Untitled

2023-03-03

Reading in the data

Read in the data, and set the data types to the correct types Drop index number from the pandas dataframe and NCBI.tax.ID

```
# Connect to the DB
conn <- dbConnect(SQLite(),"../../../s16.sqlite")
# List of all the tables
dbListTables(conn)
```

```
##  [1] "bacdive"                         "bacdiveByspecies2gcf"
##  [3] "gcf2species"                     "ribdif_bacdive_joined"
##  [5] "ribdif_info"                     "s16full_sequence"
##  [7] "species"                         "species2V1V9sequence"
##  [9] "species2V3V4sequence"            "species2s16full_sequence"
## [11] "species_gcf2species_ribdif_info" "sqlite_sequence"
## [13] "taxInfoFull"                     "v1v9sequence"
## [15] "v3v4sequence"
```

```
D_tmp <- dbGetQuery(conn, "SELECT * FROM ribdif_bacdive_joined")
D_tmp <- tibble(D_tmp, .name_repair ="universal")
```

```
## New names:
## * ‘polymyxin b‘ -> ‘polymyxin.b‘
## * ‘penicillin g‘ -> ‘penicillin.g‘
## * ‘pipemidic acid‘ -> ‘pipemidic.acid‘
## * ‘actinomycin d‘ -> ‘actinomycin.d‘
## * ‘sodium dodecyl sulfate‘ -> ‘sodium.dodecyl.sulfate‘
## * ‘sodium chloride‘ -> ‘sodium.chloride‘
## * ‘cefotaxime sodium‘ -> ‘cefotaxime.sodium‘
## * ‘nalidixic acid‘ -> ‘nalidixic.acid‘
## * ‘clavulanic acid‘ -> ‘clavulanic.acid‘
## * ‘co-trimoxazole‘ -> ‘co.trimoxazole‘
## * ‘spiramycin II‘ -> ‘spiramycin.II‘
## * ‘NCBI tax ID‘ -> ‘NCBI.tax.ID‘
## * ‘strain designation‘ -> ‘strain.designation‘
## * ‘gram stain‘ -> ‘gram.stain‘
## * ‘oxygen tolerance‘ -> ‘oxygen.tolerance‘
## * ‘PH range‘ -> ‘PH.range‘
## * ‘GC-content‘ -> ‘GC.content‘
## * ‘Total samples‘ -> ‘Total.samples‘
## * ‘soil counts‘ -> ‘soil.counts‘
## * ‘aquatic counts‘ -> ‘aquatic.counts‘
## * ‘animal counts‘ -> ‘animal.counts‘
## * ‘plant counts‘ -> ‘plant.counts‘
```

```
D_tmp <- mutate(D_tmp, across(antibiotics:PH.range, factor))
D_tmp <- select(D_tmp, !c(NCBI.tax.ID,strain.designation))

# Chainging AR with no annotation to PNR
D_tmp <- mutate(D_tmp, antibiotics = ifelse(is.na(antibiotics), "PNR", "R"))
```

Splitting up data

```
set.seed(25022023)
# Adding ID as a column
D_tmp %<>% mutate(ID = row_number(species))
# Randomly selecting the training/exploration data with seed set
D <- D_tmp %>% slice_sample(prop = 0.7)
# Assigning the rest of the data to the test dataset
D_test <- anti_join(D_tmp, D, by = "ID")

# Checking if its correctly split up
percent_in_test <- nrow(D_test)/(nrow(D)+nrow(D_test))
percent_in_test
```

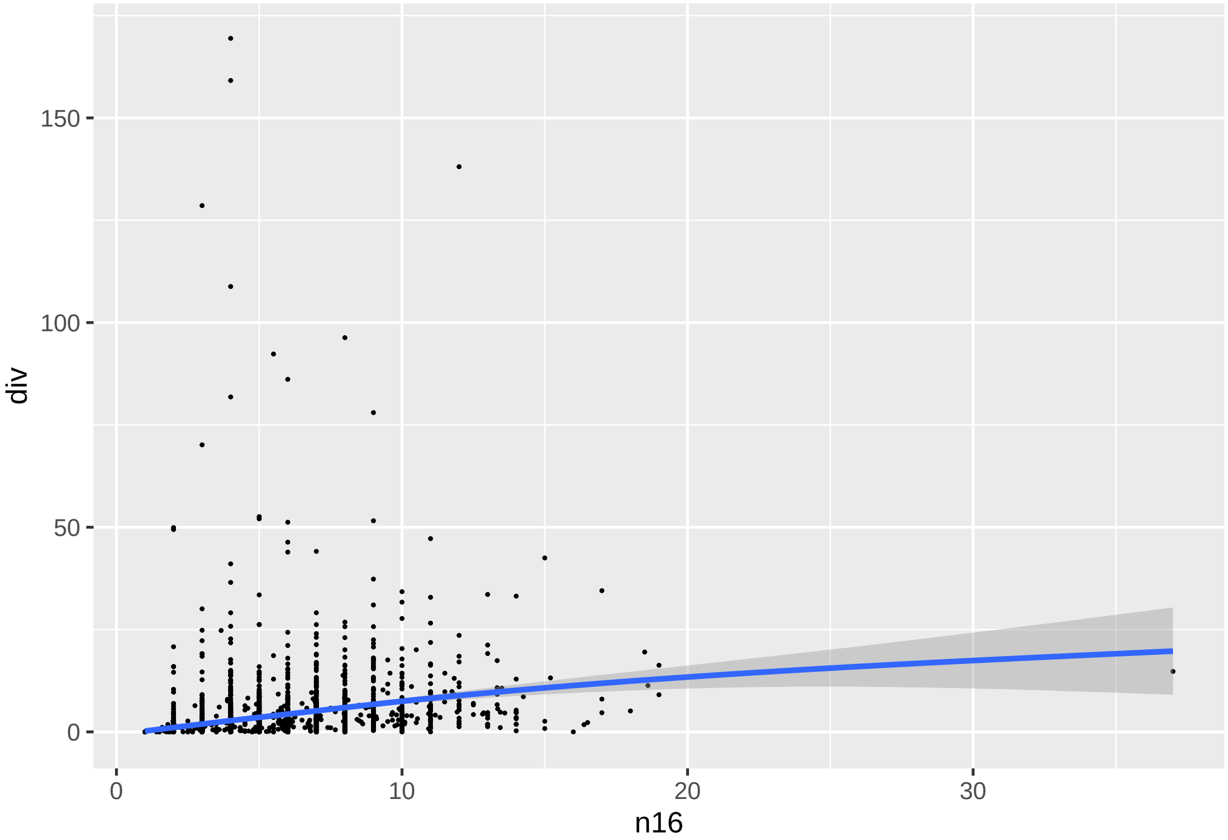
```
## [1] 0.300025
```

Modeling

The goal of this part is to build a model which takes into consideration n16 and taxonomic relationships as it seems they might have a big impact. We could either: 1) Remove all positions with n16 = 0, as they are not going to include any information about the relationship between bacterial ecology and div. Here the intercept could now be set to (0,0) or not 2) Fit a model with a varying intercept 3) Fit a model with 0,0 as intercept as described in the next paragraph Below we can see div against n16. The main takeway is that we have a lot small values and a few large, therefore there is an arguemt for applying a transformation to both axis.

```
ggplot(D,aes(x=n16, y=div)) +
  geom_point(size = 0.2) +
  geom_smooth()
```

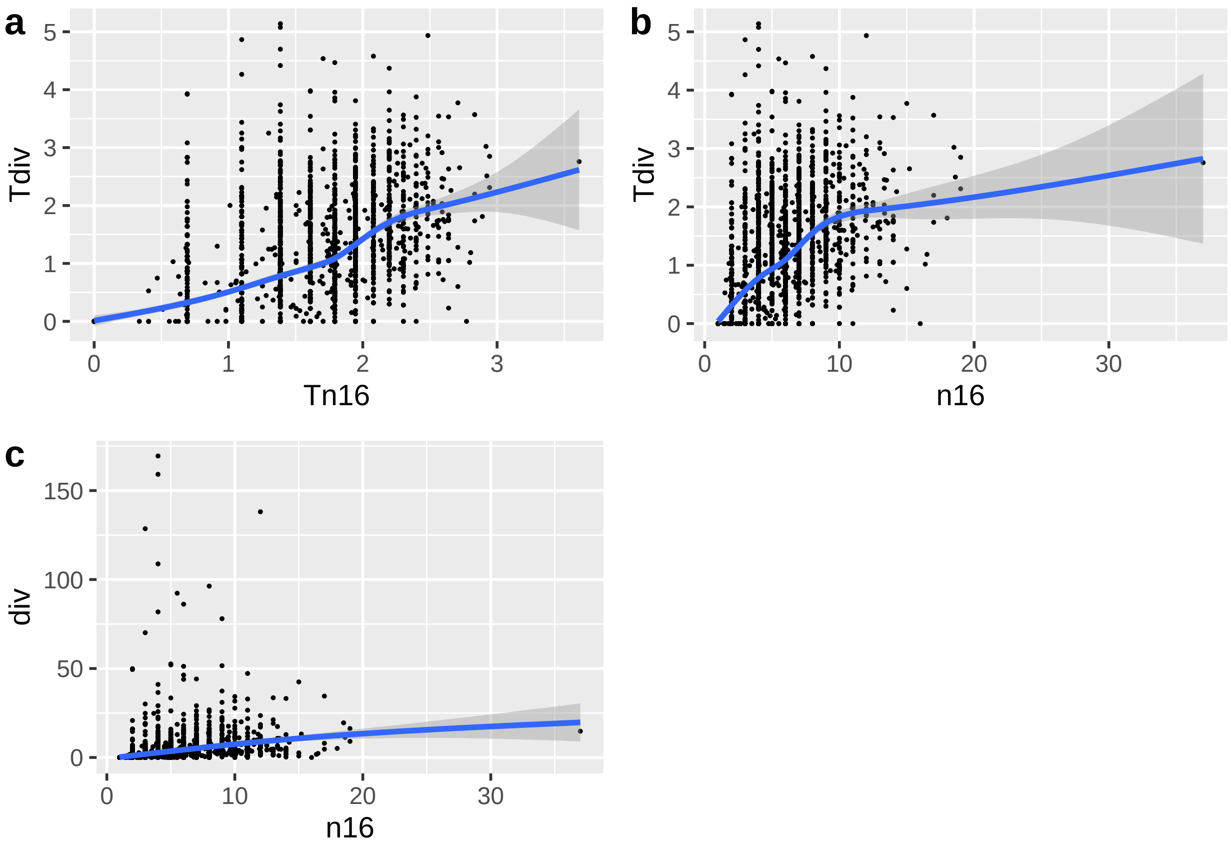
```
## ‘geom_smooth()‘ using method = ’gam’ and formula = ’y ~ s(x, bs = "cs")’
```



Transformations

```
Dt <- D %>%
  mutate(Tn16=log(n16), Tdiv=log1p(div))

# Plotting the different transformations
p1 <- ggplot(Dt,aes(x=Tn16, y=Tdiv)) +
  geom_point(size=0.2) +
  geom_smooth()

p2 <- ggplot(Dt,aes(x=n16, y=Tdiv)) +
  geom_point(size=0.2) +
  geom_smooth()

p3 <- ggplot(Dt,aes(x=n16, y=div)) +
  geom_point(size=0.2) +
  geom_smooth()

plot_grid(p1,p2,p3,labels ="auto")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

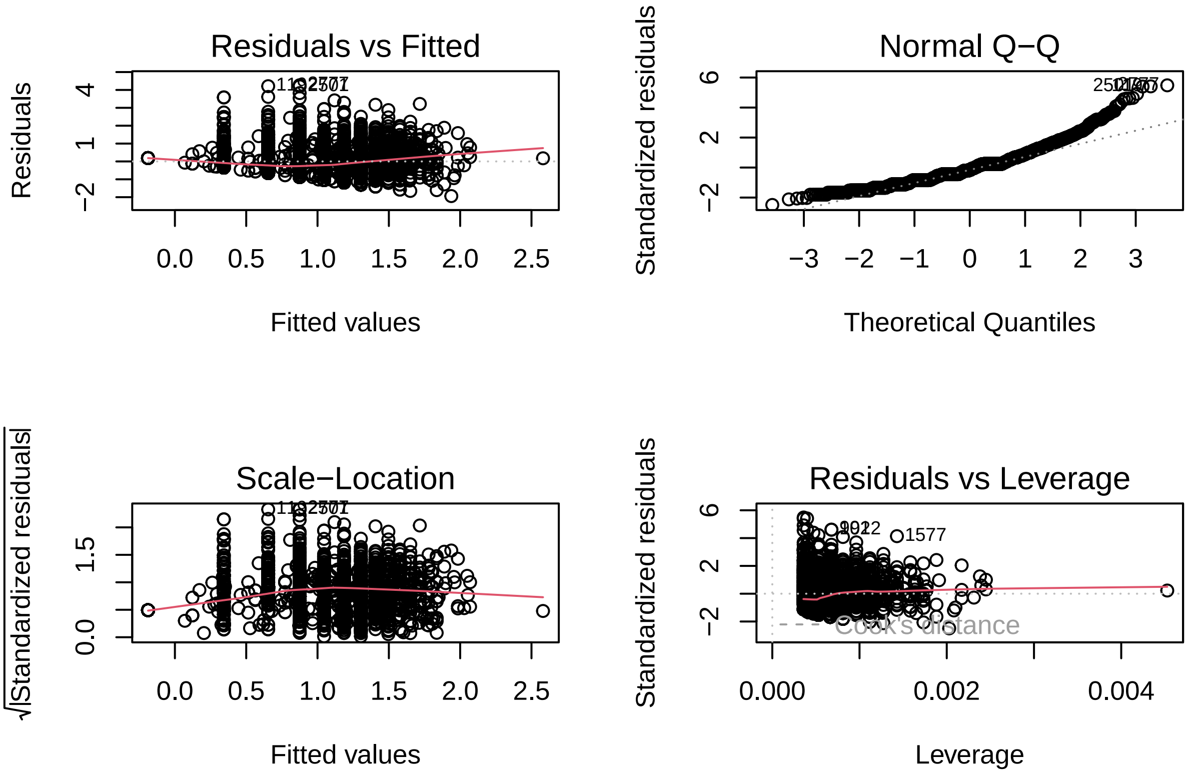


## Looking at residuals for each

```
a <- lm(Tdiv ~ Tn16  ,Dt)
b <- lm(Tdiv ~ n16  ,Dt)
c <- lm(div ~ n16 ,Dt)
```

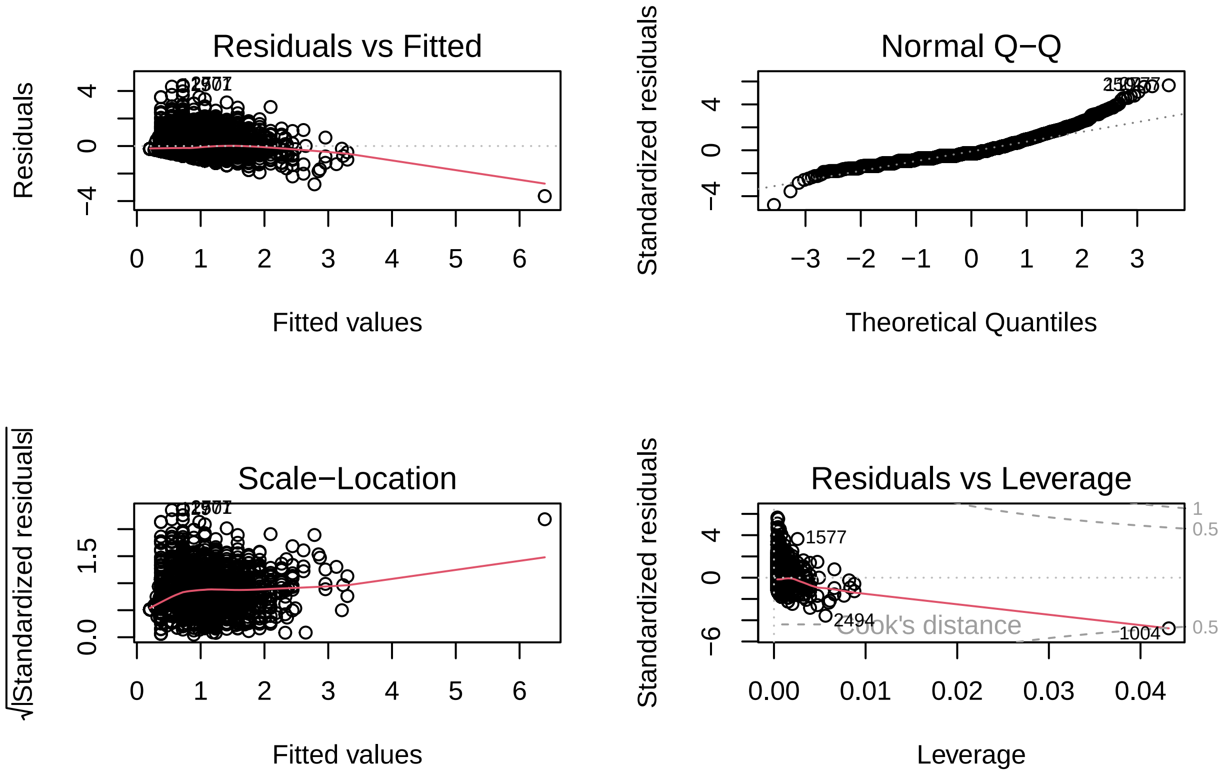
a)

```
par(mfrow=c(2,2))
plot(a)
```



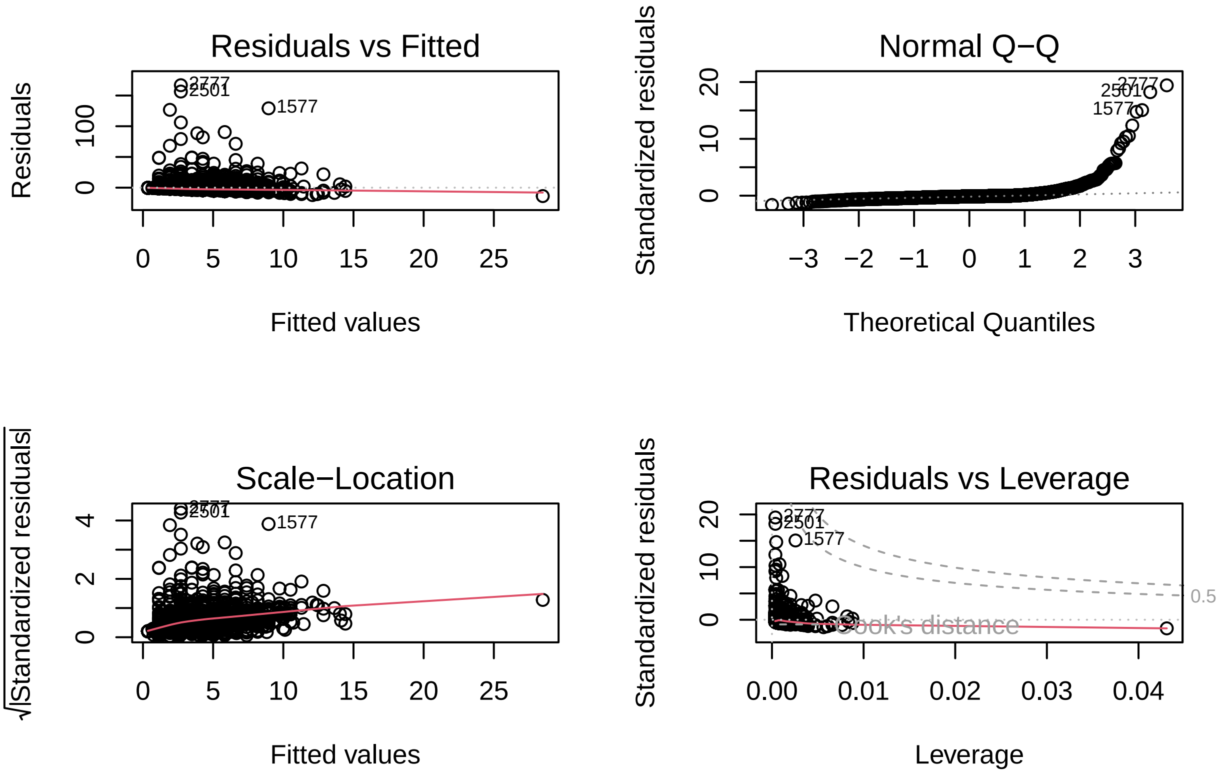
b)

```
par(mfrow=c(2,2))
plot(b)
```



c)

```
par(mfrow=c(2,2))
plot(c)
```

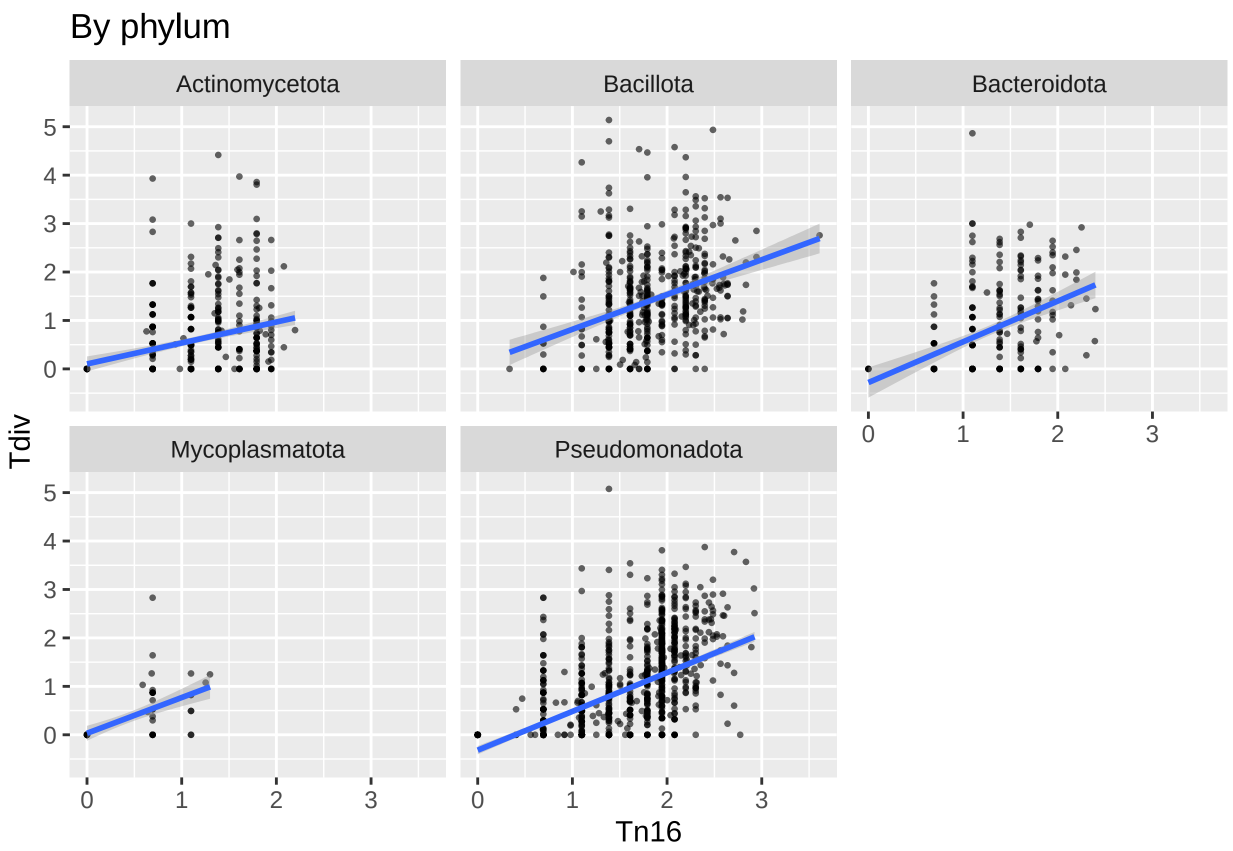


Based on this im going to go with the transformation of both sides. Since these transformations decrease the leverage of larger numbers

Taxonomic information in the model

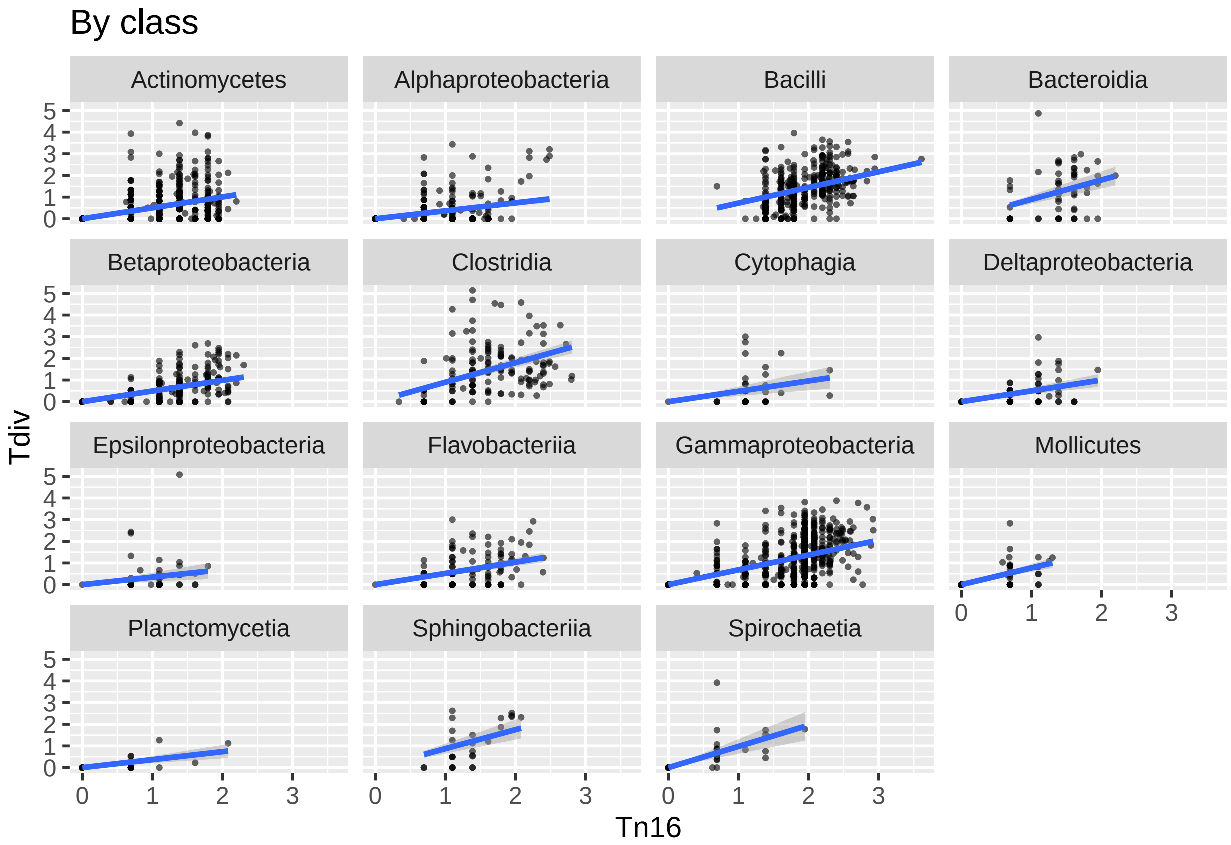
Lets try and add taxonomic information to the mdoel Lets first visualize the phylums with over 20 entries Here it seems that there could be some gain in including phylum in the model, as it seems to have an effect

```r
# Plotting for different phylum
Dt %>%
  group_by(phylum) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 50) %>%
  ggplot(aes(x=Tn16, y=Tdiv)) +
    geom_point(size=0.5, alpha=0.6) +
    theme(legend.position="none") +
    facet_wrap(~phylum) +
    geom_smooth(method=lm, formula = "y~x") +
    ggtitle("By phylum")
```



Lets also have a look for class,

```
library(magrittr)
# Plotting for different orders
Dt %>%
  group_by(class) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20) %>%
  ggplot(aes(x=Tn16, y=Tdiv)) +
    geom_point(size=0.5, alpha=0.6) +
    theme(legend.position="none") +
    facet_wrap(~class) +
    geom_smooth(method=lm, formula = "y~x+0") +
    ggtitle("By class")
```



Here it's hard to see how much information we lose by just including phylum instead of order. Therefore lets try and remove the effect of phylum by plotting the residuals of a simple model

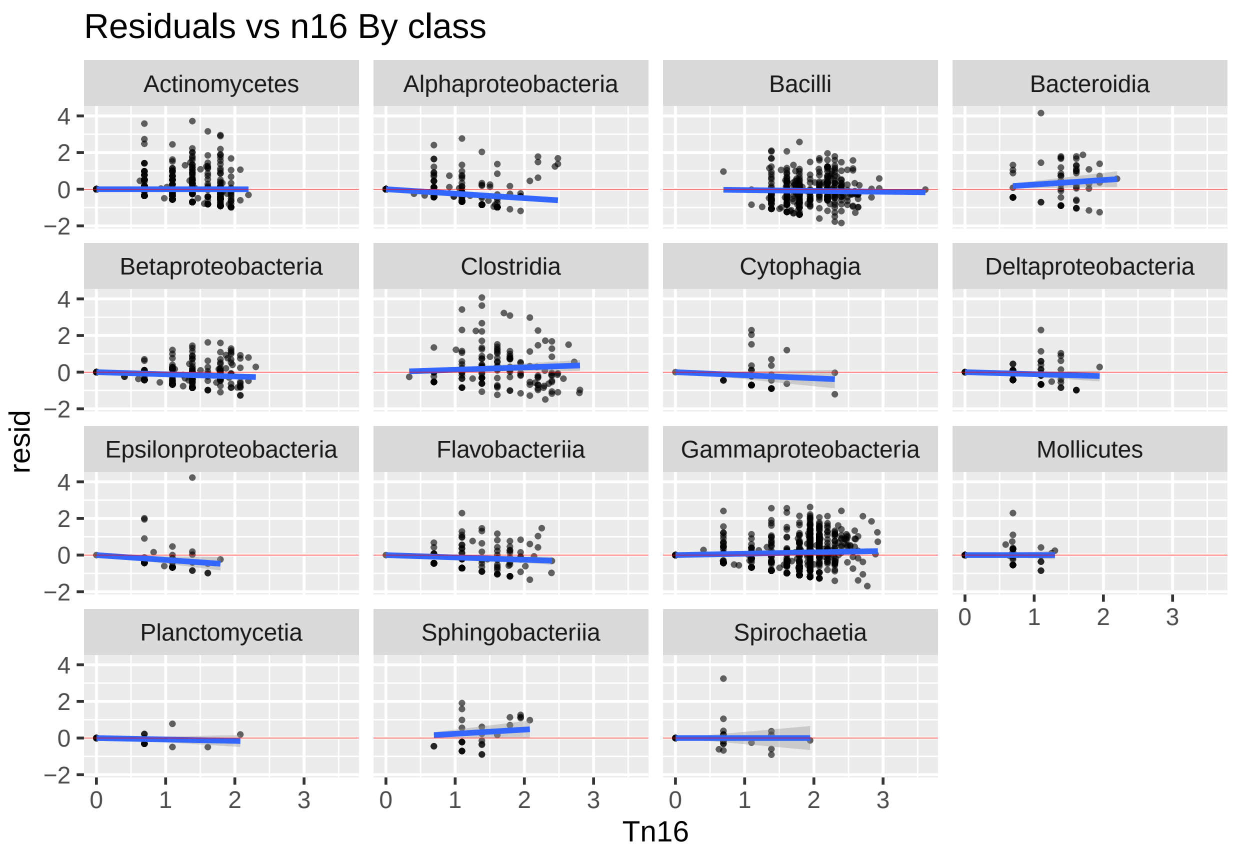
```
fitTaxPhylum <- lm(Tdiv ~ 0 + Tn16 + Tn16:factor(phylum)  ,Dt)
summary(fitTaxPhylum)
```

```
##
## Call:
## lm(formula = Tdiv ~ 0 + Tn16 + Tn16:factor(phylum), data = Dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8448 -0.5536 -0.1236  0.3196  4.2307
##
## Coefficients: (6 not defined because of singularities)
##                                        Estimate Std. Error t value Pr(>|t|)
## Tn16                                    0.01997    0.52411   0.038   0.9696
## Tn16:factor(phylum)Acidobacteriota           NA         NA      NA       NA
## Tn16:factor(phylum)Actinomycetota       0.48564    0.52487   0.925   0.3549
## Tn16:factor(phylum)Aquificota           1.04606    0.74121   1.411   0.1583
## Tn16:factor(phylum)Atribacterota        0.73973    1.23051   0.601   0.5478
## Tn16:factor(phylum)Bacillota            0.74936    0.52441   1.429   0.1531
## Tn16:factor(phylum)Bacteroidota         0.62609    0.52532   1.192   0.2334
## Tn16:factor(phylum)Caldisericota             NA         NA      NA       NA
## Tn16:factor(phylum)Calditrichota             NA         NA      NA       NA
## Tn16:factor(phylum)Chlamydiota          0.86991    0.68126   1.277   0.2017
## Tn16:factor(phylum)Chlorobiota              0.10484    0.64191   0.163   0.8703
## Tn16:factor(phylum)Chloroflexota            1.74112    0.79222   2.198   0.0280
## Tn16:factor(phylum)Chrysiogenota            0.84671    0.87641   0.966   0.3341
## Tn16:factor(phylum)Cyanobacteriota          1.24310    0.76457   1.626   0.1041
## Tn16:factor(phylum)Deferribacterota         0.11785    0.70679   0.167   0.8676
## Tn16:factor(phylum)Deinococcota             0.54957    0.55091   0.998   0.3186
## Tn16:factor(phylum)Dictyoglomerota          0.35988    0.94574   0.381   0.7036
## Tn16:factor(phylum)Elusimicrobiota               NA         NA      NA       NA
## Tn16:factor(phylum)Fibrobacterota           1.90851    0.87641   2.178   0.0295
## Tn16:factor(phylum)Fusobacteriota           0.78454    0.53492   1.467   0.1426
## Tn16:factor(phylum)Gemmatimonadota         -0.01997    1.23051  -0.016   0.9871
## Tn16:factor(phylum)Kiritimatiellota        -0.01997    1.23051  -0.016   0.9871
## Tn16:factor(phylum)Mycoplasmatota           0.75417    0.54729   1.378   0.1683
## Tn16:factor(phylum)Nitrospirota             1.23480    1.23051   1.003   0.3157
## Tn16:factor(phylum)Planctomycetota          0.42520    0.55739   0.763   0.4456
## Tn16:factor(phylum)Pseudomonadota           0.58985    0.52431   1.125   0.2607
## Tn16:factor(phylum)Rhodothermota                 NA         NA      NA       NA
## Tn16:factor(phylum)Spirochaetota            0.95724    0.55206   1.734   0.0830
## Tn16:factor(phylum)Synergistota             1.17997    0.68456   1.724   0.0849
## Tn16:factor(phylum)Thermodesulfobacteriota  1.23480    1.23051   1.003   0.3157
## Tn16:factor(phylum)Thermomicrobiota         0.98727    0.94574   1.044   0.2966
## Tn16:factor(phylum)Thermotogota             0.52067    0.58231   0.894   0.3713
## Tn16:factor(phylum)Verrucomicrobiota             NA         NA      NA       NA
##
## Tn16
## Tn16:factor(phylum)Acidobacteriota
## Tn16:factor(phylum)Actinomycetota
## Tn16:factor(phylum)Aquificota
## Tn16:factor(phylum)Atribacterota
## Tn16:factor(phylum)Bacillota
## Tn16:factor(phylum)Bacteroidota
## Tn16:factor(phylum)Caldisericota
## Tn16:factor(phylum)Calditrichota
## Tn16:factor(phylum)Chlamydiota
## Tn16:factor(phylum)Chlorobiota
## Tn16:factor(phylum)Chloroflexota           *
## Tn16:factor(phylum)Chrysiogenota
## Tn16:factor(phylum)Cyanobacteriota
## Tn16:factor(phylum)Deferribacterota
## Tn16:factor(phylum)Deinococcota
## Tn16:factor(phylum)Dictyoglomerota
## Tn16:factor(phylum)Elusimicrobiota
## Tn16:factor(phylum)Fibrobacterota          *
## Tn16:factor(phylum)Fusobacteriota
## Tn16:factor(phylum)Gemmatimonadota
## Tn16:factor(phylum)Kiritimatiellota
## Tn16:factor(phylum)Mycoplasmatota
## Tn16:factor(phylum)Nitrospirota
## Tn16:factor(phylum)Planctomycetota
## Tn16:factor(phylum)Pseudomonadota
## Tn16:factor(phylum)Rhodothermota
## Tn16:factor(phylum)Spirochaetota           .
## Tn16:factor(phylum)Synergistota            .
## Tn16:factor(phylum)Thermodesulfobacteriota
## Tn16:factor(phylum)Thermomicrobiota
## Tn16:factor(phylum)Thermotogota
## Tn16:factor(phylum)Verrucomicrobiota
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7717 on 2775 degrees of freedom
## Multiple R-squared:  0.6208, Adjusted R-squared:  0.6171
## F-statistic: 168.2 on 27 and 2775 DF,  p-value: < 2.2e-16
```

```
res <- Dt %>%
  add_residuals(fitTaxPhylum)
```

```
## Warning in predict.lm(model, data): prediction from a rank-deficient fit may be
## misleading
```

```
res %>%
  group_by(class) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20) %>%
  ggplot(aes(x=Tn16, y=resid)) +
    geom_point(size=0.5, alpha=0.6) +
    theme(legend.position="none") +
    facet_wrap(~class) +
    geom_smooth(method=lm, formula = "y~x+0") +
    geom_ref_line(h=0, col = "red", size = 0.1) +
    ggtitle("Residuals vs n16 By class")
```



The residuals seems ok distributed. It seems to make sense stay at the phylum level just based on this But checking the amount of entries in both it seems that they are about the same. So this effect could be due to each phylum just having one class.

```
print("class:")
```

```
## [1] "class:"
```

```
res %>%
  group_by(class) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20) %>%
  nrow()
```

```
## [1] 2591
```

```
print("phylum:")
```

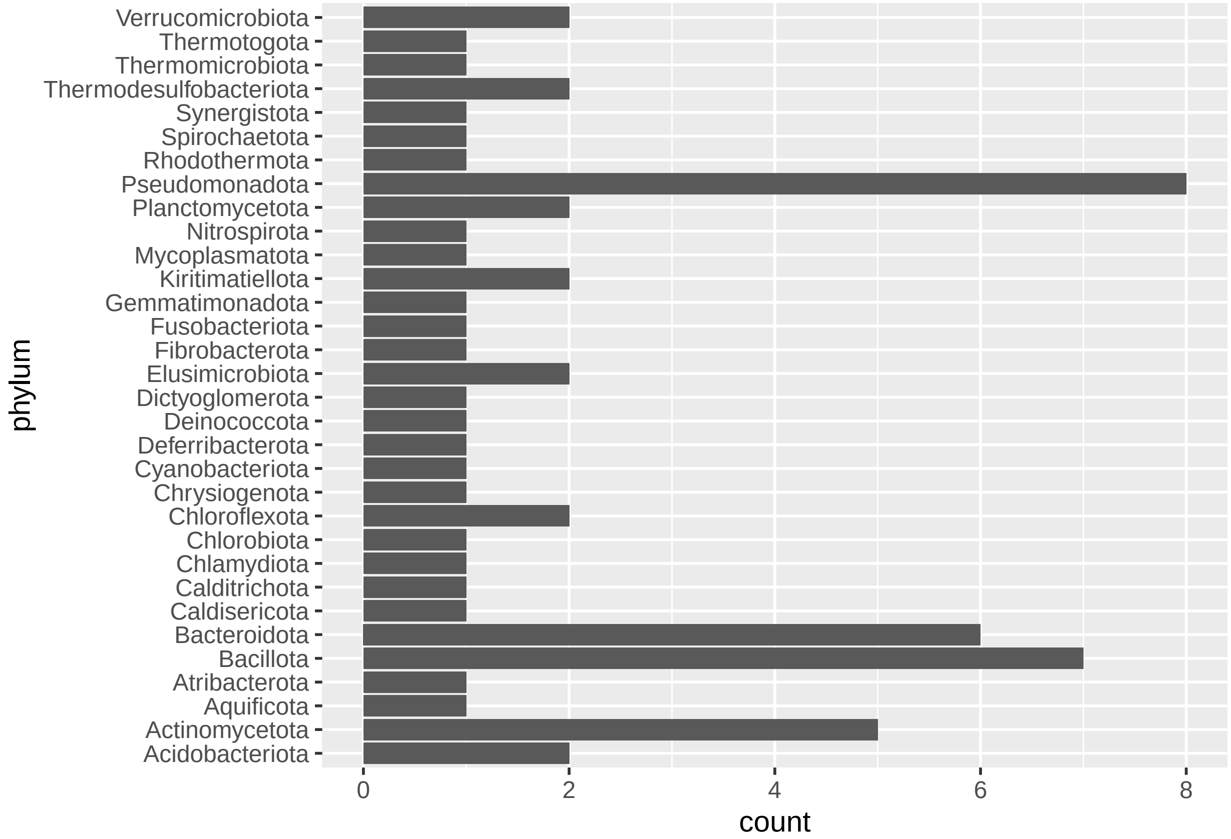
```
## [1] "phylum:"
```

```
res %>%
  group_by(phylum) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20) %>%
  nrow()
```

```
## [1] 2678
```

Lets check

```
res %>%
  group_by(phylum) %>%
  reframe(uClass = unique(class)) %>%
  ggplot() + geom_bar(aes(x=phylum)) +coord_flip()
```



This seems to be the case We could also do the same for order. Here there is more varibilty. But it will argue that we get closer to just predicting the datapoints directly instead of the tendency. Therefore i am going to just keep the model with including the phylum level

```
fitTaxPhylum <- lm(Tdiv ~ 0 + Tn16 + Tn16:phylum ,Dt)
summary(fitTaxPhylum)
```

```
##
## Call:
## lm(formula = Tdiv ~ 0 + Tn16 + Tn16:phylum, data = Dt)
##
## Residuals:
```

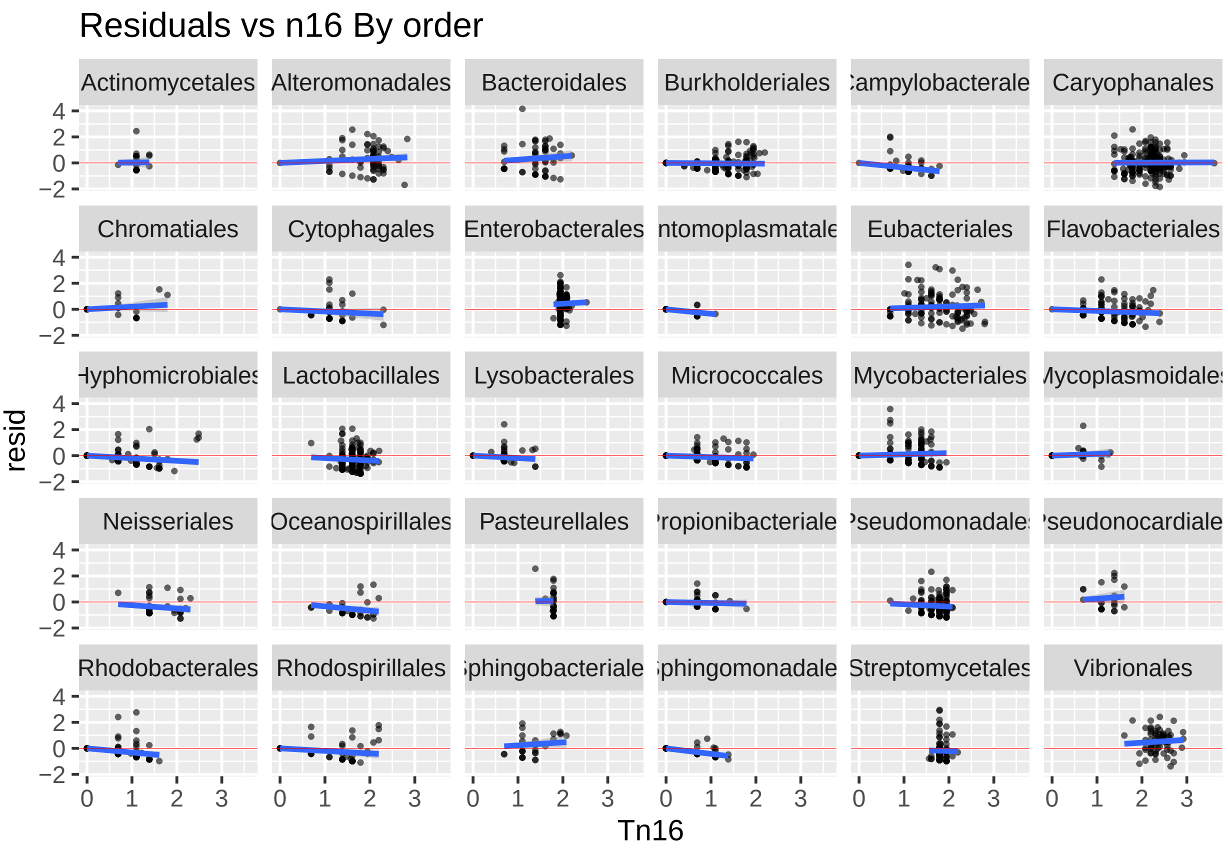
```
##     Min      1Q  Median      3Q     Max 
## -1.8448 -0.5536 -0.1236  0.3196  4.2307 
## 
## Coefficients: (6 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)  
## Tn16                                0.01997    0.52411   0.038   0.9696  
## Tn16:phylumAcidobacteriota               NA         NA      NA       NA  
## Tn16:phylumActinomycetota           0.48564    0.52487   0.925   0.3549  
## Tn16:phylumAquificota               1.04606    0.74121   1.411   0.1583  
## Tn16:phylumAtribacterota            0.73973    1.23051   0.601   0.5478  
## Tn16:phylumBacillota                0.74936    0.52441   1.429   0.1531  
## Tn16:phylumBacteroidota             0.62609    0.52532   1.192   0.2334  
## Tn16:phylumCaldisericota                 NA         NA      NA       NA  
## Tn16:phylumCalditrichota                 NA         NA      NA       NA  
## Tn16:phylumChlamydiota              0.86991    0.68126   1.277   0.2017  
## Tn16:phylumChlorobiota              0.10484    0.64191   0.163   0.8703  
## Tn16:phylumChloroflexota            1.74112    0.79222   2.198   0.0280 *
## Tn16:phylumChrysiogenota            0.84671    0.87641   0.966   0.3341  
## Tn16:phylumCyanobacteriota          1.24310    0.76457   1.626   0.1041  
## Tn16:phylumDeferribacterota         0.11785    0.70679   0.167   0.8676  
## Tn16:phylumDeinococcota             0.54957    0.55091   0.998   0.3186  
## Tn16:phylumDictyoglomerota          0.35988    0.94574   0.381   0.7036  
## Tn16:phylumElusimicrobiota               NA         NA      NA       NA  
## Tn16:phylumFibrobacterota           1.90851    0.87641   2.178   0.0295 *
## Tn16:phylumFusobacteriota           0.78454    0.53492   1.467   0.1426  
## Tn16:phylumGemmatimonadota         -0.01997    1.23051  -0.016   0.9871  
## Tn16:phylumKiritimatiellota        -0.01997    1.23051  -0.016   0.9871  
## Tn16:phylumMycoplasmatota           0.75417    0.54729   1.378   0.1683  
## Tn16:phylumNitrospirota             1.23480    1.23051   1.003   0.3157  
## Tn16:phylumPlanctomycetota          0.42520    0.55739   0.763   0.4456  
## Tn16:phylumPseudomonadota           0.58985    0.52431   1.125   0.2607  
## Tn16:phylumRhodothermota                 NA         NA      NA       NA  
## Tn16:phylumSpirochaetota            0.95724    0.55206   1.734   0.0830 .
## Tn16:phylumSynergistota             1.17997    0.68456   1.724   0.0849 .
## Tn16:phylumThermodesulfobacteriota  1.23480    1.23051   1.003   0.3157  
## Tn16:phylumThermomicrobiota         0.98727    0.94574   1.044   0.2966  
## Tn16:phylumThermotogota             0.52067    0.58231   0.894   0.3713  
## Tn16:phylumVerrucomicrobiota             NA         NA      NA       NA  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.7717 on 2775 degrees of freedom
## Multiple R-squared:  0.6208, Adjusted R-squared:  0.6171 
## F-statistic: 168.2 on 27 and 2775 DF,  p-value: < 2.2e-16
```

```
res <- Dt %>%
  add_residuals(fitTaxPhylum)
```

```
## Warning in predict.lm(model, data): prediction from a rank-deficient fit may be
## misleading
```

```
res %>%
  group_by(order) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20) %>%
  ggplot(aes(x=Tn16, y=resid)) +
    geom_point(size=0.5, alpha=0.6) +
    theme(legend.position="none") +
    facet_wrap(~order) +
    geom_smooth(method=lm, formula = "y~x+0") +
    geom_ref_line(h=0, col = "red", size = 0.1) +
    ggtitle("Residuals vs n16 By order")
```

Tag dem der er store og gå længere ned for at se hvordan dist er ift dem



Looking at tax + n16 model

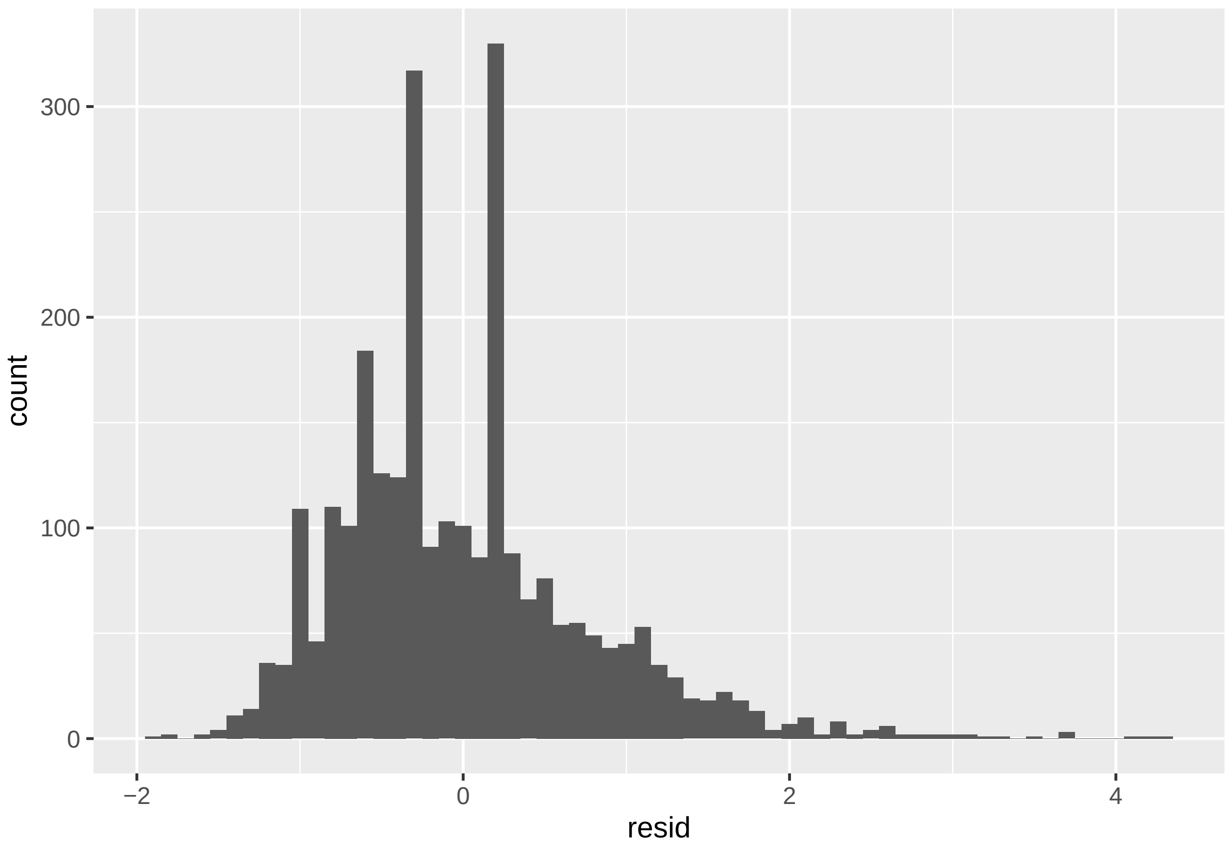
```
Dt <- Dt %>%
  group_by(phylum) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20)
# Lets add it to the model
Dt
```

```
## # A tibble: 2,678 x 108
##    species       antib~1 linco~2 novob~3 kanam~4 ampic~5 genta~6 neomy~7 strep~8
##    <chr>         <chr>   <fct>   <fct>   <fct>   <fct>   <fct>   <fct>   <fct>
##  1 Yersinia pes~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  2 Methylobacte~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  3 Elizabethkin~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  4 Advenella mi~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  5 Corynebacter~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  6 Carnobacteri~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  7 Suicoccus ac~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  8 Rathayibacte~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  9 Syntrophothe~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 10 Zhongshania ~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## # ... with 2,668 more rows, 99 more variables: chloramphenicol <fct>,
## #   rifampicin <fct>, polymyxin.b <fct>, erythromycin <fct>, bacitracin <fct>,
## #   penicillin <fct>, tetracycline <fct>, aztreonam <fct>, cefalotin <fct>,
## #   cefazolin <fct>, cefotaxime <fct>, fosfomycin <fct>, imipenem <fct>,
## #   linezolid <fct>, mezlocillin <fct>, moxifloxacin <fct>,
## #   nitrofurantoin <fct>, norfloxacin <fct>, nystatin <fct>, ofloxacin <fct>,
## #   oxacillin <fct>, penicillin.g <fct>, pipemidic.acid <fct>, ...
```

```
fitTaxPhylum <- lm(Tdiv ~ Tn16 + Tn16:phylum ,Dt)
summary(fitTaxPhylum)
```

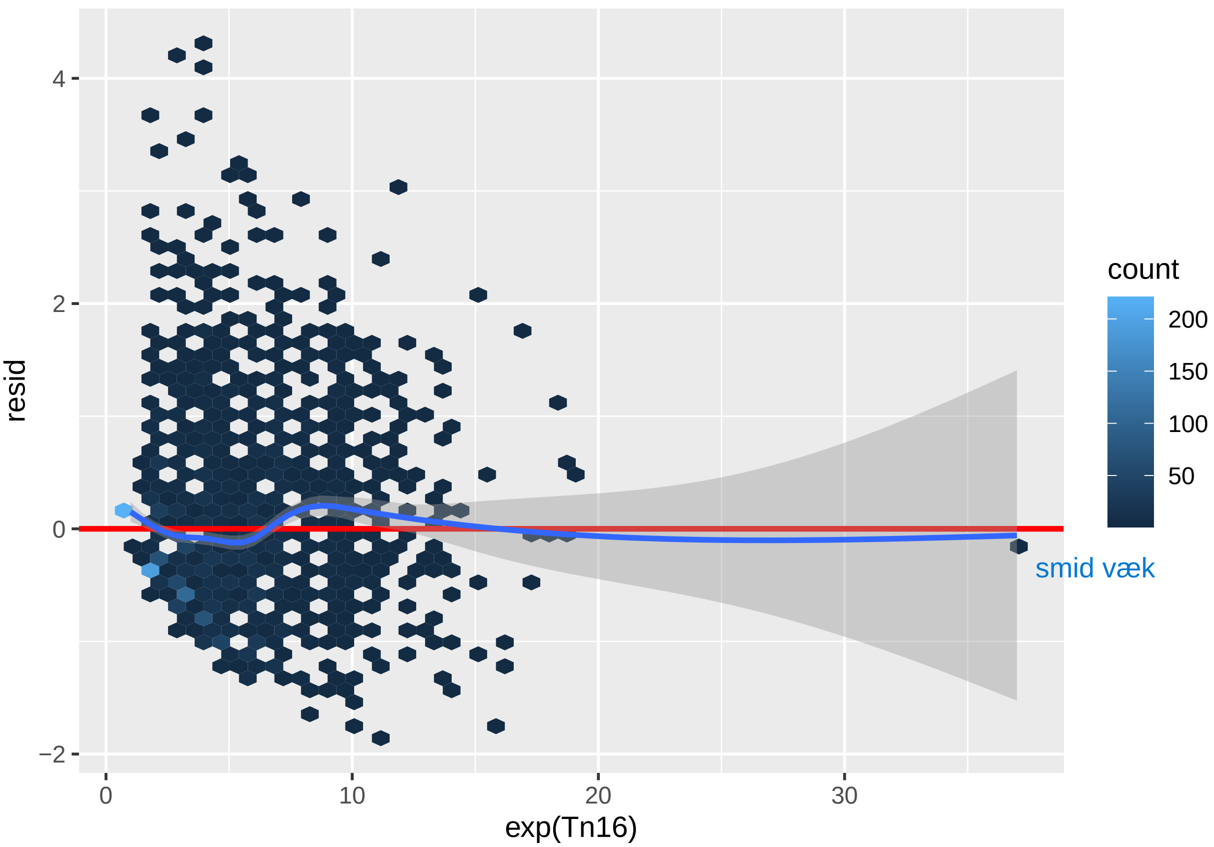
```
## 
## Call:
## lm(formula = Tdiv ~ Tn16 + Tn16:phylum, data = Dt)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.8830 -0.5197 -0.1216  0.3282  4.2583 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                -0.16335    0.03572  -4.574 5.01e-06 ***
## Tn16                        0.62174    0.03794  16.388  < 2e-16 ***
## Tn16:phylumBacillota        0.23166    0.03394   6.826 1.08e-11 ***
## Tn16:phylumBacteroidota     0.13725    0.04547   3.019  0.00256 ** 
## Tn16:phylumMycoplasmatota   0.34380    0.16151   2.129  0.03338 *  
## Tn16:phylumPlanctomycetota -0.02833    0.19262  -0.147  0.88307    
## Tn16:phylumPseudomonadota   0.08601    0.03197   2.690  0.00718 ** 
## Tn16:phylumSpirochaetota    0.51157    0.17655   2.898  0.00379 ** 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.7746 on 2670 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3138 
## F-statistic: 175.9 on 7 and 2670 DF,  p-value: < 2.2e-16
```

```
res <- Dt %>%
  add_residuals(fitTaxPhylum)

res %>%
  ggplot(aes(resid)) + geom_histogram(bins = 40,binwidth = 0.1)
```



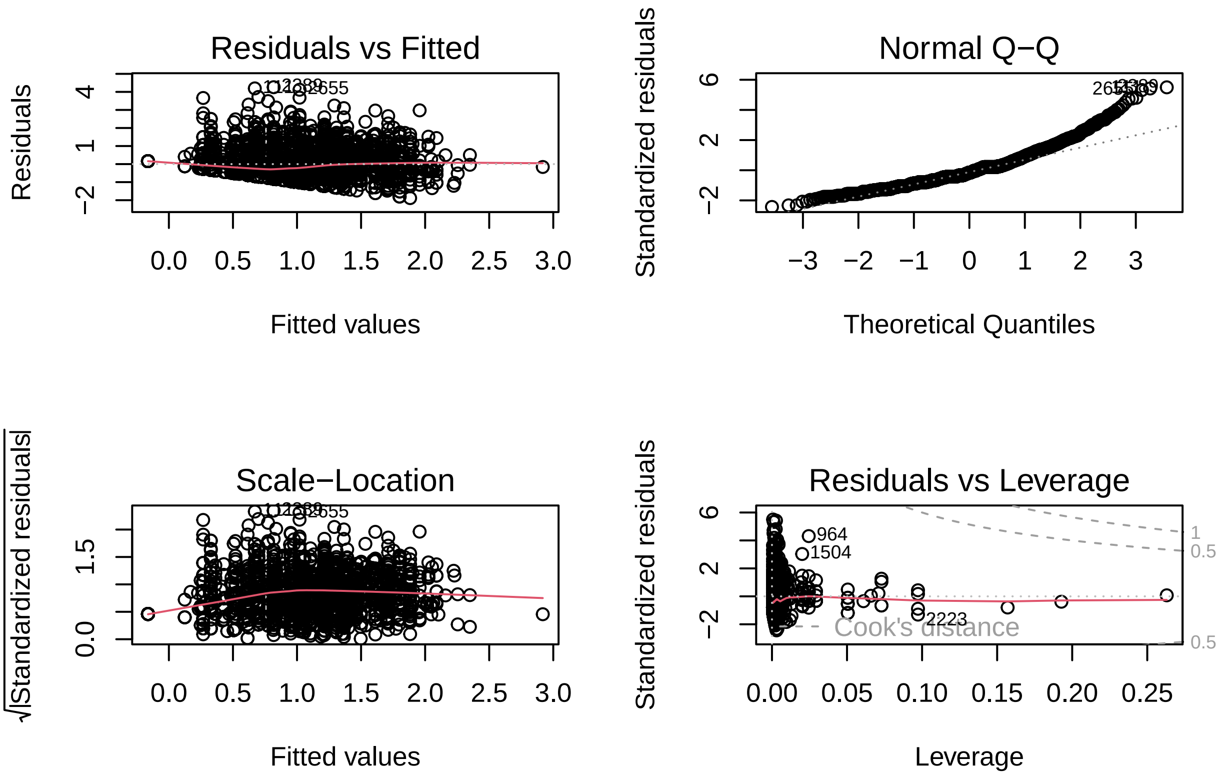
```
res %>%
  ggplot(aes(x=exp(Tn16) ,y=resid)) +
  geom_hex(bins=50) +
  geom_ref_line(h=0, col = "red", size = 1) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



smid væk

```
par(mfrow=c(2,2))
plot(fitTaxPhylum )
```



```
# We have on with a very large amount of n16
#Dsub %>% filter(n16>20)

#I checked and its also high here
#https://www.arb-silva.de/search/
#Tumebacillus avium
# we observe more var at the start in the res since they are predicting wrong
```

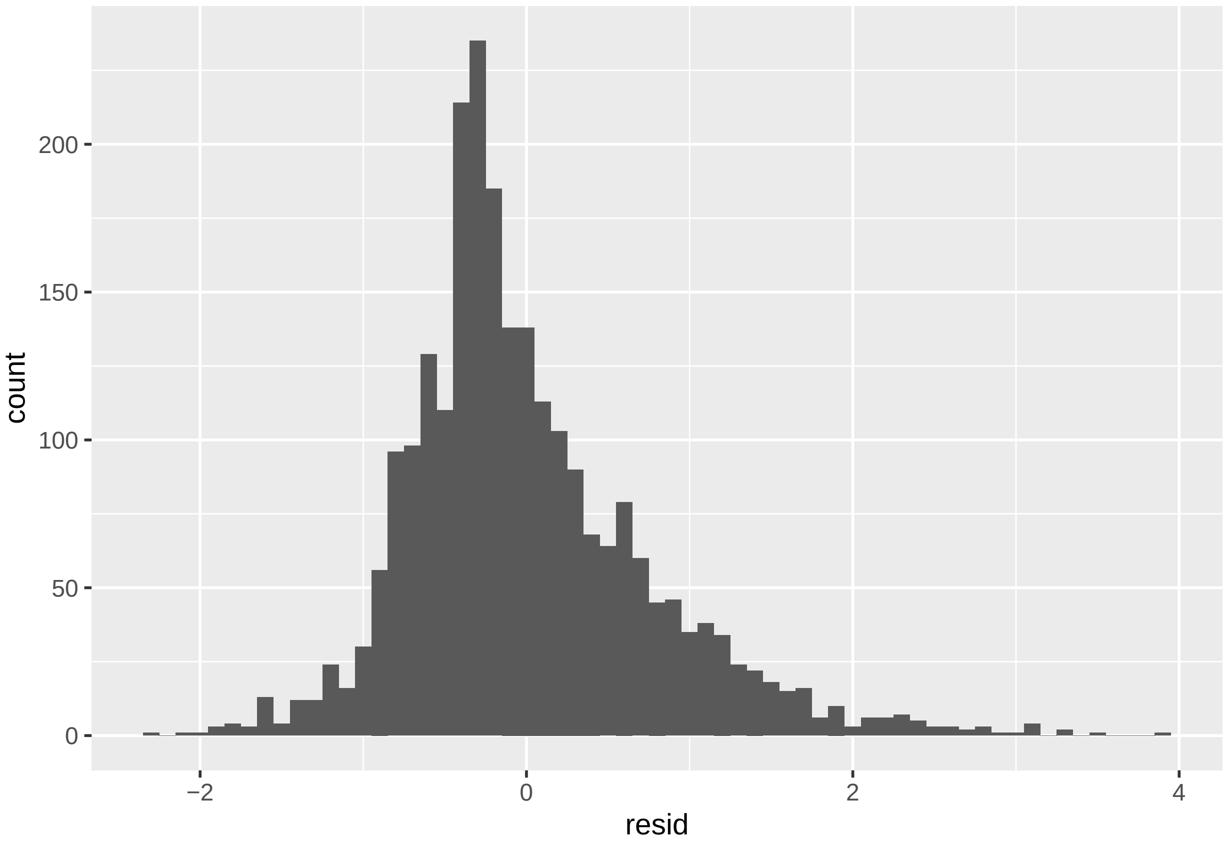
We can observe that we tend to overestimate the div on genera with larger amount of #16s. And we tend to underestimate div for genera with samller amounts of #16s. Therefore we still have some unexplained variance in the model

Lets also have a look for order

```
Dt <- Dt %>%
  filter(n16 > 1)
# Lets add it to the model
Dt
```

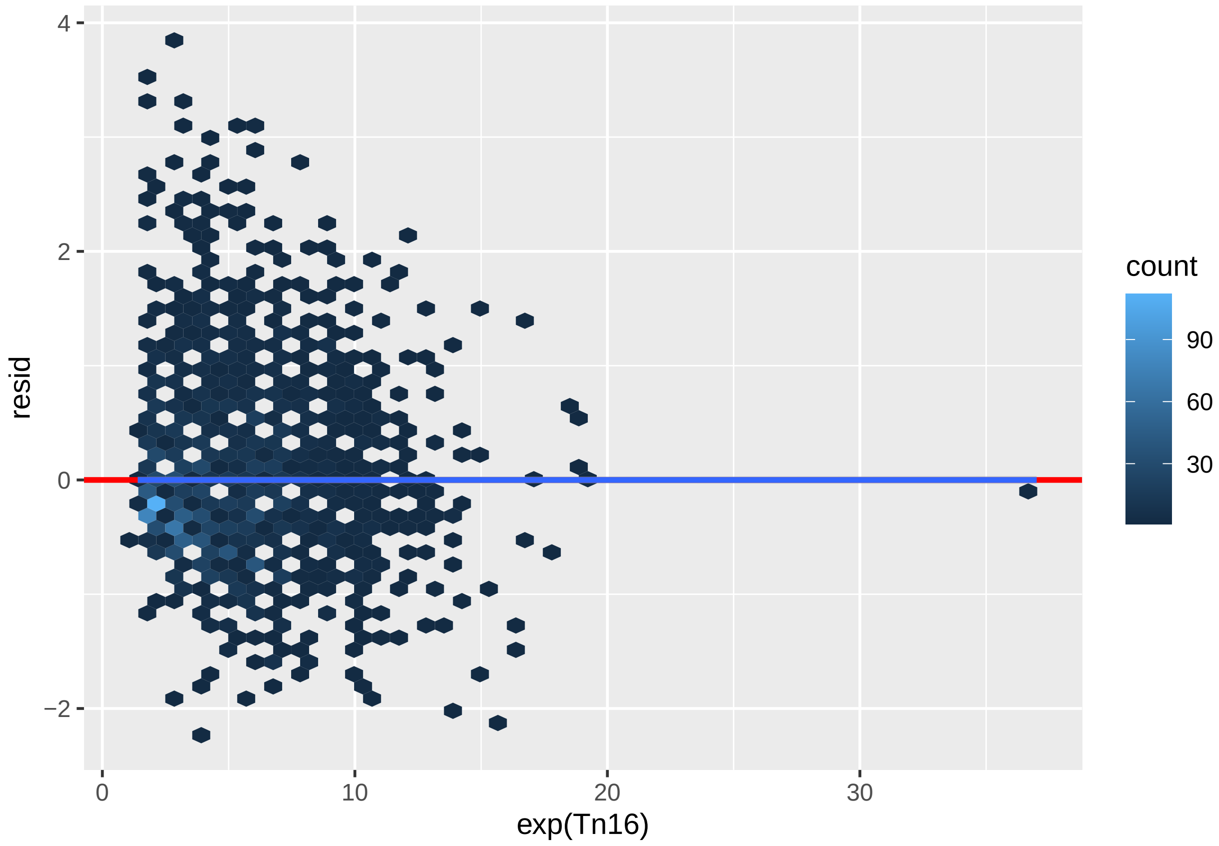
```
## # A tibble: 2,457 x 108
##    species        antib~1 linco~2 novob~3 kanam~4 ampic~5 genta~6 neomy~7 strep~8
##    <chr>          <chr>   <fct>   <fct>   <fct>   <fct>   <fct>   <fct>   <fct>
##  1 Yersinia pes~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  2 Methylobacte~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  3 Elizabethkin~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  4 Advenella mi~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  5 Corynebacter~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  6 Carnobacteri~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  7 Suicoccus ac~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  8 Rathayibacte~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
##  9 Syntrophothe~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 10 Zhongshania ~ PNR     <NA>    <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## # ... with 2,447 more rows, 99 more variables: chloramphenicol <fct>,
## #   rifampicin <fct>, polymyxin.b <fct>, erythromycin <fct>, bacitracin <fct>,
## #   penicillin <fct>, tetracycline <fct>, aztreonam <fct>, cefalotin <fct>,
## #   cefazolin <fct>, cefotaxime <fct>, fosfomycin <fct>, imipenem <fct>,
## #   linezolid <fct>, mezlocillin <fct>, moxifloxacin <fct>,
## #   nitrofurantoin <fct>, norfloxacin <fct>, nystatin <fct>, ofloxacin <fct>,
## #   oxacillin <fct>, penicillin.g <fct>, pipemidic.acid <fct>, ...
```

```
fitTaxOrder <- lm(Tdiv ~ Tn16 + Tn16:order ,Dt)
res <- Dt %>%
  add_residuals(fitTaxOrder)

res %>%
  ggplot(aes(resid)) + geom_histogram(bins = 40,binwidth = 0.1)
```

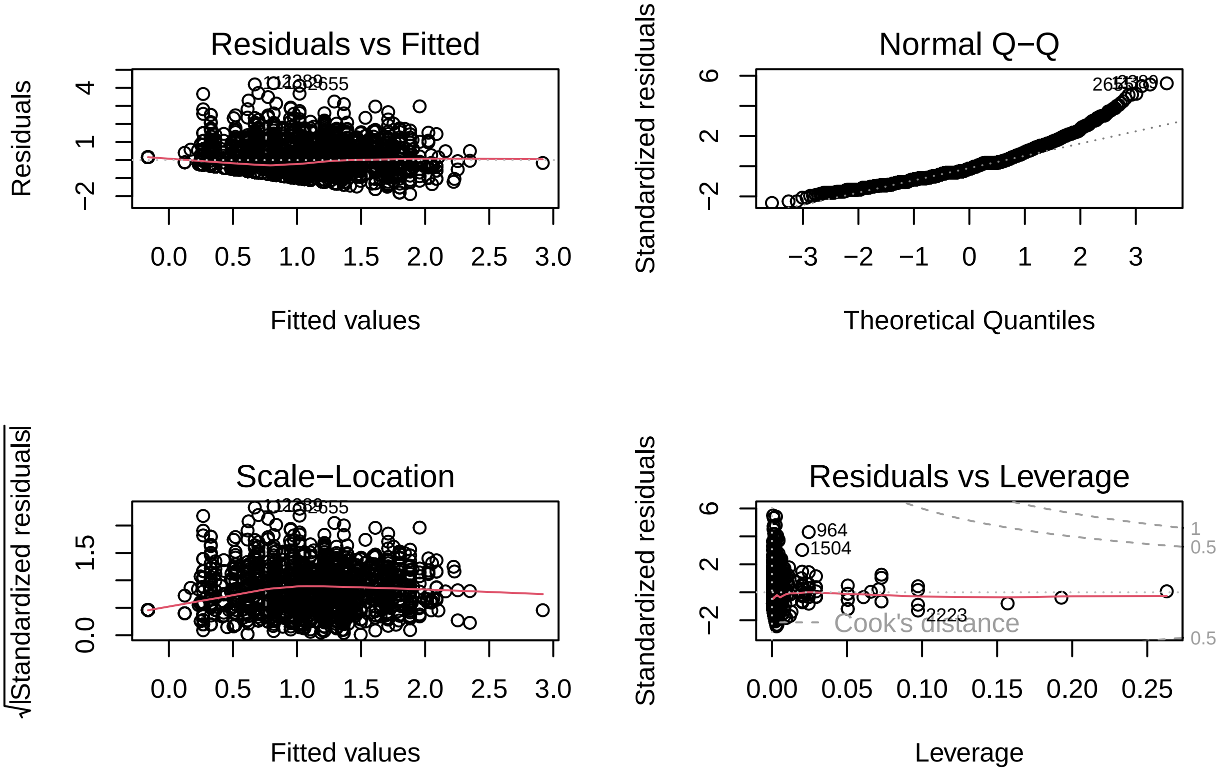


```
res %>%
  ggplot(aes(x=exp(Tn16) ,y=resid)) +
  geom_hex(bins=50) +
  geom_ref_line(h=0, col = "red", size = 1) +
  geom_smooth()
```

```
## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
par(mfrow=c(2,2))
plot(fitTaxPhylum )
```



While some of the plots look ok, we can see that most of the orders have few entries (~half having below 10 entries)

```
# Under 10
print("under 10")
```

```
## [1] "under 10"
```

```
Dt %>%
  group_by(order) %>%
  summarise(n = n()) %>%
  filter(n < 10) %>%
  nrow
```

```
## [1] 46
```

```
# Over or equal to 10
print("over or equal to 10")
```

```
## [1] "over or equal to 10"
```

```
Dt %>%
  group_by(order) %>%
  summarise(n = n()) %>%
  filter(n >= 10) %>%
  nrow
```

```
## [1] 47
```

SAMPLING

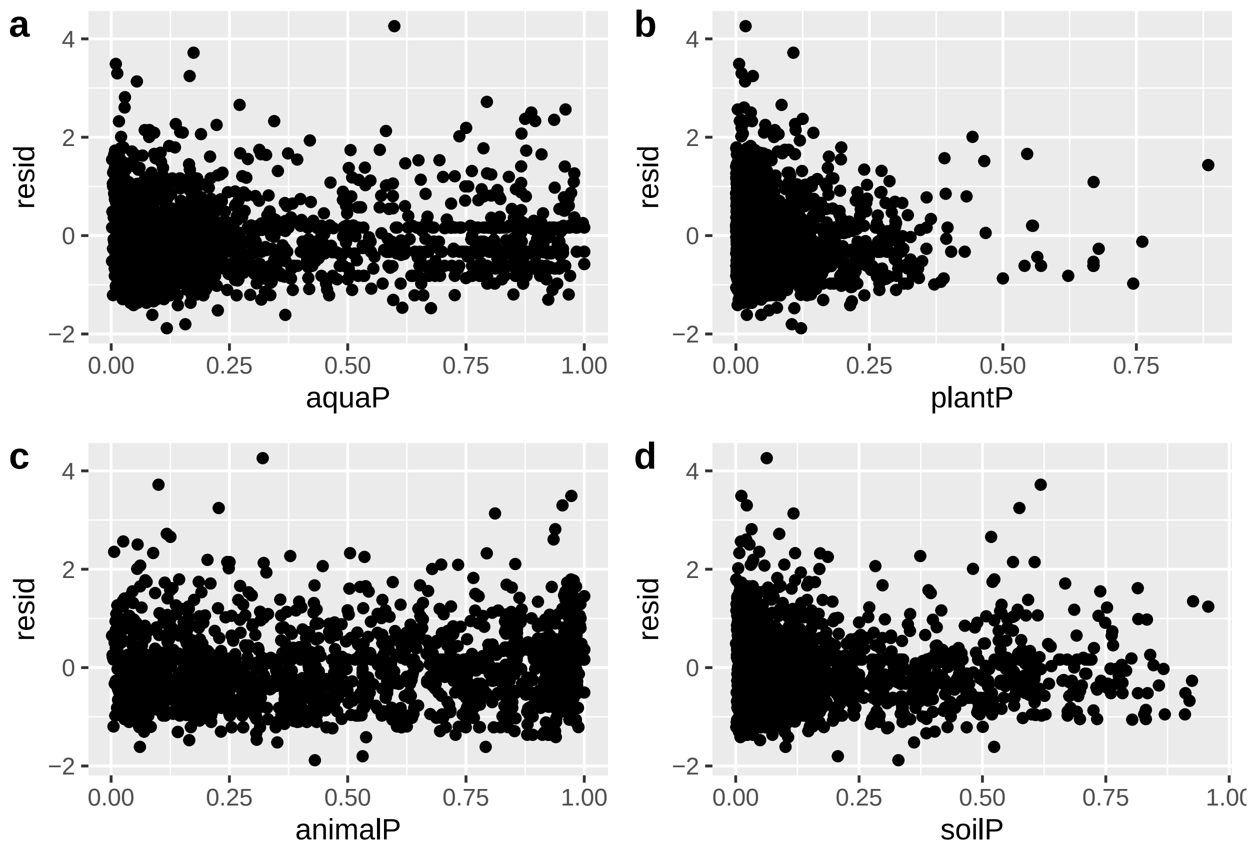
Lets try and get an idea about the effect of where it is samples from

```
Dt <- D %>%
  mutate(Tn16=log(n16), Tdiv=log1p(div))
Dt <- Dt %>%
  group_by(phylum) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n > 20)

Denv <- Dt %>% mutate(aquaP = aquatic.counts/Total.samples ,
              animalP = animal.counts/Total.samples,
              plantP = plant.counts/Total.samples,
              soilP = soil.counts/Total.samples)

fitTaxPhylum <- lm(Tdiv ~ Tn16 + Tn16:phylum ,Denv)
res <- Denv %>%
  add_residuals(fitTaxPhylum)
summary(fitTaxPhylum)
```

```
##
## Call:
## lm(formula = Tdiv ~ Tn16 + Tn16:phylum, data = Denv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8830 -0.5197 -0.1216  0.3282  4.2583
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.16335    0.03572  -4.574 5.01e-06 ***
## Tn16                          0.62174    0.03794  16.388  < 2e-16 ***
## Tn16:phylumBacillota          0.23166    0.03394   6.826 1.08e-11 ***
## Tn16:phylumBacteroidota       0.13725    0.04547   3.019  0.00256 **
## Tn16:phylumMycoplasmatota     0.34380    0.16151   2.129  0.03338 *
## Tn16:phylumPlanctomycetota -0.02833    0.19262  -0.147  0.88307
## Tn16:phylumPseudomonadota     0.08601    0.03197   2.690  0.00718 **
## Tn16:phylumSpirochaetota      0.51157    0.17655   2.898  0.00379 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7746 on 2670 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3138
## F-statistic: 175.9 on 7 and 2670 DF,  p-value: < 2.2e-16
```

```
p1 <- ggplot(res, aes(x=aquaP, y=resid)) +
  geom_point()
p2 <- ggplot(res, aes(x=plantP, y=resid)) +
  geom_point()
p3 <- ggplot(res, aes(x=animalP, y=resid)) +
  geom_point()
p4 <- ggplot(res, aes(x=soilP, y=resid)) +
  geom_point()
plot_grid(p1, p2, p3, p4,labels ="auto")
```

```
## Warning: Removed 951 rows containing missing values (‘geom_point()‘).

## Warning: Removed 1028 rows containing missing values (‘geom_point()‘).

## Warning: Removed 948 rows containing missing values (‘geom_point()‘).

## Warning: Removed 975 rows containing missing values (‘geom_point()‘).
```



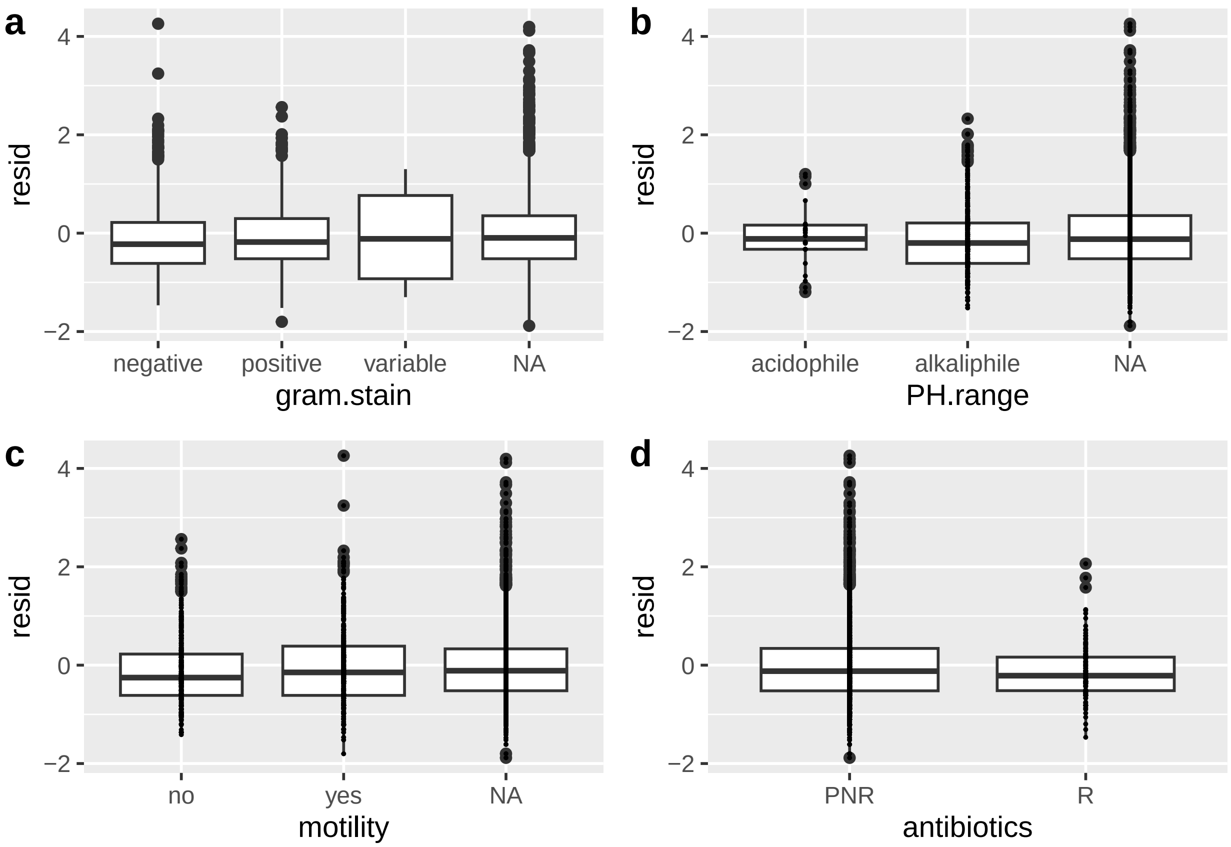
It seems that there is no difference here

Antibiotics, motility, PH, gramstain motility

```
res <- Dt %>%
  add_residuals(fitTaxPhylum)
summary(fitTaxPhylum)
```

```
##
## Call:
## lm(formula = Tdiv ~ Tn16 + Tn16:phylum, data = Denv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8830 -0.5197 -0.1216  0.3282  4.2583
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.16335    0.03572  -4.574 5.01e-06 ***
## Tn16                          0.62174    0.03794  16.388  < 2e-16 ***
## Tn16:phylumBacillota          0.23166    0.03394   6.826 1.08e-11 ***
## Tn16:phylumBacteroidota       0.13725    0.04547   3.019  0.00256 **
## Tn16:phylumMycoplasmatota     0.34380    0.16151   2.129  0.03338 *
## Tn16:phylumPlanctomycetota -0.02833    0.19262  -0.147  0.88307
## Tn16:phylumPseudomonadota     0.08601    0.03197   2.690  0.00718 **
## Tn16:phylumSpirochaetota      0.51157    0.17655   2.898  0.00379 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7746 on 2670 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3138
## F-statistic: 175.9 on 7 and 2670 DF,  p-value: < 2.2e-16
```

```r
p1 <- ggplot(res, aes(x=gram.stain, y=resid)) +
  geom_boxplot()
p2 <- ggplot(res, aes(x=PH.range, y=resid)) +
  geom_boxplot() +
  geom_point(size=0.2)
p3 <- ggplot(res, aes(x=motility, y=resid)) +
  geom_boxplot()+
  geom_point(size=0.2)
p4 <- ggplot(res, aes(x=antibiotics, y=resid)) +
  geom_boxplot() +
  geom_point(size=0.2)
plot_grid(p1, p2, p3, p4,labels ="auto")
```



Lets test antibiotics

```
library(car)
```

```
## Indlæser krævet pakke: carData
```

```
##
## Vedhæfter pakke: 'car'
```

```
## Det følgende objekt er maskeret fra 'package:dplyr':
##
##     recode
```

```
## Det følgende objekt er maskeret fra 'package:purrr':
##
##     some
```

```
# Updating model and running ancova on it
fit_ar <- update(fitTaxPhylum, . ~ . + factor(antibiotics) + Tn16:factor(antibiotics))
Anova(fit_ar)
```

```
## Anova Table (Type II tests)
##
## Response: Tdiv
##                              Sum Sq   Df   F value    Pr(>F)
## Tn16                      696.27    1 1160.9930 < 2.2e-16 ***
## factor(antibiotics)         1.79    1    2.9922   0.08378 .
## Tn16:phylum                42.58    6   11.8334 3.855e-13 ***
## Tn16:factor(antibiotics)    0.10    1    0.1742   0.67647
## Residuals                1600.06 2668
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
drop1(fit_ar)
```

```
## Single term deletions
##
## Model:
## Tdiv ~ Tn16 + factor(antibiotics) + Tn16:phylum + Tn16:factor(antibiotics)
##                          Df Sum of Sq    RSS     AIC
## <none>                                1600.1 -1359.2
## Tn16:phylum               6    42.581 1642.6 -1300.9
## Tn16:factor(antibiotics)  1     0.104 1600.2 -1361.1
```

```
fit_ar2 <- update(fit_ar, .~. -Tn16:factor(antibiotics))
Anova(fit_ar2)
```

```
## Anova Table (Type II tests)
##
## Response: Tdiv
##                      Sum Sq   Df   F value    Pr(>F)
## Tn16                 696.27    1 1161.3524 < 2.2e-16 ***
## factor(antibiotics)    1.79    1    2.9932   0.08373 .
## Tn16:phylum           42.49    6   11.8109 4.102e-13 ***
## Residuals           1600.17 2669
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

It seems there is no significant effect of antibiotics

Lets have a look at specific antibiotics

Formatting daTA

Lets check for more specifc types of AR First getting the subset of the data with AR resistence info about the Antibiotics which target the 16s rRNA

```
# Getting the ones which are actually targeting 16S
# Reading them from ARtarget16s.csv
target16S <- read_csv2("../data/ARtarget16s.csv",show_col_types = FALSE,col_names = FALSE)
```

```
## i Using "’,’" as decimal and "’.’" as grouping mark. Use ‘read_delim()‘ for more control.
```

```
targetvector <- as.array(target16S$X1)
found_16S <- as.array(colnames(select(D_tmp,lincomycin:spiramycin.II)))
intersect <- intersect(targetvector,found_16S)
D_ar <- select(Dt, all_of(intersect), Tn16, Tdiv, phylum)
```

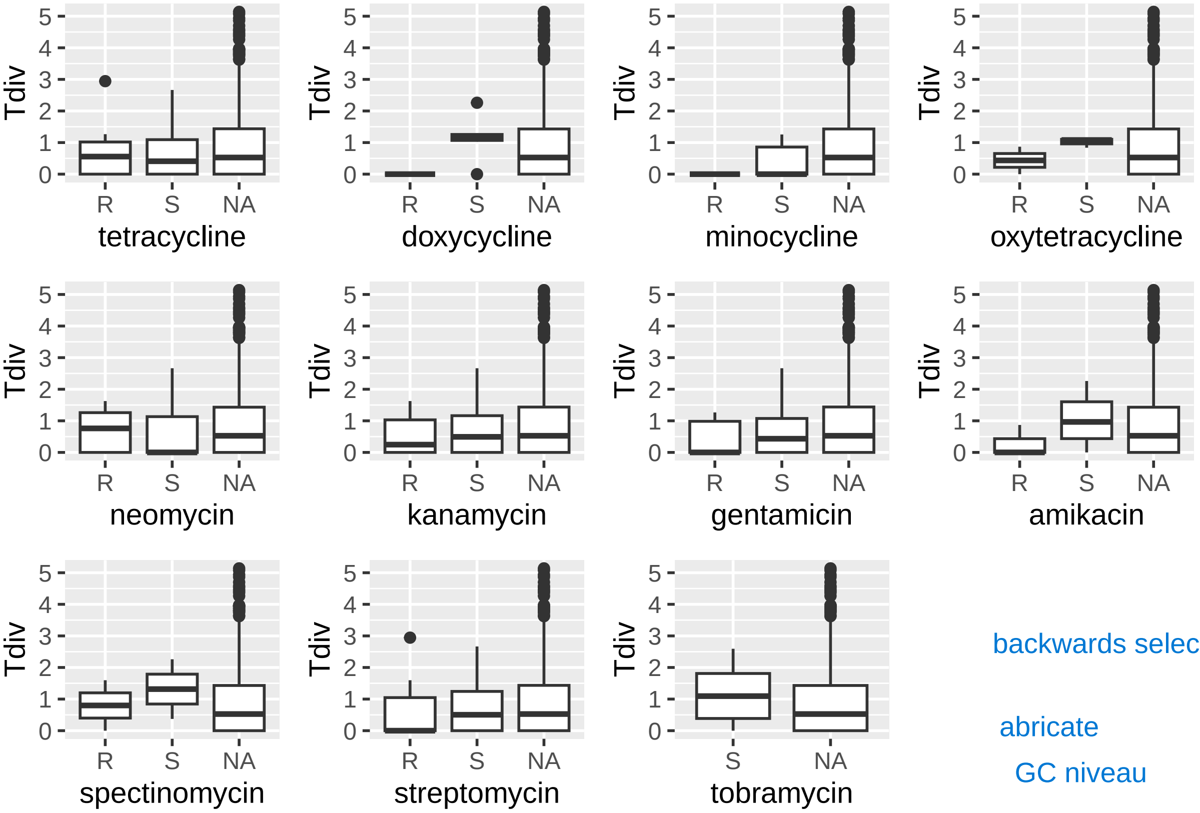
Different types

Div Now lets look at some plots firstly for div

```
library(patchwork)
```

```
##
## Vedhæfter pakke: 'patchwork'

## Det følgende objekt er maskeret fra 'package:cowplot':
##
##     align_plots
```

```
plotlist = list()
for(i in seq_along(intersect)){
  antibiotic = intersect[i]
  p <- ggplot(D_ar)+
    geom_boxplot(aes(x=.data[[antibiotic]], y=Tdiv))
  plotlist = c(plotlist, list(p))
}

wrap_plots(plotlist)
```



backwards selection

abricate

GC niveau

beeswarm

random forrest

Lets test it

```
update(fitTaxPhylum, . ~ . + minocycline + Tn16:minocycline) %>%
  Anova()
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Anova Table (Type II tests)
##
## Response: Tdiv
##                   Sum Sq Df F value  Pr(>F)
## Tn16             0.28053  1  1.5584 0.27997
## minocycline      0.05356  1  0.2975 0.61443
## Tn16:phylum      0.87904  1  4.8834 0.09164 .
## Tn16:minocycline          0
## Residuals        0.72002  4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
filter(D_ar, !is.na(minocycline ))
```

```
## # A tibble: 9 x 14
##   tetracycline doxycyc~1 minoc~2 oxyte~3 neomy~4 kanam~5 genta~6 amika~7 spect~8
##   <fct>        <fct>     <fct>   <fct>   <fct>   <fct>   <fct>   <fct>   <fct>
## 1 S            S         S       <NA>    S       S       S       <NA>    <NA>
## 2 <NA>         <NA>      R       <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 3 R            S         S       <NA>    R       R       R       <NA>    <NA>
## 4 S            <NA>      S       <NA>    S       S       S       <NA>    <NA>
## 5 S            <NA>      S       <NA>    S       S       S       <NA>    <NA>
## 6 R            R         S       <NA>    S       S       S       S       <NA>
## 7 S            S         S       <NA>    S       S       S       S       <NA>
## 8 <NA>         <NA>      R       <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## 9 <NA>         <NA>      R       <NA>    <NA>    <NA>    <NA>    <NA>    <NA>
## # ... with 5 more variables: streptomycin <fct>, tobramycin <fct>, Tn16 <dbl>,
## #   Tdiv <dbl>, phylum <fct>, and abbreviated variable names 1: doxycycline,
## #   2: minocycline, 3: oxytetracycline, 4: neomycin, 5: kanamycin,
## #   6: gentamicin, 7: amikacin, 8: spectinomycin
```

```
update(fitTaxPhylum, . ~ . + streptomycin + Tn16:streptomycin) %>%
  Anova()
```

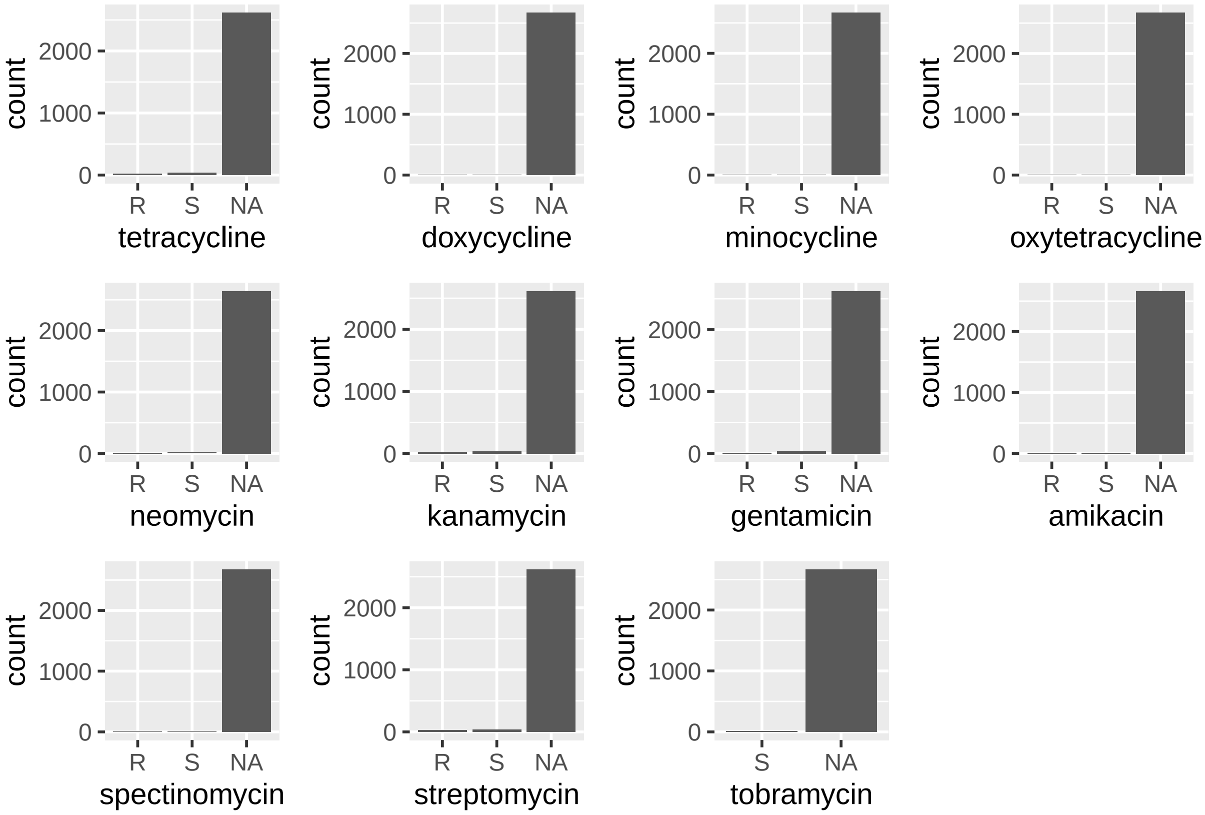
```
## Anova Table (Type II tests)
##
## Response: Tdiv
##                    Sum Sq Df F value    Pr(>F)
## Tn16              11.3299  1 31.1849 8.231e-07 ***
## streptomycin       0.1767  1  0.4864  0.488590
## Tn16:phylum        5.4189  4  3.7288  0.009548 **
## Tn16:streptomycin  0.0913  1  0.2513  0.618267
## Residuals         19.2556 53
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#update(fitTaxPhylum, . ~ . + tobramycin + Tn16:tobramycin) %>%
#  Anova()
```

```
plotlist = list()
for(i in seq_along(intersect)){
  antibiotic = intersect[i]
  p <- ggplot(D_ar)+
    geom_bar(aes(x=.data[[antibiotic]],na.rm = TRUE))
  plotlist = c(plotlist, list(p))
}
```

```
## Warning in geom_bar(aes(x = .data[[antibiotic]], na.rm = TRUE)): Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
## Ignoring unknown aesthetics: na.rm
```

```
wrap_plots(plotlist)
```



Lets have a look at interactions