DATA VIZ - HOMEWORK VI

NATHANIEL ASIEDU SAKYI

2023-04-12

Loading Complete Data Set into R

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

```
## [1] 43  6
```

```
## [1] "SUMOvar"      "X10.x.copies" "Replicate.1"  "Replicate.2"  "Replicate.3"
## [6] "Average.Cq"
```

Comments: The data consists of 43 observations and 6 variables which are information about gene variant transcriptions, across three replications of each variant.

Inspecting The Unique Groups of the Data

```
## [1] 6 5 4 3 2 1
```

Extracting the Groups Within Data for Visualization

Comment: Another column named "group" which identifies the six different groups was created and added as above.

Table 1: First few observations

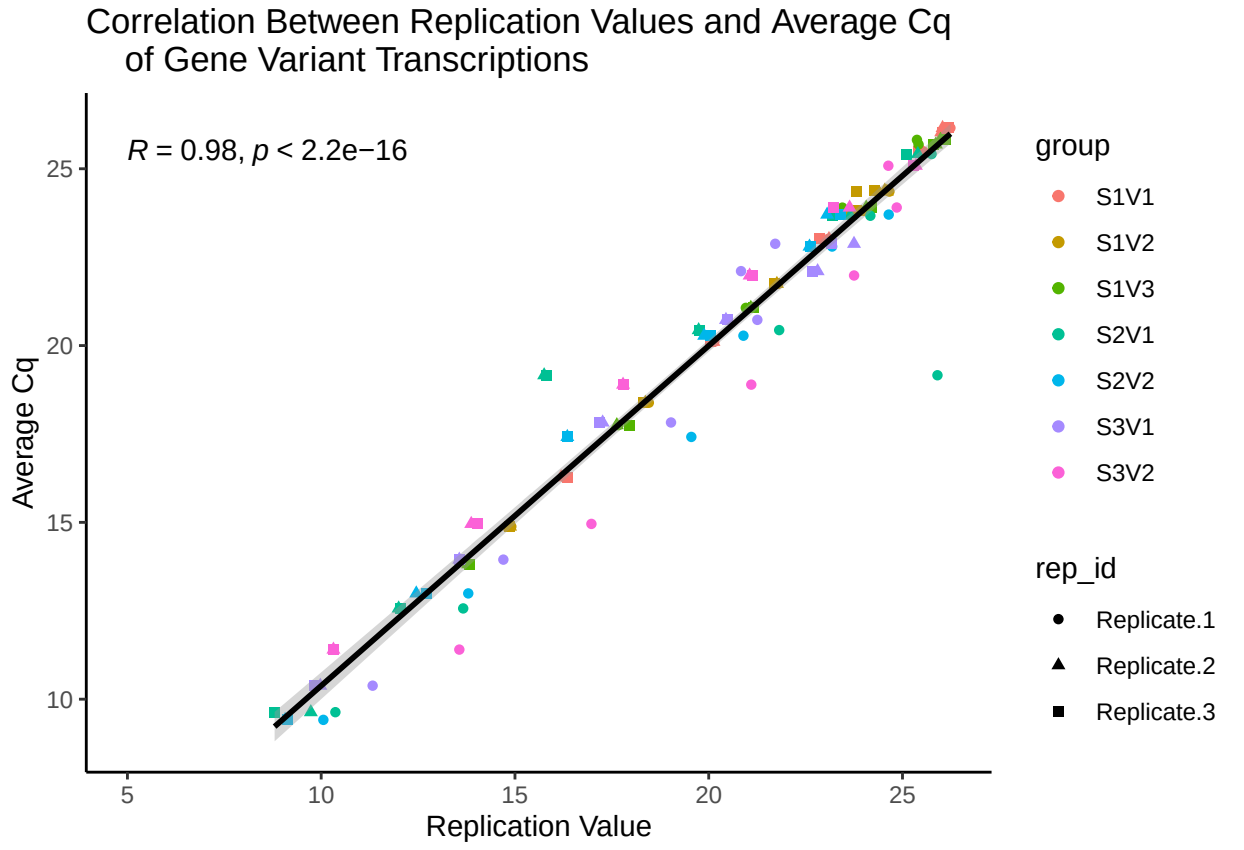
SUMOvar	group	Replicate.1	Replicate.2	Replicate.3	Average.Cq
S1V1-10^6	S1V1	16.27132	16.19231	16.36603	16.27655
S1V1-10^5	S1V1	20.14263	20.12184	20.05466	20.10638
S1V1-10^4	S1V1	23.07819	23.10269	22.86079	23.01389
S1V1-10^3	S1V1	25.53921	25.51511	25.41548	25.48993
S1V1-10^2	S1V1	26.05758	25.99988	26.04024	26.03257
S1V1-10^1	S1V1	26.23620	26.03428	26.19077	26.15375

Table 2: First few observations - Melted Data

SUMOvar	group	Average.Cq	rep_id	rep_value
S1V1-10^6	S1V1	16.27655	Replicate.1	16.27132
S1V1-10^6	S1V1	16.27655	Replicate.2	16.19231
S1V1-10^6	S1V1	16.27655	Replicate.3	16.36603
S1V1-10^5	S1V1	20.10638	Replicate.1	20.14263
S1V1-10^5	S1V1	20.10638	Replicate.2	20.12184
S1V1-10^5	S1V1	20.10638	Replicate.3	20.05466

Melting Data For Visualizing Distributions Across The Three Replications

Visualizing the Association Between the Replication Values and Average Cq For The Different Gene Variant Transcriptions



Comments: It can be seen there exists a very strong linear correlation between the replication value and the Average Cq. The linear correlation coefficient stands statistically significant at .98. The positive correlation between the two quantities is very strong for all three replicates; as well as for all seven groups of the gene variant transcriptions. although we see one outlier for replicate 1 of the S2V1 gene variant.

Appendix

```
knitr::opts_chunk$set(echo = F, warning = FALSE, message = FALSE, cache = F)

library(stringr)
library(dplyr)
library(plotly)
#library(hrbrthemes)
library(kableExtra)
library(knitr)
library(tinytex)
library(tibble)
library(ggrepel)
library("reshape2")
# change default ggplot theme
theme_set(theme_classic())

dat <- read.csv("serialdat.csv", header = T)
class(dat); dim(dat)
names(dat)

unique(dat[-43,]$X10.x.copies, na.rm=T)
dat1 <- dat[-43,]%>%
  select(-X10.x.copies)%>%
  mutate(group = sapply(str_split(SUMOvar,'-'), function(x) {x[1]}),
    .after="SUMOvar")

head(dat1) %>%
  kable(booktabs=T, linesep="",
    caption = "First few observations")

library(tidyr)

df <- dat1 %>%
  gather(key = 'Replicate', value = 'Value',
    -SUMOvar,-group,-Average.Cq)

df <- df %>% dplyr::filter(Value == 1) %>%
  select(SUMOvar, group, Replicate, Average.Cq)

df_tall <- dat1 %>%
  pivot_longer(starts_with("Replicate"),
    values_to = "rep_value", names_to = "rep_id")


head(df_tall)  %>%
  kable(booktabs=T, linesep="",
    caption = "First few observations - Melted Data")
library(ggpubr)
ggplot(data = df_tall, mapping = aes(x = rep_value, y = Average.Cq)) +
  geom_point(mapping = aes(col = group, shape = rep_id)) +
  geom_smooth(method = 'lm', col = "black") +
  stat_cor(label.x = 5) +
```

```
  ylab("Average Cq") +
  xlab("Replication Value") +
  ggtitle("Correlation Between Replication Values and Average Cq
    of Gene Variant Transcriptions")
```