DATA VIZ - HOMEWORK V

NATHANIEL ASIEDU SAKYI

2023-03-29

Contents

Table 1: First few observations

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

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
```

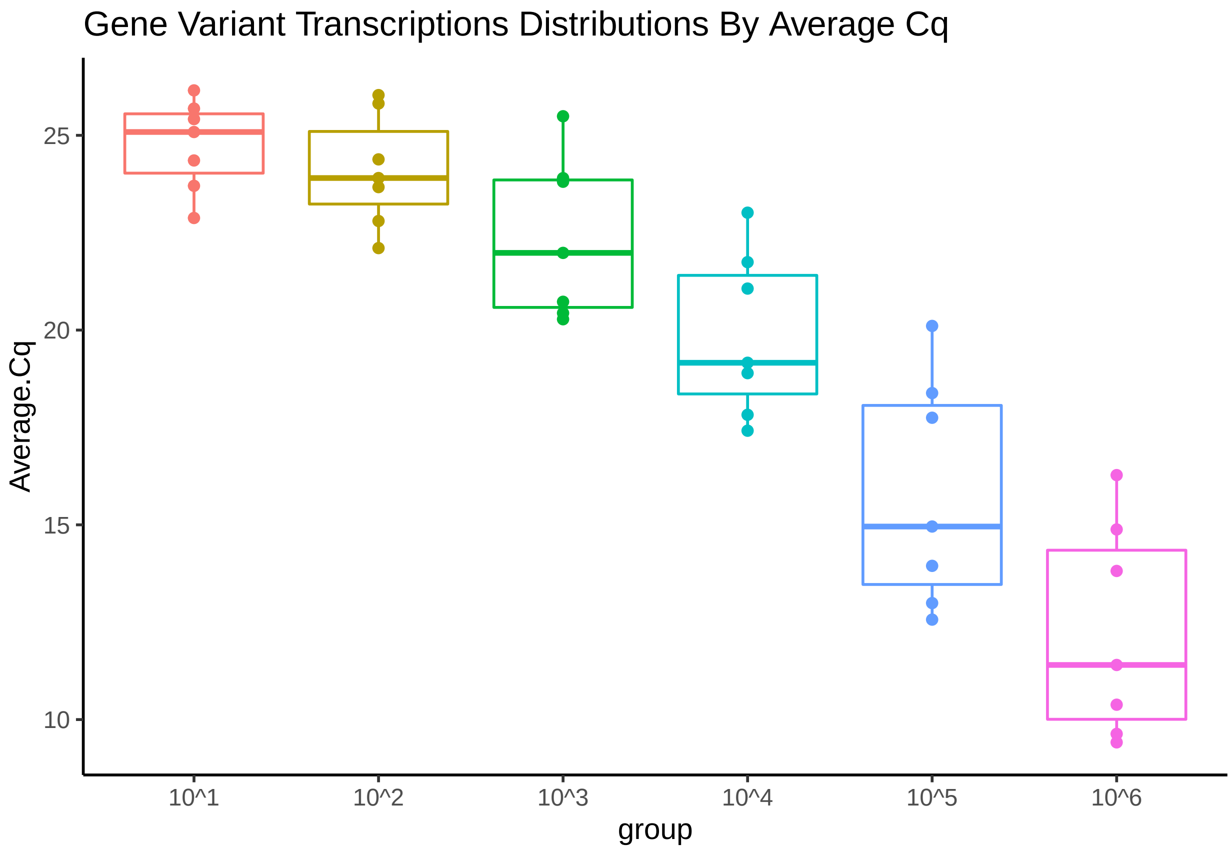
Extracting the Groups Within Data for Visualization

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

Table 2: First few observations - Melted Data

SUMOvar	group	Average.Cq	rep_id	rep_value
S1V1-10ˆ6	10ˆ6	16.27655	Replicate.1	16.27132
S1V1-10ˆ6	10ˆ6	16.27655	Replicate.2	16.19231
S1V1-10ˆ6	10ˆ6	16.27655	Replicate.3	16.36603
S1V1-10ˆ5	10ˆ5	20.10638	Replicate.1	20.14263
S1V1-10ˆ5	10ˆ5	20.10638	Replicate.2	20.12184
S1V1-10ˆ5	10ˆ5	20.10638	Replicate.3	20.05466

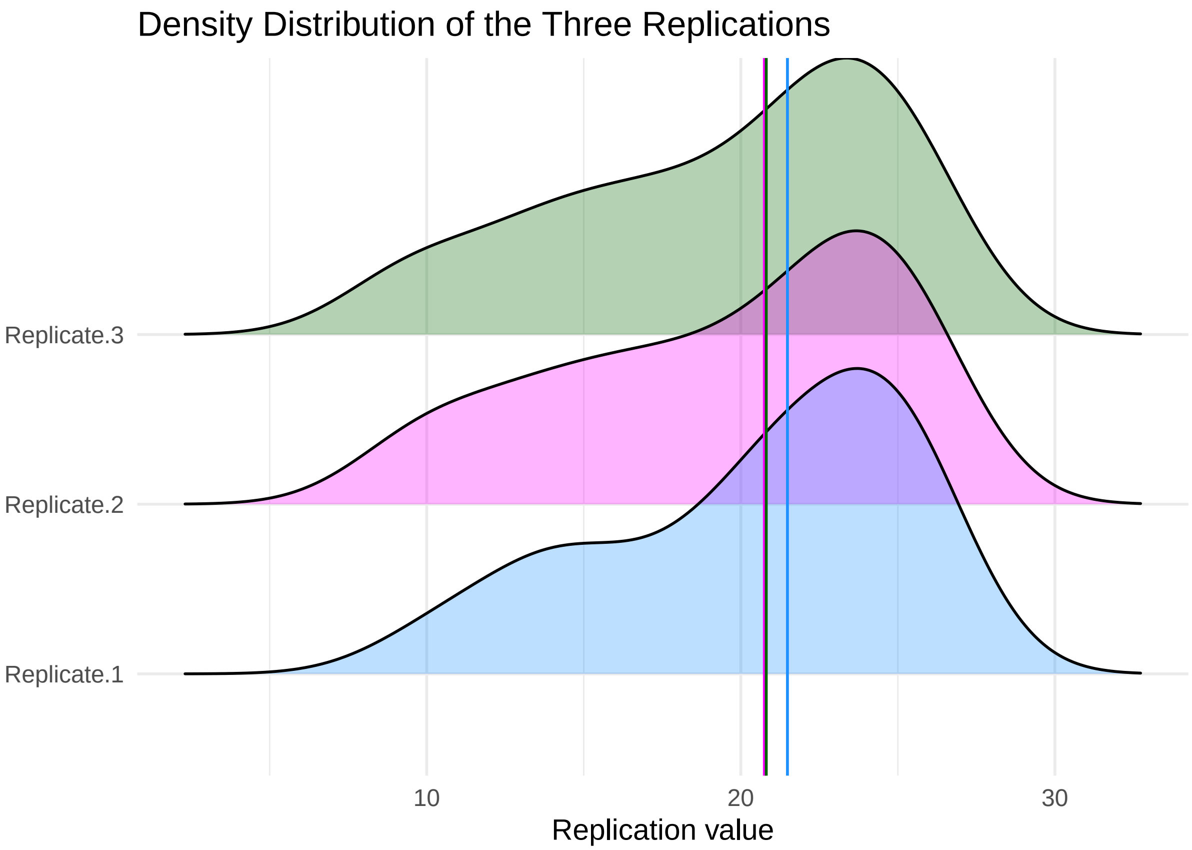
Visualizing The Distribution of Gene Variants By Average Cq.



Comments: It can be seen from the above that, although there exist significant variations among each of the six groups, the variation within the 10^3 group is relatively least among the groups. Also, whereas group 10^1 has the largest Average Cq, group 10^6 has the least Average Cq.

Melting Data For Visualizing Distributions Across The Three Replications

Visualizing Distributions Across The Three Replications



Comments: The plot above reveals the distribution of the gene variant transcriptions across the three replications. It can be clearly seen from the above that, the three distributions are each skewed to the left. The similarity of the skewness implies all three replications are identically distributed. As a result, the median was a chosen measure of location for comparison. It could be seen that, Replicate.1 has the highest median among the three. And the medians for the other two seem to overlap.

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[2]}),
    .after="SUMOvar")

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

ggplot(dat1, aes(group, Average.Cq, group=group, color=group)) +
  geom_boxplot(show.legend = F) +
  geom_point(show.legend = F)+
  ggtitle("Gene Variant Transcriptions Distributions By Average Cq")
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")
# %>%
#   kable_classic()
library(ggridges)

rep_median <- df_tall %>% group_by(rep_id) %>%
  summarise(med=median(rep_value))%>%
  pull(med)
rep_cols <- c("dodgerblue", "magenta", "darkgreen")
ggplot(df_tall, aes(x = rep_value, y = rep_id)) +
  theme_minimal() +
  scale_fill_manual(values = rep_cols)+
  geom_density_ridges(aes(fill = rep_id), alpha = 0.3) +
  geom_vline(xintercept = rep_median[1], color=rep_cols[1]) +
   geom_vline(xintercept = rep_median[2], color=rep_cols[2]) +
   geom_vline(xintercept = rep_median[3], color=rep_cols[3]) +
  labs(title = "Density Distribution of the Three Replications",
    y = "", x = "Replication value") +
    theme(legend.position = "none")
```