MSB7102 Mini-project, Semester I, 2021

kikomeko Aloysious

6/23/2021

MAKERERE UNIVERSITY

COLLEGE OF HEALTH SCIENCES
SCHOOL OF BIOMEDICAL SCIENCES
DEPARTMENT OF IMMUNOLOGY AND MOLECULAR BIOLOGY
MASTER OF SCIENCE IN BIOINFORMATICS
COURSE UNIT: Bioconductor & R for Boioinformatics
COURSE CODE: MSB 7102
STUDENT NAME: Kikomeko Aloysious
Reg. No. 2019/HD07/24856U
Student No. 1900724856
EMAIL: kikomekoa2013@gmail.com

**

1. Import the data described above into R, **

```
t1d_metadata <- read.csv("diabimmune_16s_t1d_metadata.csv",
              stringsAsFactors = FALSE)
```

Change structure of categorical variables to factors

```
t1d_metadata$Gender <- as.factor(t1d_metadata$Gender)
t1d_metadata$Case_Control <- as.factor(t1d_metadata$Case_Control)
t1d_metadata$Delivery_Route <- as.factor(t1d_metadata$Delivery_Route)
```

i. provide descriptive summaries of the subject data (using appropriate graphics and statistical summary measures) given in the diabimmune_16s_t1d_metadata.csv file.

```
summary(t1d_metadata)
```

```
##   Sample_ID          Subject_ID         Case_Control    Gender
##  Length:777         Length:777         case   :260   female:412
##  Class :character   Class :character   control:517   male  :365
##  Mode  :character   Mode  :character
##
##
##
##   Delivery_Route Age_at_Collection
##  cesarian: 66    Min.   :   6.0
##  vaginal :711    1st Qu.: 229.0
##                  Median : 452.0
##                  Mean   : 482.9
##                  3rd Qu.: 702.0
##                  Max.   :1233.0
```

Cross tabulation Gender Vs Age at enrollment

```
addmargins(xtabs(~Gender + Case_Control, data = t1d_metadata))
```

```
##         Case_Control
## Gender   case control Sum
##   female  142     270 412
##   male    118     247 365
##   Sum     260     517 777
```

The table above show that, more females (412) were involved in the study than the males (365). A total of 260 respondents (cases), 142 were female and 118 were male. Out of the 517 respondents (Controls), 270 were female and 365 were male

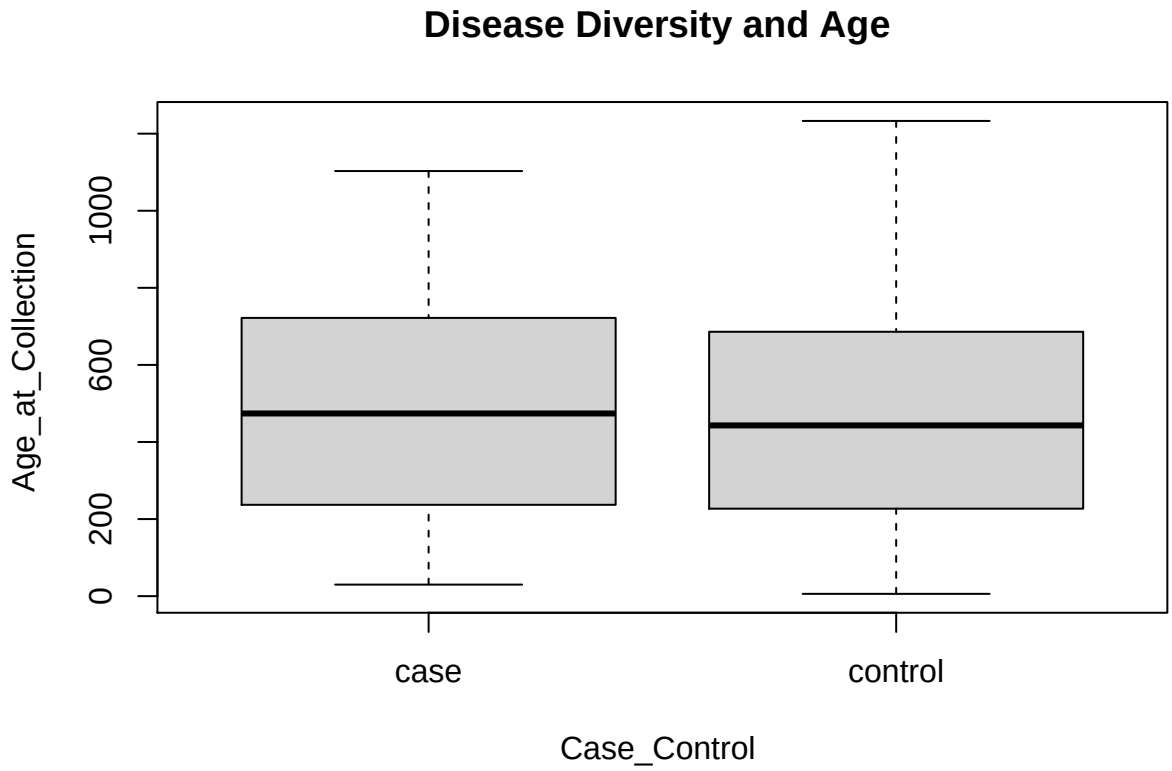
```
addmargins(xtabs(~Delivery_Route+Case_Control, data = t1d_metadata))
```

```
##                Case_Control
## Delivery_Route case control Sum
##        cesarian    0      66  66
##        vaginal   260     451 711
##        Sum       260     517 777
```

from the table above, its noted that no baby born by Cesarian section was included in the sample as a case, but 66 babies born by cesarian section were sampled as control elements. 260 babies were

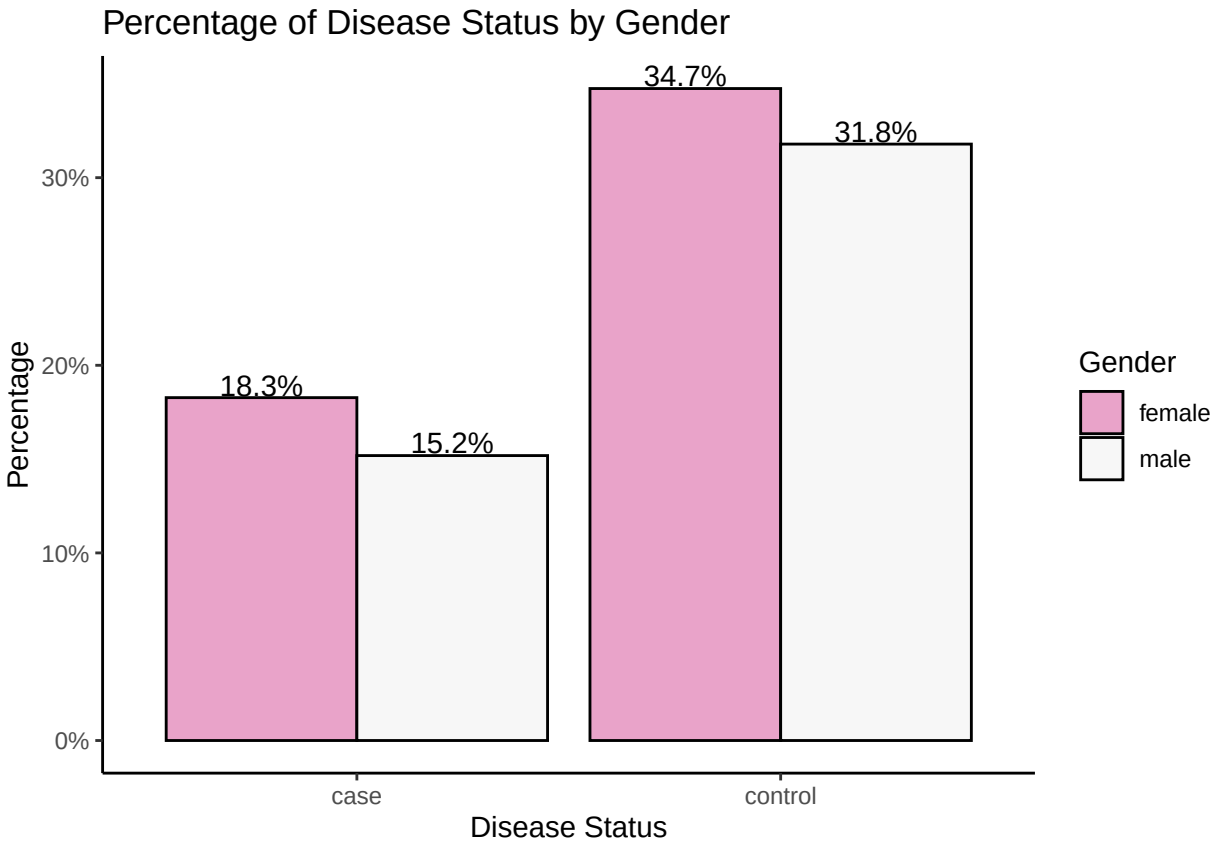
Graphics

```
plot(t1d_metadata$Case_Control, t1d_metadata$Age_at_Collection, main="Disease Diversity and Age ",
    xlab="Case_Control ", ylab="Age_at_Collection ", pch=19)
```



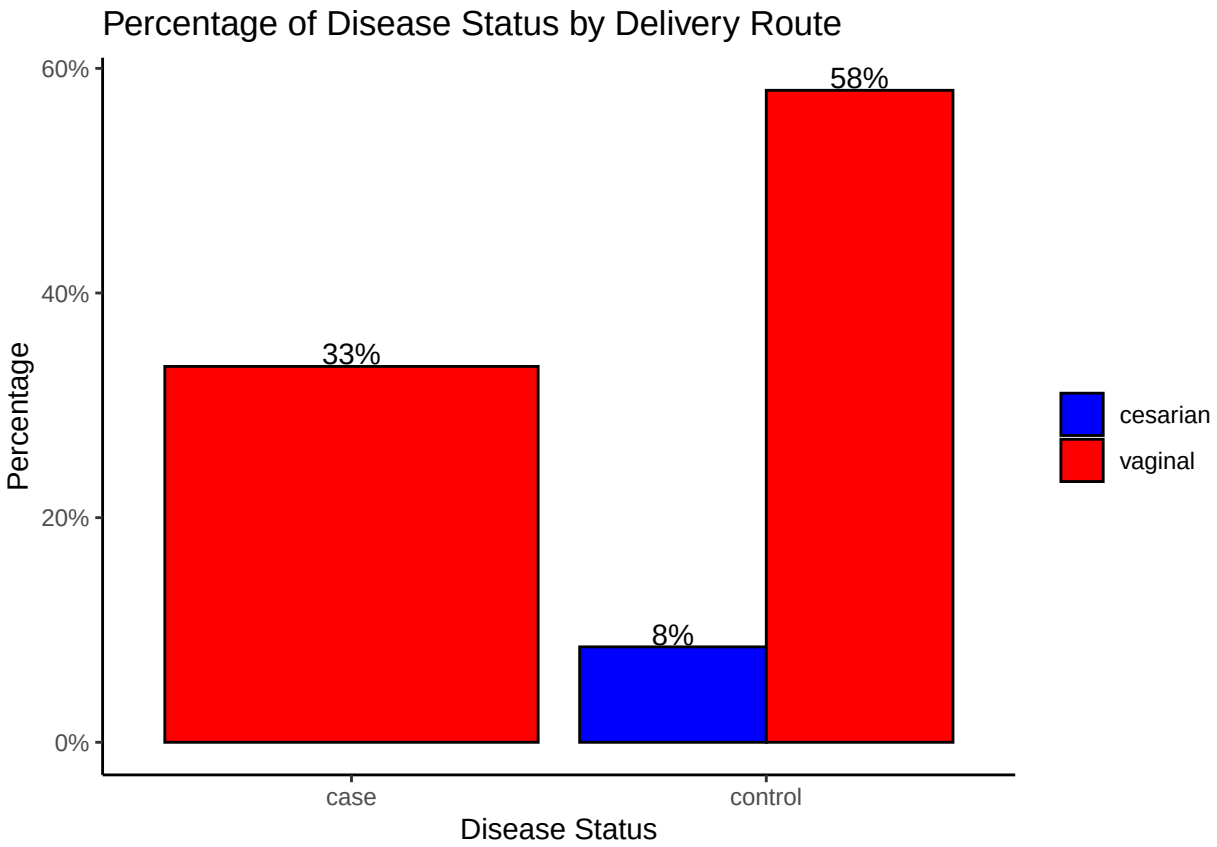
From the box plot above, it can be seen that, the age at collection for both cases and control groups, were almost at the same level of the lower quartile, Although, the study elements under cases had a higer median value and a higher upper quartile value but with a lower value in the upper and higher value for the lower whiskers. unlike in the control groups somewhat a lower median value, and lower upper quartile, they had a higher value in the upper whisker lower value in the lower whisker.

```
library (ggplot2)
bar_chart1 <- ggplot(t1d_metadata, aes(x =Case_Control ,y= ((..count..)/sum(..count..)), fill= Gender))
  geom_bar(position = "dodge", color= "Black")+
  scale_y_continuous(labels=scales::percent)+
  labs(title = "Percentage of Disease Status by Gender",
       y= "Percentage", x= "Disease Status" )+
  geom_text(stat= "count", aes(label= scales::percent((..count..)/sum(..count..))), position = position
  theme_classic()+
  scale_fill_brewer(palette = "PiYG")
  bar_chart1
```



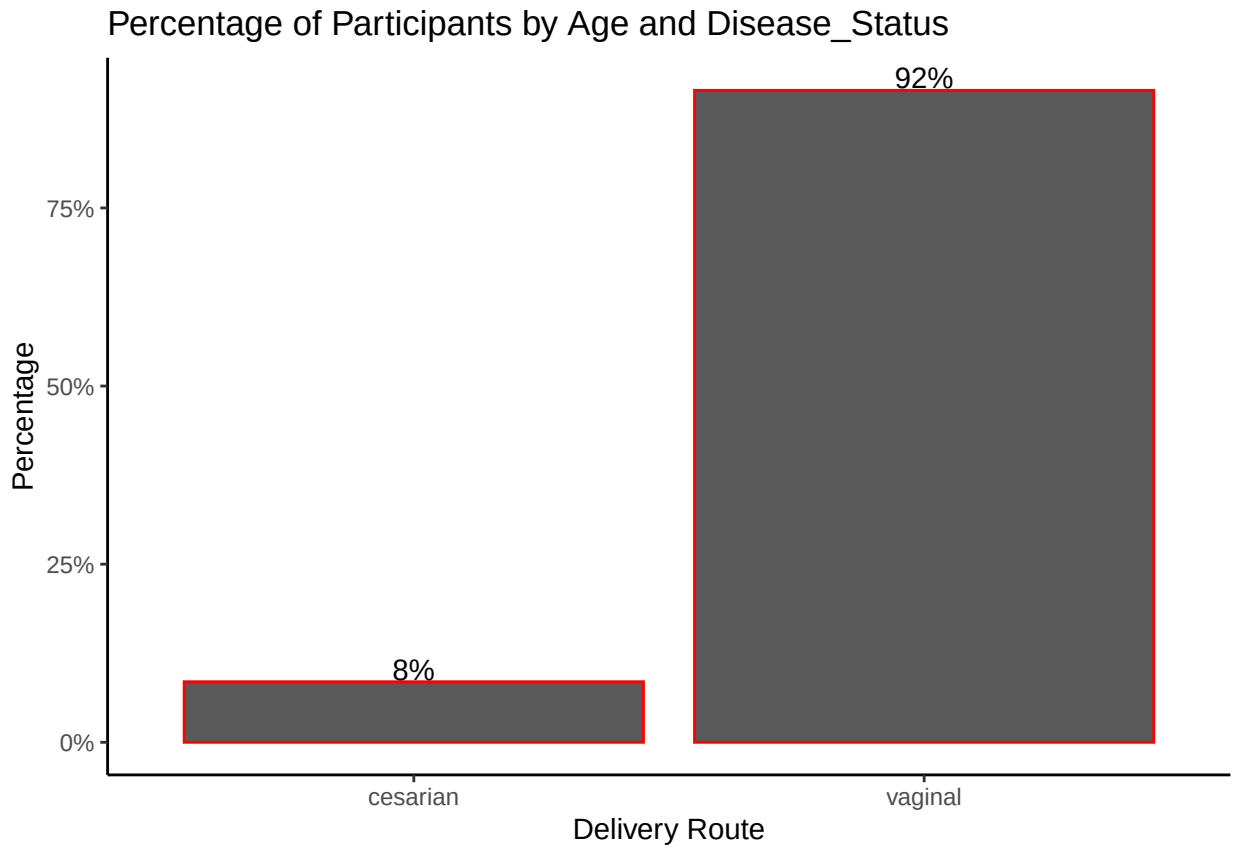
From the graph above, among the cases, 18.3% were female and 15.2% were male. Among the study control elements, 34.7% were female and 31.8% were male.

```
bar_chart2 <- ggplot(t1d_metadata, aes(x =Case_Control ,y= (..count..)/sum(..count..), fill= Delivery_R
  geom_bar(position = "dodge", color= "Black")+
  scale_y_continuous(labels=scales::percent)+
  labs(title = "Percentage of Disease Status by Delivery Route",
       y= "Percentage", x= "Disease Status" )+
  geom_text(stat= "count", aes(label= scales::percent((..count..)/sum(..count..))), position = position
  theme_classic()+
  scale_fill_manual(name = "", values = c("blue", "red"))
  bar_chart2
```



The graph above shows that, 33% of the study population were cases and had been produced through vaginal delivery, 8% of the study population among controls, had been produced through cesarian section, and 58% among control study elements had been produced through the vagianal route.

```
bar_chart3 <- ggplot(t1d_metadata, aes(x =Delivery_Route ,y= (..count..)/sum(..count..), fill= Age_at_C
  geom_bar(position = "dodge",color= "red")+
  scale_y_continuous(labels=scales::percent)+
  labs(title = "Percentage of Participants by Age and Disease_Status",
       y= "Percentage", x= "Delivery Route")+
  geom_text(stat= "count", aes(label= scales::percent((..count..)/sum(..count..))), position = position
  theme_classic()+
  scale_fill_manual(name = "", values = ("black"))
  bar_chart3
```



Its evident that, only 8% of the study population were born by cesarian section and 92% were born by vaginal delivery.

ii. 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.

```
df1 <- xtabs(~Gender + Case_Control, data = t1d_metadata)

chisq.test(df1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df1
## X-squared = 0.30687, df = 1, p-value = 0.5796
```

```
fisher.test(df1)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df1
## p-value = 0.543
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
```

```
##  0.8074825 1.5018846
## sample estimates:
## odds ratio
##   1.100742
```

The Chi-squared probability value of 0.5796, like the Fisher's Exact Test p-value (0.543), which is greater than 5% level of significance, is insignificant at the 95% level of significance, implying that gender and case_Control were independent variables.

```
df2 <- xtabs(~Delivery_Route + Case_Control, data = t1d_metadata)

chisq.test(df2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df2
## X-squared = 34.649, df = 1, p-value = 3.949e-09
```

```
fisher.test(df2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df2
## p-value = 6.86e-13
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0000000 0.1010148
## sample estimates:
## odds ratio
##          0
```

The Chi-squared probability value of 3.949e-09, like the Fisher's Exact Test p-value (6.86e-13), which is less than 5% level of significance, is significant at the 95% level of significance, implying that Delivery_route and case_Control were dependent variables.

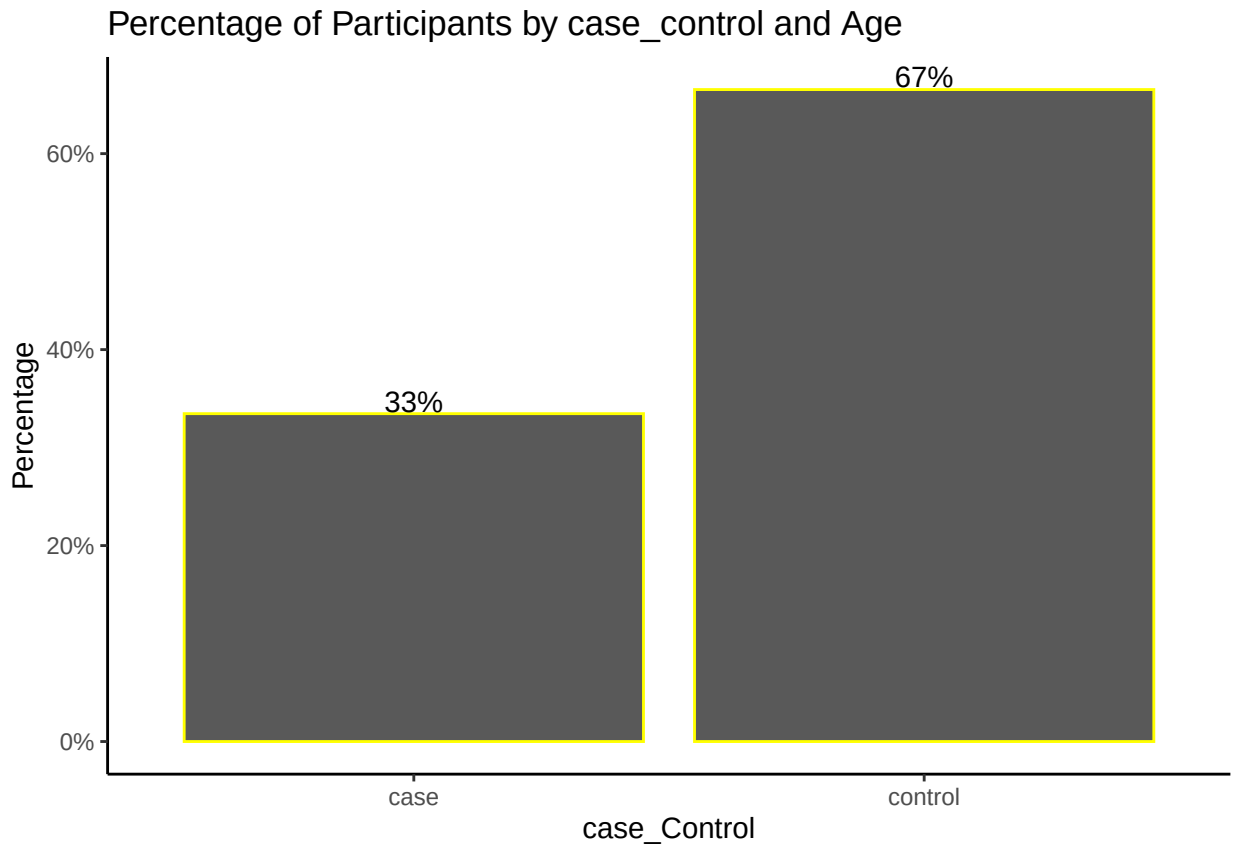
```
df3 <- xtabs(~Age_at_Collection+ Delivery_Route , data = t1d_metadata)

t.test(df3)
```

```
##
##  One Sample t-test
##
## data:  df3
## t = 28.078, df = 1085, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.6654712 0.7654681
## sample estimates:
## mean of x
## 0.7154696
```

The Chi-squared probability value (0.5186), like the Fisher's Exact Test p-value (0.5195), which is greater than 5% level of significance, is insignificant at the 95% level of significance, implying that gender and case_Control were independent variables.

```
bar_chart4 <- ggplot(t1d_metadata, aes(x =Case_Control ,y= (..count..)/sum(..count..), fill= Age_at_Coll
  geom_bar(position = "dodge",color= "yellow")+
  scale_y_continuous(labels=scales::percent)+
  labs(title = "Percentage of Participants by case_control and Age",
       y= "Percentage", x= "case_Control")+
  geom_text(stat= "count", aes(label= scales::percent((..count..)/sum(..count..))), position = position_
  theme_classic()+
  scale_fill_manual(name = "", values = c("black"))
  bar_chart4
```



Its evident that, the biggest proportion of the respondents 67%, were considered as controls irrespective of the age, while 33% of the study population were cases in the study.

```
model = glm(t1d_metadata$Case_Control ~ t1d_metadata$Gender + t1d_metadata$Delivery_Route + t1d_metadata
summary(model)
```

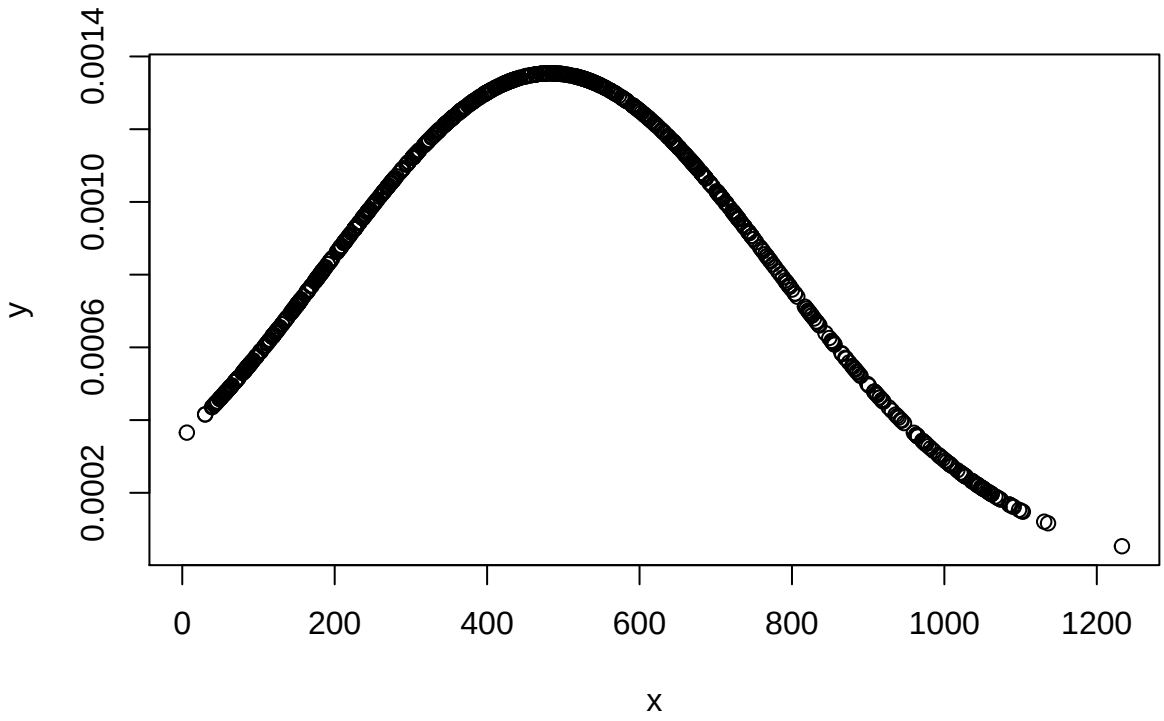
```
##
## Call:
## glm(formula = t1d_metadata$Case_Control ~ t1d_metadata$Gender +
##     t1d_metadata$Delivery_Route + t1d_metadata$Age_at_Collection,
##     family = binomial)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5032  -1.3854   0.9051   0.9596   1.0564
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.763e+01  4.862e+02   0.036    0.971
## t1d_metadata$Gendermale           1.375e-01  1.566e-01   0.878    0.380
## t1d_metadata$Delivery_Routevaginal -1.702e+01  4.862e+02  -0.035    0.972
## t1d_metadata$Age_at_Collection    -2.614e-04  2.648e-04  -0.987    0.324
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 990.52  on 776  degrees of freedom
## Residual deviance: 932.07  on 773  degrees of freedom
## AIC: 940.07
##
## Number of Fisher Scoring iterations: 16
```

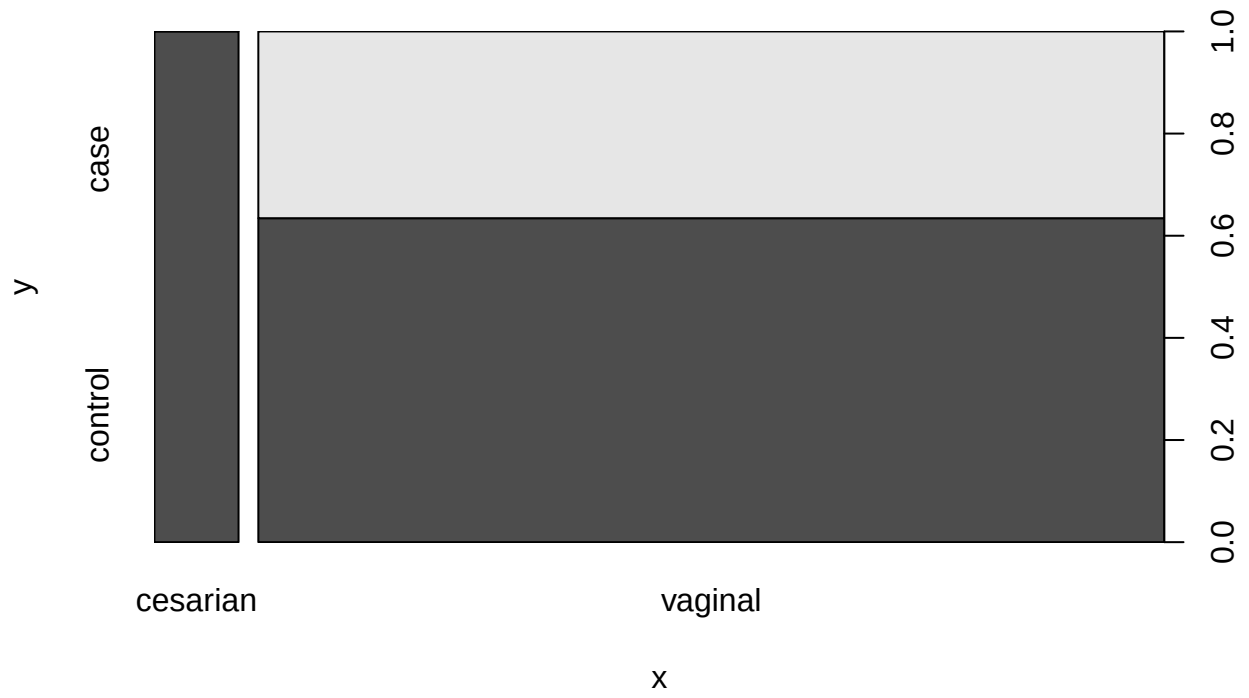
```
anova(model, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: t1d_metadata$Case_Control
##
## Terms added sequentially (first to last)
##
##
##                                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                              776     990.52
## t1d_metadata$Gender             1    0.397       775     990.13    0.5285
## t1d_metadata$Delivery_Route     1   57.083       774     933.04 4.178e-14 ***
## t1d_metadata$Age_at_Collection  1    0.973       773     932.07    0.3239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
x <- t1d_metadata$Age_at_Collection
y <- dnorm(t1d_metadata$Age_at_Collection, mean(t1d_metadata$Age_at_Collection), sd(t1d_metadata$Age_at
plot(x, y)
```




```
plot(t1d_metadata$Delivery_Route, t1d_metadata$Case_Control)
```



From the linear regression model above, we can conclude that, inclusion the study elements was significantly depended on the deliverly route unlike the Age_at_enrollment and gender.

The normal distribution curve indicates that, the age of most of the study elements were 600 days of life old and bellow. This is due to the fact that, most of the data is left skewed.

The stacked bar plot shows that, all study elements delivered by cesarian section were considered as controls, and only those produced through spontaneous vaginal delivered, were used as both cases and controls.

```
model1 = anova(model, test = "Chisq")
summary(model1)
```

```
##        Df       Deviance         Resid. Df       Resid. Dev       Pr(>Chi)
##  Min.   :1   Min.   : 0.3974   Min.   :773.0   Min.   :932.1   Min.   :0.0000
##  1st Qu.:1   1st Qu.: 0.6853   1st Qu.:773.8   1st Qu.:932.8   1st Qu.:0.1619
##  Median :1   Median : 0.9732   Median :774.5   Median :961.6   Median :0.3239
##  Mean   :1   Mean   :19.4845   Mean   :774.5   Mean   :961.4   Mean   :0.2841
##  3rd Qu.:1   3rd Qu.:29.0281   3rd Qu.:775.2   3rd Qu.:990.2   3rd Qu.:0.4262
##  Max.   :1   Max.   :57.0829   Max.   :776.0   Max.   :990.5   Max.   :0.5285
##  NA's   :1   NA's   :1                                         NA's   :1
```

```
model1 = anova(model, test = "Cp")
summary(model1)
```

```
##        Df       Deviance         Resid. Df       Resid. Dev          Cp
##  Min.   :1   Min.   : 0.3974   Min.   :773.0   Min.   :932.1   Min.   :939.0
```

```
##  1st Qu.:1   1st Qu.: 0.6853   1st Qu.:773.8   1st Qu.:932.8   1st Qu.:939.8
##  Median :1   Median : 0.9732   Median :774.5   Median :961.6   Median :966.3
##  Mean   :1   Mean   :19.4845   Mean   :774.5   Mean   :961.4   Mean   :966.4
##  3rd Qu.:1   3rd Qu.:29.0281   3rd Qu.:775.2   3rd Qu.:990.2   3rd Qu.:992.9
##  Max.   :1   Max.   :57.0829   Max.   :776.0   Max.   :990.5   Max.   :994.1
##  NA's   :1   NA's   :1
```

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).

```
library(tidyverse)
taxa.data <- read_tsv("diabimmune_t1d_16s_otu_table.txt", skip = 1)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   ConsensusLineage = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
otu.data <- taxa.data %>% select(1:778)
taxonomy <- taxa.data %>% select(`#OTU ID`, ConsensusLineage) %>%
separate(ConsensusLineage, c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "Species"), sep =
sampledata <- t1d_metadata
```

Required to have similar row names in OTU and taxonomy table

Change OTU and taxonomy dataframes to matrices

```
otu.data_matx <- as.matrix(otu.data)
taxonomy_matx <- as.matrix(taxonomy)
```

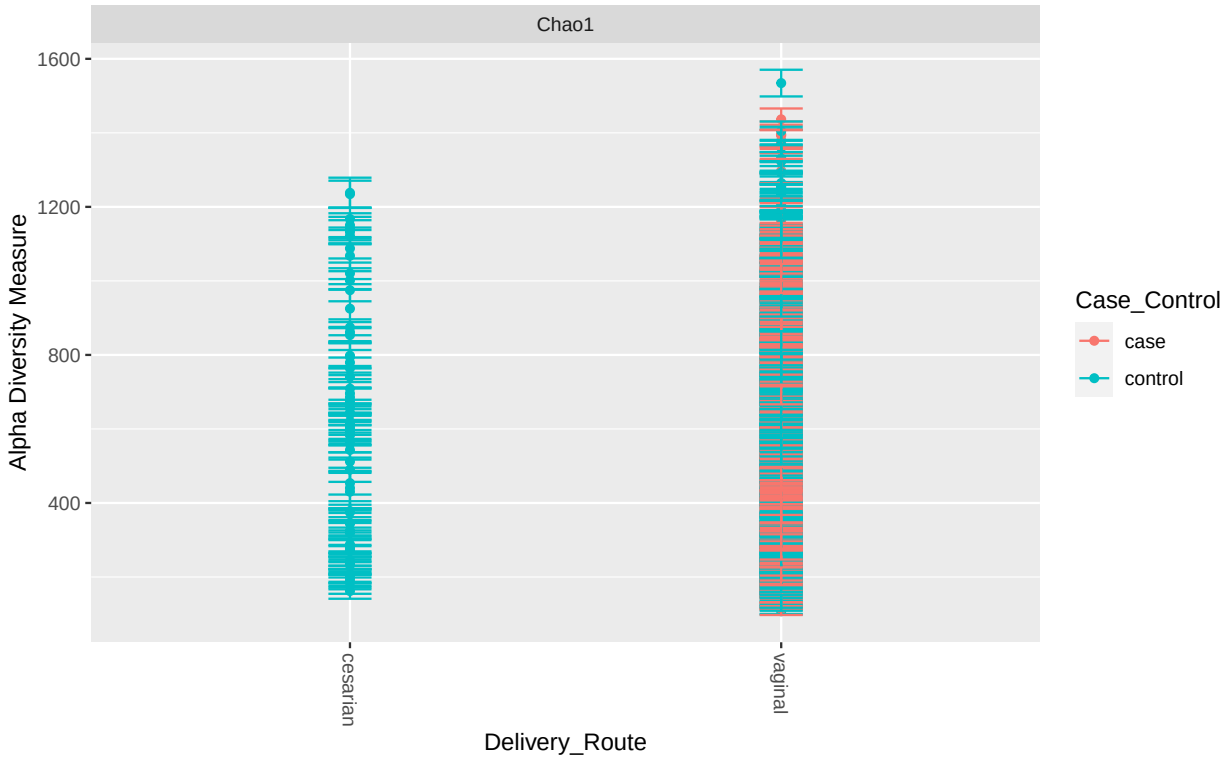
create phyloseq object

Validity checks

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

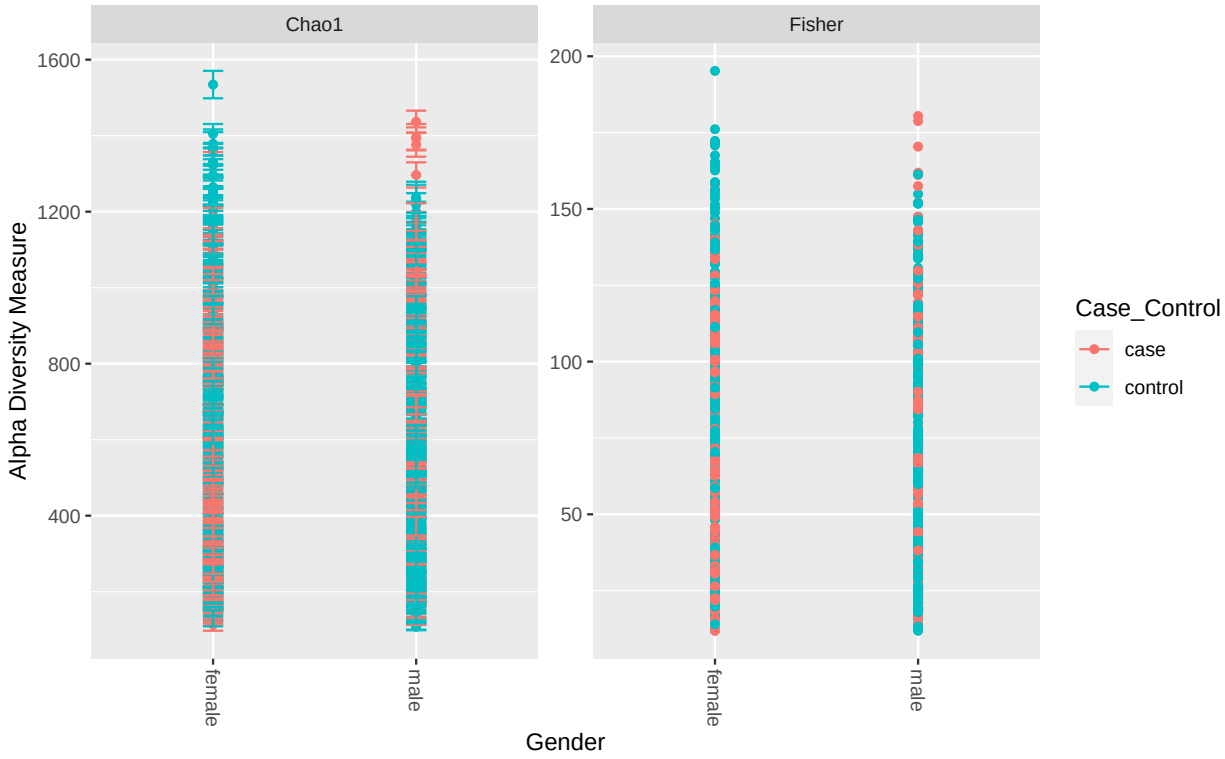
Alpha Diversity

```
Fig.A <- plot_richness(phyloseq.objectA,x = "Delivery_Route", color="Case_Control", measures= "Chao1")
Fig.A
```



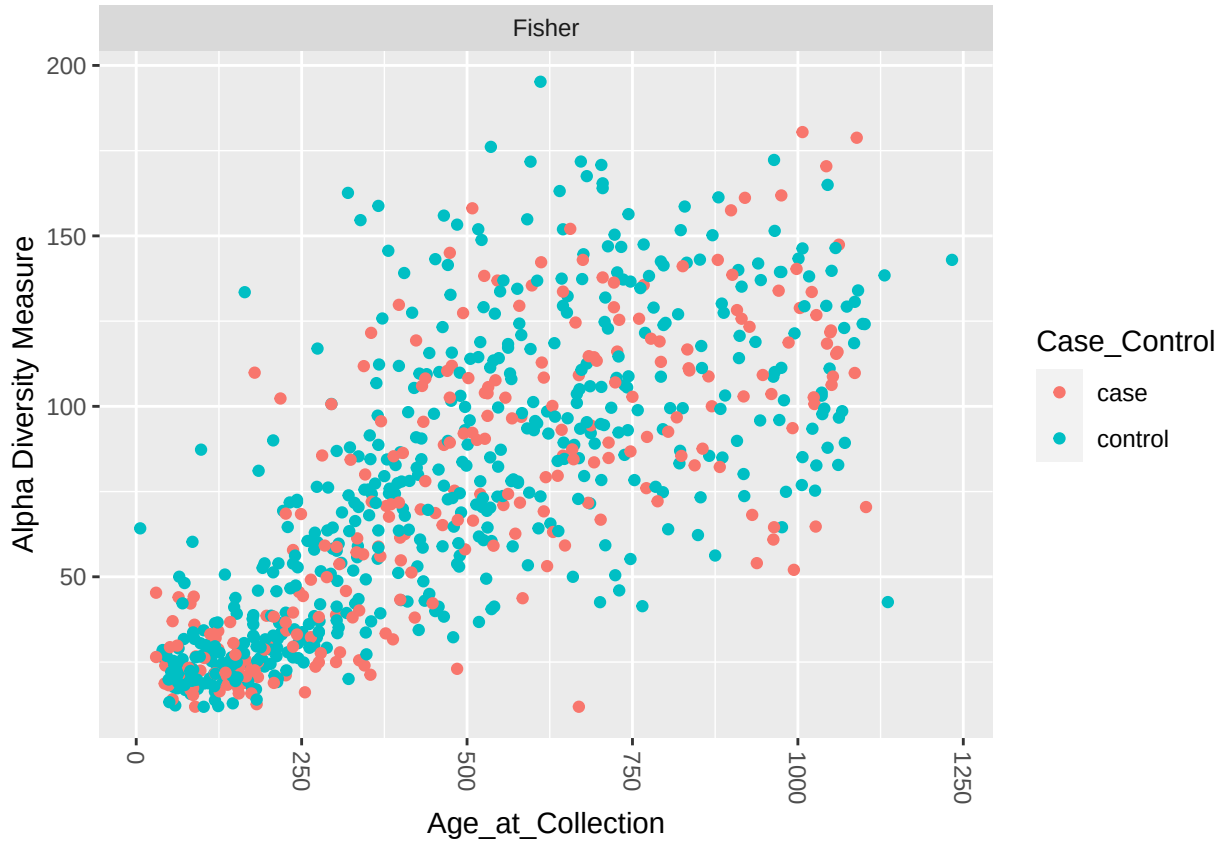
Its evident that all there were no cases involved in the sample that had been delivered by cesarian section. impling that all cases involved in the studay had been born by vaginal route.

```
Fig.B <- plot_richness(phyloseq.objectA,x = "Gender", color="Case_Control", measures= c("Chao1", "fishe
Fig.B
```



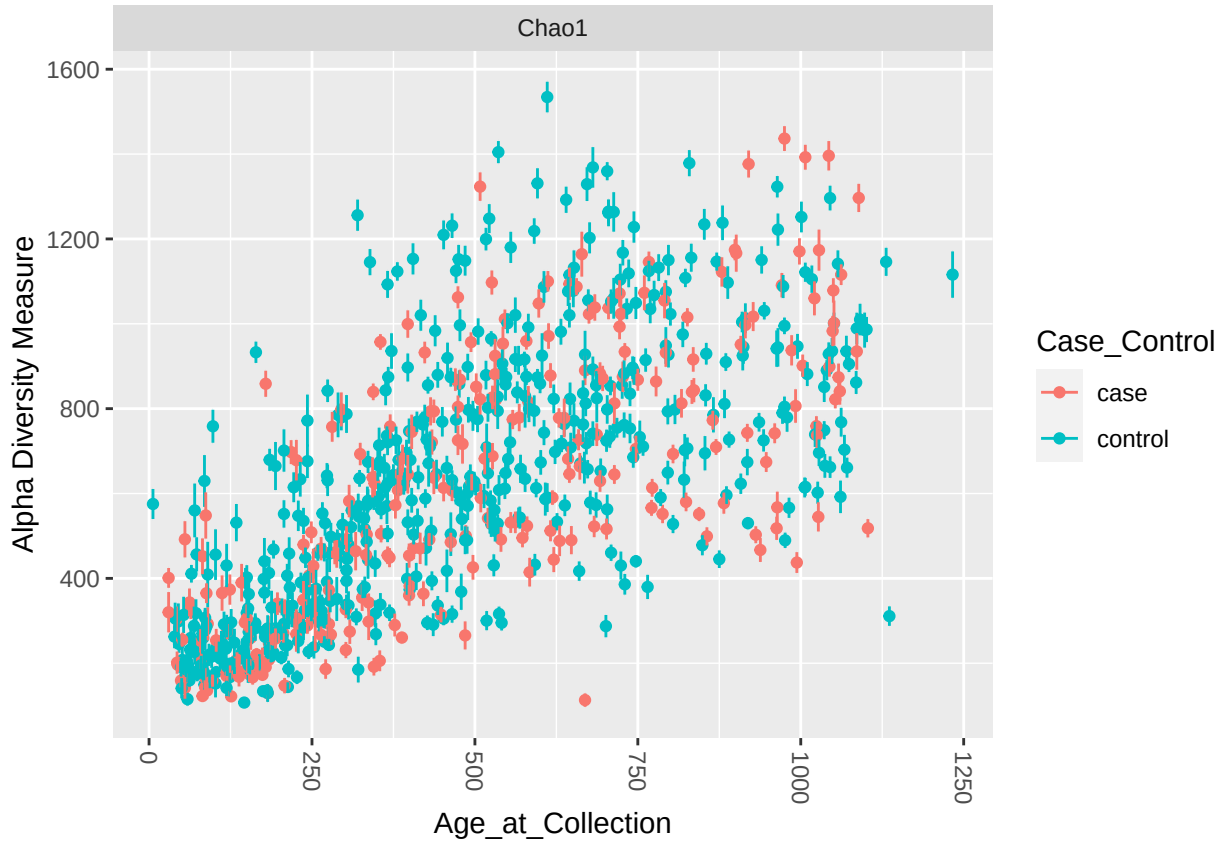
The distribution shows that, the male and female were equally involved in the study both as cases and controls.

```
Fig.C <- plot_richness(phyloseq.objectA,x = "Age_at_Collection", color="Case_Control", measures= c("fis
Fig.C
```



Its evident that, most of the elements involved in the study were bellow 750 days of life and fall within 0 to 150 measure of diversity.

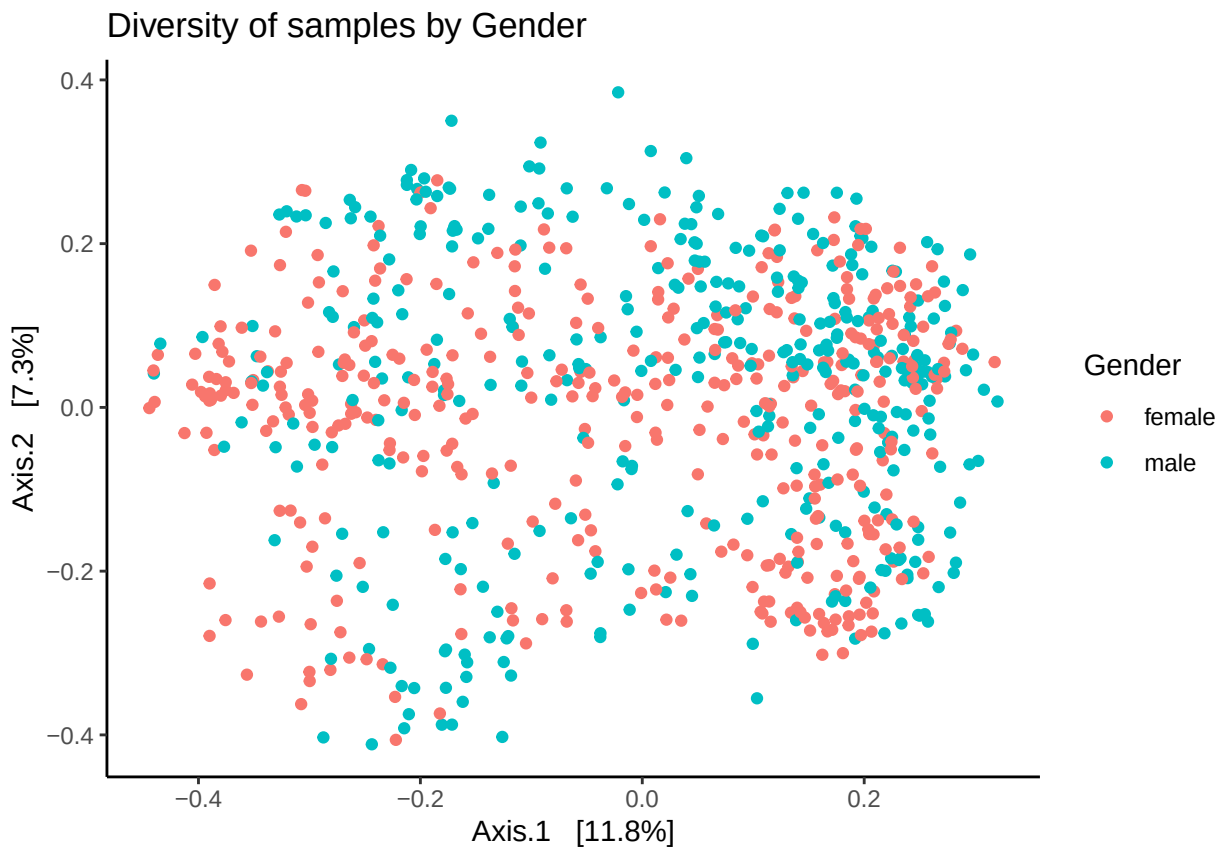
```
Fig.D <- plot_richness(phyloseq.objectA,x = "Age_at_Collection", color="Case_Control", measures= "Chao1"
Fig.D
```



like the Fisher test, the choa1 test of richness also shows that, most of the elements involved in the study were bellow 750 days of life and fall within 0 to 150 measure of diversity.

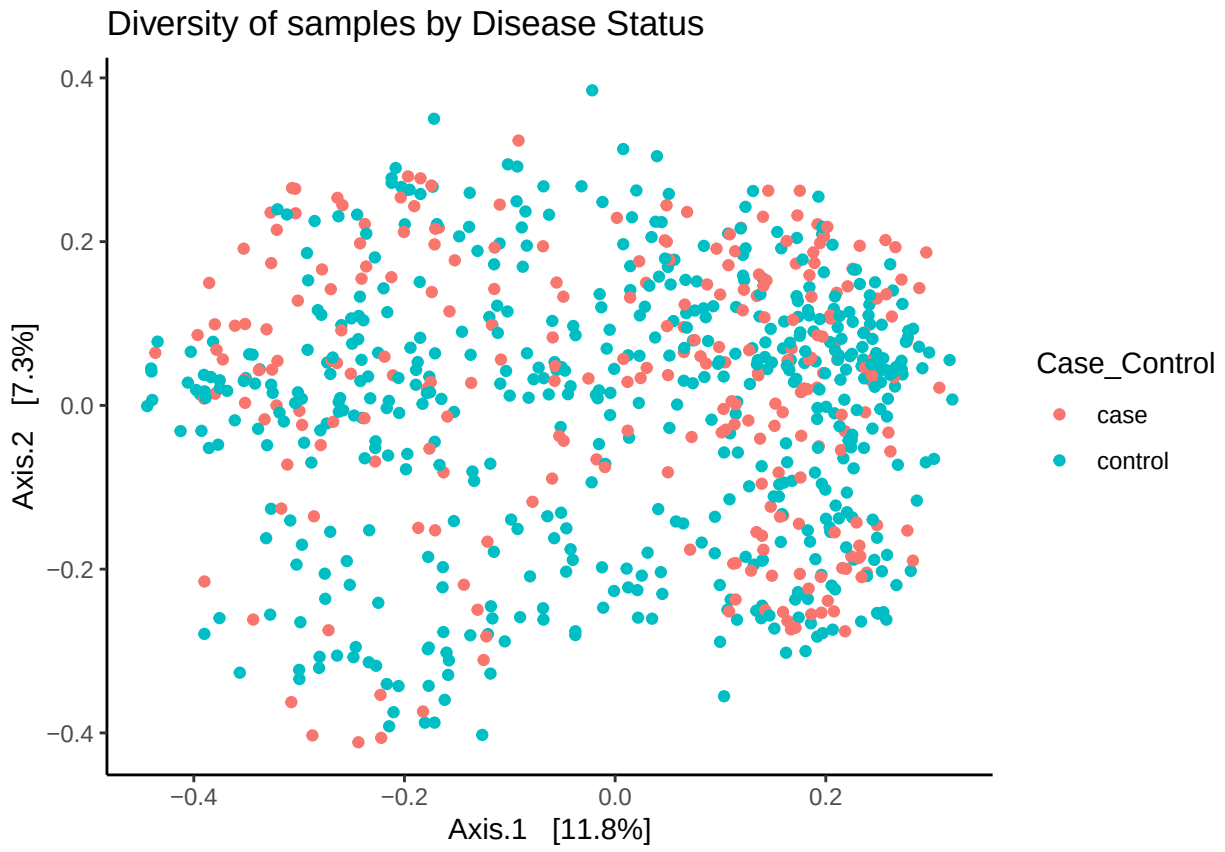
Ordination Plots

```
gf <- ordinate(phyloseq.objectA, "PCoA", "bray")%>%
plot_ordination(phyloseq.objectA, .,color = "Gender", title = "Diversity of samples by Gender")+
  theme_classic()
gf
```




```
library(phyloseq)
ordinate <- ordinate(phyloseq.objectA, "PCoA", "bray")

gf2 <- plot_ordination(phyloseq.objectA, ordinate,color = "Case_Control", title = "Diversity of samples
  theme_classic()
gf2
```

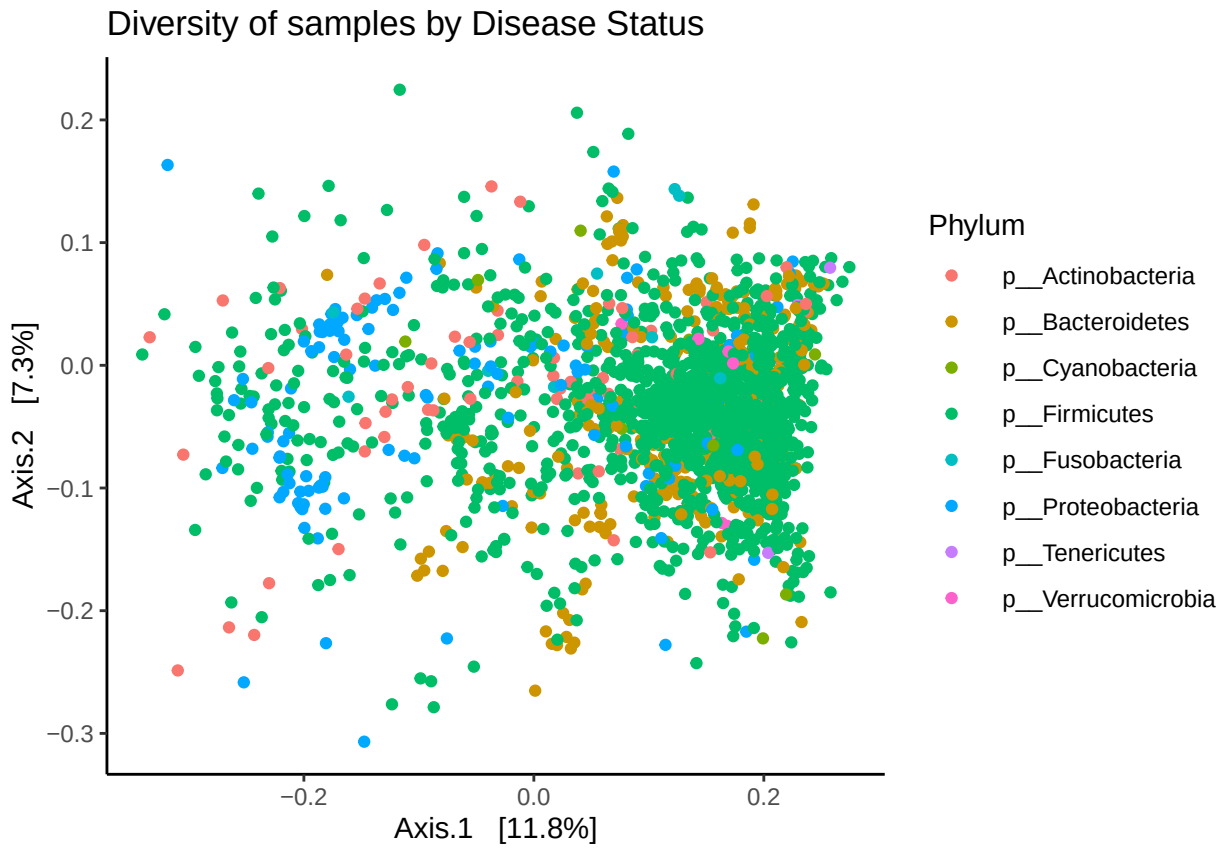



```
library(phyloseq)
ordinate1 <-ordinate(phyloseq.objectA, "NMDS", "bray")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.181114
## Run 1 stress 0.1862791
## Run 2 stress 0.1840098
## Run 3 stress 0.1852188
## Run 4 stress 0.1928129
## Run 5 stress 0.186378
## Run 6 stress 0.1833052
## Run 7 stress 0.1825132
## Run 8 stress 0.1828612
## Run 9 stress 0.1875457
## Run 10 stress 0.1821221
## Run 11 stress 0.1837213
## Run 12 stress 0.1839571
## Run 13 stress 0.1865901
## Run 14 stress 0.1883869
## Run 15 stress 0.1851983
## Run 16 stress 0.1855685
## Run 17 stress 0.1845901
## Run 18 stress 0.1834561
## Run 19 stress 0.1887101
```

```
## Run 20 stress 0.1845662
## *** No convergence -- monoMDS stopping criteria:
##      1: no. of iterations >= maxit
##      6: stress ratio > sratmax
##     13: scale factor of the gradient < sfgrmin
```

```
gf3 <-plot_ordination(phyloseq.objectA, ordinate,type="taxa", color="Phylum", title = "Diversity of samp
  theme_classic()
gf3
```



4. Perform differential abundance using DEseq2

DESeq: Creating a DESeq object

```
phyloseq_objectB <- phyloseq(otu_table(otu.data_matx + 1, taxa_are_rows=TRUE), tax_table(taxonomy_matx)
sample_data(phyloseq_objectB) <- sample_data1

casecontrol = phyloseq_to_deseq2(phyloseq_objectB, ~ Case_Control)
```

```
## converting counts to integer mode

##   it appears that the last variable in the design formula, 'Case_Control',
```

```
##   has a factor level, 'control', which is not the reference level. we recommend
##   to use factor(...,levels=...) or relevel() to set this as the reference level
##   before proceeding. for more information, please see the 'Note on factor levels'
##   in vignette('DESeq2').
```

DESeq test

```
casecontrol = DESeq(casecontrol, test="Wald", fitType="parametric")
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## -- note: fitType='parametric', but the dispersion trend was not well captured by the
##    function: y = a/x + b, and a local regression fit was automatically substituted.
##    specify fitType='local' or 'mean' to avoid this message next time.

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 991 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

Results table

```
res = results(casecontrol, cooksCutoff = FALSE)
alpha = 0.01
casetab = res[which(res$padj < alpha), ]
casetab = cbind(as(casetab, "data.frame"), as(tax_table(phyloseq_objectB)[rownames(casetab), ], "matrix
```

plot_theme

```
theme_set(theme_bw())
scale_fill_discrete <- function(palname = "Set1", ...) {
  scale_fill_brewer(palette = palname, ...)
}
```

Phylum order

```
x = tapply(casetab$log2FoldChange, casetab$Phylum, function(x) max(x))
x = sort(x, TRUE)
casetab$Phylum = factor(as.character(casetab$Phylum), levels=names(x))
```

Genus order

```
x = tapply(casetab$log2FoldChange, casetab$Genus, function(x) max(x))
x = sort(x, TRUE)
casetab$Genus = factor(as.character(casetab$Genus), levels=names(x))
ggplot(casetab, aes(x=Genus, y=log2FoldChange, color=Phylum)) + geom_point(size=6) +
  theme(axis.text.x = element_text(angle = -90, hjust = 0, vjust=0.5))
```