Data Cleaning for Gene Expression

Athena Xiourouppa

The following document explains the data cleaning procedures performed on the spreadsheet Karl Berator provided on the 1st of March, 2023, located in `raw-data` under `WTF-IISfD data.xlsx`.

Excel Cleaning

Before importing the spreadsheet into R, we:

- deleted the figures on pages `GL-CsE` and `GL-bNo`
- removed the summary statistics on page `GL-CsE`
- shifted data cells to top-left corner to avoid blank rows when reading into R
- made cell line and treatment type names consistent
- removed cell line and treatment type names from spreadsheet and copied separately.

R Cleaning

We now read the spreadsheet `modified-WTF-IISfD data.xlsx` under `data` into R:

```
# Load necessary packages
pacman::p_load(tidyverse, readxl, reshape2, stringr, viridis, gt, webshot2)

# Obtain names of sheets
sheet_names <- excel_sheets(here::here("data/modified-WTF-IISfD-data.xlsx"))

# Read in each sheet
wtf_iisfd <- lapply(sheet_names,
                    function(n) read_excel(here::here("data/modified-WTF-IISfD-data.xlsx"),
                                           sheet=n))
```

```r
# Create more useful names using data
better_names <- c("WT-P-1", "WT-P-2",
                  "WT-A42-1", "WT-A42-2",
                  "C101-P-1", "C101-P-2",
                  "C101-A42-1", "C101-A42-2")

# Name sheets
names(wtf_iisfd) <- better_names

# Show tibble
print(wtf_iisfd)
```

One trial has a missing value in the gene expression column. Firstly, we label it as NA, and then impute with the mean.

```r
wtf_iisfd$`WT-A42-2`$gene_expression[6] <- NA
wtf_iisfd$`WT-A42-2`$gene_expression[6] <- mean(wtf_iisfd$`WT-A42-2`$gene_expression,
                                                na.rm=TRUE)
```

As there are two trials per cell line and treatment, we merge the separated trial tibbles and take an average of gene expression for every concentration:

```r
WT_P <- right_join(wtf_iisfd$`WT-P-1`, wtf_iisfd$`WT-P-2`, by="conc")
WT_A42 <- right_join(wtf_iisfd$`WT-A42-1`, wtf_iisfd$`WT-A42-2`, by="conc")
CT101_P <- right_join(wtf_iisfd$`C101-P-1`, wtf_iisfd$`C101-P-2`, by="conc")
CT101_A42 <- right_join(wtf_iisfd$`C101-A42-1`, wtf_iisfd$`C101-A42-2`, by="conc")

# Save into new list
grouped_trials <- list(WT_P, WT_A42, CT101_P, CT101_A42)
names(grouped_trials) <- c("WT_P", "WT_A42", "CT101_P", "CT101_A42")
```

Tabular Summary

We now summarise the average gene expression for every concentration.

```r
for(i in 1:4){
  grouped_trials[[i]] <-
    grouped_trials[[i]] |>
    transmute(conc, mean_ge = rowMeans(across(-conc)))
```

```
}

summary_trials <- tibble(conc = grouped_trials[[1]]$conc,
                         WT_P_mean = grouped_trials[[1]]$mean_ge,
                         WT_A42_mean = grouped_trials[[2]]$mean_ge,
                         CT101_P_mean = grouped_trials[[3]]$mean_ge,
                         CT101_A42_mean = grouped_trials[[4]]$mean_ge)
head(summary_trials)
```

```
# A tibble: 6 x 5
   conc WT_P_mean WT_A42_mean CT101_P_mean CT101_A42_mean
  <dbl>     <dbl>       <dbl>        <dbl>          <dbl>
1     0      4.99        9.40         5.58           10.2
2     1      5.76       12.4          5.30           13.3
3     2      5.25       15.0          8.90           16.4
4     3      6.9        18.3         10.3            17.8
5     4      5.58       22.5         12.7            21.8
6     5      6.56       24.2         12.1            26.3
```

Create nicer table for Wild-Type cell summary using `gt`:

```
summary_trials |>
  select(c(conc, WT_P_mean, WT_A42_mean)) |>
  gt() |>
  cols_label(conc = "Concentration (mg/mL)",
             WT_P_mean = "Placebo",
             WT_A42_mean = "Activating Factor 42") |>
  tab_header(title = md("**Summary of Gene Expression in Wild-Type Cells**")) |>
  fmt_number(columns = c(WT_P_mean, WT_A42_mean),
             decimals = 2) |>
  cols_align("center") |>
  tab_spanner(columns = c(WT_P_mean, WT_A42_mean),
              label = "Average Gene Expression") |>
  gtsave(here::here("tabs/2023-03-03-data-cleaning/wt-summary.png"))
```

Create nicer table for 101-Type cell summary using `gt`:

```
summary_trials |>
  select(c(conc, CT101_P_mean, CT101_A42_mean)) |>
  gt() |>
```

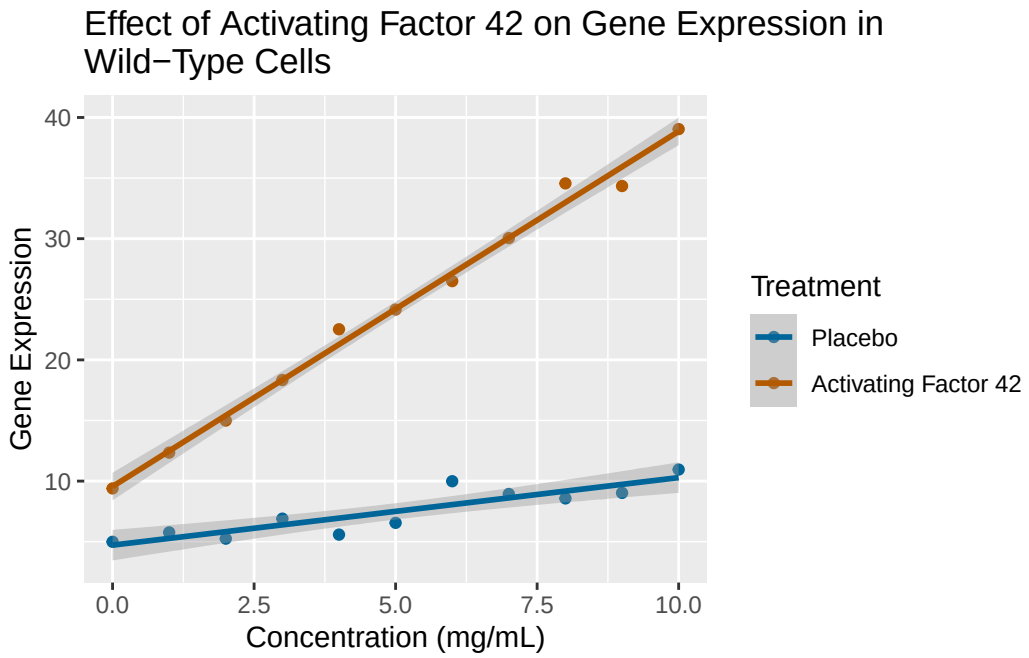
```
  cols_label(conc = "Concentration (mg/mL)",
             CT101_P_mean = "Placebo",
             CT101_A42_mean = "Activating Factor 42") |>
  tab_header(title = md("**Summary of Gene Expression in 101-Type Cells**")) |>
  fmt_number(columns = c(CT101_P_mean, CT101_A42_mean),
             decimals = 2) |>
  cols_align("center") |>
  tab_spanner(columns = c(CT101_P_mean, CT101_A42_mean),
              label = "Average Gene Expression") |>
  gtsave(here::here("tabs/2023-03-03-data-cleaning/101-summary.png"))
```

Figures

We now plot the summaries as line graphs for each cell line and treatment type:

```
# Treatment on Wild cells
summary_trials |>
  melt(id="conc") |>
  filter(str_detect(variable, "WT")) |>
  ggplot(aes(conc, value, col=variable)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(x = "Concentration (mg/mL)",
       y = "Gene Expression",
       title = "Effect of Activating Factor 42 on Gene Expression in \nWild-Type Cells"
       ) +
  harrypotter::scale_color_hp_d("Ravenclaw",
                                labels = c("Placebo", "Activating Factor 42"),
                                name = "Treatment")
```

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

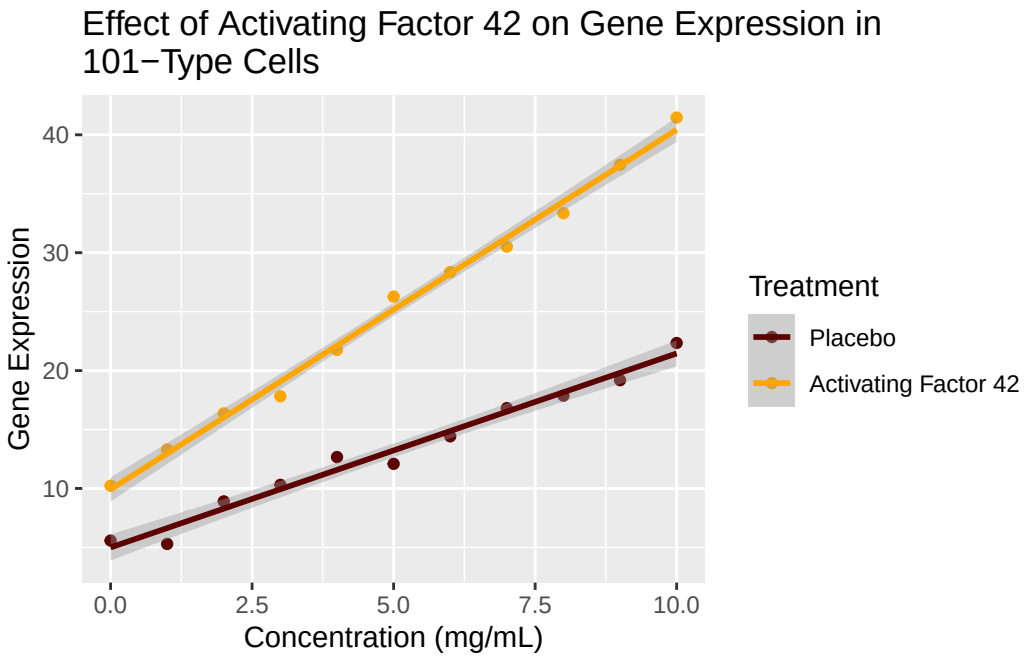



```
ggsave(here::here("figs/2023-03-03-data-cleaning/wt-linegraph.png"))
```

```
Saving 5.5 x 3.5 in image
`geom_smooth()` using formula = 'y ~ x'
```

```
# Treatment on 101 Cells
summary_trials |>
  melt(id="conc") |>
  filter(str_detect(variable, "CT101")) |>
  ggplot(aes(conc, value, col=variable)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(x = "Concentration (mg/mL)",
       y = "Gene Expression",
       title = "Effect of Activating Factor 42 on Gene Expression in \n101-Type Cells"
       ) +
    harrypotter::scale_color_hp_d("Gryffindor",
                                  labels = c("Placebo", "Activating Factor 42"),
                                  name = "Treatment")
```

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




```
ggsave(here::here("figs/2023-03-03-data-cleaning/101-linegraph.png"))
```

```
Saving 5.5 x 3.5 in image
`geom_smooth()` using formula = 'y ~ x'
```

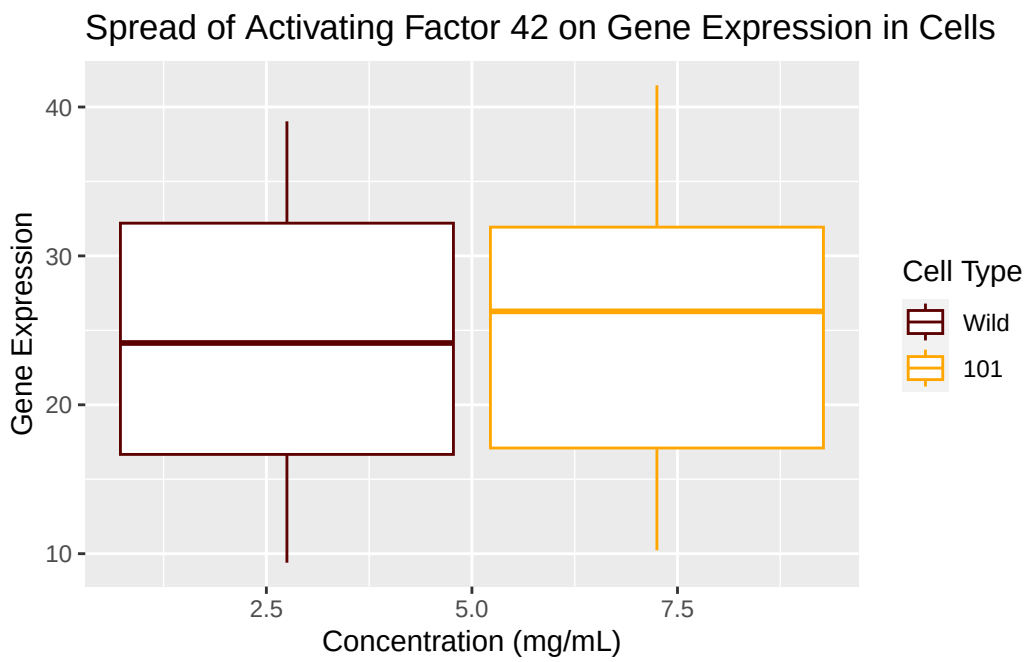
We also display the spread of values for each cell line and treatment combination

```
# Placebo on cells
summary_trials |>
  melt(id="conc") |>
  filter(str_detect(variable, "P")) |>
  ggplot(aes(conc, value, col=variable)) +
  geom_boxplot() +
  labs(x = "Concentration (mg/mL)",
       y = "Gene Expression",
       title = "Spread of Placebo on Gene Expression in Cells"
       ) +
    harrypotter::scale_color_hp_d("Ravenclaw",
                                  labels = c("Wild", "101"),
                                  name = "Cell Type")
```


```
ggsave(here::here("figs/2023-03-03-data-cleaning/placebo-boxplot.png"))
```

```
Saving 5.5 x 3.5 in image
```

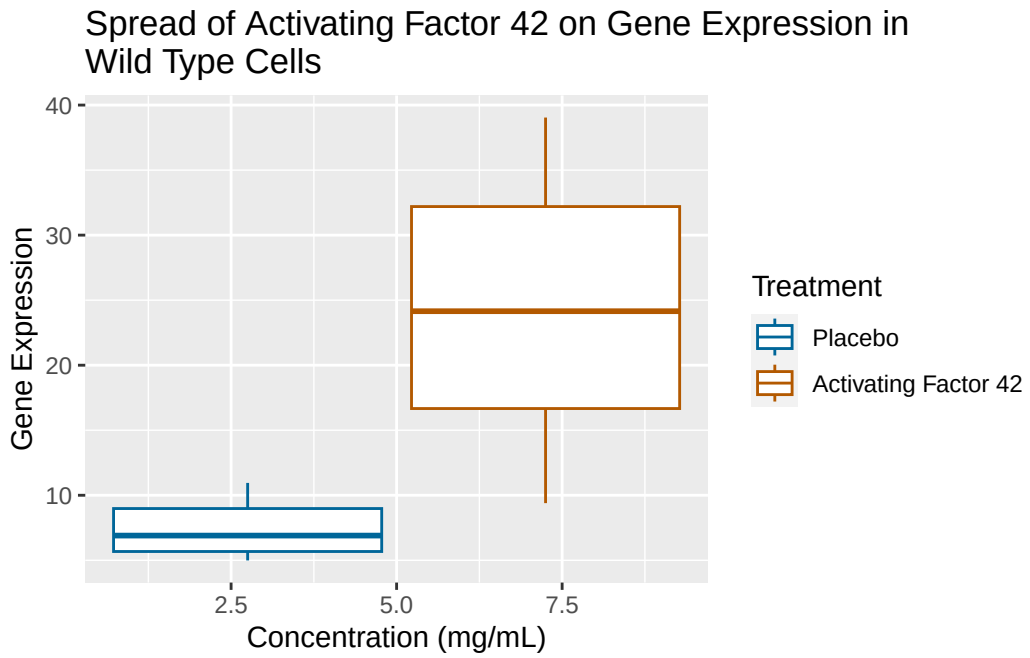
```
# A42 on cells
summary_trials |>
  melt(id="conc") |>
  filter(str_detect(variable, "A42")) |>
  ggplot(aes(conc, value, col=variable)) +
  geom_boxplot() +
  labs(x = "Concentration (mg/mL)",
       y = "Gene Expression",
       title = "Spread of Activating Factor 42 on Gene Expression in Cells"
       ) +
  harrypotter::scale_color_hp_d("Gryffindor",
                                labels = c("Wild", "101"),
                                name = "Cell Type")
```




```
ggsave(here::here("figs/2023-03-03-data-cleaning/a42-boxplot.png"))
```

```
Saving 5.5 x 3.5 in image
```

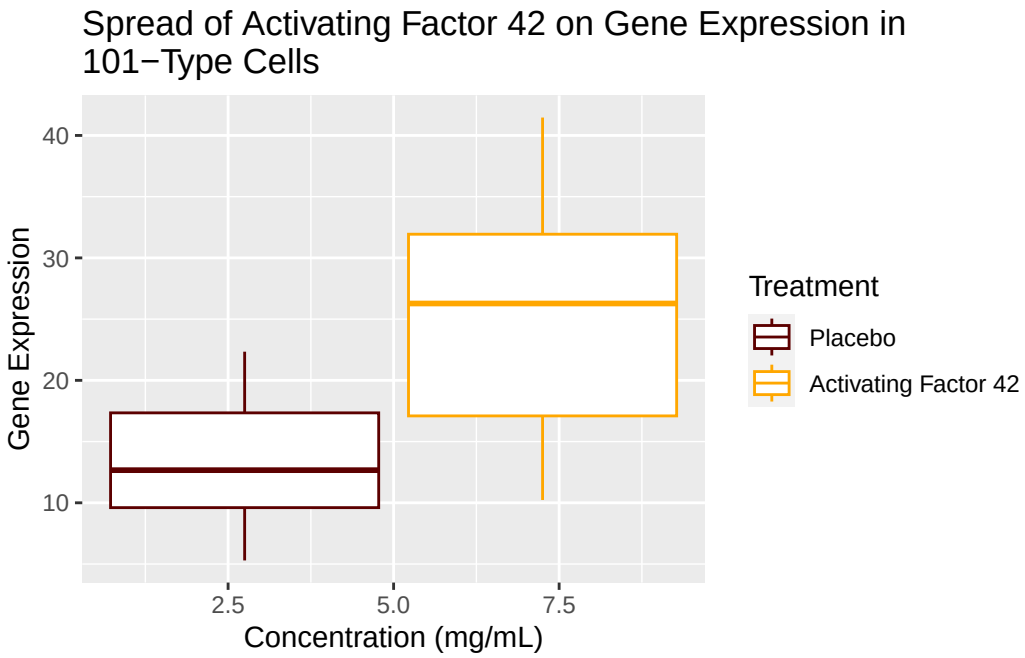
```
# Treatment on WT cells
summary_trials |>
  melt(id="conc") |>
  filter(str_detect(variable, "WT")) |>
  ggplot(aes(conc, value, col=variable)) +
  geom_boxplot() +
  labs(x = "Concentration (mg/mL)",
       y = "Gene Expression",
       title = "Spread of Activating Factor 42 on Gene Expression in \nWild Type Cells"
       ) +
  harrypotter::scale_color_hp_d("Ravenclaw",
                                labels = c("Placebo", "Activating Factor 42"),
                                name = "Treatment")
```




```
ggsave(here::here("figs/2023-03-03-data-cleaning/wt-boxplot.png"))
```

```
Saving 5.5 x 3.5 in image
```

```
# Treatment on 101 cells
summary_trials |>
  melt(id="conc") |>
  filter(str_detect(variable, "CT101")) |>
  ggplot(aes(conc, value, col=variable)) +
  geom_boxplot() +
  labs(x = "Concentration (mg/mL)",
       y = "Gene Expression",
       title = "Spread of Activating Factor 42 on Gene Expression in \n101-Type Cells"
       ) +
  harrypotter::scale_color_hp_d("Gryffindor",
                                labels = c("Placebo", "Activating Factor 42"),
                                name = "Treatment")
```




```
ggsave(here::here("figs/2023-03-03-data-cleaning/101-boxplot.png"))
```

```
Saving 5.5 x 3.5 in image
```