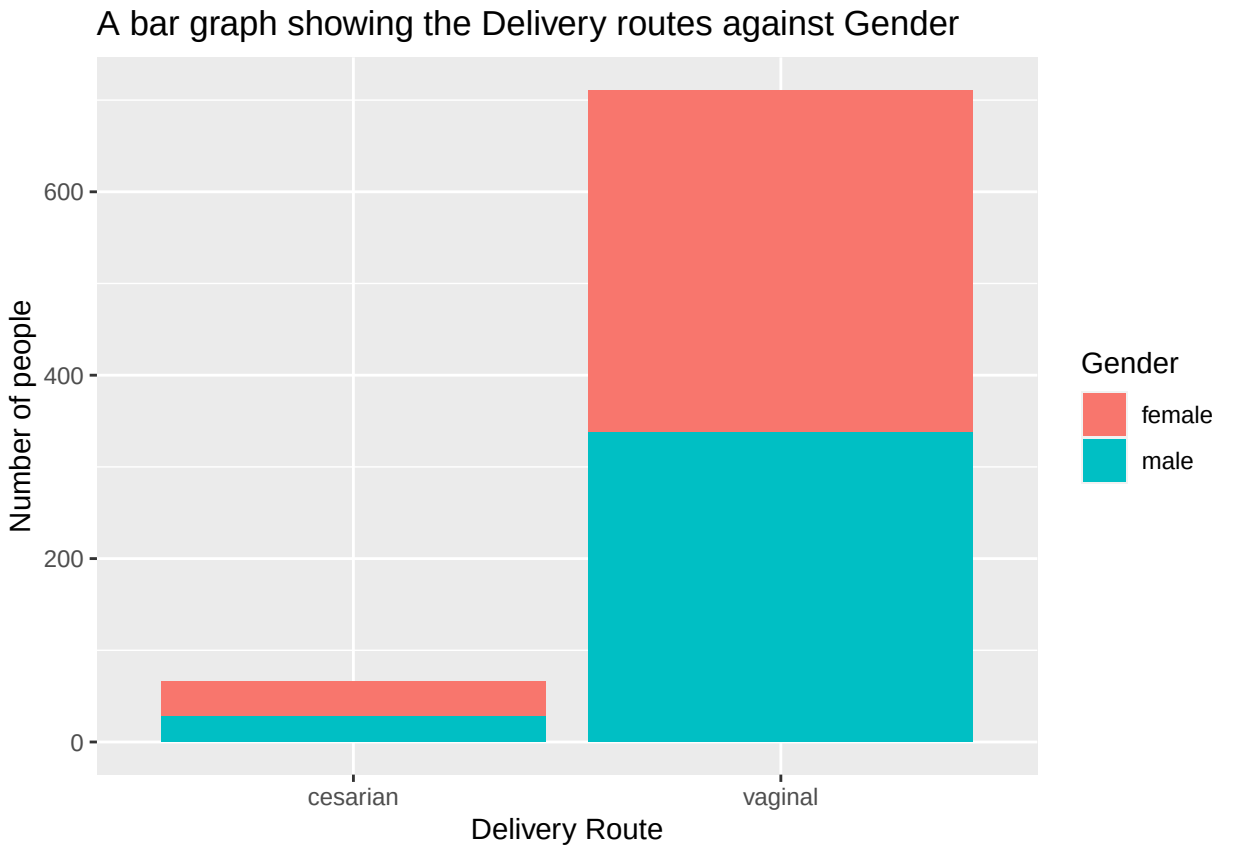
1.1.4 The Delivery Route against the Gender

```
##
##    case control
##     260     517
```

```
table(data$Gender)
```

```
##
## female   male
##    412    365
```

```
qplot(data$Delivery_Route, fill = data$Gender) + geom_bar() + labs(title = "A bar graph showing the Deli
```



1.2.1. Appropriate tests to check for association/independency between disease status and other variables. Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

Chi-square test examines whether rows and columns of a contigency table are statistically significatly associated

- Null hypothesis(H0): the row and the column variables of the contigency table are independent
- Alternate hypothesis(H1): row and column variables are dependent

```
tbDelivery <- table(data$Case_Control, data$Delivery_Route)  #Generate a contigency table of Disease st
chisq.test(tbDelivery)
```

Test the hypothesis whether the disease status is independent of their Delivery Mode at .05 significance level

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbDelivery
## X-squared = 34.649, df = 1, p-value = 3.949e-09
```

As the p-value 3.949e-09 is less than the .05 significance level, we reject the null hypothesis that the Disease Control status is independent of the Delivery Mode. This means that the two variables are somewhat related

```
tbGender <- table(data$Case_Control, data$Gender)  #Generate a contigency table of Disease status and G
chisq.test(tbGender)
```

Test the hypothesis whether the disease status is independent of their Gender at .05 significance level.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbGender
## X-squared = 0.30687, df = 1, p-value = 0.5796
```

As the p-value 0.5796 is greater than the .05 significance level, we do not reject the null hypothesis that the Disease Control status is independent of the Gender. It indicates strong evidence for the null hypothesis and reject the alternative hypothesis. This means that there's no relationship between the two variables being studied

```
tbAge <- table(data$Case_Control, data$Age_at_Collection)  #Generate a contigency table of Disease stat
chisq.test(tbAge)
```

Test the hypothesis whether the disease status is independent of their Age_at_Collection at .05 significance level.

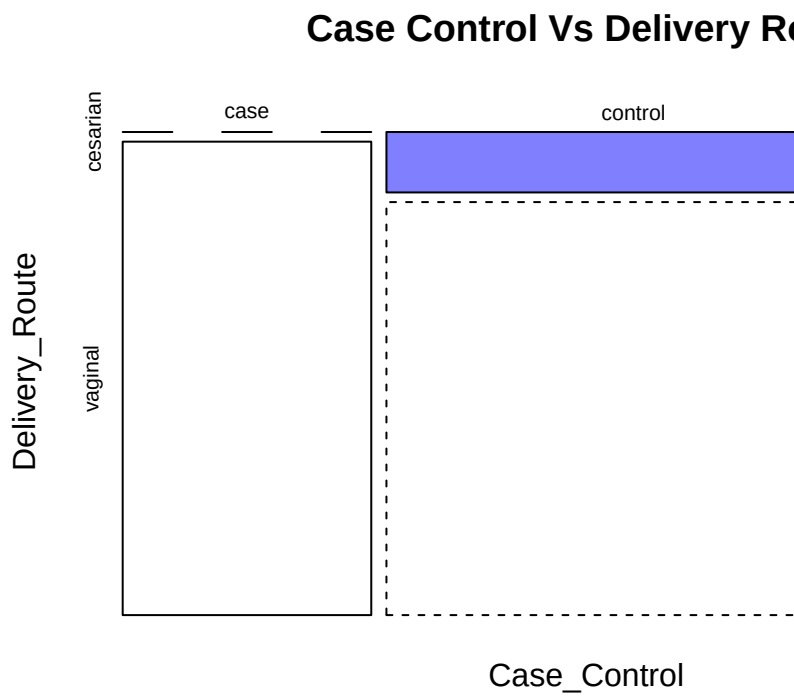
```
## Warning in chisq.test(tbAge): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbAge
## X-squared = 532.07, df = 542, p-value = 0.6115
```

As the p-value 0.6115 is greater than the .05 significance level, we do not reject the null hypothesis that the Disease Control status is independent of the Gender. It indicates strong evidence for the null hypothesis and reject the alternative hypothesis. This means that there's no relationship between the two variables being studied

1.2.2. Mosaic Plots. To further visualise the tests of association/independecy, they can be plotted on a mosaic plot, which can give us insights on how the two variables are related. Units are in standard deviations, so a residual greater than 2 or less than -2 represents a significant departure from independence. A mosaic plot maps standardised residuals to cells, positive sign is for blue and negative sign is for red ones. Blue means there are more observations in the cell than would be expected under the null model(independence). Red means there are fewer observations than would have been expected

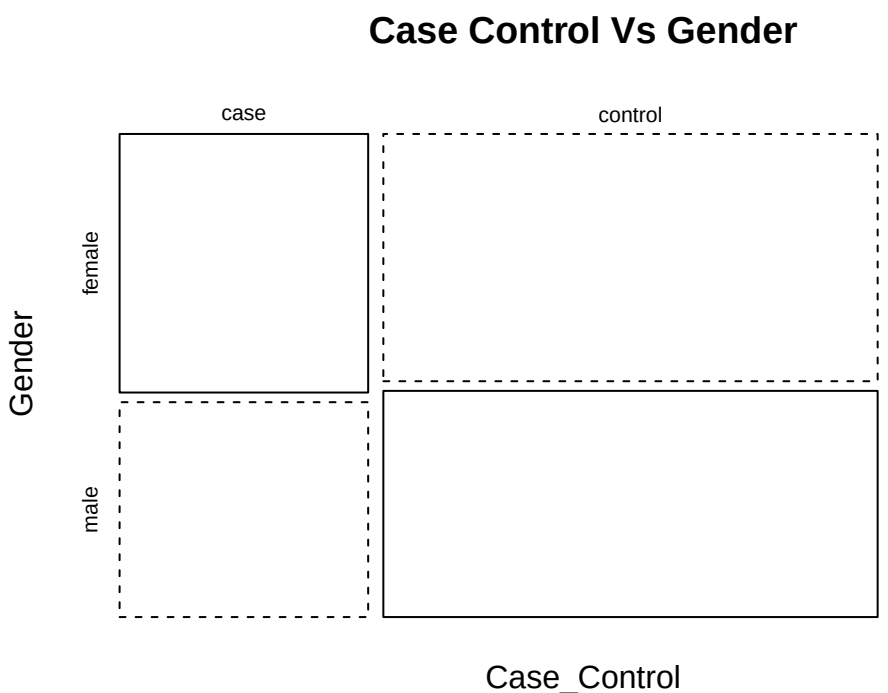
```
library(graphics)
mosaicplot(~ Case_Control + Delivery_Route, data = data, main = "Case Control Vs Delivery Route", shade
```



A mosaic plot for Case Control vs Delivery Route

From the plot, you will notice that the "control" cases are related to the method of delivery. Infact, most if not all of the contolled cases were delivered by cesarian method. Its slso important to note that there was a p-value less than .05 significance level when a chi-square test was conducted. In other words, from the both tests, it is confident to say that there will be "controls" whenever the Delivery Route is cesarian.

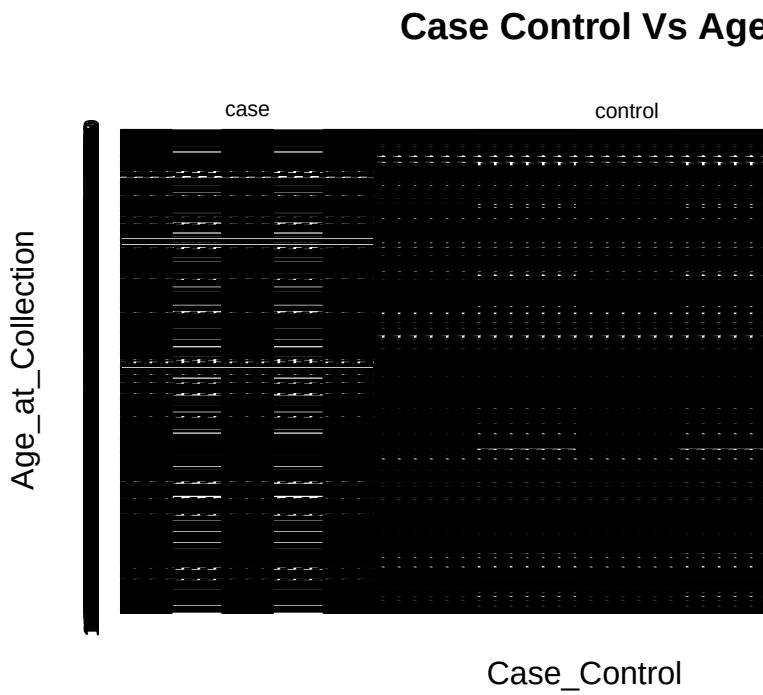
```
mosaicplot(~ Case_Control + Gender, data = data, main = "Case Control Vs Gender", shade = TRUE)
```



A mosaic plot for Case Control vs Gender

The plot indicates no significant relationship when you look at the standardised residuals

```
mosaicplot(~ Case_Control + Age_at_Collection, data = data, main = "Case Control Vs Age", shade = TRUE)
```



A mosaic plot for Case Control vs Age at Collection

The plot indicates no significant relationship when you look at the standardised residuals

Question 2

Phyloseq

Using phyloseq, create a phyloseq object. this will comprise the OTU abundance, taxonomy(provided in the .txt file) and sample data(provided in the .csv file)

```
otuTable <- read.table("otu_table")  #Importing the OTU table
head(otuTable, n = 1)
```

```
##         G37016 G36918 G37044 G37009 G37029 G37035 G36982 G36984 G37030 G37031
## 4333897     12      1     12      0      0     22      0      0      0     35
##         G36906 G37028 G37014 G37010 G36944 G36902 G37051 G36964 G36951 G37025
## 4333897      1     27     36      4      1      0      3     24      0      0
##         G36930 G36935 G37041 G36959 G36905 G36917 G36992 G36921 G37004 G36966
## 4333897      0     43     29     14      1      2      3      2     14      5
##         G36936 G37015 G37034 G36953 G36974 G37039 G36933 G37052 G36925 G37046
## 4333897      0      0      0      0      0      0      3      0      1      4
##         G37053 G36998 G37045 G37049 G37019 G36988 G37037 G37005 G36928 G36923
## 4333897     39      0      5      8      1      7      7      4      3      0
##         G37033 G37017 G37012 G36952 G37011 G36957 G37021 G36937 G37047 G37032
## 4333897     35      1     14      0     19     26     23      1      0      7
##         G36932 G37043 G37054 G36948 G36972 G36968 G36911 G36987 G37042 G36960
## 4333897      6     38      0      4      0      1      0     11     17      1
##         G37040 G37027 G37024 G37023 G37036 G37026 G37050 G36973 G36991 G36904
## 4333897      0      1      0      8      4      5     17      0      1      0
##         G36958 G36993 G37020 G36910 G36927 G37018 G37048 G36940 G37013 G37038
## 4333897     32     10      0      0      0      0      0      4      0      7
```

```
##         G36938 G37022 G35534 G35535 G35361 G35364 G35433 G35528 G35381 G35416
## 4333897      0      8     34      0      0      2     30      7      3     13
##         G35477 G35390 G35536 G35514 G35525 G35395 G35480 G35409 G35391 G35437
## 4333897      1     56      0      0      9      3     18      0      0      4
##         G35511 G35448 G35510 G35470 G35403 G35424 G35398 G35365 G35442 G35476
## 4333897      1      3      0      1      9     21      8      3      4      0
##         G35515 G35351 G35484 G35495 G35473 G35418 G35529 G35532 G35537 G35520
## 4333897      0      2      0     16      0      0      1     64      1      0
##         G35370 G35517 G35417 G35465 G35350 G35421 G35531 G35444 G35443 G35349
## 4333897     60     17     17      1     38      0      2     40     17     10
##         G35523 G35487 G35386 G35521 G35475 G35490 G35422 G35526 G35498 G35348
## 4333897     26      1      0     11      0      7      1     27     16      0
##         G35423 G35374 G35486 G35524 G35394 G35359 G35419 G35356 G35519 G35505
## 4333897     42     24      3      9      4      1      0      0      1      4
##         G35447 G35458 G35397 G35462 G35392 G35464 G35497 G35373 G35453 G35352
## 4333897      0      0      0      0      1     21     25      4      1      9
##         G35366 G35530 G35512 G35372 G35407 G35455 G35481 G35355 G35469 G35513
## 4333897      5      7      0      7      0      0     11      6      0      0
##         G35387 G35527 G35516 G35463 G35441 G35414 G35412 G35388 G35354 G35533
## 4333897      0     20     57      1      9     27      8      1     51     14
##         G35406 G35467 G35522 G35404 G35445 G36173 G36145 G36196 G36219 G36299
## 4333897      0      0      6      1      7      0      3      7     80      0
##         G36137 G36273 G36201 G36165 G36202 G36226 G36269 G36135 G36250 G36225
## 4333897      0      4      0      8      0      0     10      0      0      0
##         G36272 G36285 G36259 G36229 G36155 G36251 G36297 G36170 G36241 G36224
## 4333897      0      0      9      0      0      0      0      1      0      3
##         G36169 G36266 G36164 G36215 G36195 G36205 G36246 G36150 G36237 G36148
## 4333897      0      0      2     62      0      1      0      0      5      0
##         G36213 G36232 G36240 G36249 G36298 G36176 G36214 G36230 G36320 G36245
## 4333897     17      3      0      1      0      0      6      0     12      1
##         G36161 G36172 G36212 G36180 G36153 G36210 G36303 G36193 G36151 G36248
## 4333897      2      4     12     39      0      6      0      3      0      0
##         G36239 G36270 G36158 G36233 G36311 G36175 G36267 G36183 G36227 G36322
## 4333897      0      0     80      0      0      0      1      0      0      0
##         G36182 G36256 G36163 G36287 G36157 G36262 G36200 G36160 G36185 G36162
## 4333897      0      0      2      0      0      5      0      1      0      0
##         G36204 G36318 G36194 G36244 G36197 G36264 G36317 G36159 G36234 G36206
## 4333897      0      0      1      0      2      0     12     36      0      0
##         G36257 G36216 G36208 G36309 G36146 G36192 G36284 G36291 G36142 G36138
## 4333897      1      0      0      0     17      5      5      0     12      2
##         G36203 G36282 G36290 G36274 G36314 G36258 G36147 G36235 G36143 G36308
## 4333897      5      0      0      3      1      0      0      0      1      0
##         G36209 G36166 G36268 G36156 G36179 G36261 G36154 G36207 G36286 G36281
## 4333897      9      0      4      0      2      5      0     23      0      0
##         G36316 G36247 G36283 G36141 G36263 G36304 G36292 G36136 G36174 G36300
## 4333897      0      0      0      2      1      0      0      3      0      0
##         G36187 G36288 G36177 G36168 G36171 G36289 G36186 G36313 G36255 G36140
## 4333897      0      0      1      0      0      0      0      0      0      0
##         G36211 G36315 G36184 G36295 G36144 G36296 G36242 G36139 G36222 G36181
## 4333897      0      0      4      6      4      0      0      0      0      0
##         G36310 G36199 G36307 G36191 G36260 G36217 G36276 G36280 G36238 G36321
## 4333897      0      1      9      0     66      0      0      0      0      0
##         G36243 G36279 G36254 G36223 G36236 G36133 G36231 G36198 G36189 G36252
## 4333897      0      0      0      0      0      2      0     79     22     14
```

```
##         G36221 G36293 G36149 G36220 G36294 G36277 G36190 G36167 G36228 G36301
## 4333897      0      7      0      0     16      0     29      0      0      0
##         G36319 G36305 G36265 G36275 G36253 G36188 G36152 G36134 G36302 G36312
## 4333897     23     11      0      2      4      3      5      2      5      0
##         G36271 G36430 G36510 G36437 G36444 G36544 G36421 G36513 G36502 G36472
## 4333897      0      7      9      0      1     14      0     82      0      0
##         G36494 G36432 G36555 G36493 G36441 G36479 G36448 G36515 G36554 G36484
## 4333897      8     10      3      1     11      0      0      0     10      1
##         G36485 G36475 G36460 G36456 G36435 G36497 G36521 G36451 G36540 G36433
## 4333897      0     23      0      0      9      0      0     30      0      9
##         G36528 G36491 G36429 G36471 G36537 G36546 G36481 G36542 G36529 G36408
## 4333897      2      6      3      8      0      0     11      0      0     29
##         G36523 G36509 G36495 G36545 G36506 G36467 G36452 G36556 G36420 G36478
## 4333897      0     43      0      2      1      5     19      0     28      0
##         G36449 G36533 G36487 G36536 G36499 G36511 G36527 G36445 G36453 G36534
## 4333897     33      1      0      1      3     10      0      0     12      0
##         G36462 G36549 G36516 G36501 G36424 G36532 G36547 G36505 G36464 G36507
## 4333897      0      0      0      0     16      0      0      0      0      8
##         G36476 G36480 G36474 G36434 G36535 G36431 G36470 G36406 G36436 G36439
## 4333897      2      0      0      8      0     46      1      0      0     15
##         G36443 G36450 G36550 G36551 G36473 G36543 G36459 G36438 G36522 G36465
## 4333897      0     10     14     12      2      2      0      0      0      9
##         G36519 G36461 G36482 G36514 G36518 G36496 G36553 G36498 G36538 G36457
## 4333897      0      8      0     11      0      2      0      0      0      0
##         G36557 G36520 G36508 G36490 G36463 G36530 G36531 G36488 G36419 G36492
## 4333897     11      0      7      6      0      3      2      0     22      0
##         G36466 G36525 G36500 G36454 G36413 G36446 G36486 G36477 G36417 G36455
## 4333897      1      9      0      0     13      0      0      0      0      2
##         G36517 G36558 G36552 G36524 G36503 G36423 G36526 G36440 G36541 G36458
## 4333897      0      7     70      4      0      0      0      1     42      0
##         G36489 G36512 G36548 G36442 G36468 G36422 G36886 G36796 G36889 G36818
## 4333897      0     20      1      2      0     16      6     59    116     37
##         G36779 G36824 G36878 G36881 G36859 G36834 G36899 G36856 G36797 G36895
## 4333897     30      8     26      0      2     90      2      0      1      0
##         G36898 G36816 G36888 G36756 G36845 G36809 G36864 G36840 G36799 G36860
## 4333897      0      1      0     23      4     97      0      0     16      1
##         G36786 G36882 G36822 G36827 G36748 G36865 G36897 G36866 G36820 G36829
## 4333897     29     30    108     55      9      4      0      0      4     74
##         G36879 G36884 G36894 G36861 G36815 G36784 G36791 G36846 G36873 G36766
## 4333897     84      1      3      1     23      3      8      3      4     51
##         G36765 G36769 G36778 G36764 G36863 G36787 G36896 G36868 G36885 G36839
## 4333897      0      2      7      4      1      2     12      3      9     20
##         G36777 G36835 G36762 G36753 G36877 G36763 G36854 G36773 G36808 G36795
## 4333897      8     40      0     23      1      1      0      2     10     14
##         G36788 G36871 G36759 G36789 G36819 G36847 G36893 G36780 G36872 G36875
## 4333897     12      0      0     18      2     46      0      0      3      0
##         G36883 G36761 G36760 G36807 G36798 G36862 G36751 G36855 G36870 G36783
## 4333897      1     26      3     23      0      2      3      1     19     13
##         G36772 G36843 G36874 G36810 G36867 G36900 G36771 G36776 G36851 G36830
## 4333897      0     35    213      7      4      1      5     10      0     11
##         G36858 G36891 G36826 G36831 G36849 G36757 G36844 G36752 G36813 G36850
## 4333897      1      5      1      0     12     17      0     15     16    153
##         G36887 G36880 G36785 G36803 G36857 G36750 G36837 G36814 G36838 G36892
## 4333897      0      4     12     34      0      0     18     20     24      1
```

```
##         G36811 G36890 G36823 G36755 G36848 G35986 G35959 G35971 G36018 G36008
## 4333897     15      0     46      2      1    114      0    118      2     26
##         G35957 G35967 G35933 G36033 G35990 G35974 G35906 G35955 G36002 G36050
## 4333897      0      0     21      1     22      1      0      0     25      0
##         G36029 G35958 G36054 G36004 G35978 G36034 G35868 G36016 G36000 G36045
## 4333897    203      0      2      1    124    213     41     10      1      3
##         G35996 G35972 G35942 G36024 G36040 G35976 G36048 G35989 G35926 G36026
## 4333897      8     48      0      0    144      3      0      0      2     85
##         G35936 G35878 G35945 G36020 G36013 G35982 G35934 G35975 G35991 G35940
## 4333897      7      6      2      0      6      0      6     26     47     41
##         G36003 G36023 G36031 G36043 G35898 G35963 G35932 G35994 G36021 G35892
## 4333897      0      0     49     29      1      3      8     32      0      2
##         G36032 G36042 G35980 G35937 G35947 G35902 G35953 G35962 G35946 G35992
## 4333897      0    132      2     19      7      0      0      0      7     19
##         G35999 G35985 G35941 G35961 G35950 G36052 G36039 G36012 G36015 G35993
## 4333897      0     58      0     17     56      0    163     34      9     82
##         G35869 G35966 G35876 G36007 G36019 G36010 G35954 G36038 G35983 G36051
## 4333897     35      4     12     60     12      5      0     37     31     24
##         G36053 G36006 G35987 G35925 G36049 G36025 G35960 G36041 G35984 G35948
## 4333897      1     22     22      2      0     31      9    180      1     11
##         G36047 G35881 G36037 G35924 G35894 G35977 G35995 G36001 G36017 G35939
## 4333897     51     57    162      8     18      0     51     51     12      2
##         G36046 G35969 G35891 G36036 G35998 G35979 G35997 G35951 G35965 G35908
## 4333897      1     45      0      9      1     23      1      1      0      0
##         G36028 G35944 G36022 G35952 G36044 G35901 G35970 G35915 G36035 G35956
## 4333897     54     30     18      1      1     24    106     15      0      2
##         G35874 G35867 G36306 G35928 G36011 G36005 G35368 G35504 G35938 G35988
## 4333897      1     53     42     96     22     45      4      4     58      0
##         G35923 G35518 G35449 G35900 G36030 G35981 G35973
## 4333897     10      4     43      0     32      3     36
```

```
taxaTable <- read.table("taxa_table") #Importing the Taxa table
head(taxaTable, n = 1)
```

```
##               Domain.            Phylum.                  Class.
## 4333897; k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria;
##                          Order.                Family. Genus. Species
## 4333897; o__Enterobacteriales; f__Enterobacteriaceae;   g__;     s__
```

```
dim(otuTable)
```

```
## [1] 2240  777
```

```
dim(taxaTable)
```

```
## [1] 2240    7
```

```
class(otuTable)
```

```
## [1] "data.frame"
```

```
class(taxaTable)
```

```
## [1] "data.frame"
```

```
#Converting the Taxa and OTU Table into a Matrix
mtaxaTable <- as.matrix(taxaTable)
motuTable <- as.matrix(otuTable)
```

```
class(mtaxaTable)
```

```
## [1] "matrix" "array"
```

```
class(motuTable)
```

```
## [1] "matrix" "array"
```

```
#Data Cleansing. This is done to have consistent data across all the matrices
#This will involve making sure that the OTU/taxa row names match. Currently they dont as taxa have a tr

#rownames(mtaxaTable)[rownames(mtaxaTable) == "4333897;"] = "4333897"

head(mtaxaTable, n=1)
```

```
##          Domain.        Phylum.              Class.
## 4333897; "k__Bacteria;" "p__Proteobacteria;" "c__Gammaproteobacteria;"
##          Order.                  Family.                  Genus. Species
## 4333897; "o__Enterobacteriales;" "f__Enterobacteriaceae;" "g__;" "s__"
```

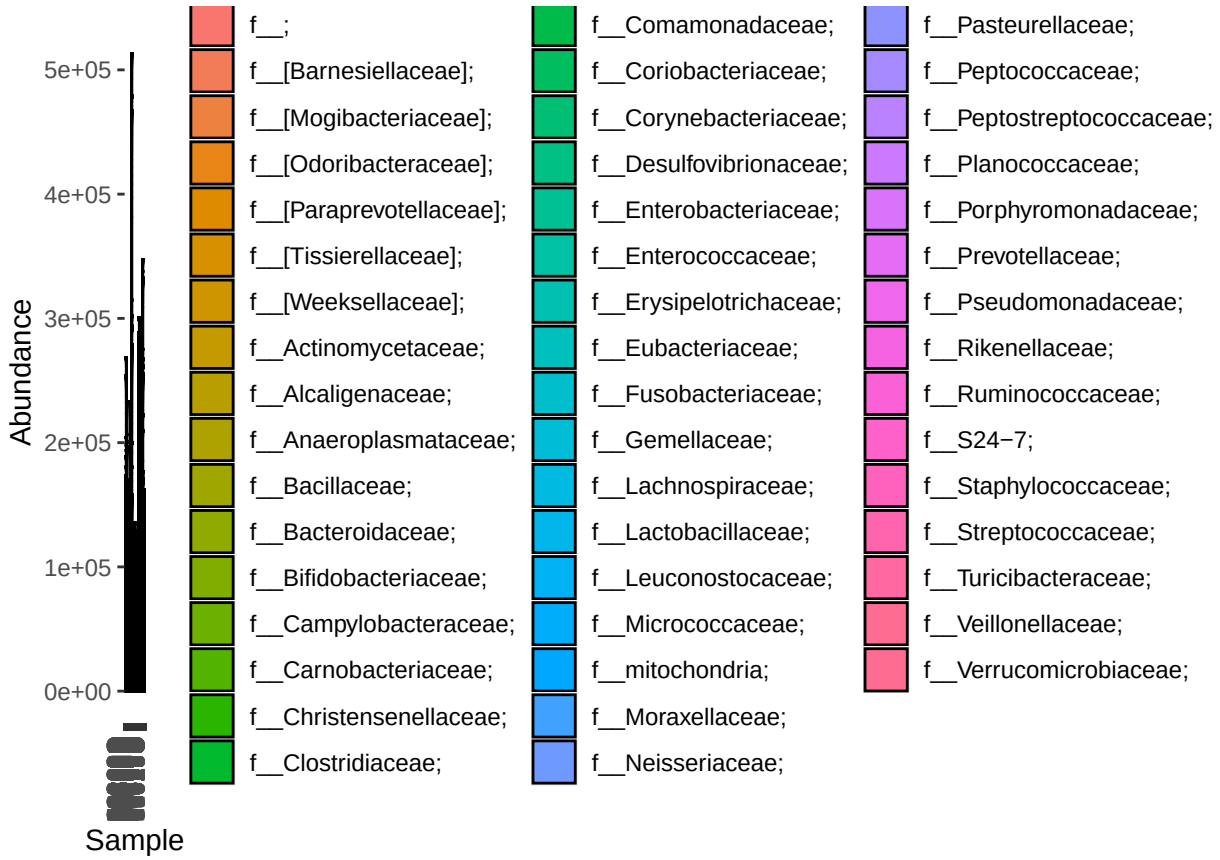
```
tnames <- rownames(mtaxaTable)  #Extract rownames from the matrix
ntnames <- gsub(x = tnames, pattern = ";", replacement = "")  #Remove the ; from the extracted rownames
#ntnames
rownames(mtaxaTable) <- ntnames  #Set the new rownames
head(mtaxaTable, n=1)
```

```
##         Domain.        Phylum.              Class.
## 4333897 "k__Bacteria;" "p__Proteobacteria;" "c__Gammaproteobacteria;"
##         Order.                  Family.                  Genus. Species
## 4333897 "o__Enterobacteriales;" "f__Enterobacteriaceae;" "g__;" "s__"
```

```
library(phyloseq)
library(ggplot2)

#Tell phyloseq to load them into a phyloseq object
OTU = otu_table(motuTable, taxa_are_rows = TRUE)
TAX = tax_table(mtaxaTable)

#OTU
#TAX

#Generating the phyloseq object
physeq = phyloseq(OTU, TAX)
physeq
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 2240 taxa and 777 samples ]
## tax_table()   Taxonomy Table:    [ 2240 taxa by 7 taxonomic ranks ]
```

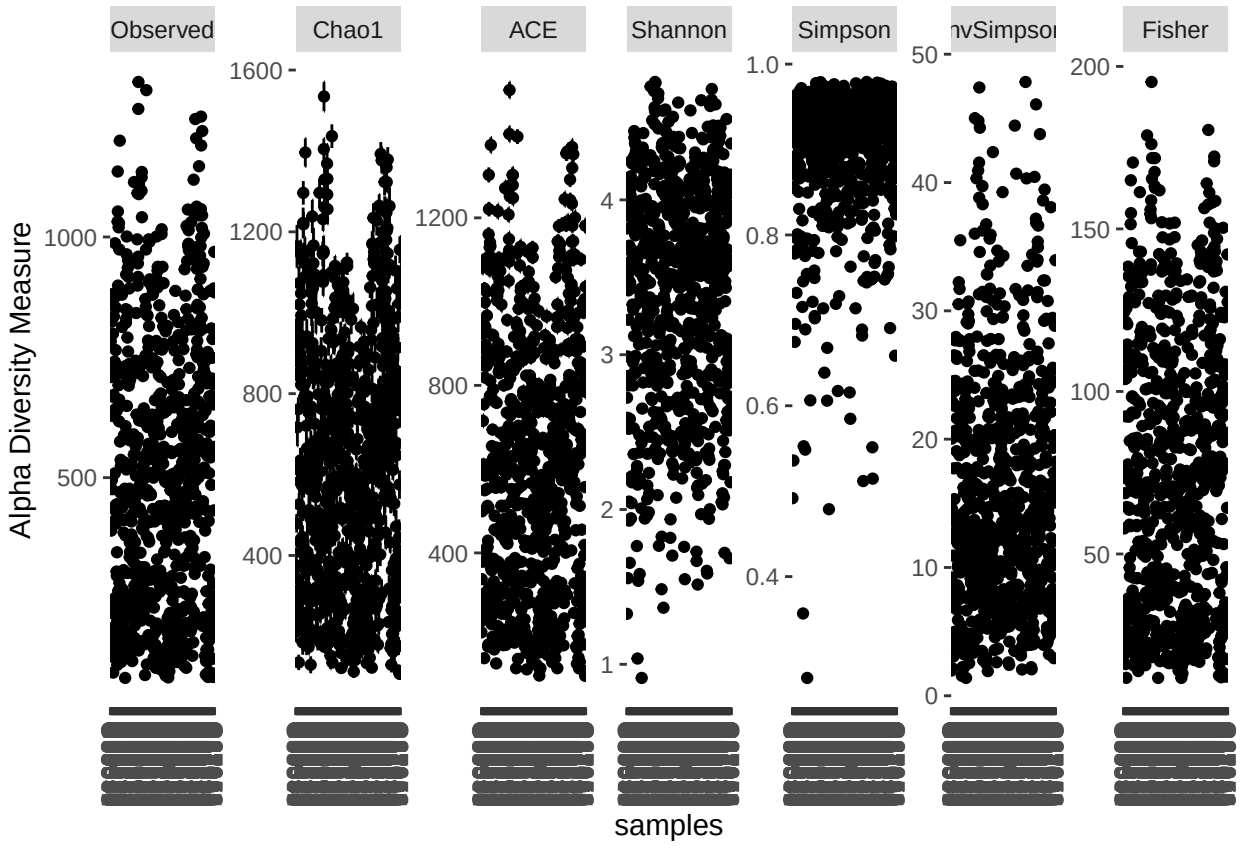
```
#Plotting the phyloseq
plot_bar(physeq, fill = "Family.")
```



Question 3.

Generate Alpha diversity plots and ordination plots. Examine any observed patterns by delivery mode, gender and disease status

```
plot_richness(physeq)  #Defauit plot produced by the plot_richness function
```




```
#plot_richness(physeq = physeq, x ="Case_Control")
```

Alpha diversity comparison between the gender in cases and control

```
#plot_richness(physeq = physeq, x ="Case_Control")
```

Alpha diversity comparison between the Delivery Mode and Case Control

Question 4

Perform differential abundance using DESEq2