Abstract

Jannik Hannemann

2024-12-25

Introduction

Given a certain data set, existing of 4 different cell types, 9 different genes, 3 different experimental conditions and their measured CT values (in 6 instances), we are interested in the variance of the CT values for each gene depending on the different experimental conditions and the different cell lines. Like this, the scientist group can choose a gene from this experiment as their reference gene for the real time PCRs which have to be conducted in the future. This chosen gene should be a stable one and the science group can choose, whether they need a gene that is stable by the different cell lines or by the different experimental conditions.

The Objective of this project in R is the graphical representation of these two variances for each gene in a scatter plot.

Editing the data

```
#read all data
data <- read_csv2("GeneDataAllCellLinesNewFormat.csv")
```

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

```
## Rows: 648 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (3): Gene, Cell Line, Experimental Condition
## dbl (1): CT Value
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#change the variable names to a usable format in R
names(data)<-str_replace_all(names(data), c(" " = "."))

#show the mean of each gene in each cell line and experimental condition to get a first
#overview
data %>%
  group_by(Cell.Line, Gene, Experimental.Condition) %>%
  summarise(mean(CT.Value, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Cell.Line', 'Gene'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 108 x 4
## # Groups:   Cell.Line, Gene [36]
##    Cell.Line Gene  Experimental.Condition `mean(CT.Value, na.rm = TRUE)`
##    <chr>     <chr> <chr>                                           <dbl>
##  1 A549      18s   24h HX                                           8.02
##  2 A549      18s   24h NX                                           7.69
##  3 A549      18s   72h HX                                           7.82
##  4 A549      ActB  24h HX                                          17.3
##  5 A549      ActB  24h NX                                          17.0
##  6 A549      ActB  72h HX                                          16.9
##  7 A549      B2M   24h HX                                          22.7
##  8 A549      B2M   24h NX                                          22.4
##  9 A549      B2M   72h HX                                          22.1
## 10 A549      GAPDH 24h HX                                          18.6
## # i 98 more rows
```

```
#get the data that is affected by the different cell lines
varianceDataCellLine <- data %>%
  #group the different "sets" of data
  group_by(Cell.Line, Gene, Experimental.Condition) %>%
  #filter the data to only get the normoxic data, which means it will not be affected
  #by the experimental condition
  filter(Experimental.Condition == "24h NX") %>%
  #group the data by the different genes
  group_by(Gene) %>%
  summarise(
    #calculate the standard deviation of the CT values for each gene
    standardDeviationByCellLine = sd(CT.Value, na.rm = TRUE),
    #calculate the mean of the CT values for each gene
    meanByCellLine = mean(CT.Value, na.rm = TRUE),
    #calculate the coefficient of variation for each gene
    coefficientOfVariationCellLine = standardDeviationByCellLine / meanByCellLine,
    #calculate the minimum of the confidence interval for the coefficient of variation
    confIntMinCellLine = sdCI(CT.Value)$conf.int[1] / meanByCellLine,
    #calculate the maximum of the confidence interval for the coefficient of variation
    confIntMaxCellLine = sdCI(CT.Value)$conf.int[2] / meanByCellLine
  )

#get the data that is affected by the different Experimental Conditions
varianceDataExpCond <- data %>%
  #group the different "sets" of data
  group_by(Cell.Line, Gene) %>%
  #first get the data in groups for each different cell line,
  #because we want to clear out the effect of the different cell lines
  mutate(
    #calculate the standard deviation of the CT values for each gene and each cell line
    standardDeviationByExpCond = sd(CT.Value, na.rm = TRUE),
    #calculate the mean of the CT values for each gene and each cell line
    meanByExpCond = mean(CT.Value, na.rm = TRUE),
    #calculate the coefficient of variation for each gene and each cell line
    coefficientOfVariationExpCond = standardDeviationByExpCond / meanByExpCond,
    #calculate the minimum of the confidence interval for the coefficient of variation
    confIntMinExpCond = sdCI(CT.Value)$conf.int[1] / meanByExpCond,
    #calculate the maximum of the confidence interval for the coefficient of variation
```

```
    confIntMaxExpCond = sdCI(CT.Value)$conf.int[2] / meanByExpCond
  ) %>%
  #group the data by the different genes to get the means of all the values
  #for each gene (like this we clear out the effect of the different cell lines)
  group_by(Gene) %>%
  summarise(
    #calculate the mean of the coefficient of variation for each gene
    coefficientOfVariationExpCond = mean(coefficientOfVariationExpCond, na.rm = TRUE),
    #calculate the minimum of the confidence interval for the coefficient of variation
    confIntMinExpCond = mean(confIntMinExpCond, na.rm = TRUE),
    #calculate the maximum of the confidence interval for the coefficient of variation
    confIntMaxExpCond = mean(confIntMaxExpCond, na.rm = TRUE)
  )

varianceDataCellLine
```

```
## # A tibble: 9 x 6
##   Gene   standardDeviationByCellLine meanByCellLine coefficientOfVariationCell~1
##   <chr>                        <dbl>          <dbl>                        <dbl>
## 1 18s                          0.537           7.89                       0.0680
## 2 ActB                         0.312          16.9                        0.0185
## 3 B2M                          0.873          21.6                        0.0403
## 4 GAPDH                        0.570          19.5                        0.0292
## 5 PPIA                         0.543          26.0                        0.0208
## 6 RPL13a                       0.460          22.5                        0.0204
## 7 RPLP1                        0.512          20.8                        0.0246
## 8 SDHA                         0.571          24.7                        0.0231
## 9 TBP                          0.404          26.0                        0.0155
## # i abbreviated name: 1: coefficientOfVariationCellLine
## # i 2 more variables: confIntMinCellLine <dbl>, confIntMaxCellLine <dbl>
```

```
varianceDataExpCond
```

```
## # A tibble: 9 x 4
##   Gene   coefficientOfVariationExpCond confIntMinExpCond confIntMaxExpCond
##   <chr>                          <dbl>             <dbl>             <dbl>
## 1 18s                           0.0516           0.0385            0.0781
## 2 ActB                          0.0222           0.0166            0.0336
## 3 B2M                           0.0159           0.0118            0.0242
## 4 GAPDH                         0.0183           0.0136            0.0277
## 5 PPIA                          0.0167           0.0124            0.0253
## 6 RPL13a                        0.0153           0.0114            0.0231
## 7 RPLP1                         0.0165           0.0123            0.0250
## 8 SDHA                          0.0214           0.0160            0.0323
## 9 TBP                           0.0130           0.00969           0.0198
```

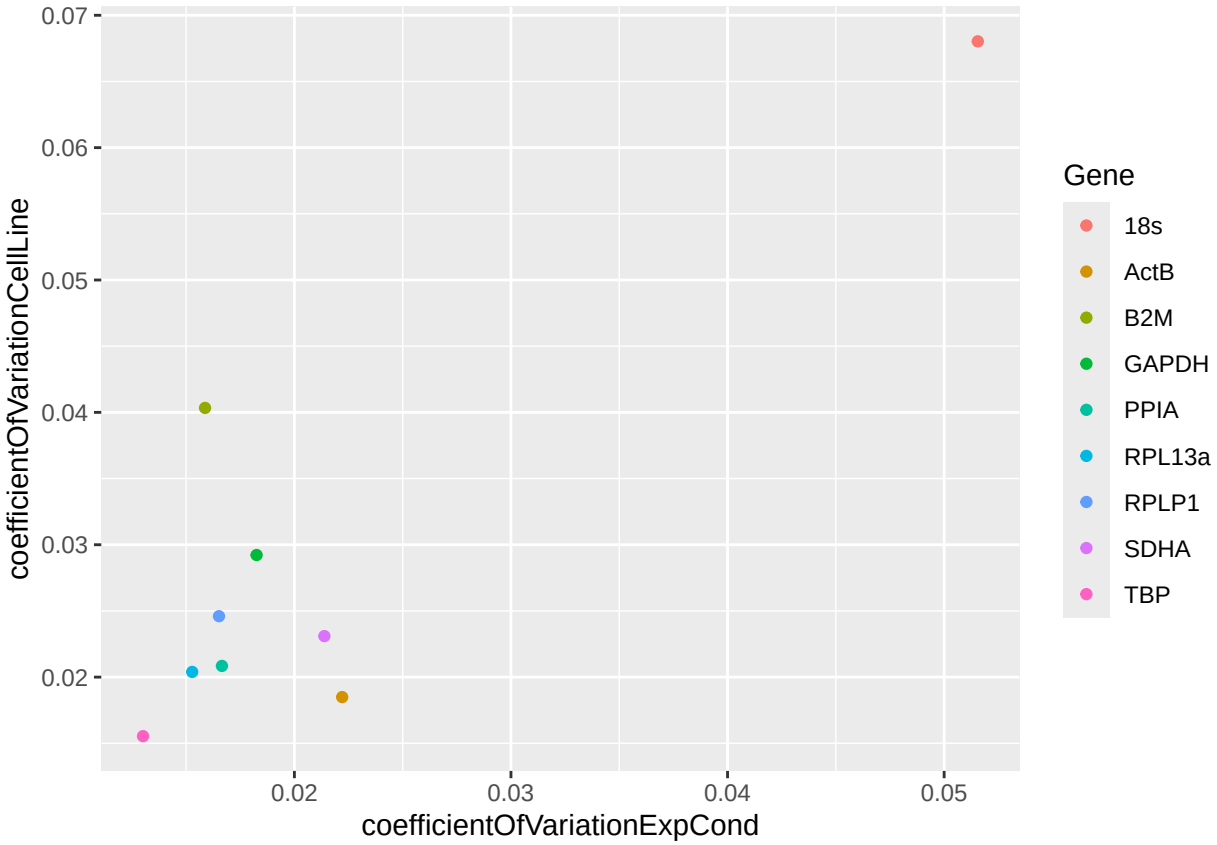
```
#group all import data in a tibble
finalData <- merge(varianceDataCellLine, varianceDataExpCond, by = "Gene")
finalData
```

```
##     Gene standardDeviationByCellLine meanByCellLine
## 1    18s                   0.5370910       7.894913
```

```
## 2   ActB                   0.3117300      16.858826
## 3    B2M                   0.8727290      21.637000
## 4  GAPDH                   0.5700562      19.507652
## 5   PPIA                   0.5428253      26.042087
## 6 RPL13a                   0.4596463      22.540609
## 7  RPLP1                   0.5122805      20.822217
## 8   SDHA                   0.5712545      24.733217
## 9    TBP                   0.4035181      25.968870
##   coefficientOfVariationCellLine confIntMinCellLine confIntMaxCellLine
## 1                     0.06803001         0.05261406         0.09628636
## 2                     0.01849061         0.01430054         0.02617071
## 3                     0.04033503         0.03119490         0.05708824
## 4                     0.02922218         0.02260029         0.04135966
## 5                     0.02084416         0.01612076         0.02950181
## 6                     0.02039192         0.01577101         0.02886173
## 7                     0.02460259         0.01902752         0.03482131
## 8                     0.02309665         0.01786283         0.03268988
## 9                     0.01553853         0.01201742         0.02199248
##   coefficientOfVariationExpCond confIntMinExpCond confIntMaxExpCond
## 1                    0.05155466       0.038494503        0.07810525
## 2                    0.02220750       0.016586651        0.03362345
## 3                    0.01587533       0.011828076        0.02416043
## 4                    0.01825517       0.013631555        0.02765269
## 5                    0.01665420       0.012409856        0.02533952
## 6                    0.01528242       0.011413951        0.02314017
## 7                    0.01651862       0.012328710        0.02504835
## 8                    0.02138449       0.015989990        0.03230032
## 9                    0.01301087       0.009694075        0.01980025
```

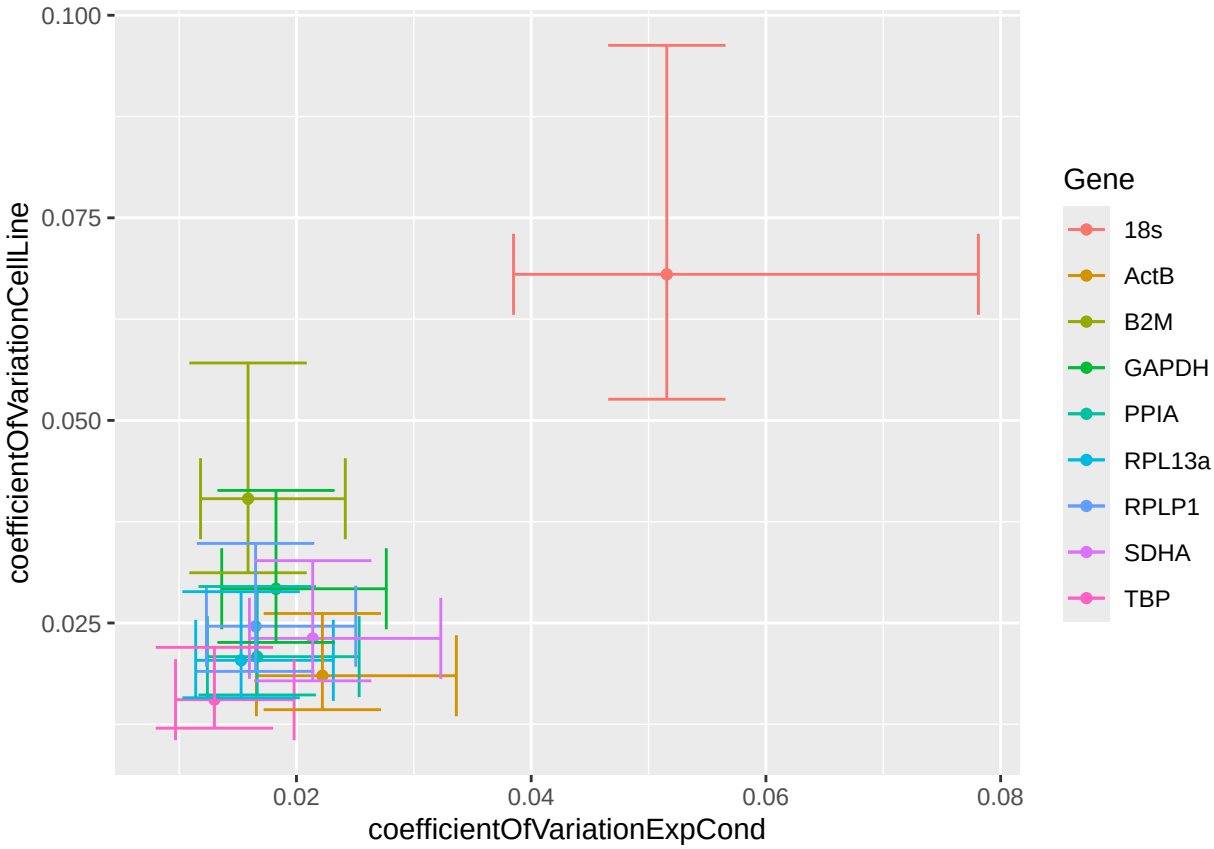
Plotting the raw data

```
ggplot(data = finalData) +
  aes(x = coefficientOfVariationExpCond, y = coefficientOfVariationCellLine, color = Gene) +
  geom_point()
```



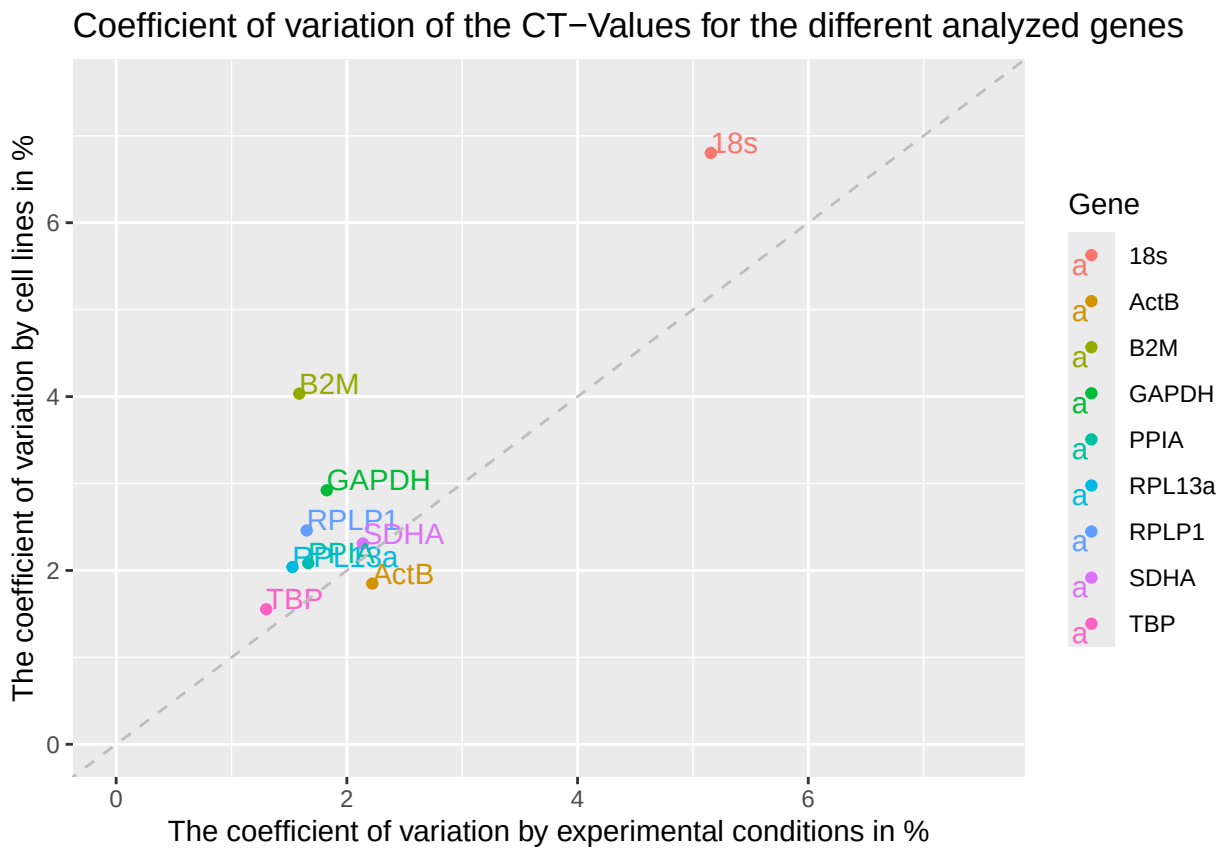
Adding error bars

```
ggplot(data = finalData) +
  aes(x = coefficientOfVariationExpCond, y = coefficientOfVariationCellLine, color = Gene) +
  geom_point() +
  geom_errorbar(aes(xmin = confIntMinExpCond, xmax=confIntMaxExpCond), width=.01) +
  geom_errorbar(aes(ymin = confIntMinCellLine, ymax=confIntMaxCellLine), width=.01)
```



Optimizing the plot (error bars are too dense for a graphical representation)

```
ggplot(data = finalData) +
  aes(
    x = coefficientOfVariationExpCond * 100,
    y = coefficientOfVariationCellLine * 100,
    color = Gene
  ) +
  geom_point() +
  geom_text(label=finalData$Gene, hjust=0, vjust=0) +
  lims(x = c(0, 7.5), y = c(0, 7.5)) +
  #add a dashed line to show the 1:1 line.
  #Any gene that is under this line means that the
  #coefficient of variation of this gene is higher by
  #the experimental conditions than by the cell lines
  geom_abline(intercept = 0, slope = 1, color = "grey", linetype = "dashed") +
  labs(
    x = "The coefficient of variation by experimental conditions in %",
    y = "The coefficient of variation by cell lines in %",
    title = "Coefficient of variation of the CT-Values for the different analyzed genes"
  )
```


Coefficient of variation of the CT−Values for the different analyzed genes
The coefficient of variation by cell lines in %
The coefficient of variation by experimental conditions in %
0
2
4
6
18s
B2M
GAPDH
RPLP1
SDHA
PPIA
RPL13a
ActB
TBP
Gene
18s
ActB
B2M
GAPDH
PPIA
RPL13a
RPLP1
SDHA
TBP