Chapter 11 Correlation and Regression Analyses

Angelo LaCommare-Soto

2025-04-16

Functional Associations in the Genome

Research Question 1: Do large genomes have higher number of genes encoding DNA polymerase III alpha subunit?

Section 1 - Importing Data

```
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```
library(tidyverse)
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# importing the gene dataset
gene <- read.table("genomics.txt",header=TRUE,sep="\t",stringsAsFactors = TRUE)
str(gene)
```

```
## 'data.frame':    1351 obs. of  7 variables:
##  $ Strain          : Factor w/ 1321 levels "[Brevibacterium]_flavum_ZL-1",..: 1230 980 1141 546 232 9
##  $ N_Gene_tot      : int  7832 5786 3224 2671 1625 2207 6239 1631 5051 2831 ...
##  $ DNApol_III_a_sub: int  2 2 1 1 1 1 2 1 1 1 ...
##  $ tRNA_synthetase : int  27 24 21 36 20 21 22 20 24 21 ...
##  $ X16rRNA         : int  6 2 3 0 1 1 5 2 2 8 ...
##  $ cellulase       : int  8 3 3 0 0 6 8 1 3 2 ...
##  $ Blactamase      : int  13 9 7 0 1 4 8 4 2 5 ...
```

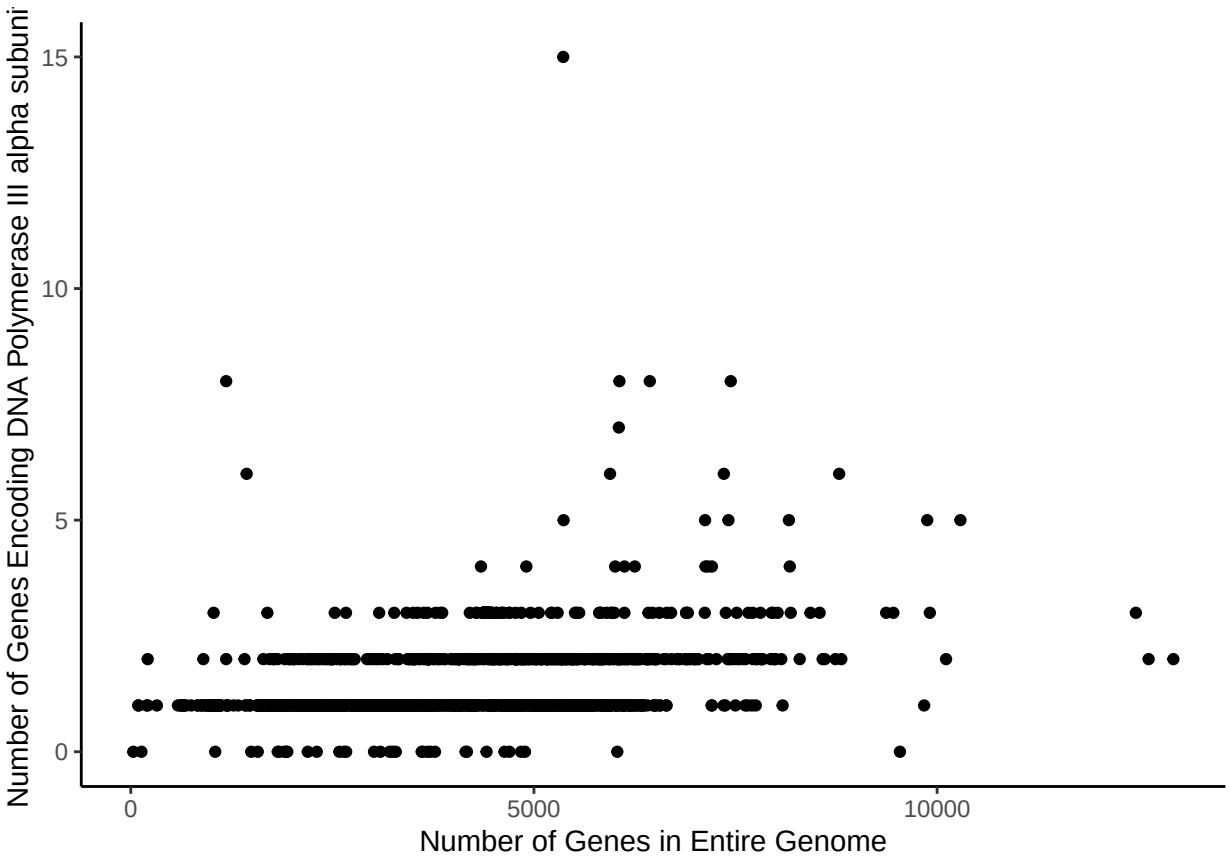
Question Answers

a. Given the question 'Do large genomes have higher number of genes encoding DNA polymerase III alpha subunit?', we are interested in the 'N_Gene_tot' and 'DNApol_III_a_sub' variables.
b. Both the 'N_Gene_tot' and 'DNApol_III_a_sub' variables are integer numeric variables, and their relationship can be shown by a scatter plot with a line of best fit.

Section 2 - Plotting the Data

Challenge 1: Plot the Data for Appropriate Visualization

```
p1<-ggplot(gene, aes(x=N_Gene_tot, y=DNApol_III_a_sub))+
  geom_point()+
  ylab("Number of Genes Encoding DNA Polymerase III alpha subunit")+
  xlab("Number of Genes in Entire Genome")+
  theme_classic()
p1
```



Question Answers

a. The data demonstrates a pattern of wide variation among genomes below about 8000 genes in size, as well as a slight shift to the right in the density of data points as genes in the genome increase.
b. I predict that there will be a very slight positive correlation between the number of genes in an entire genome and the number of genes that encode for DNA polymerase III alpha subunit.

Section 3 - Estimating the Correlation Coefficient

```
# Pearson correlation

library(rstatix)
```

```
## Warning: package 'rstatix' was built under R version 4.4.3
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
gene1_cor <- cor_test(DNApol_III_a_sub, N_Gene_tot, data=gene)

#correlation coefficient
r <- gene1_cor$cor

r
```

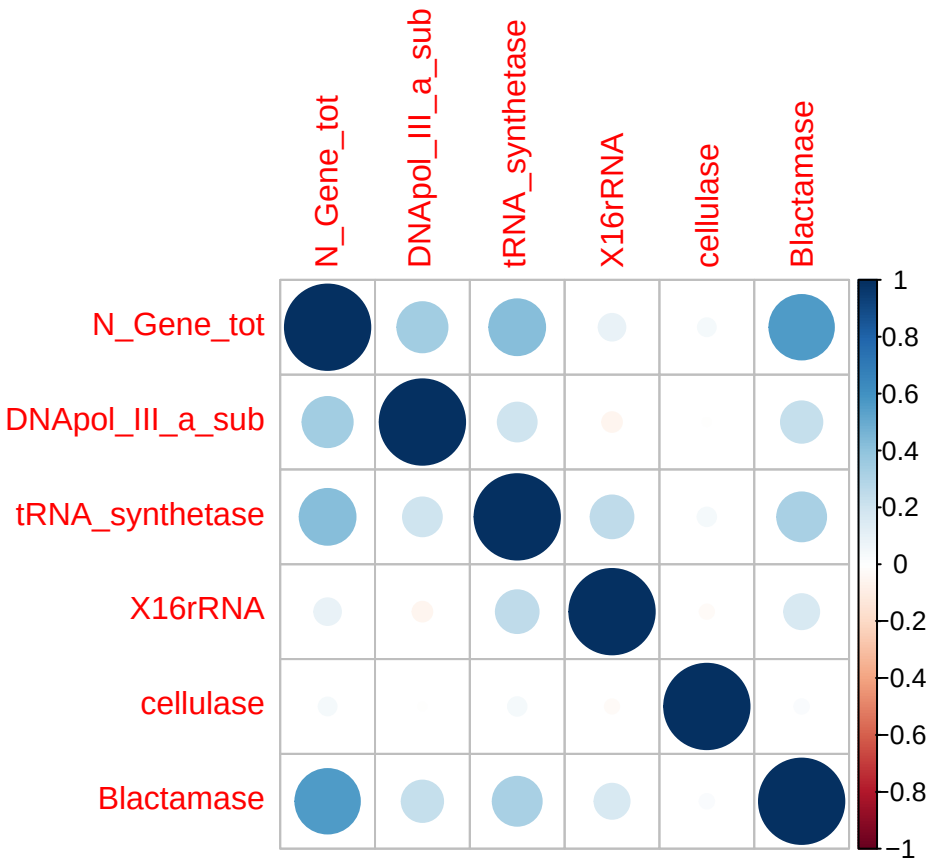
```
##  cor
## 0.34
```

```
# We can also visualize the correlation coefficient for all the possible combinations of variables. The

# converting gene into a matrix
gene2 <- as.matrix(subset(gene,select=-c(Strain)))

# correlation coefficient
gene3 <- cor(gene2,method="pearson")

# installing corrplot
# previously installed

# loading package
library(corrplot)
```

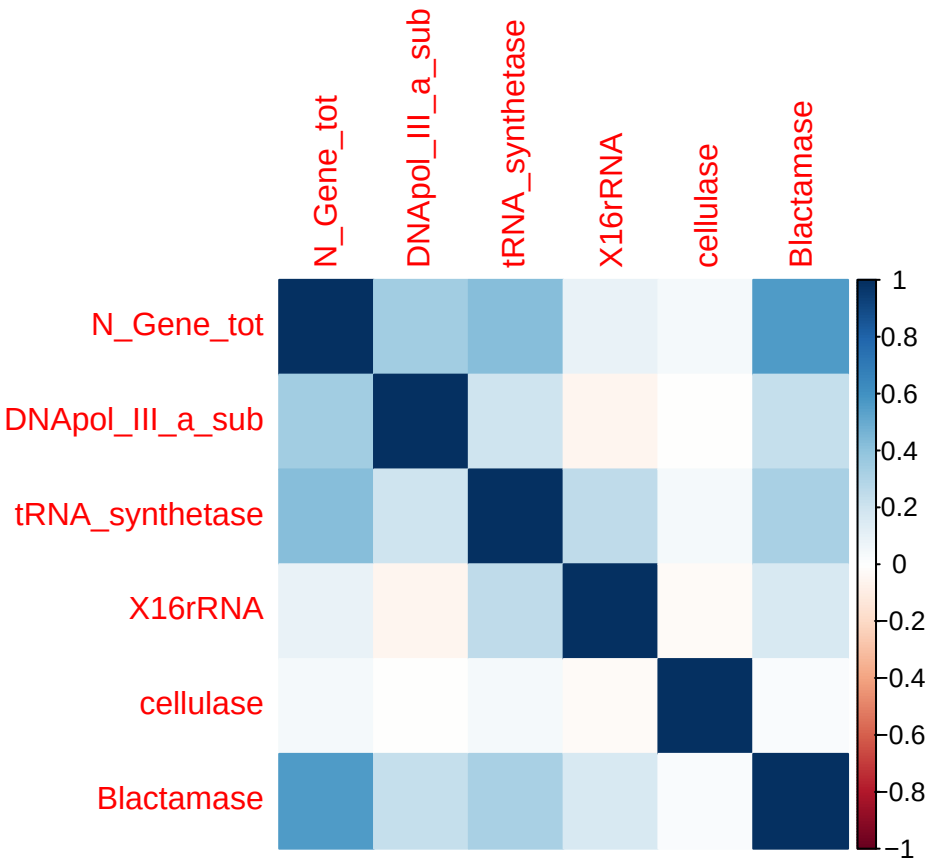
```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

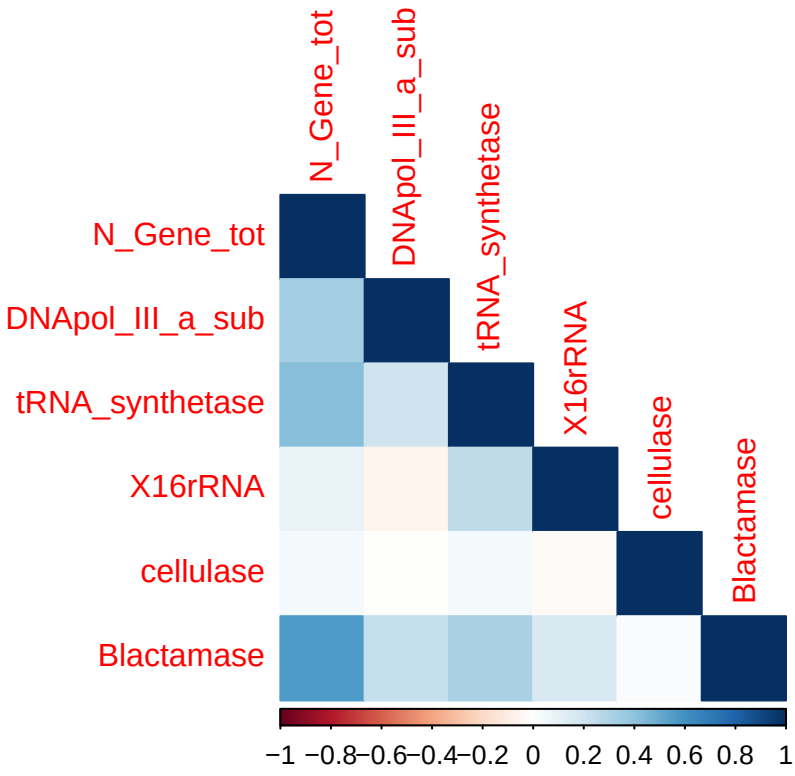
```
# "corrplot" with different visual methods
corrplot(gene3, method = "circle")
```




```
corrplot(gene3, method = "color")
```




```
corrplot(gene3, method = "color",type="lower")
```



Section 4 - Estimating the Standard Error of the Correlation Coefficient (r)

```
# sample size
n <- 1351
n
```

```
## [1] 1351
```

```
# standard error of r
r_se <- sqrt((1-r^2)/(n-2))
r_se
```

```
##         cor
## 0.02560462
```

Section 5 - Testing the Hypothesis Using the t-test

```
# t-test for correlation analysis
gene1_cor
```

```
## # A tibble: 1 x 8
##   var1             var2         cor statistic        p conf.low conf.high method
##   <chr>            <chr>      <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
## 1 DNApol_III_a_sub N_Gene_tot  0.34      13.5 5.72e-39    0.297     0.391 Pears~
```

```
# manually calculating t-statistic
t <- r/r_se
t
```

```
##      cor
## 13.27885
```

Question Answer

a. Using a statistical table for the t-distribution (statsexamples.com), we can see that the null expectation for a critical value of 0.025 is t=1.962 with 1000 degrees of freedom. As our estimated t-statistic is much more extreme at 13.48, p is less than 0.05 and we reject the null hypothesis. Additionally, the estimated p-value in the correlation test is infinitesimally small at 5.72*10^-39.

Section 6 - Estimating the Regression

```
# We can also fit a linear regression to test whether the number of genes encoding DNA polymerase III a

# linear regression
lm1 <- lm(DNApol_III_a_sub~N_Gene_tot, data=gene)

# summary of the model output
summary(lm1)
```

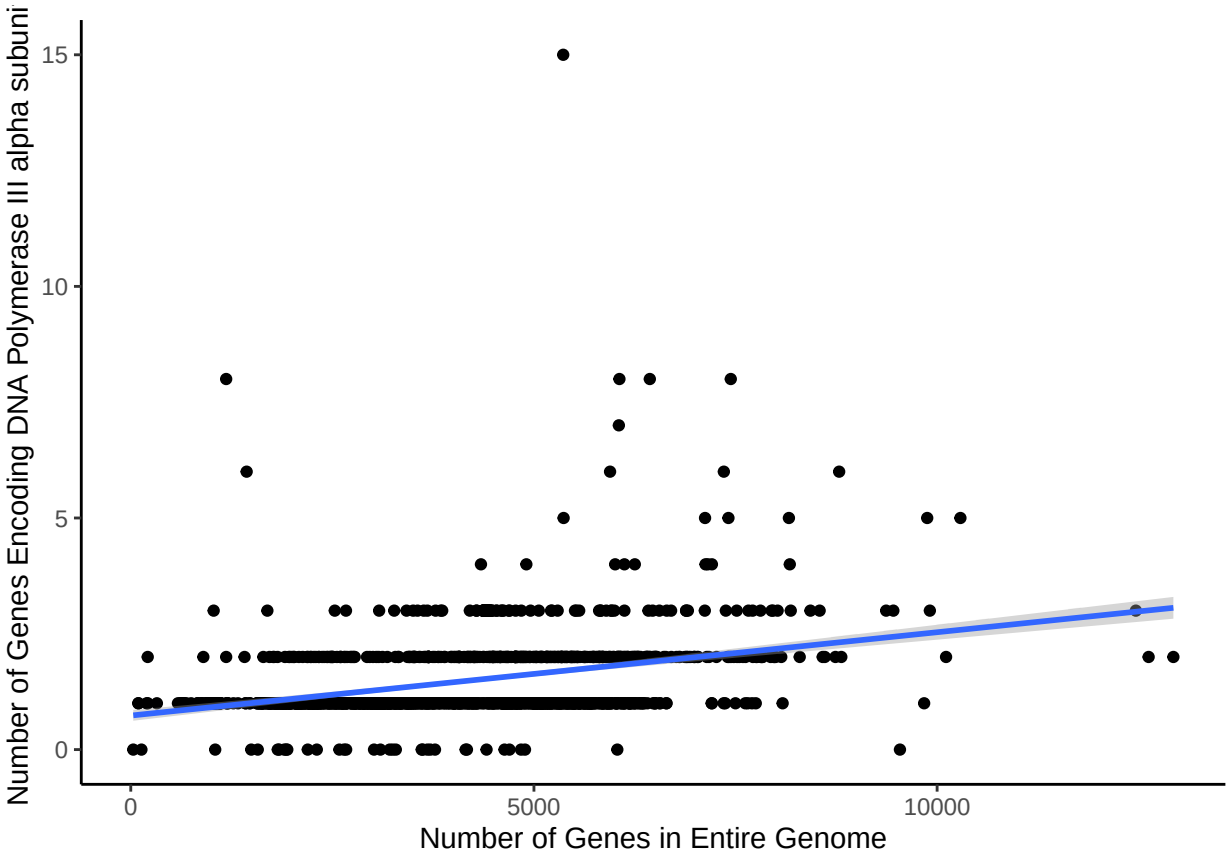
```
##
## Call:
## lm(formula = DNApol_III_a_sub ~ N_Gene_tot, data = gene)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4507 -0.5351 -0.1494  0.3034 13.3006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.346e-01  6.006e-02   12.23   <2e-16 ***
## N_Gene_tot  1.799e-04  1.334e-05   13.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8887 on 1349 degrees of freedom
## Multiple R-squared:  0.1188, Adjusted R-squared:  0.1181
## F-statistic: 181.8 on 1 and 1349 DF,  p-value: < 2.2e-16
```

Question Answers

a. The results of the linear regression model indicate that, on average, each one-integer increase in the number of total genes comprising a genome corresponds to an increase in the number of genes that encode for DNA polymerase III alpha subunit by 1.8*10^-4.
b. Thus, the linear regression formula is: y (number of genes that encode for DNA polymerase III alpha subunit) = (1.8*10^-4) x + 0.735.

```
# fitting a line to p2 to visualize patterns
p2 <- p1 + geom_smooth(method="lm",se=TRUE)

p2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Question Answer

a. The number of DNA polymerase III alpha subunit genes is predicted to increase by 1.8*10^-4 for every one genome length increase.

Challenge 2: Carry out your own analysis! The tRNA-synthetase are enzymes that attach amino acids to the tRNA. Amino acids are the building blocks of proteins and organisms need at least 20 of these enzymes; 1 for each of the 20 existent amino acids. However, some organisms have been described to have more than 20 tRNA-synthetases. Researchers suspect that having more than 20 tRNA-synthetases supports a "faster" protein production. (1) Generate a related research question and (2) test it with your new gained skills, and (3) plot your data!

Question Answers

a. Research Question: Do organisms that have higher numbers of tRNA-synthetases exhibit faster protein production by encoding more X16 rRNA subunits?

```
# Pearson correlation

tRNA <- cor_test(X16rRNA, tRNA_synthetase, data=gene)

#correlation coefficient
r1 <- tRNA$cor

r1
```

```
##  cor
## 0.25
```

```
# standard error of r
r1_se <- sqrt((1-r1^2)/(n-2))
r1_se
```

```
##        cor
## 0.02636208
```

```
# t-test for correlation analysis
tRNA
```

```
## # A tibble: 1 x 8
##   var1    var2              cor statistic        p conf.low conf.high method
##   <chr>   <chr>           <dbl>     <dbl>    <dbl>    <dbl>     <dbl> <chr>
## 1 X16rRNA tRNA_synthetase  0.25      9.51 8.54e-21    0.200     0.300 Pearson
```

```
# manually calculating t-statistic
t1 <- r1/r1_se
t1
```

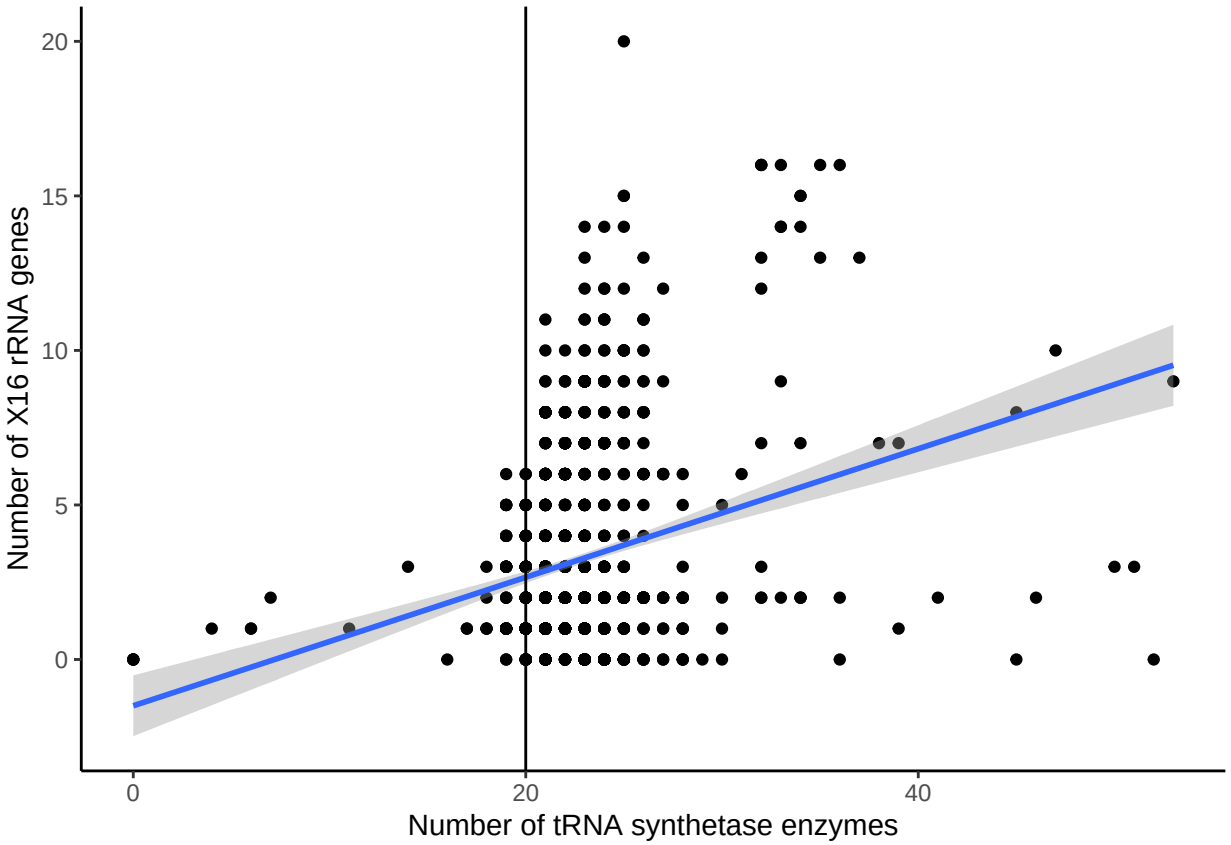
```
##      cor
## 9.483319
```

```
# linear regression
lm2 <- lm(X16rRNA~tRNA_synthetase, data=gene)

# summary of the model output
summary(lm2)
```

```
##
## Call:
## lm(formula = X16rRNA ~ tRNA_synthetase, data = gene)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.313 -1.869 -1.077  1.339 16.300
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.49686    0.50127  -2.986  0.00288 **
## tRNA_synthetase  0.20788    0.02186   9.508  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.946 on 1349 degrees of freedom
## Multiple R-squared:  0.06281,    Adjusted R-squared:  0.06211
## F-statistic: 90.41 on 1 and 1349 DF,  p-value: < 2.2e-16
```

```
p3<-ggplot(gene, aes(x=tRNA_synthetase, y=X16rRNA))+
  geom_point()+
  ylab("Number of X16 rRNA genes")+
  xlab("Number of tRNA synthetase enzymes")+
  geom_smooth(method="lm",se=TRUE)+
  geom_vline(xintercept=20)+
  theme_classic()
p3
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



b. Using a statistical table for the t-distribution (statsexamples.com), we can see that the null expectation for a critical value of 0.025 is t=1.962 with 1000 degrees of freedom. As our estimated t-statistic is much more extreme at 9.50, p is less than 0.05 and we reject the null hypothesis. Additionally, the estimated p-value in the correlation test is extremely small at 8.54*10^-21.

Discussion Question Answers

a. One would want to fit a linear regression model to their data to both determine the direction of a relationship between variables and create a predictive equation to estimate values of the dependent variable based on a hypothetical value of the independent variable.
b. A scatter plot of two associated variables presenting a low standard error of r would demonstrate a majority of data points creating a single diagonal line.
c. A strong correlation does not imply cause and effect because there could be a number of confounding variables that could be contributing to the strong correlation.