

masters eo project

2025-01-22

R Markdown

installing packages

```
#install.packages("factoextra")
#install.packages("tidyr")
#install.packages("readr")
#install.packages("ggplot2")
#install.packages("DescTools")
#install.packages('ggrepel')
#install.packages("minpack.lm")
#install.packages("drc")
#install.packages("purrr")
library(tidyr)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(DescTools)
library(ggrepel)
library(minpack.lm)
library(drc)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##   select

##
## 'drc' has been loaded.

## Please cite R and 'drc' if used for a publication,
## for references type 'citation()' and 'citation('drc')'.
```

```
##
## Attaching package: 'drc'

## The following objects are masked from 'package:stats':
##
##      gaussian, getInitial

library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(purrr)
```

we now load in the data file downloaded from the drive

```
setwd('~/Documents/Masters_project/')
HIV1_vector_data=read_csv('inputs/HIV1_vectors_collated_n0_uncleaned(Sheet1).csv')
```

```
## Rows: 1019175 Columns: 15
## -- Column specification -----
## Delimiter: ","
## chr (4): assay, cell_line, titre, condition
## dbl (11): index, batch, plate_column, replicate, screen, screen_nb, infectio...
##
## 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.
```

need to clean data, cleaning up NAs and empty rows in the original dataset, only keeping useful column we also get rid of NA

```
HIV1_vector_data=HIV1_vector_data[,c(2,5,6,7:13,15)]
HIV1_vector_data=na.omit(HIV1_vector_data)
```

here we get rid of reduced values, this normally represents dying cells which is irrelevant to our analysis. in this code if a data point with a higher volume within a replicate is less than 80% of the max for that replicate it is replaced with the maximum. This practically means that as the graph increases it plateaus at the max instead of decreasing

```
library(dplyr)
HIV1_vector_data=HIV1_vector_data%>%group_by(cell_line,screen_nb,replicate)%>%
  mutate(max_gfp=max(assay_output),max_gfp_titre=which.max(assay_output)) %>%
  mutate(assay_output=if_else((infection_volume_ul>infection_volume_ul[max_gfp_titre] & assay_output <= 0.8*max_gfp_titre),max_gfp_titre,assay_output))
```

subsetting data for each cell line, creates a list of dataframes for each cell line (makes it easier to manipulate) per cell line

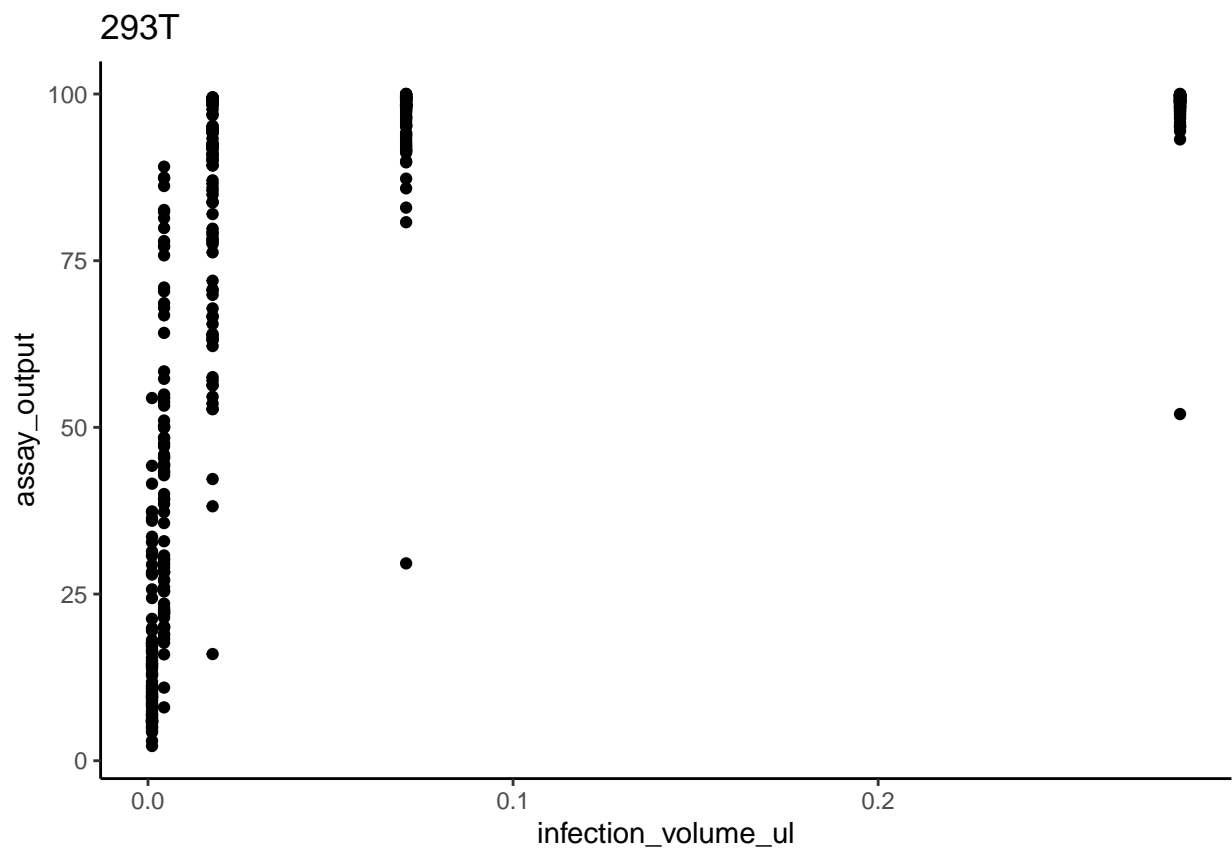
```
vector_data_per_cell_line=HIV1_vector_data%>%nest_by(cell_line,.keep = T)
```

this points all plots by titre amount and separated by screen, allows you to see general shape of data, separated by

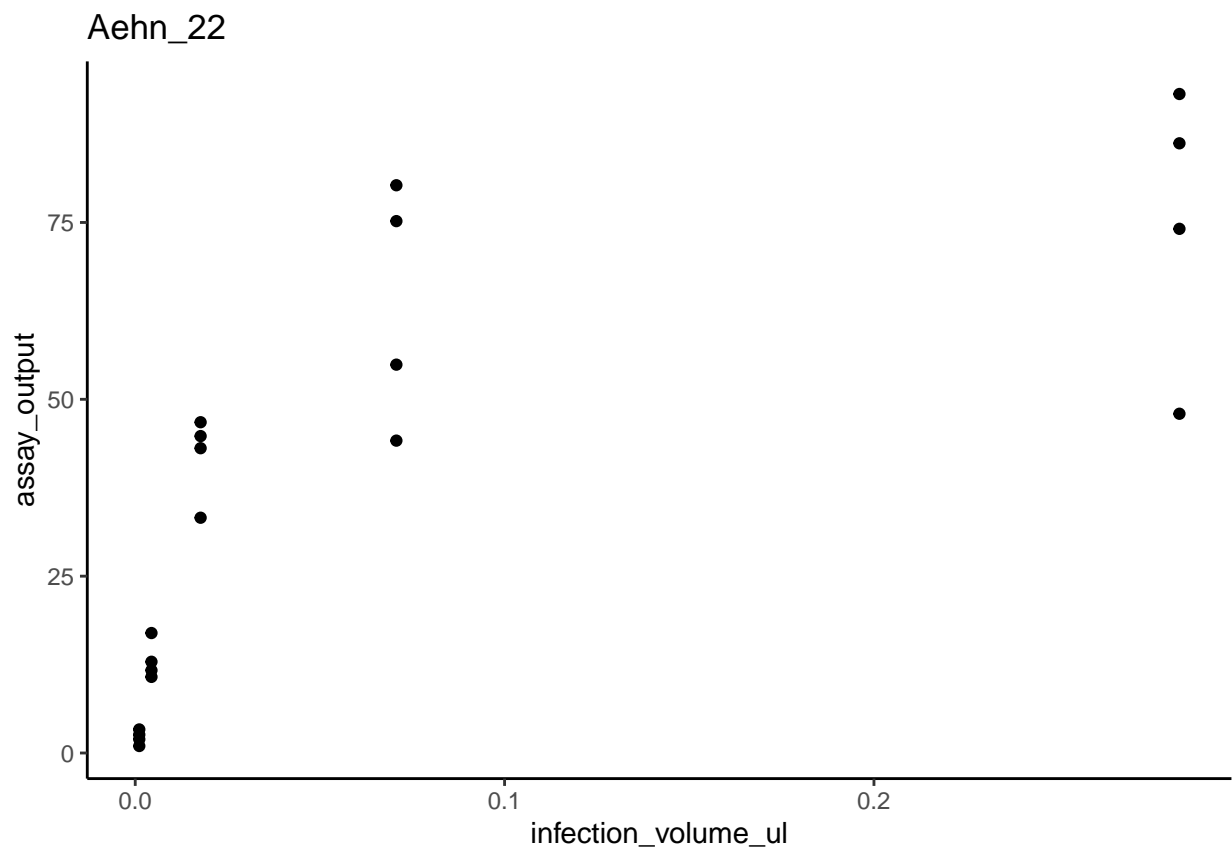
```
apply_plot=function(i){ggplot(data = i,aes(x=infection_volume_ul,y=assay_output,group = interaction(screen_nb,cell_line)))
  geom_point()+
  ggtitle(i$cell_line[1])+
  theme_classic()
}
```

```
purrr::map(vector_data_per_cell_line$data,apply_plot)
```

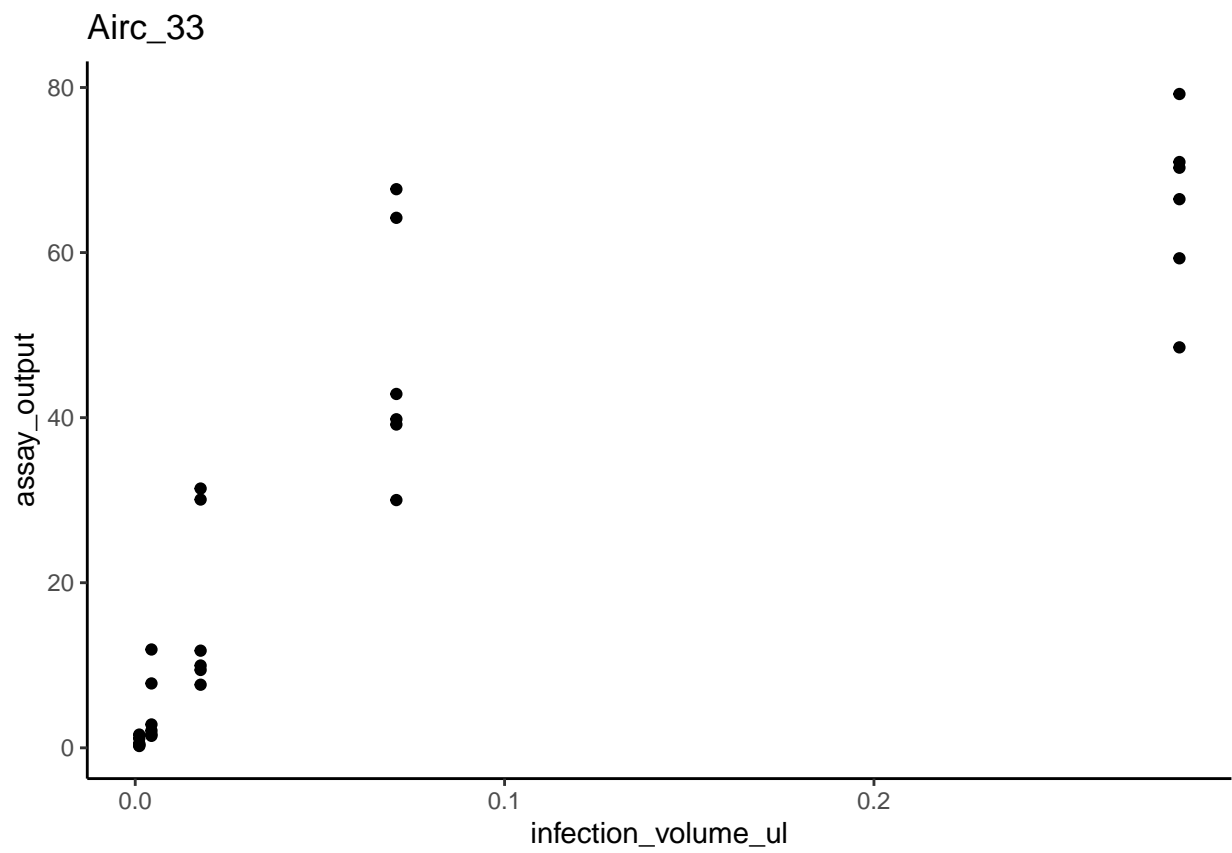
```
## [[1]]
```



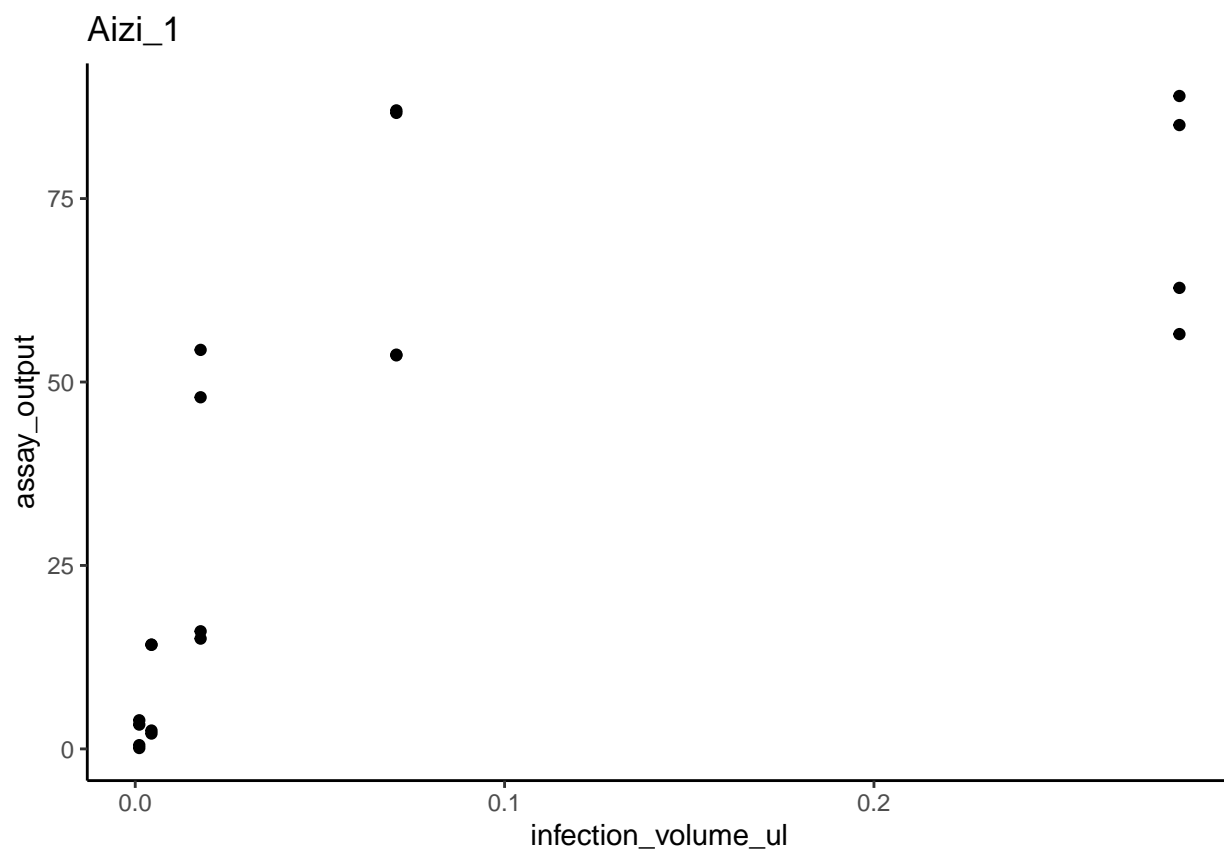
[[2]]



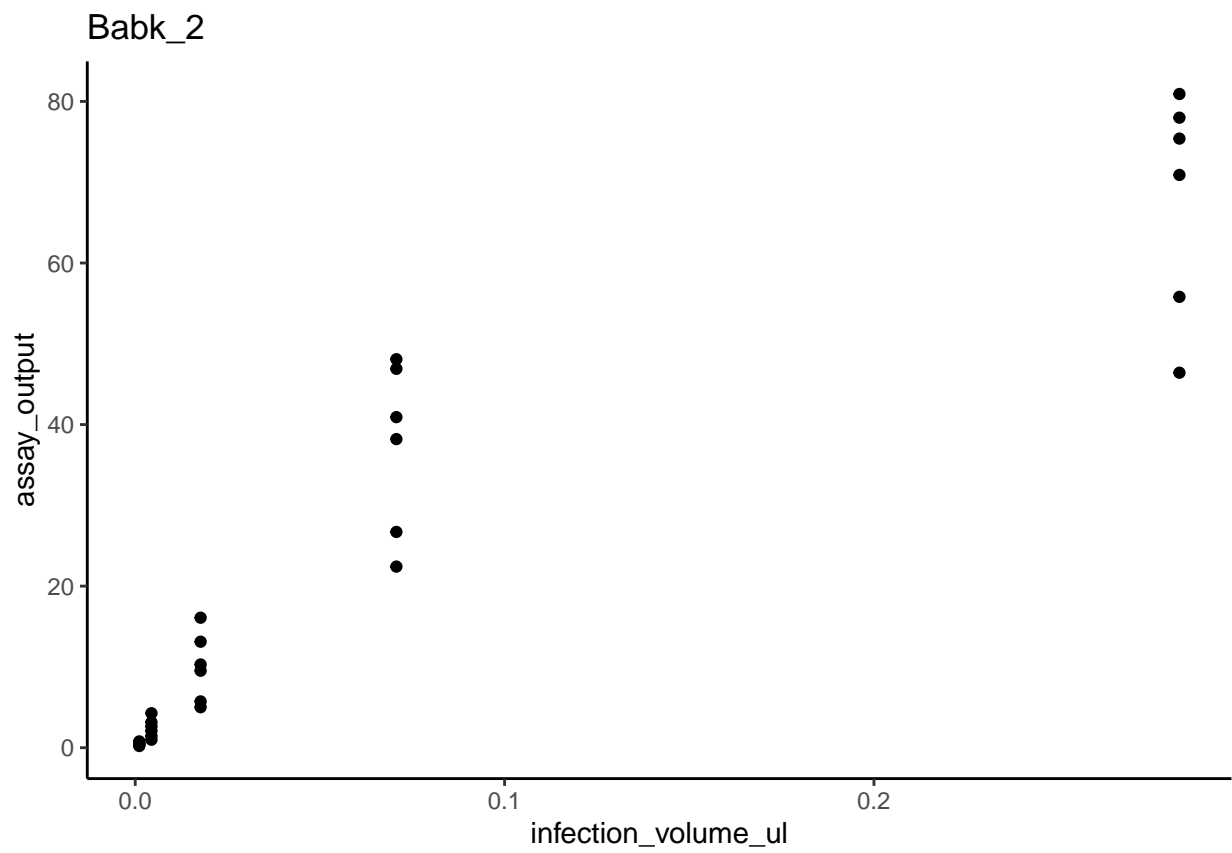
[[3]]



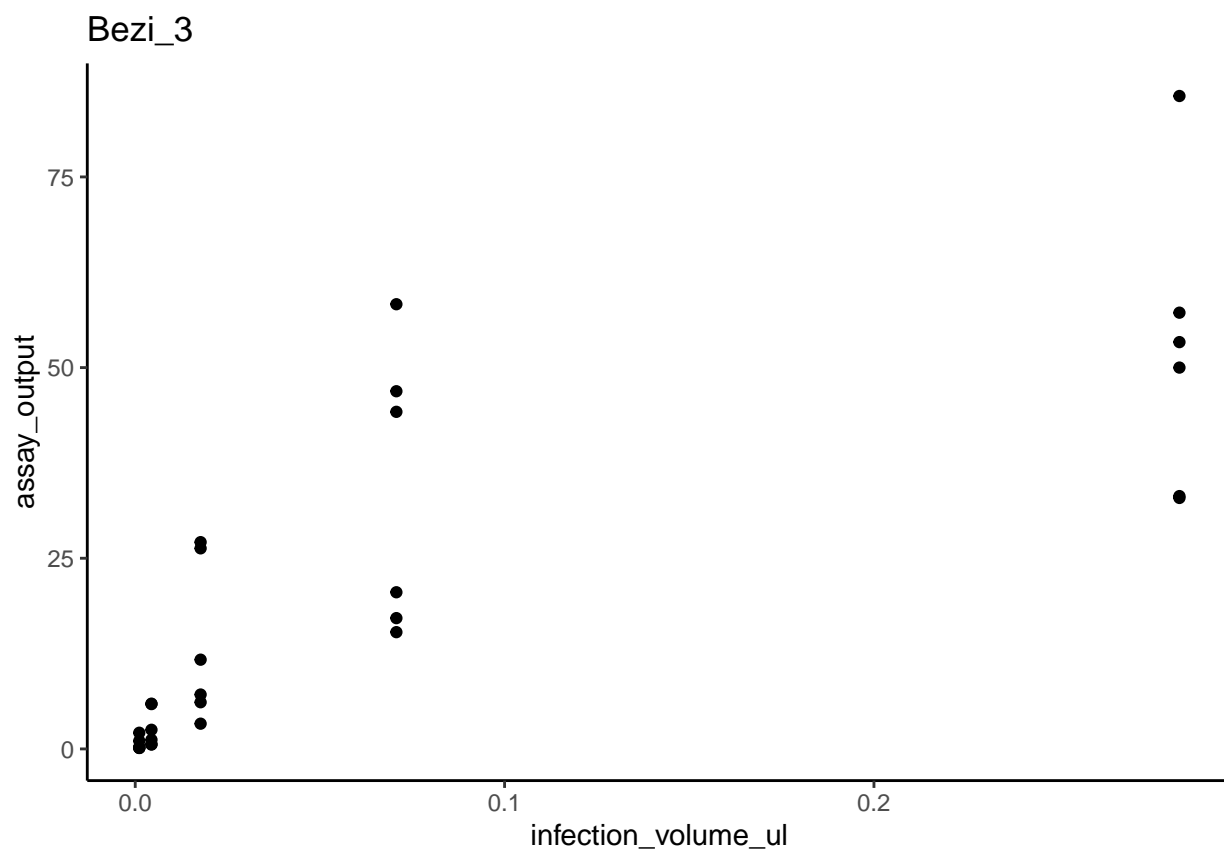
[[4]]



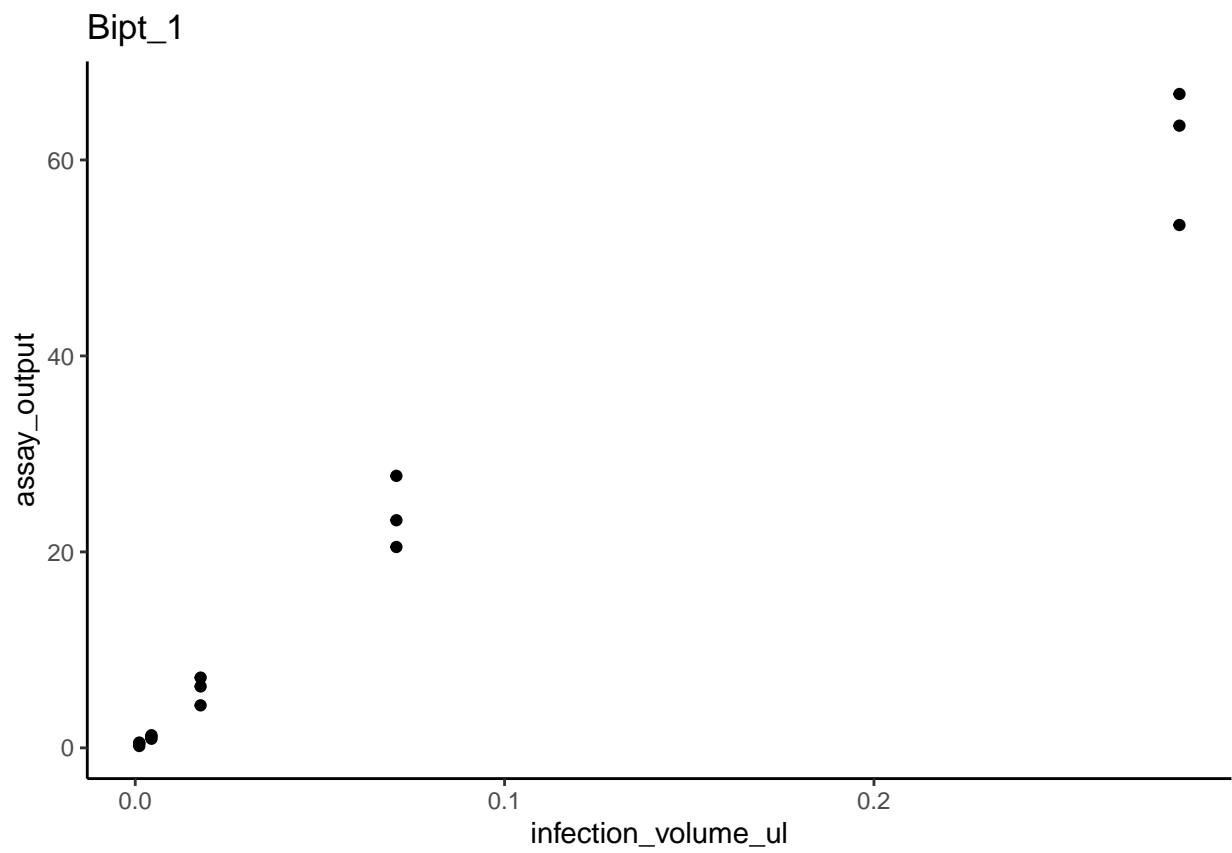
```
##  
## [[5]]
```



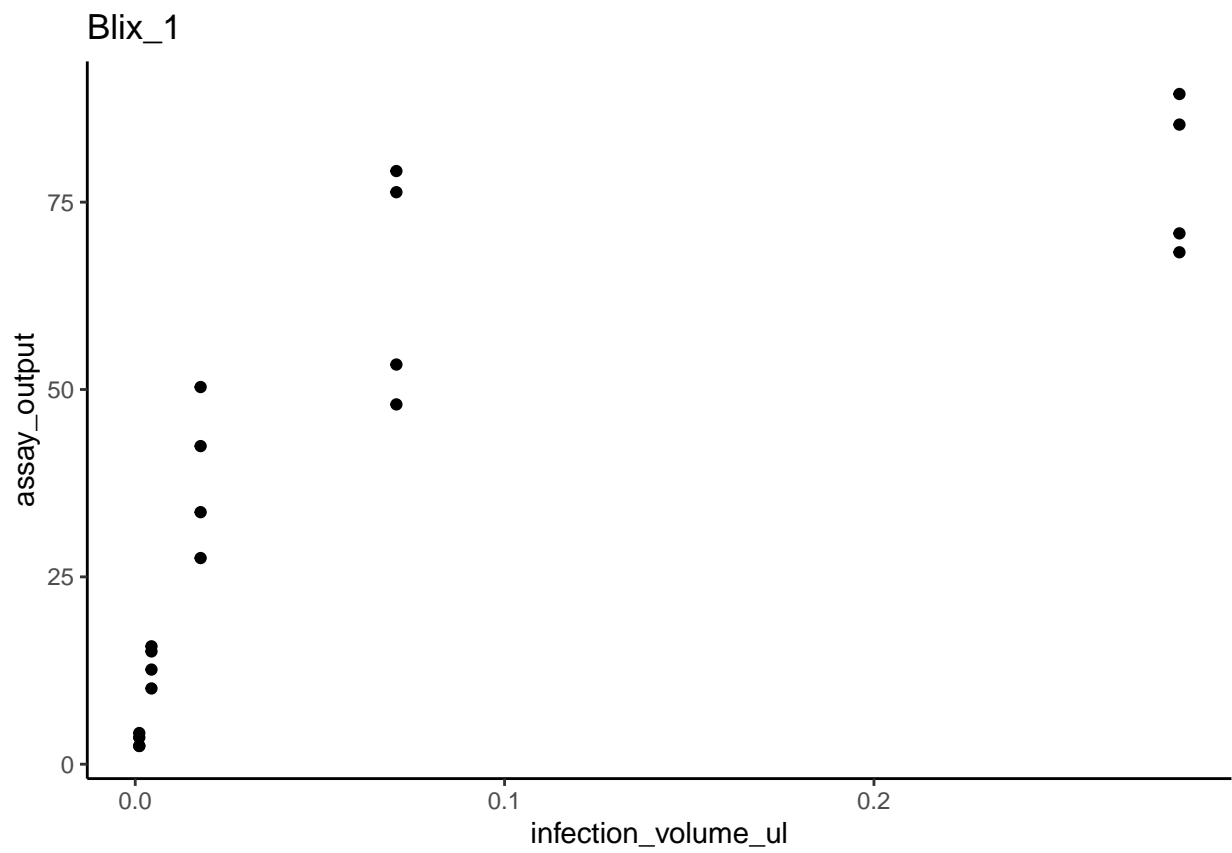
[[6]]



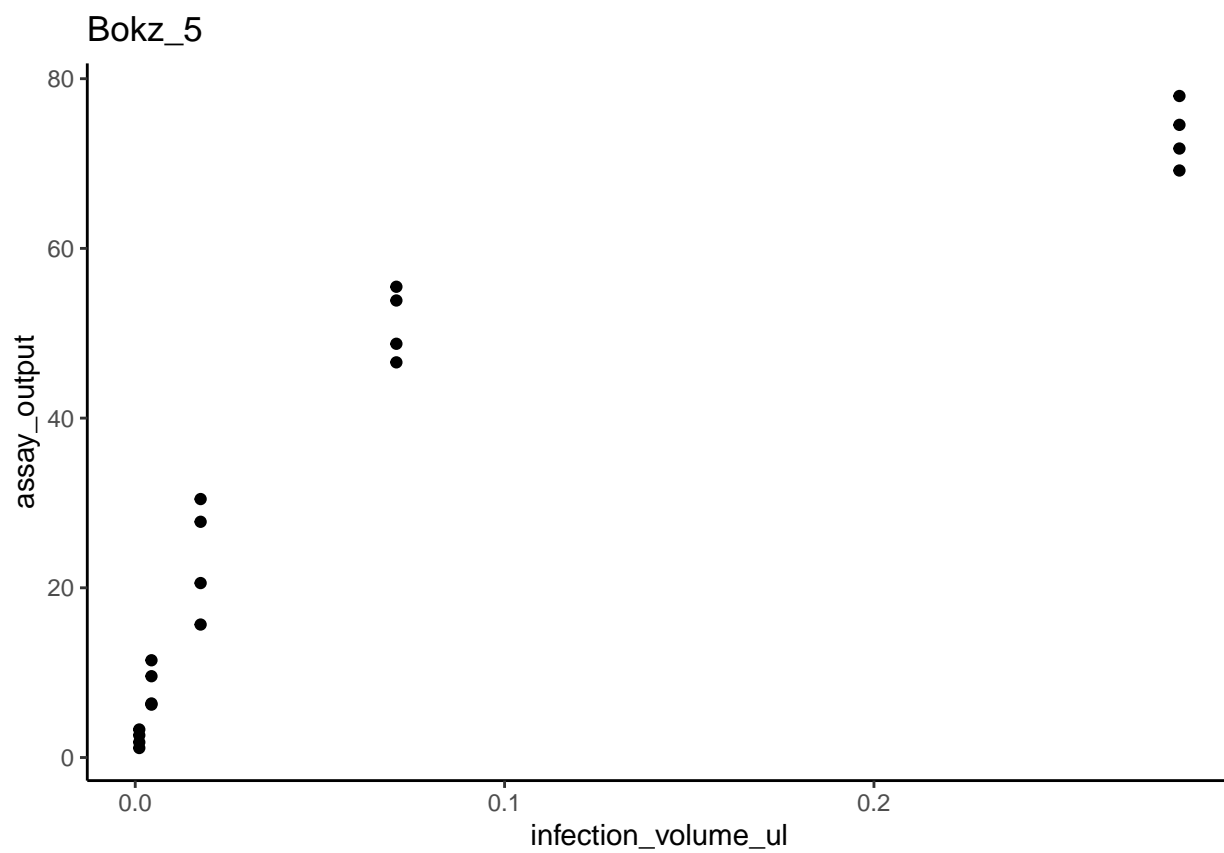
[[7]]



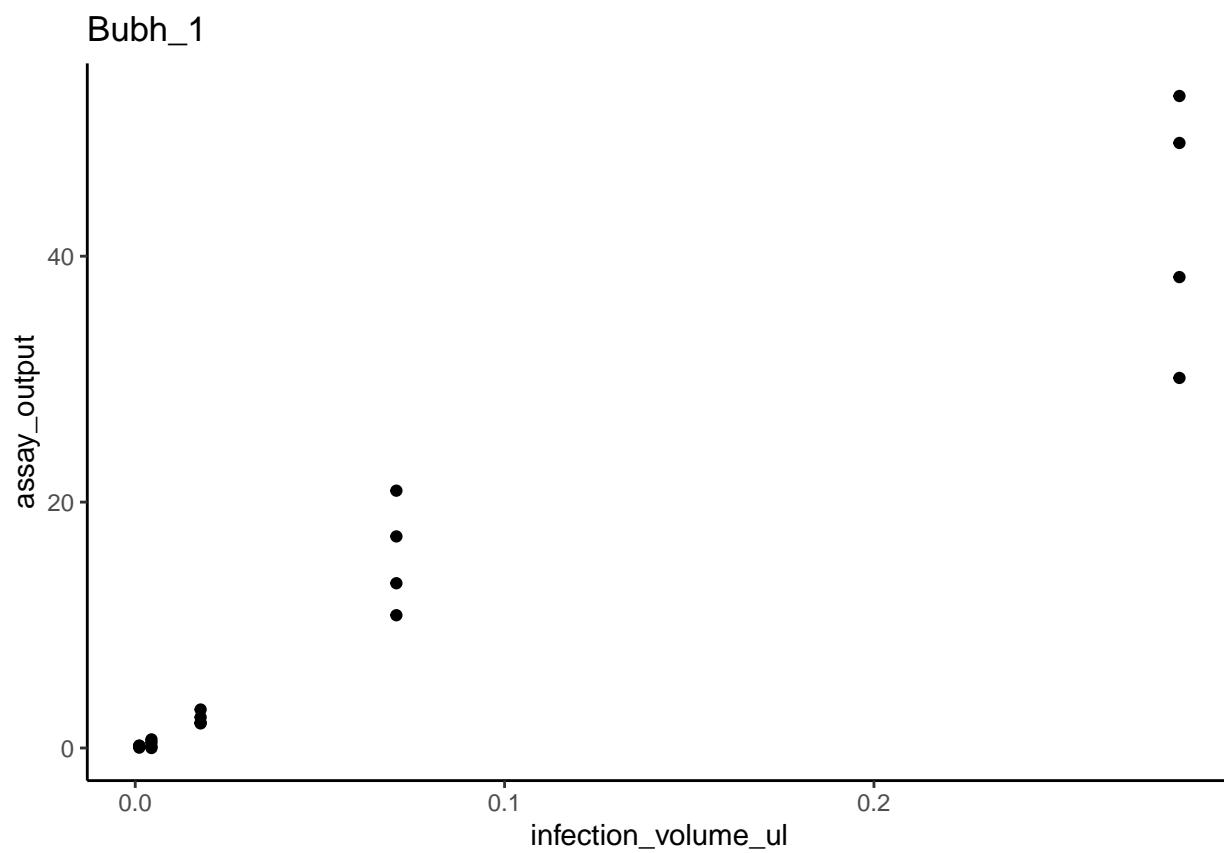
```
##  
## [[8]]
```



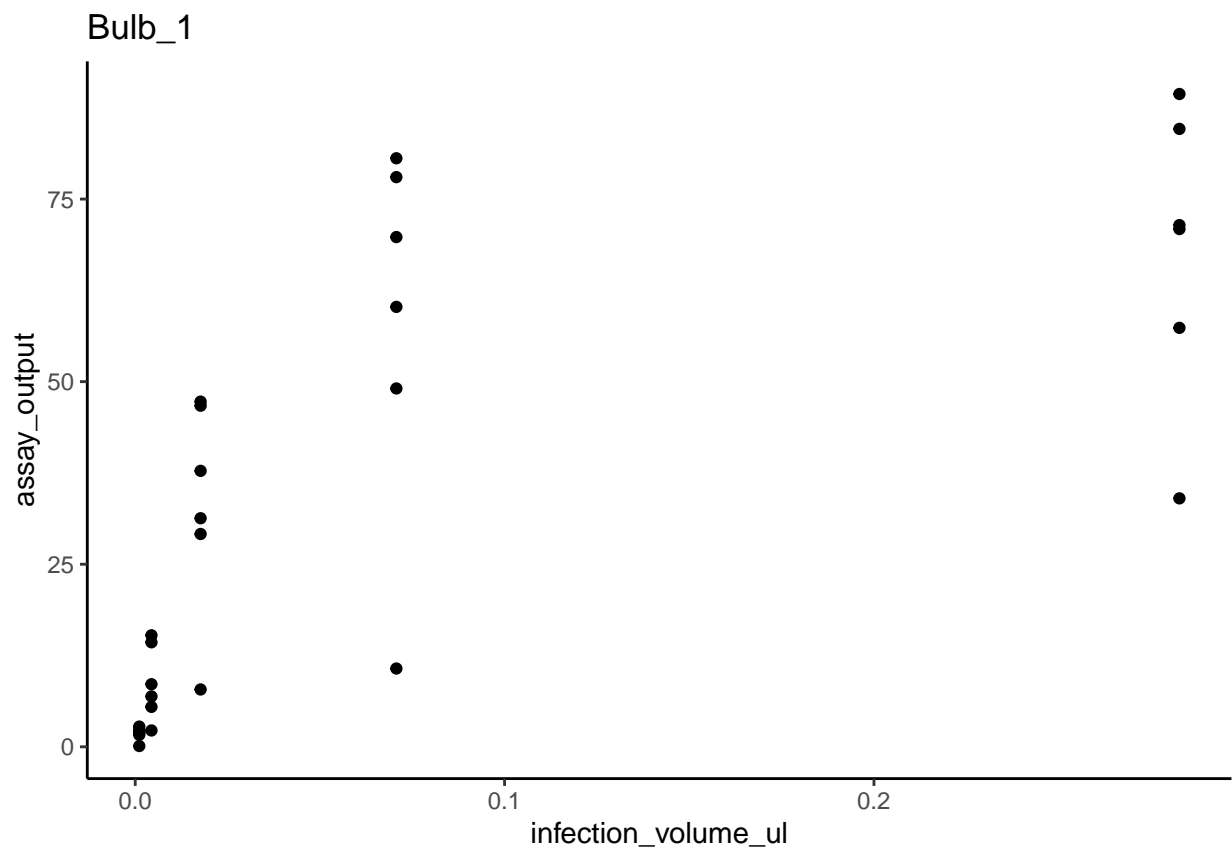
[[9]]



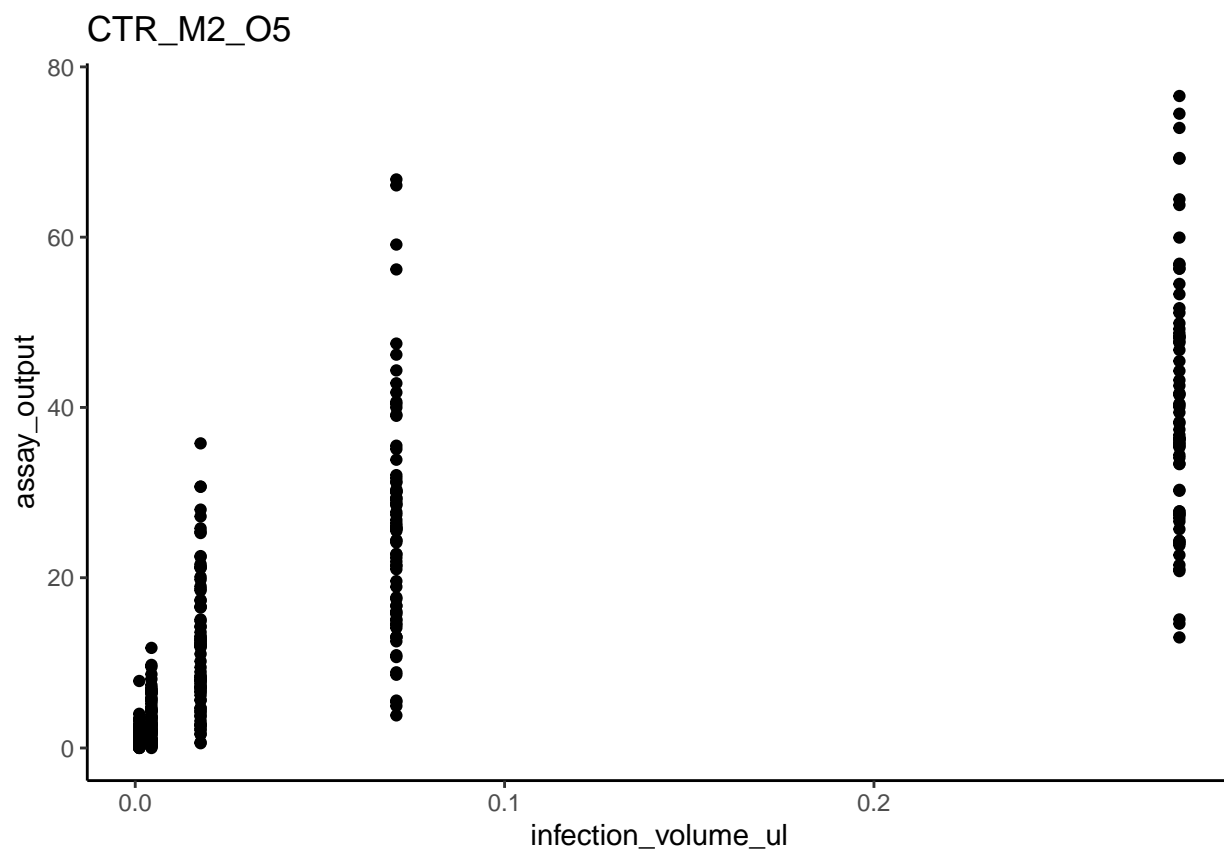
[[10]]



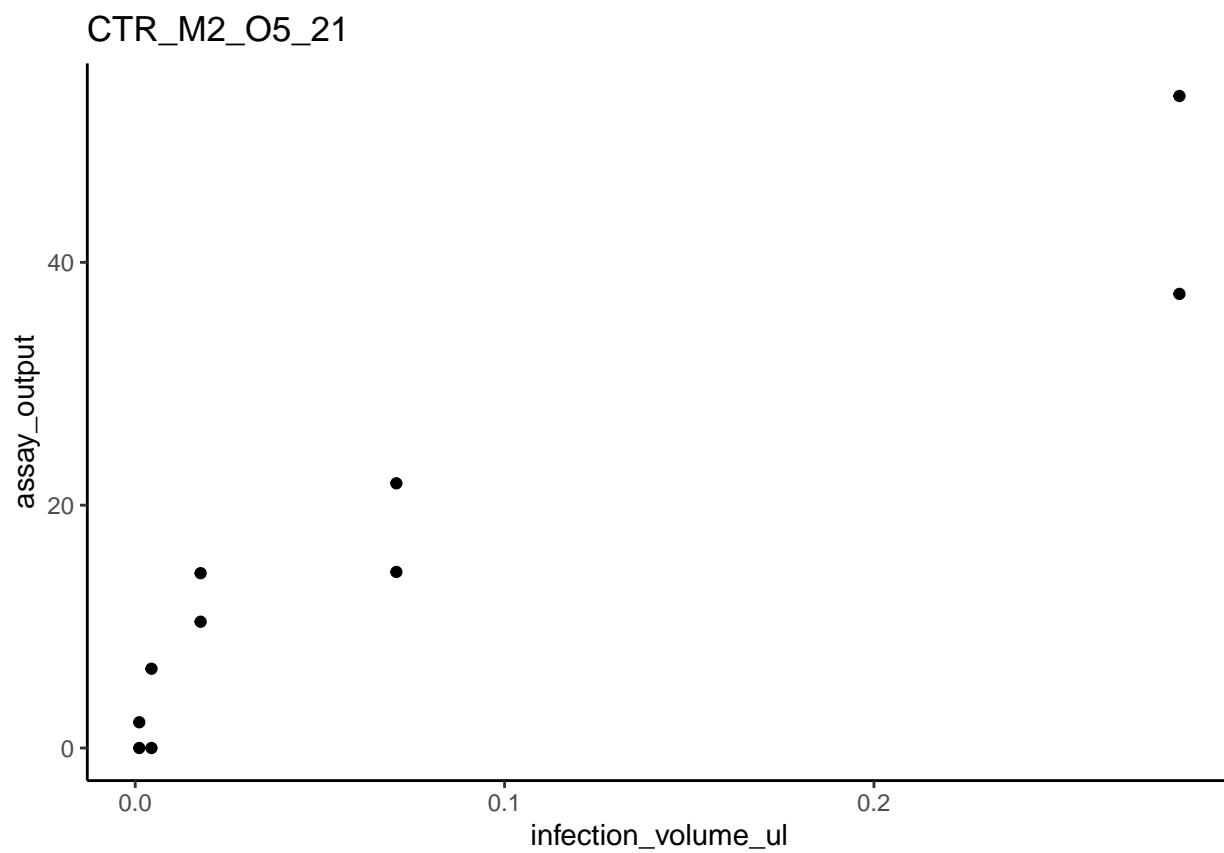
```
##  
## [[11]]
```



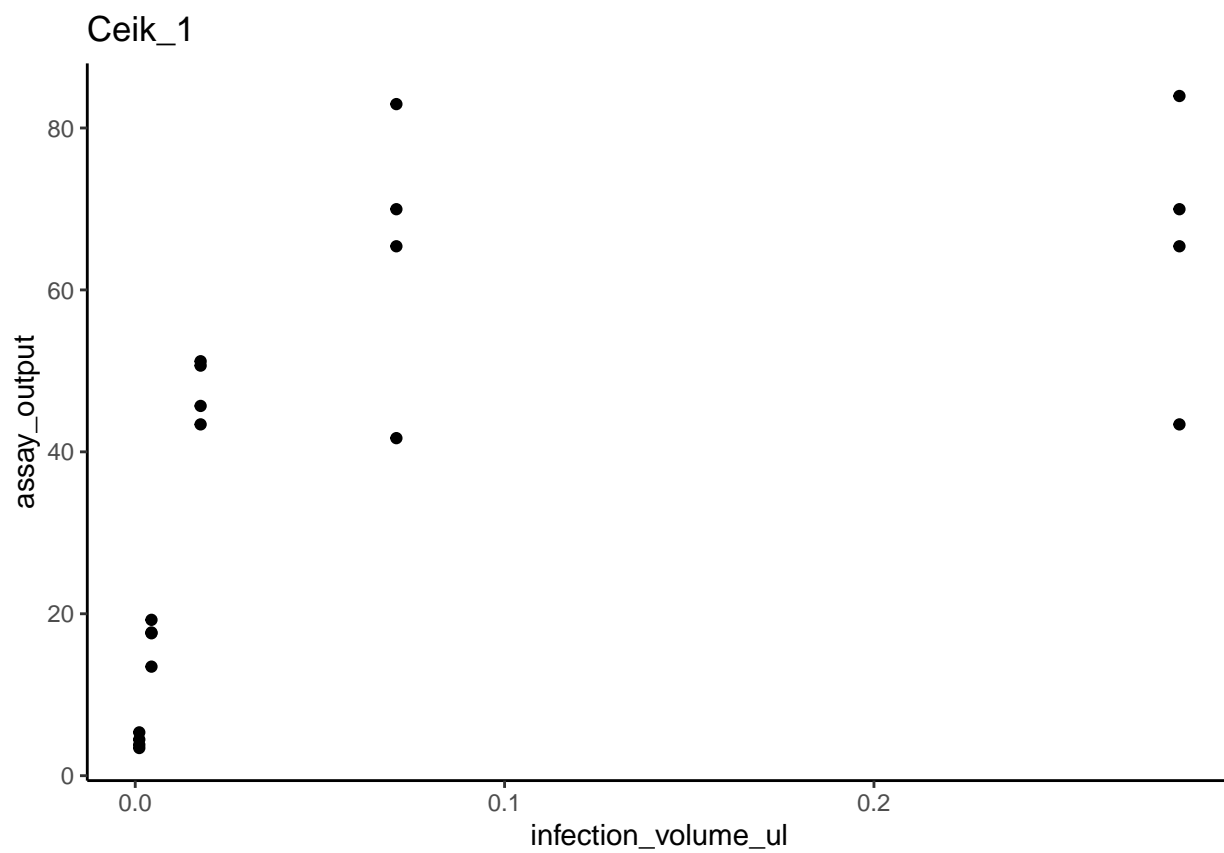
[[12]]



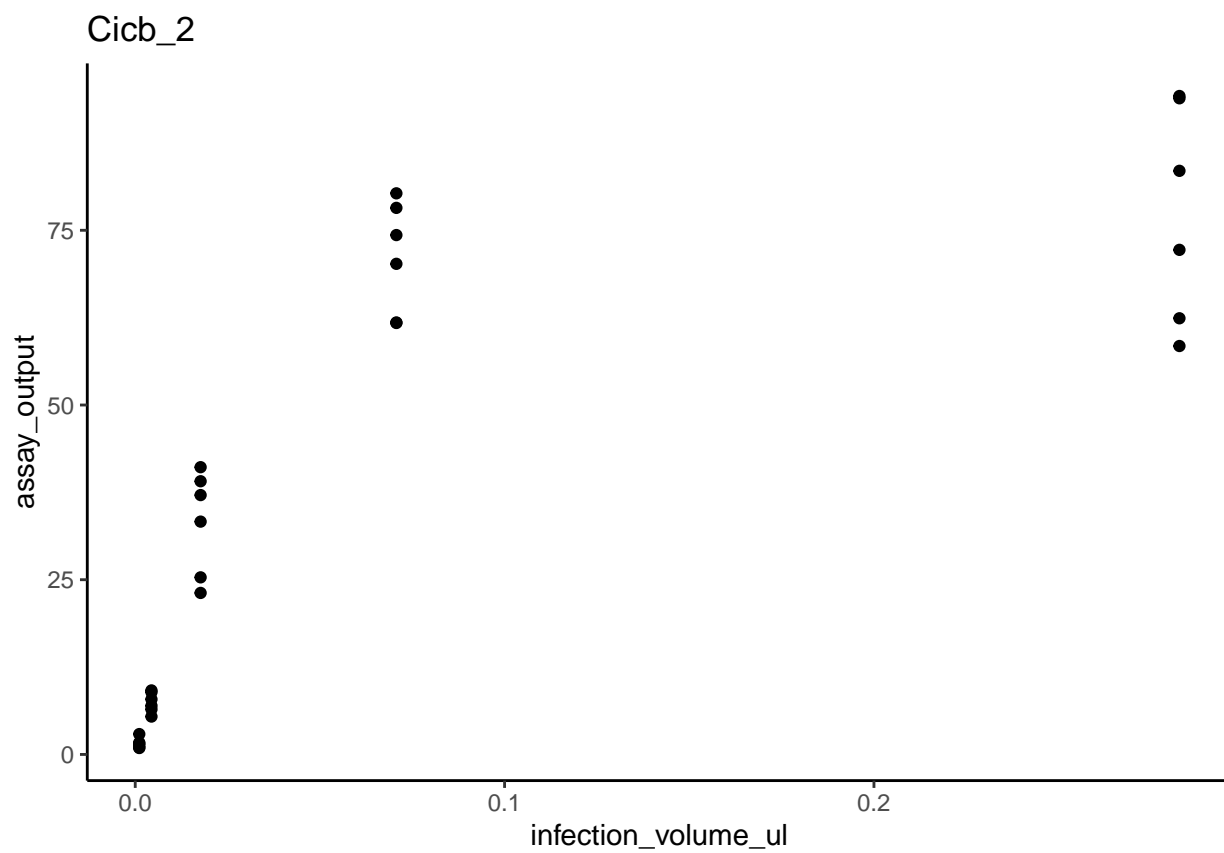
```
##  
## [[13]]
```



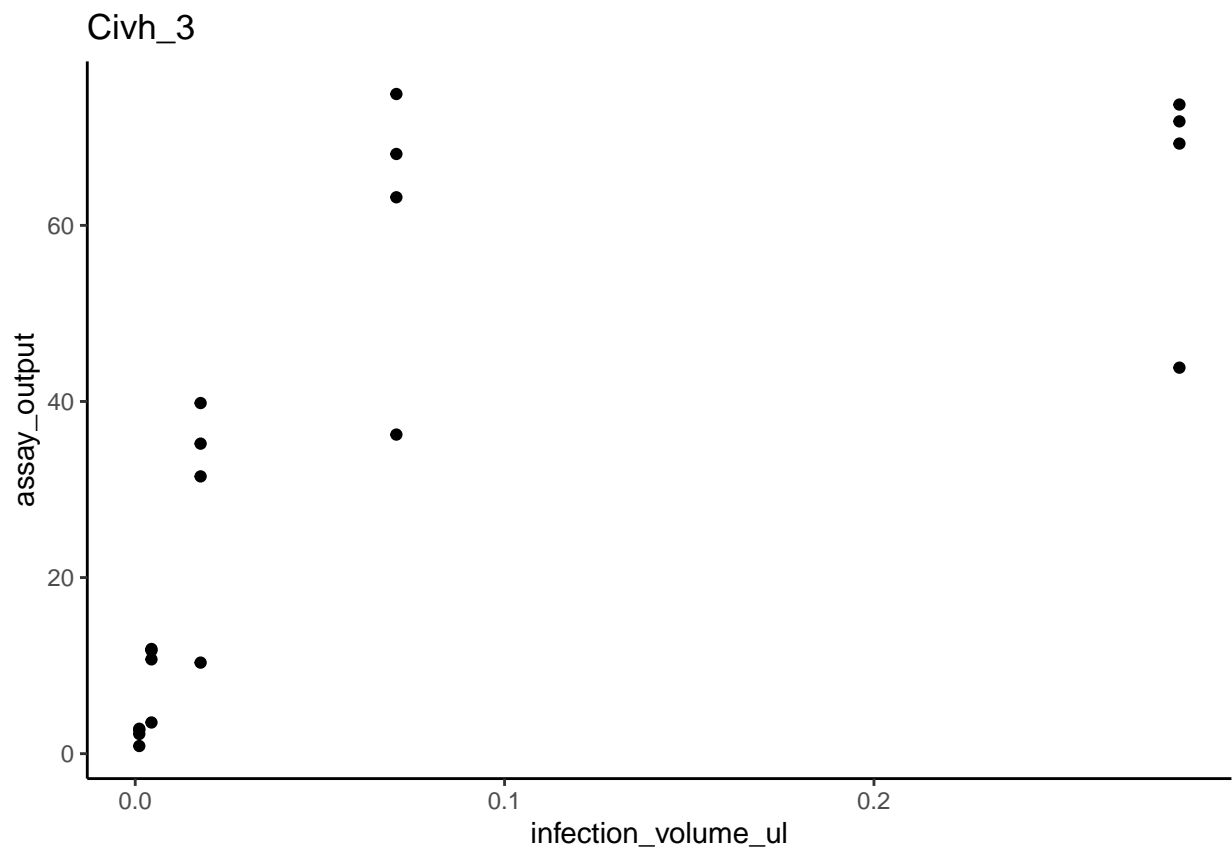
```
##  
## [[14]]
```



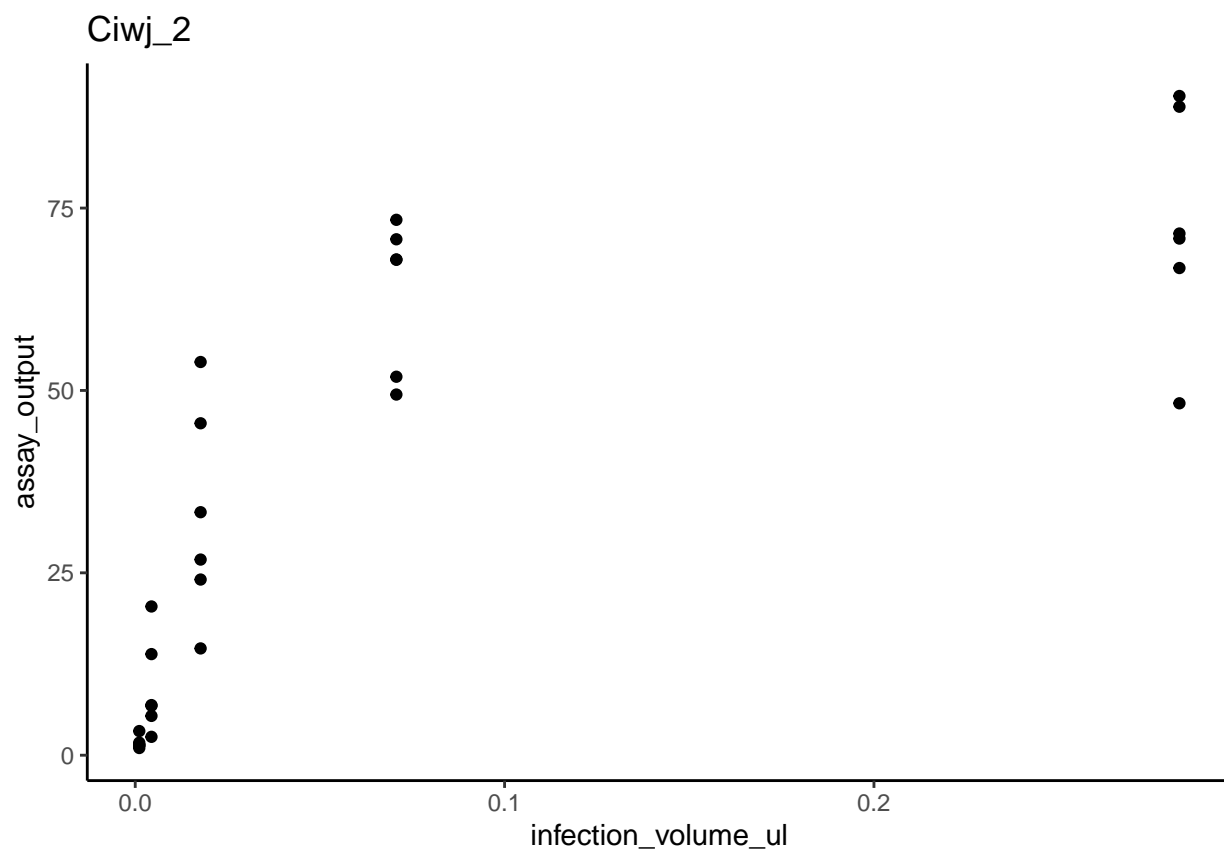
[[15]]



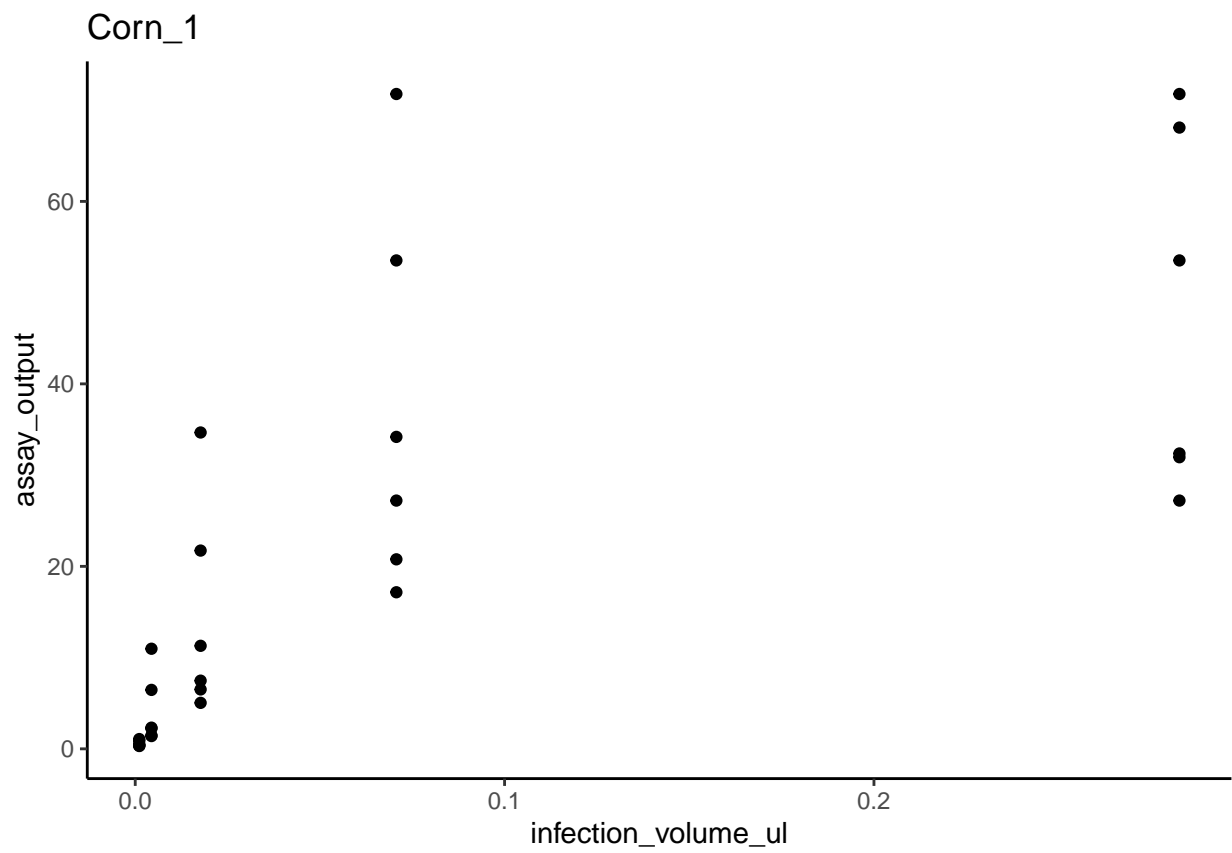
[[16]]



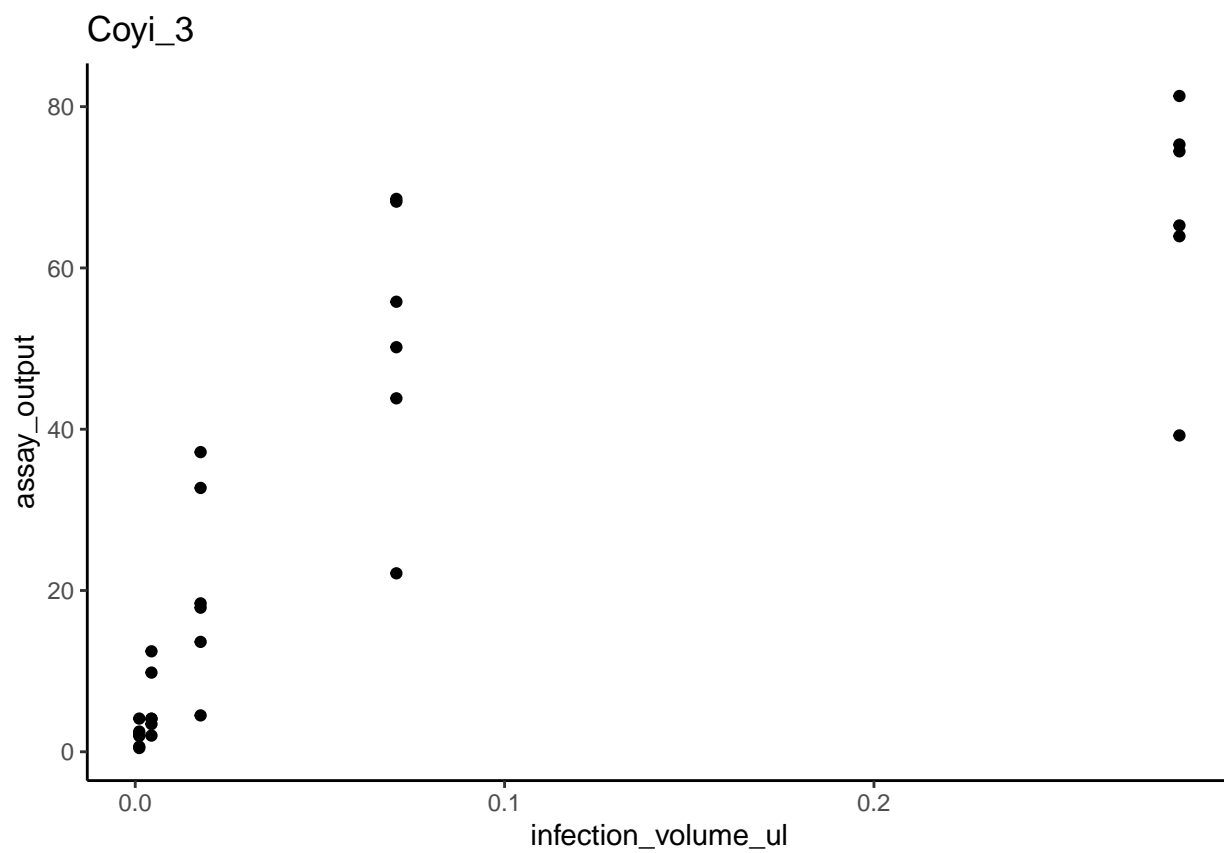
```
##  
## [[17]]
```



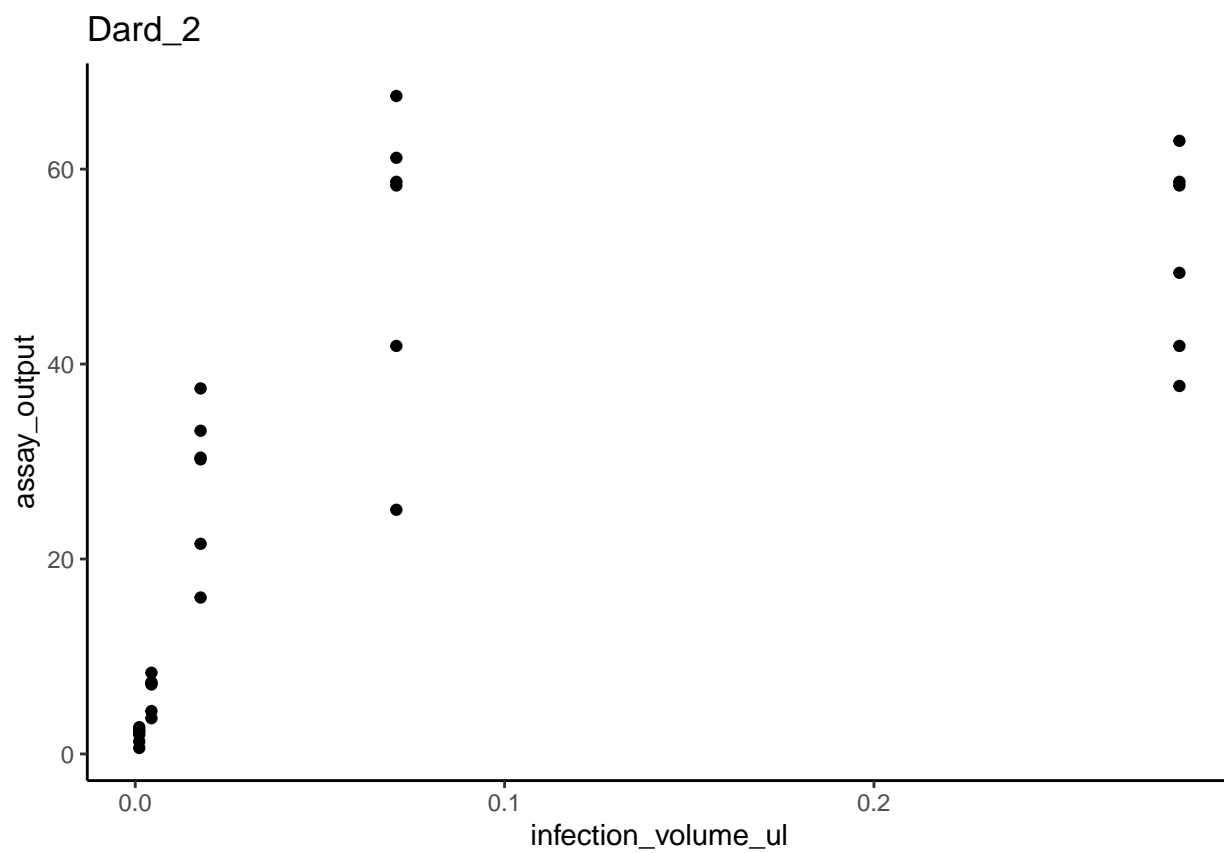
[[18]]



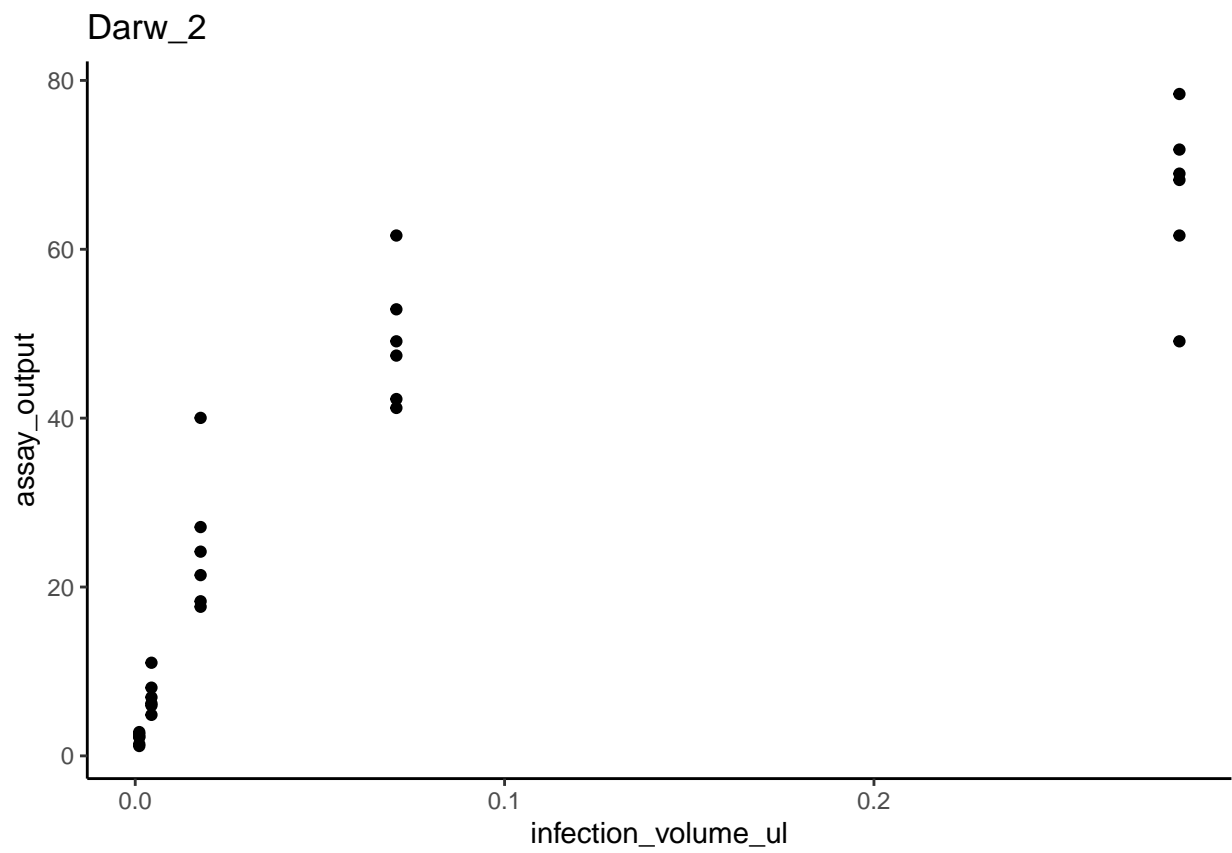
[[19]]



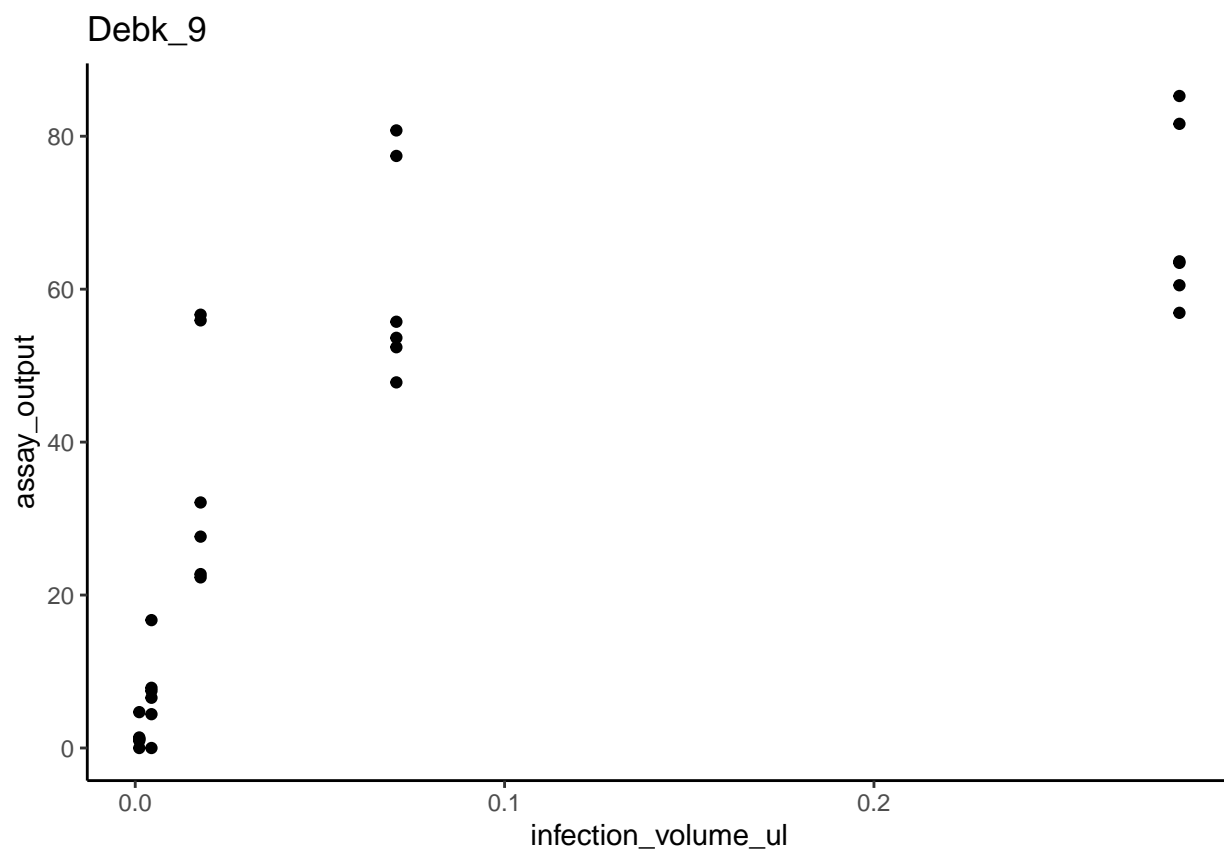
[[20]]



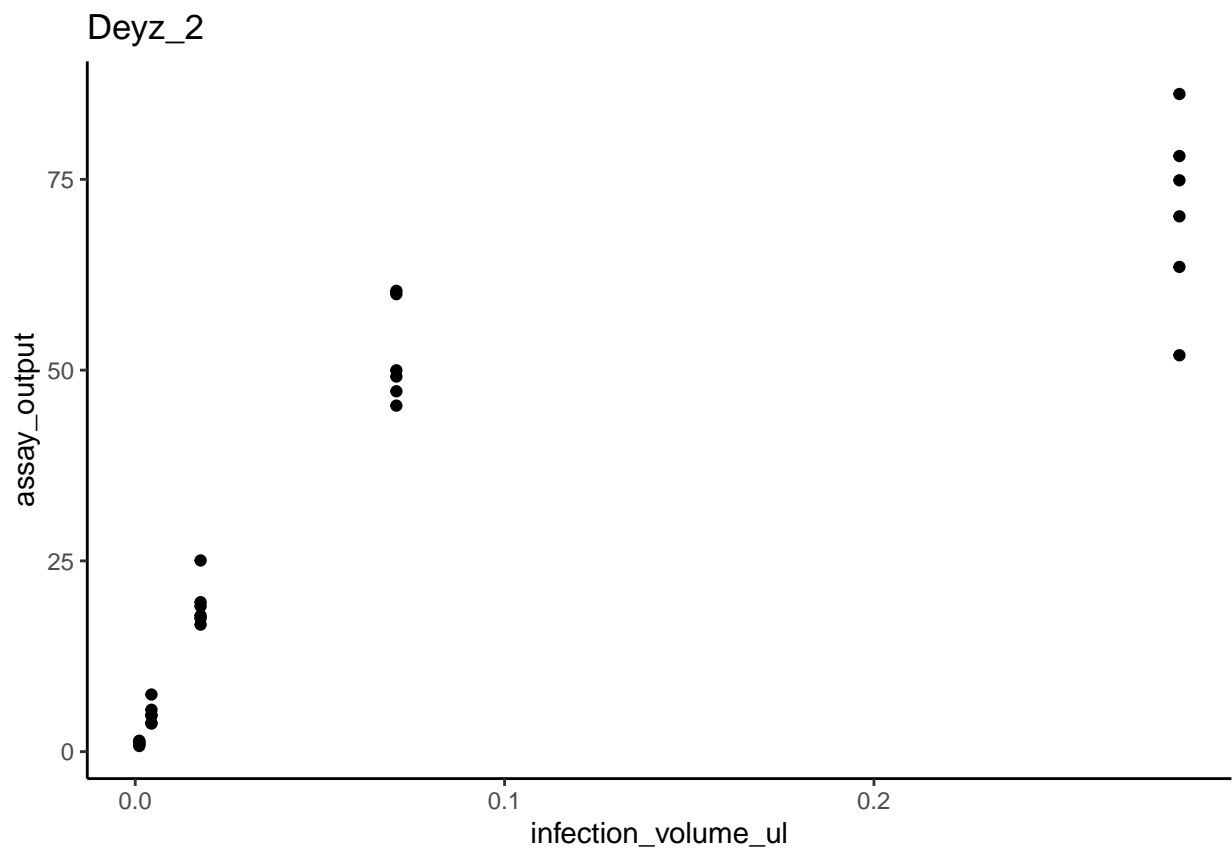
[[21]]



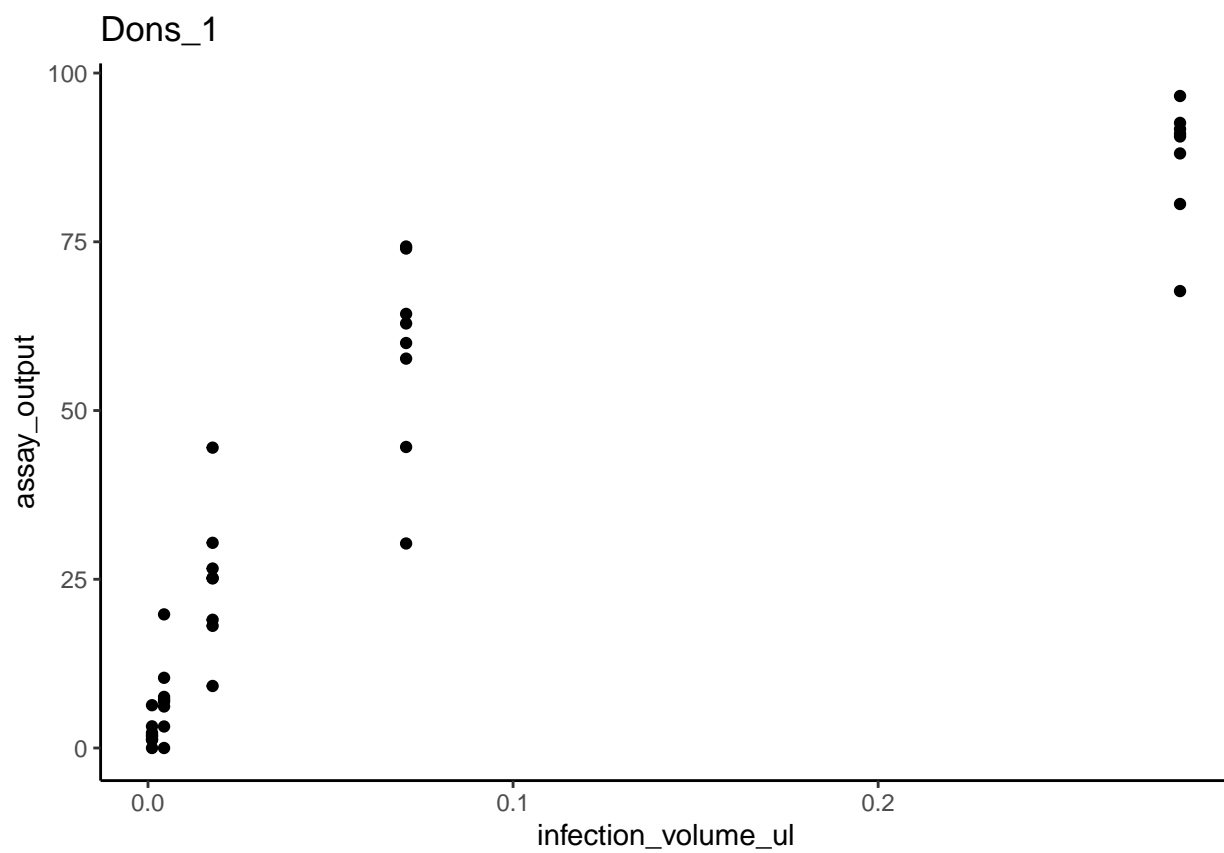
[[22]]



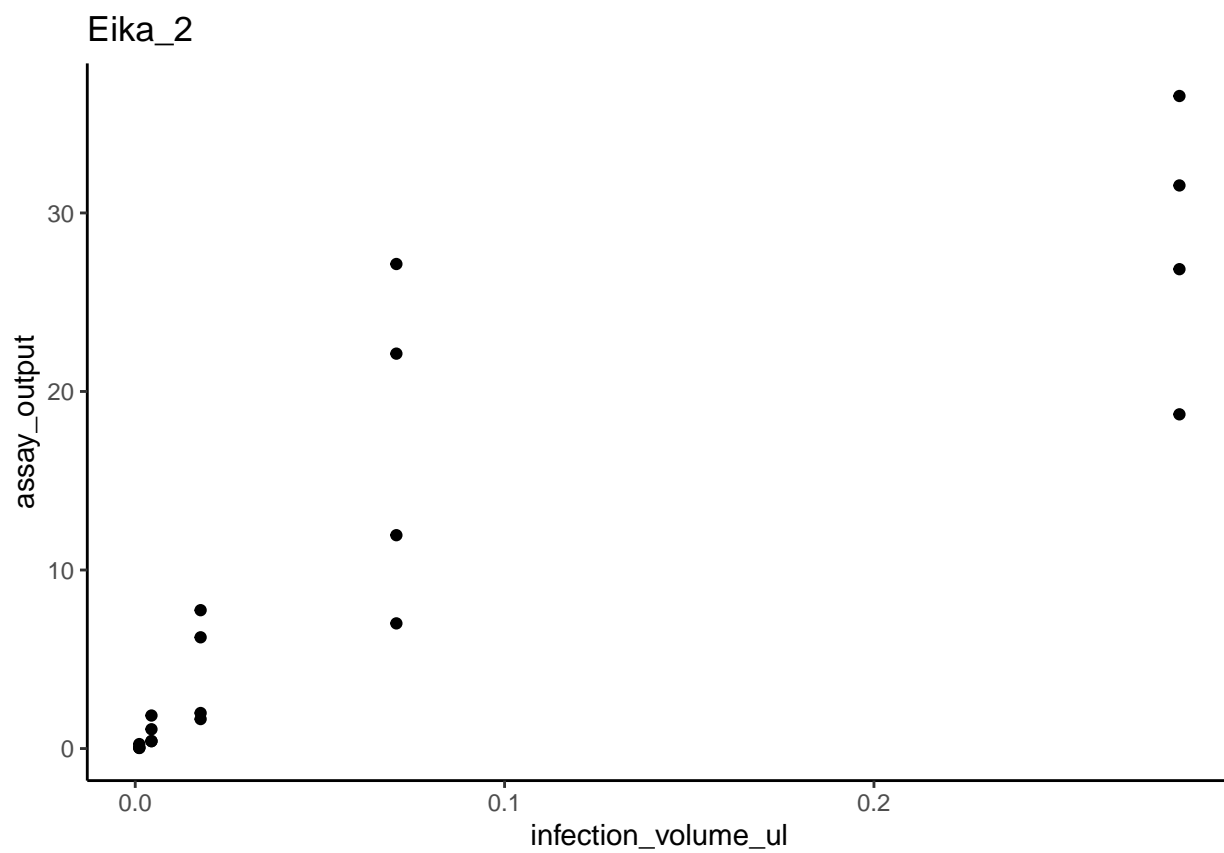
[[23]]



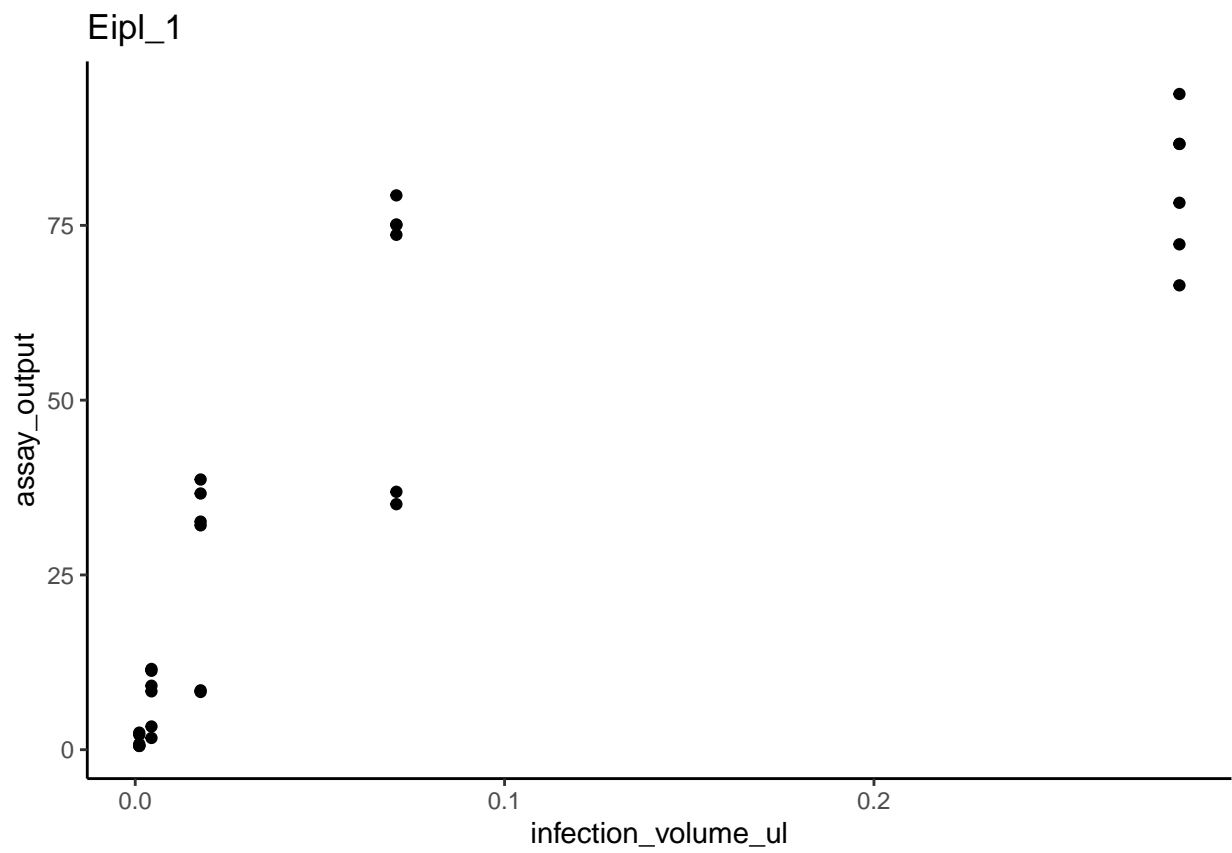
```
##  
## [[24]]
```



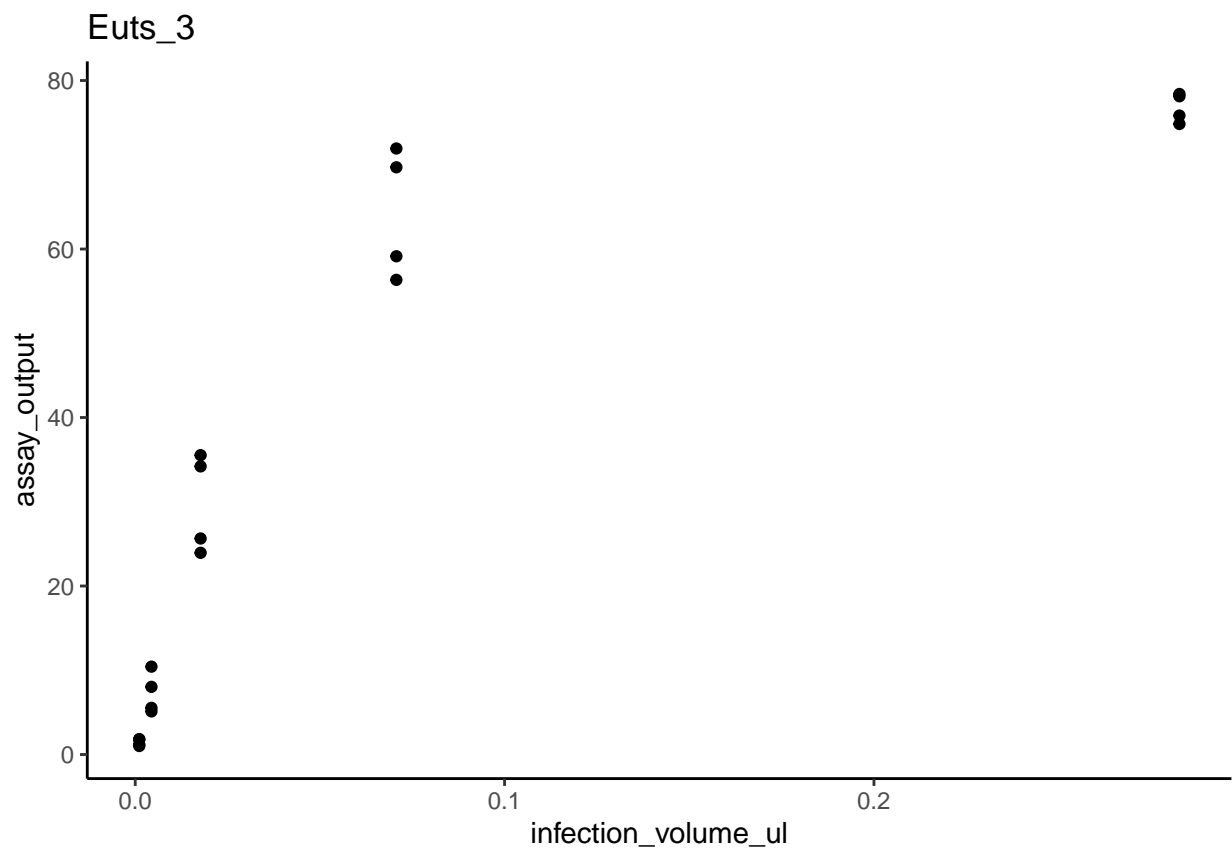
[[25]]



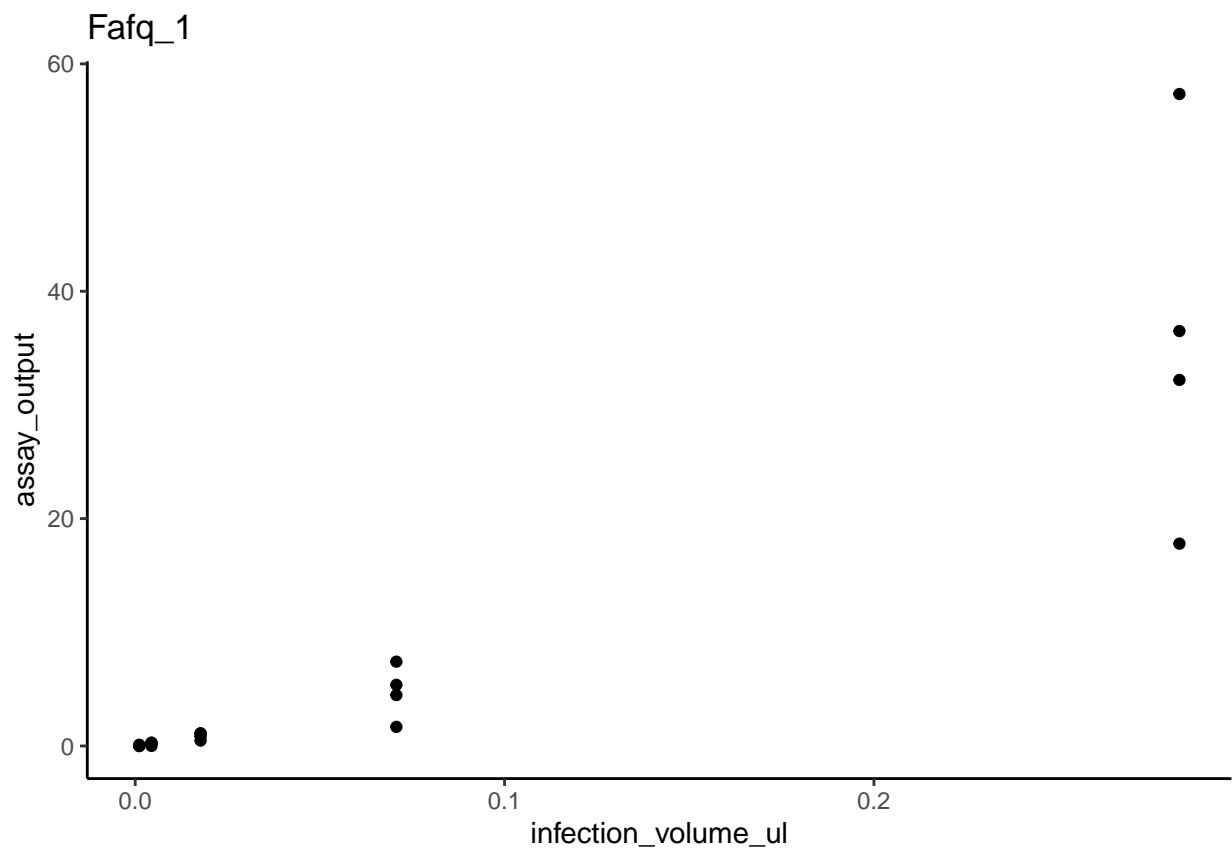
[[26]]



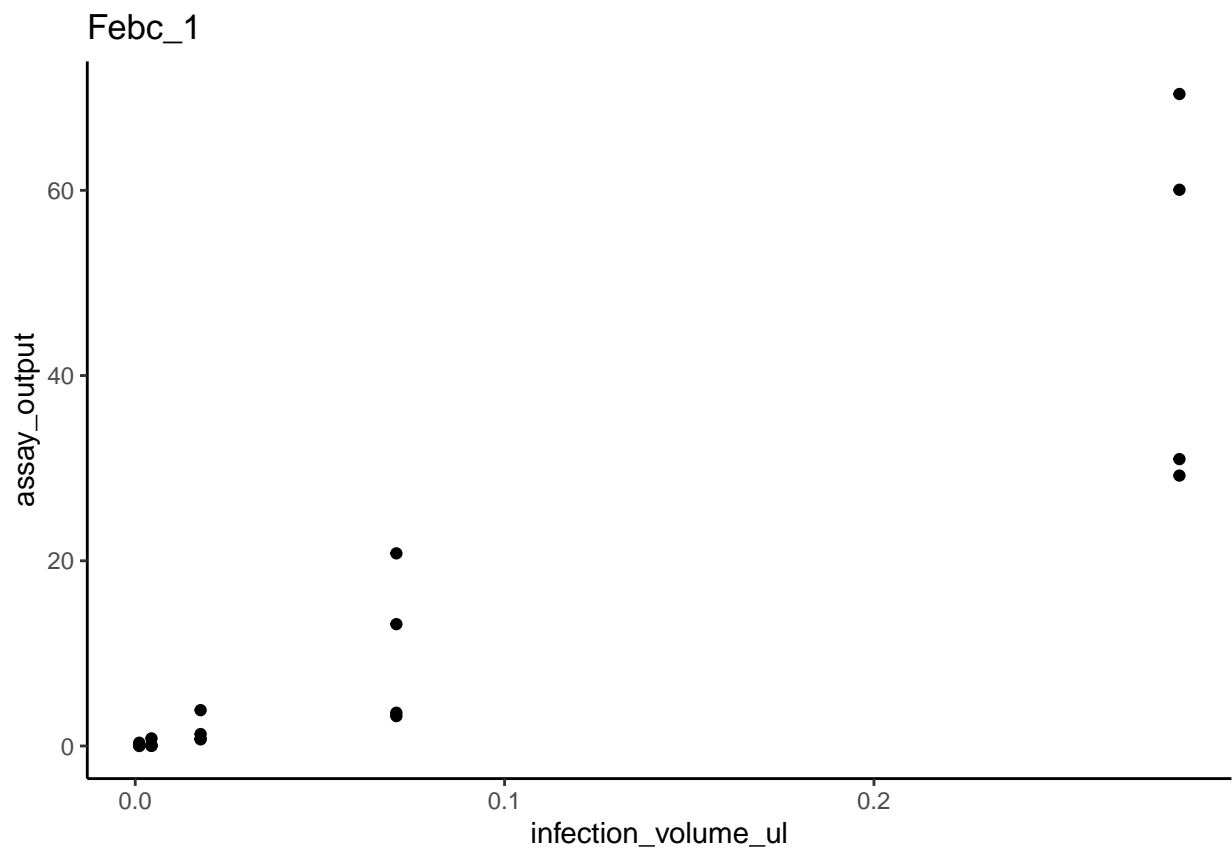
[[27]]



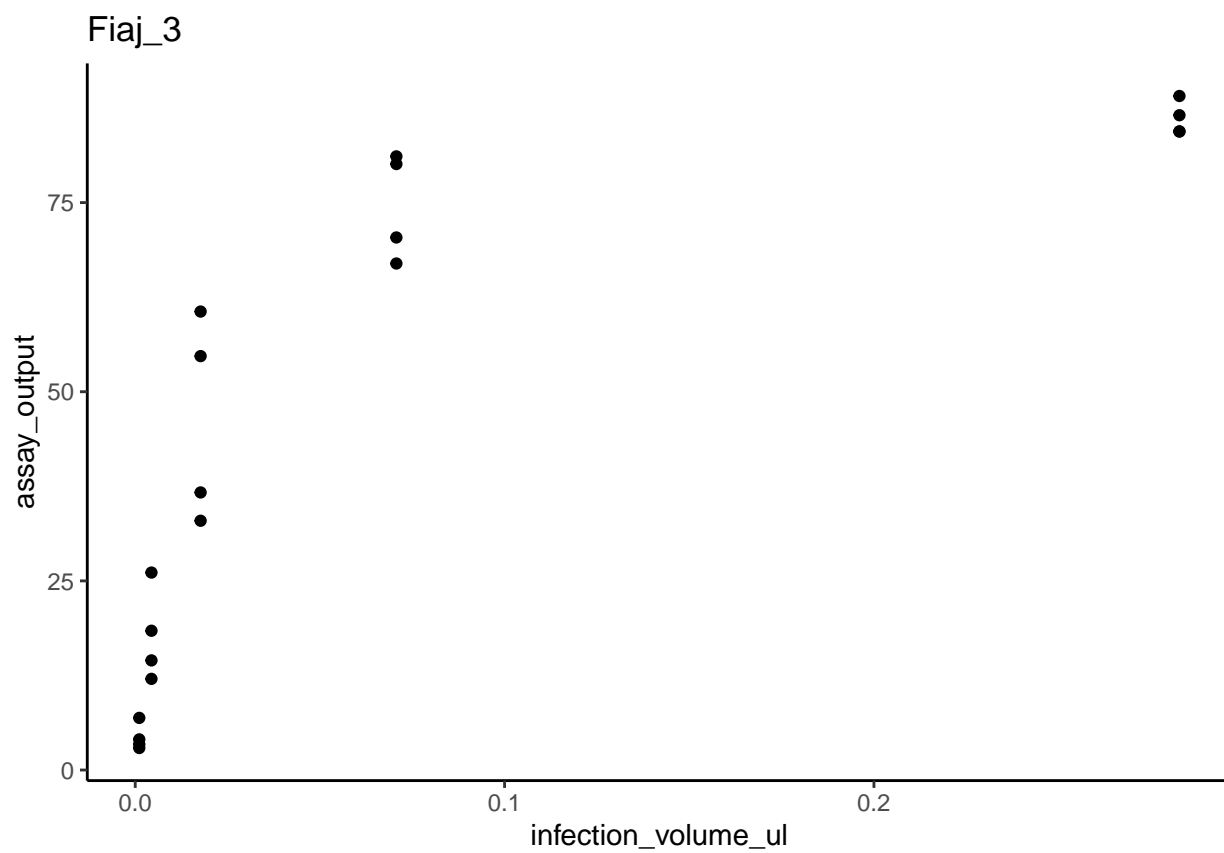
[[30]]



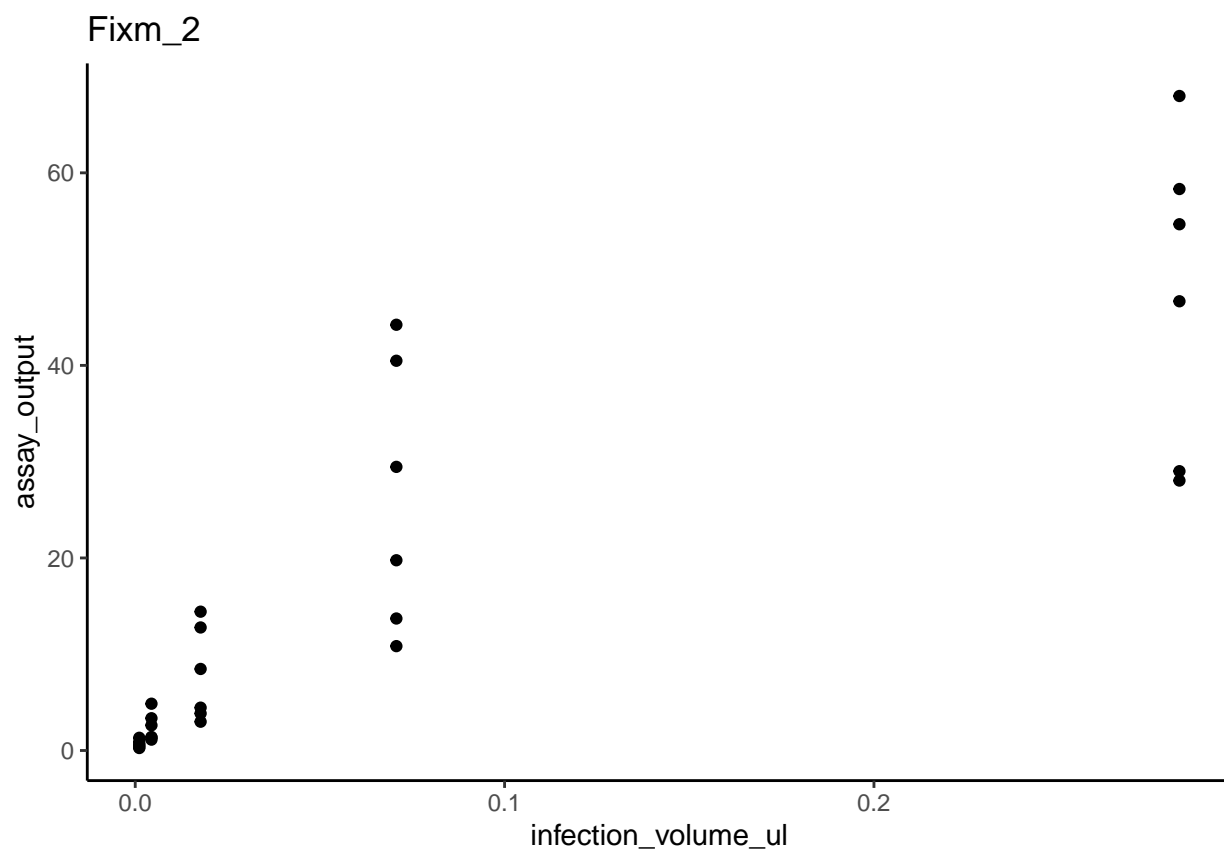
```
##  
## [[31]]
```

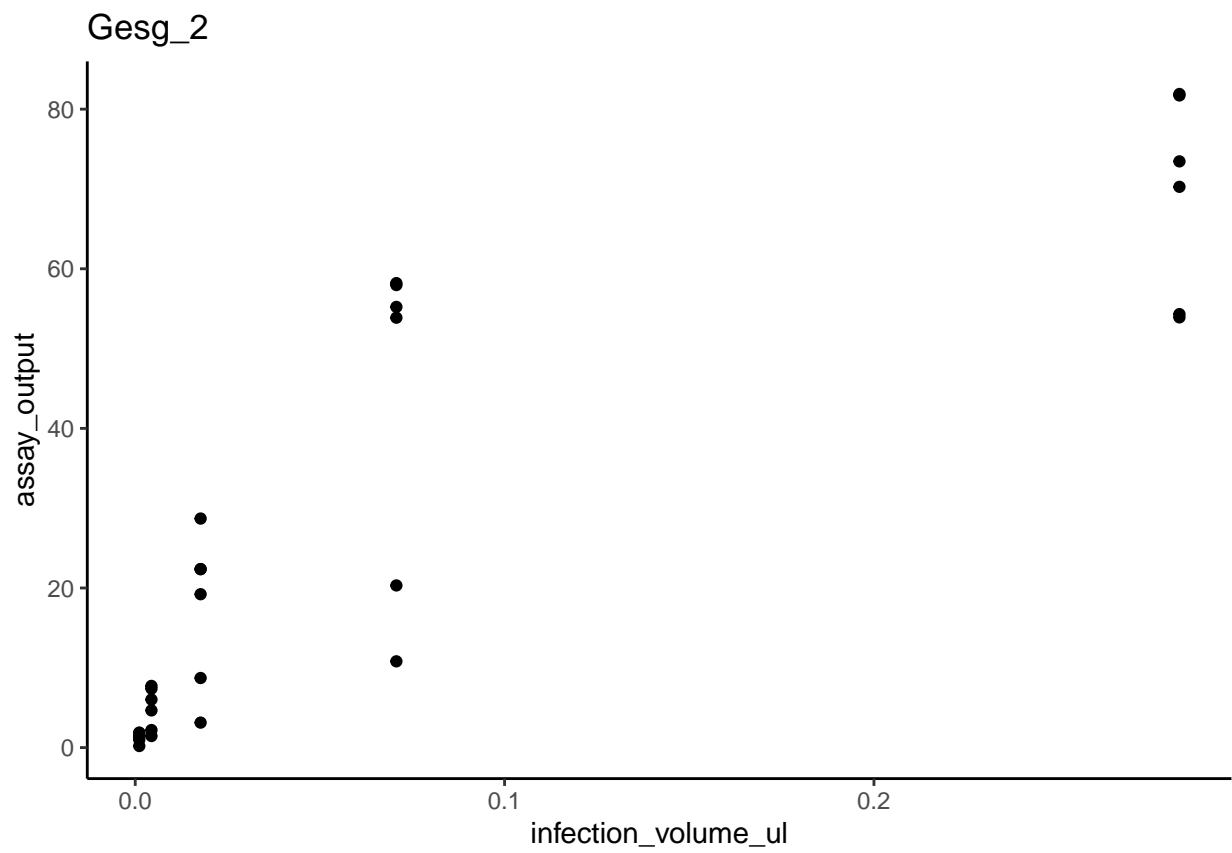
[[32]]



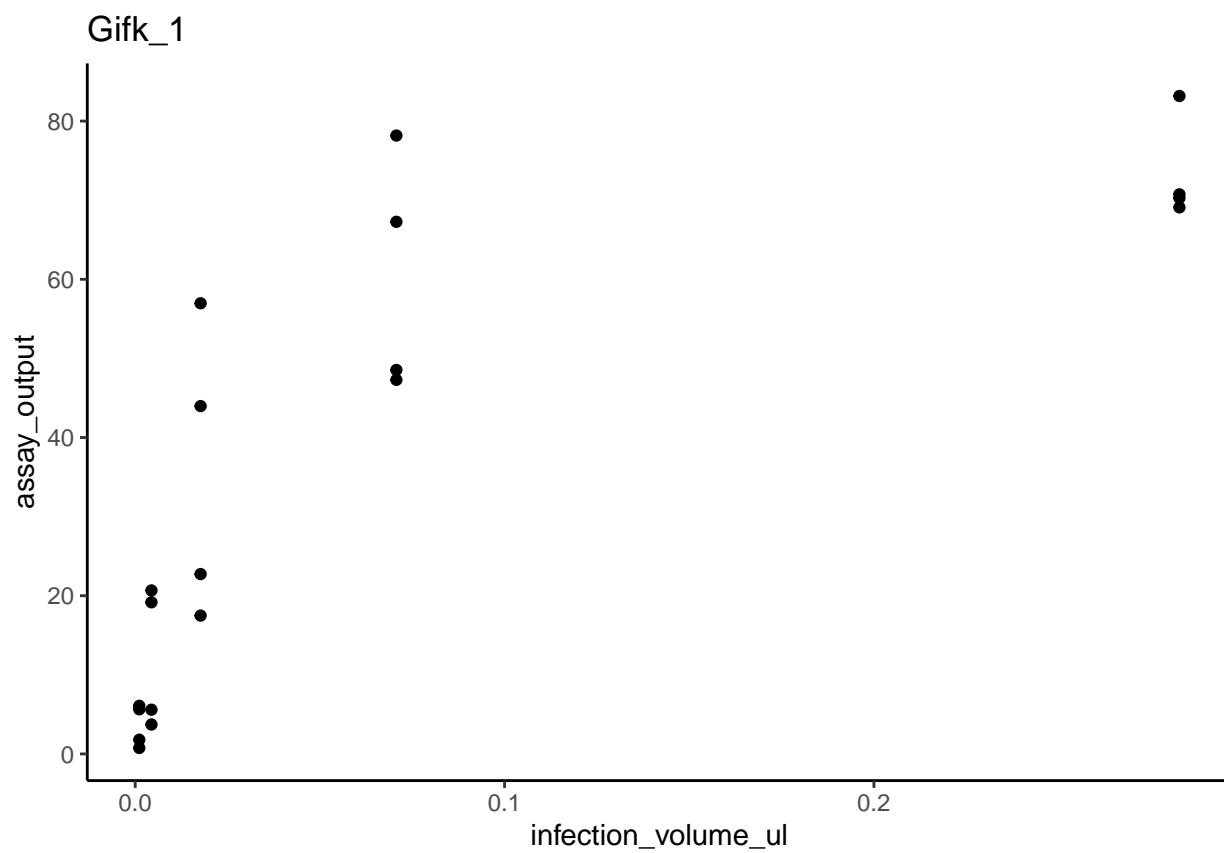
```
##  
## [[33]]
```



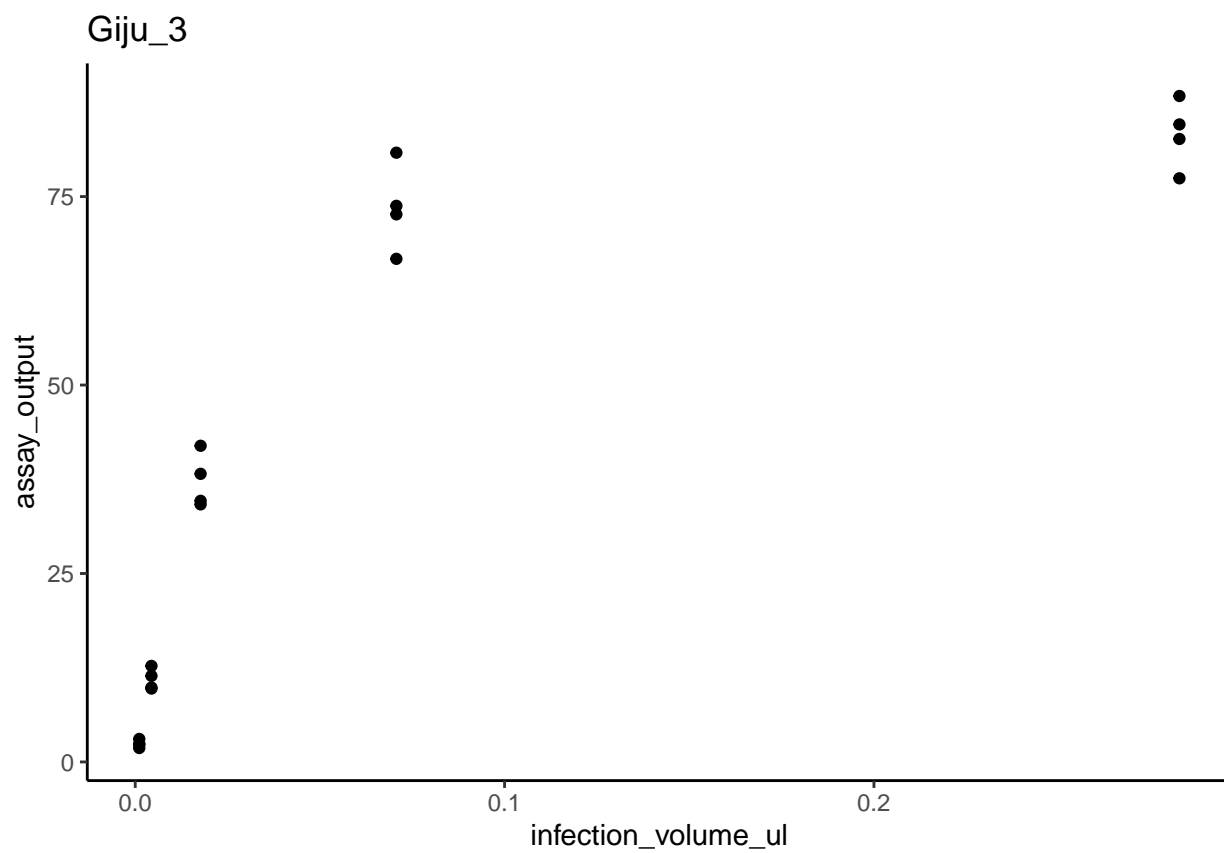
[[34]]



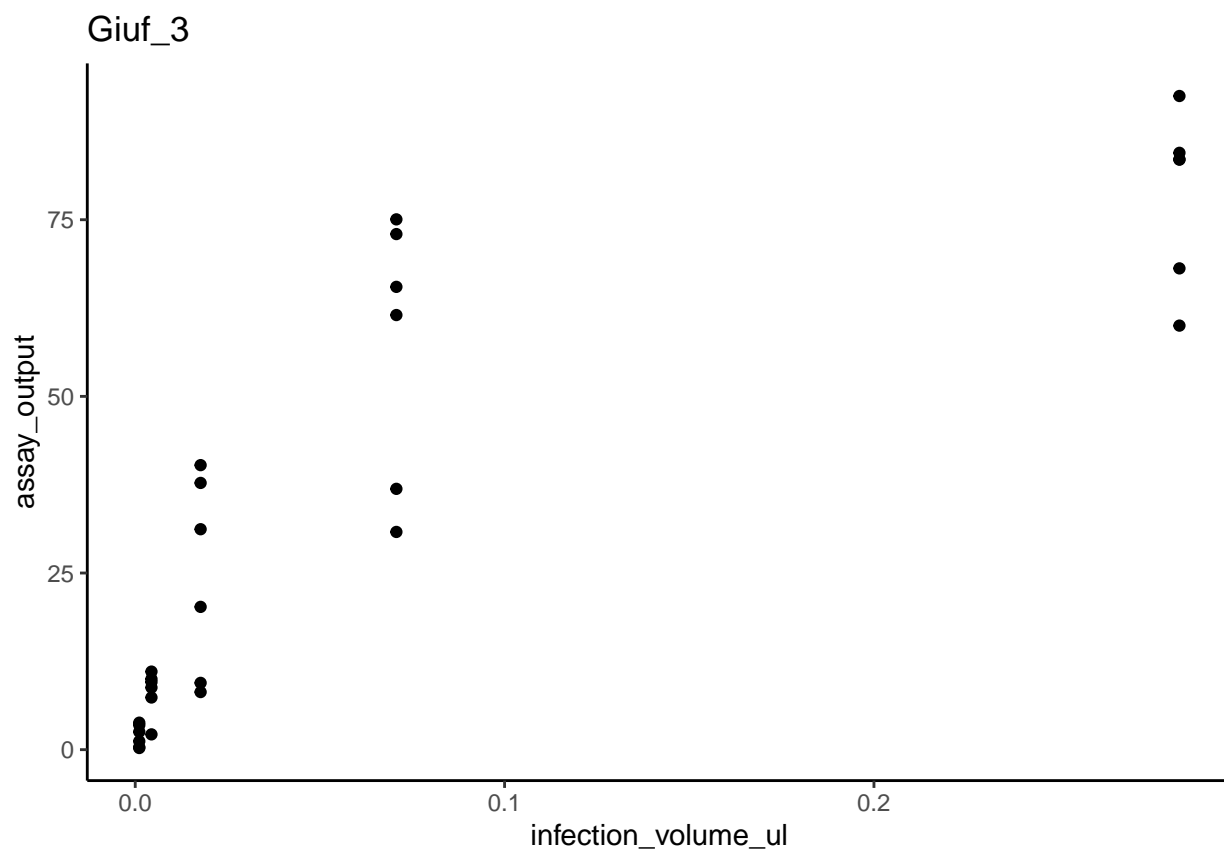
[[35]]



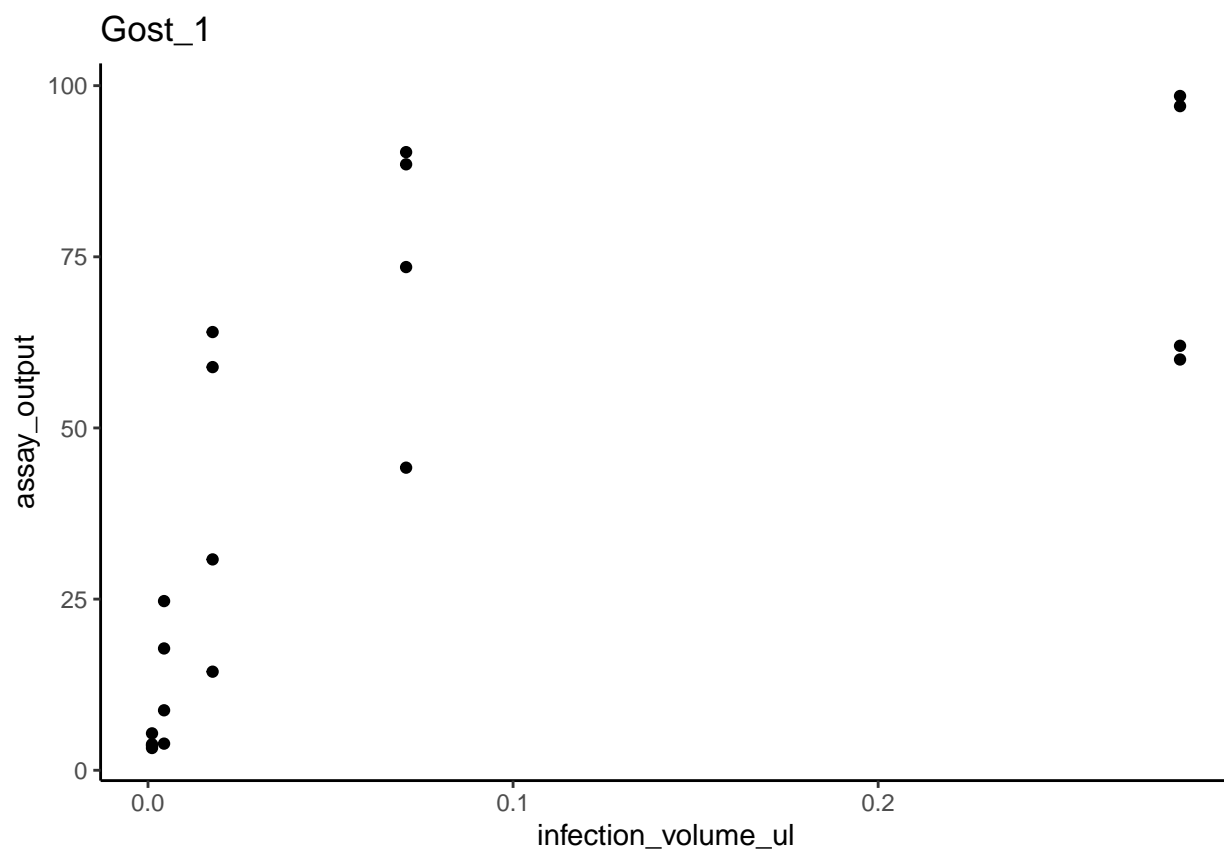
```
##  
## [[36]]
```



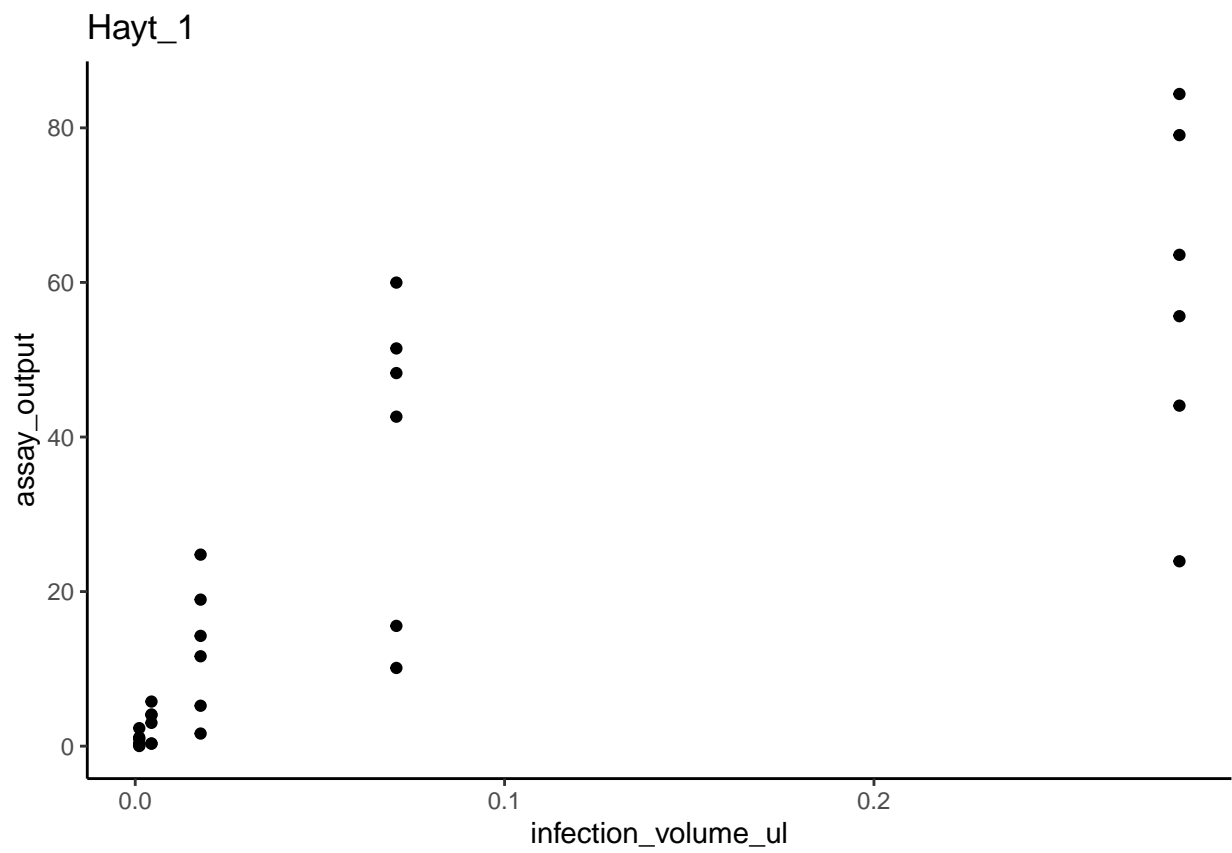
[[37]]



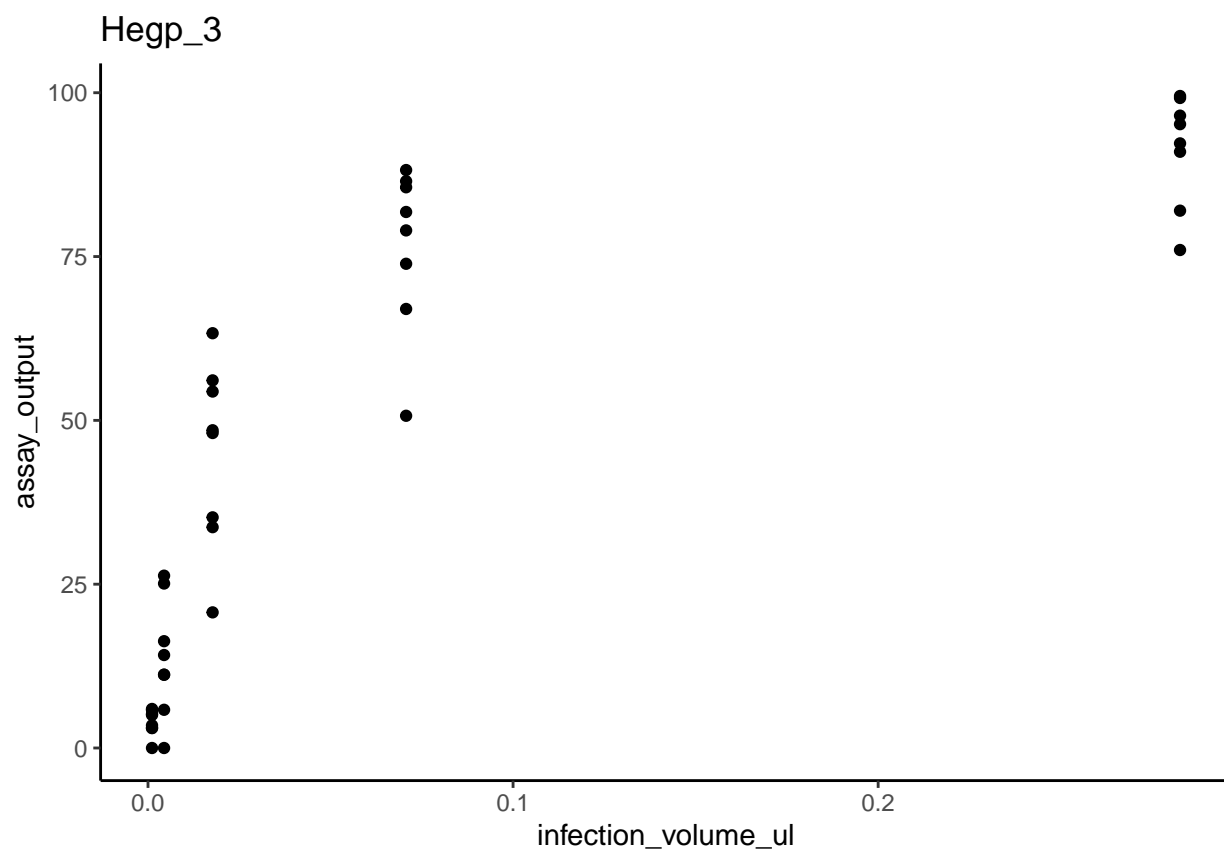
[[38]]



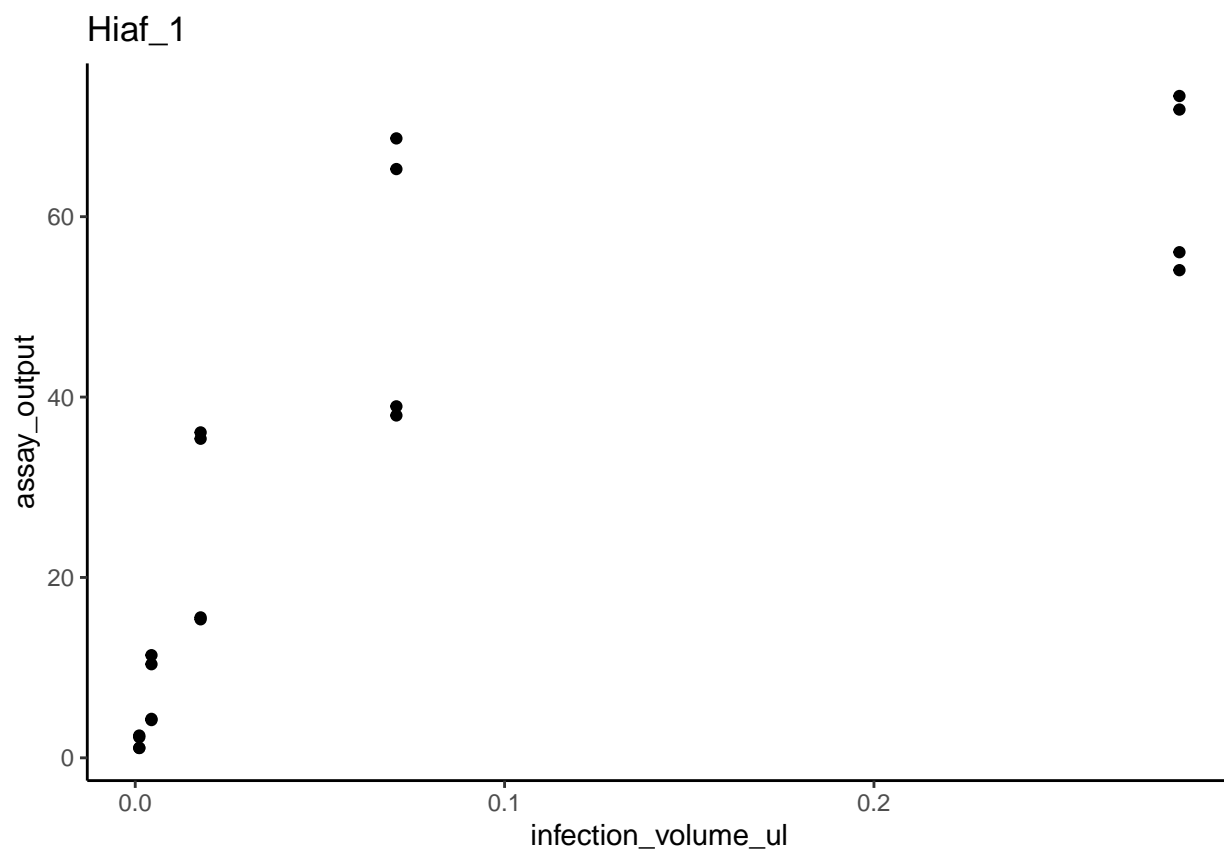
[[39]]



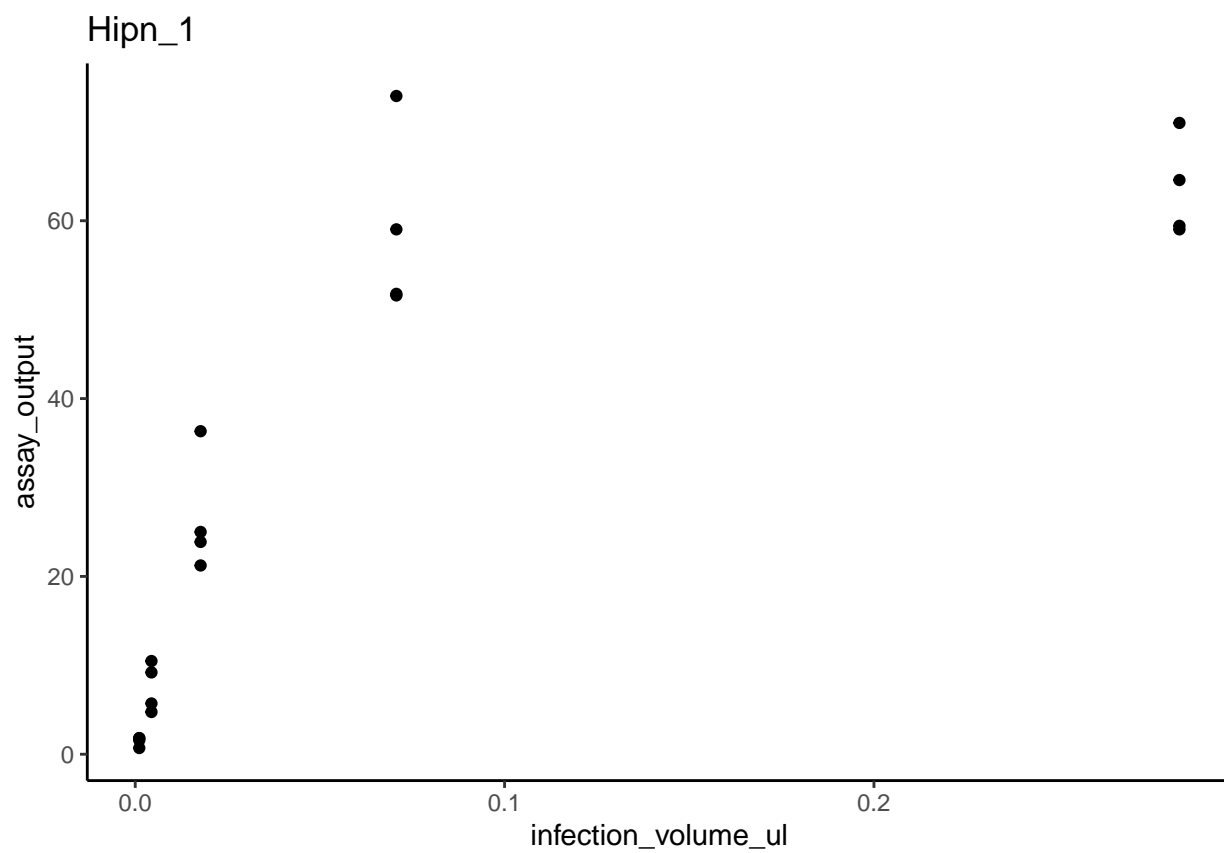
[[40]]



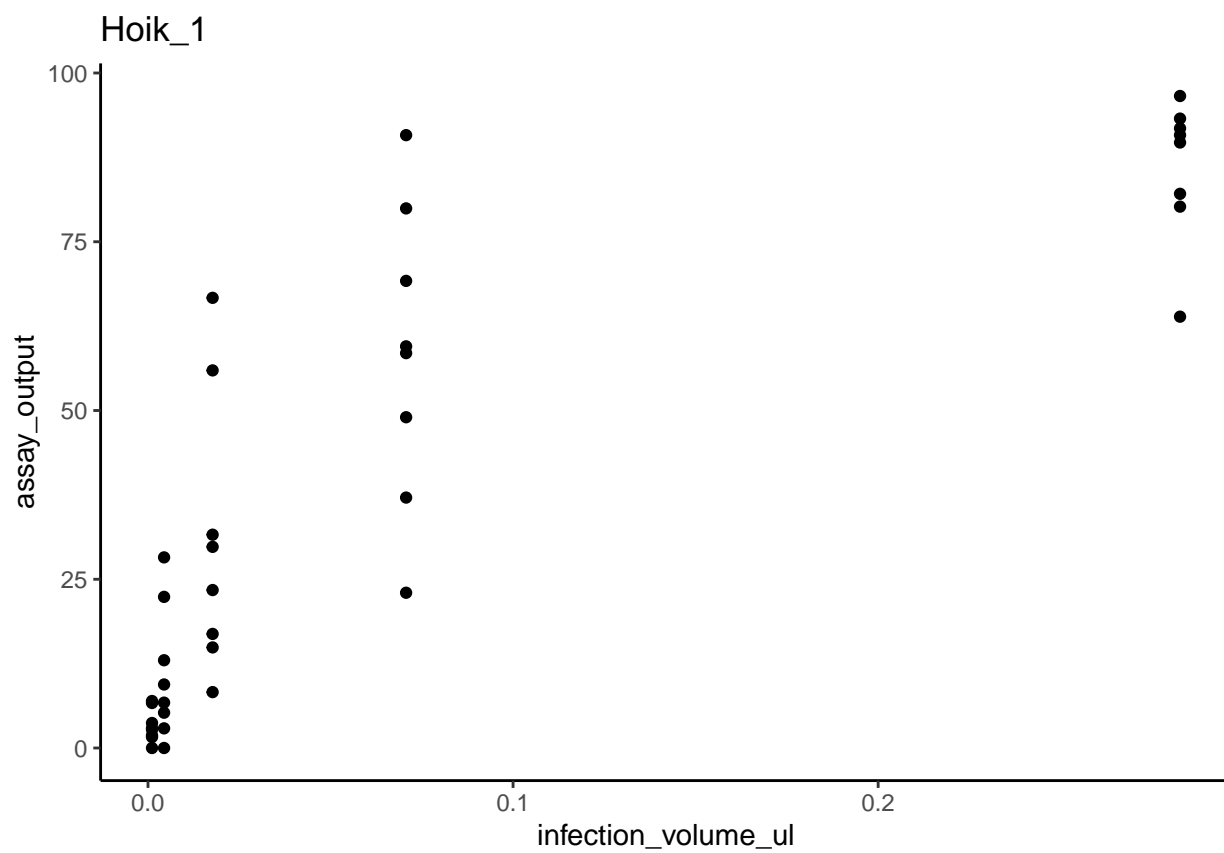
[[41]]



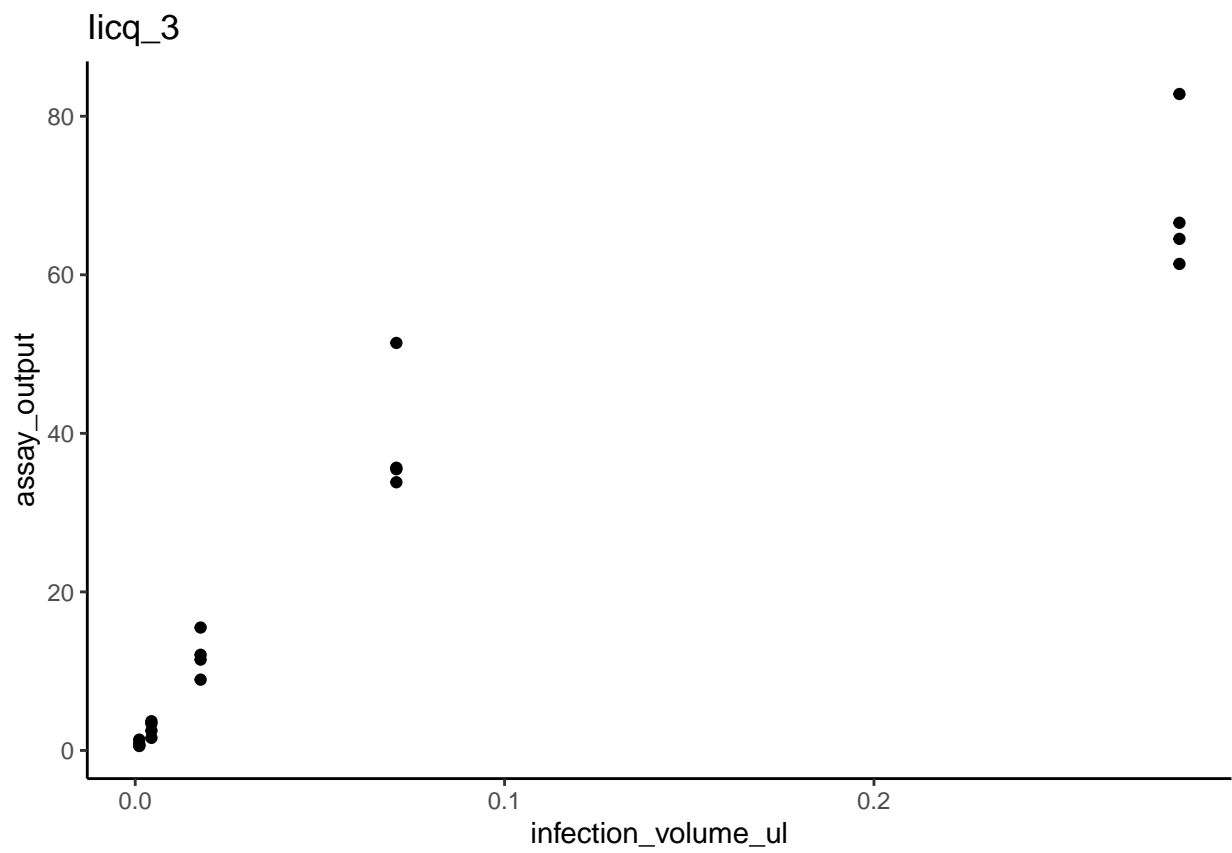
[[42]]



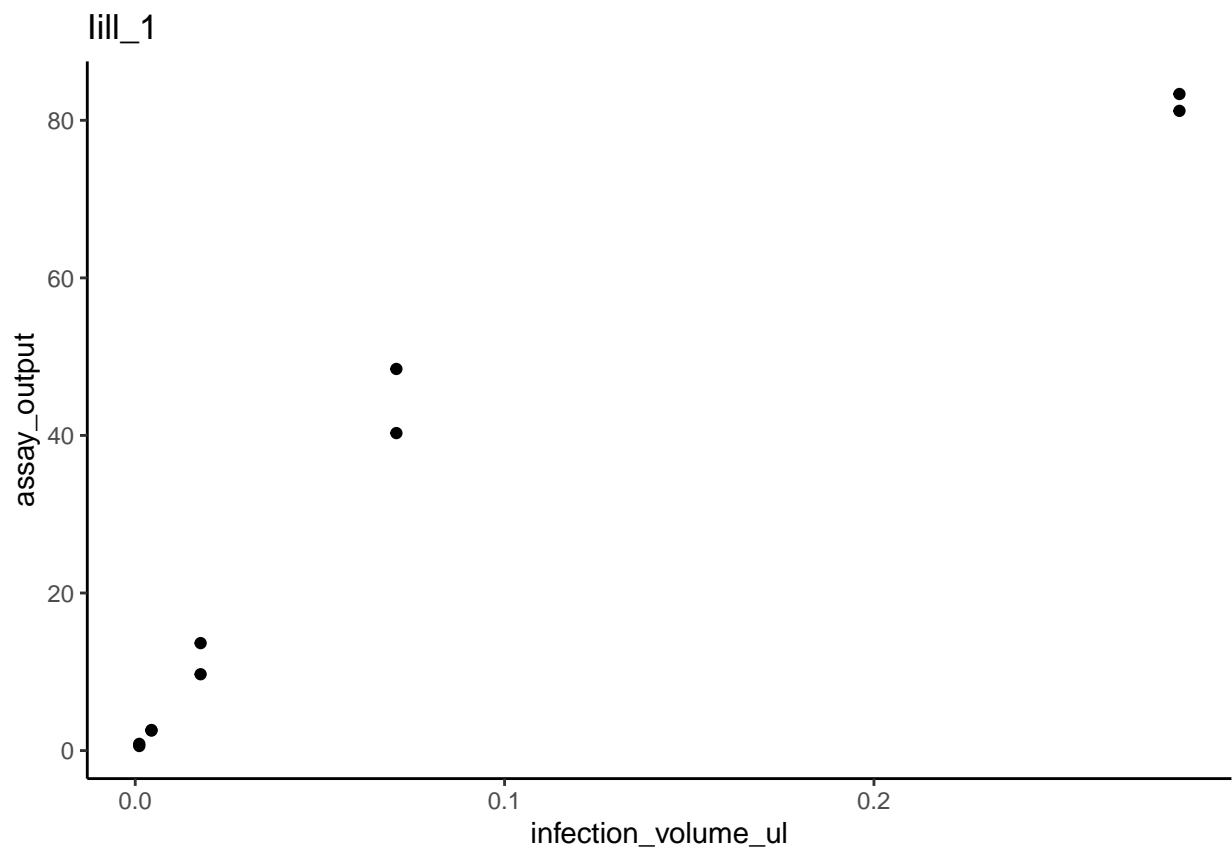
[[43]]



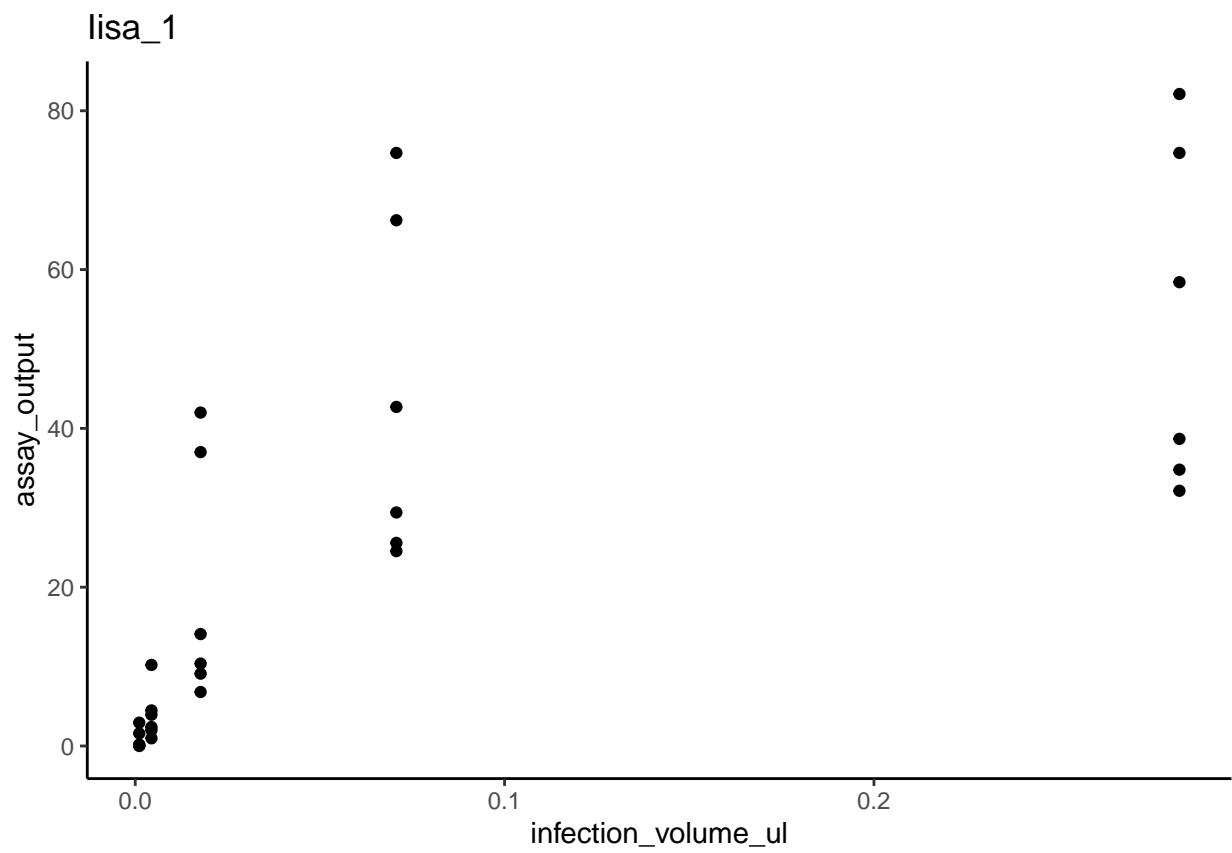
[[44]]



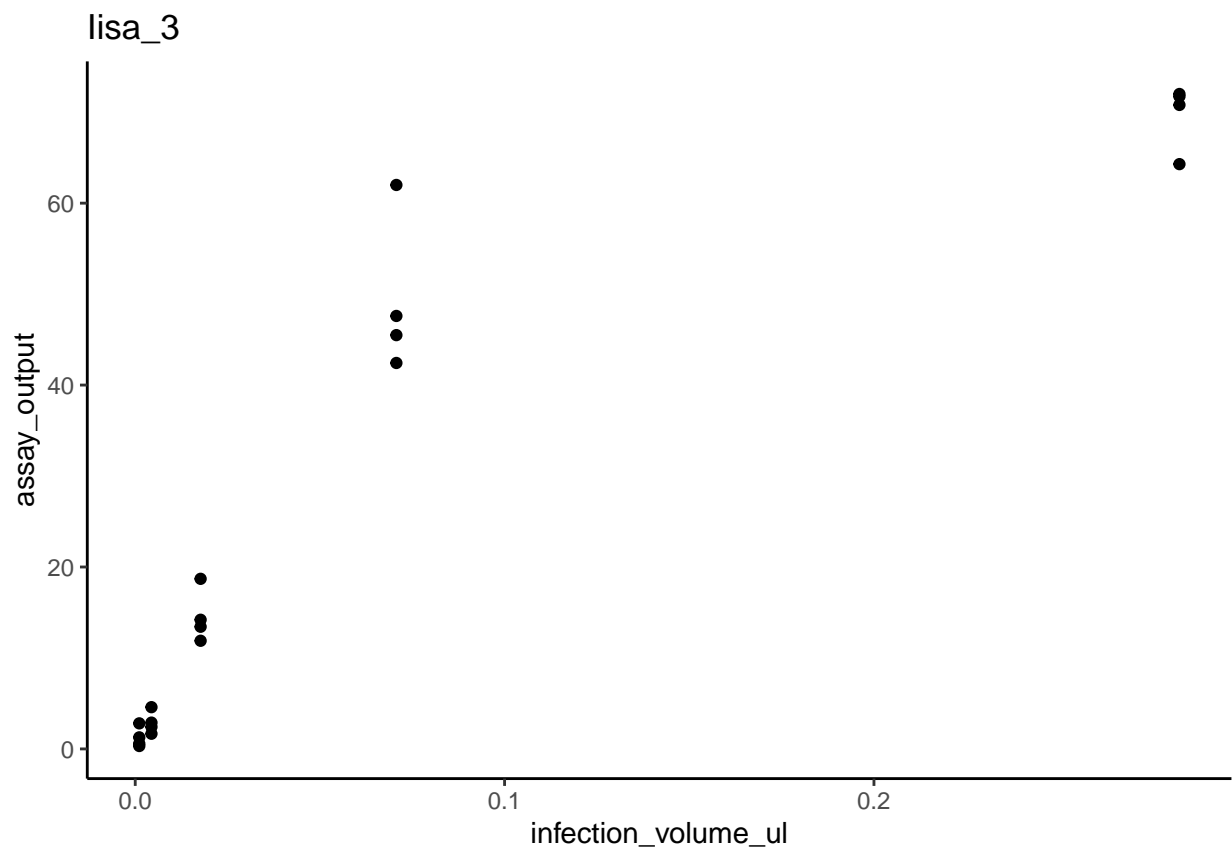
```
##  
## [[45]]
```

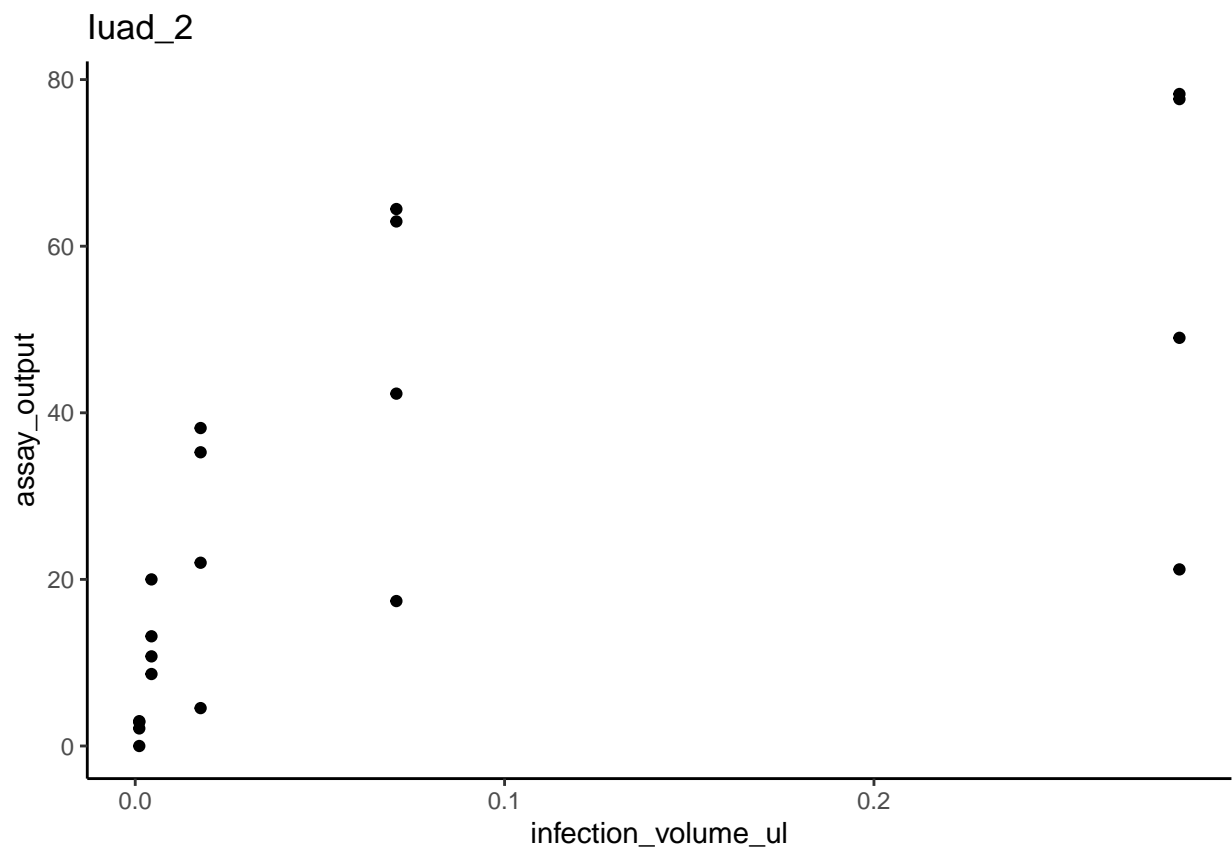


```
##  
## [[46]]
```

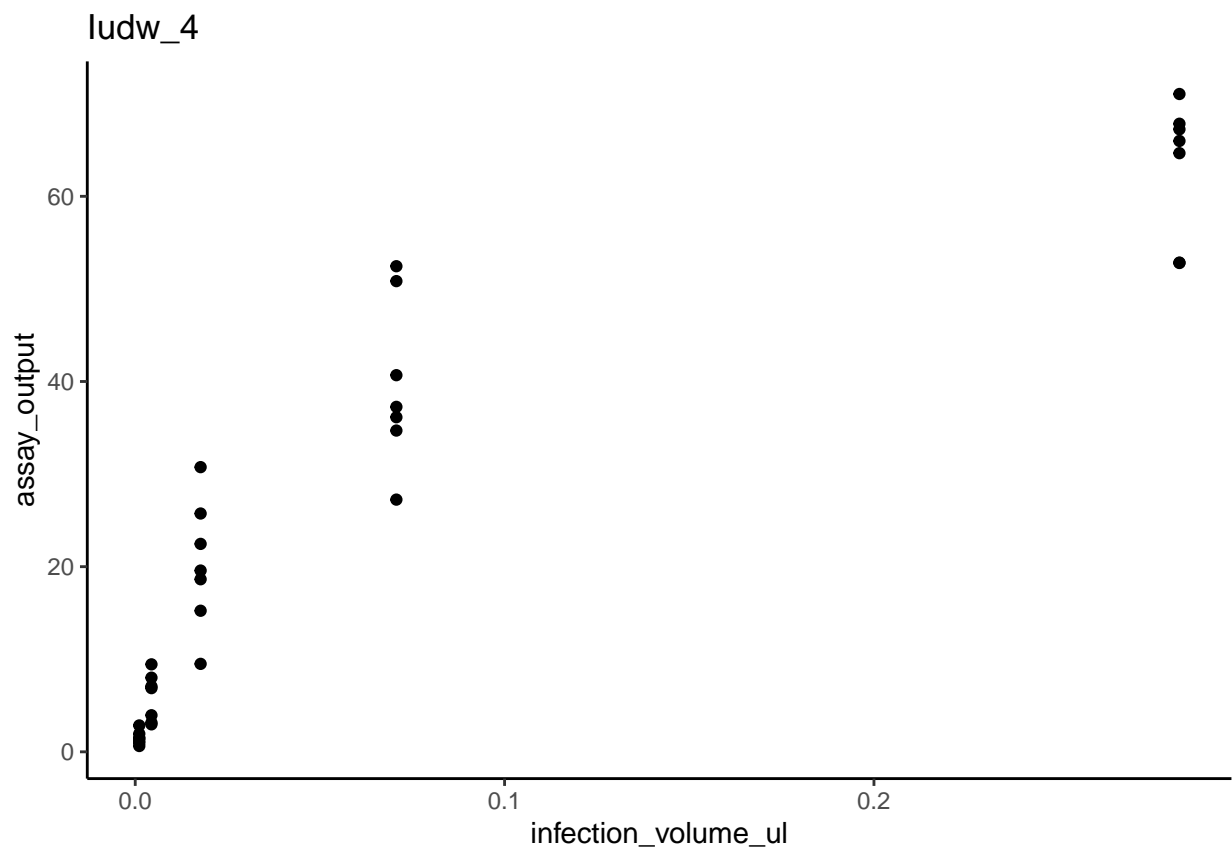


[[47]]

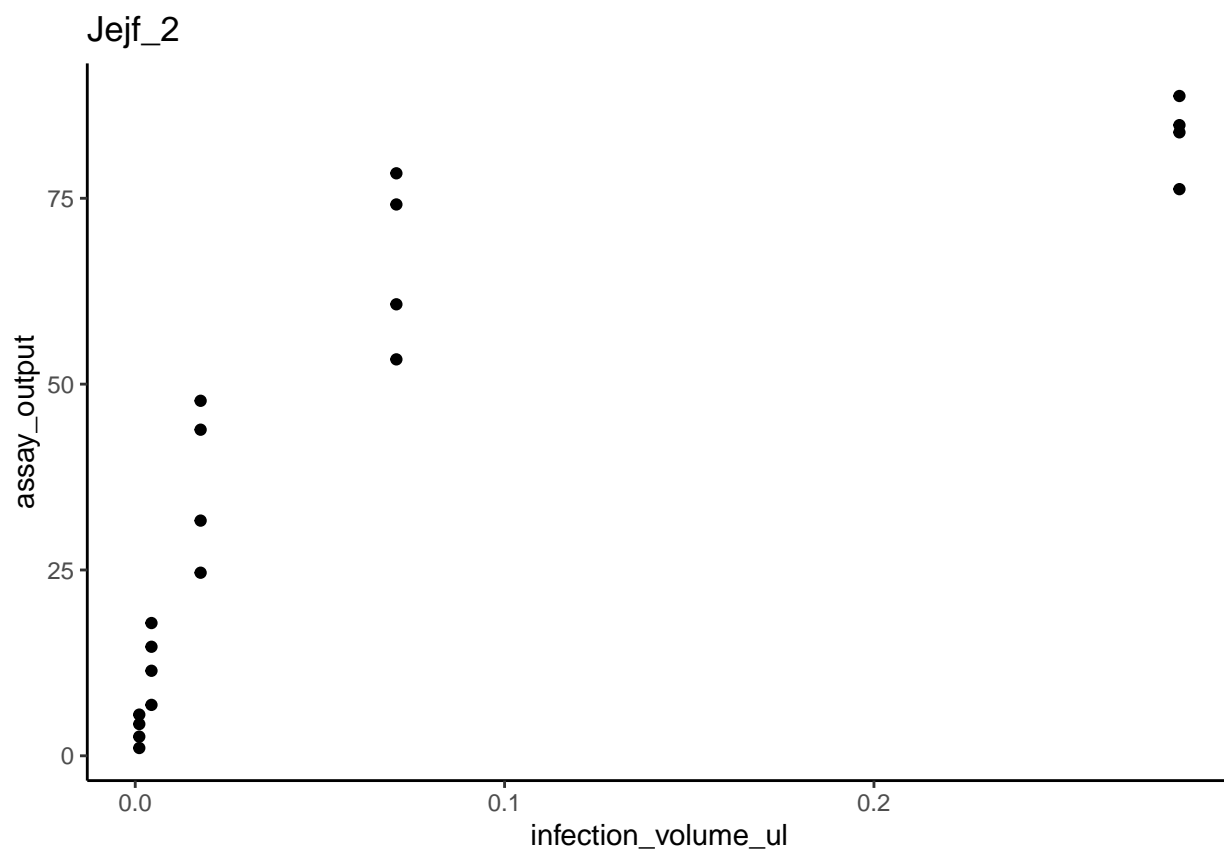




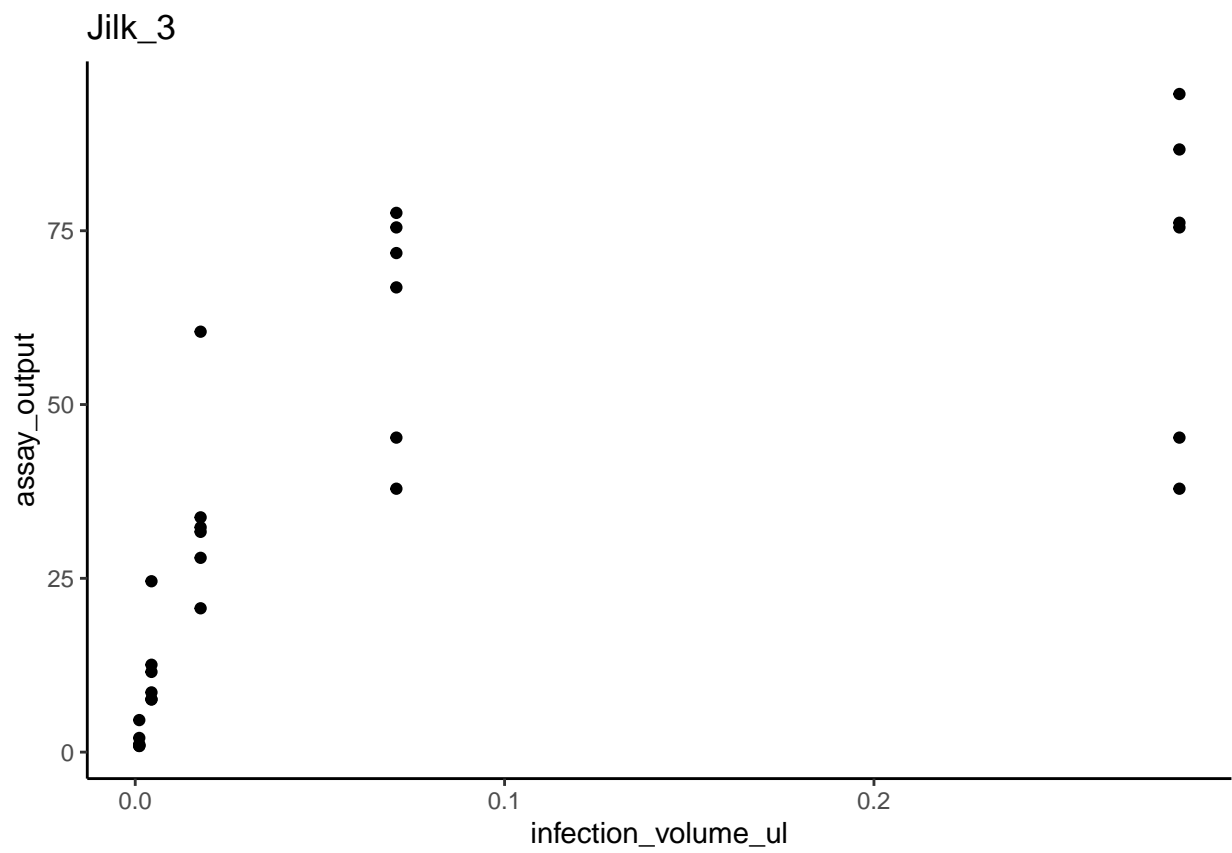
```
##  
## [[49]]
```



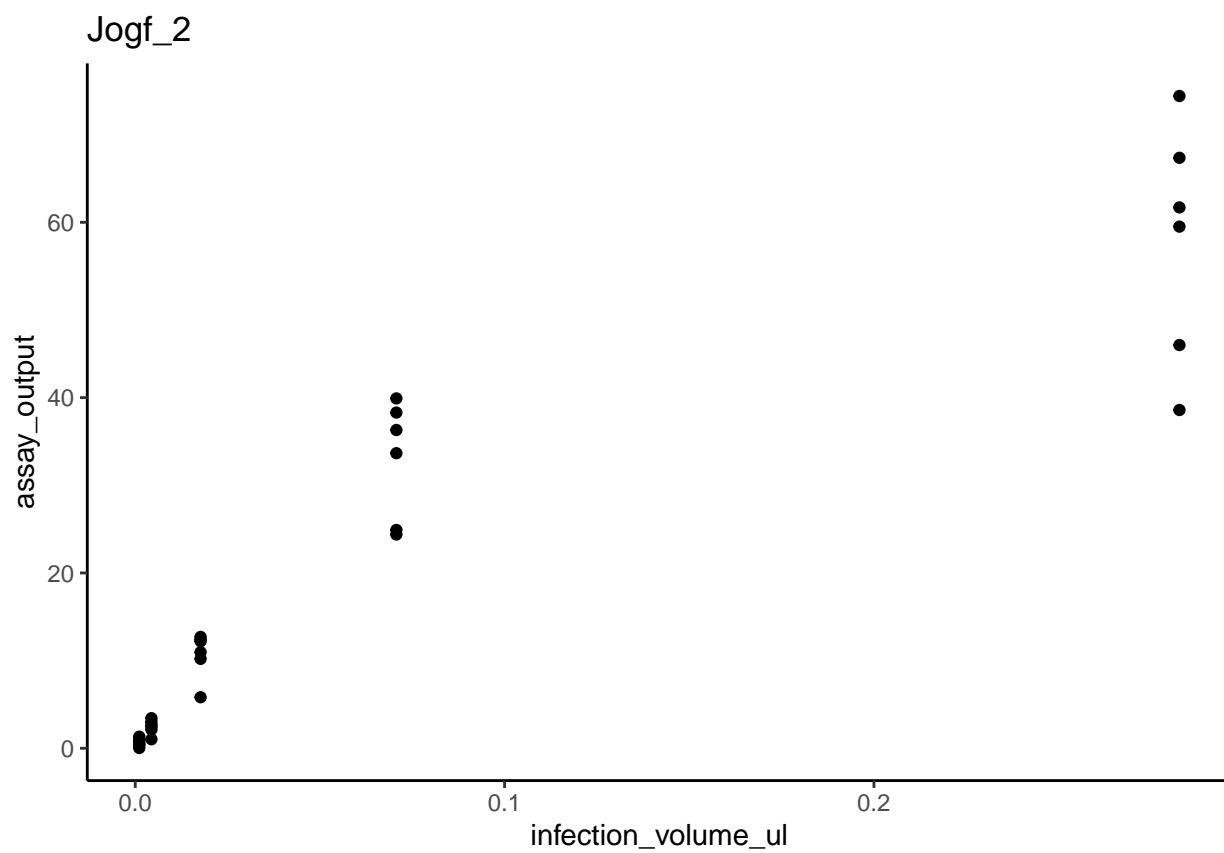
[[50]]



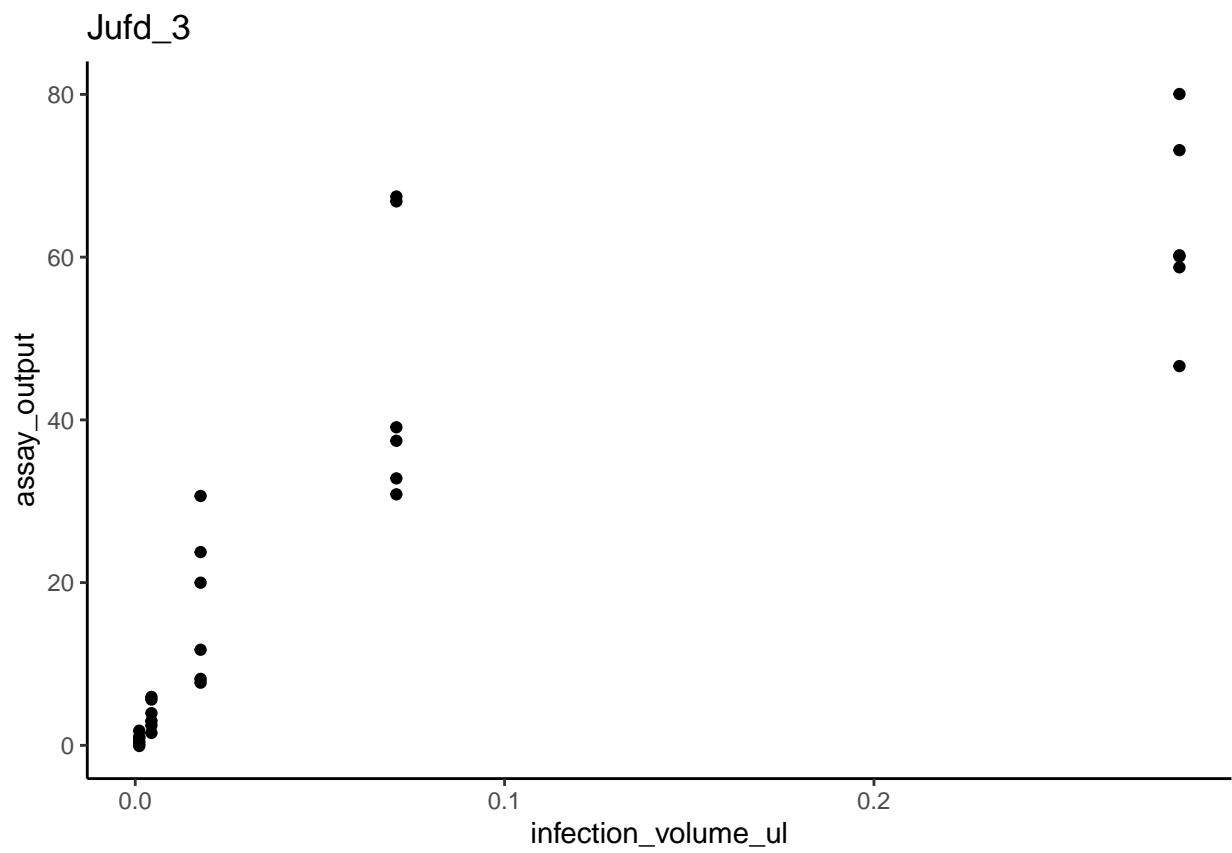
[[51]]



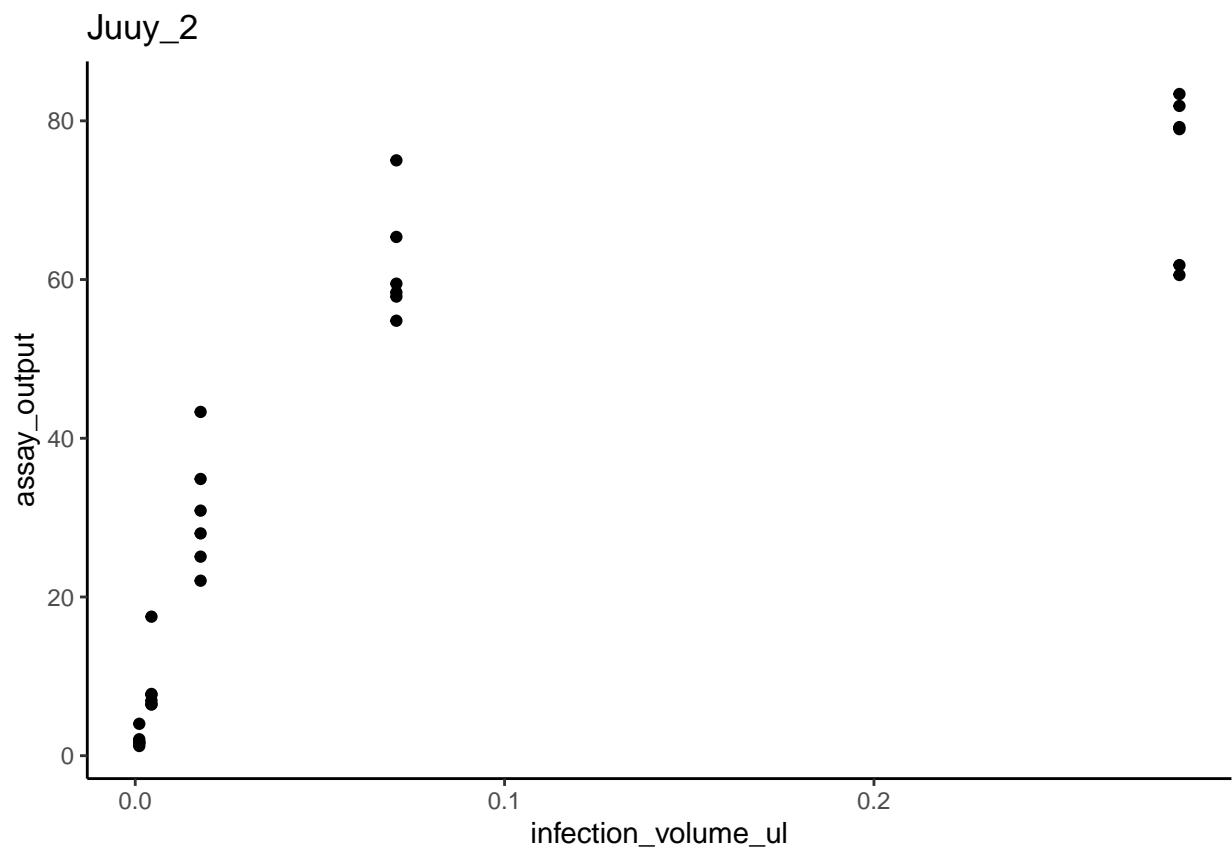
[[52]]



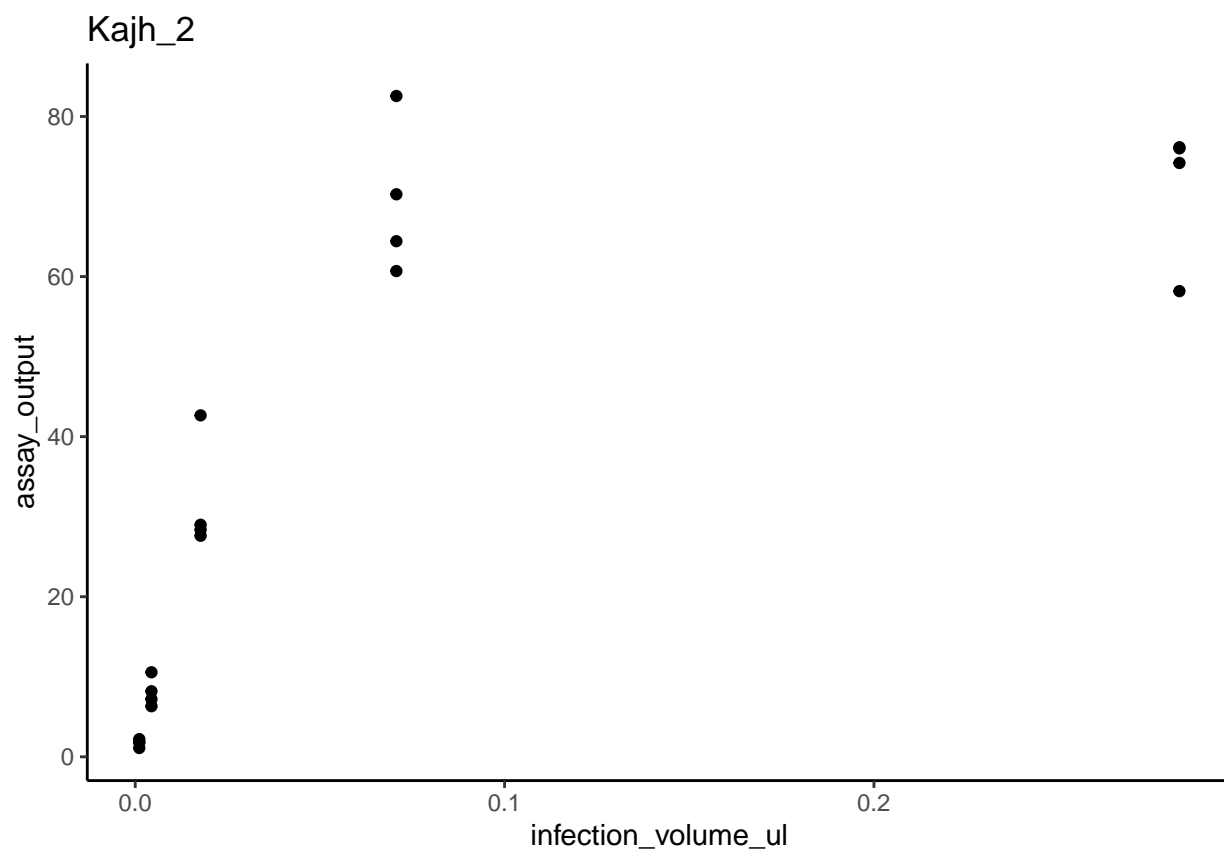
[[53]]



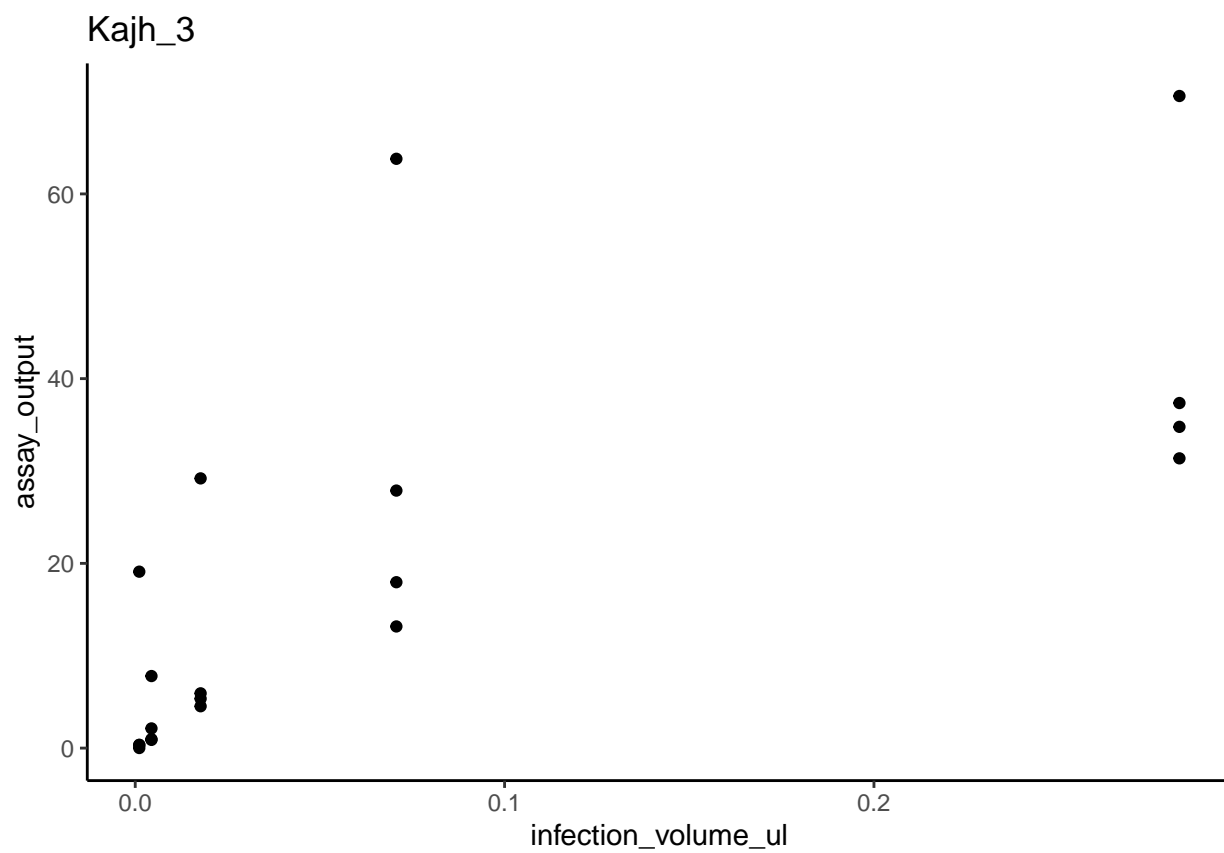
[[55]]



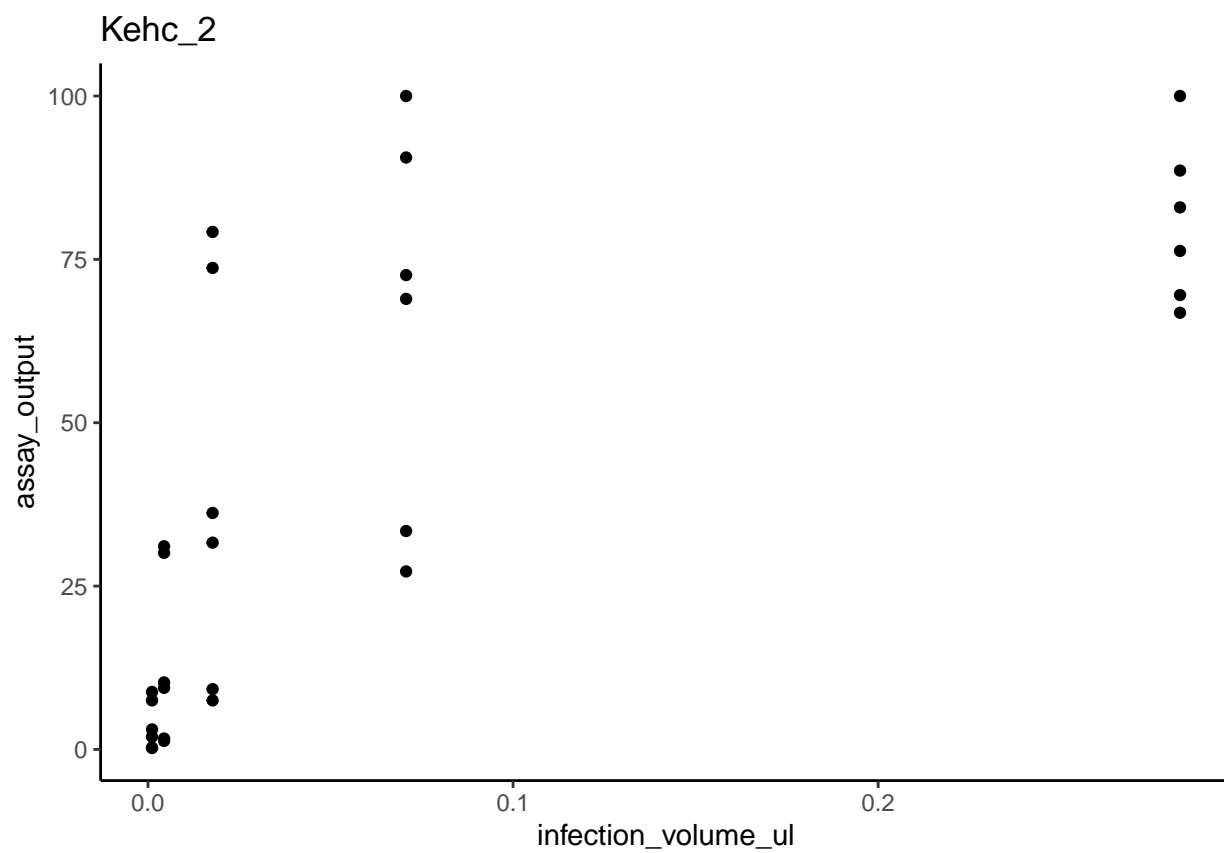
[[56]]



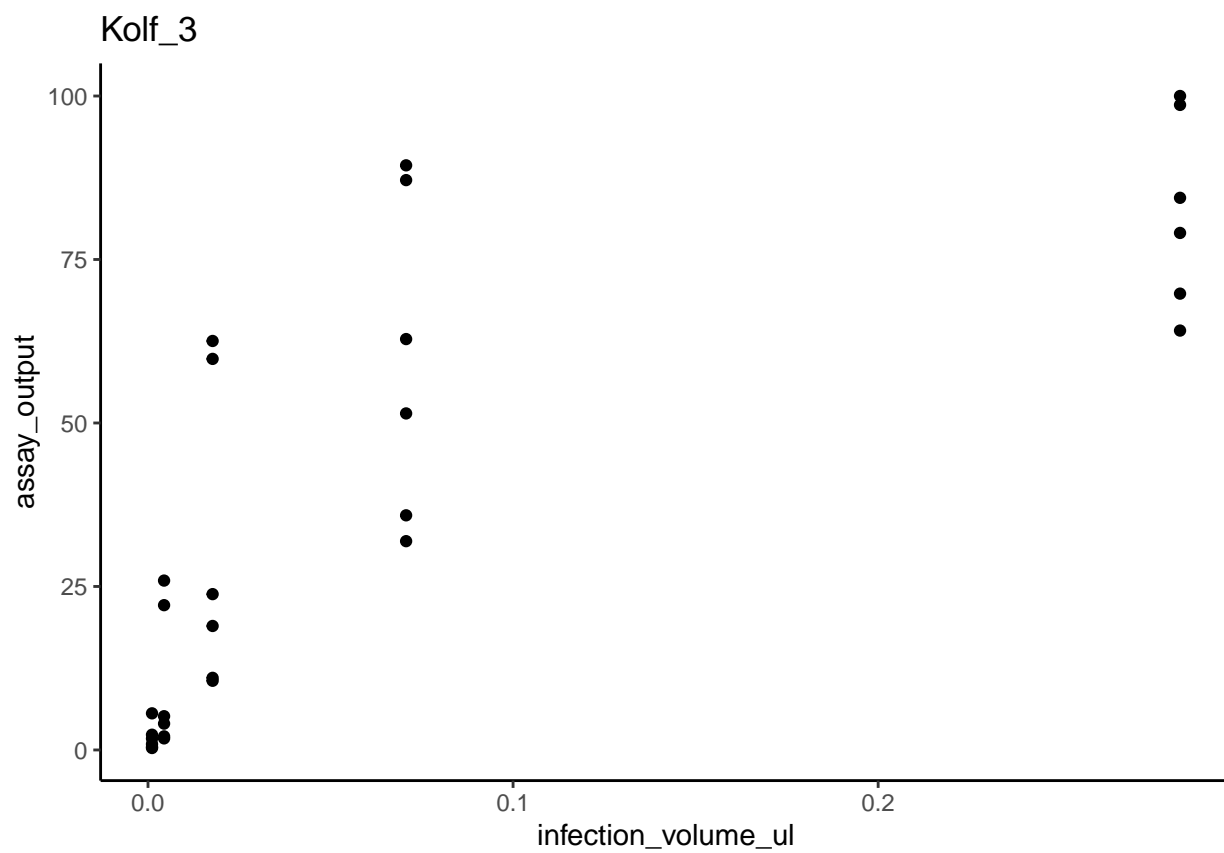
[[57]]



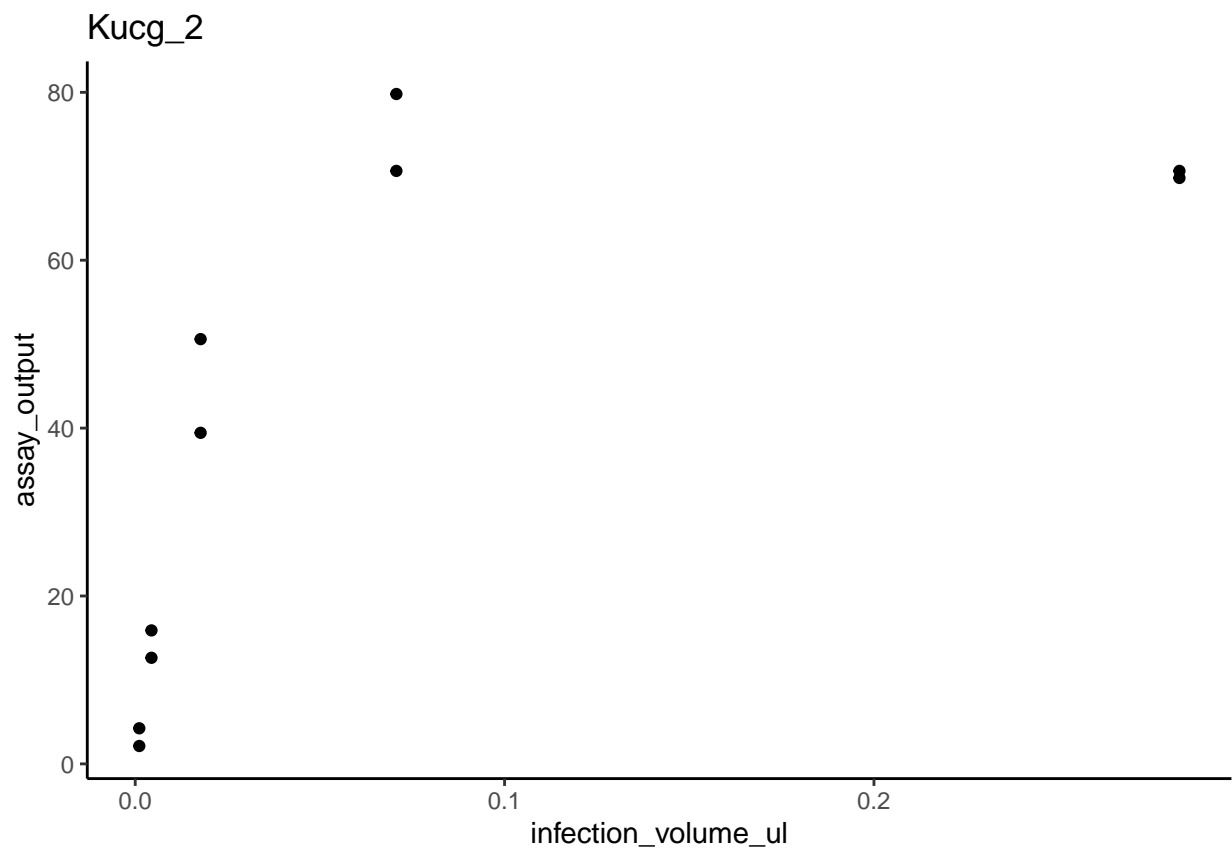
[[58]]



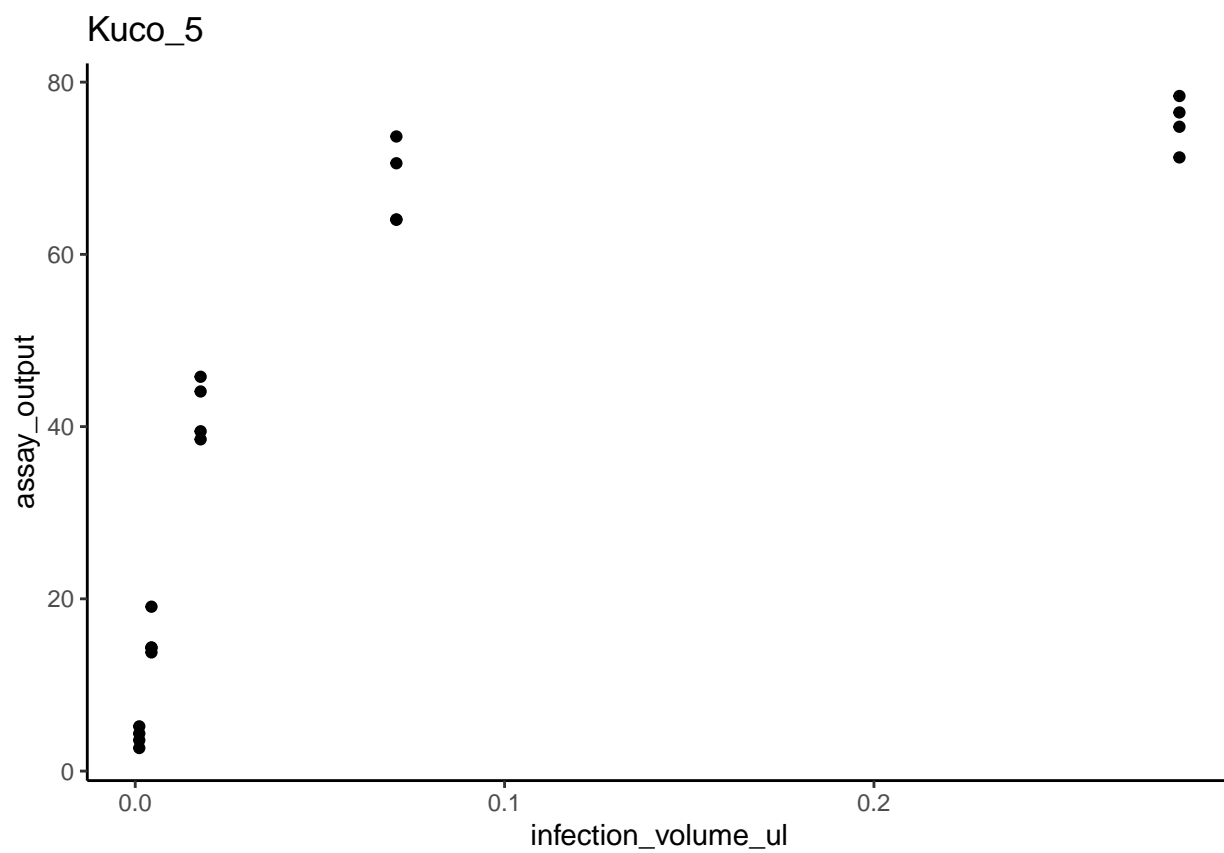
[[60]]



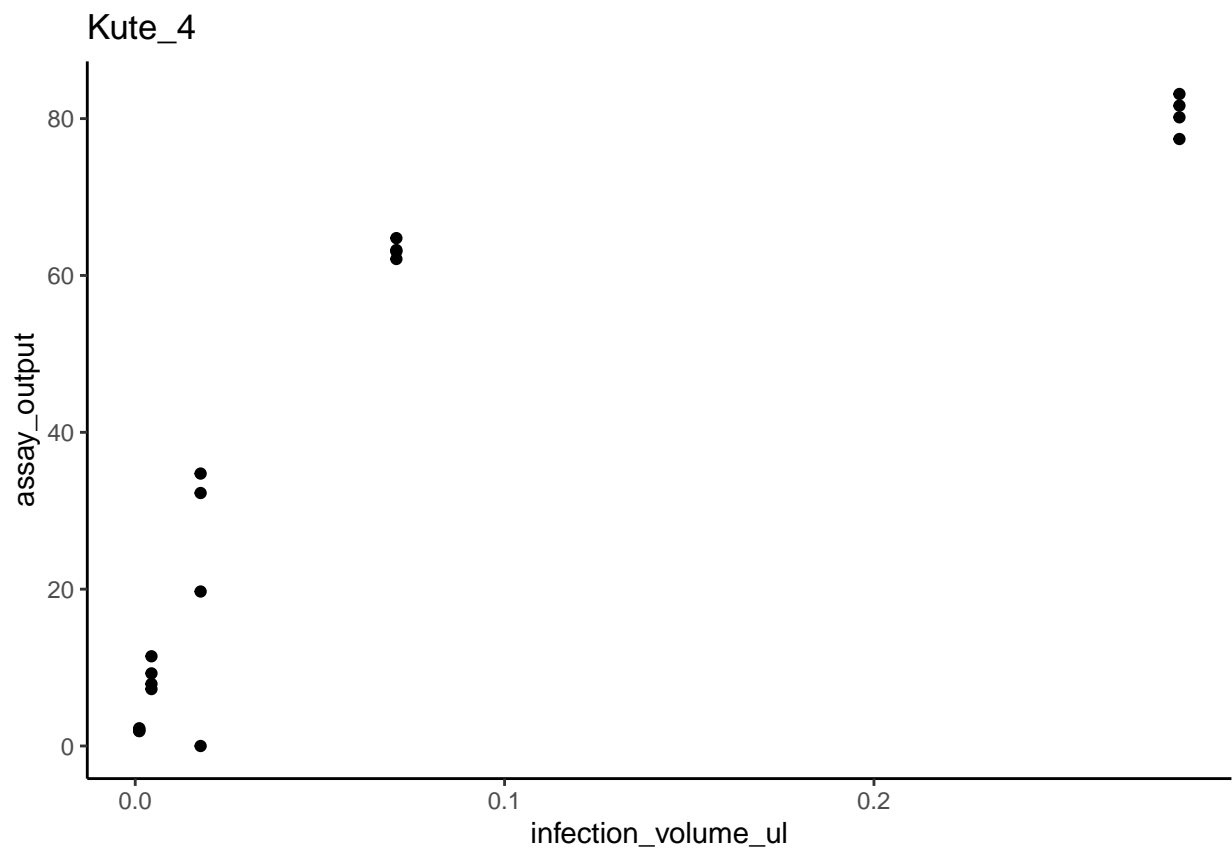
[[61]]



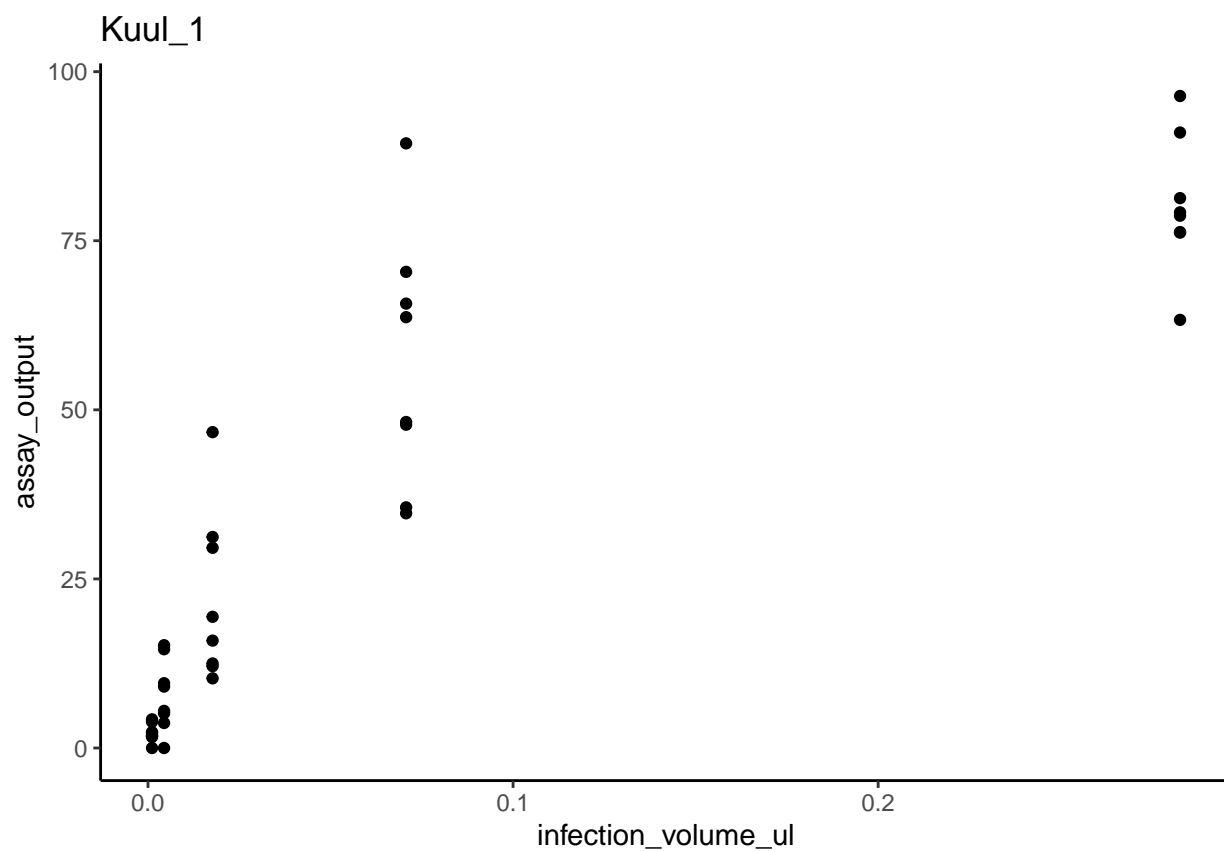
[[62]]



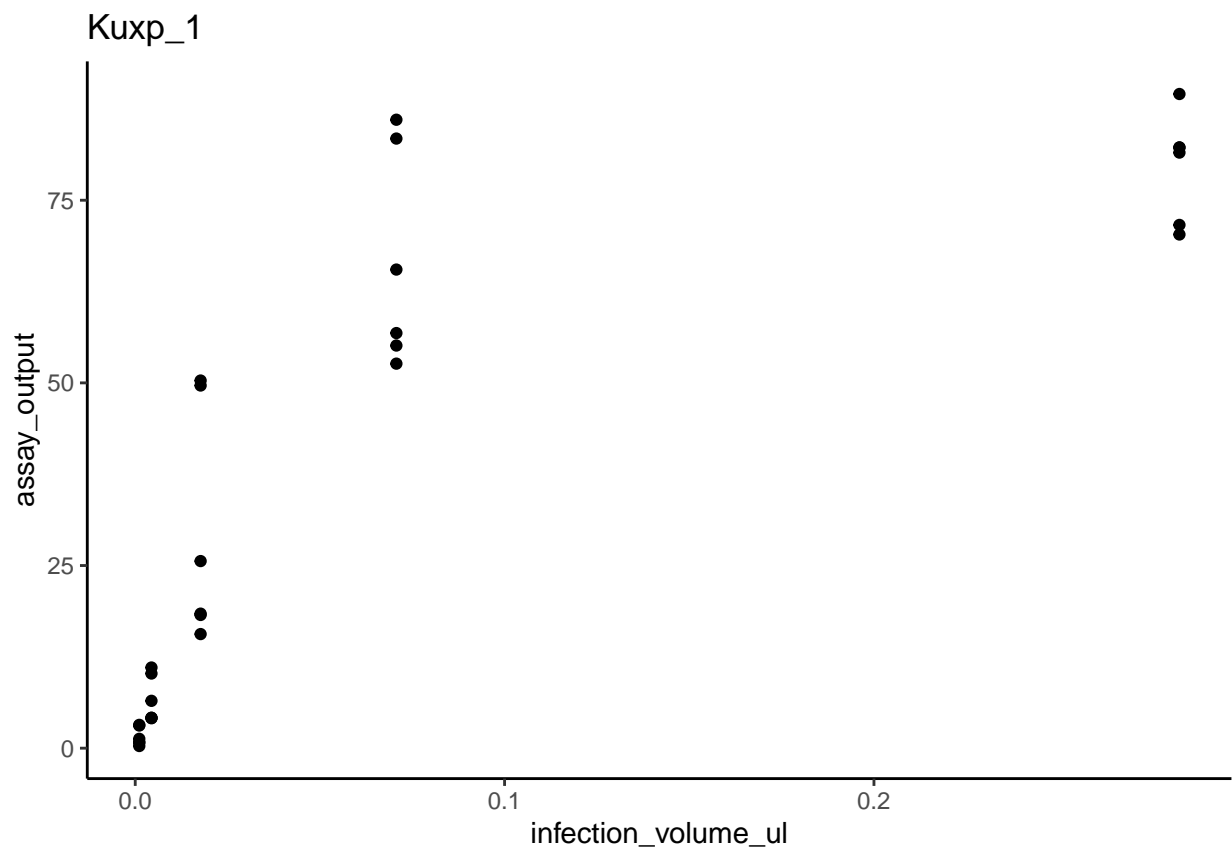
[[63]]



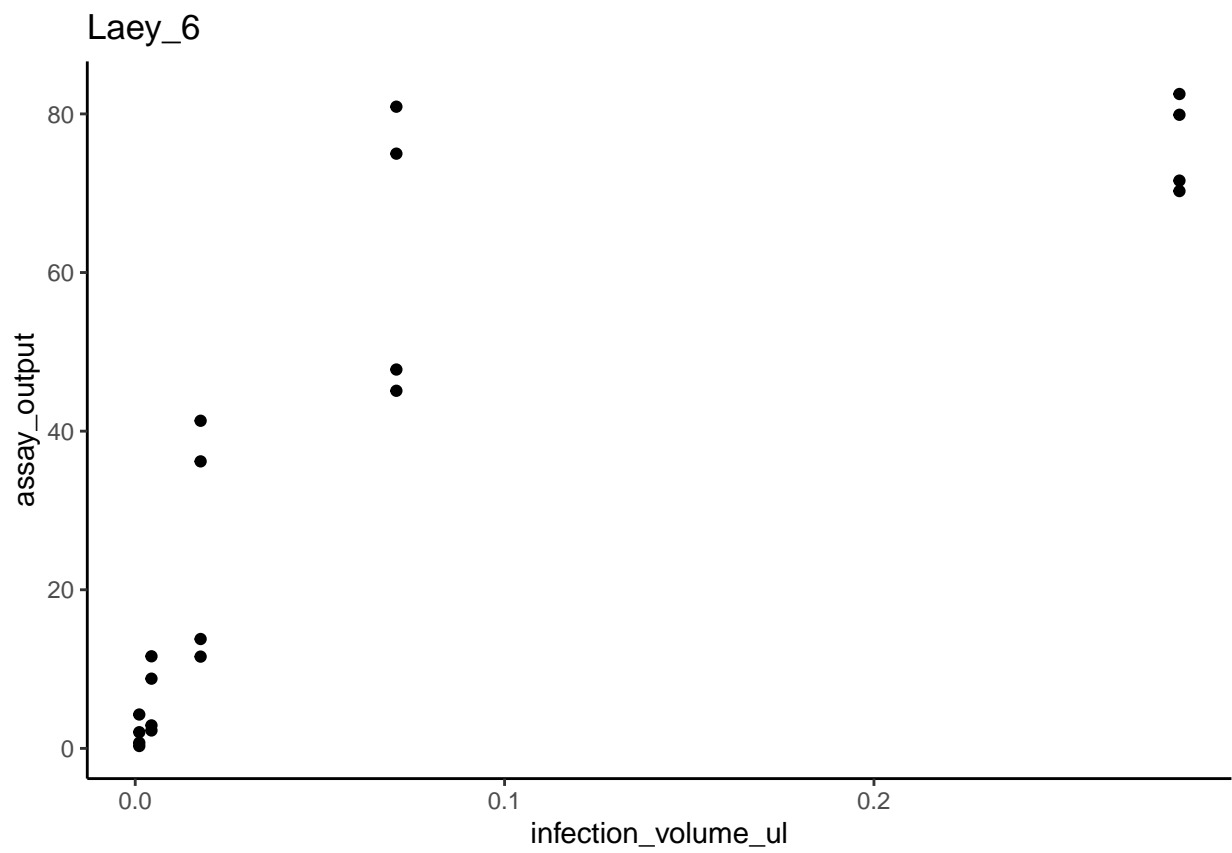
[[64]]



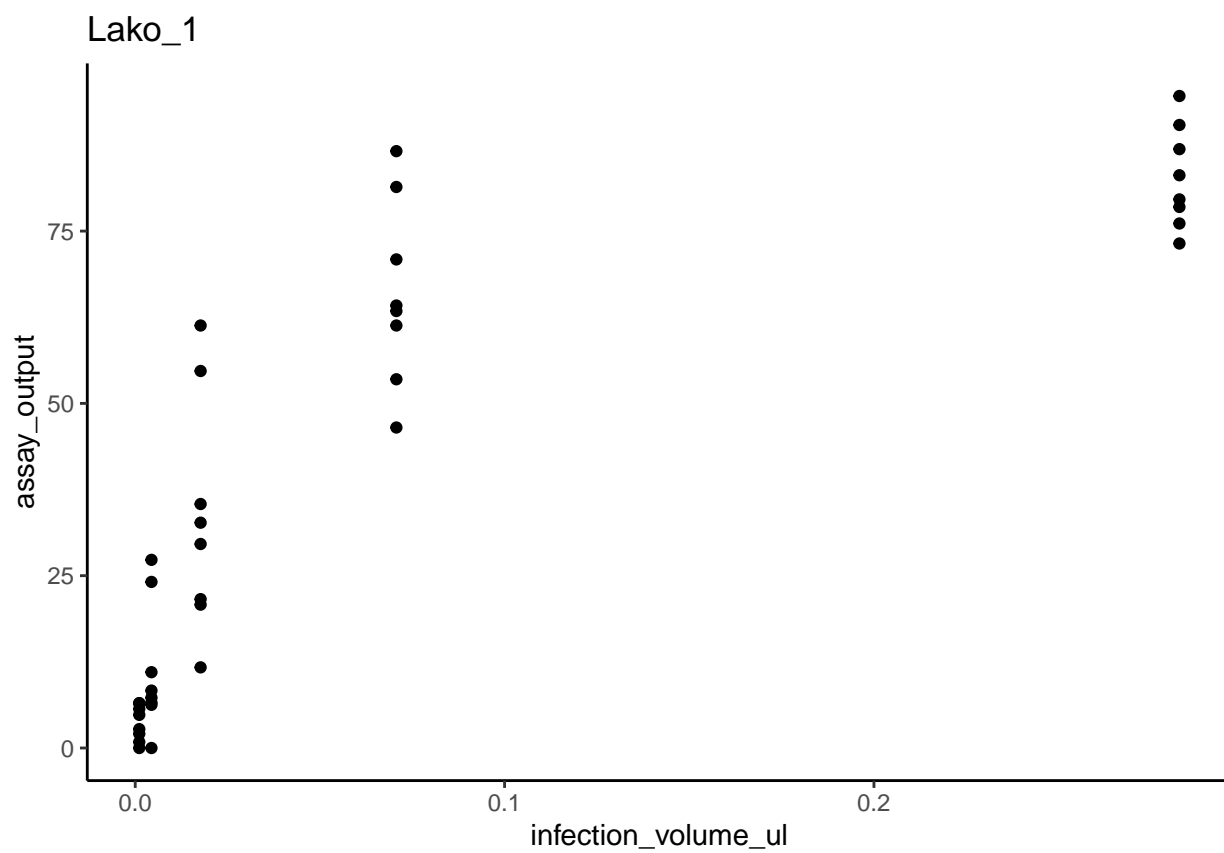
[[65]]



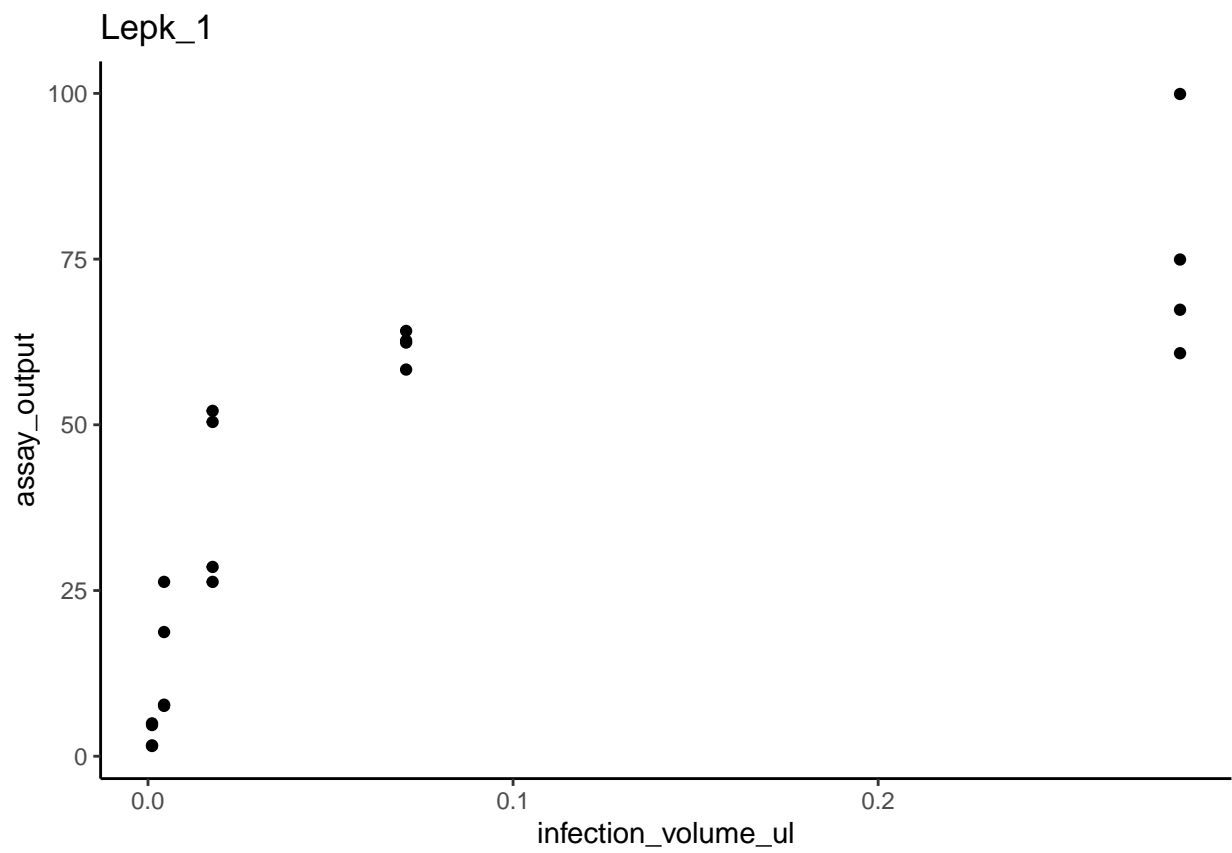
[[66]]



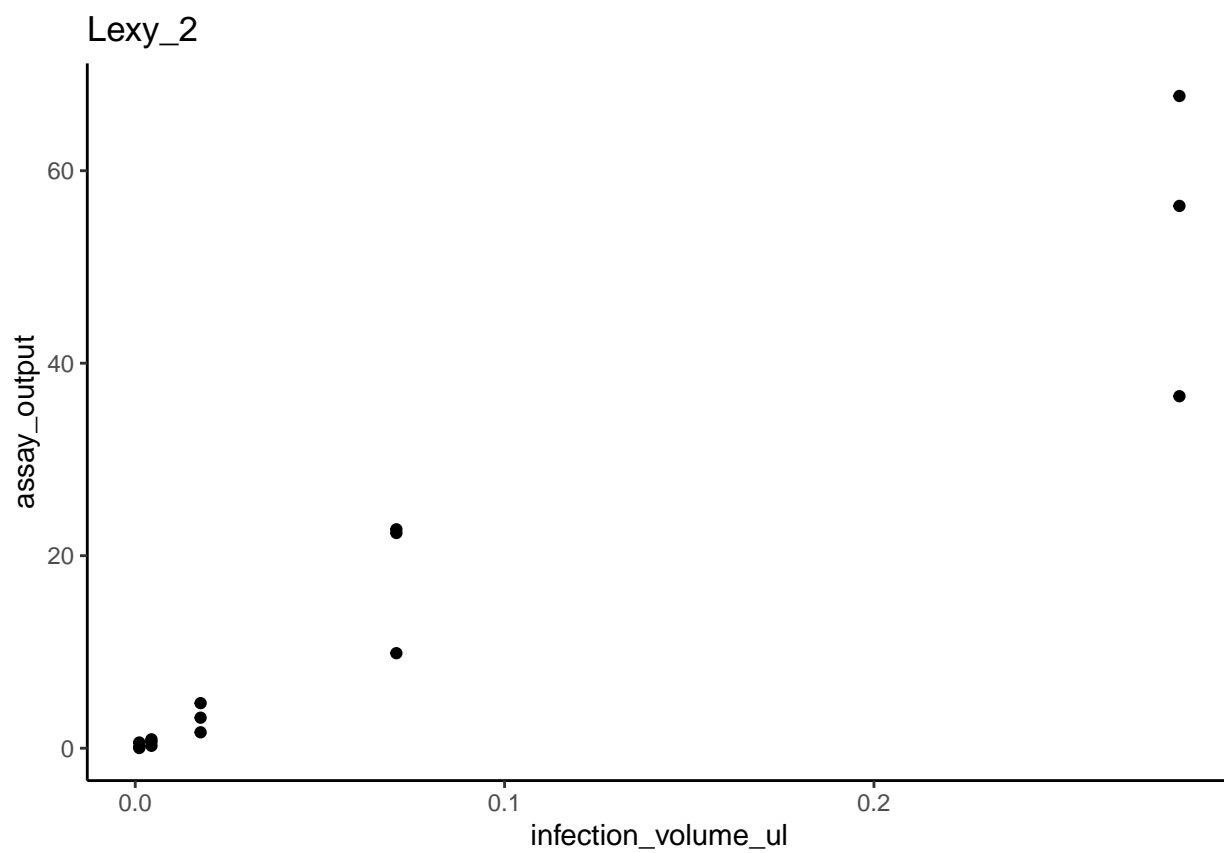
[[67]]



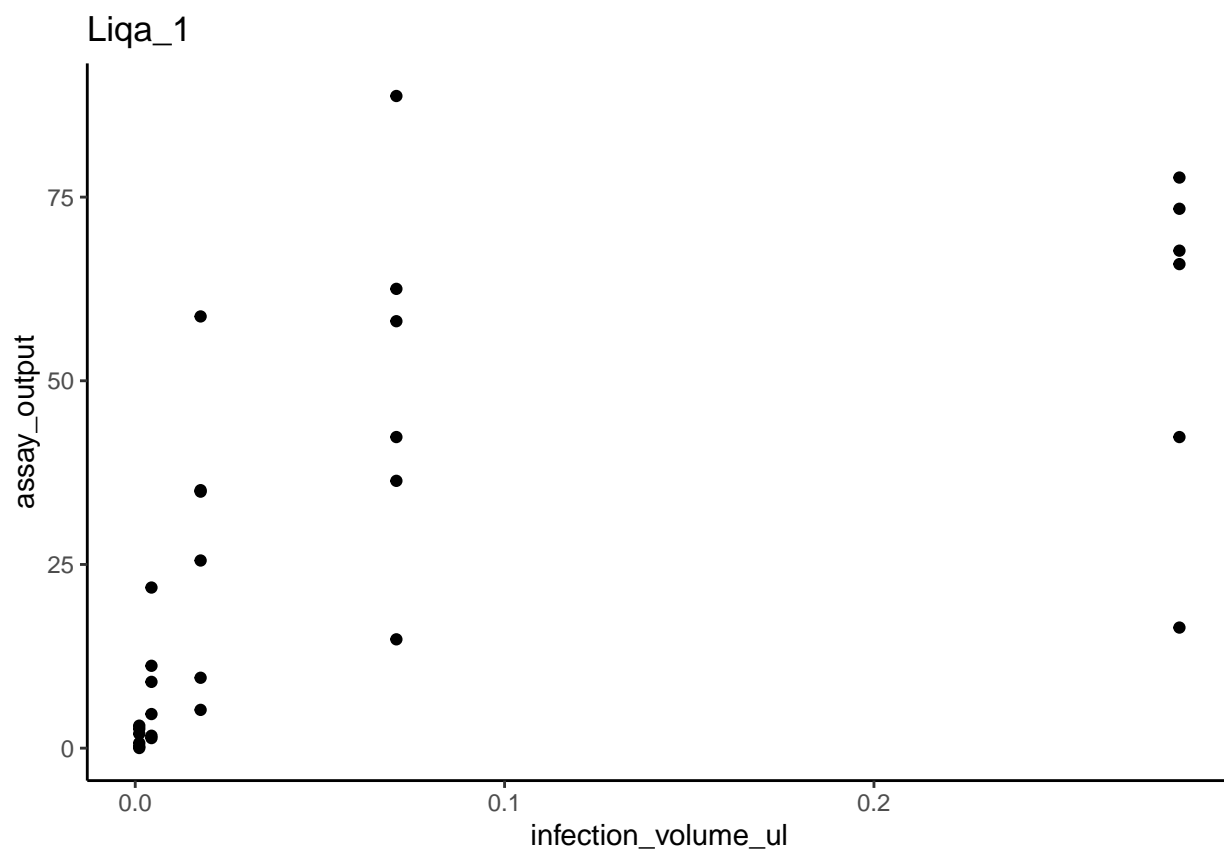
[[68]]



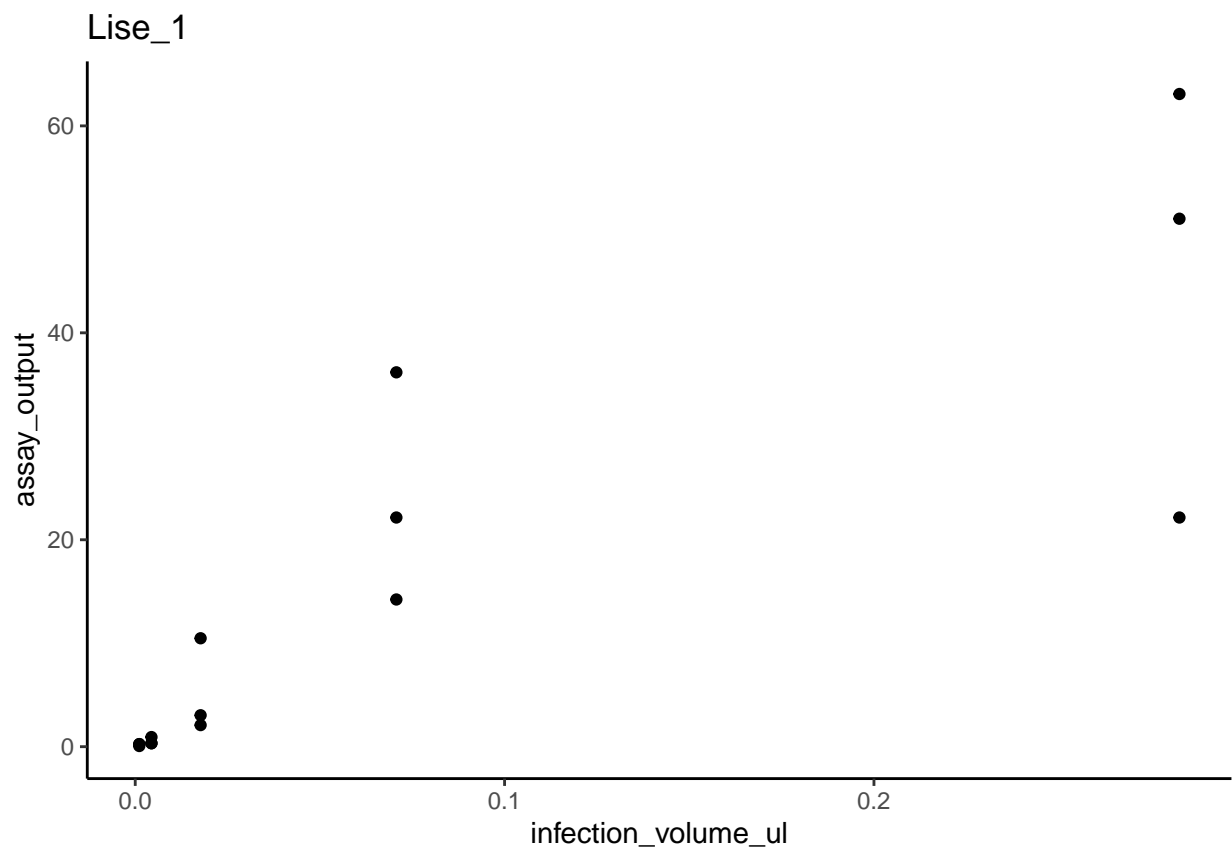
[[69]]



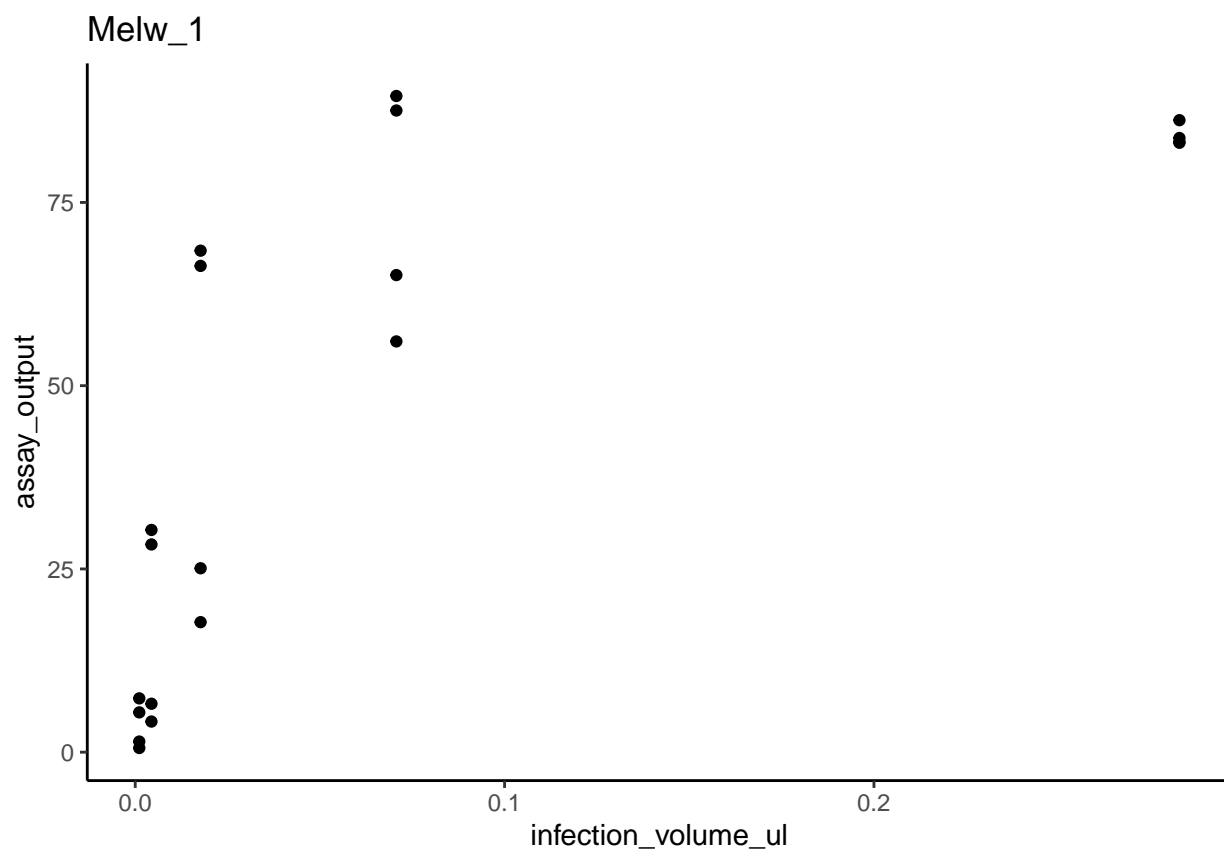
```
##  
## [[70]]
```



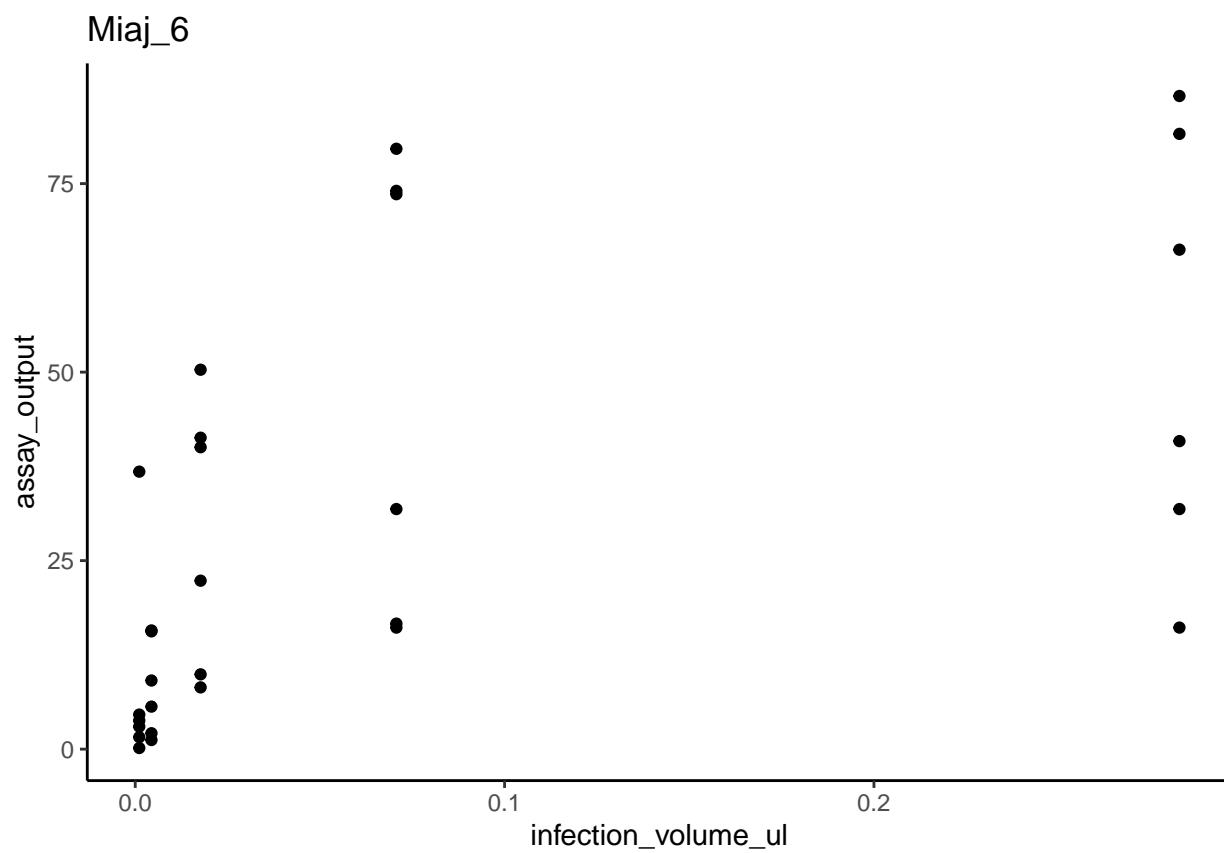
[[71]]



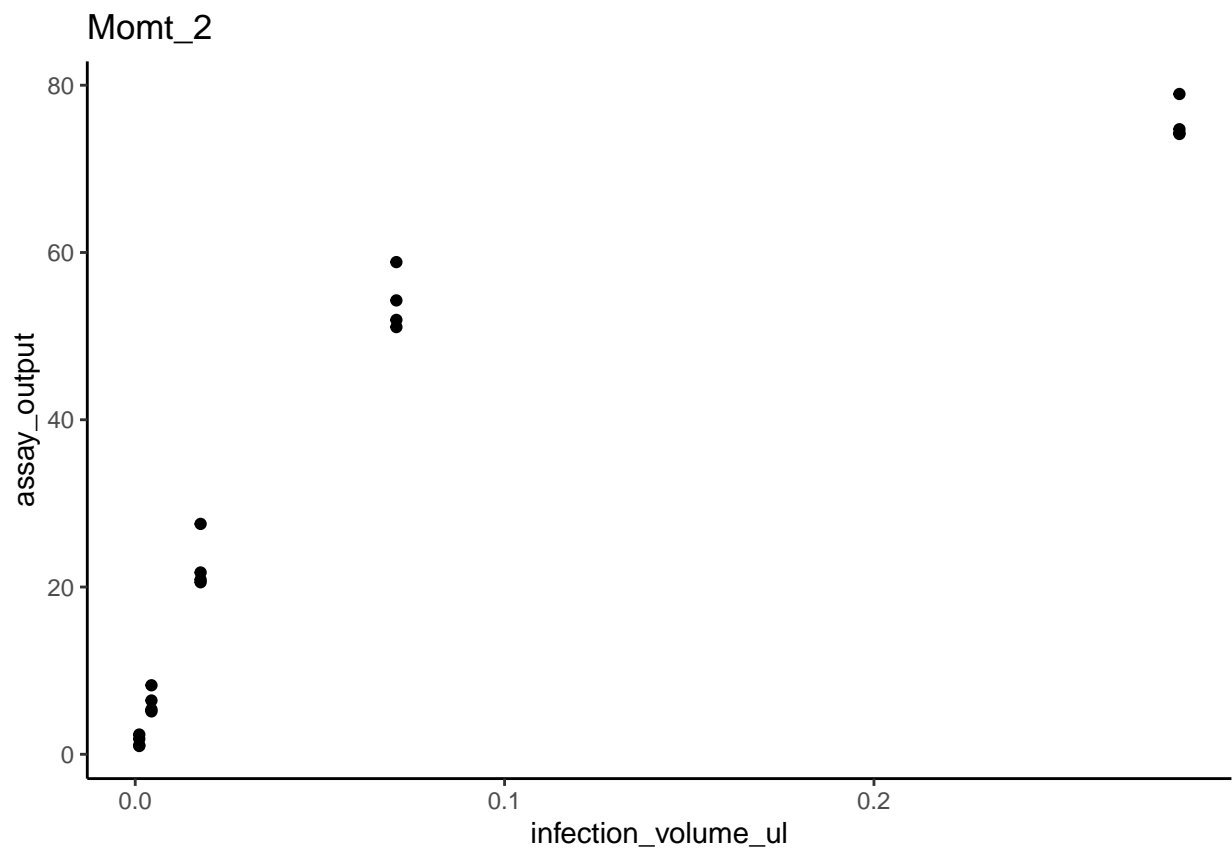
[[72]]



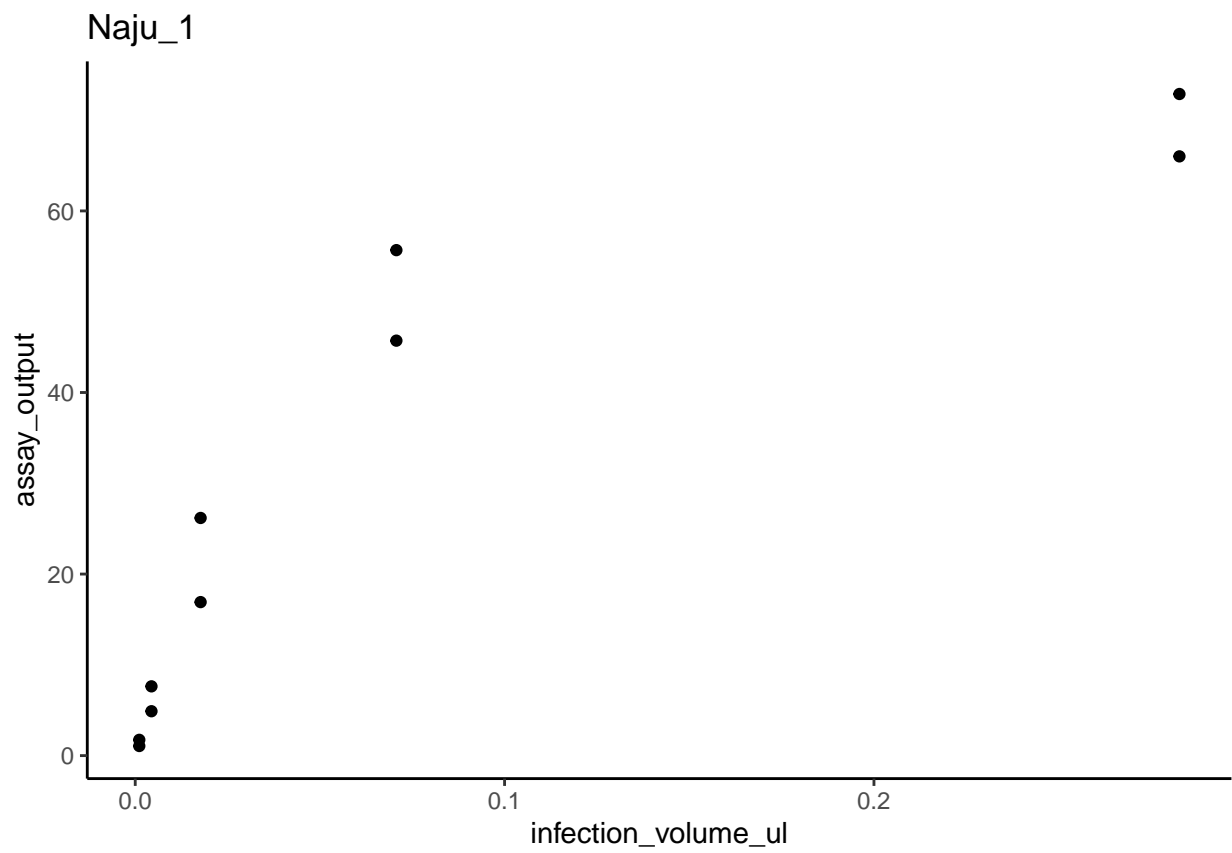
[[73]]



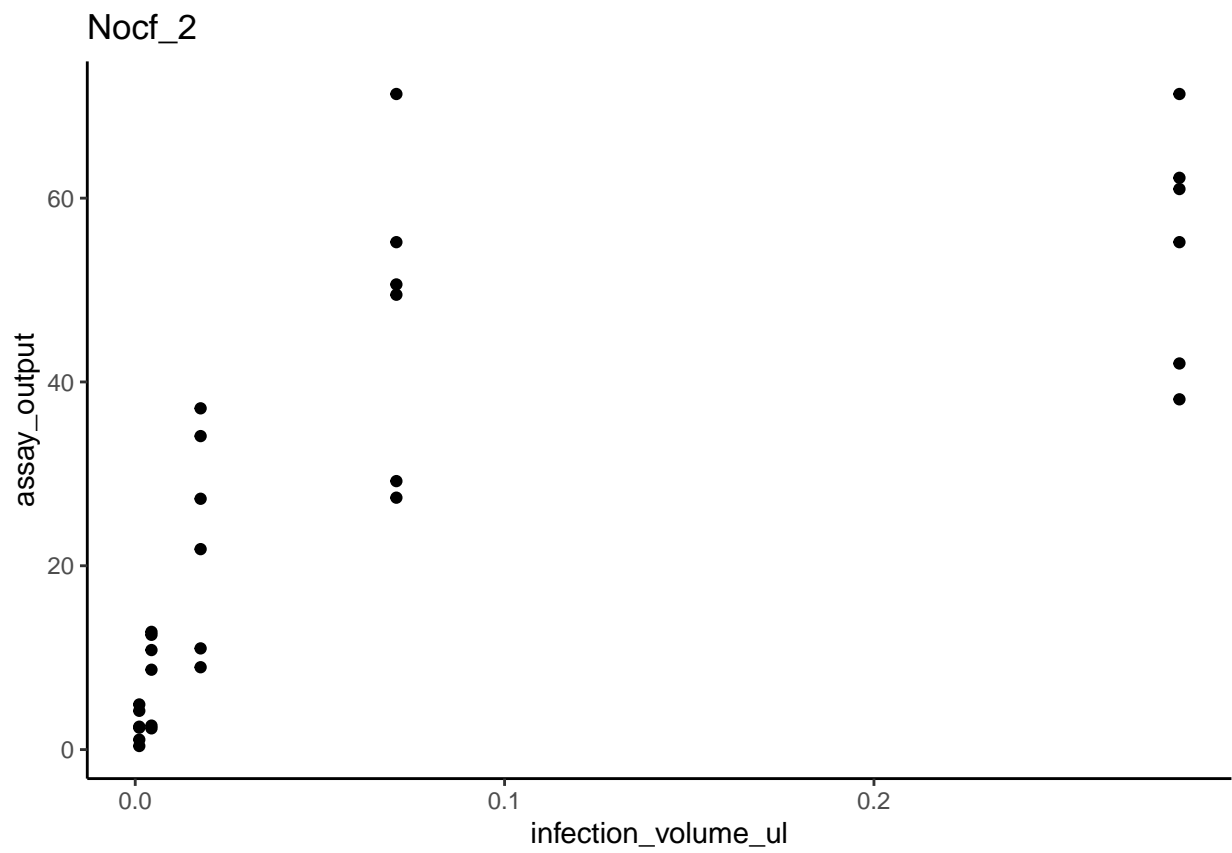
[[74]]

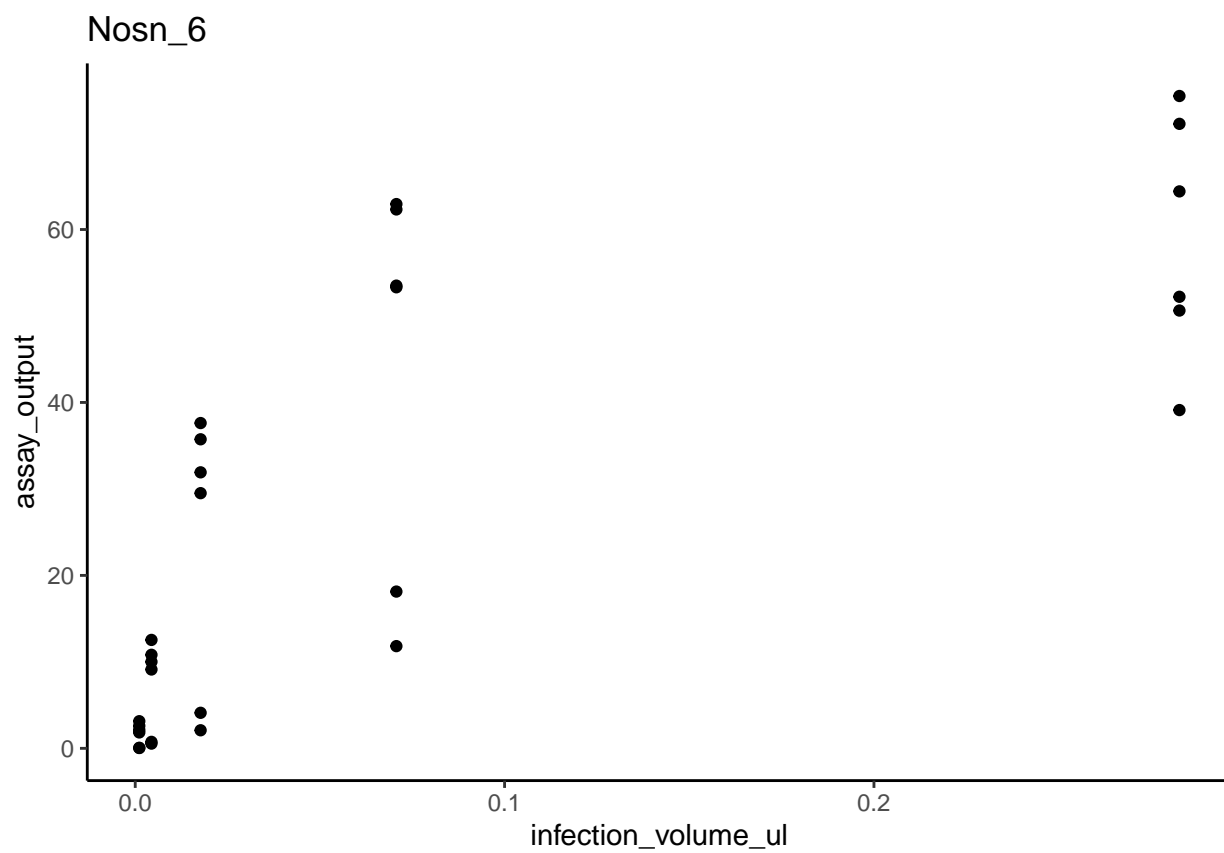


[[75]]

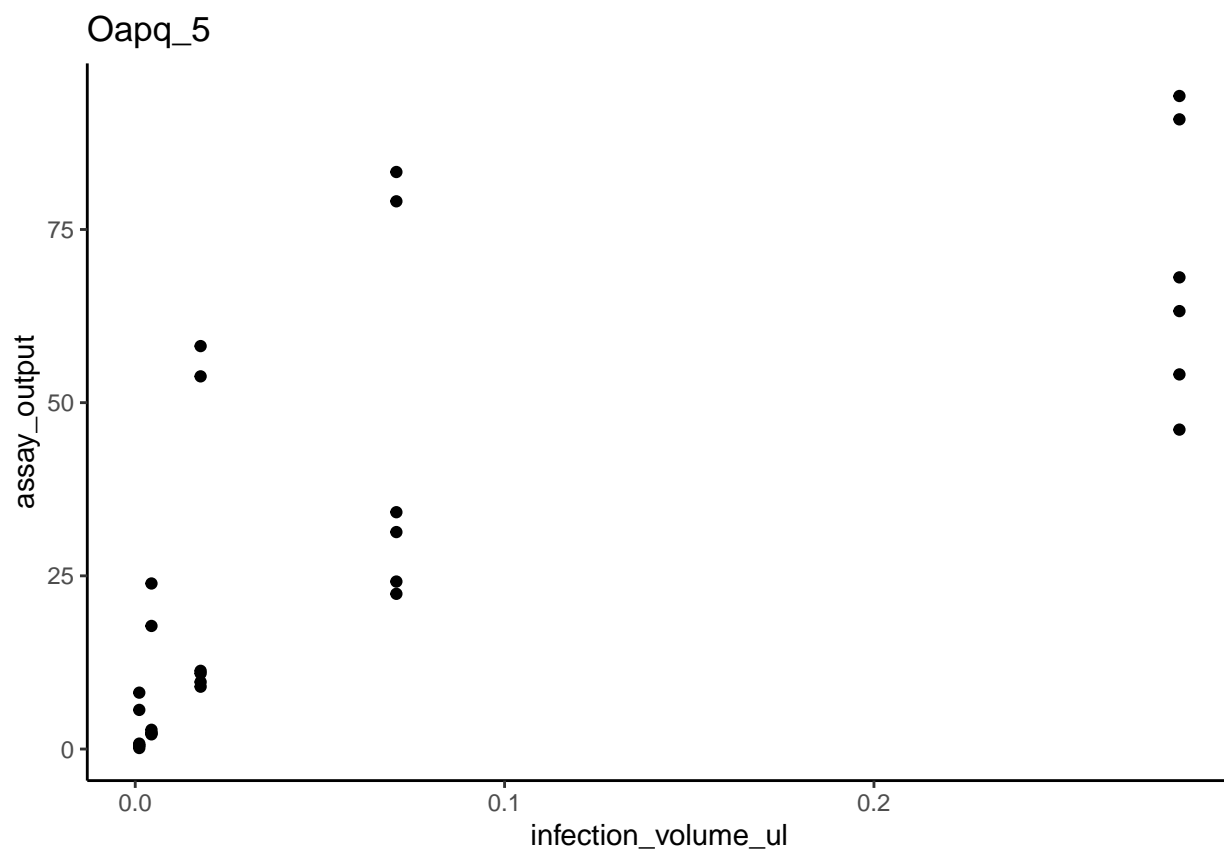


```
##  
## [[76]]
```

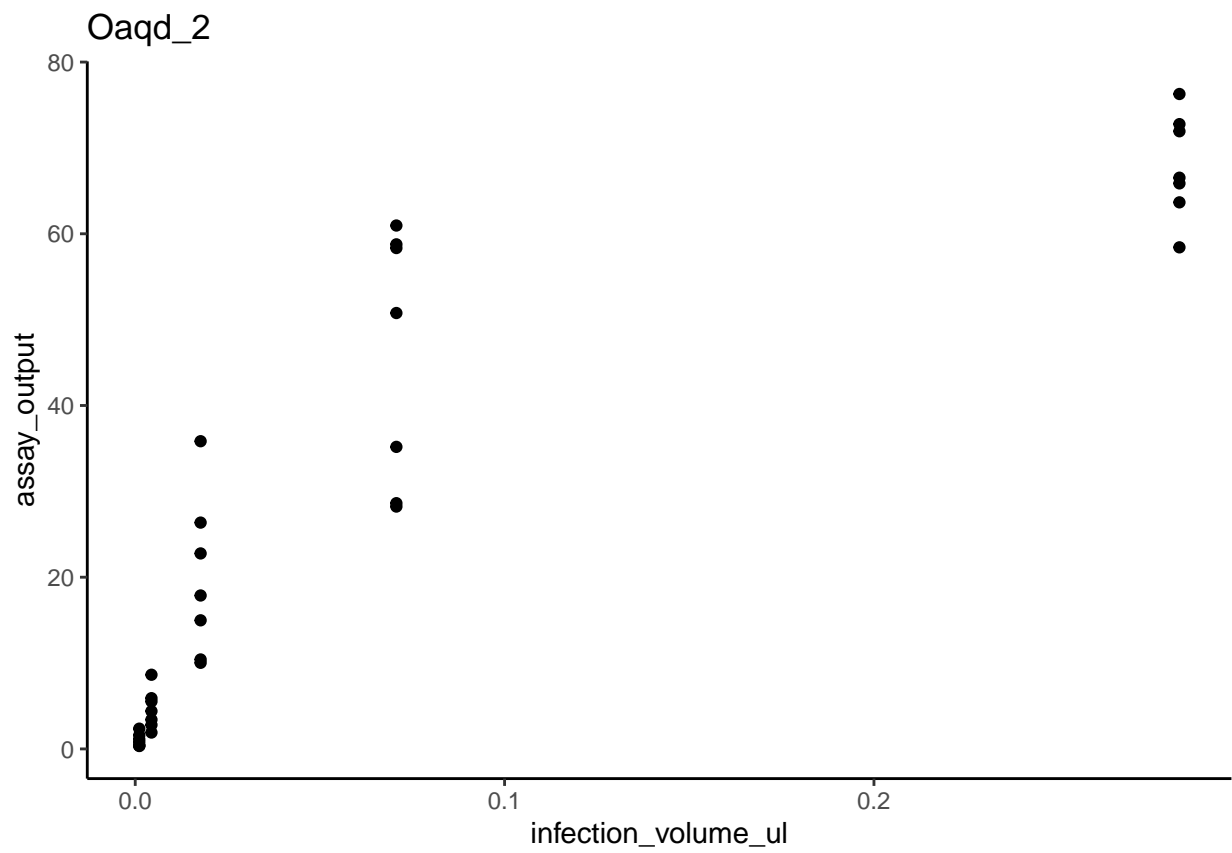




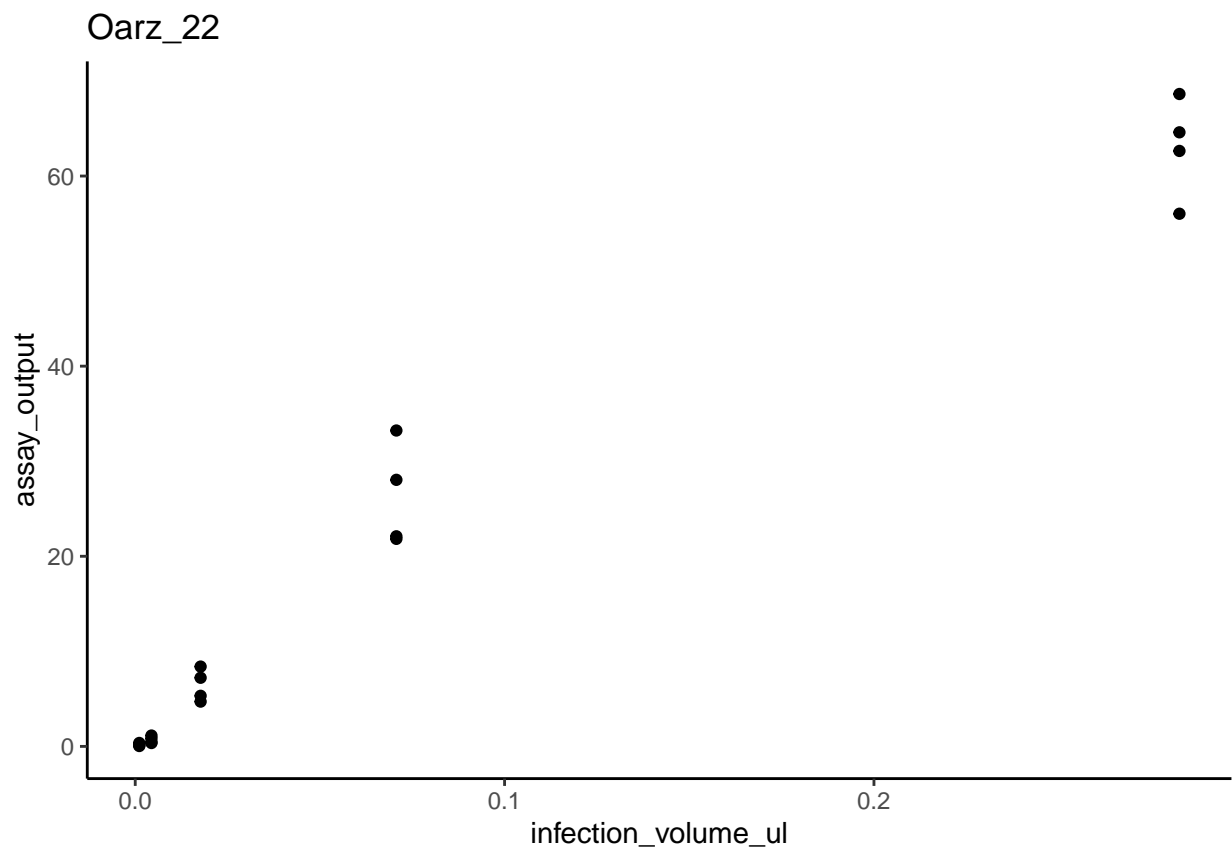
[[78]]



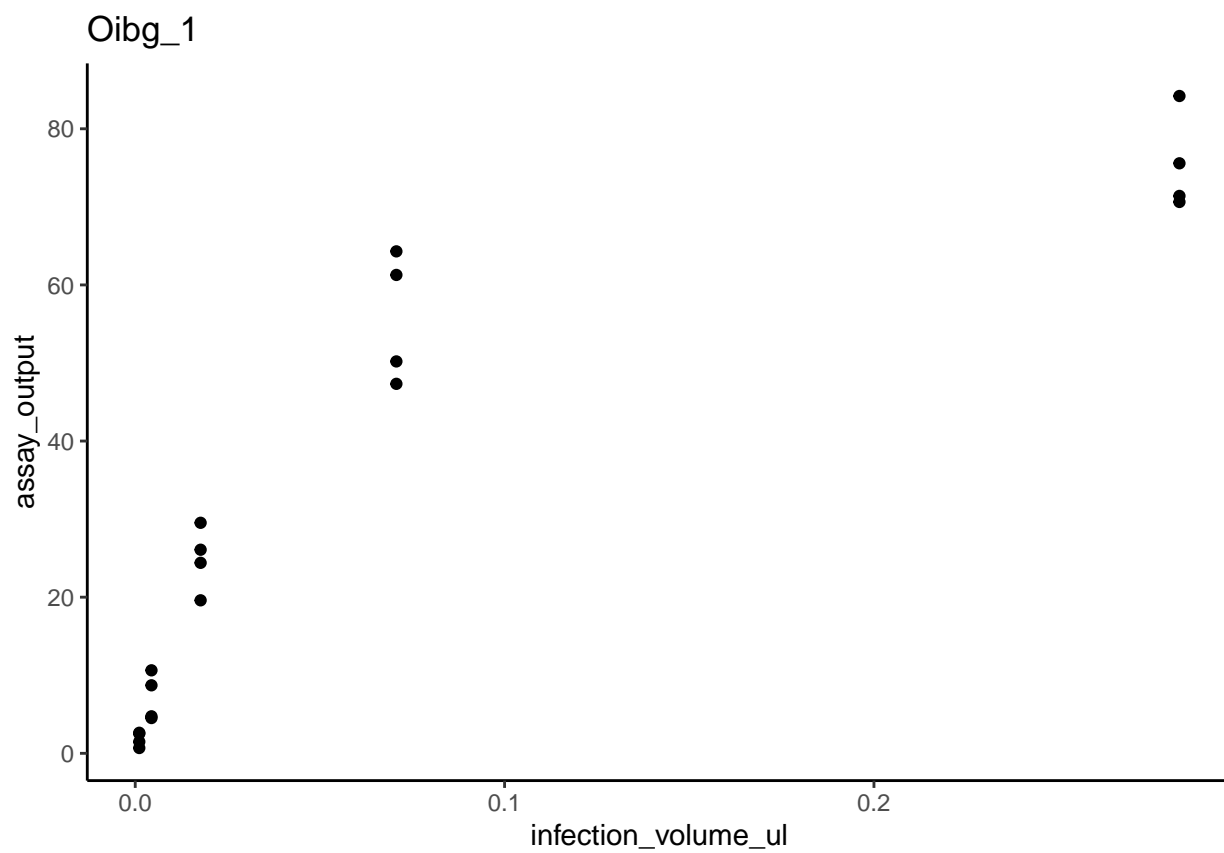
[[79]]



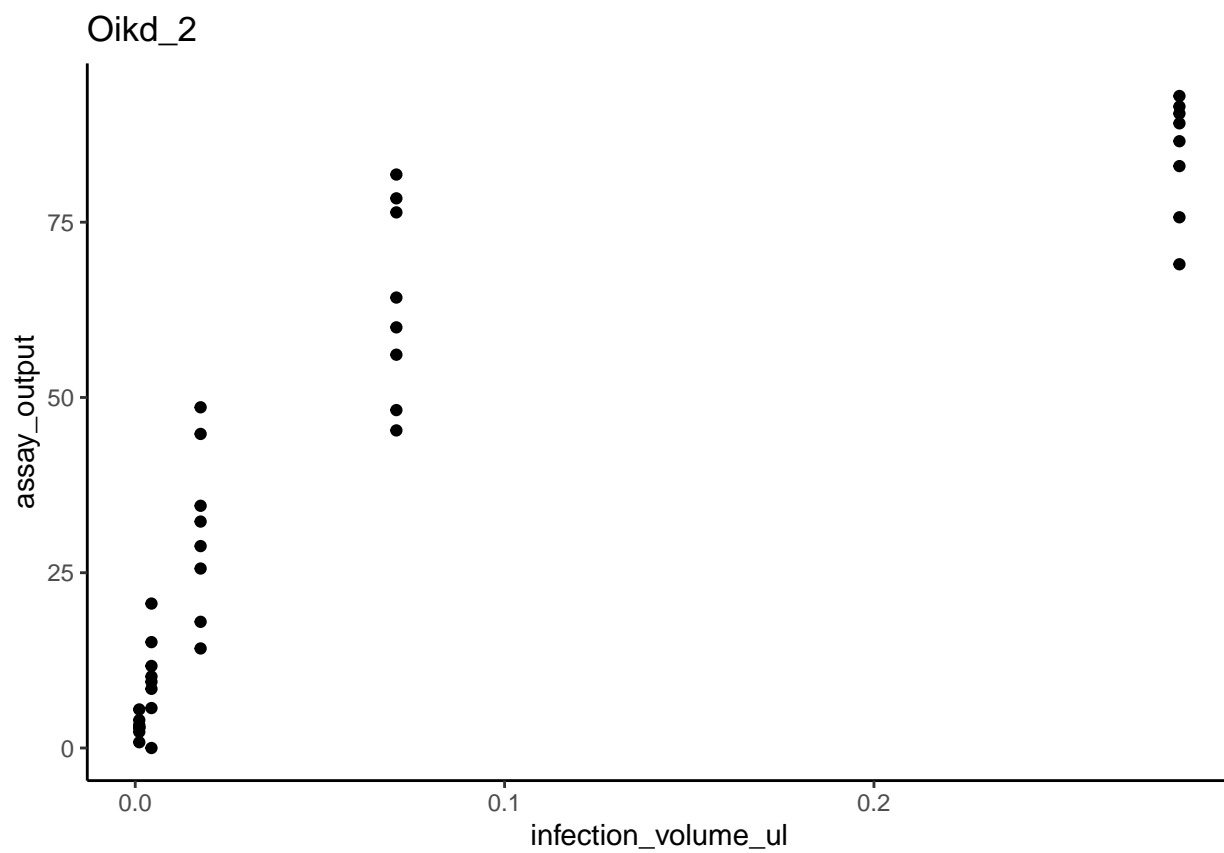
[[80]]



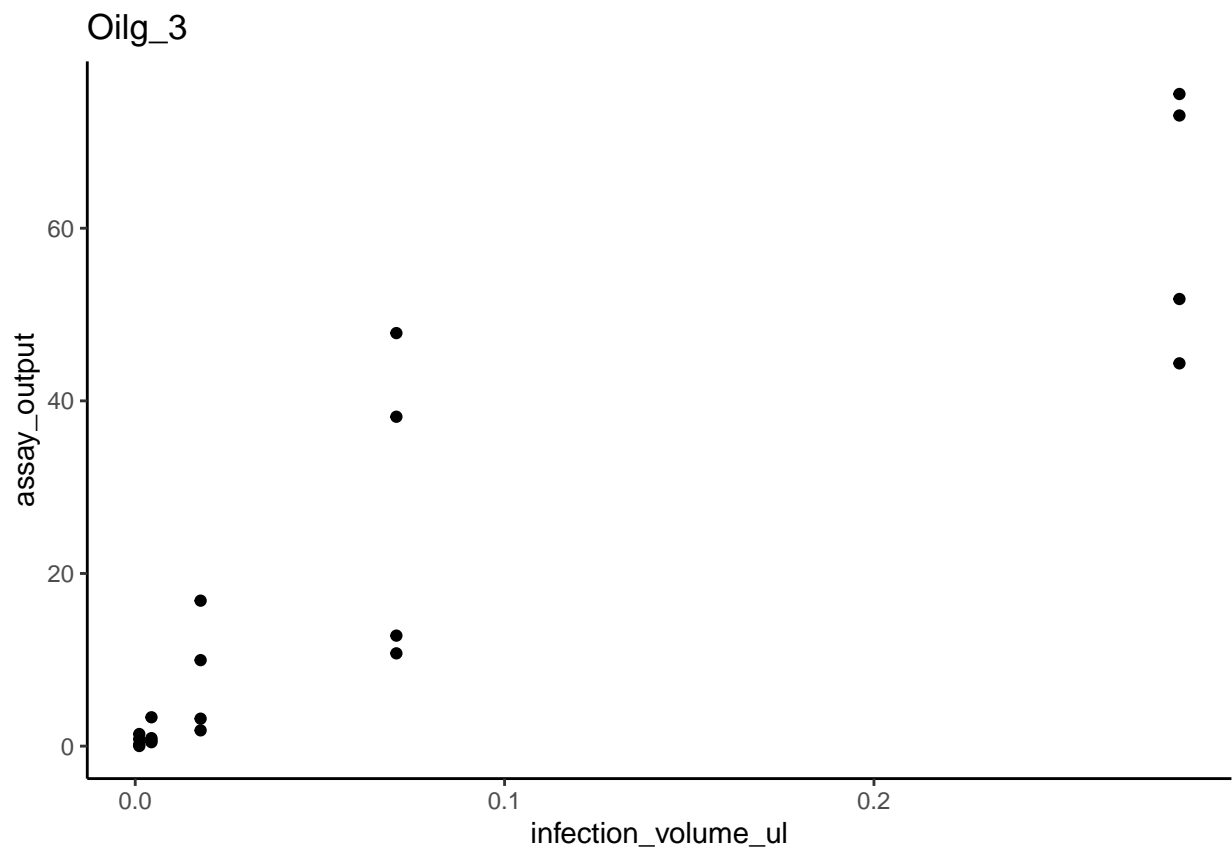
[[81]]



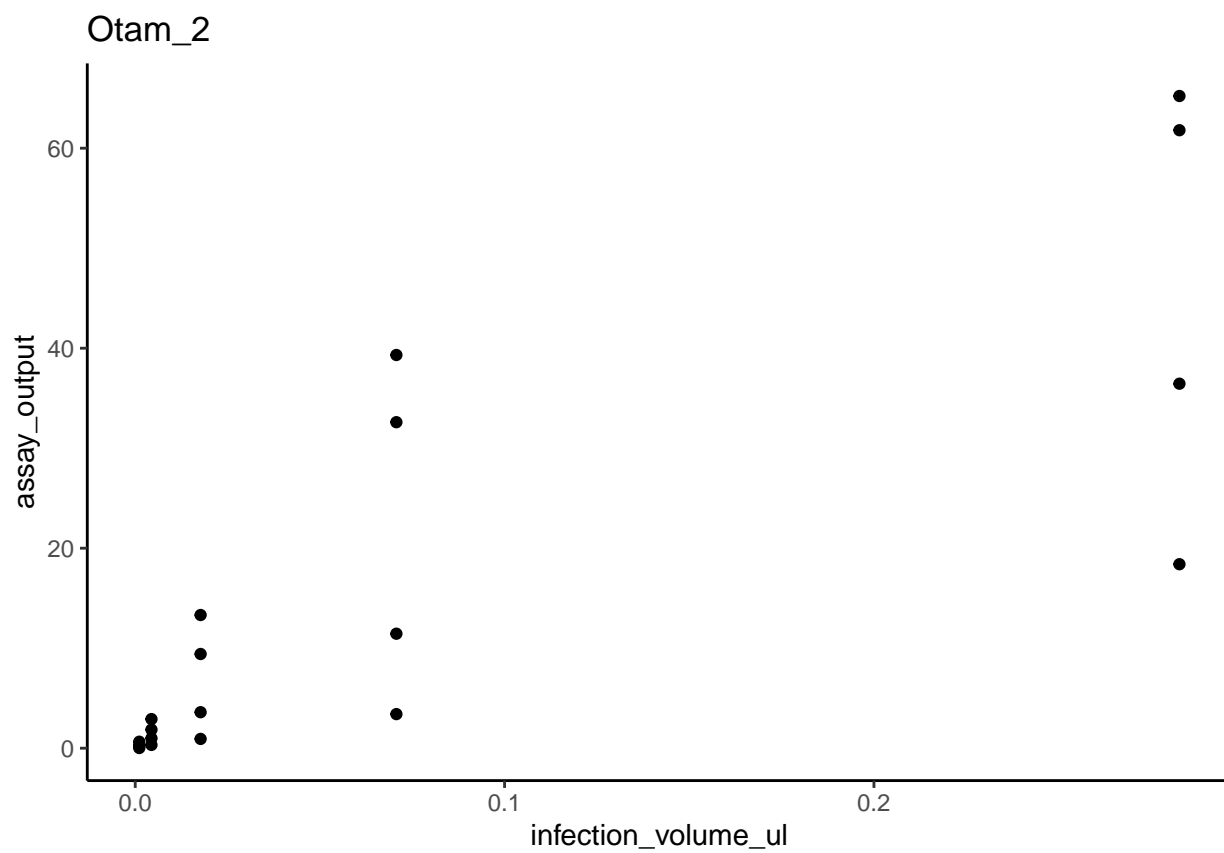
[[82]]



[[83]]

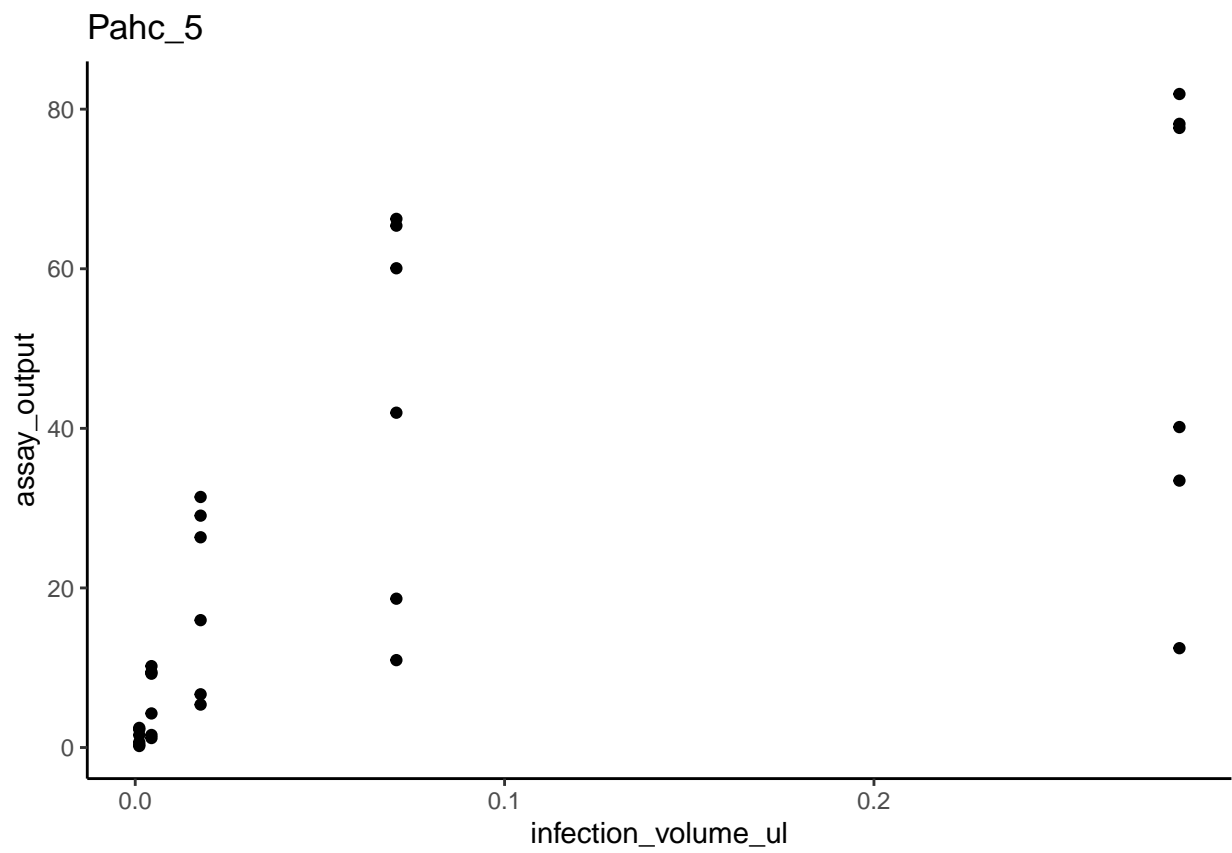


[[84]]

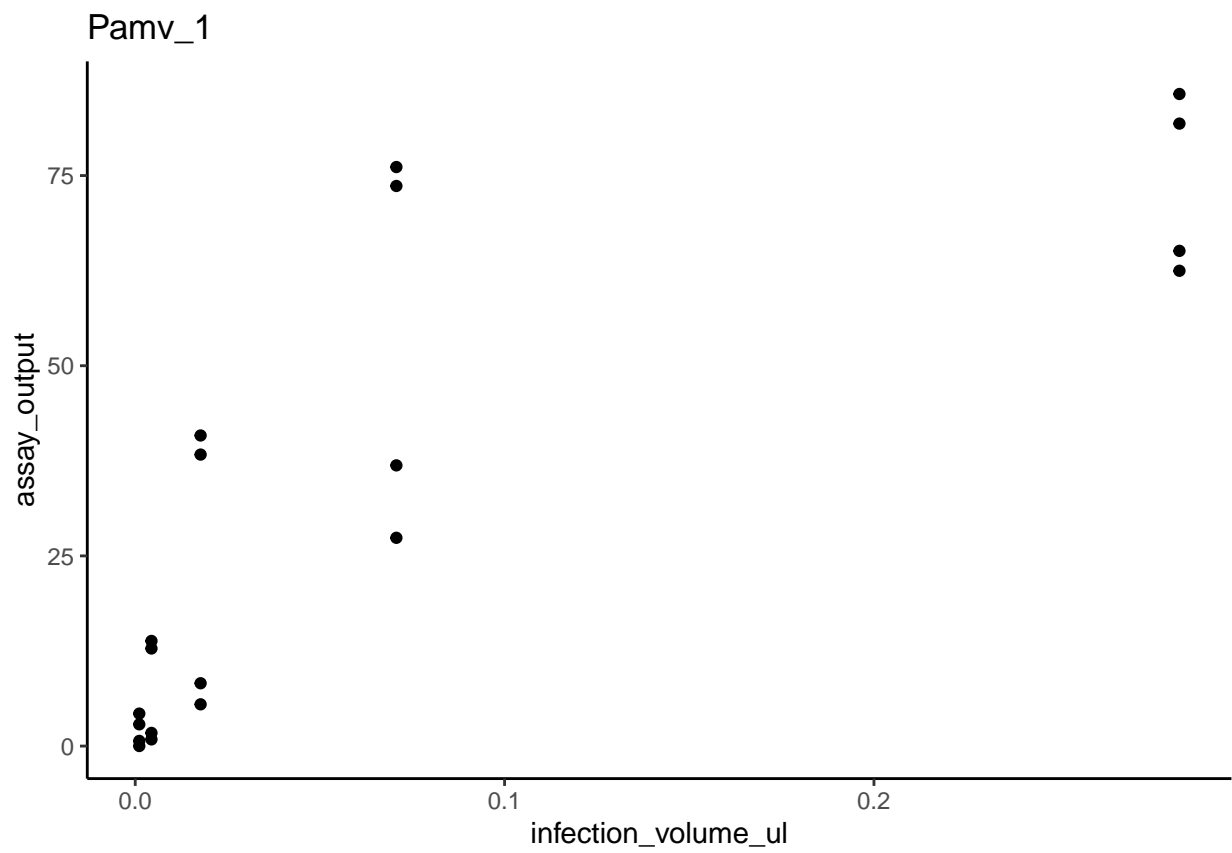


##

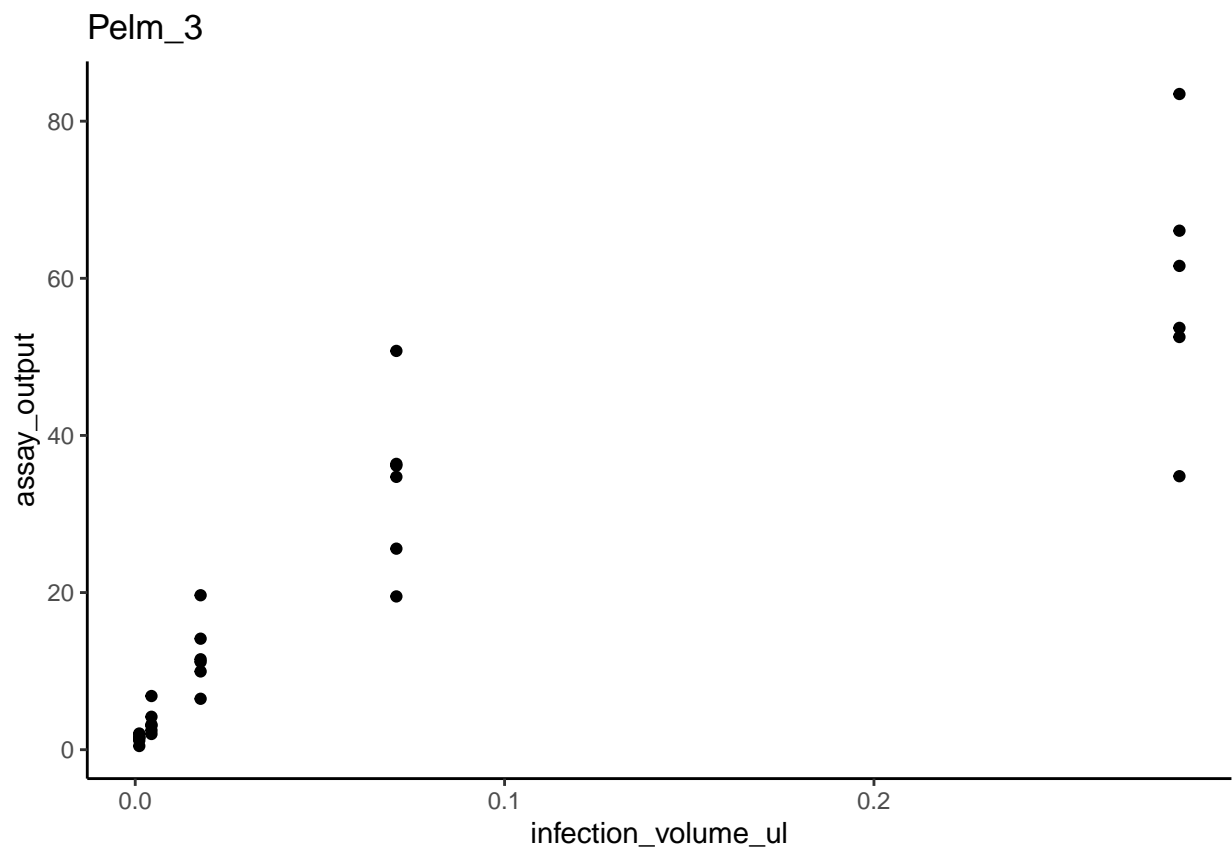
[[85]]



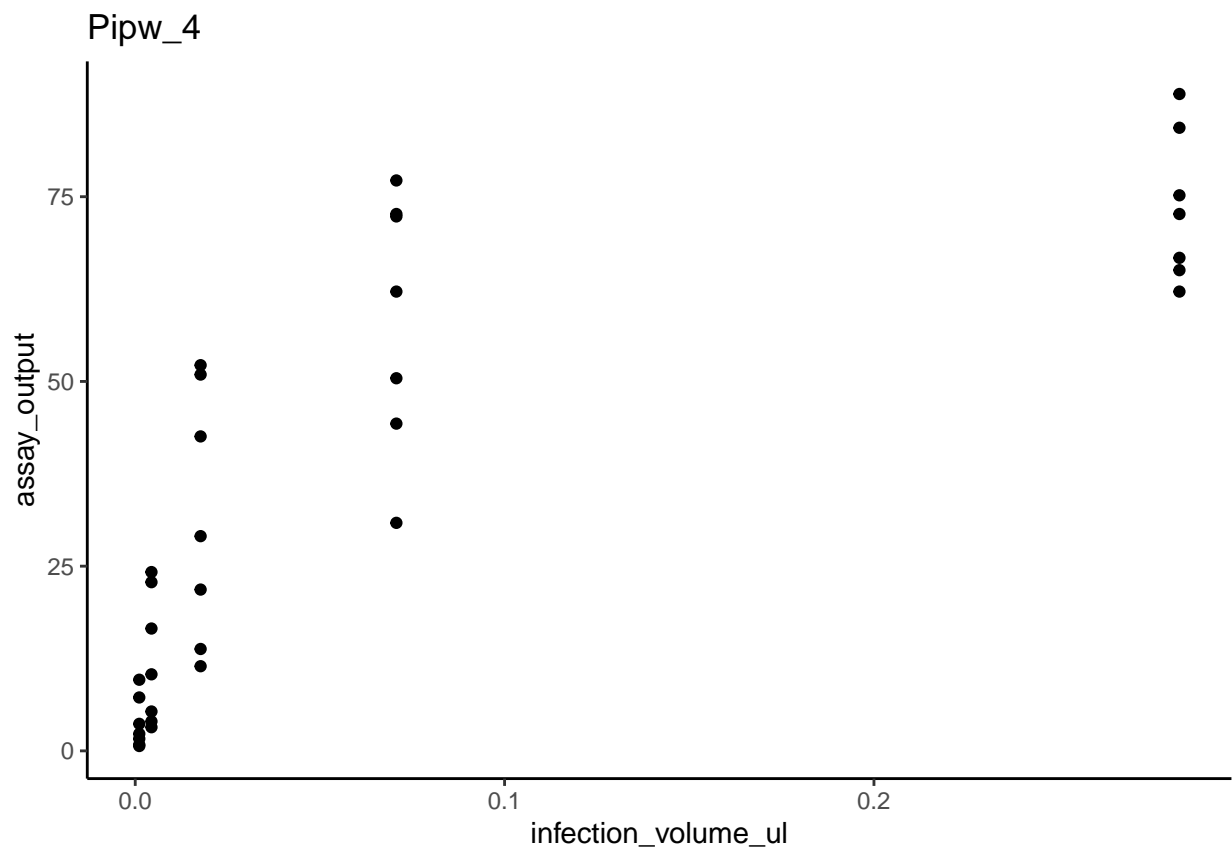
[[87]]



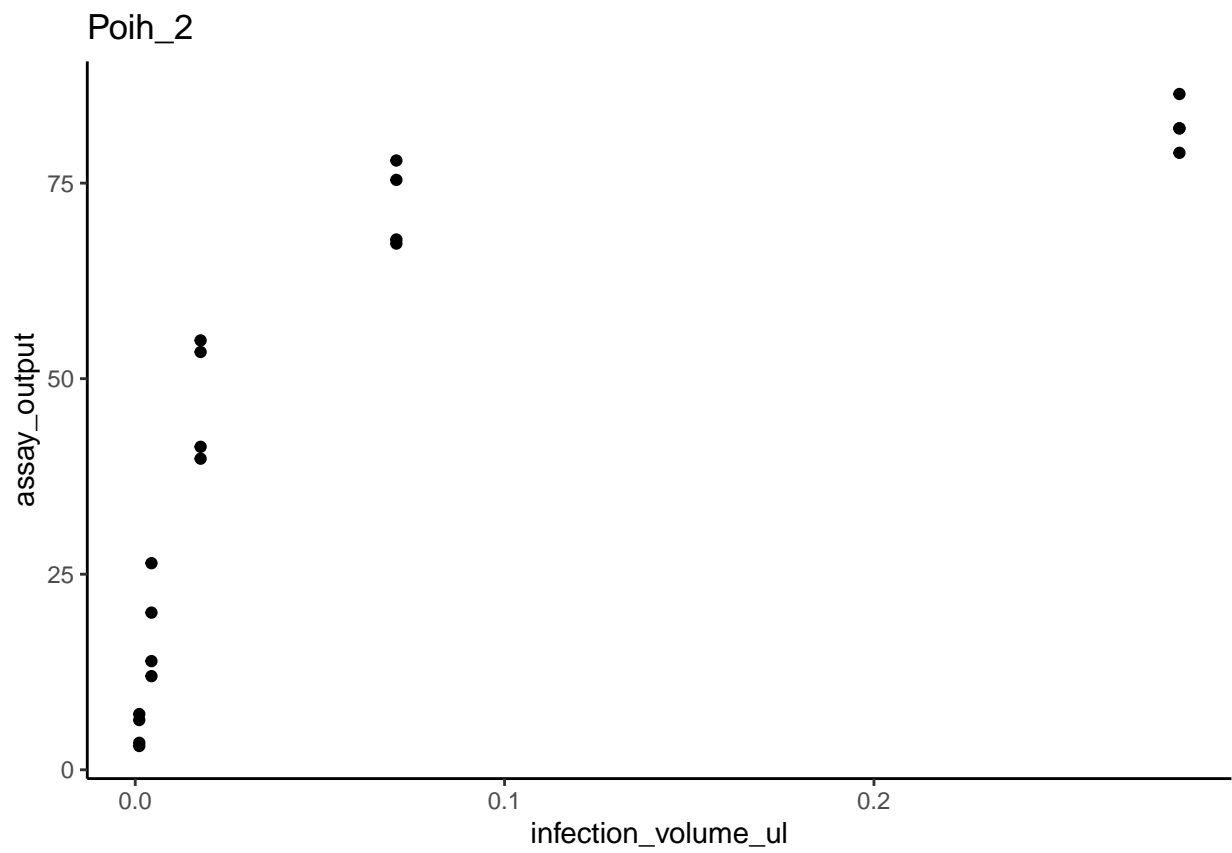
[[88]]



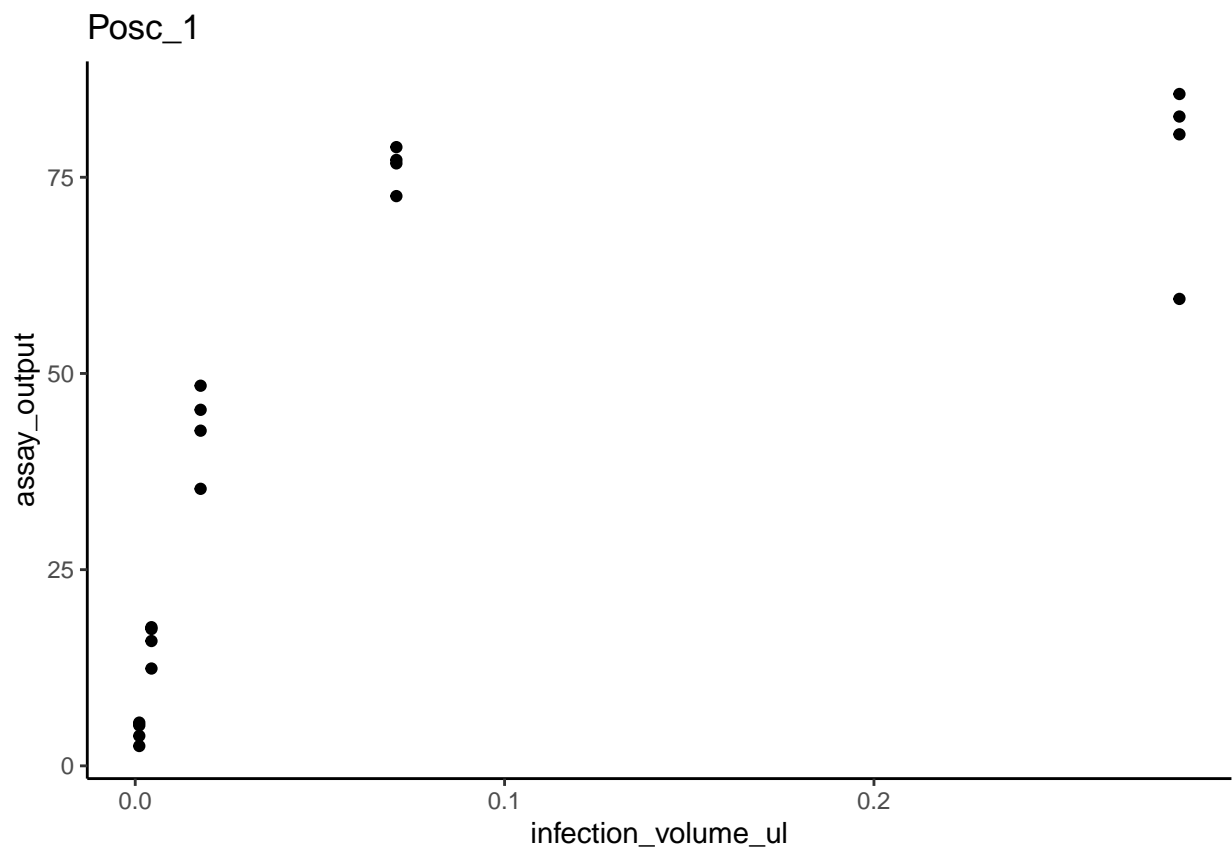
[[89]]



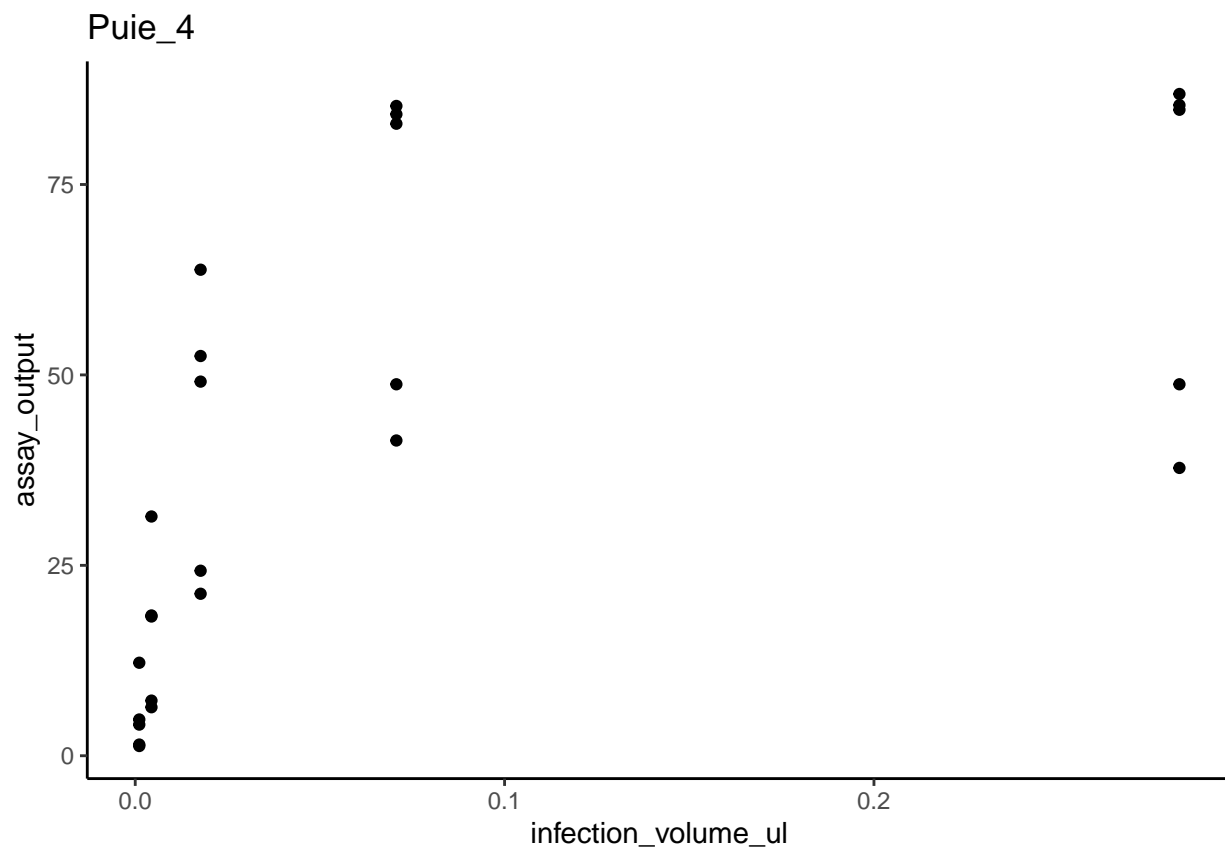
[[90]]



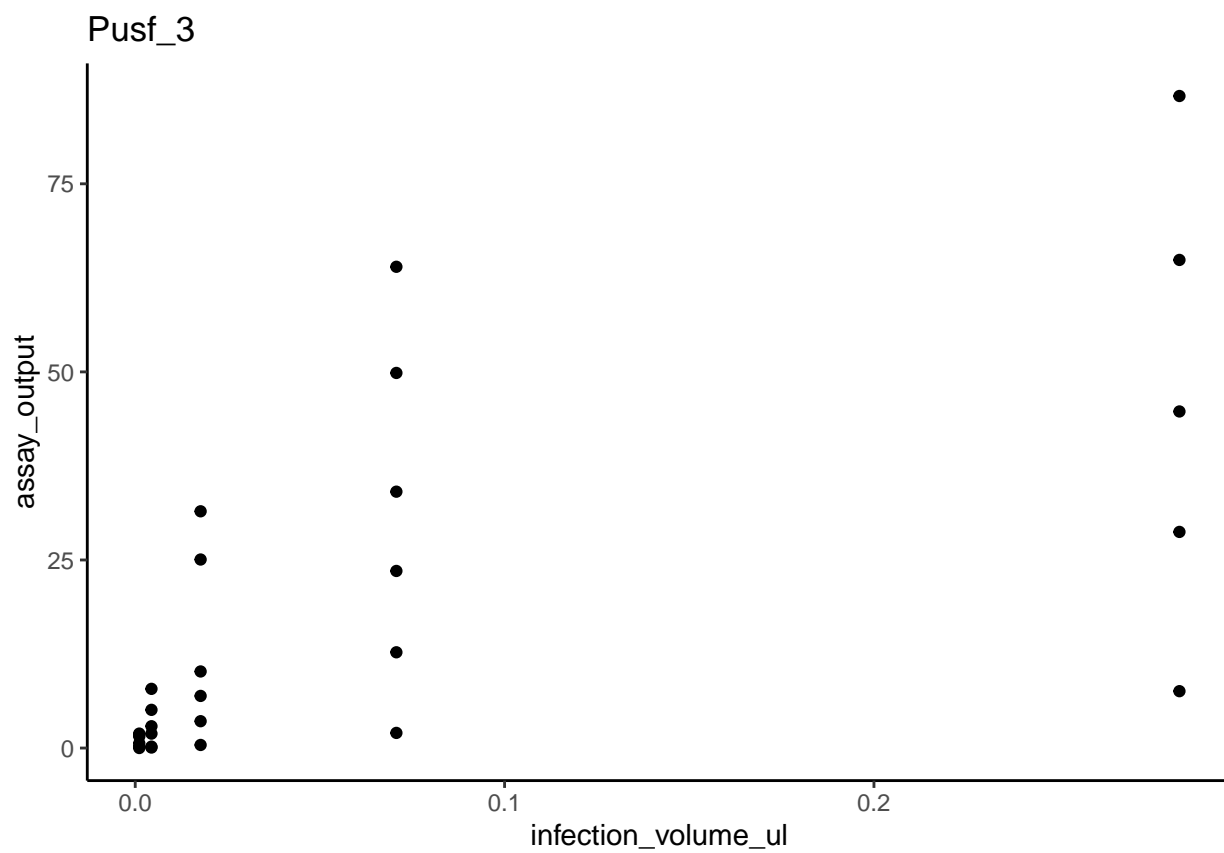
```
##  
## [[91]]
```



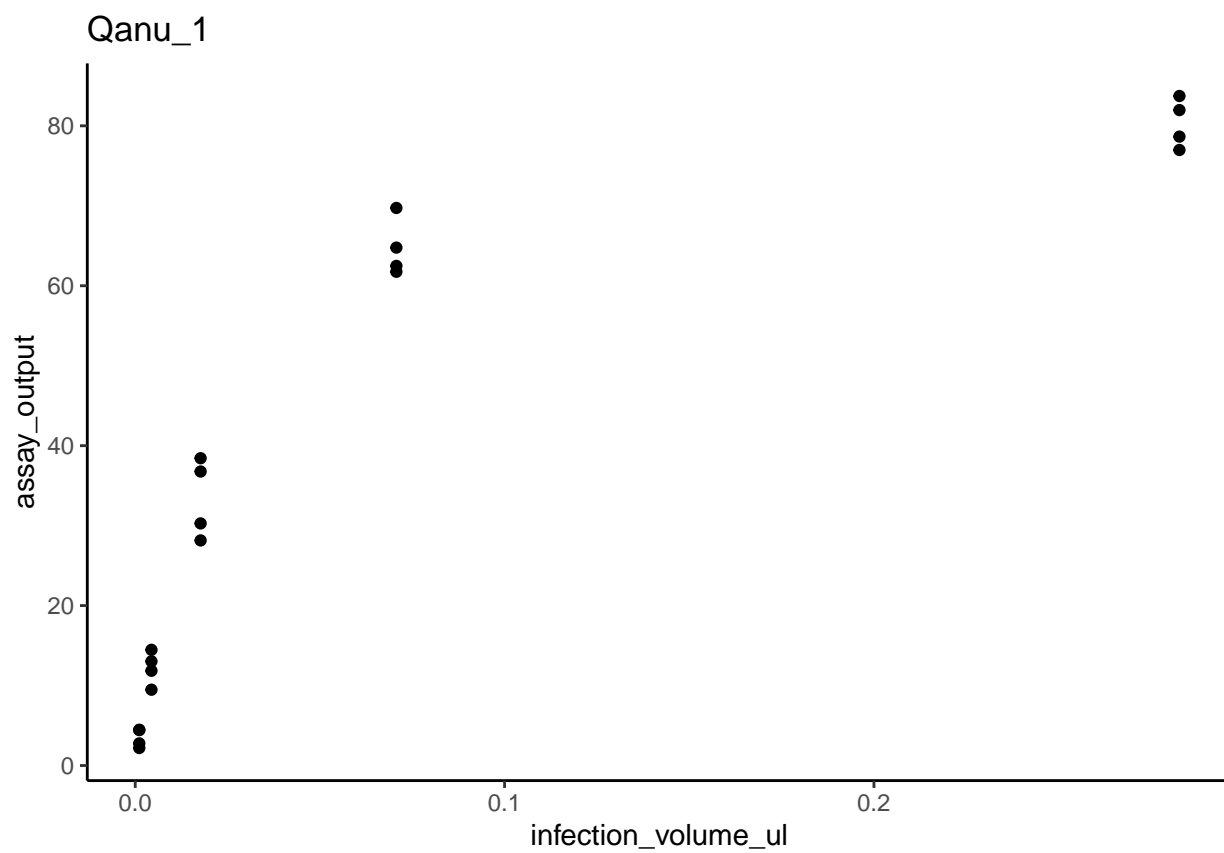
[[92]]



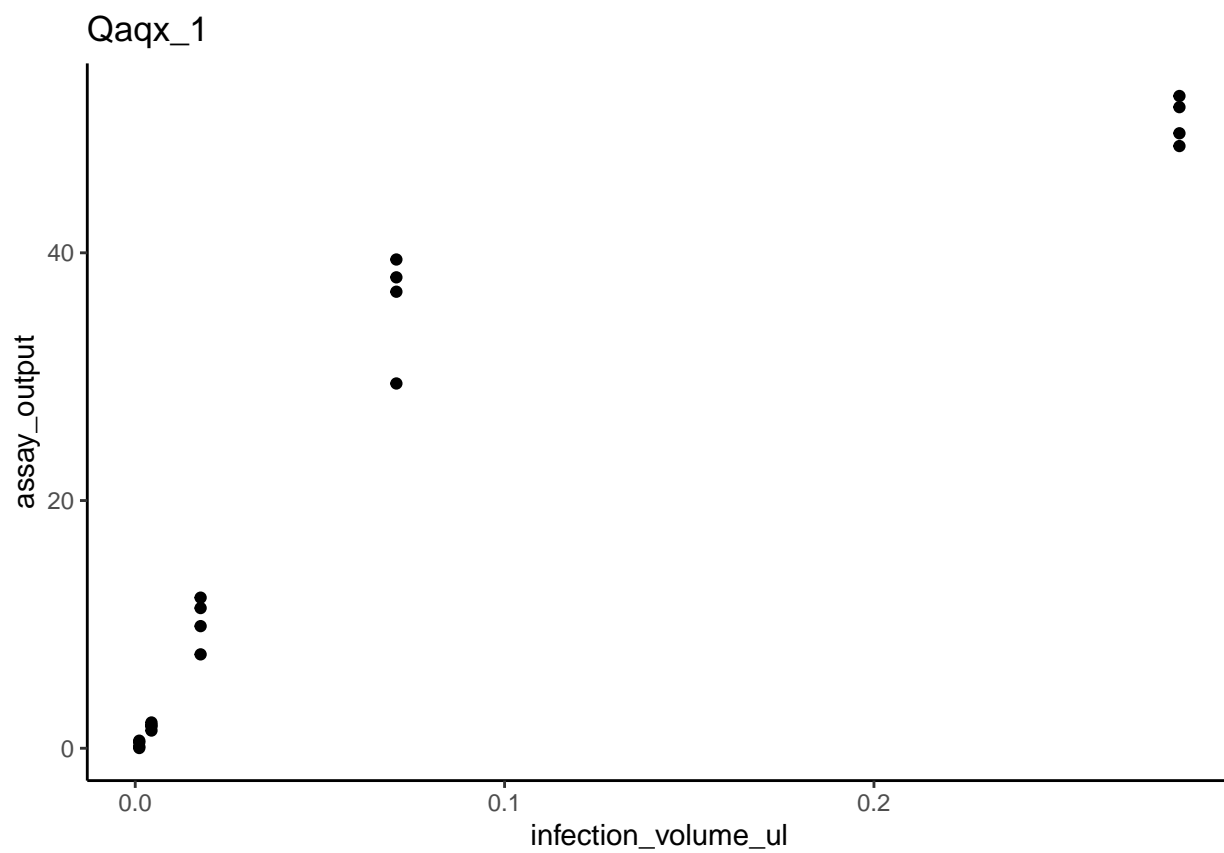
[[93]]



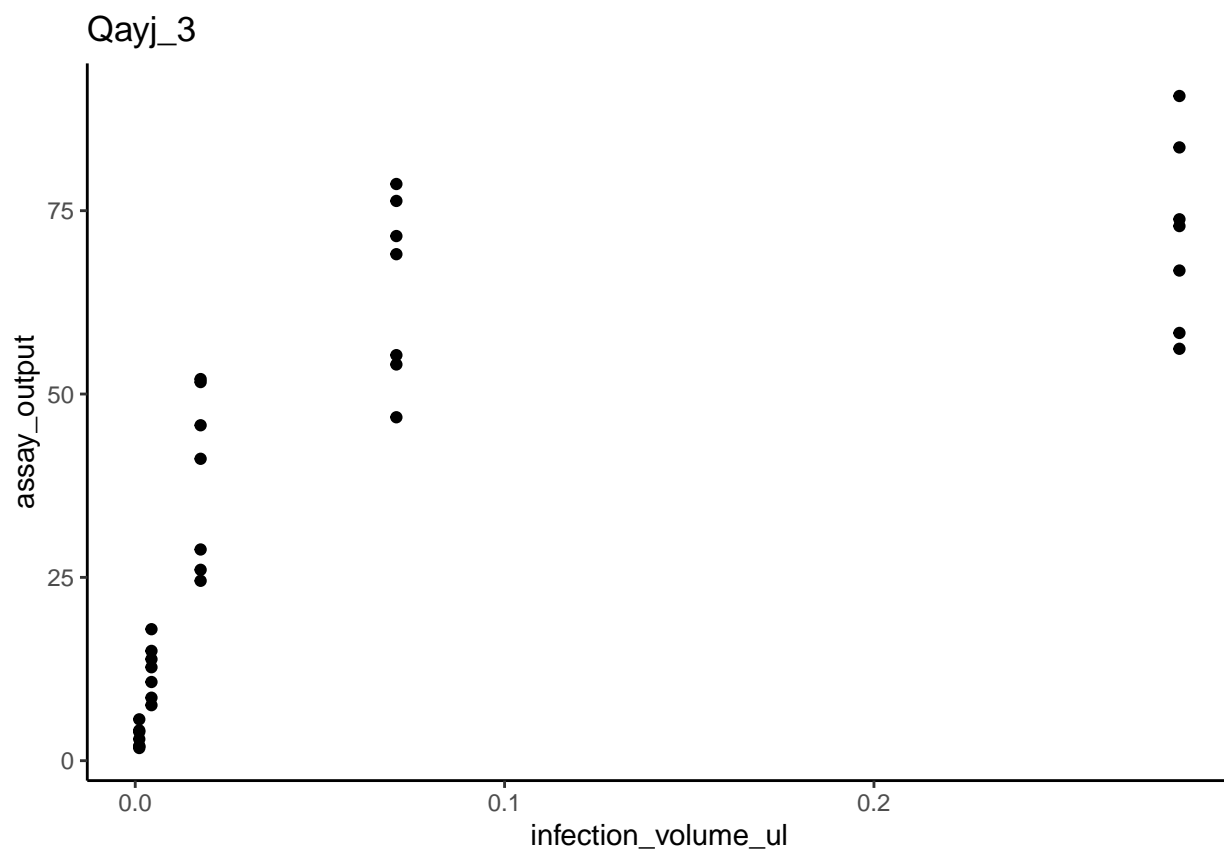
[[94]]



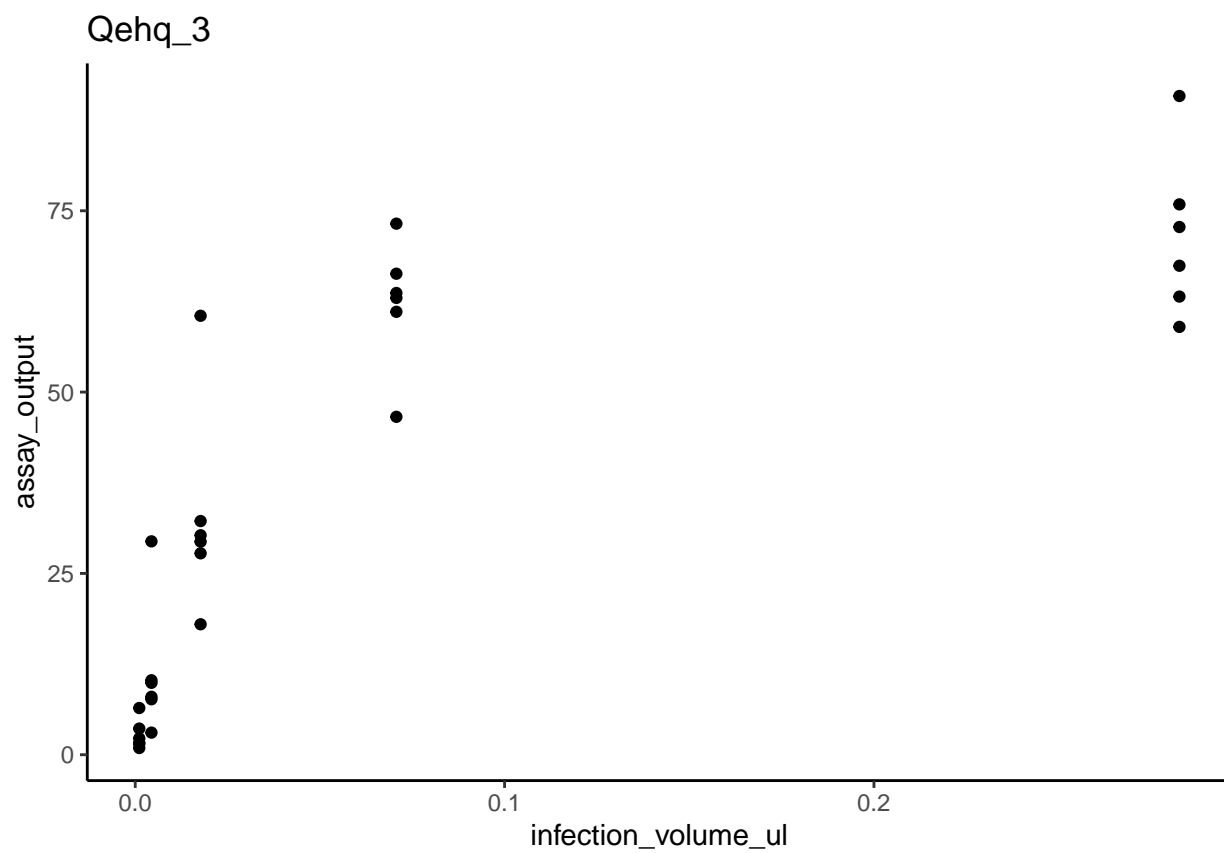
[[95]]



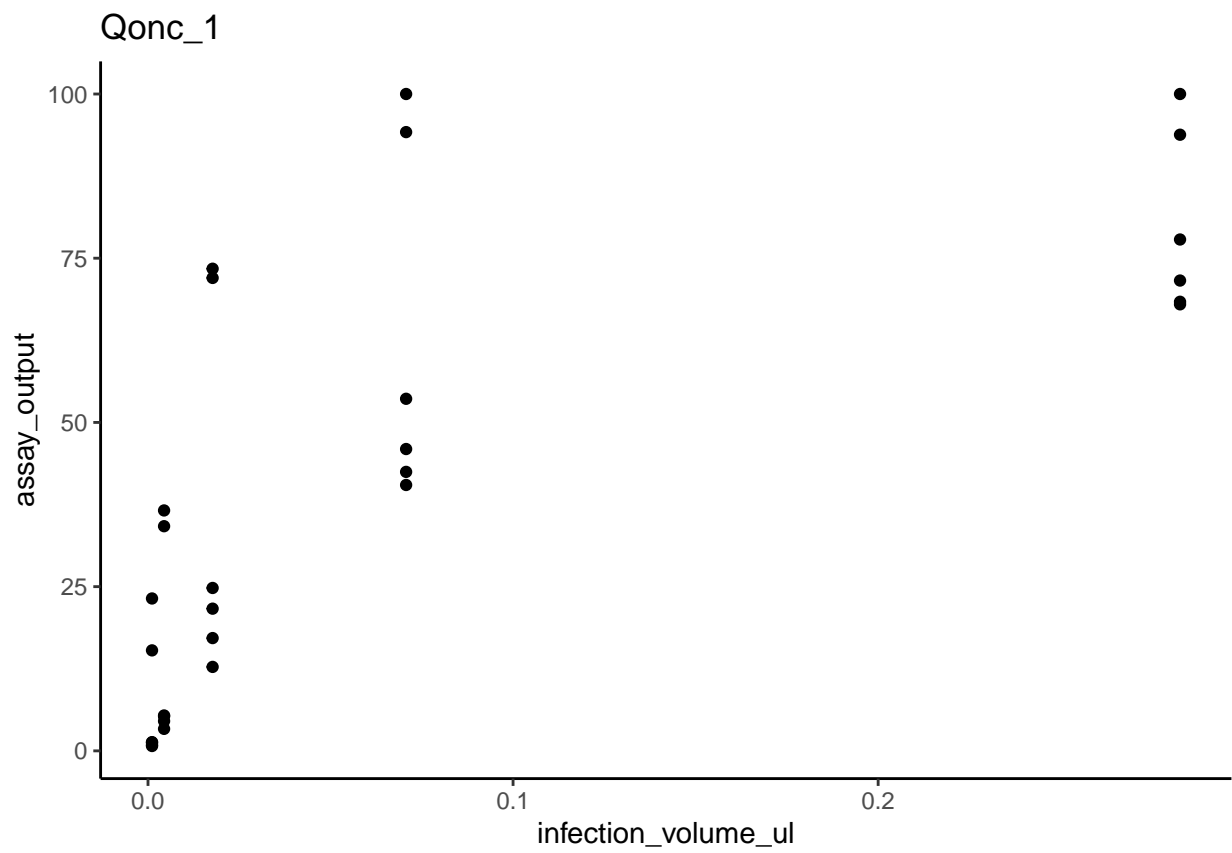
[[96]]



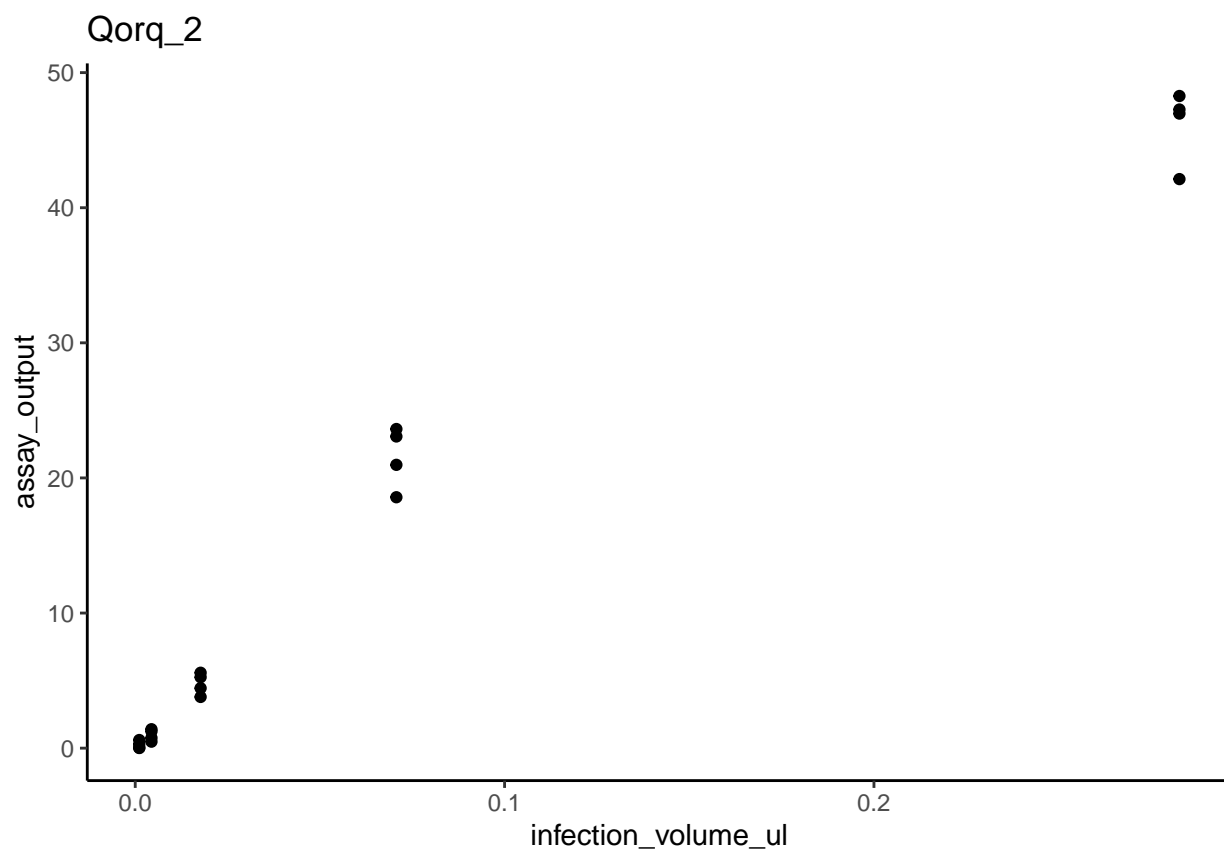
[[97]]



[[98]]

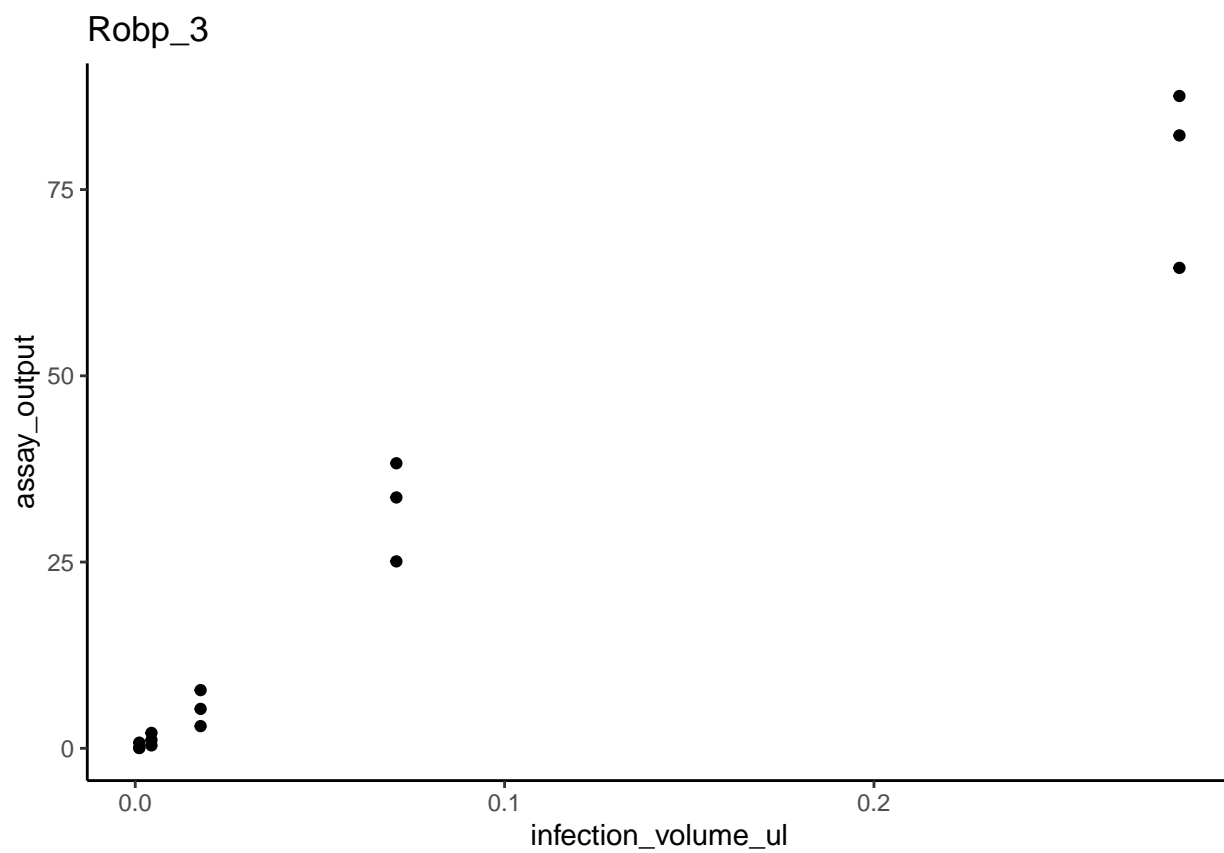


[[99]]

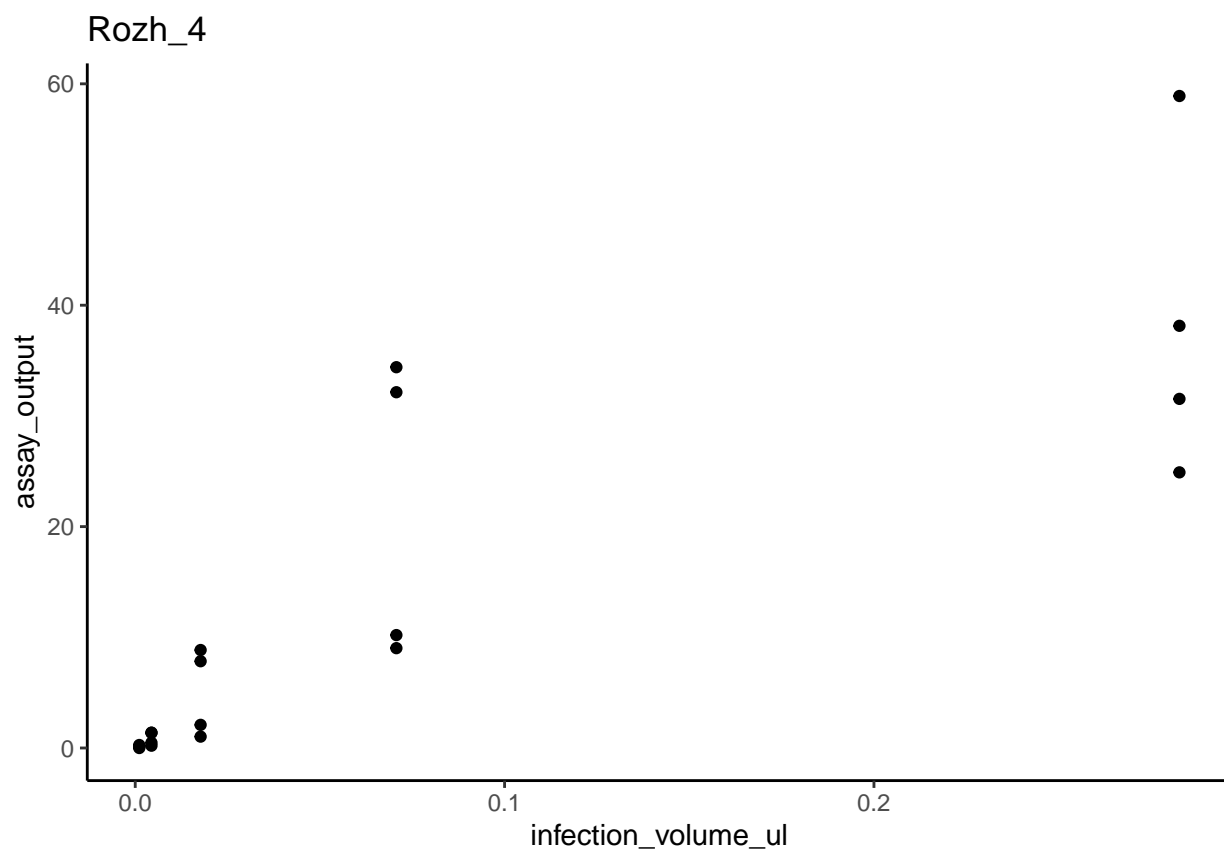


##

[[100]]

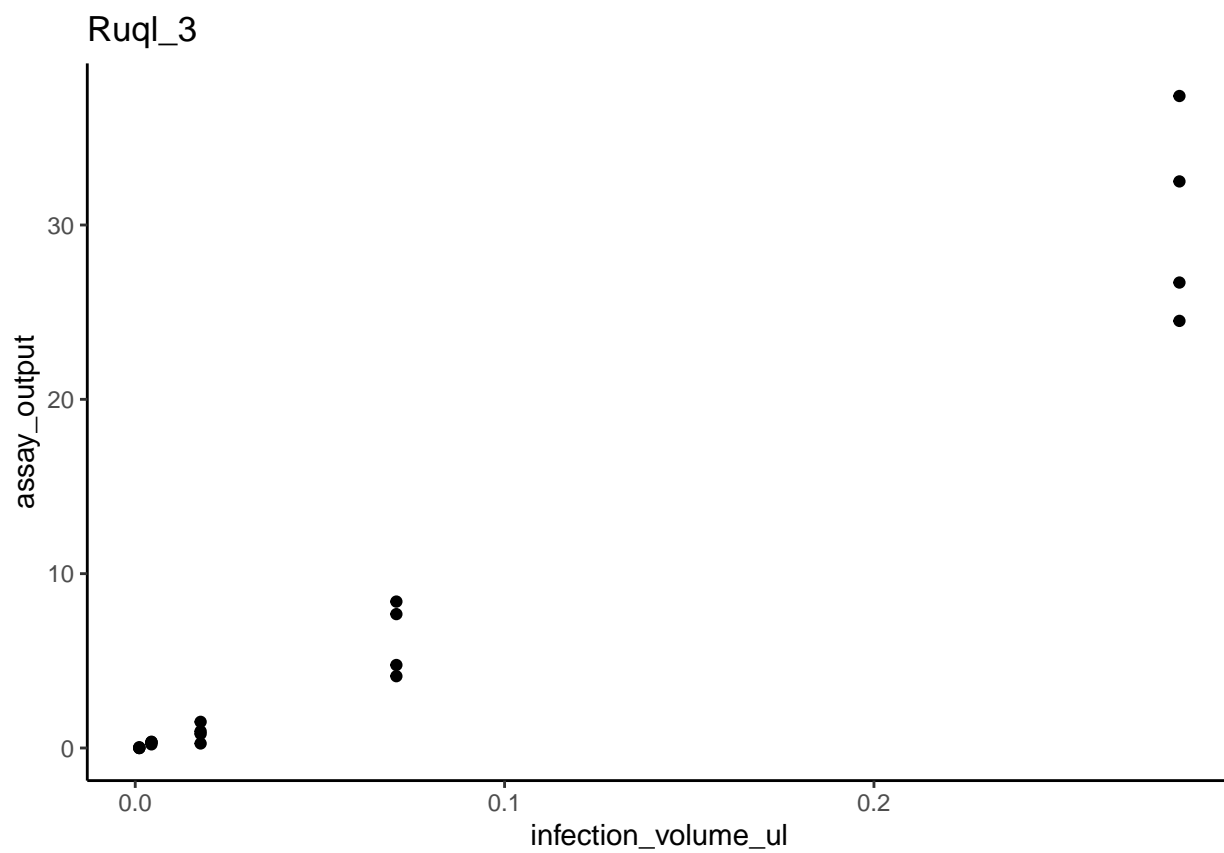


[[102]]



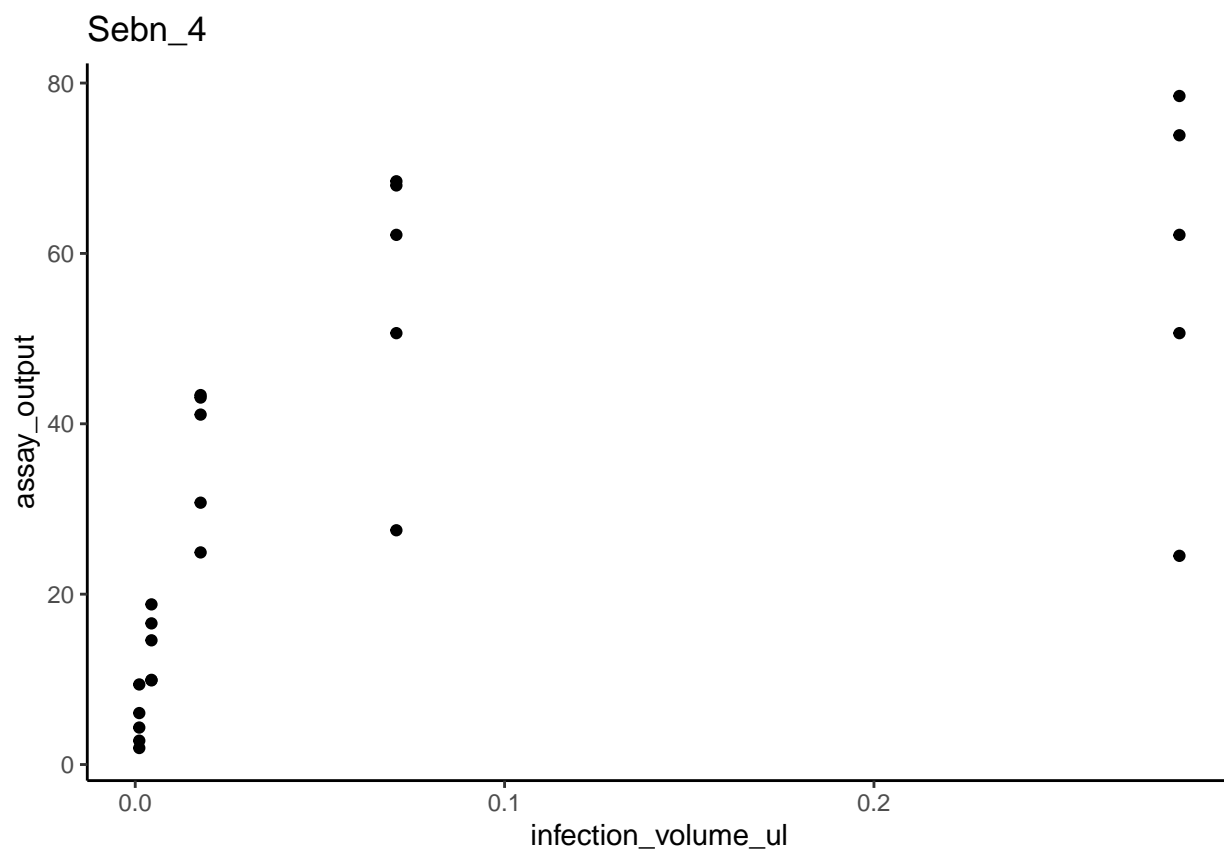
##

[[104]]



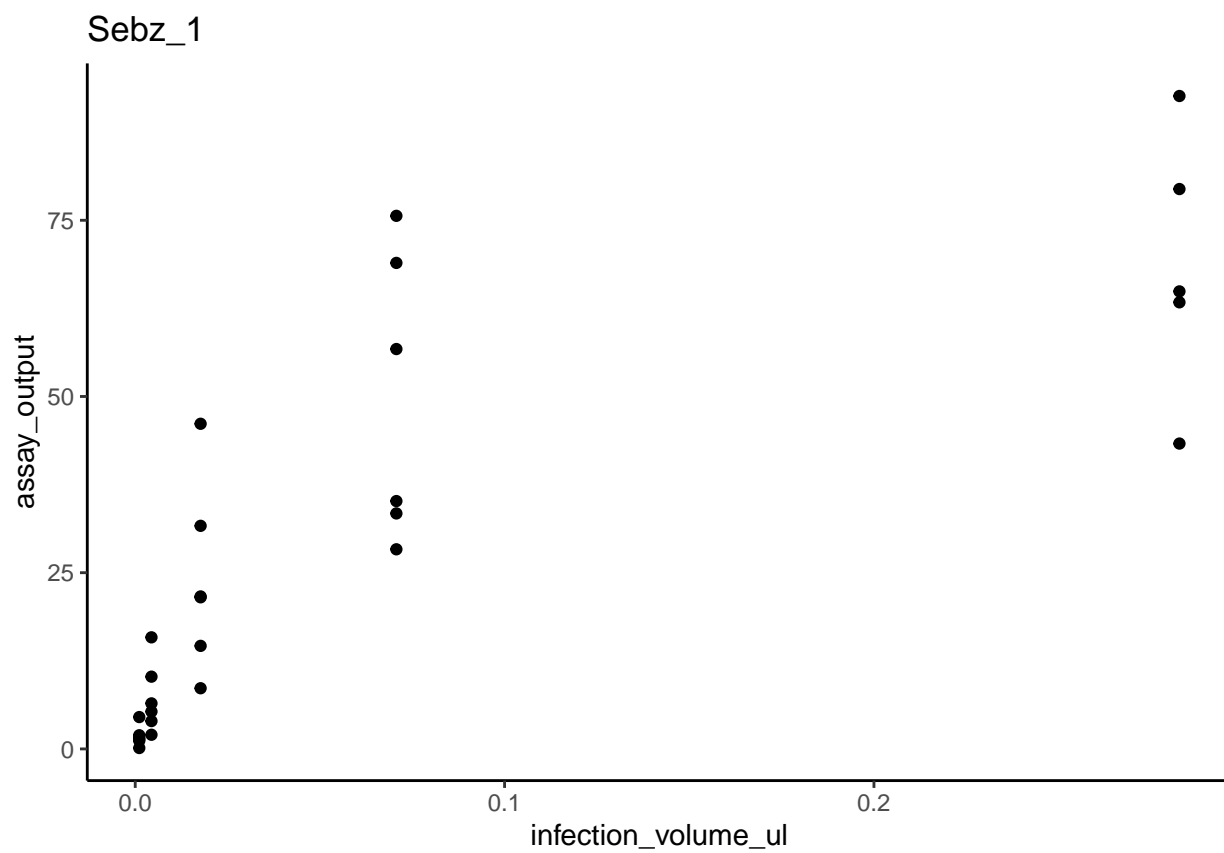
##

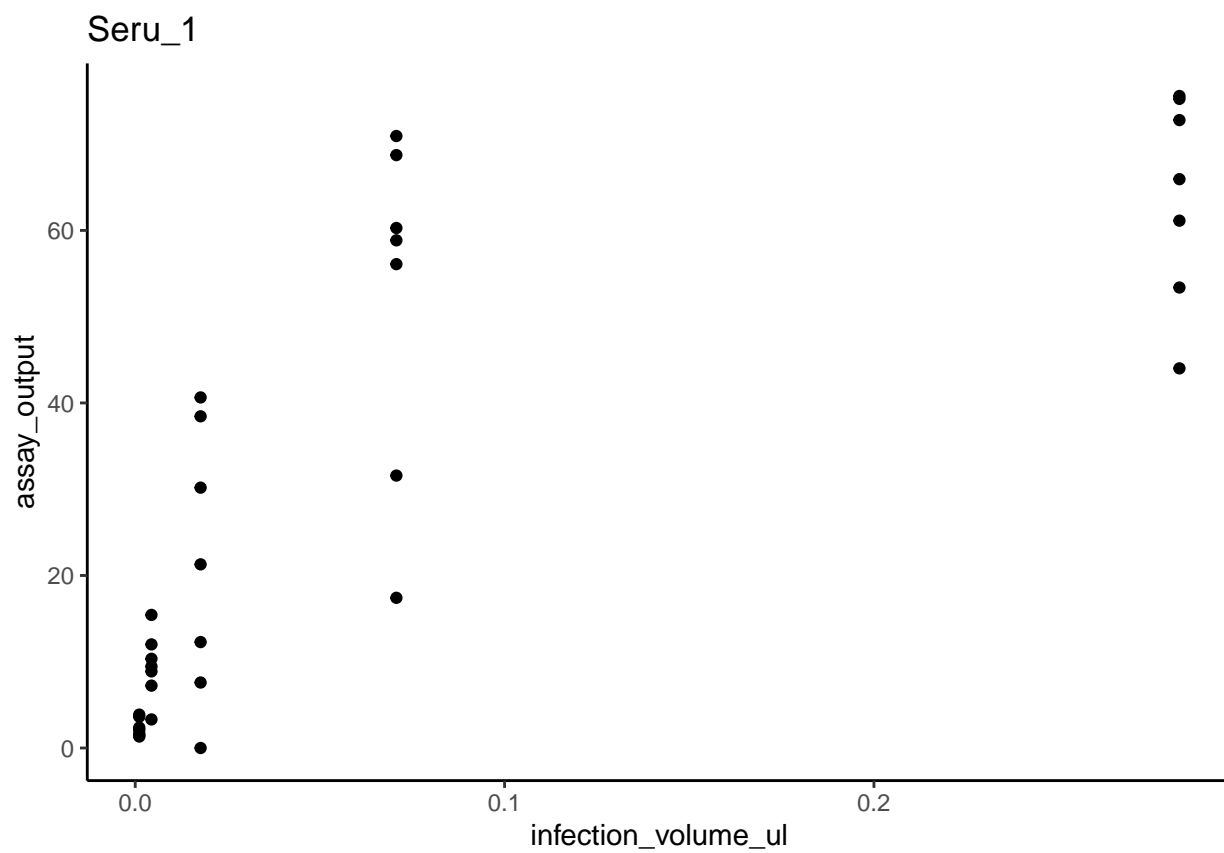
[[105]]



##

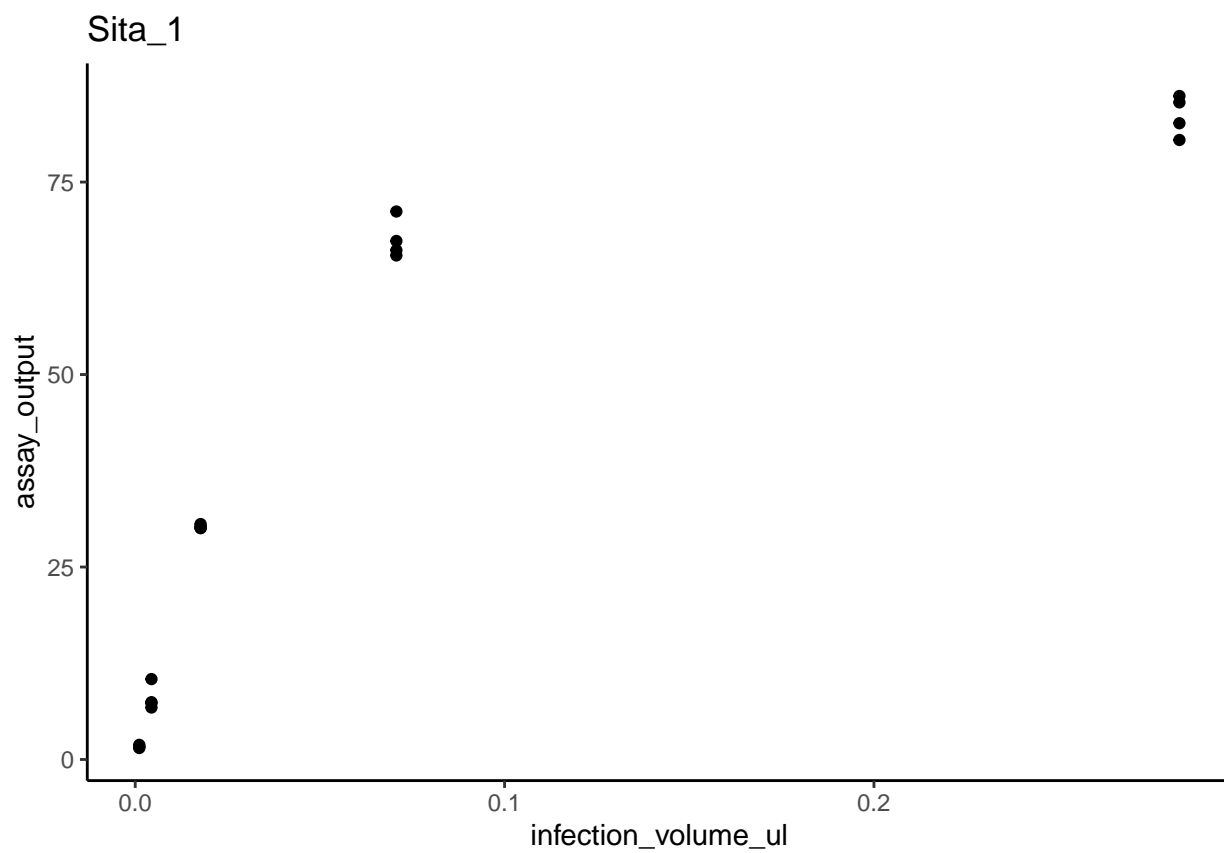
[[106]]



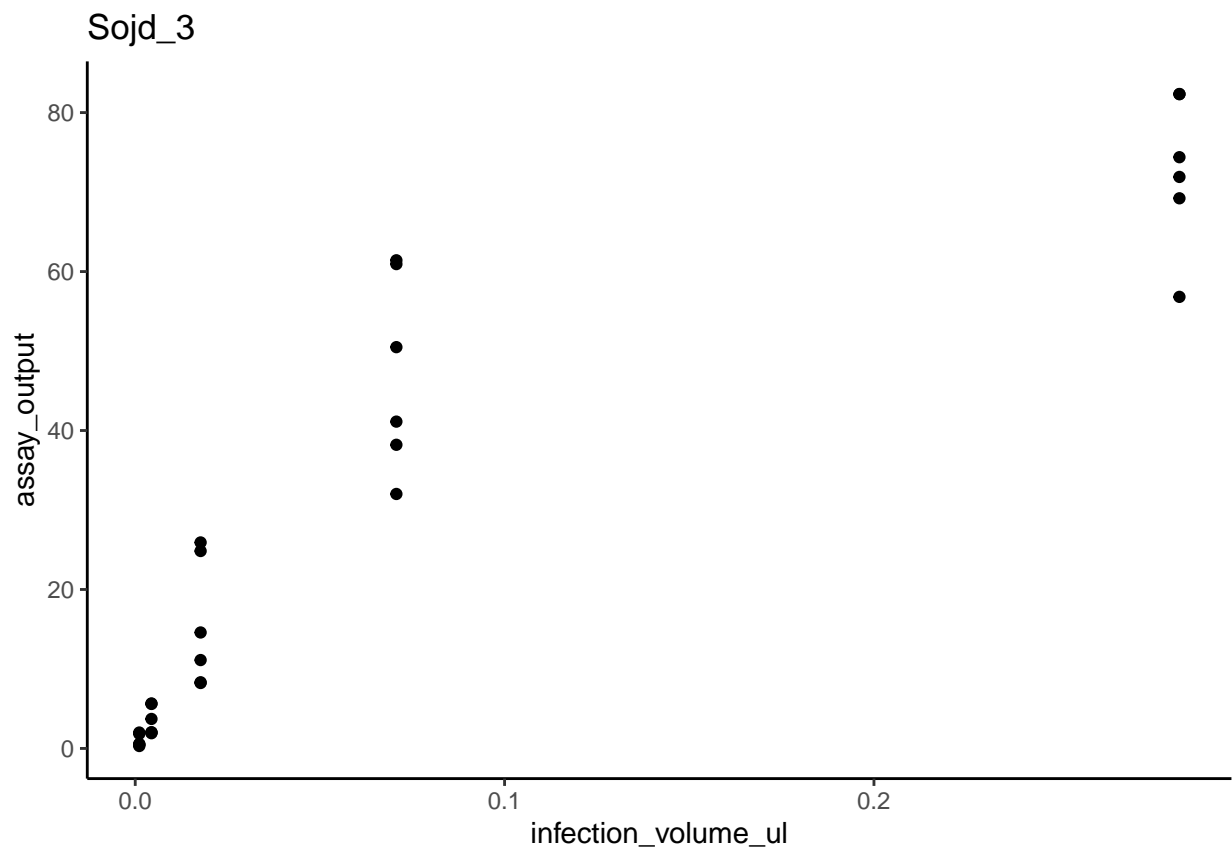


##

```
## [[109]]
```

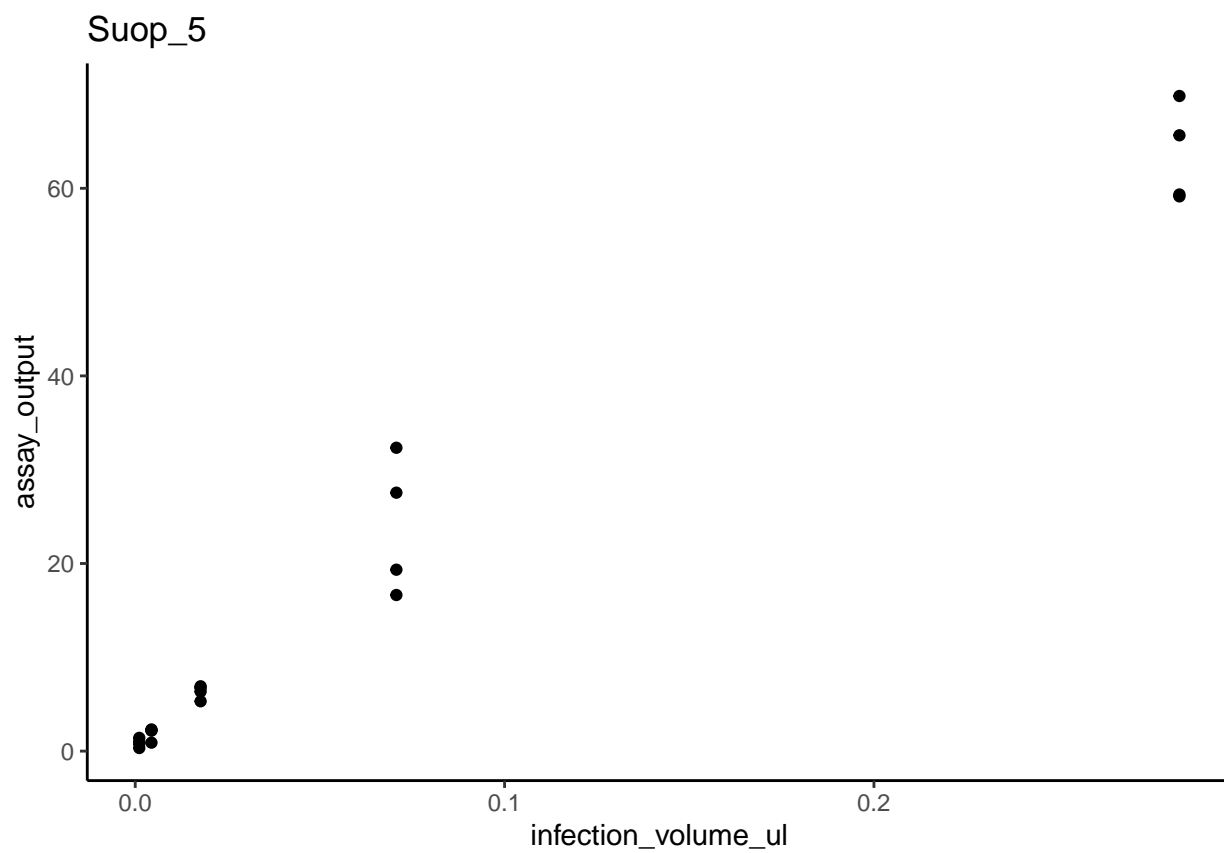


[[110]]

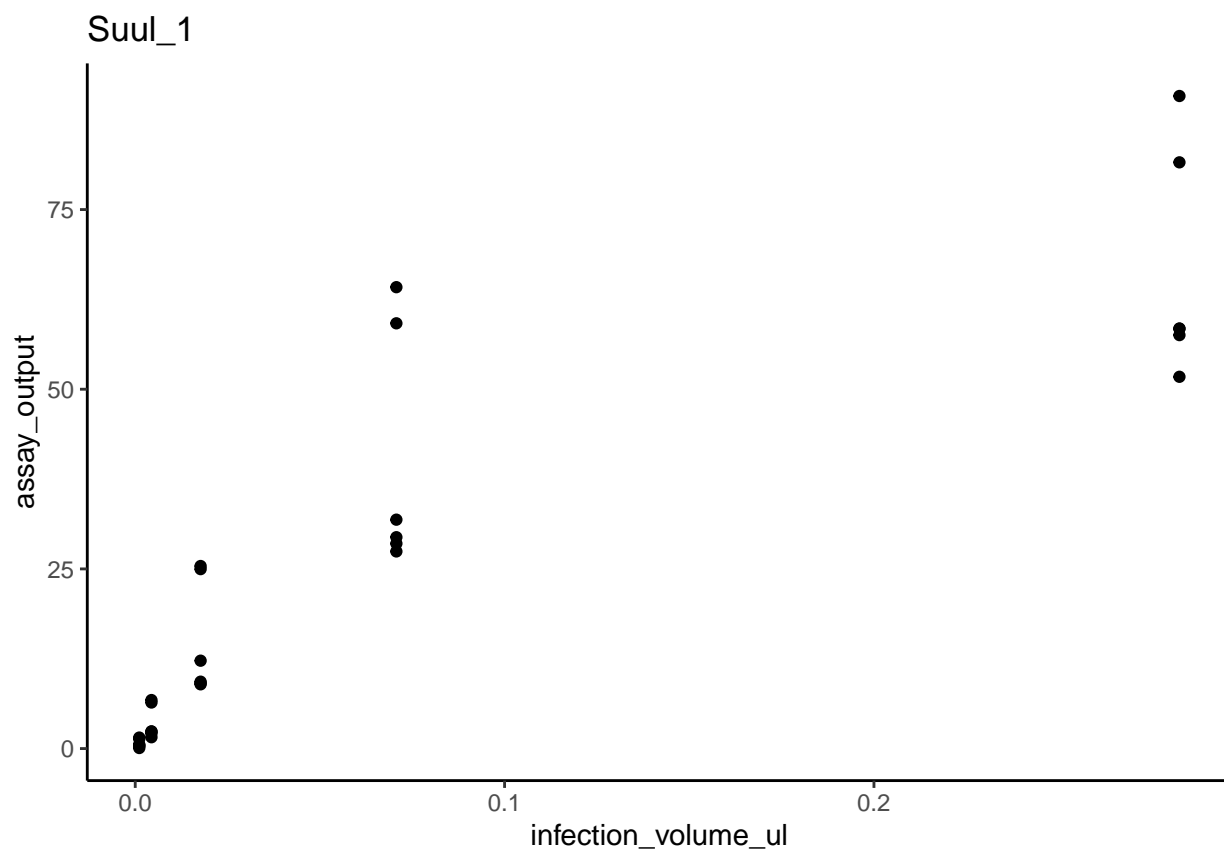


##

[[111]]

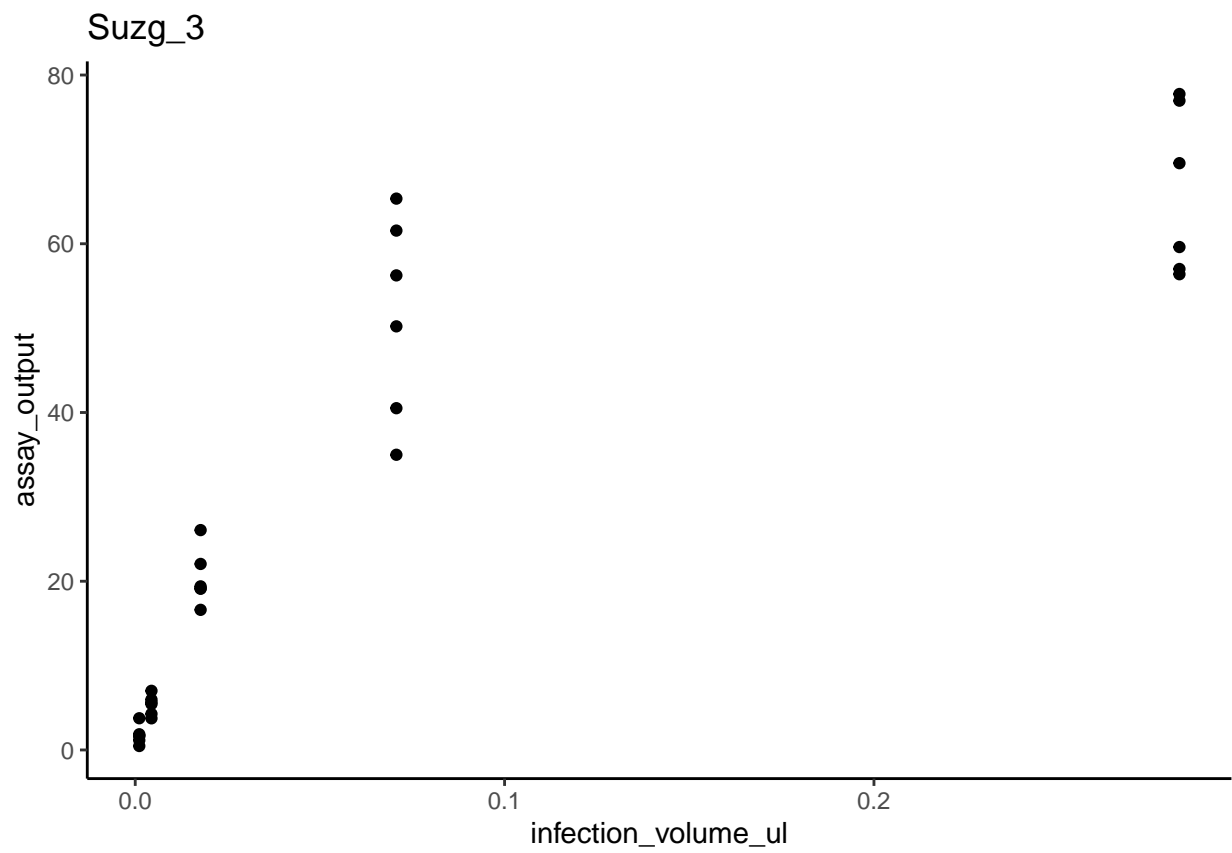


```
##  
## [[114]]
```

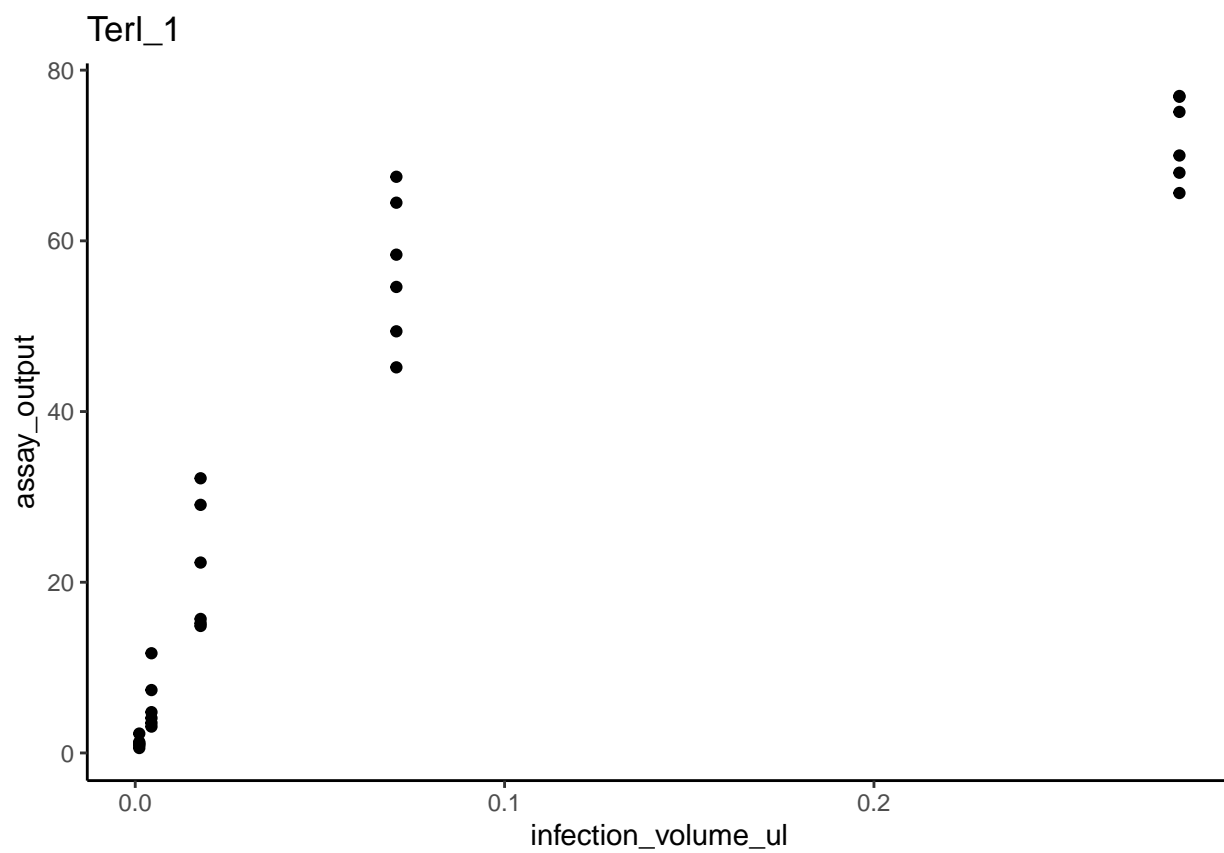


##

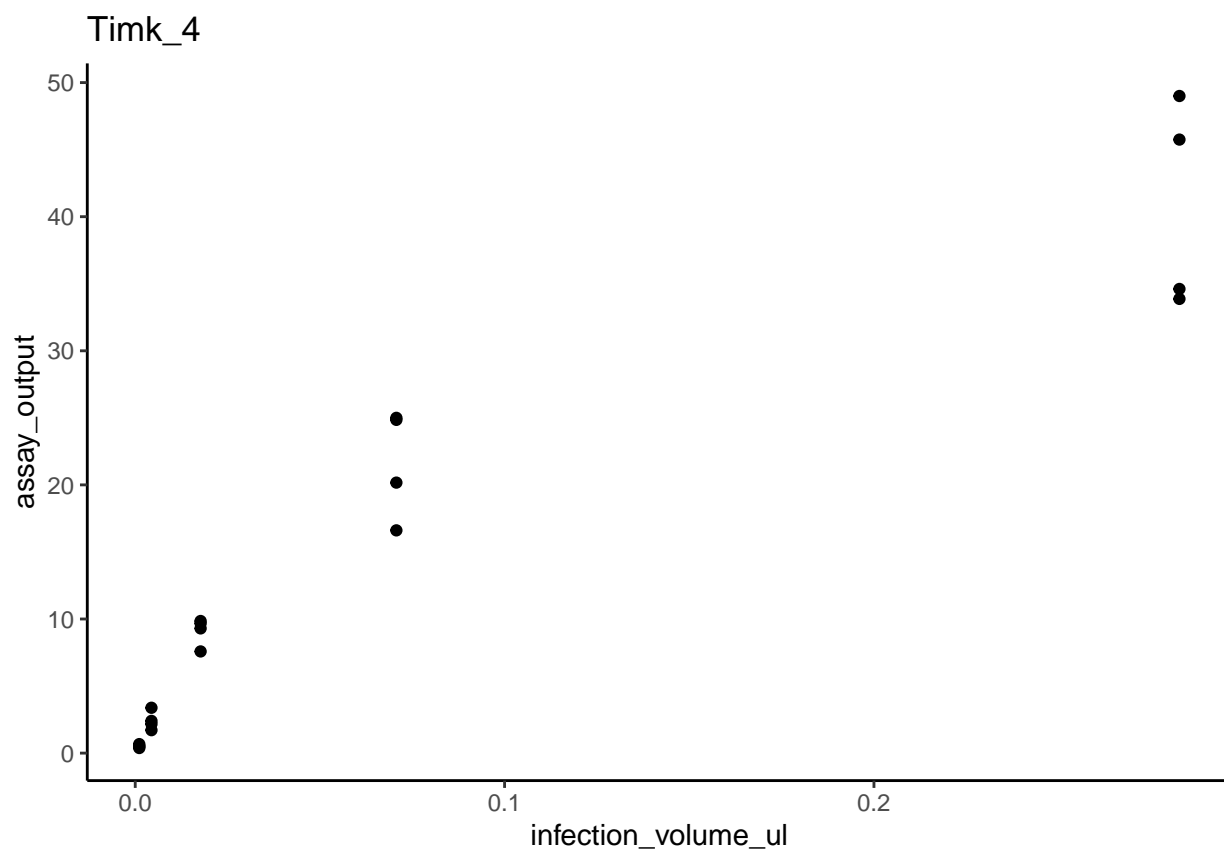
[[115]]



[[116]]

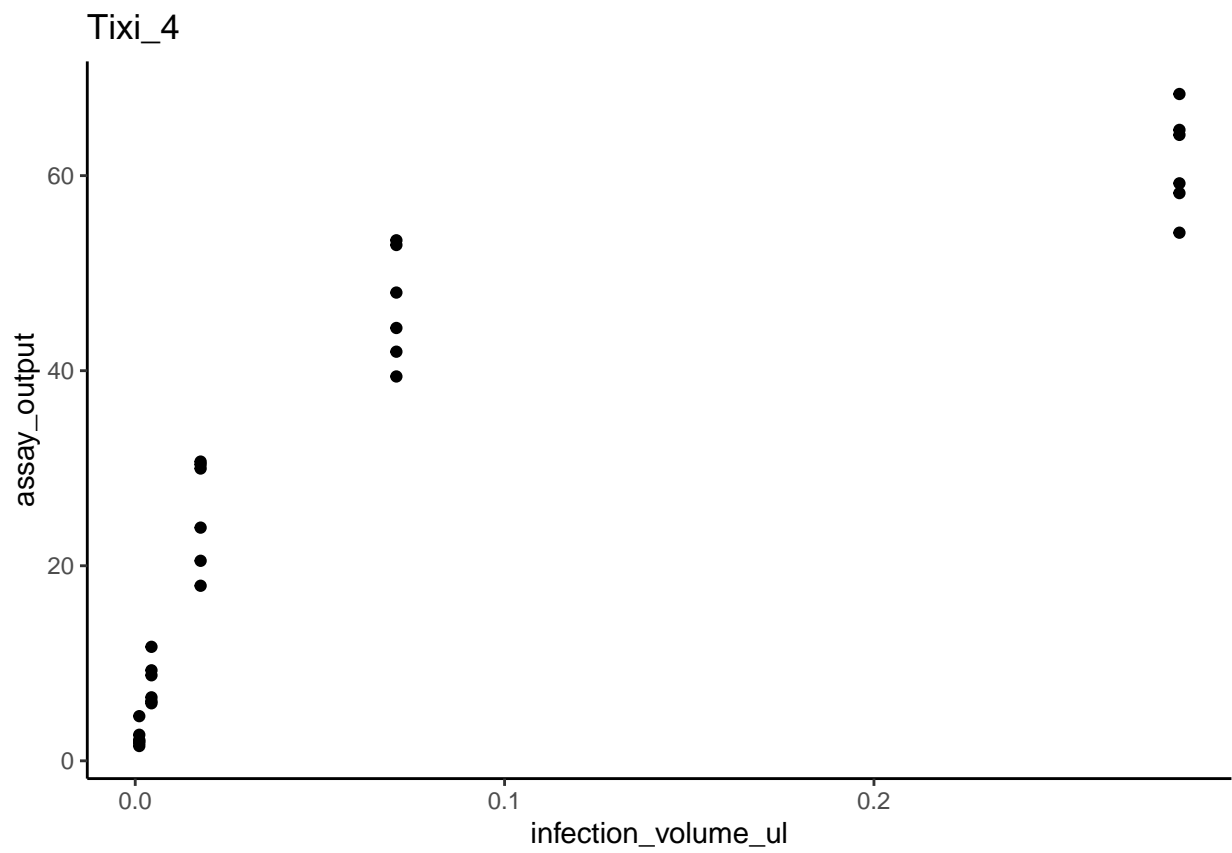


[[117]]

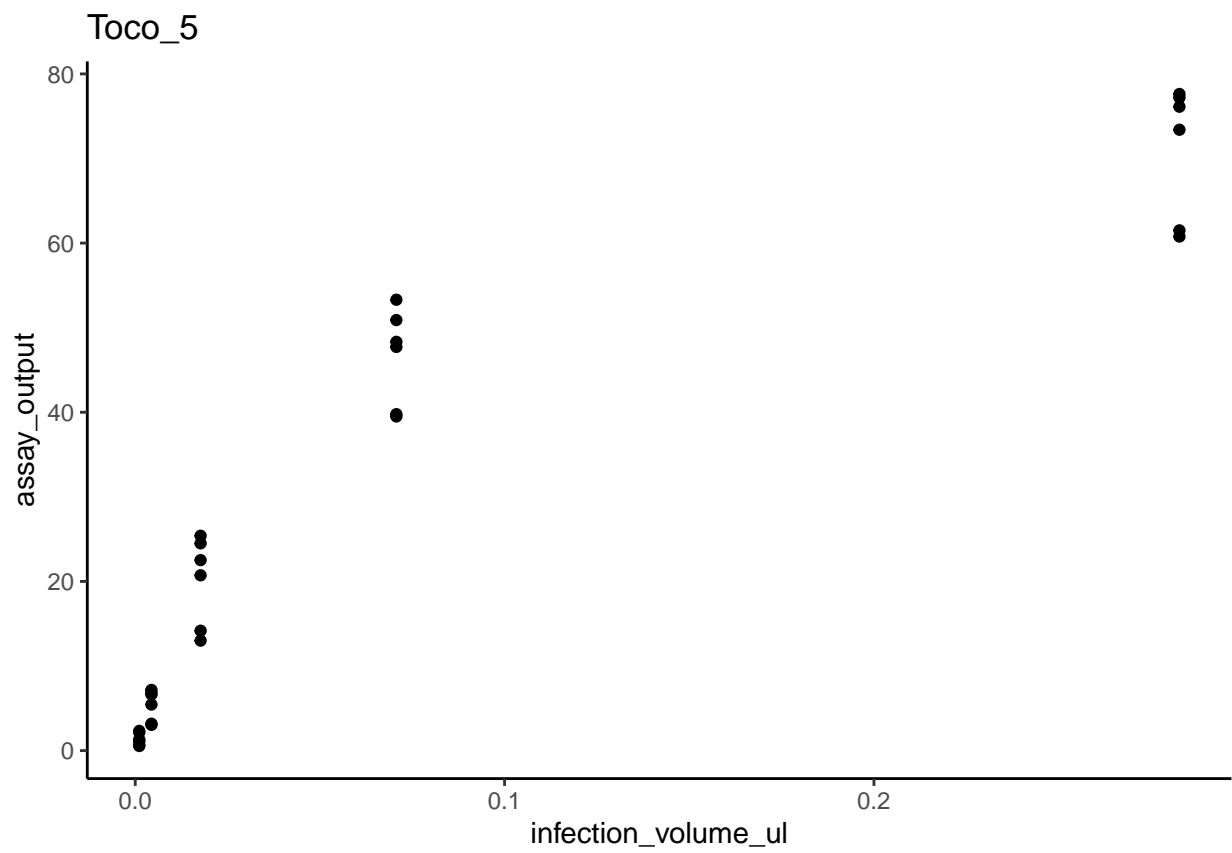


##

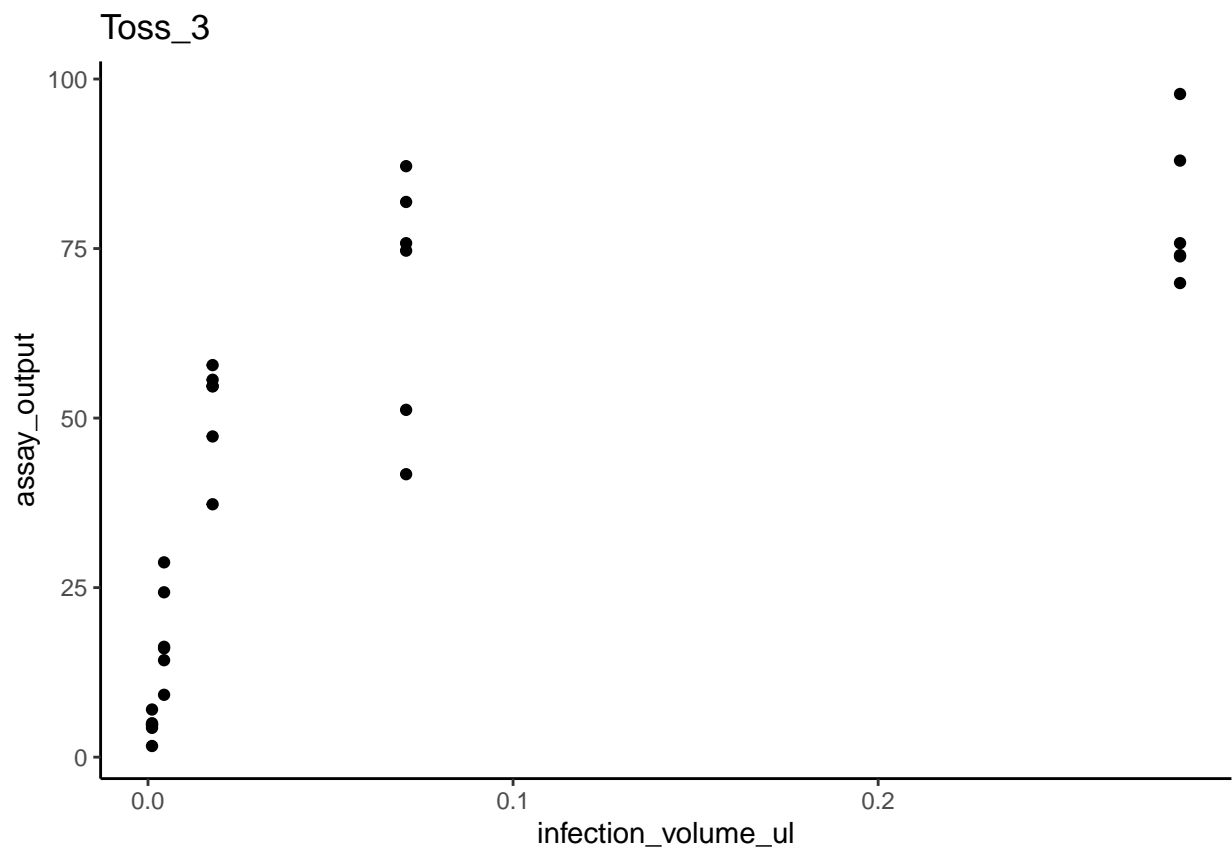
[[118]]



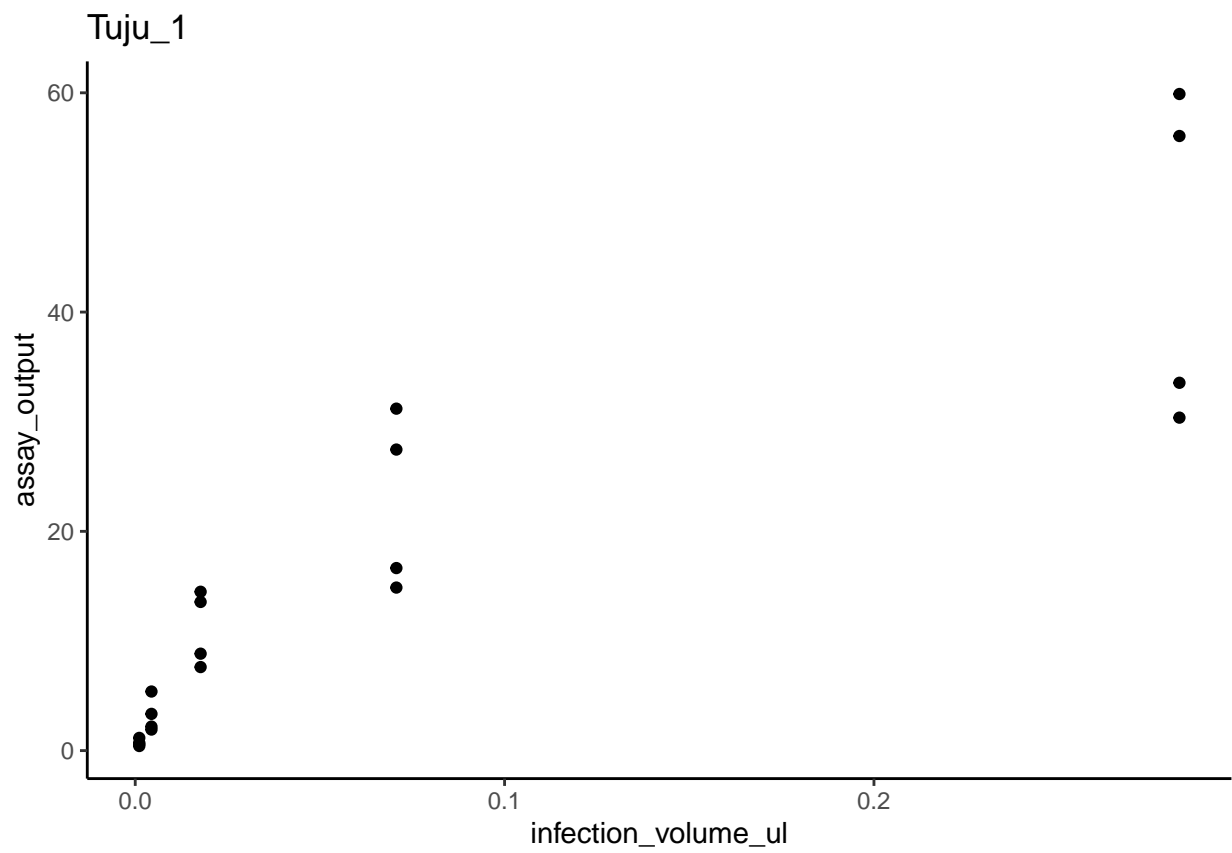
[[119]]



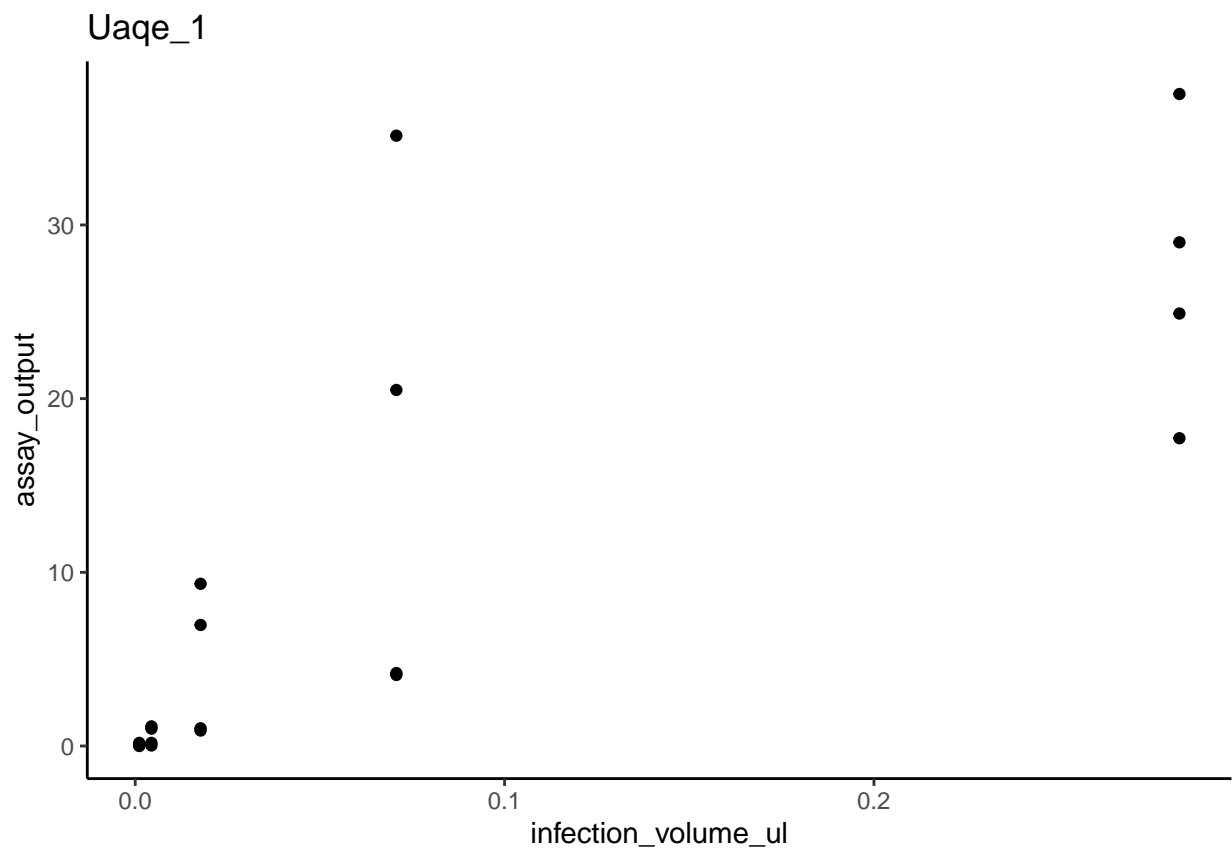
[[120]]



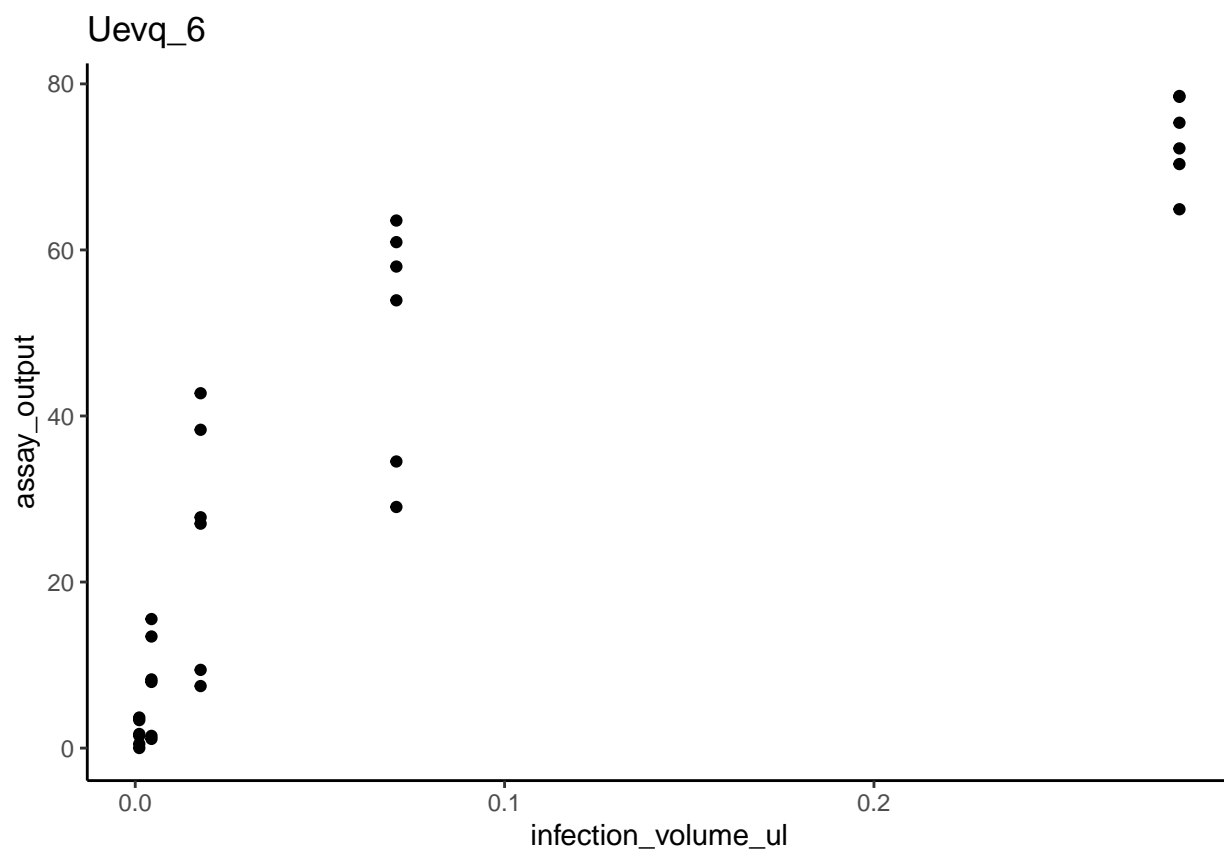
[[122]]



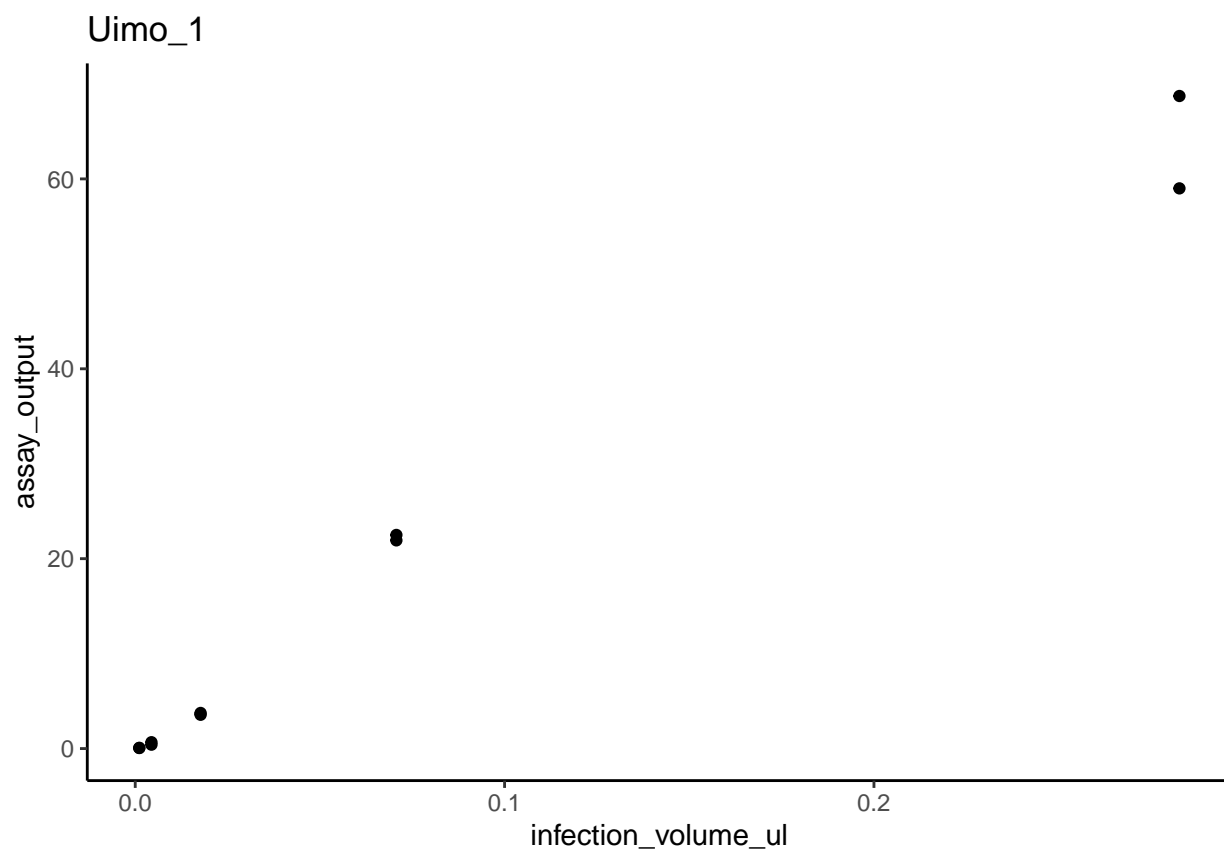
[[123]]



[[124]]

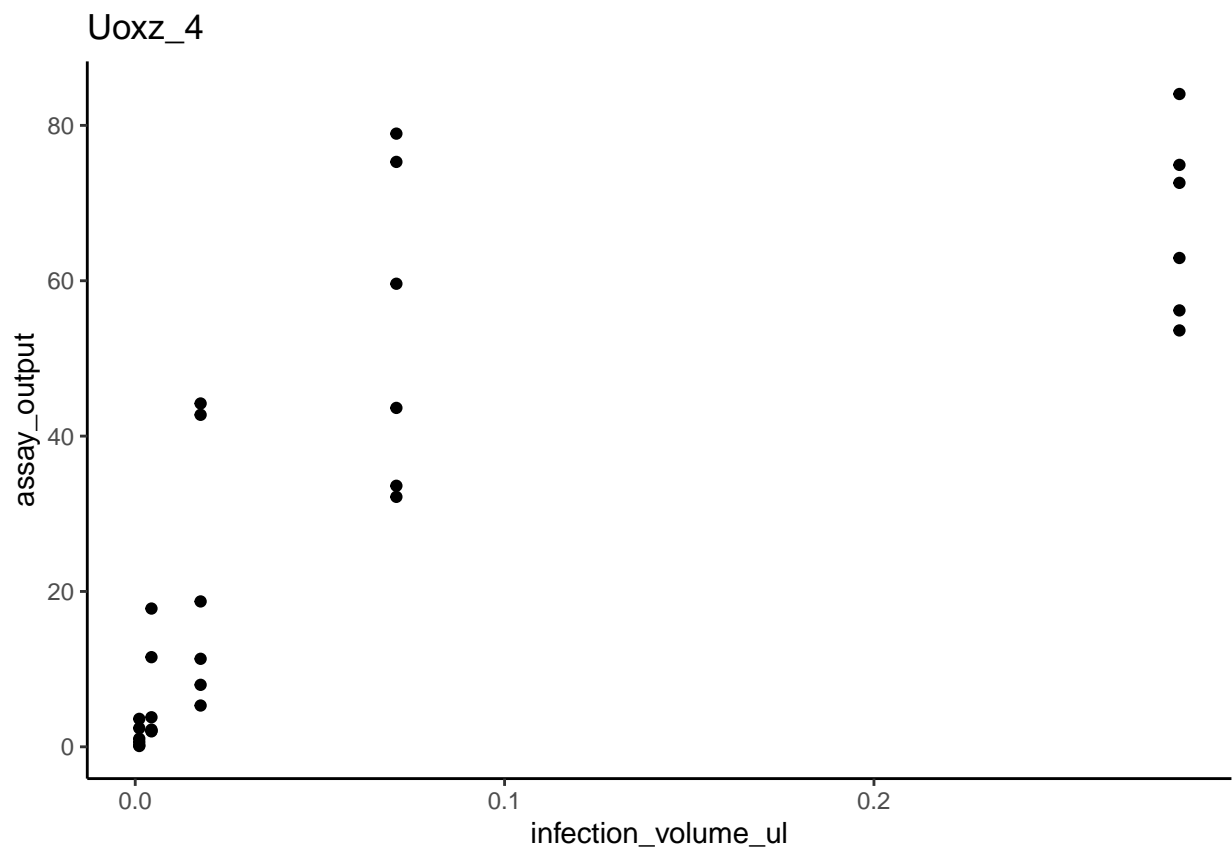


[[125]]



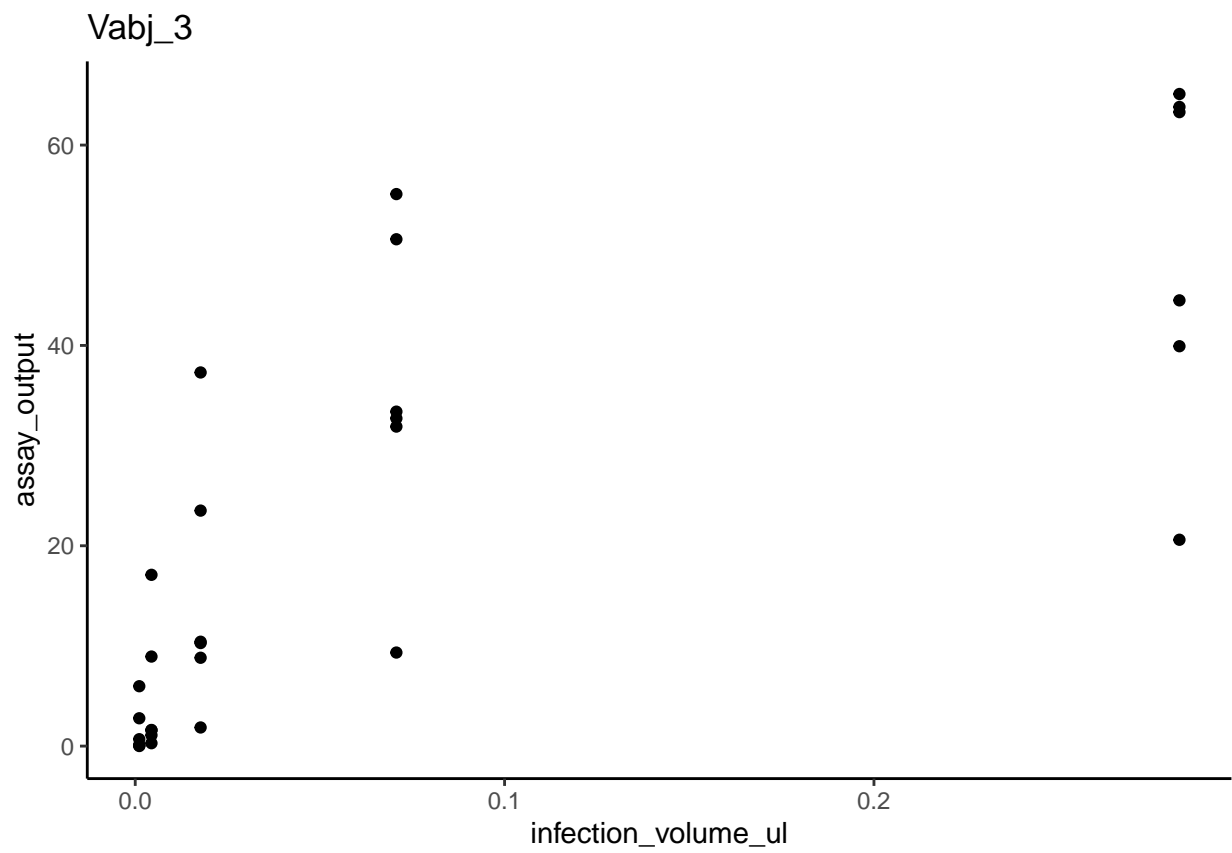
##

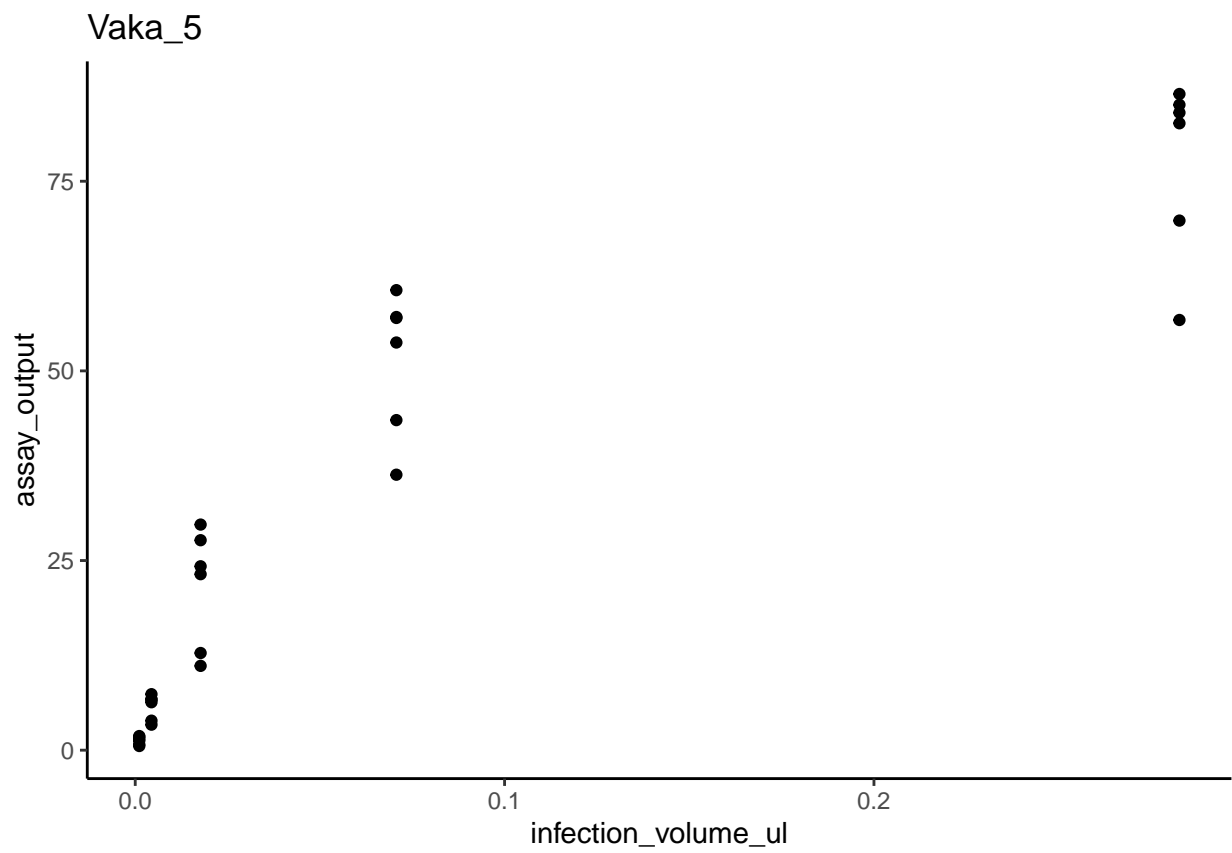
[[126]]



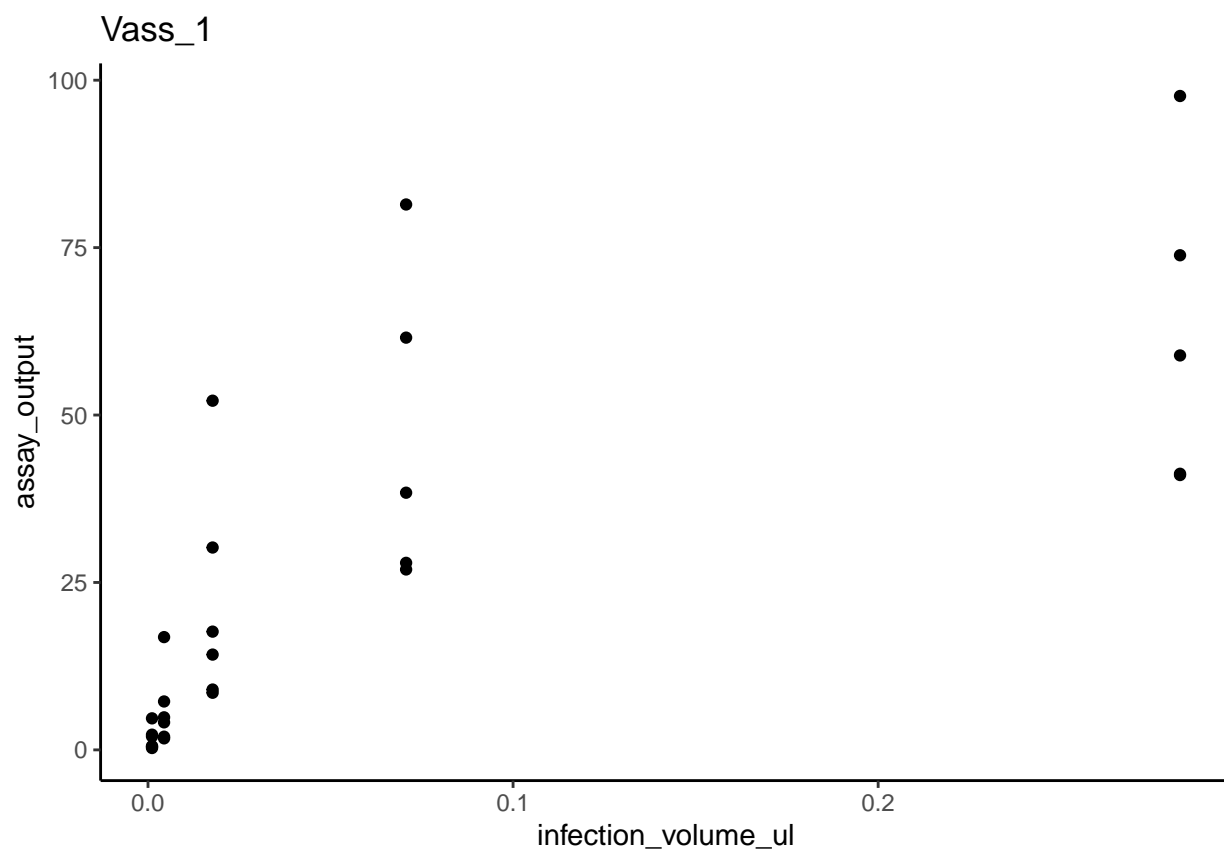
##

[[127]]

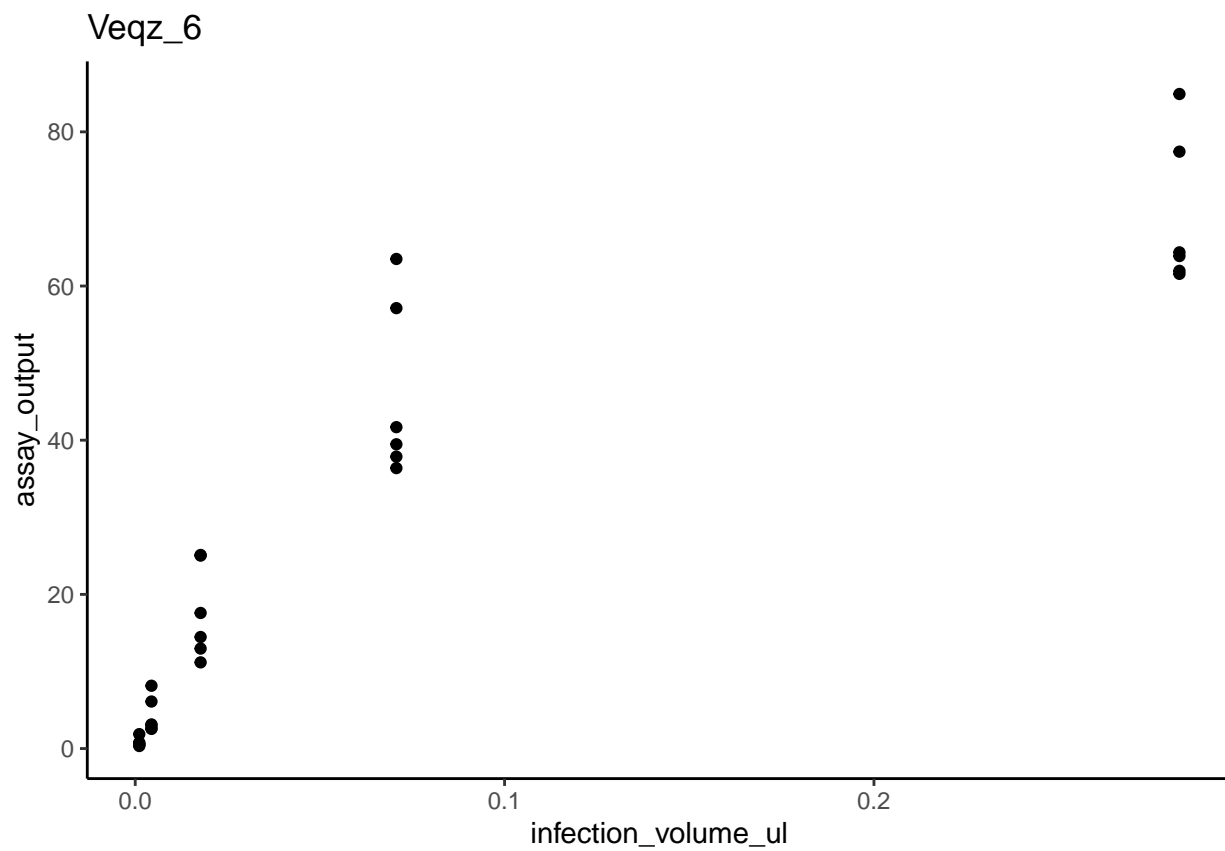




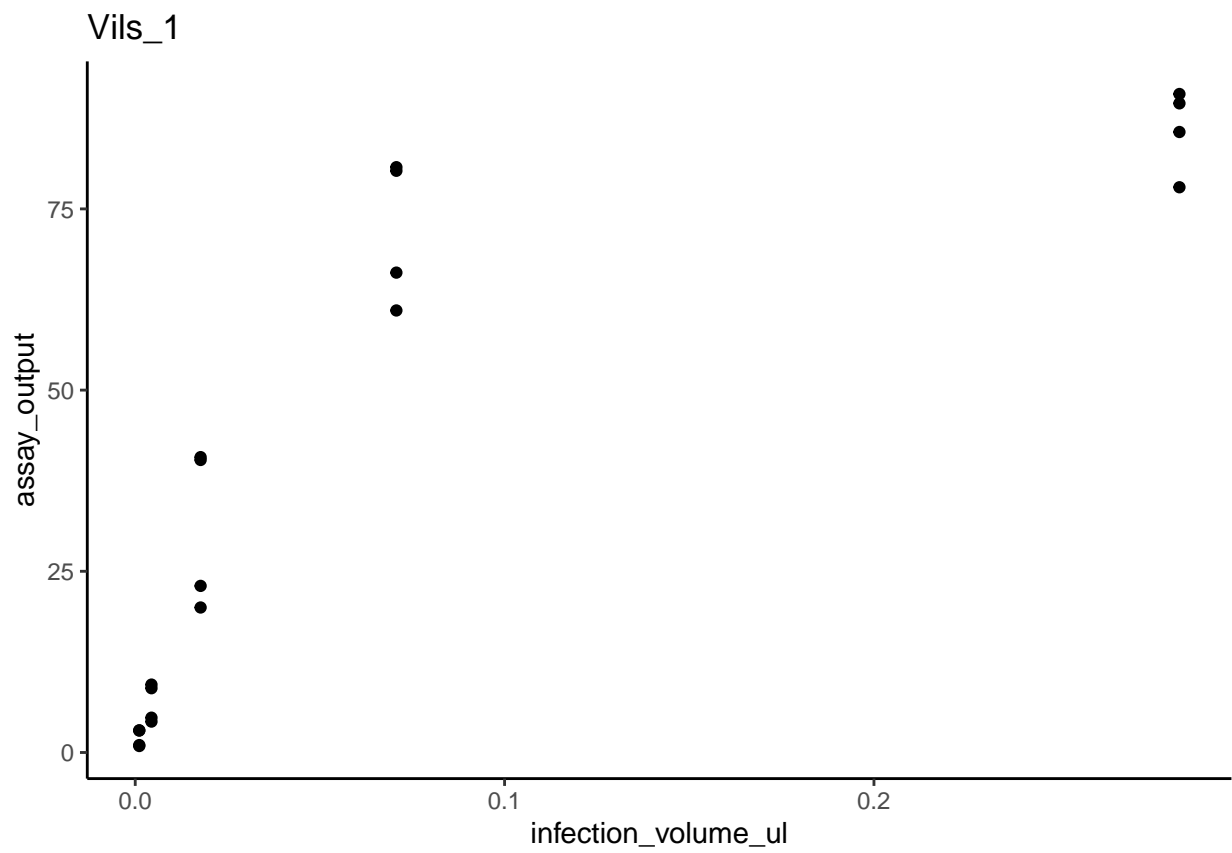
[[129]]



[[130]]

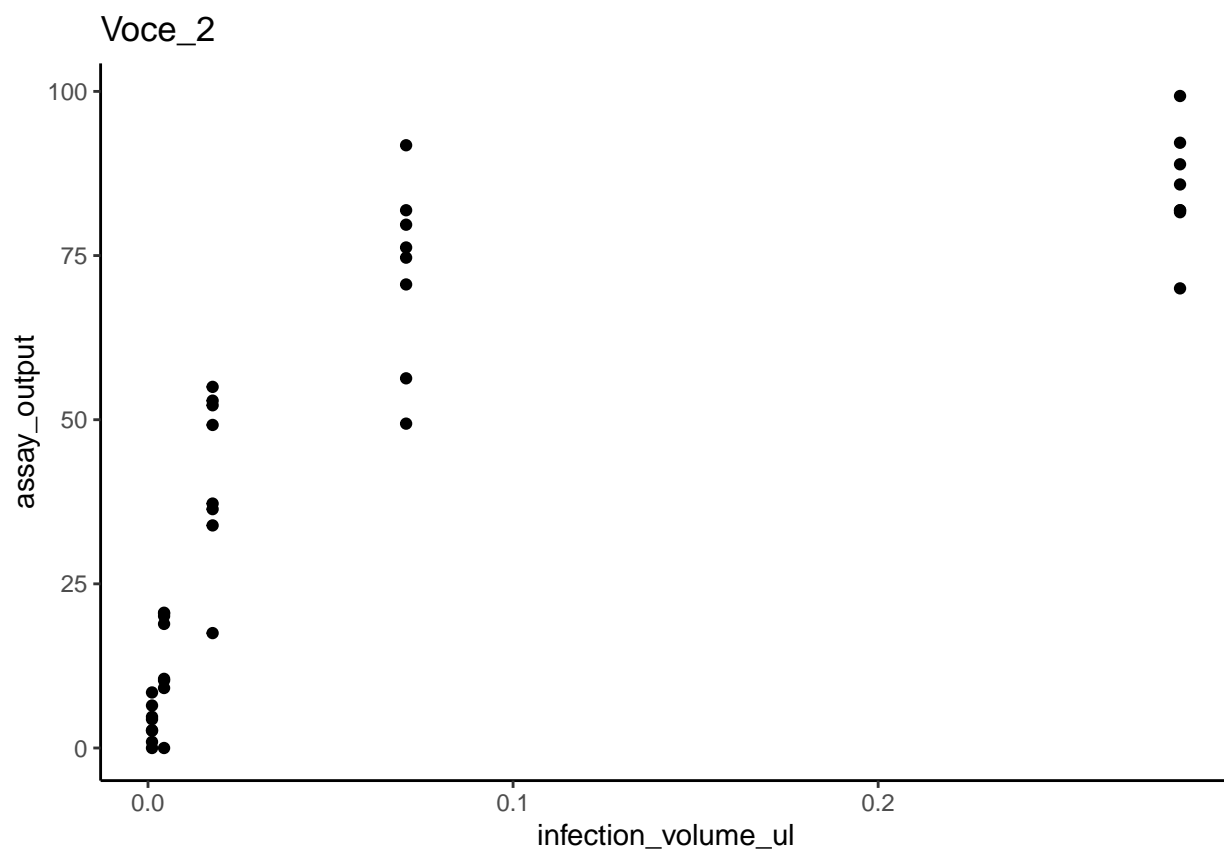


[[131]]

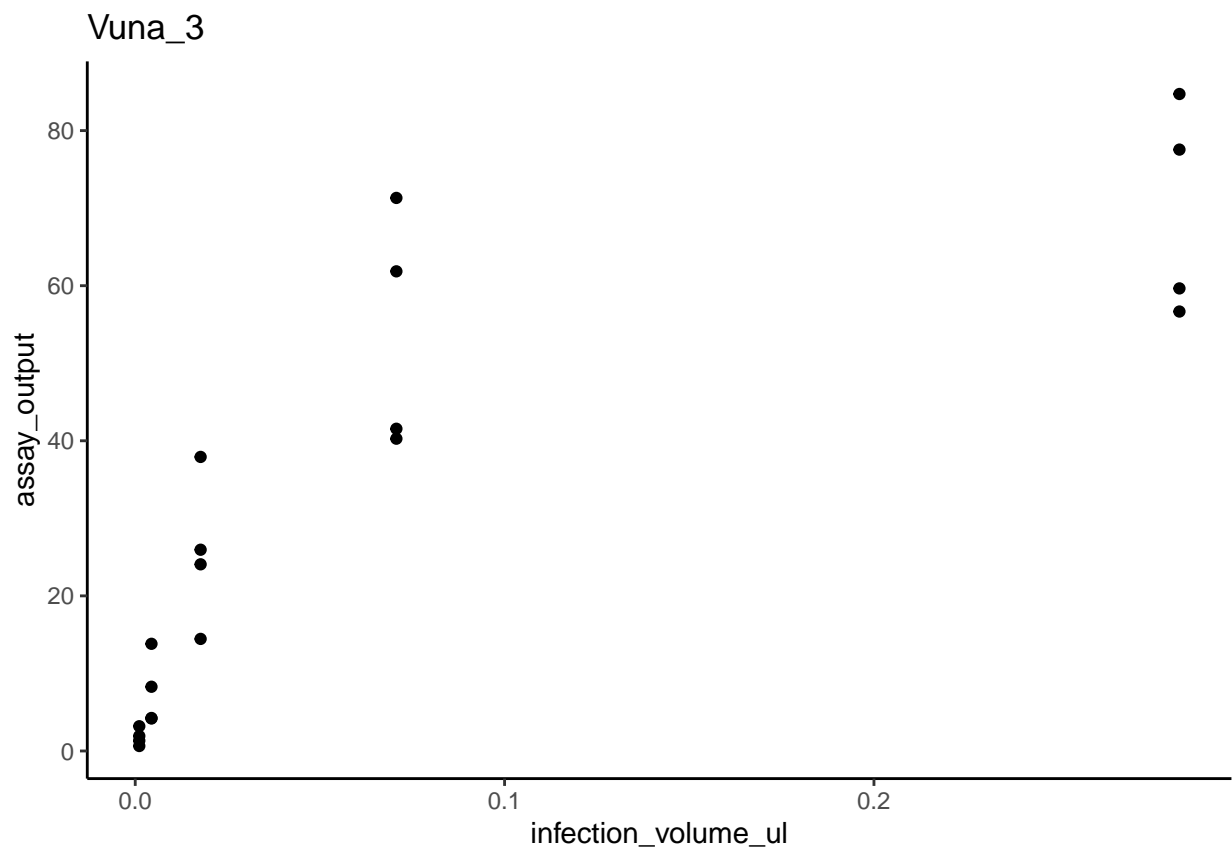


##

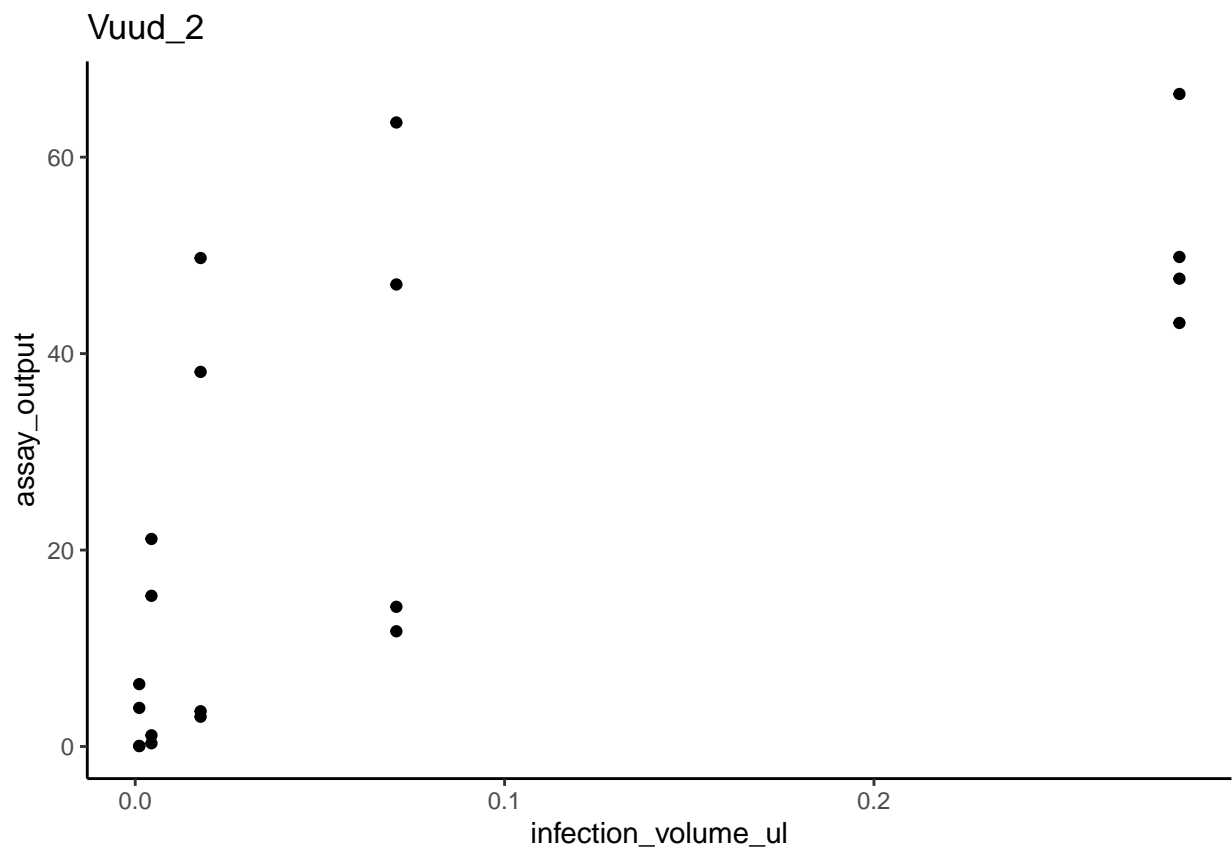
[[132]]



[[133]]

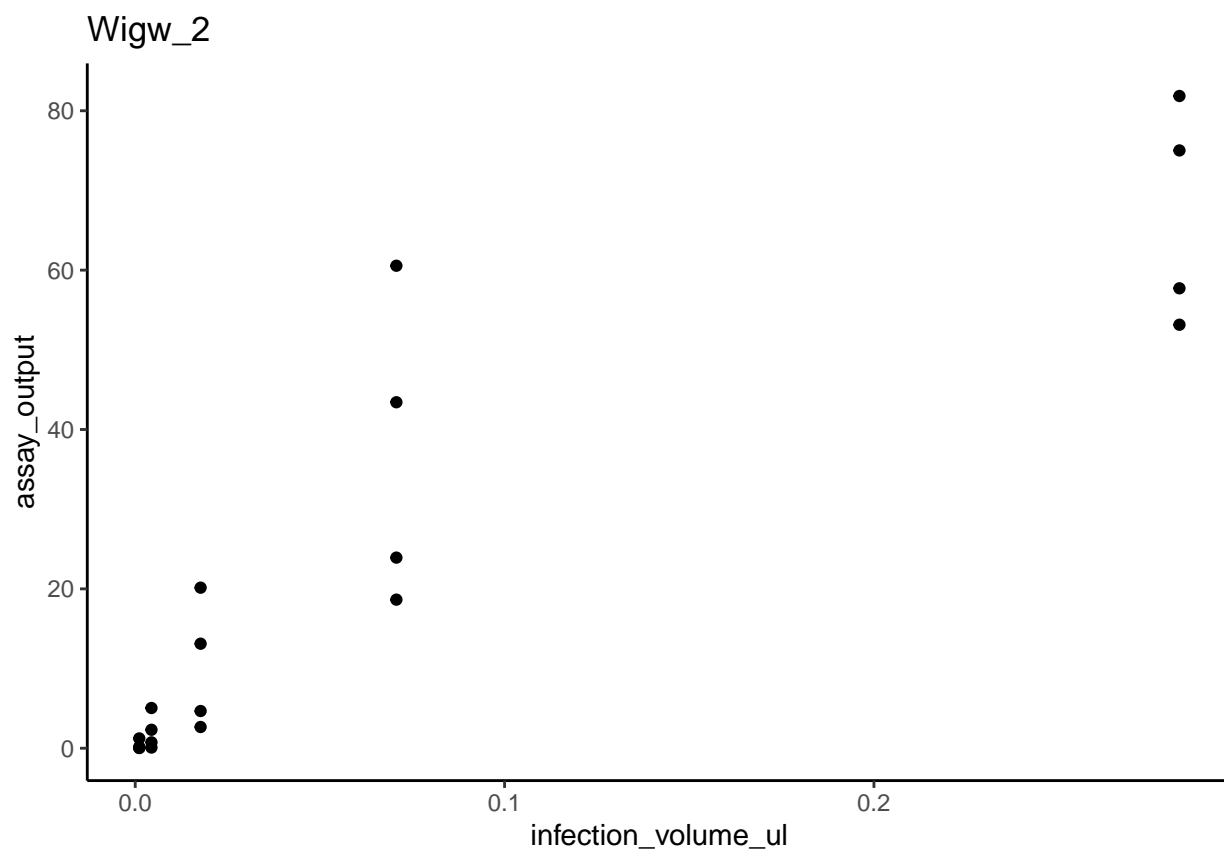


[[134]]

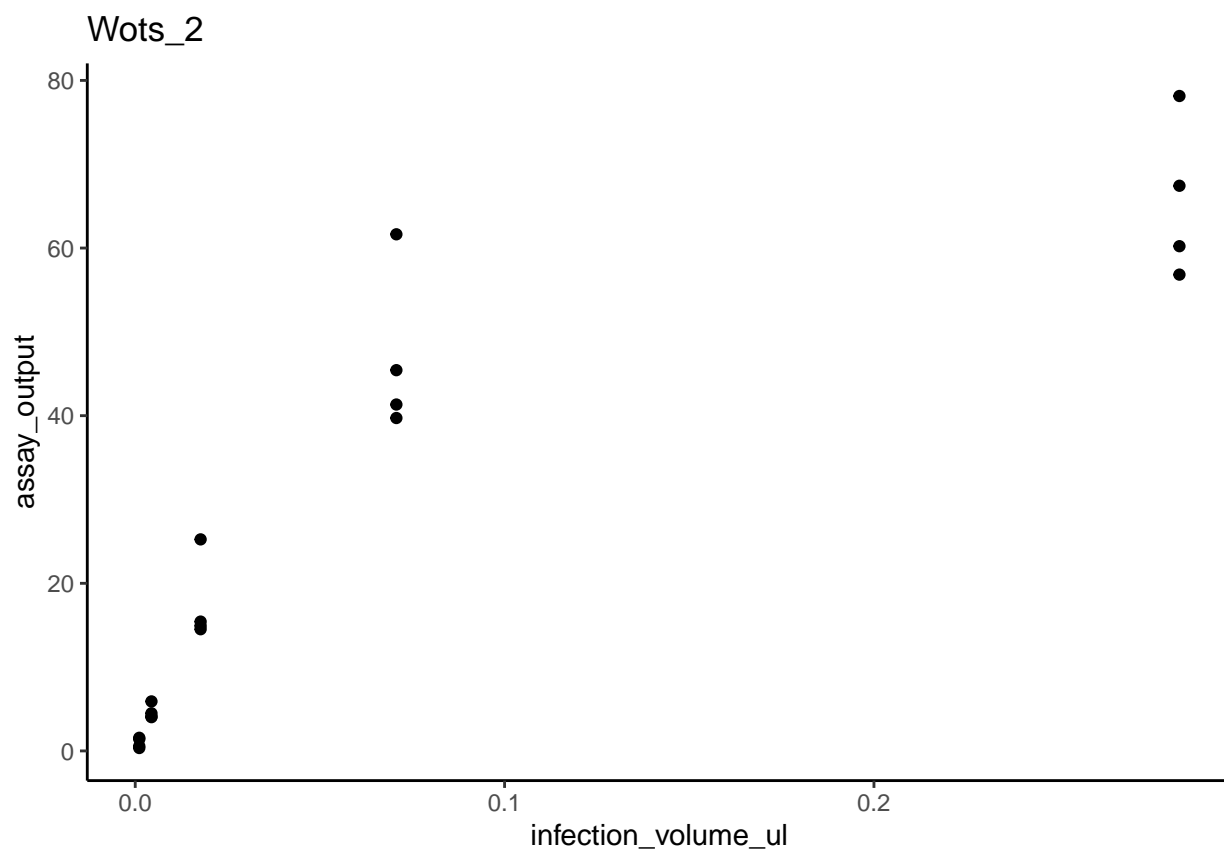


##

[[135]]

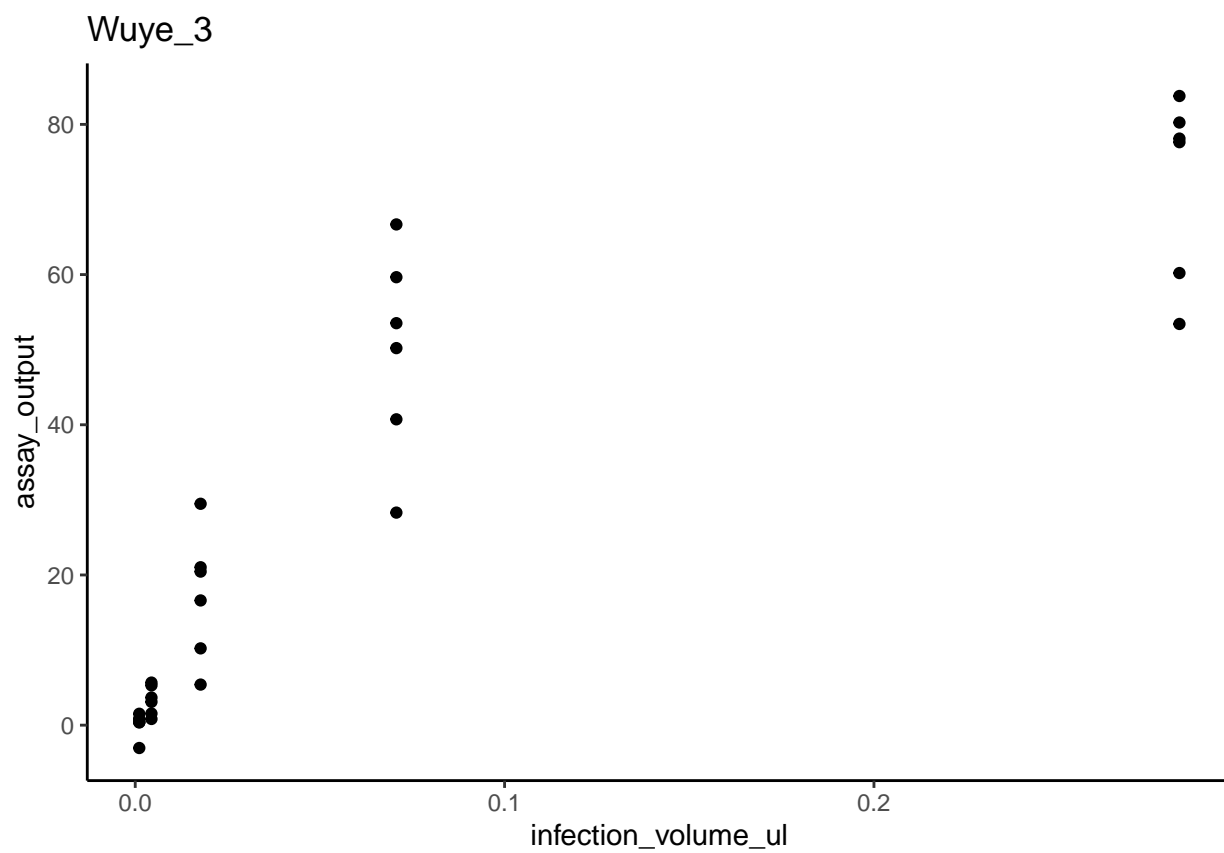


```
##  
## [[137]]
```



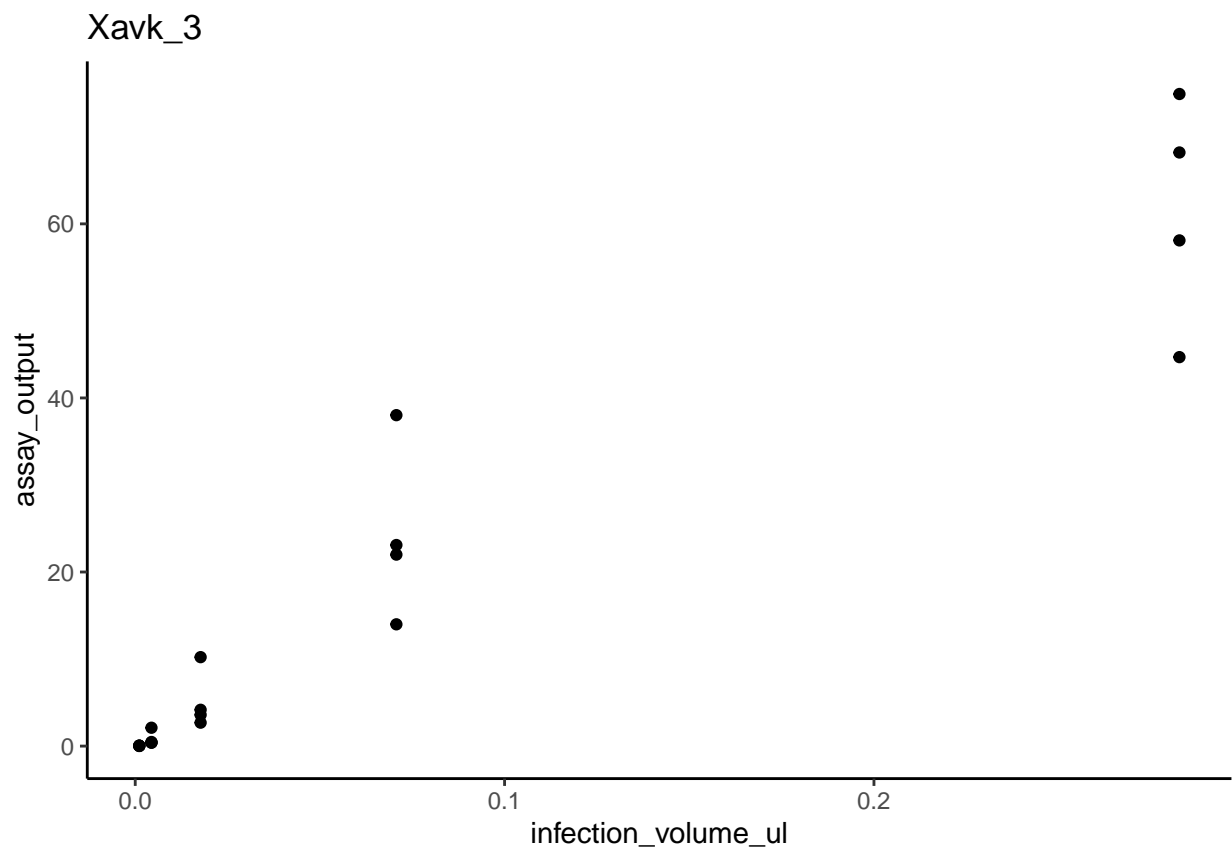
##

[[138]]

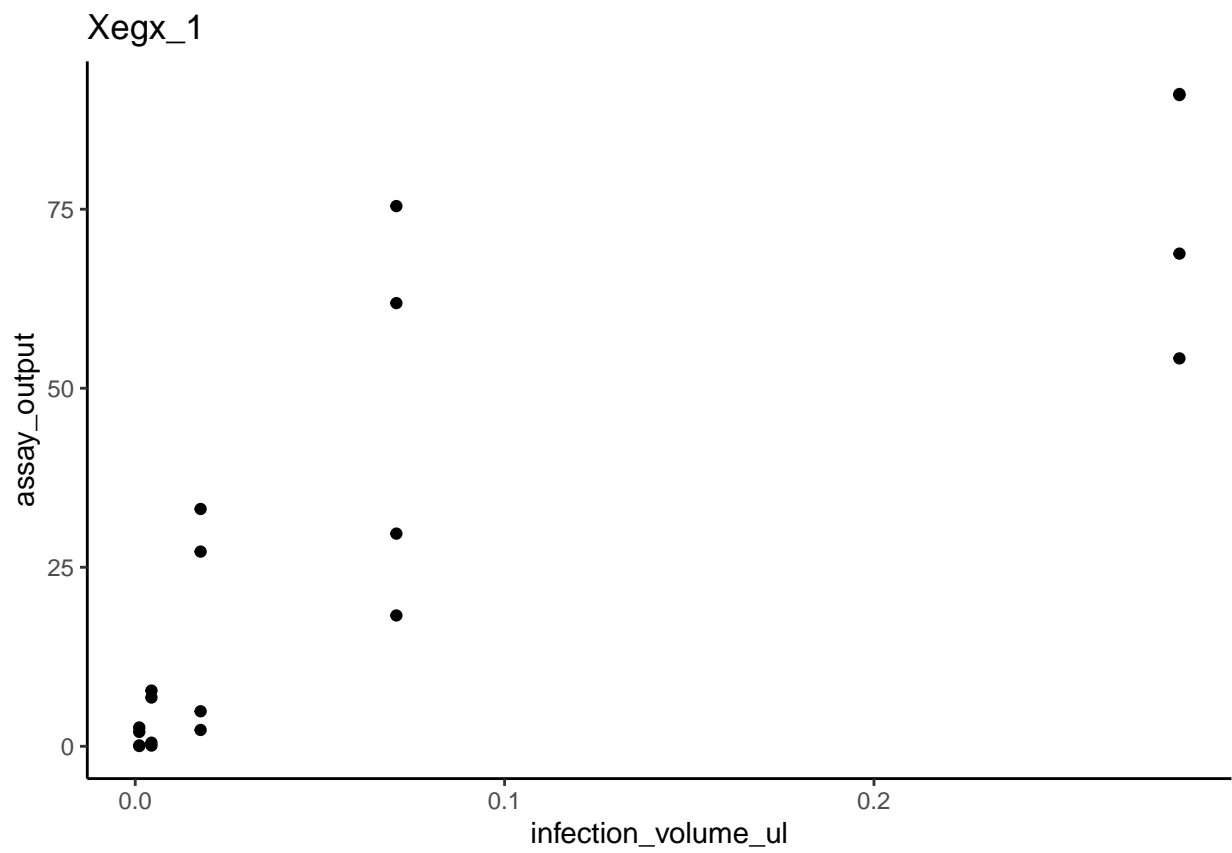


##

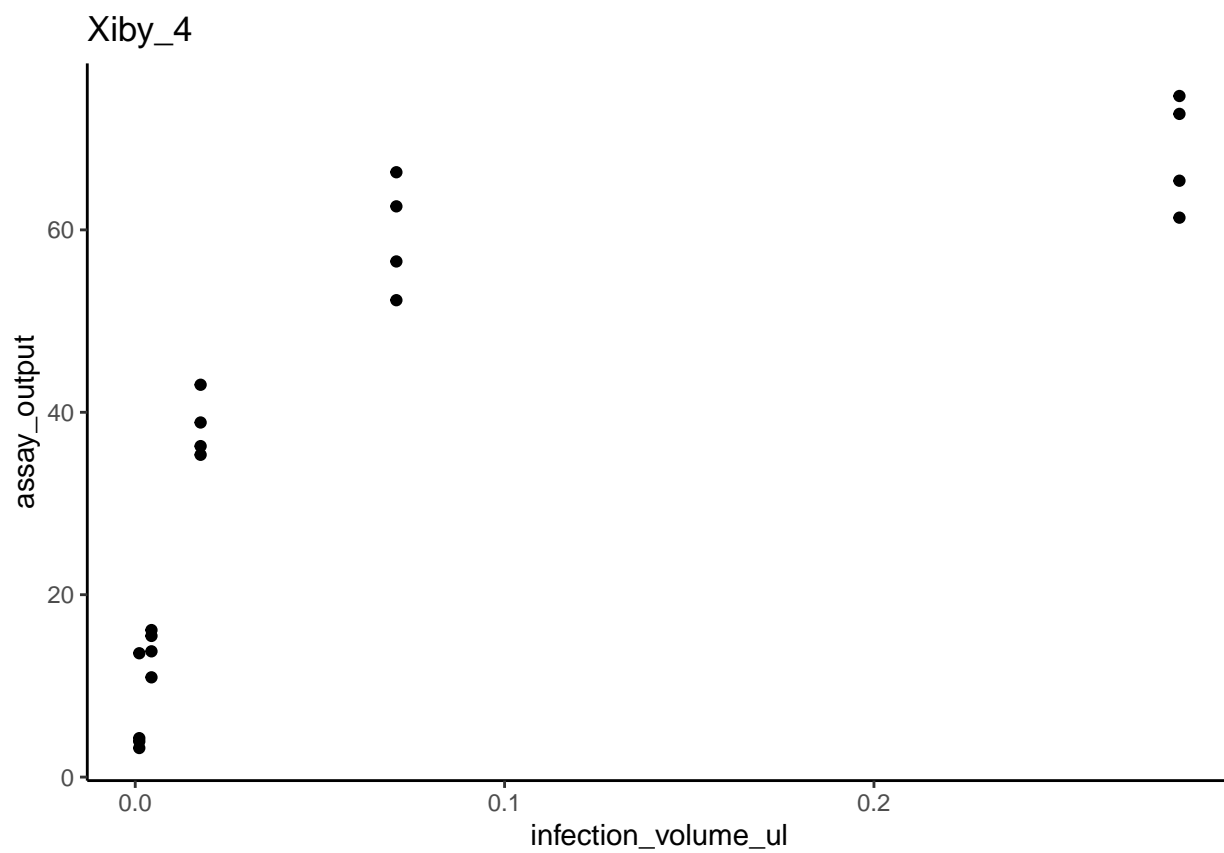
[[139]]



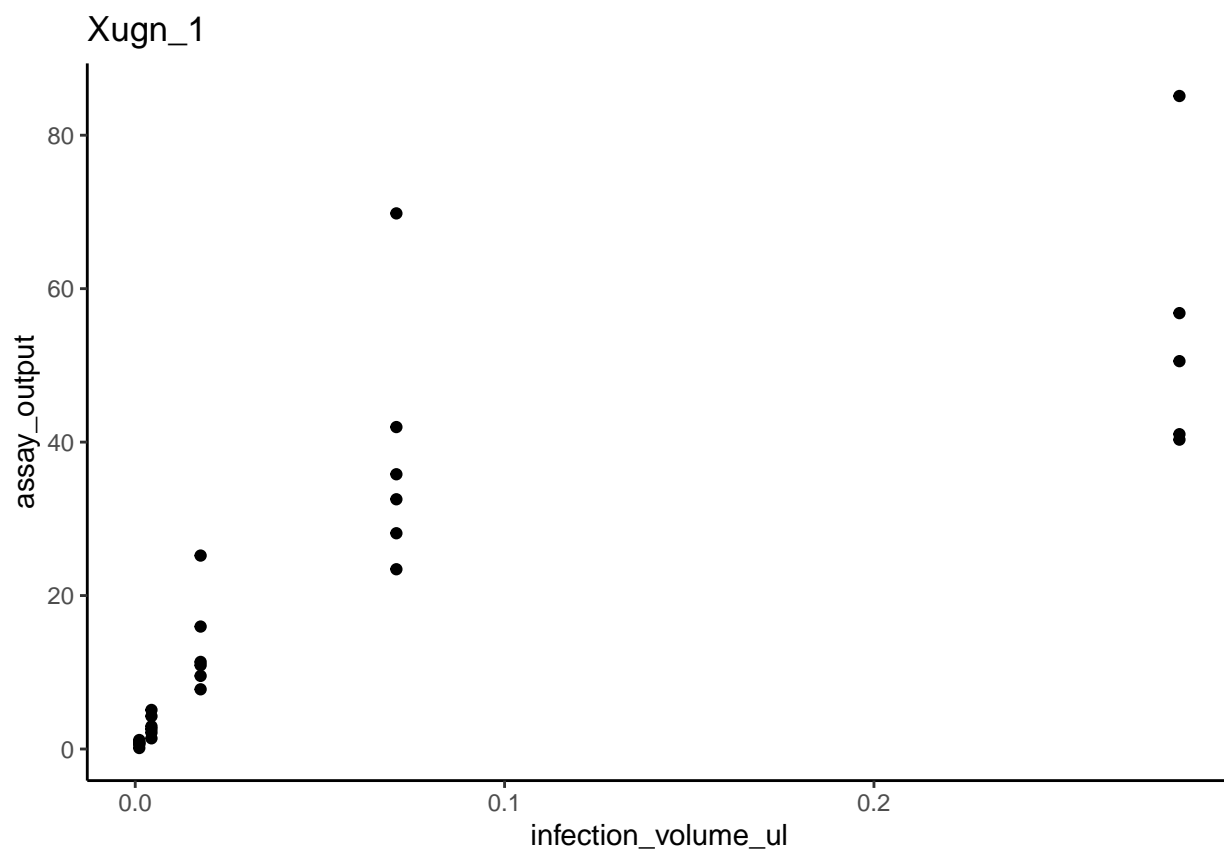
[[140]]



[[141]]

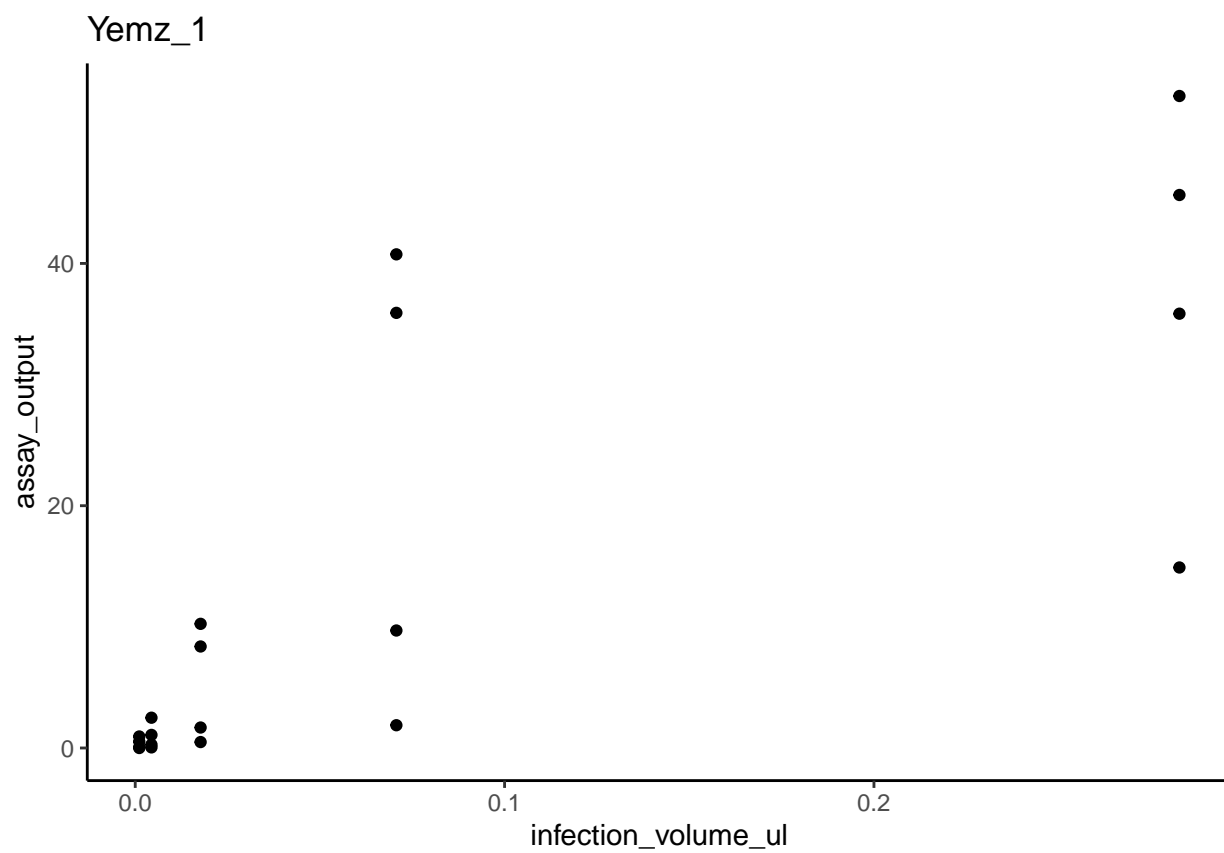


```
##  
## [[142]]
```

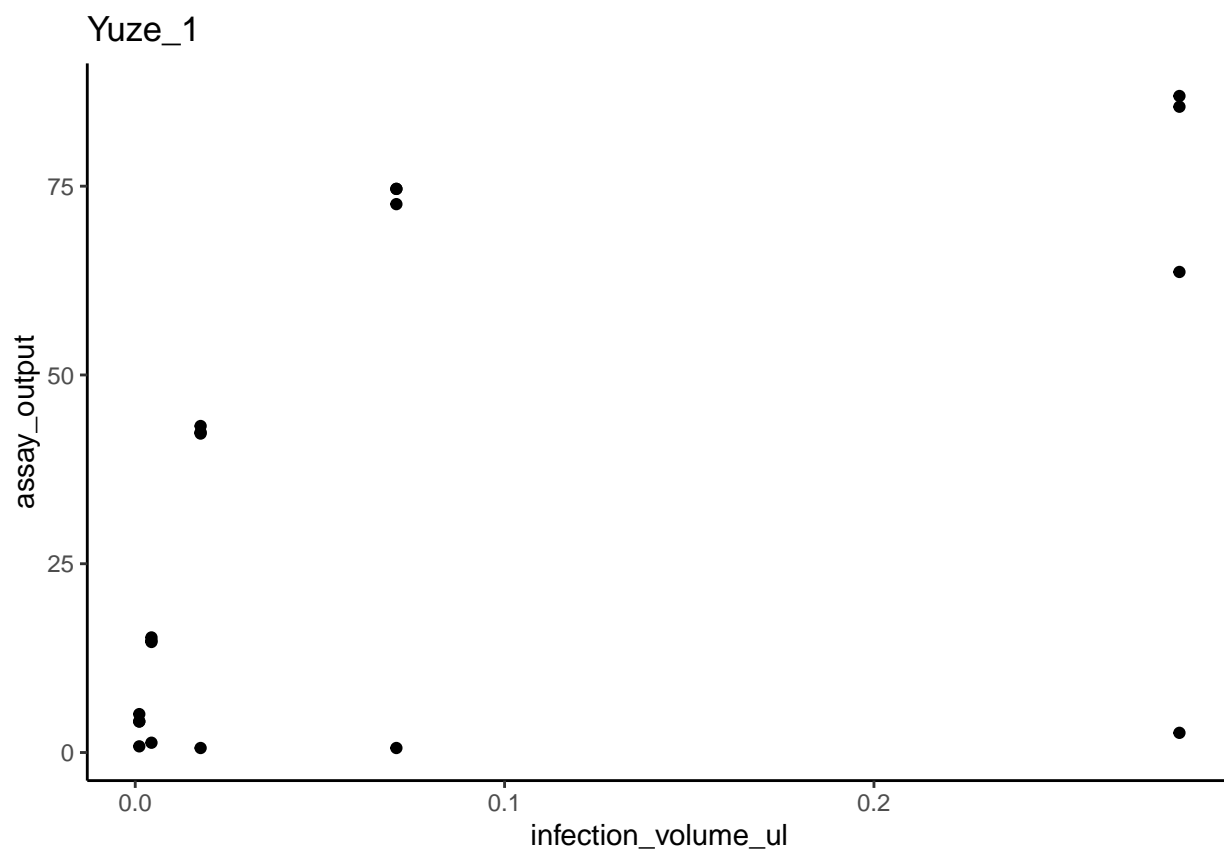
##

[[146]]



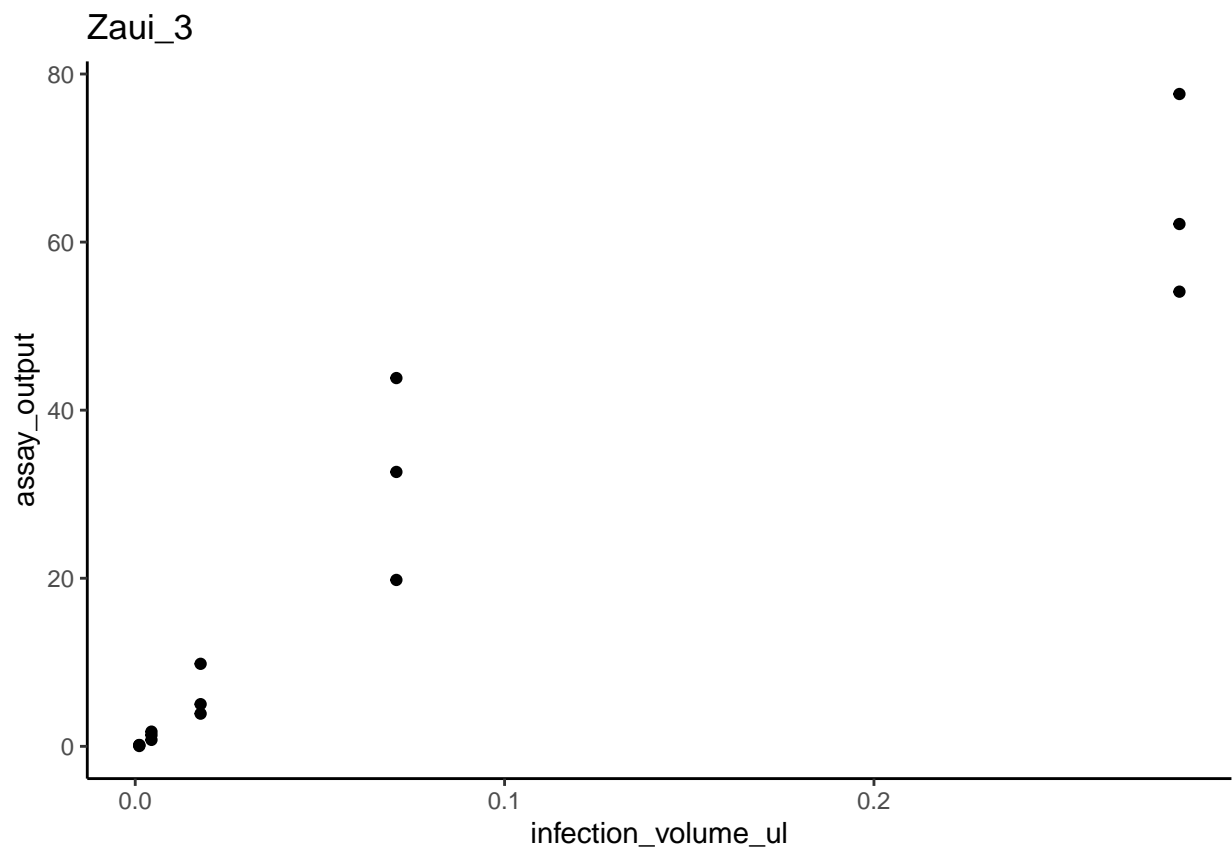
##

[[147]]

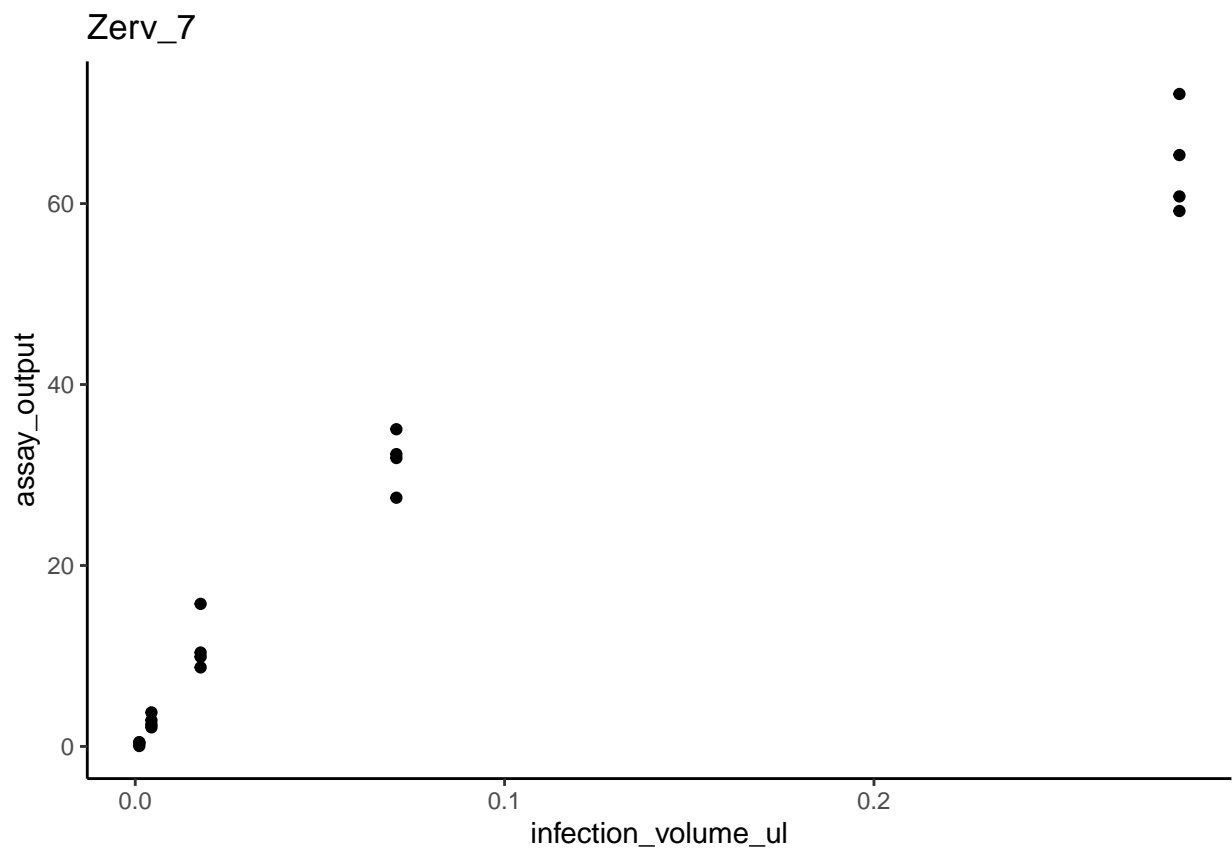


##

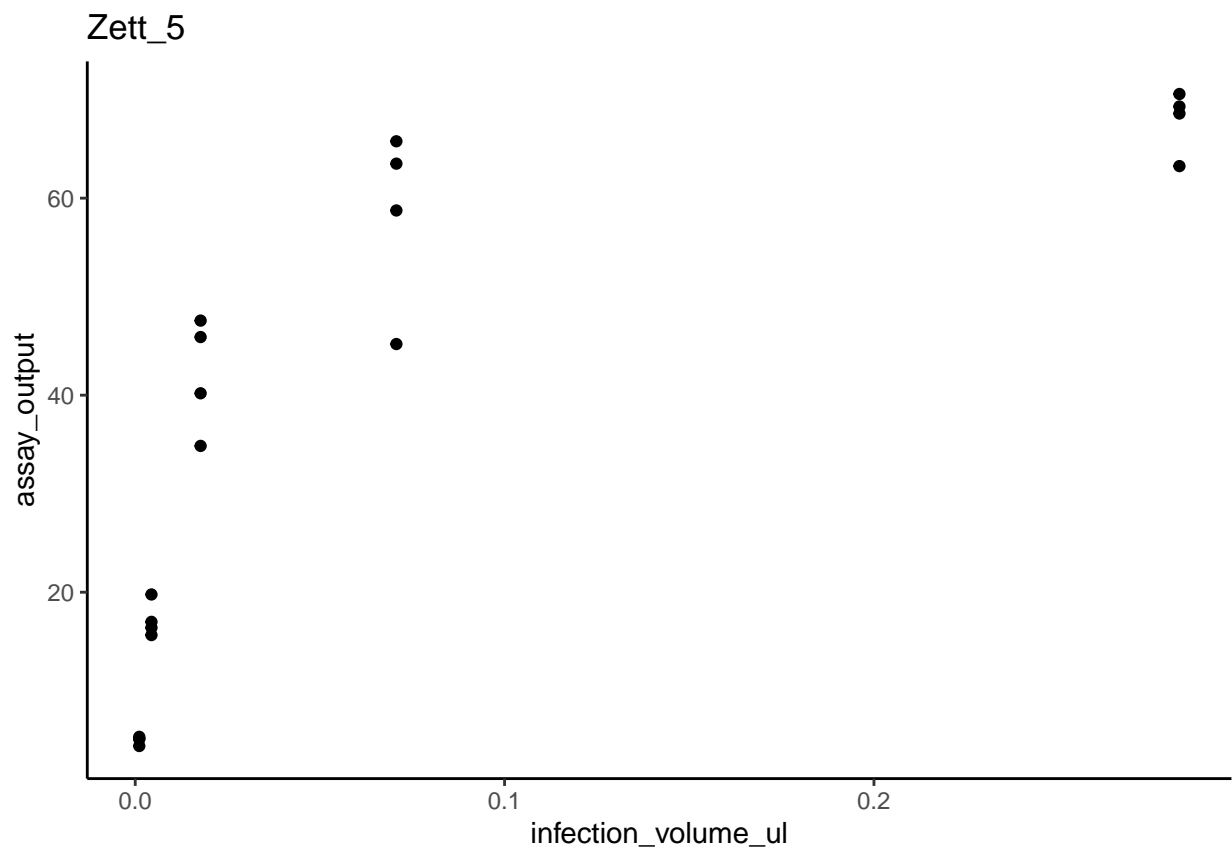
[[148]]



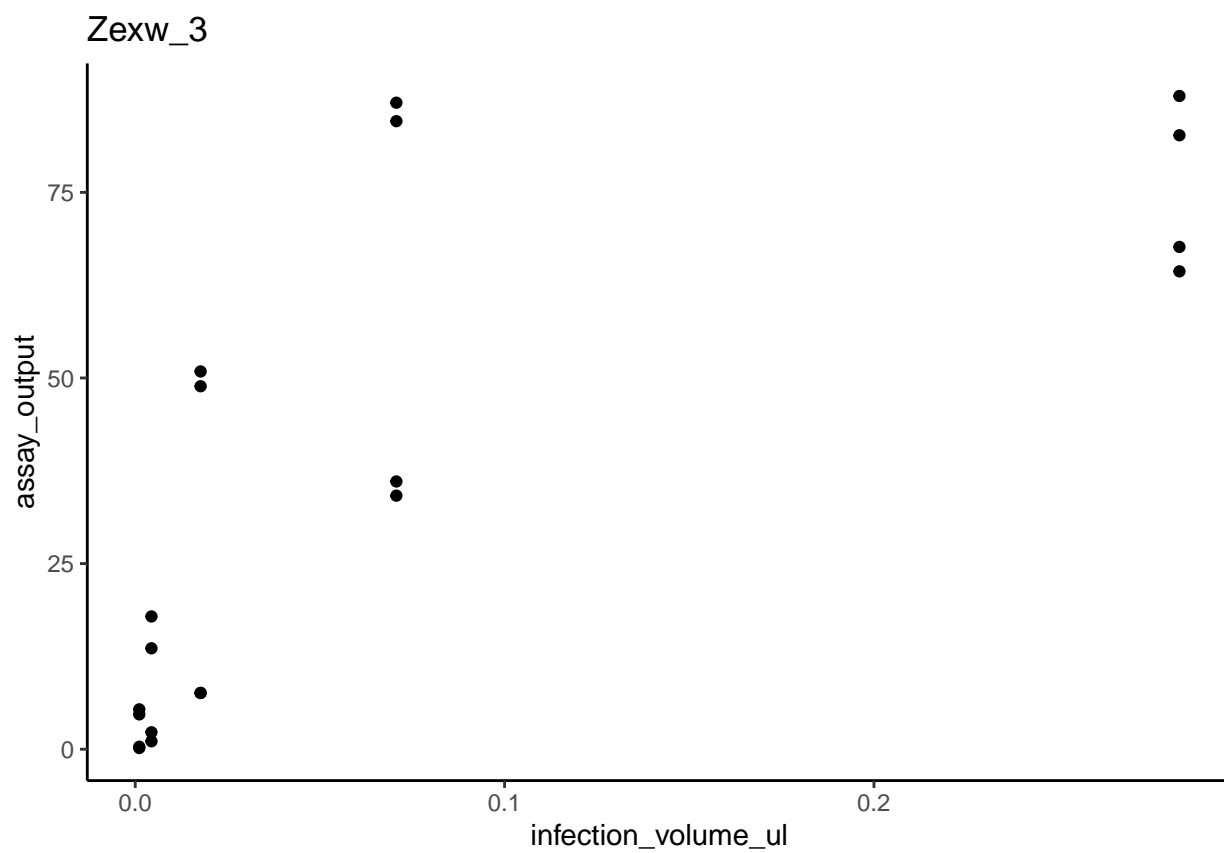
```
##  
## [[149]]
```



```
##  
## [[150]]
```

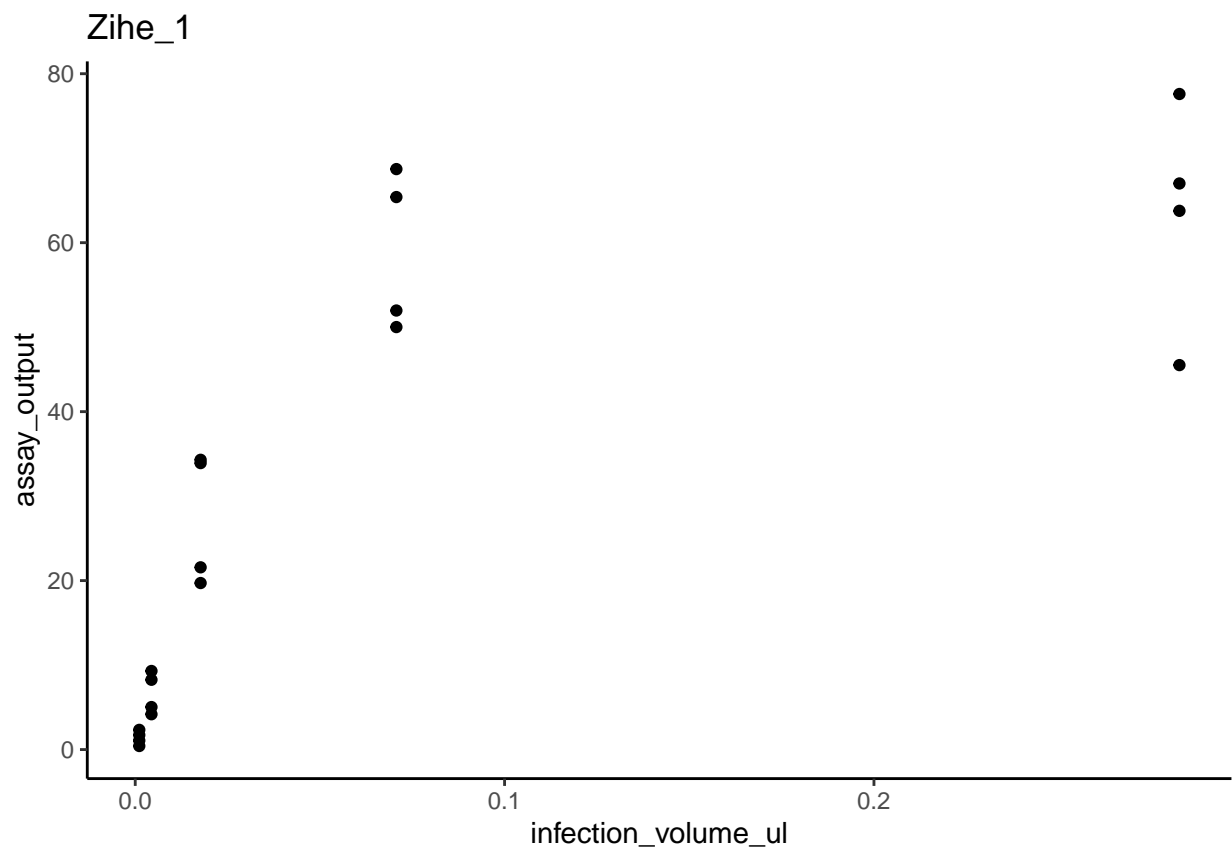


[[151]]



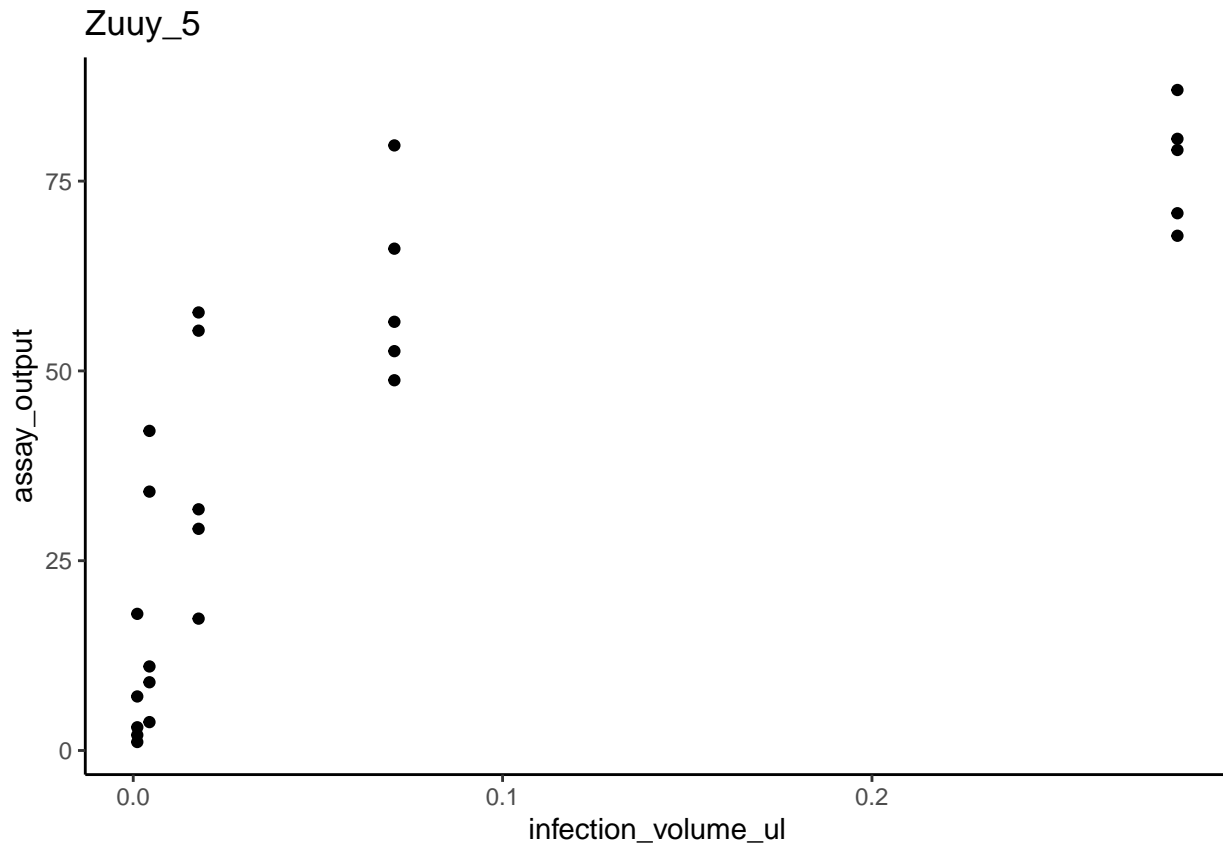
##

[[152]]



##

[[153]]



```
#names(scatter_plot_vector)=vector_data_per_cell_line$cell_line
```

making a boxplot of the assay output per cell and per titre, can use this to see distribution between screens and replicates and see the initial outliers, stat summary lets you see the upper quartile, median, lower quartile and max and min values

```
apply_plot_boxplot=function(i){ggplot(data = i,aes(y=assay_output,x=titre,))+
  geom_boxplot()+
  ggtitle(i$cell_line[1])+
  stat_summary(geom="text", fun.y=quantile,
               aes(label=sprintf("%1.1f", ..y..), color=factor(titre)),
               position=position_nudge(x=0.6), size=3.5) +
  theme_classic()
}
```

code to apply function

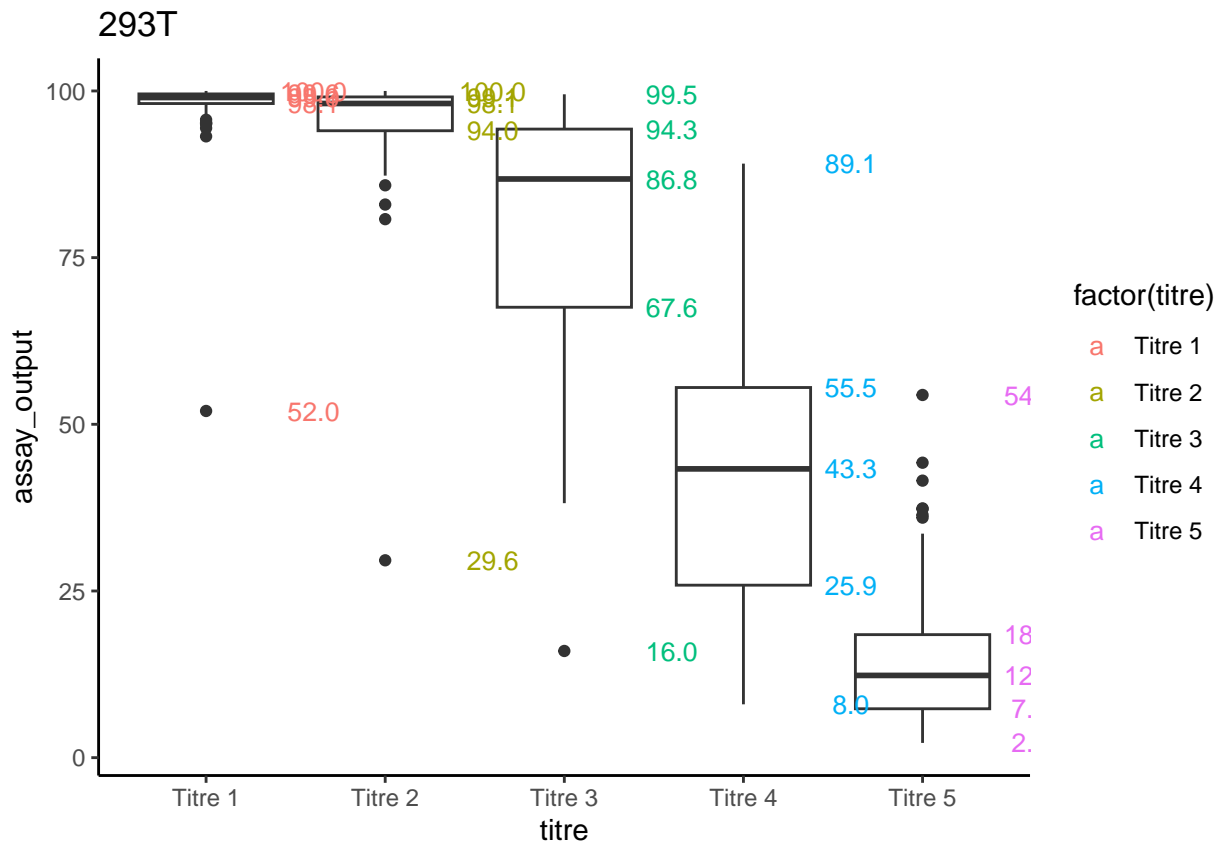
```
purrr::map(vector_data_per_cell_line$data,apply_plot_boxplot)
```

```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

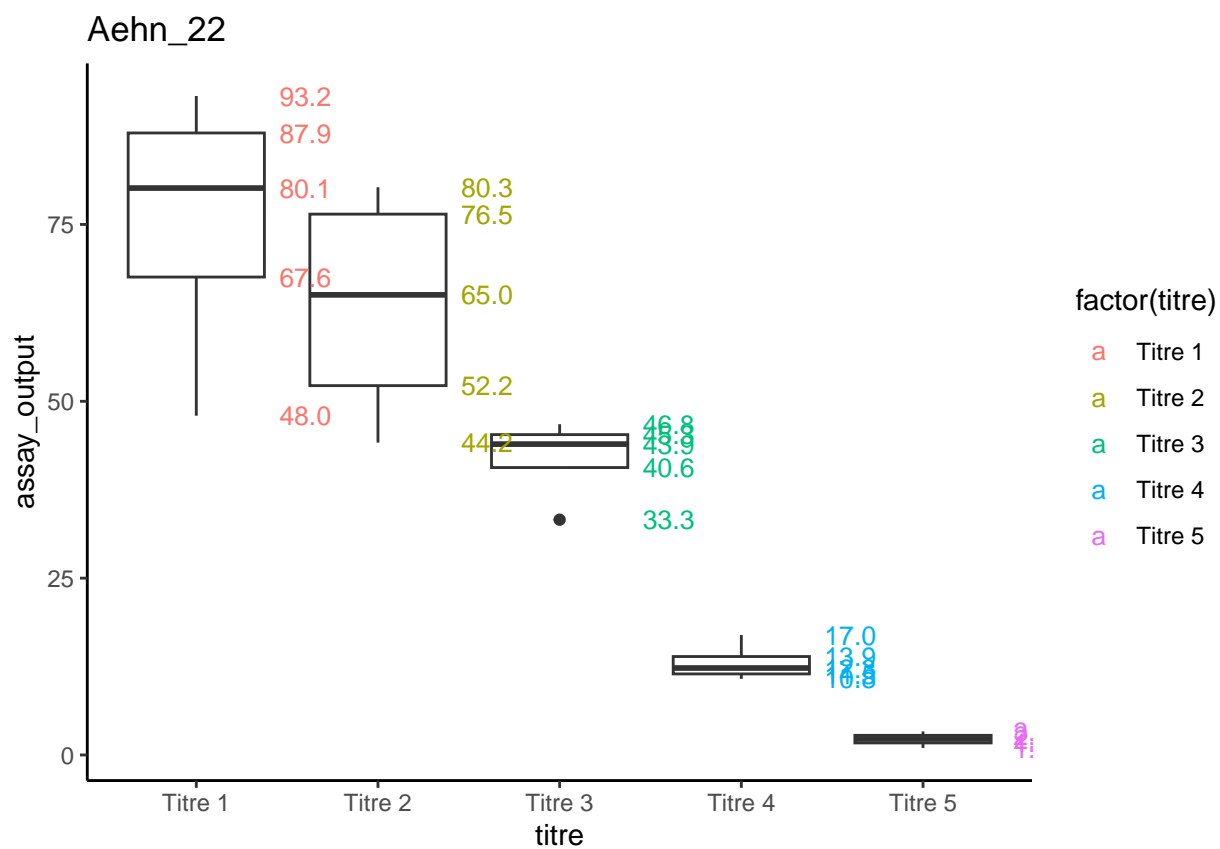
## [[1]]

## Warning: The dot-dot notation (`..y..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(y)` instead.
## This warning is displayed once every 8 hours.
```

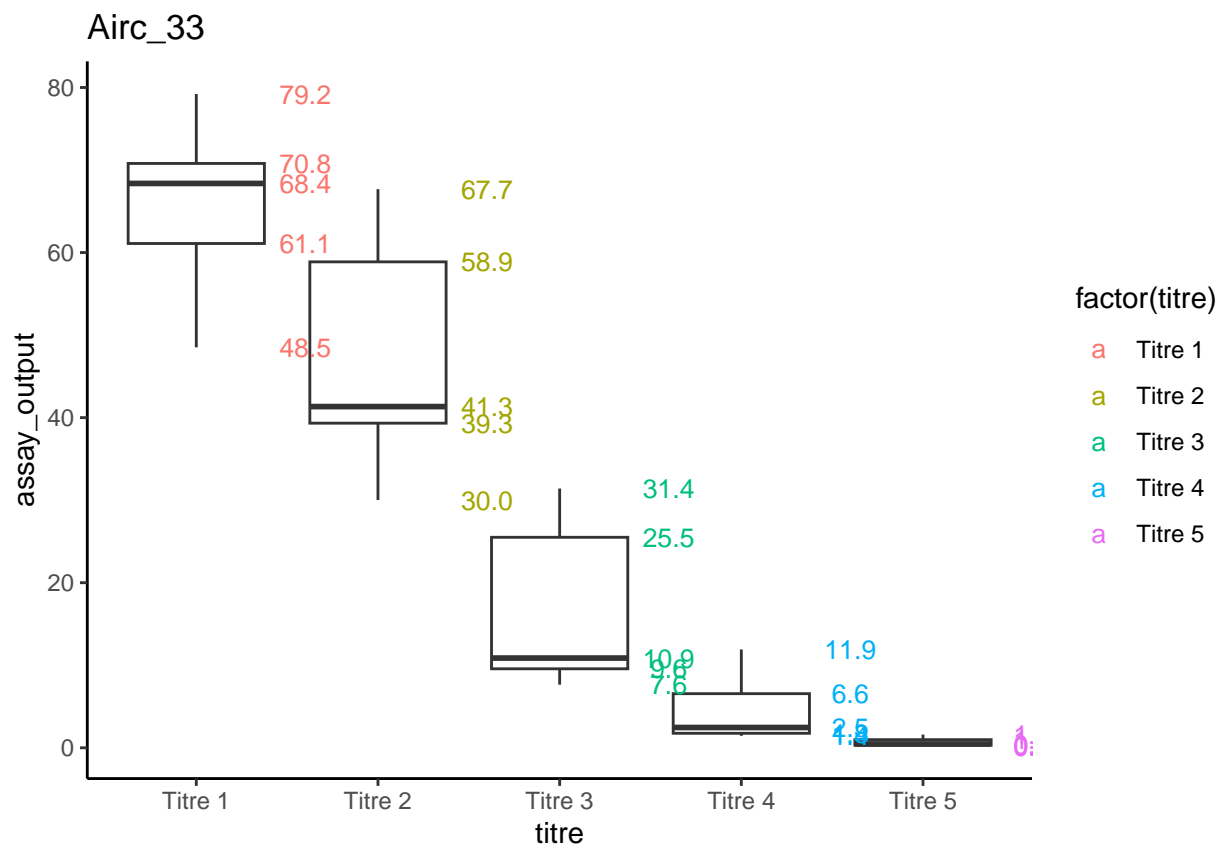
```
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



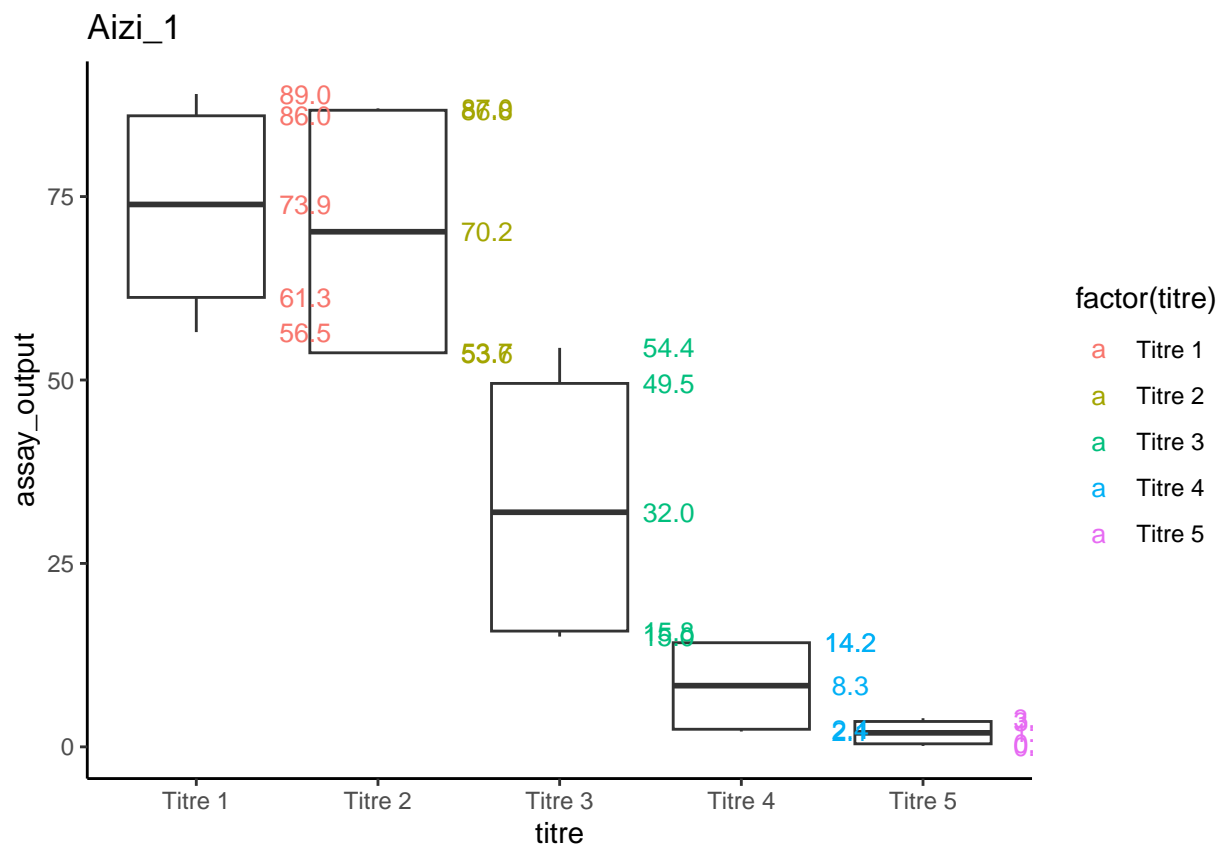
```
##
## [[2]]
```



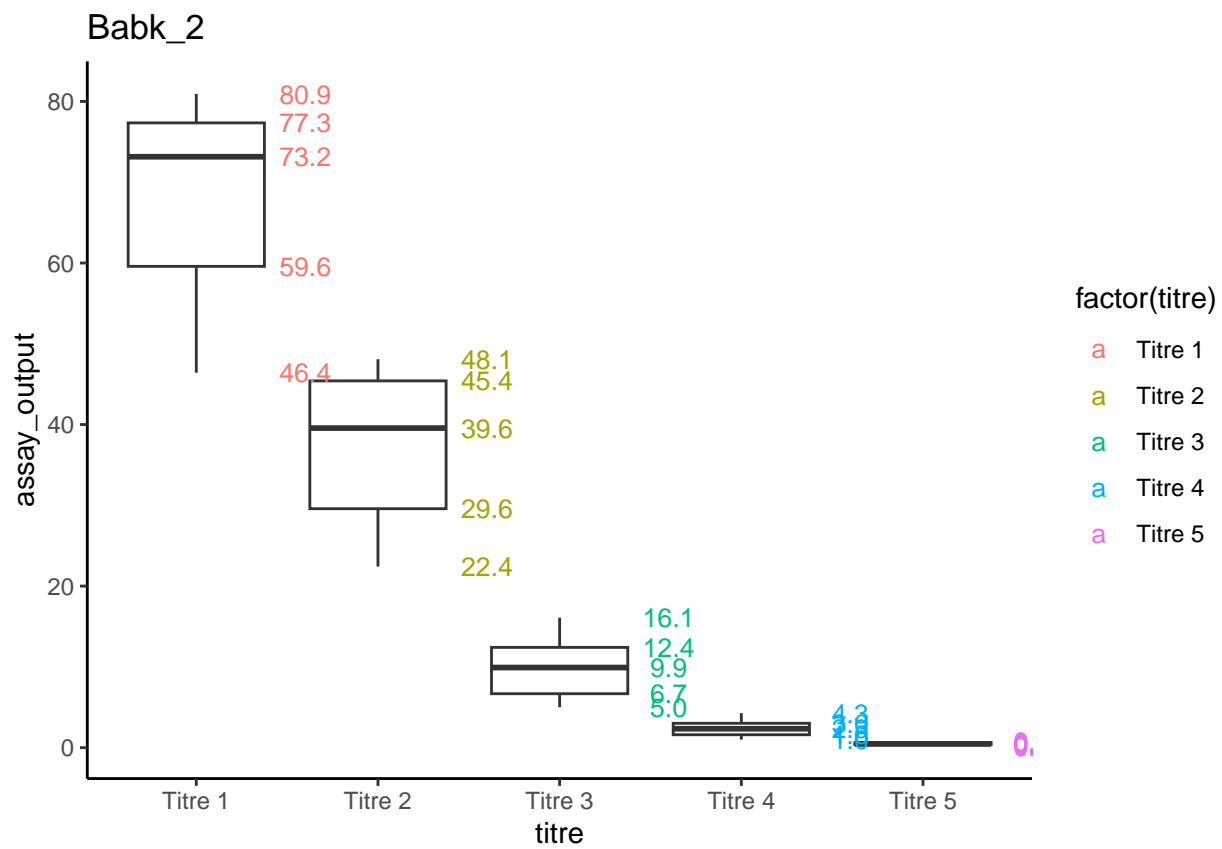
```
##
## [[3]]
```



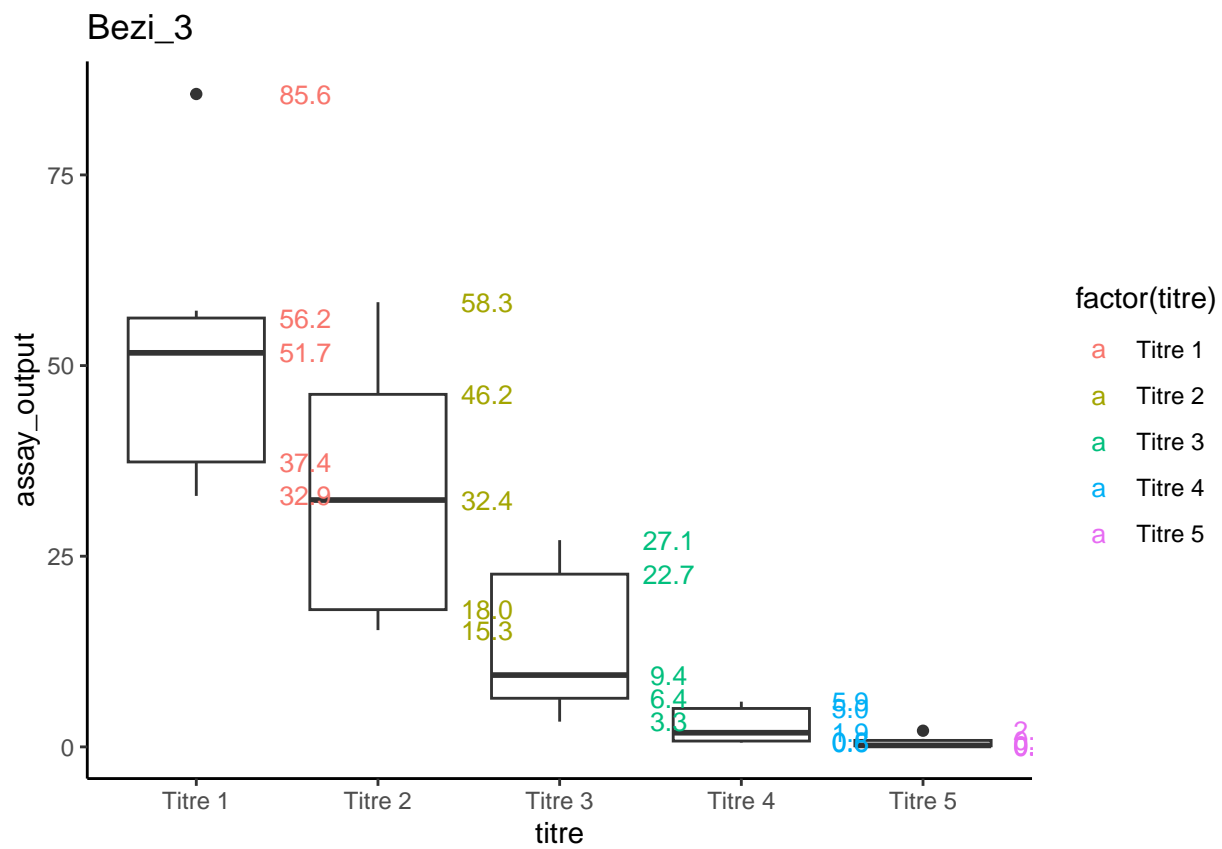
```
##
## [[4]]
```



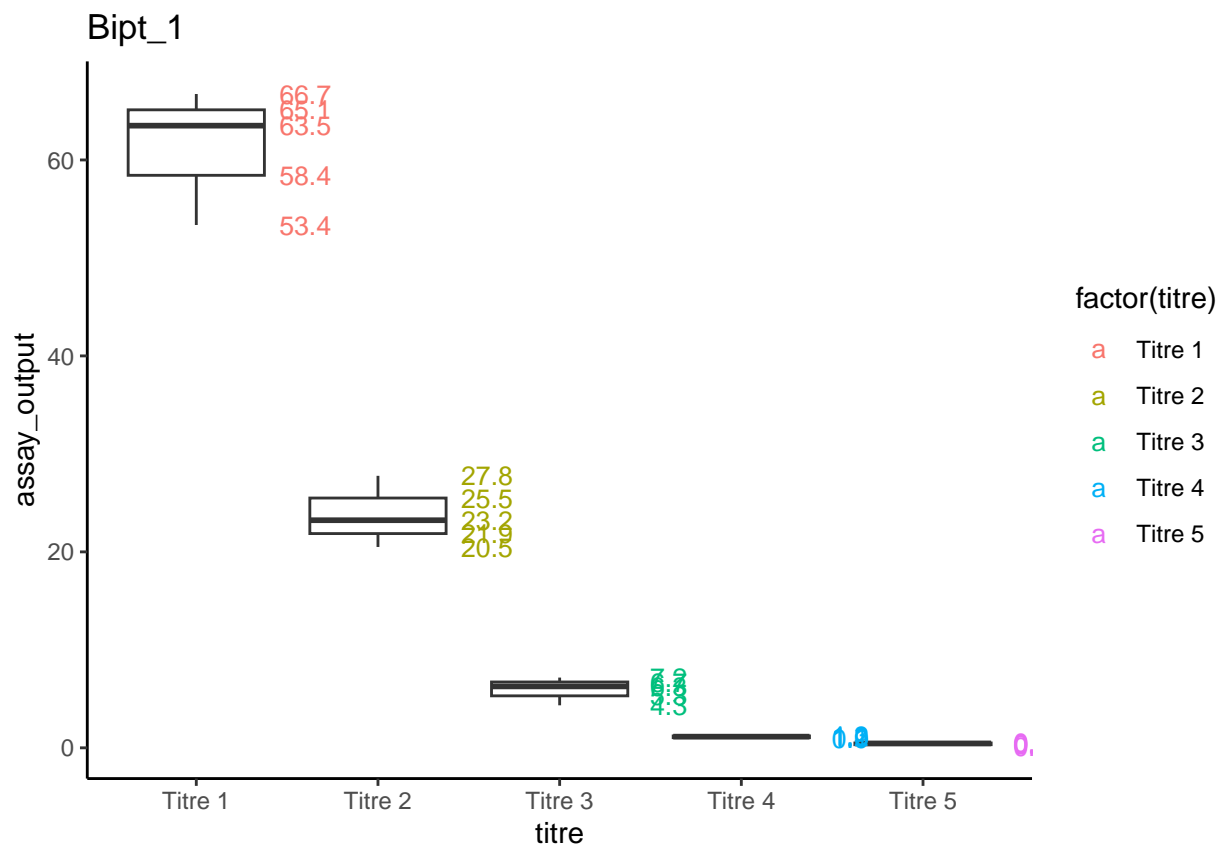
```
##
## [[5]]
```



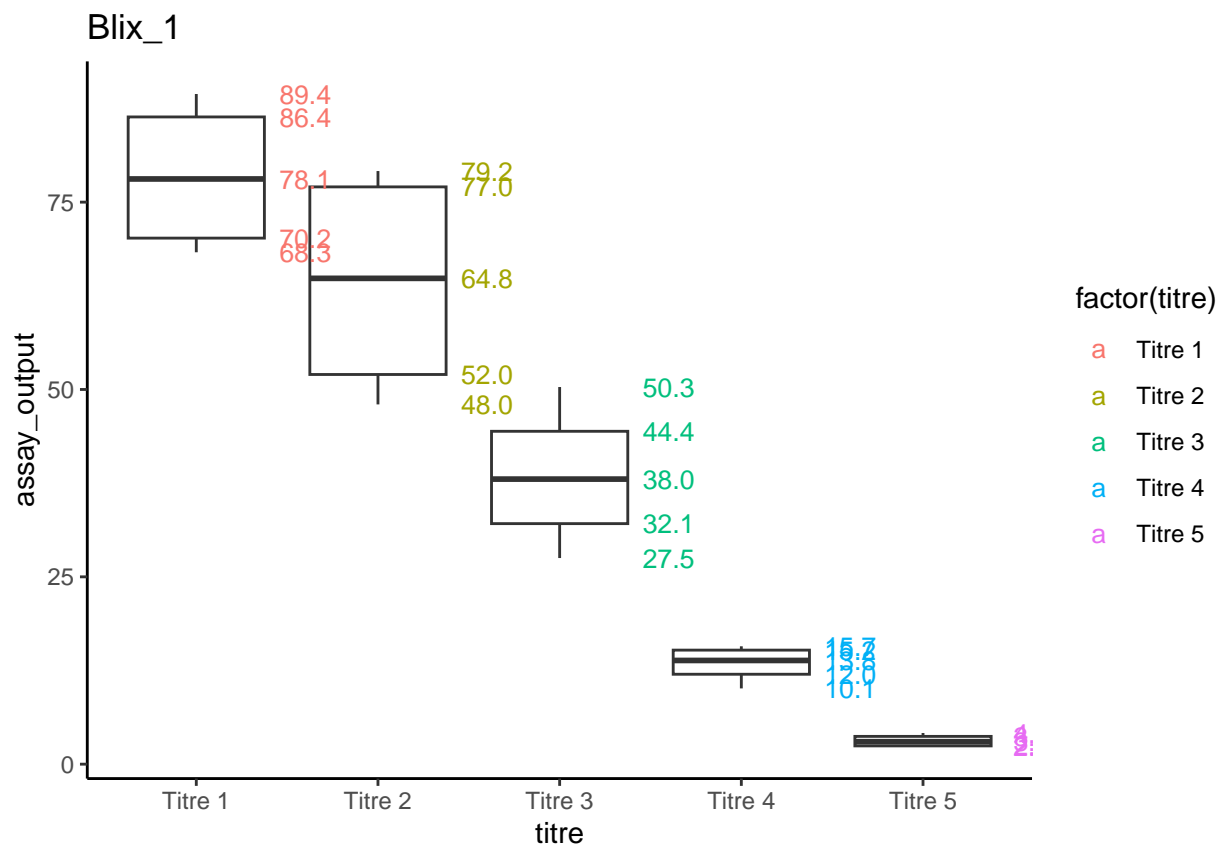
```
##
## [[6]]
```

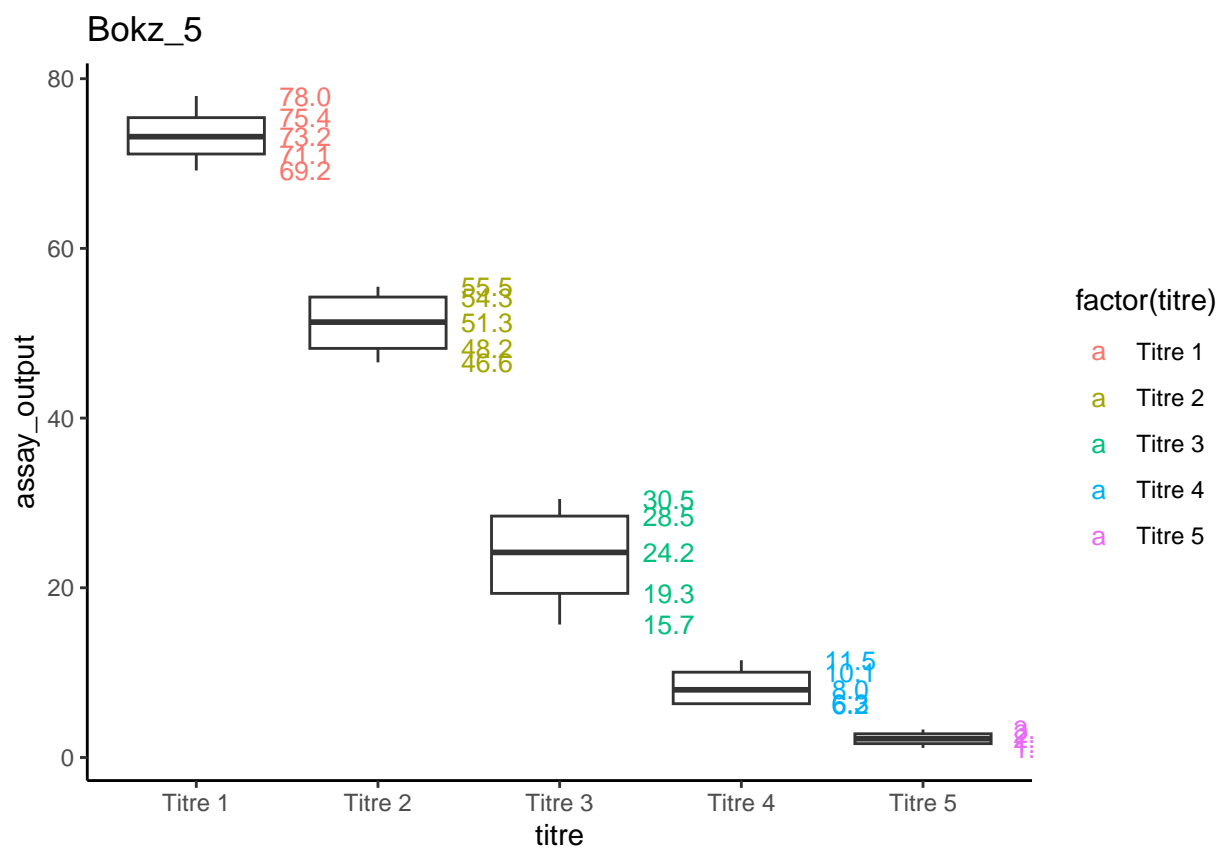
```
##
## [[7]]
```



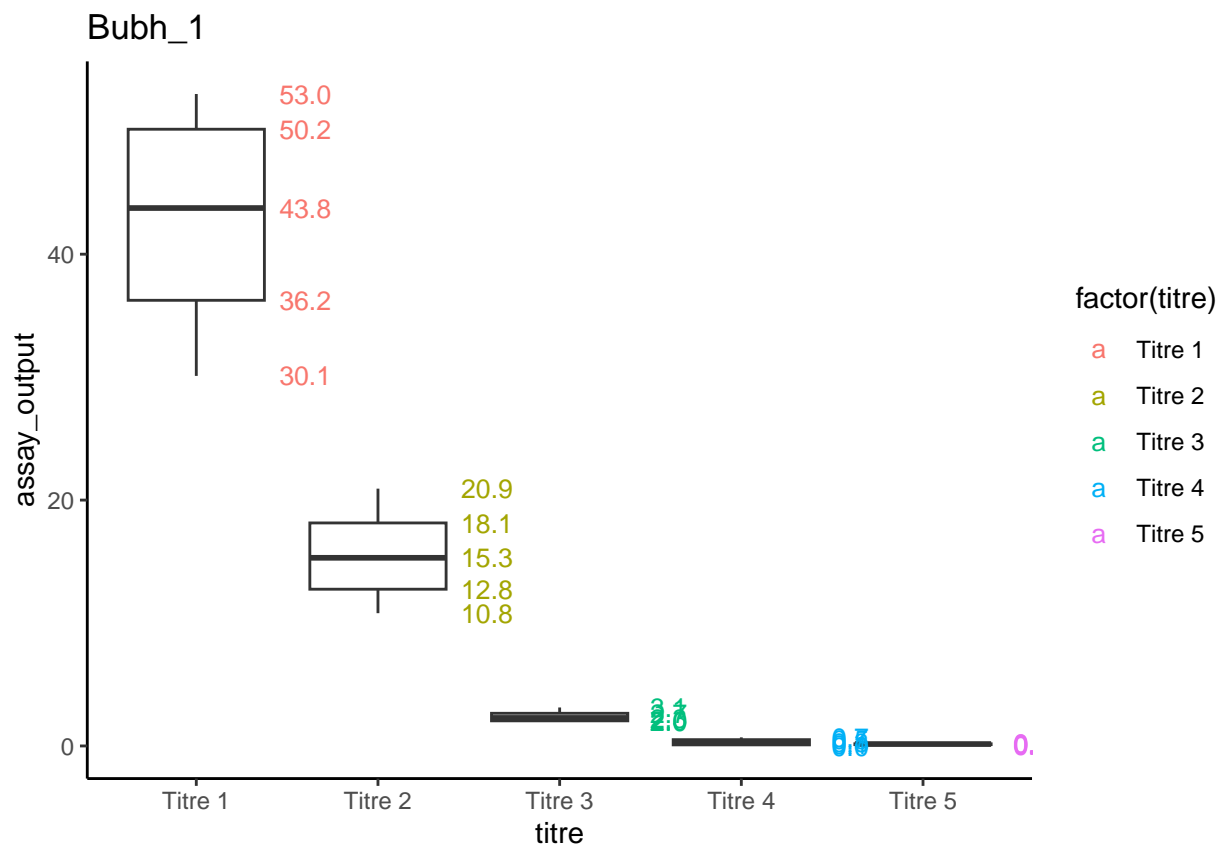
```
##
## [[8]]
```



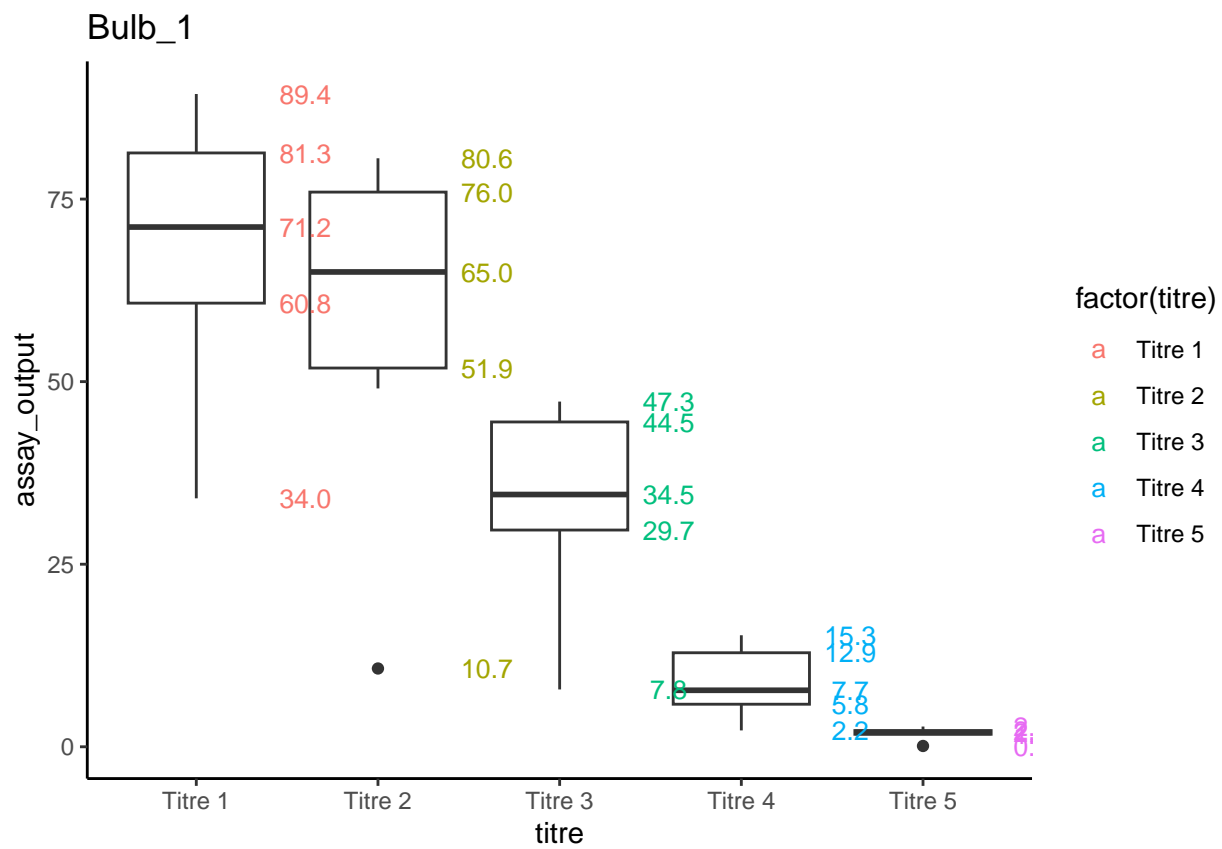
```
##
## [[9]]
```



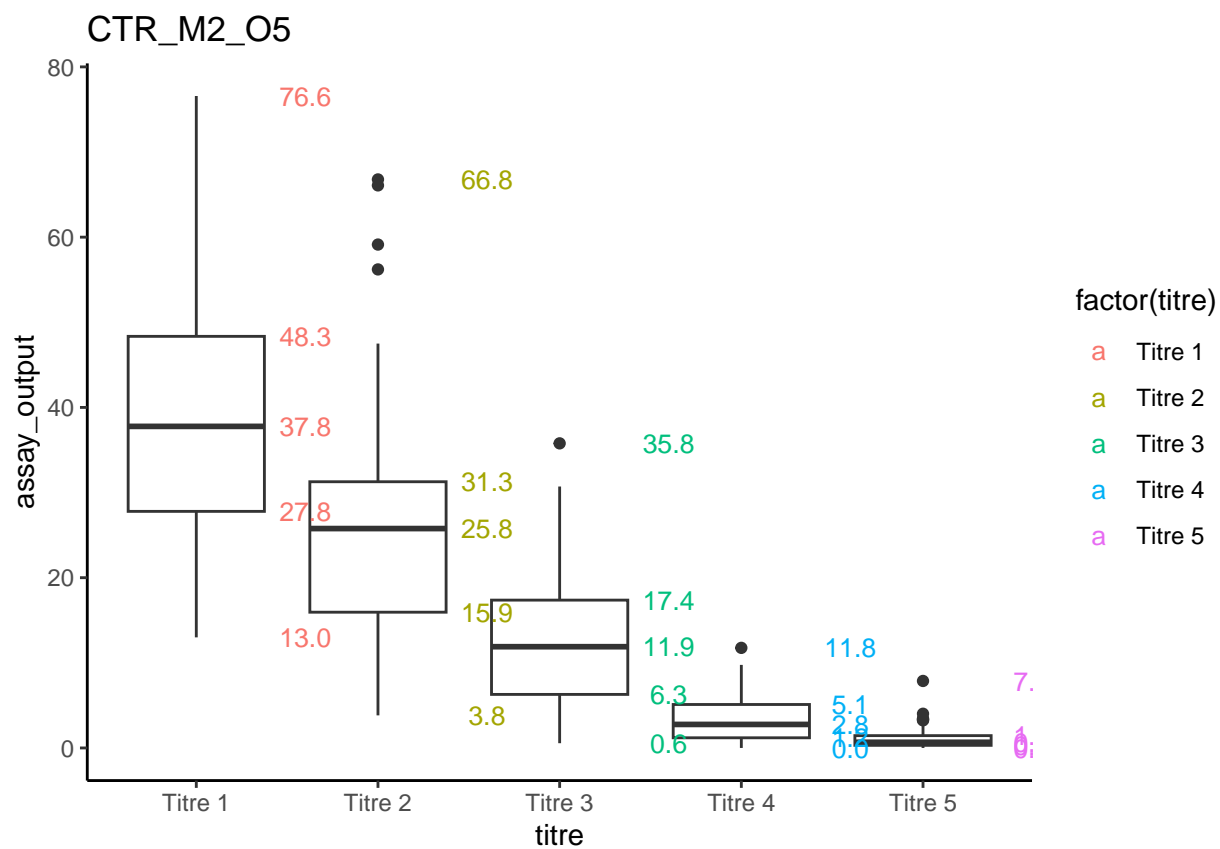
```
##  
## [[10]]
```



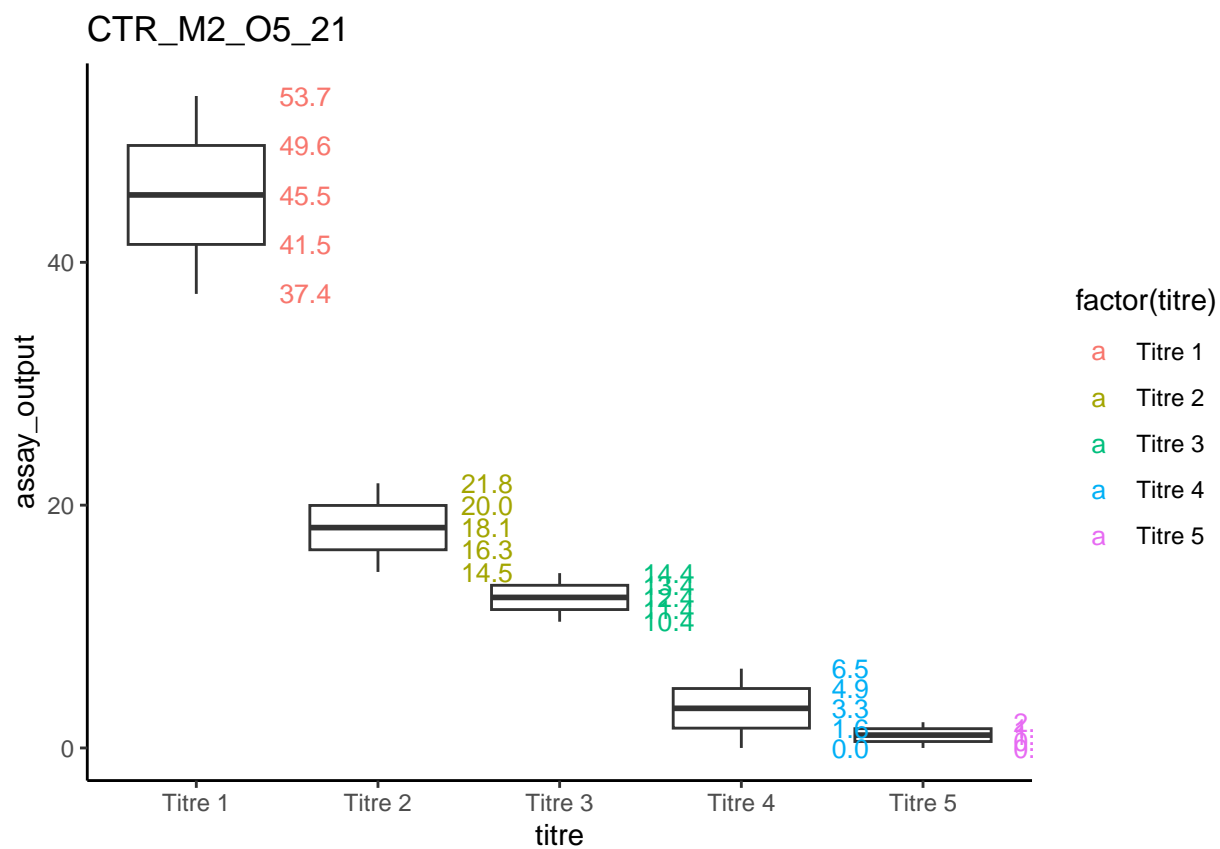
```
##
## [[11]]
```



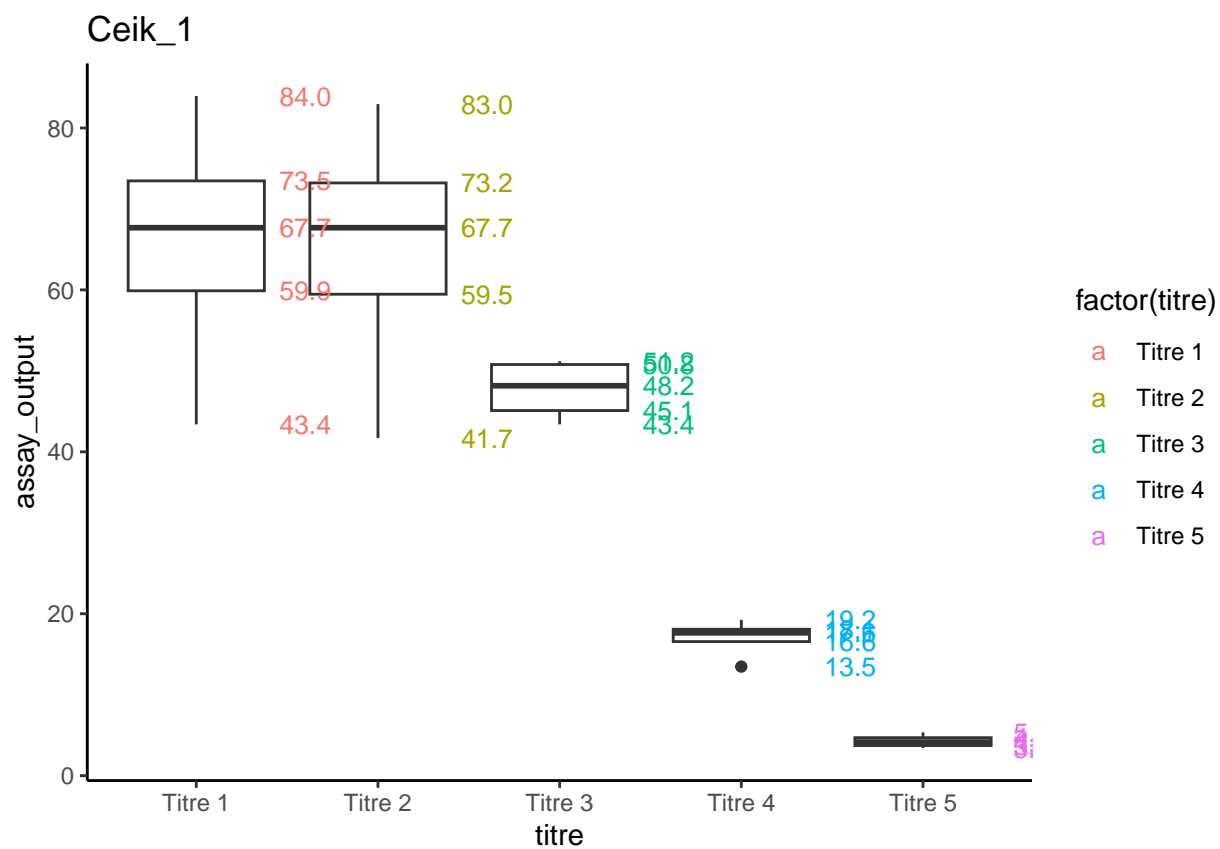
```
##
## [[12]]
```



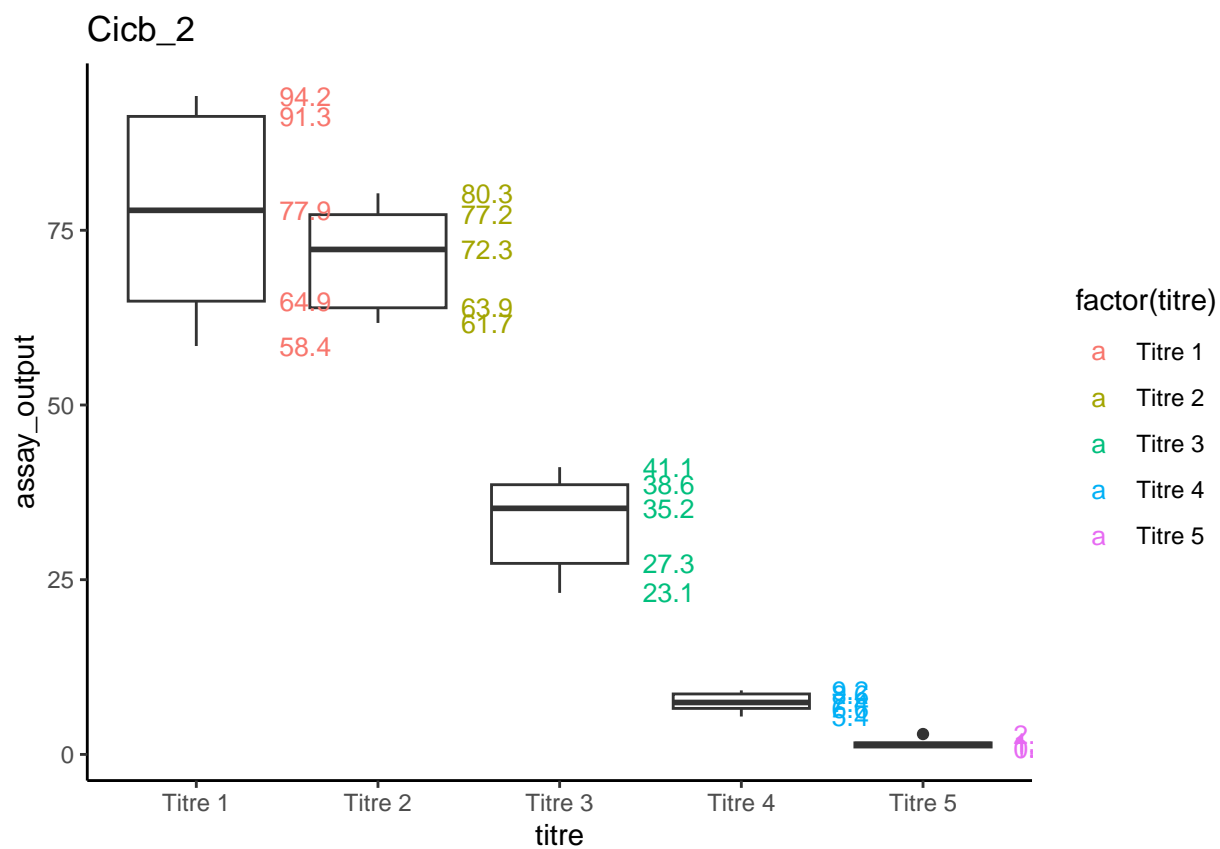
[[13]]



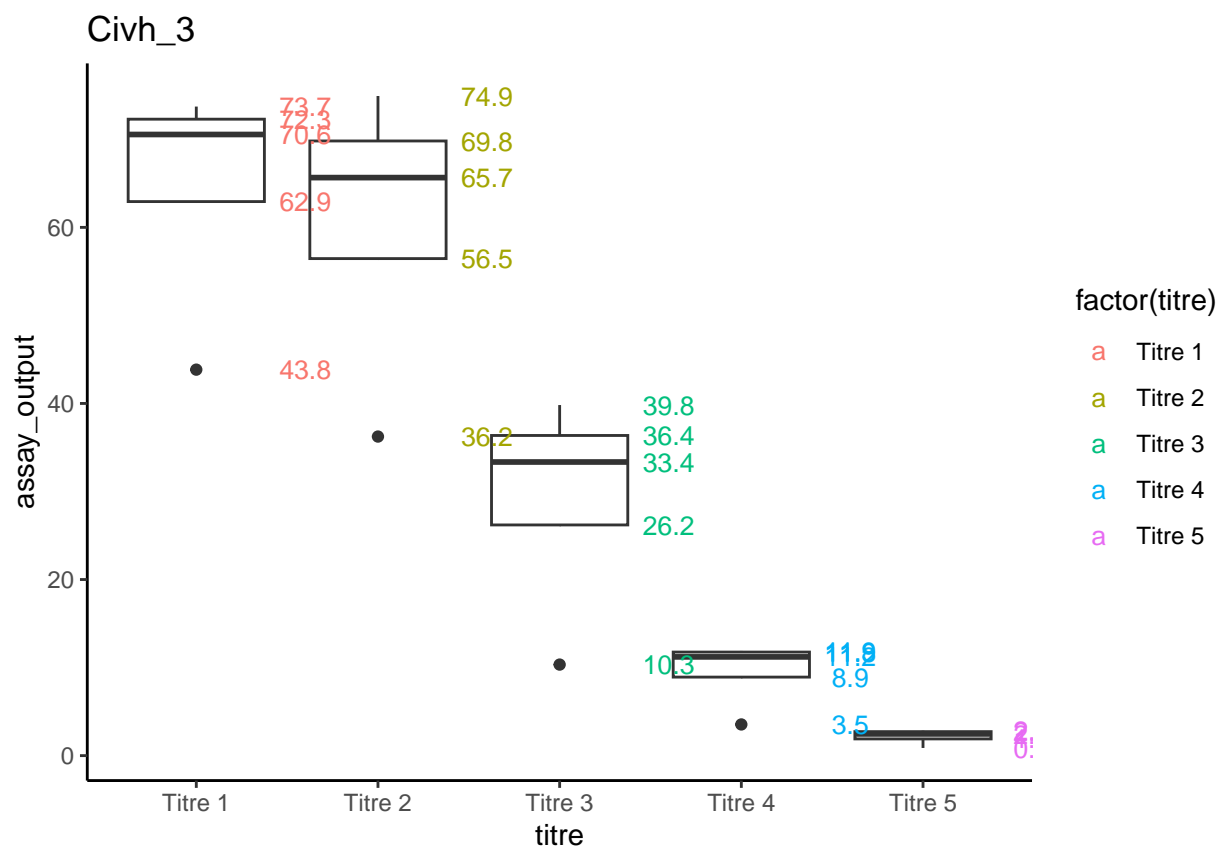
[[14]]



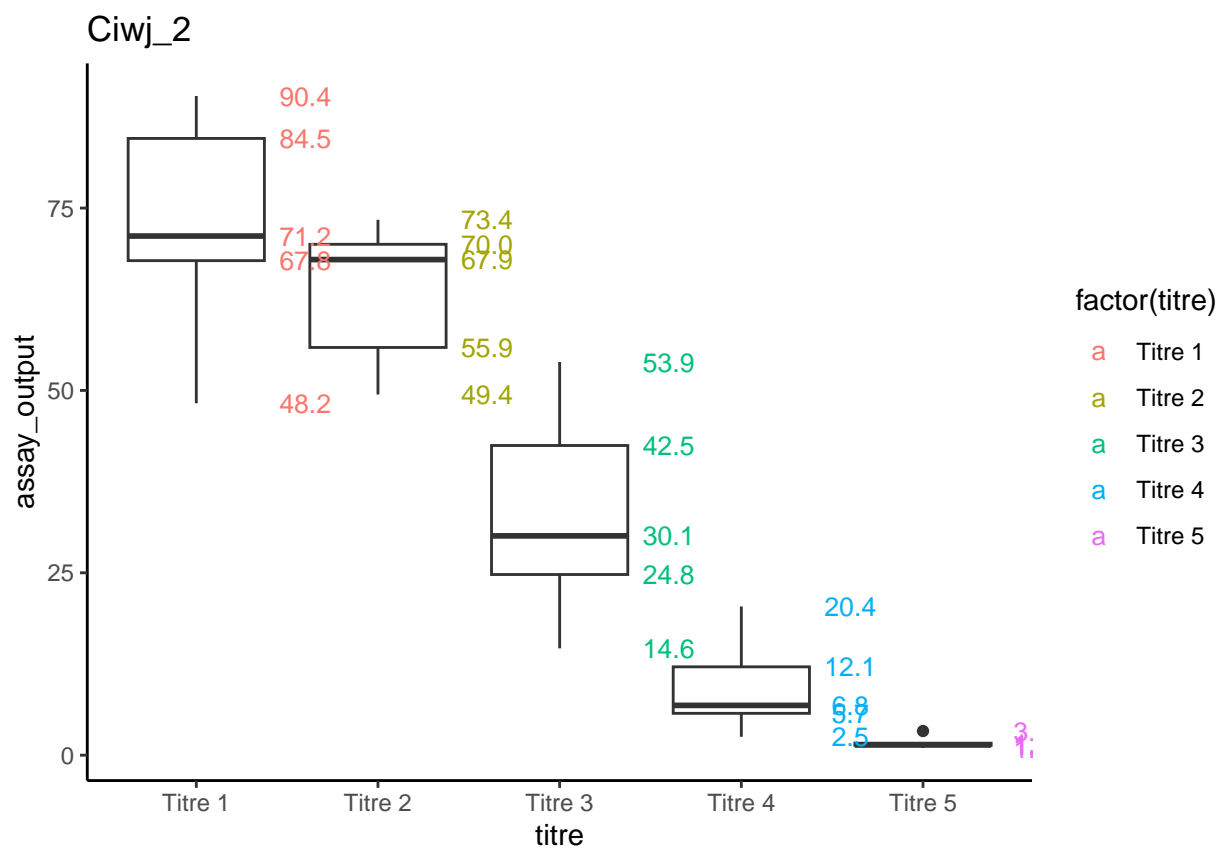
[[15]]



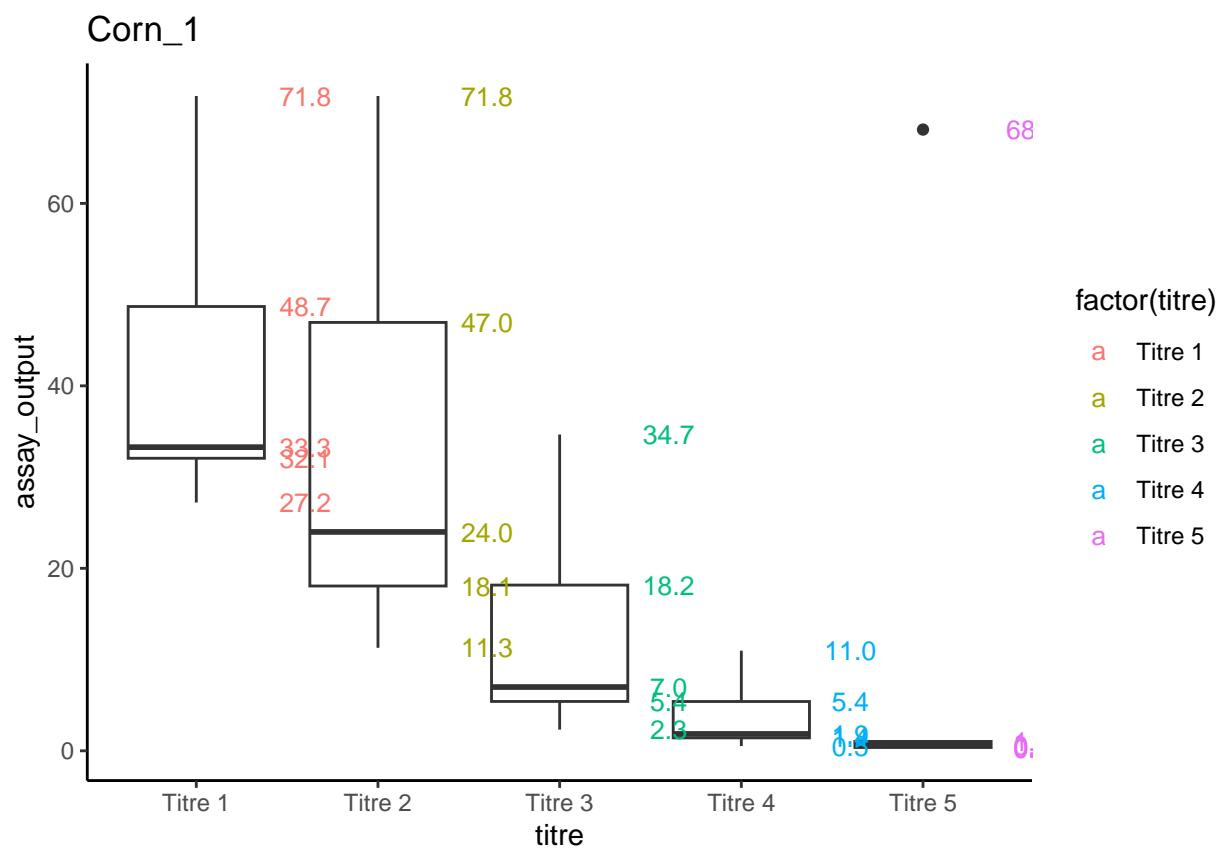
```
##
## [[16]]
```



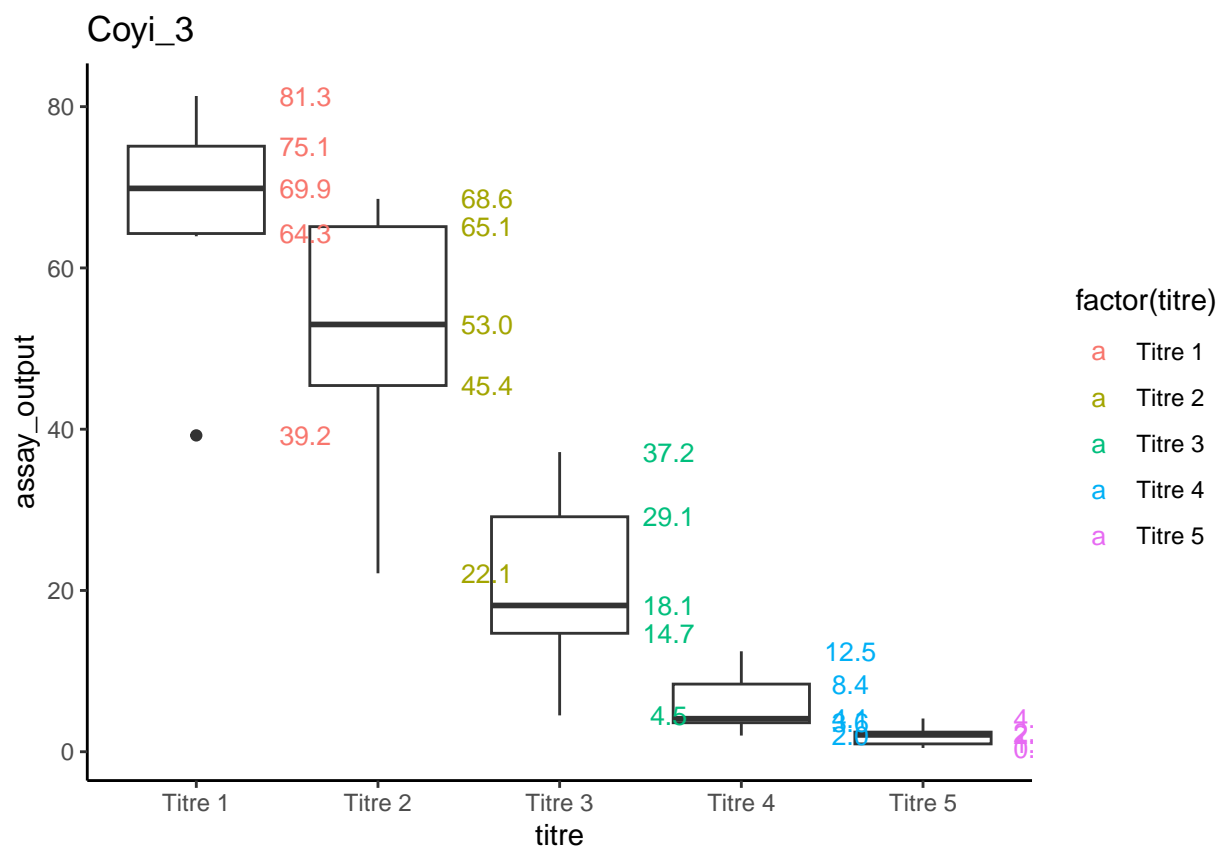
```
##
## [[17]]
```



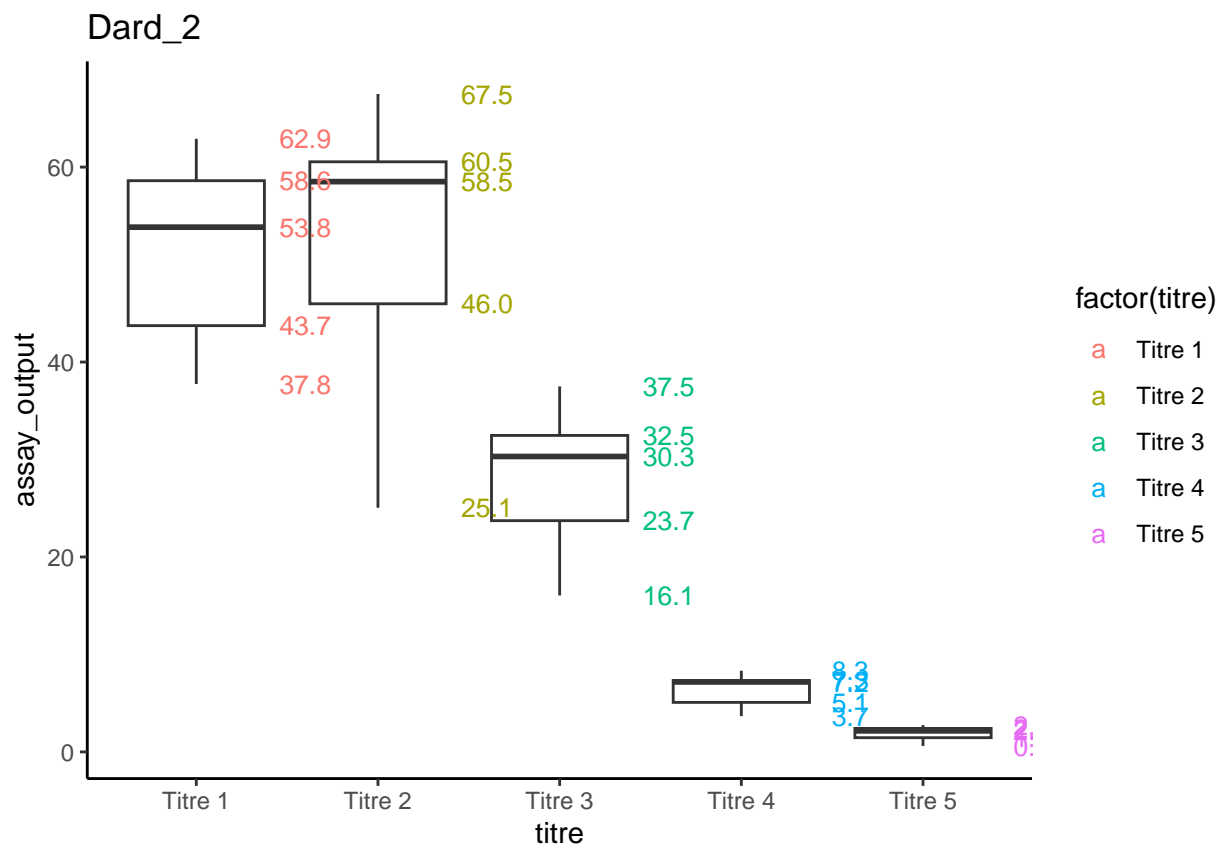
```
##
## [[18]]
```



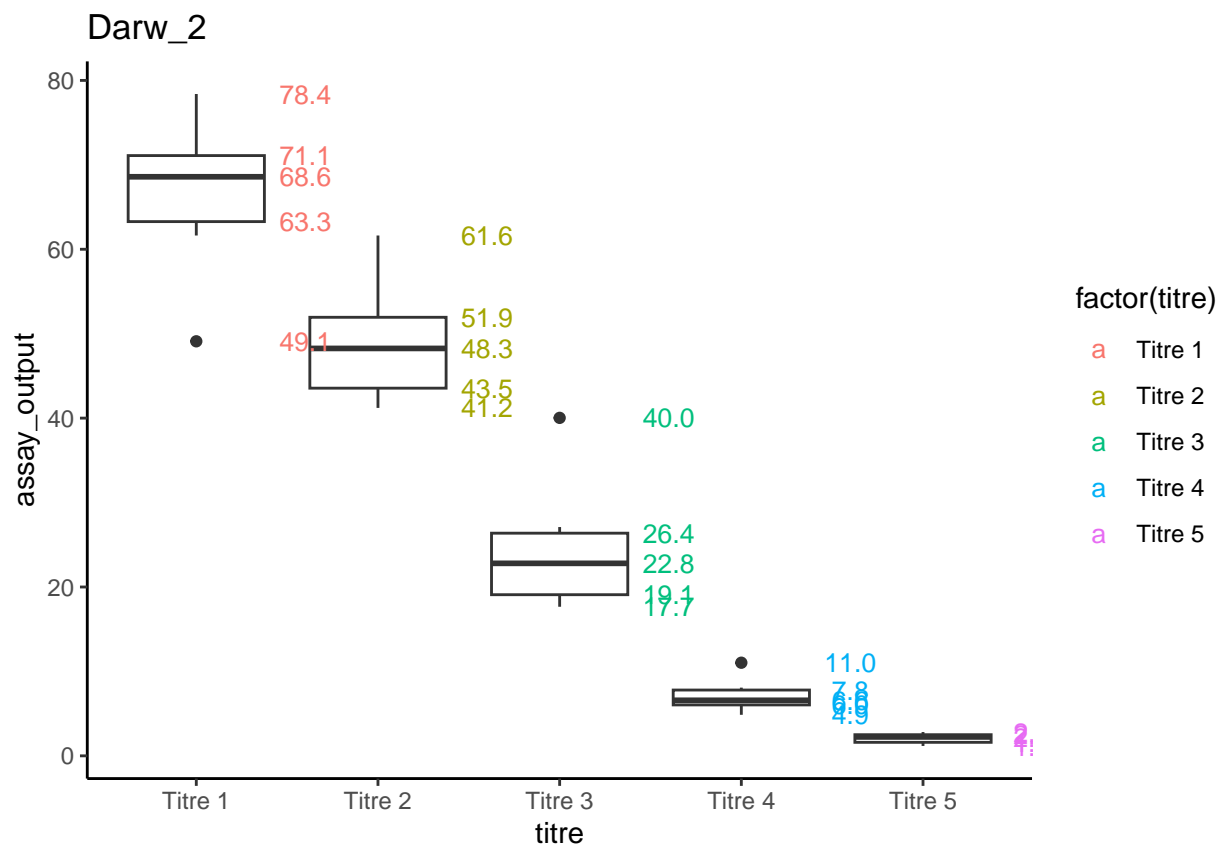
```
##
## [[19]]
```



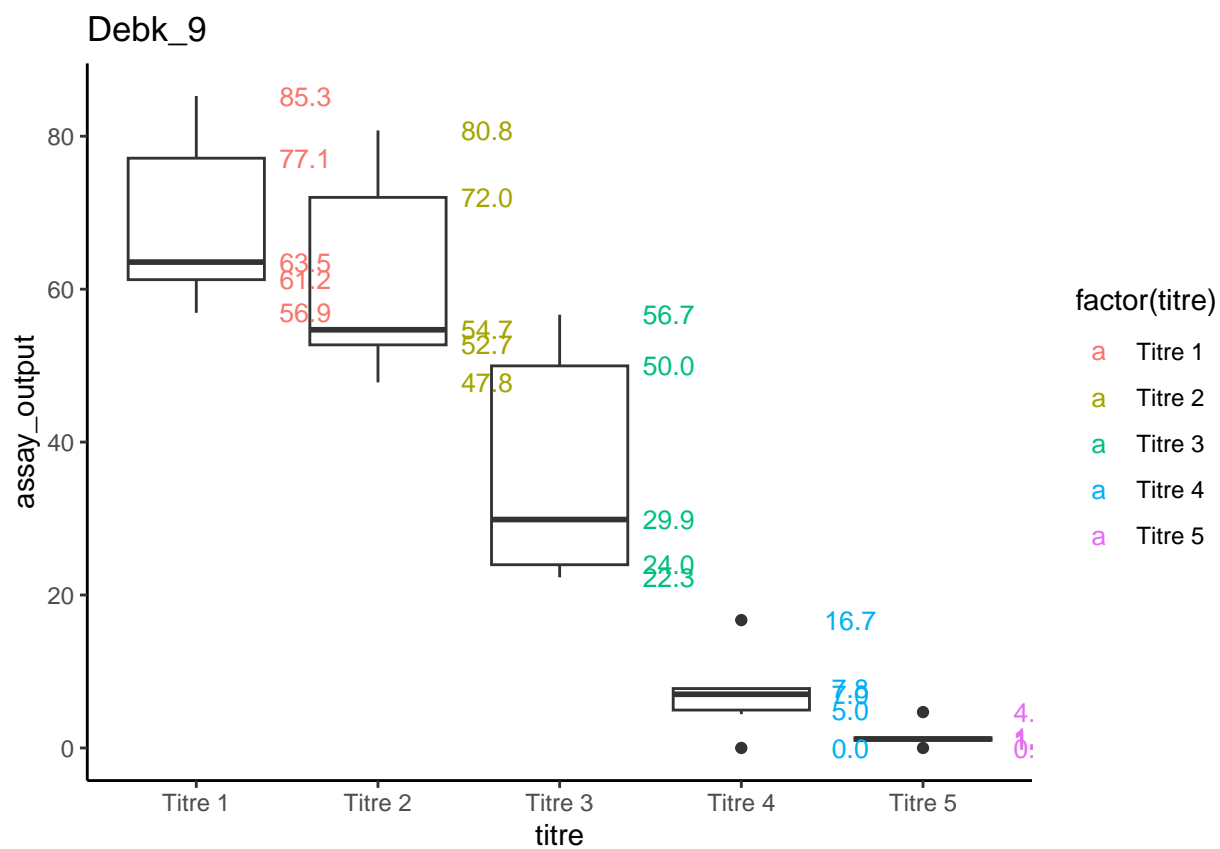
[[20]]



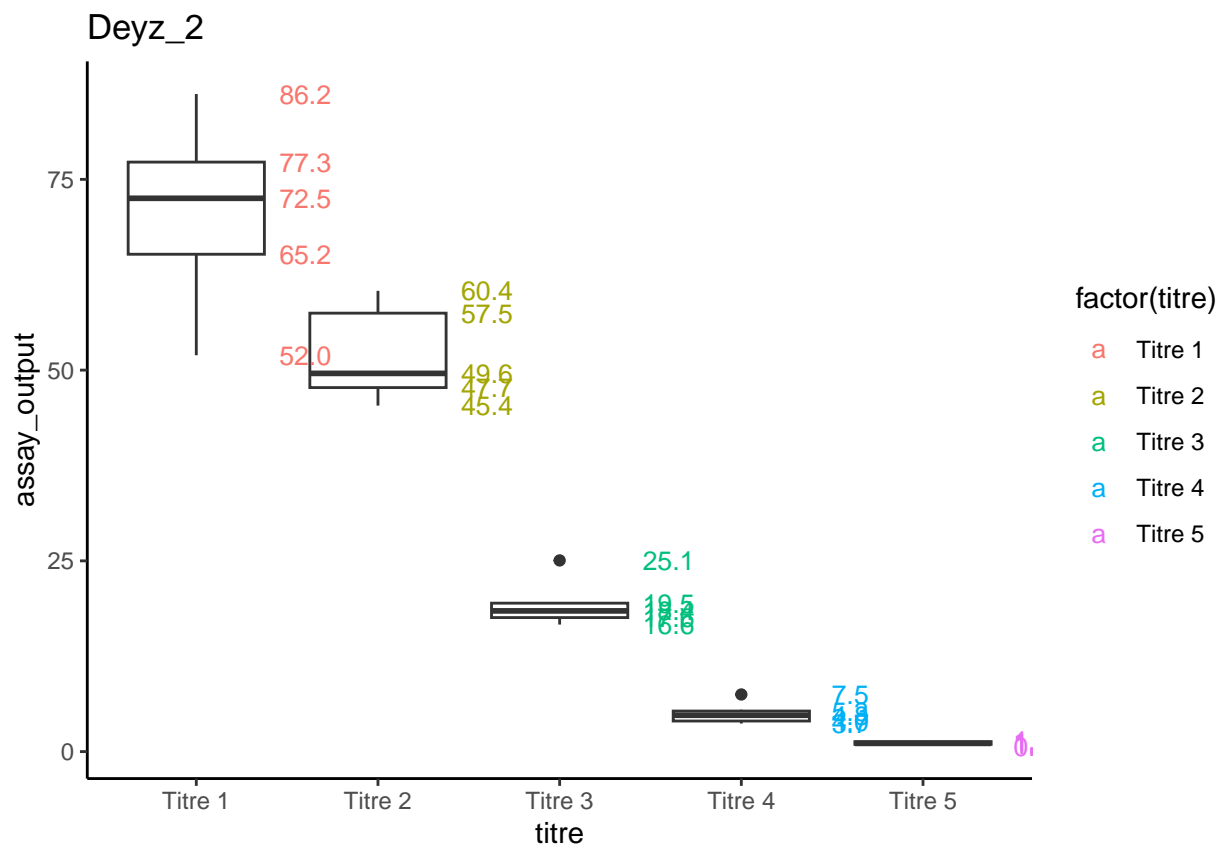
[[21]]



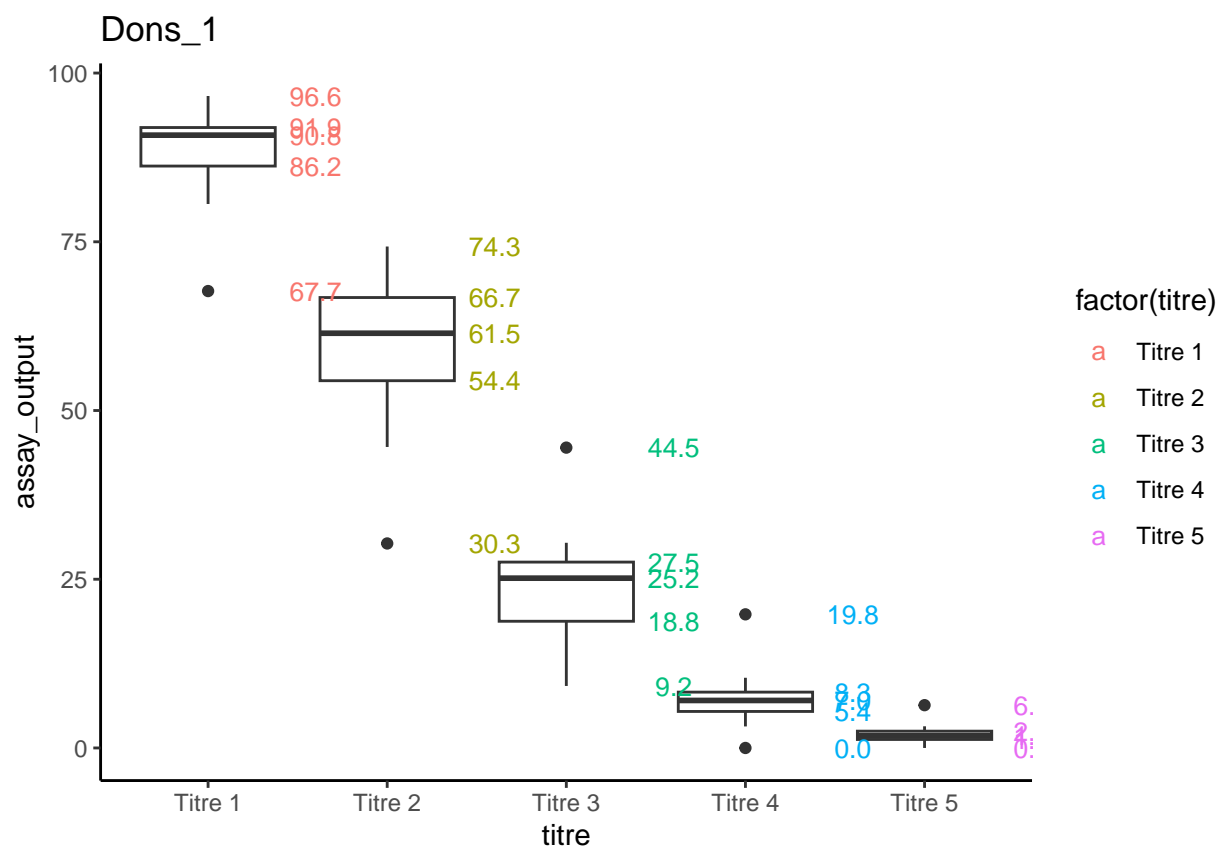
```
##
## [[22]]
```

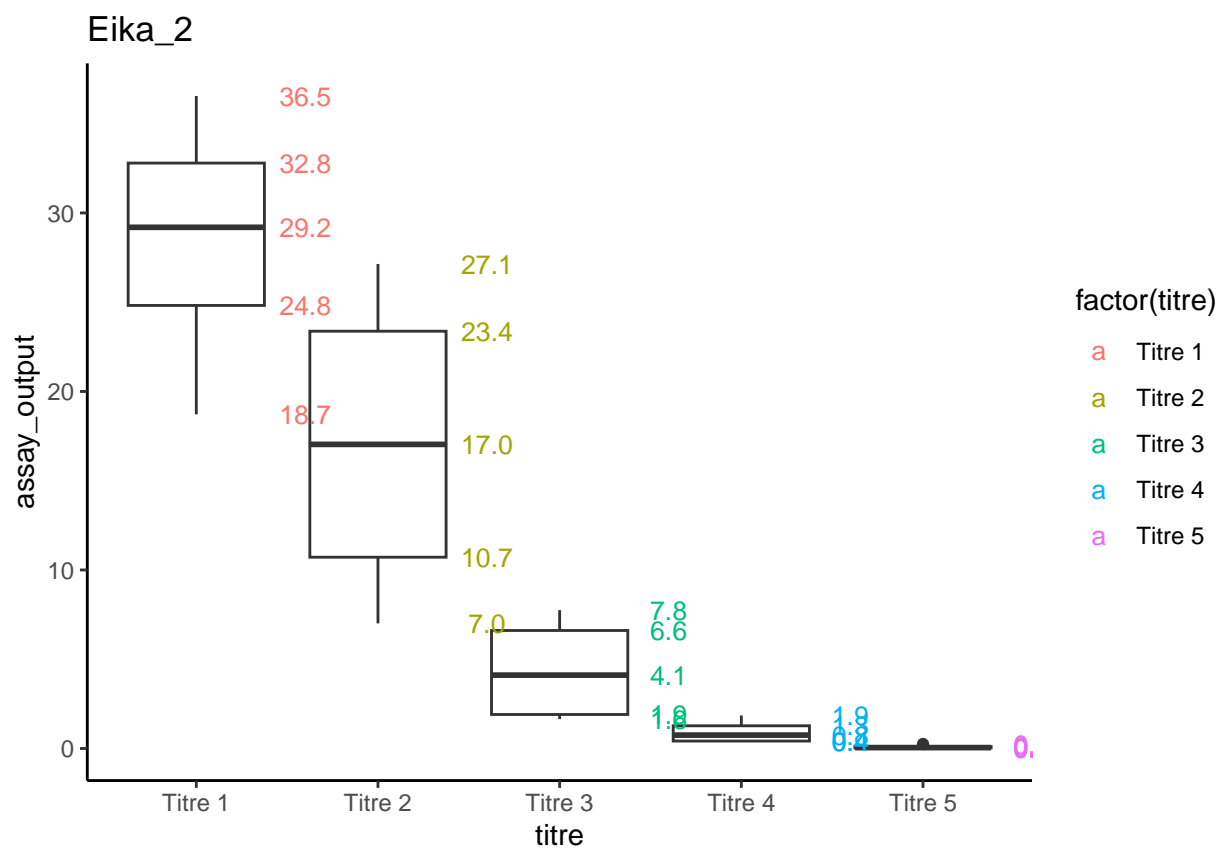
```
##
## [[23]]
```



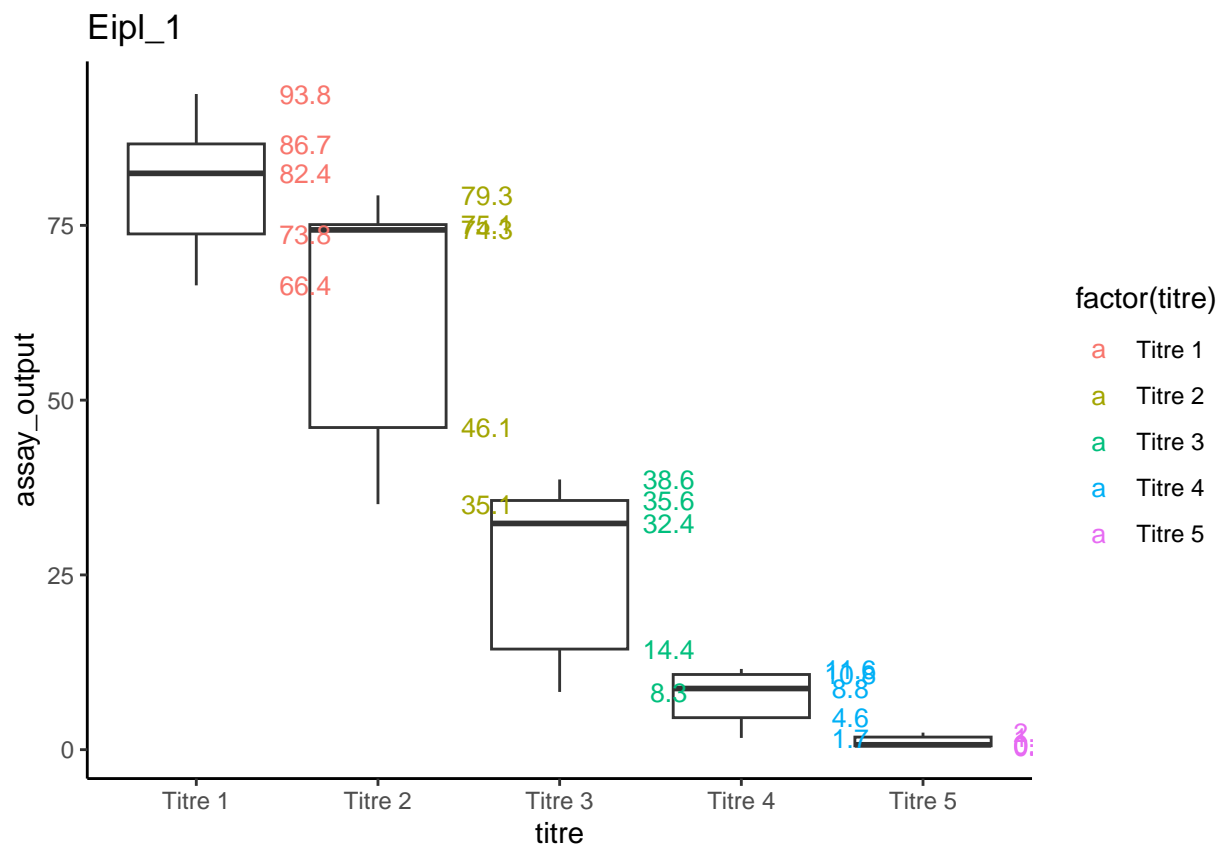
[[24]]



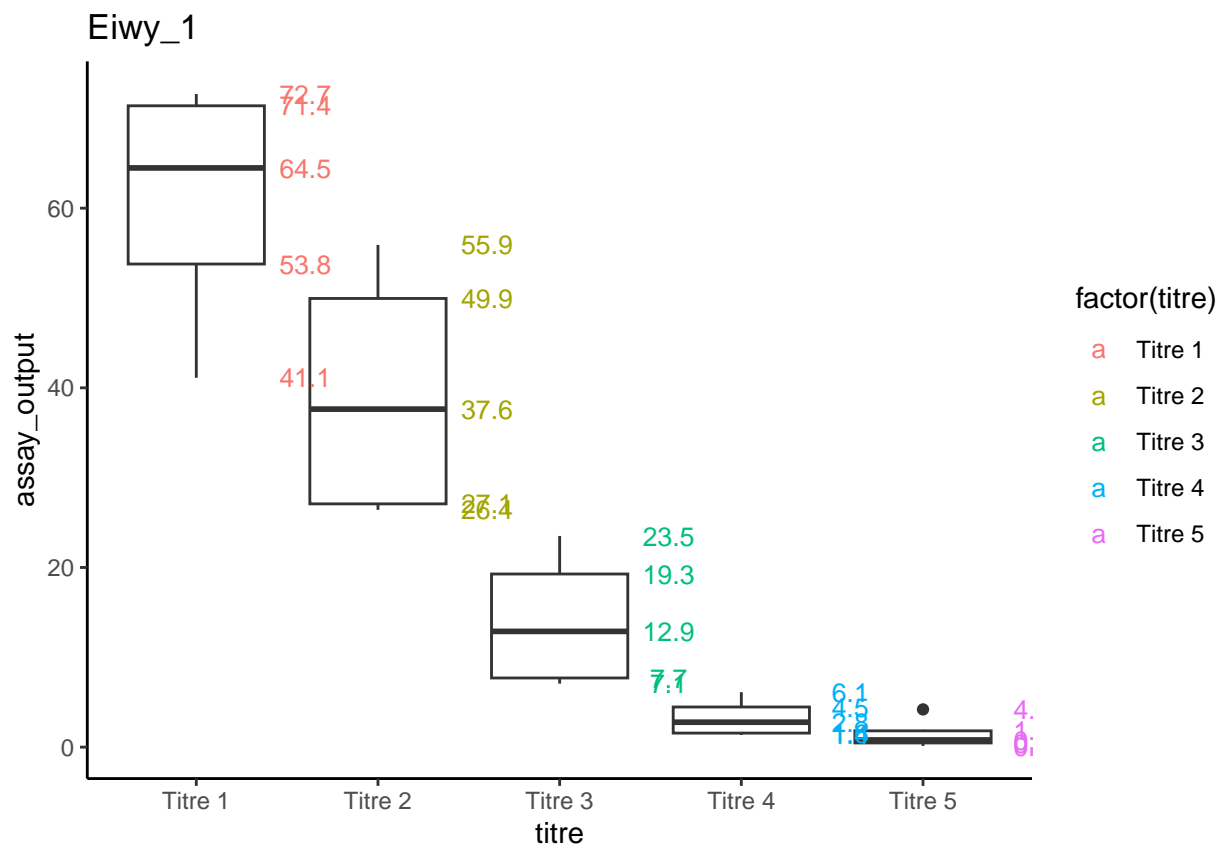
```
##
## [[25]]
```



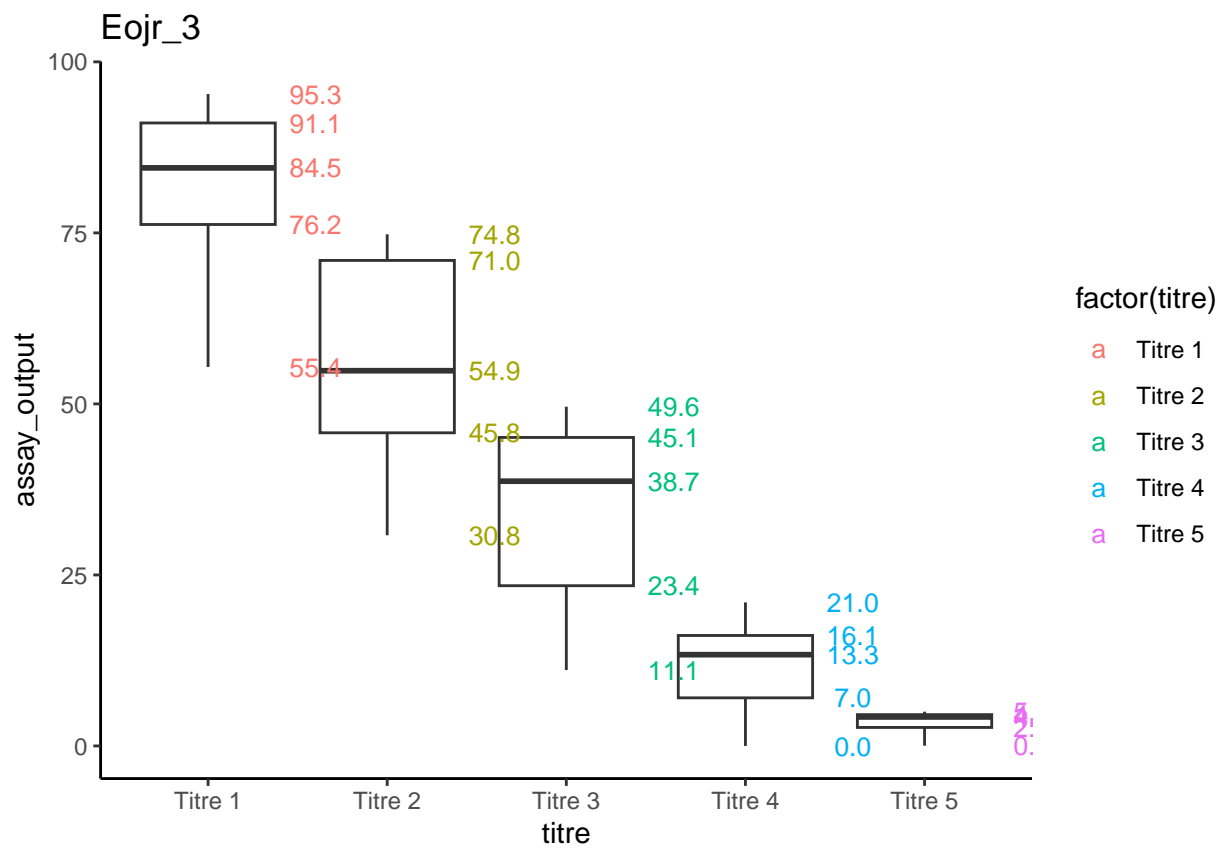
```
##
## [[26]]
```



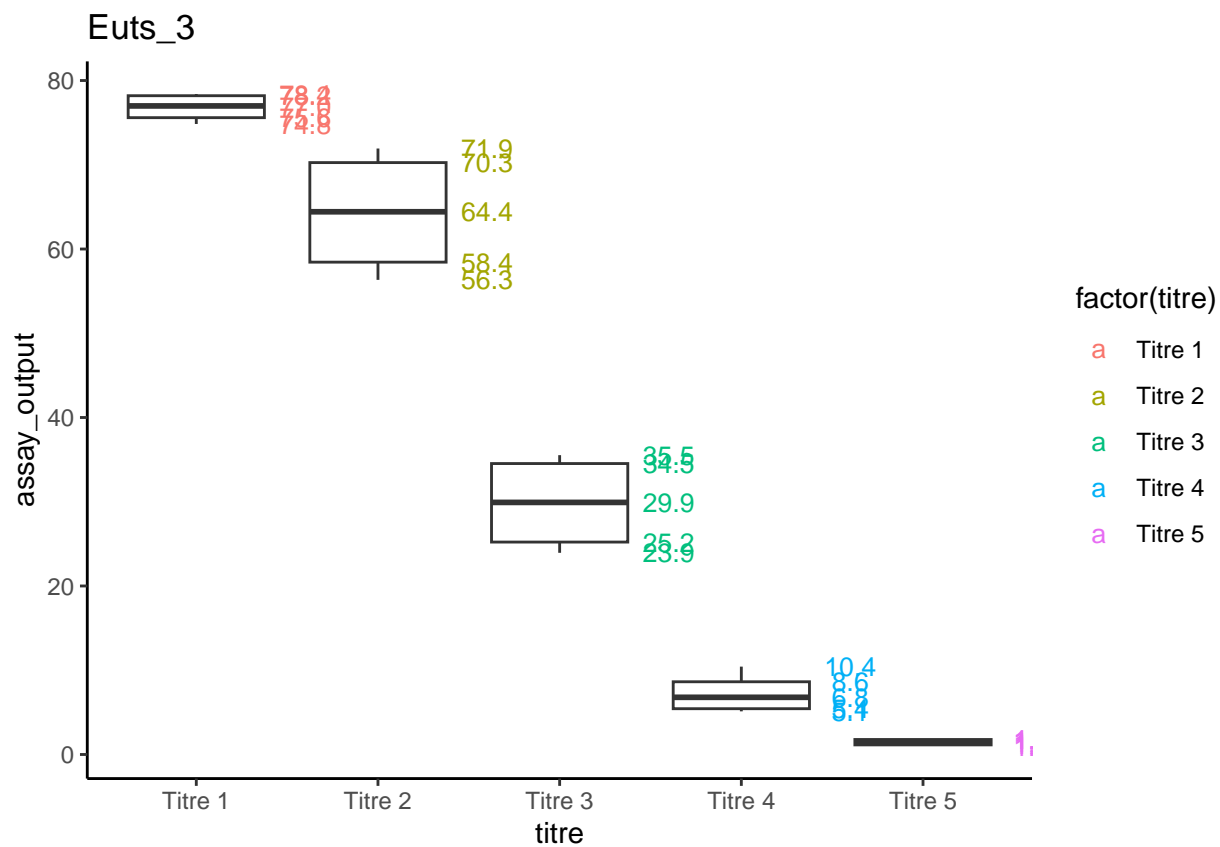
```
##
## [[27]]
```



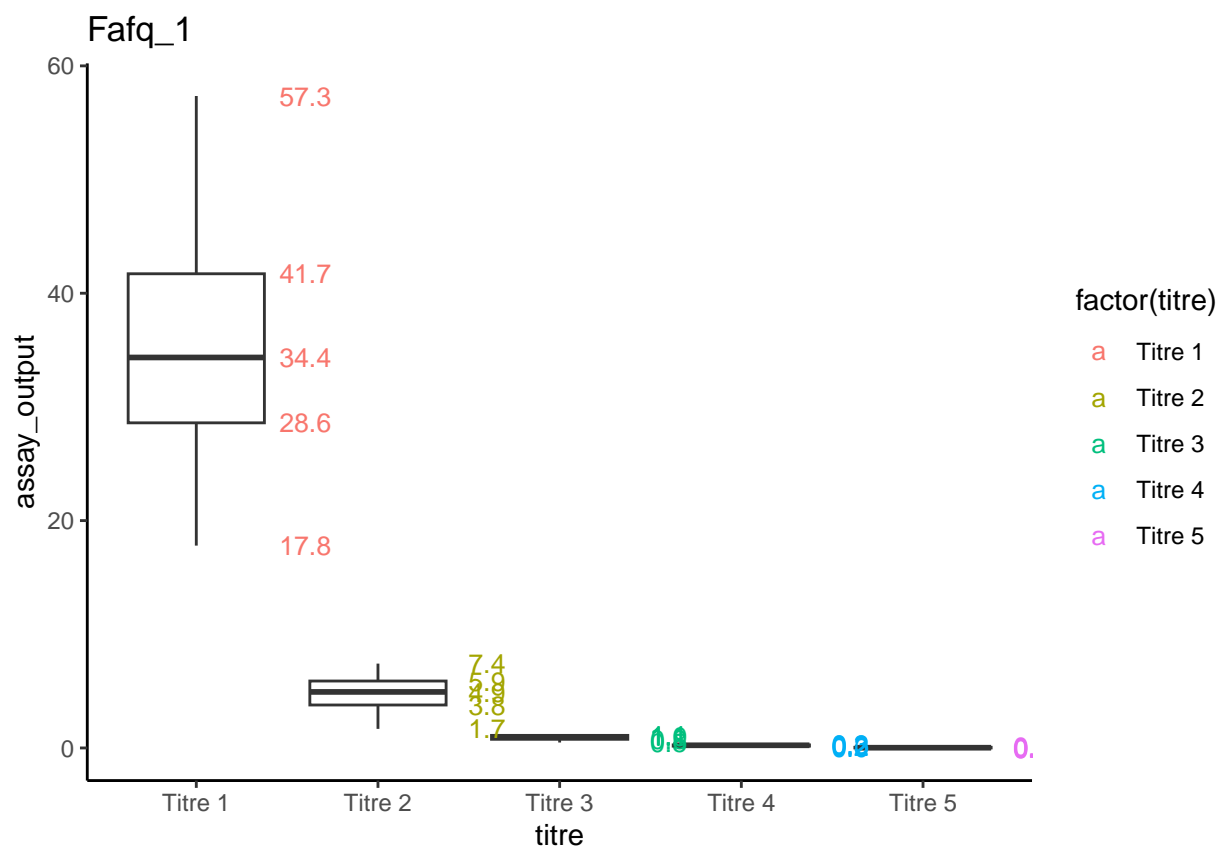
[[28]]



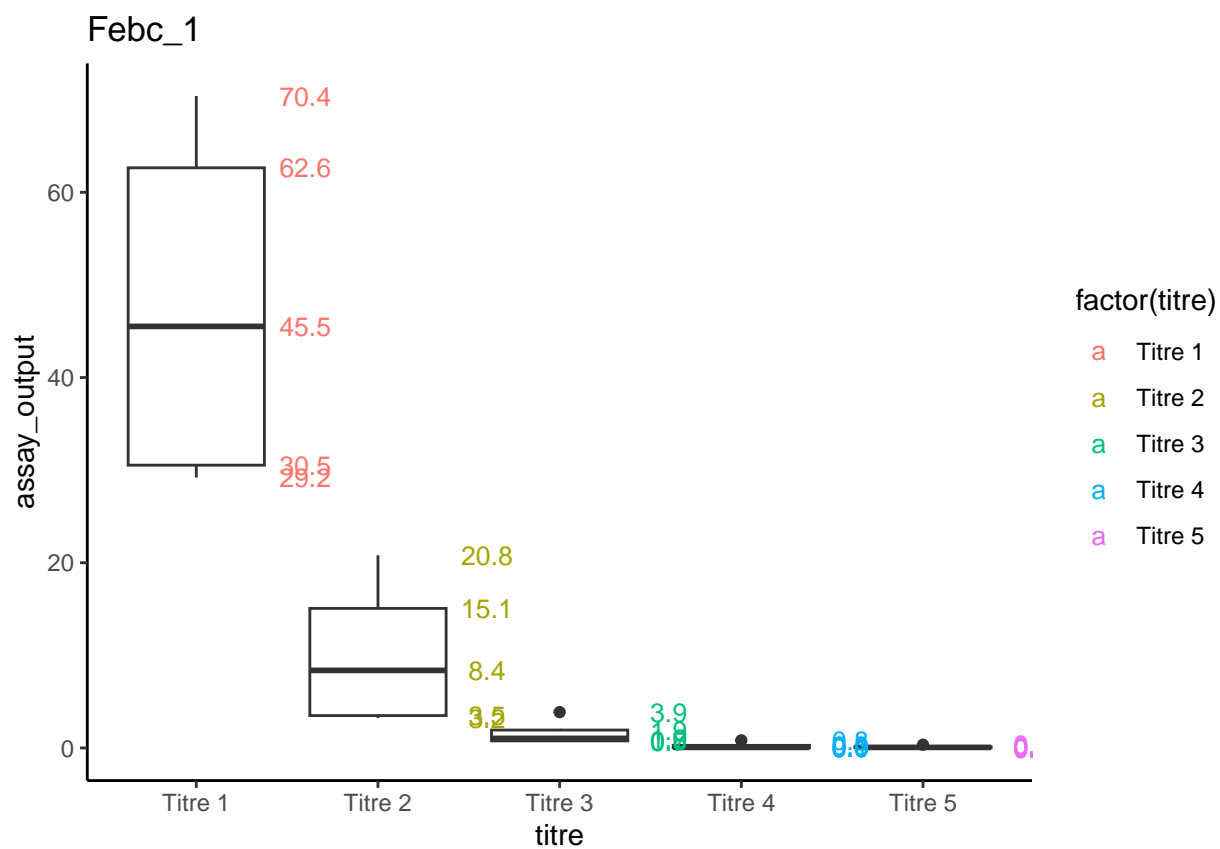
```
##
## [[29]]
```



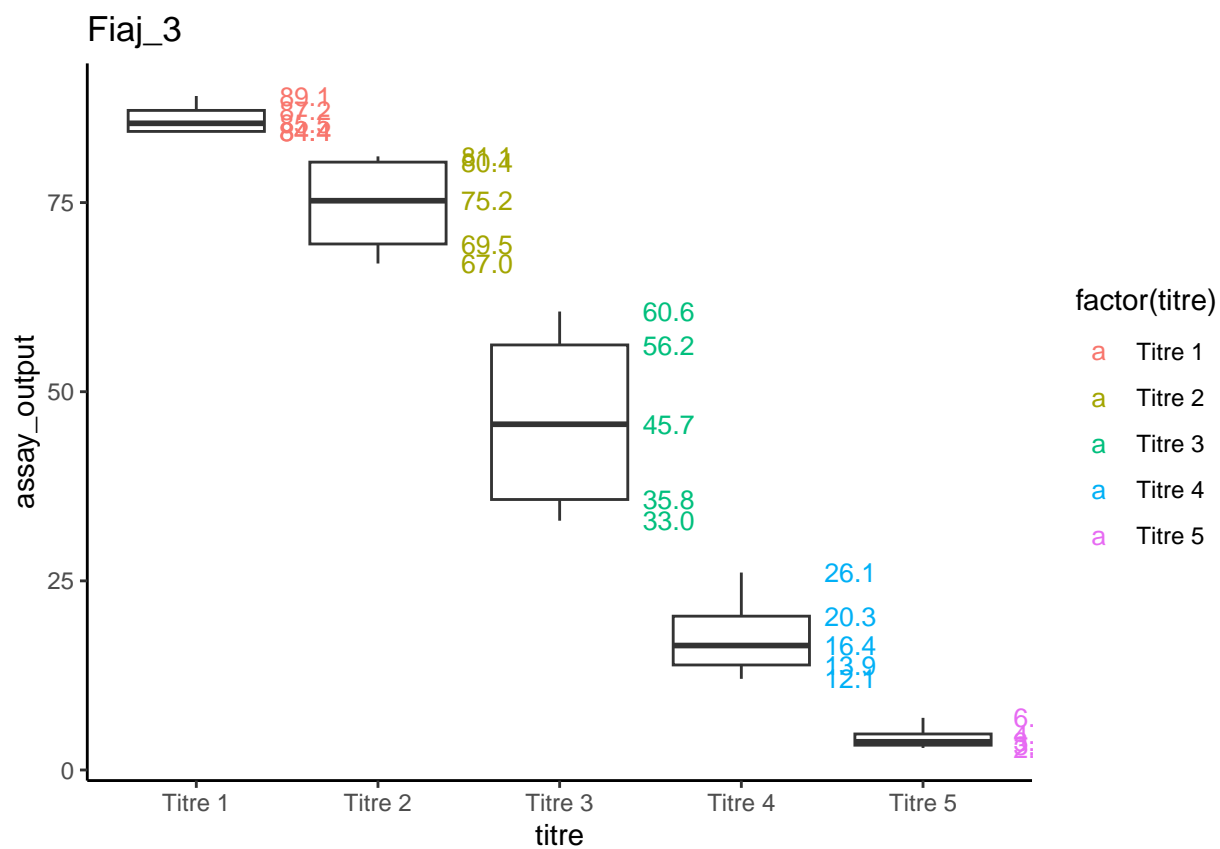
[[30]]



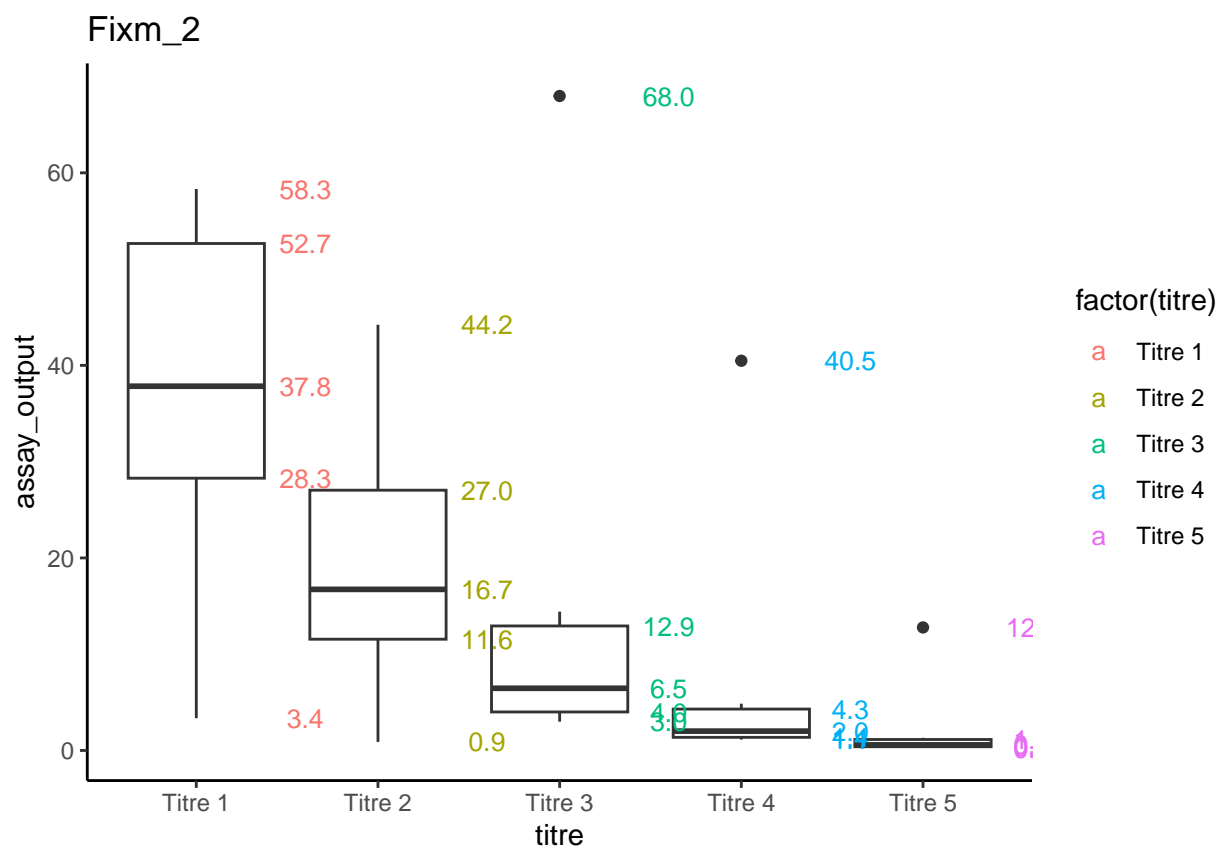
[[31]]



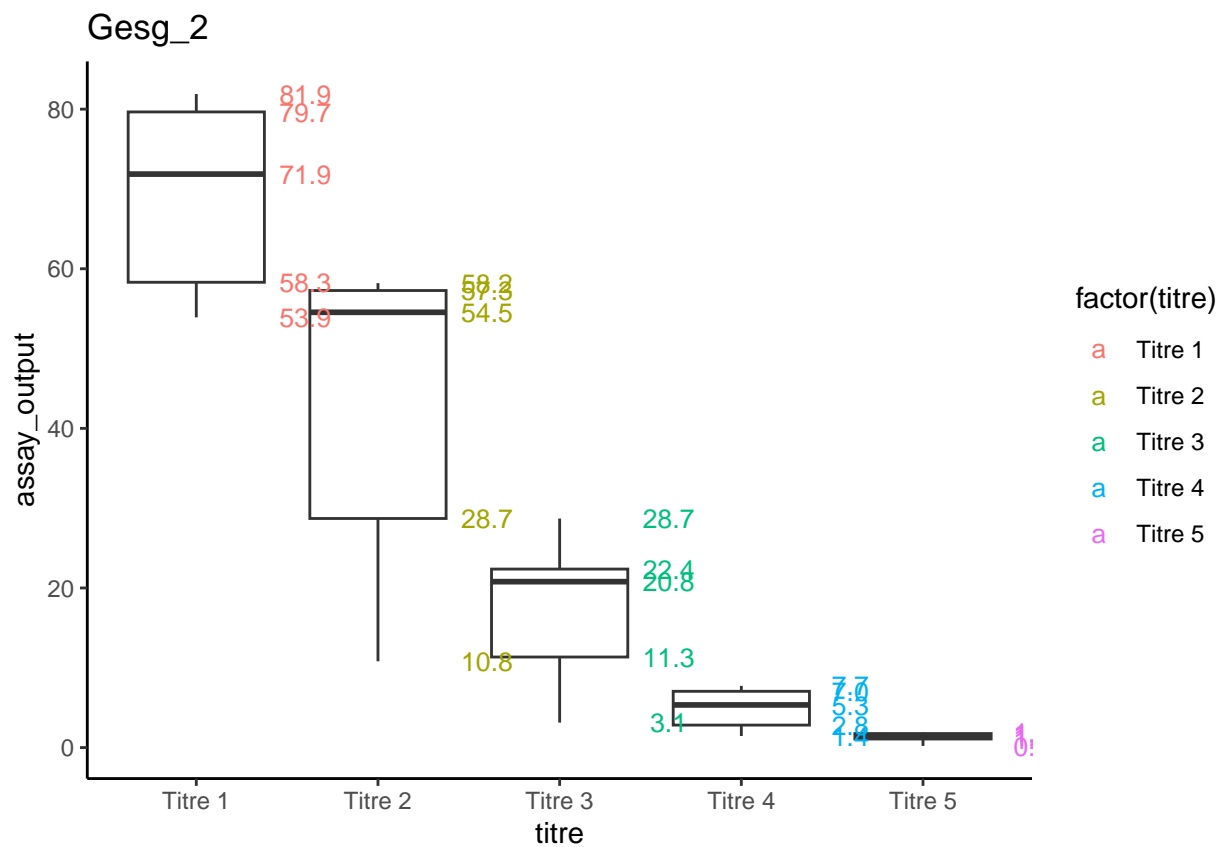
```
##
## [[32]]
```



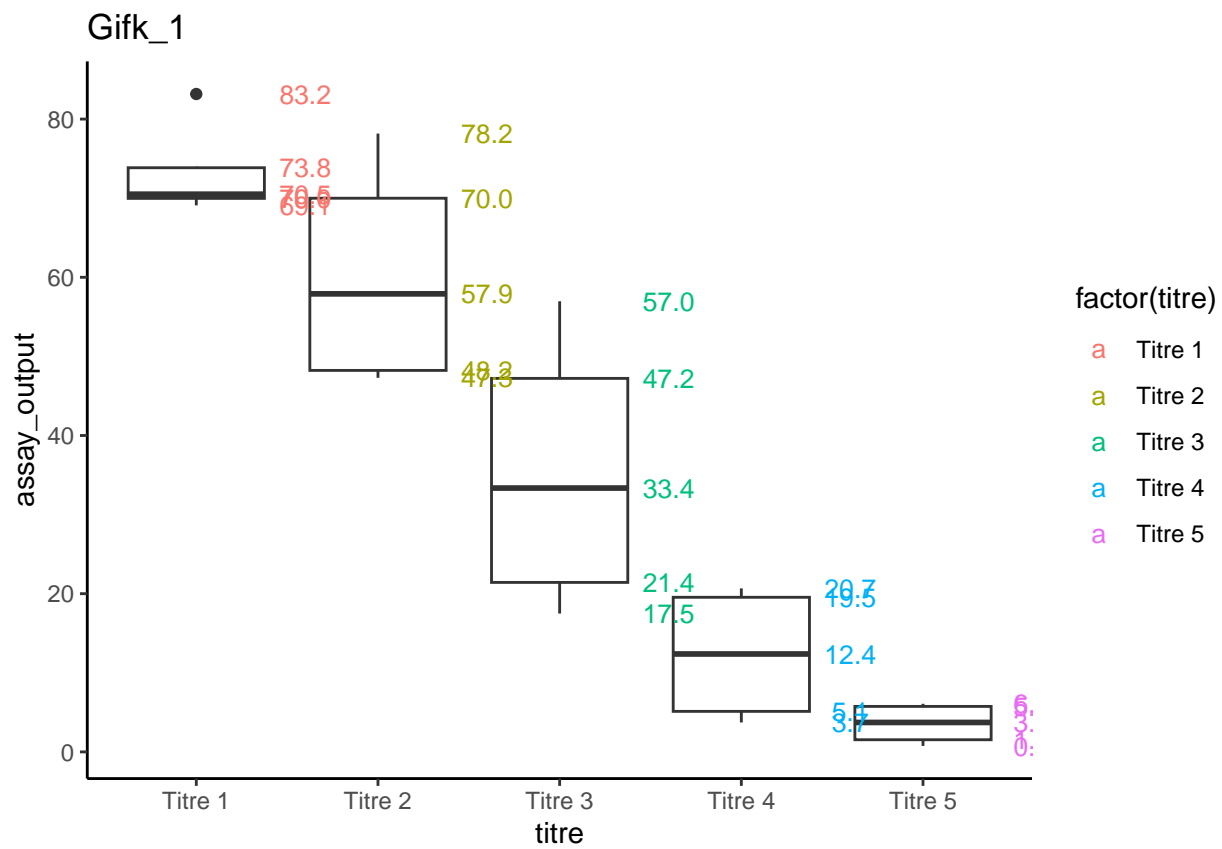
```
##
## [[33]]
```



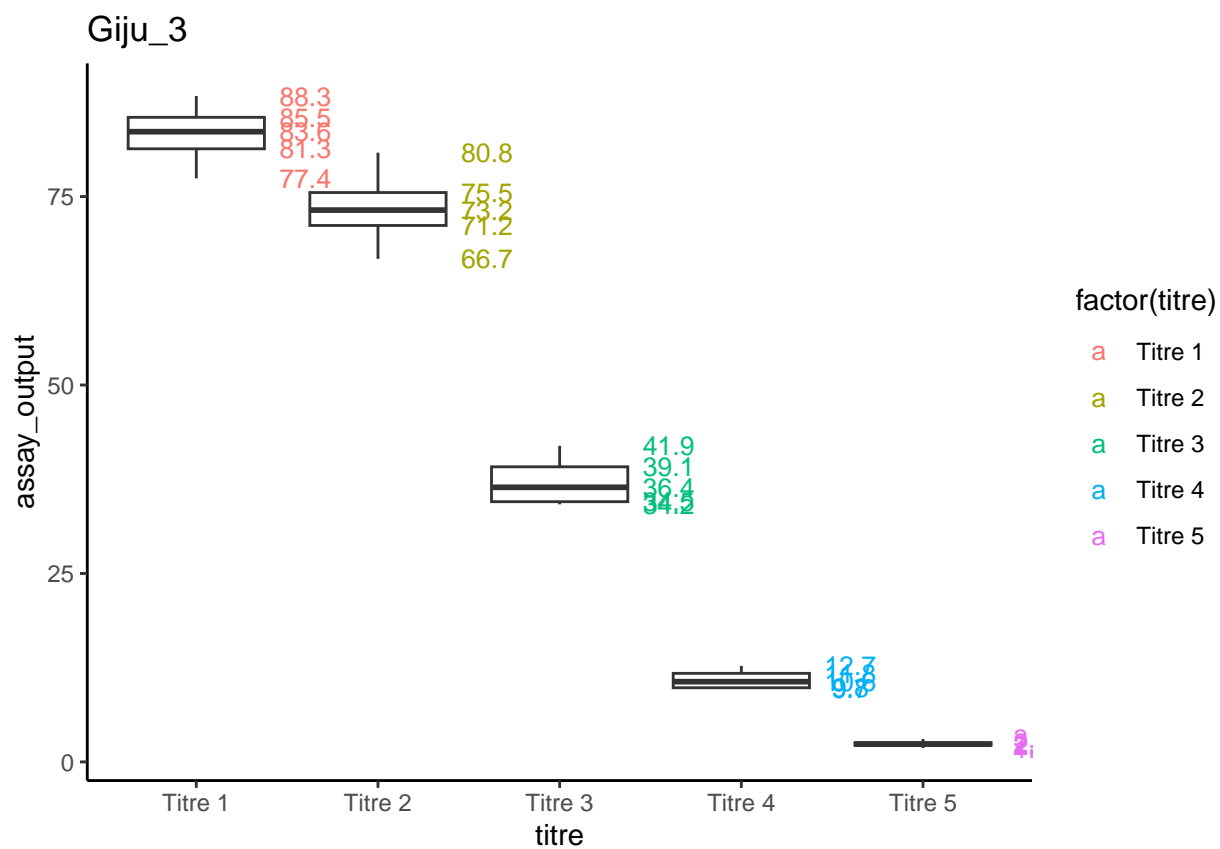
[[34]]



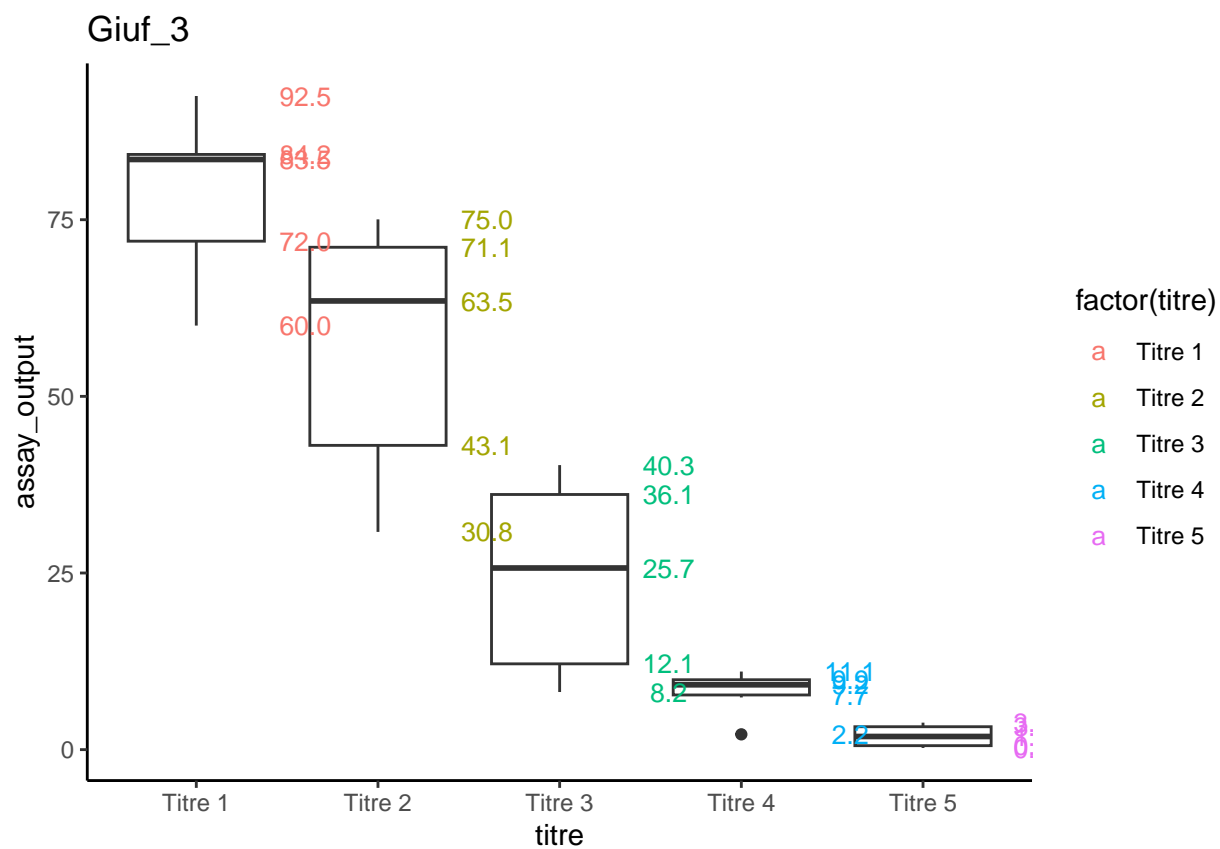
[[35]]



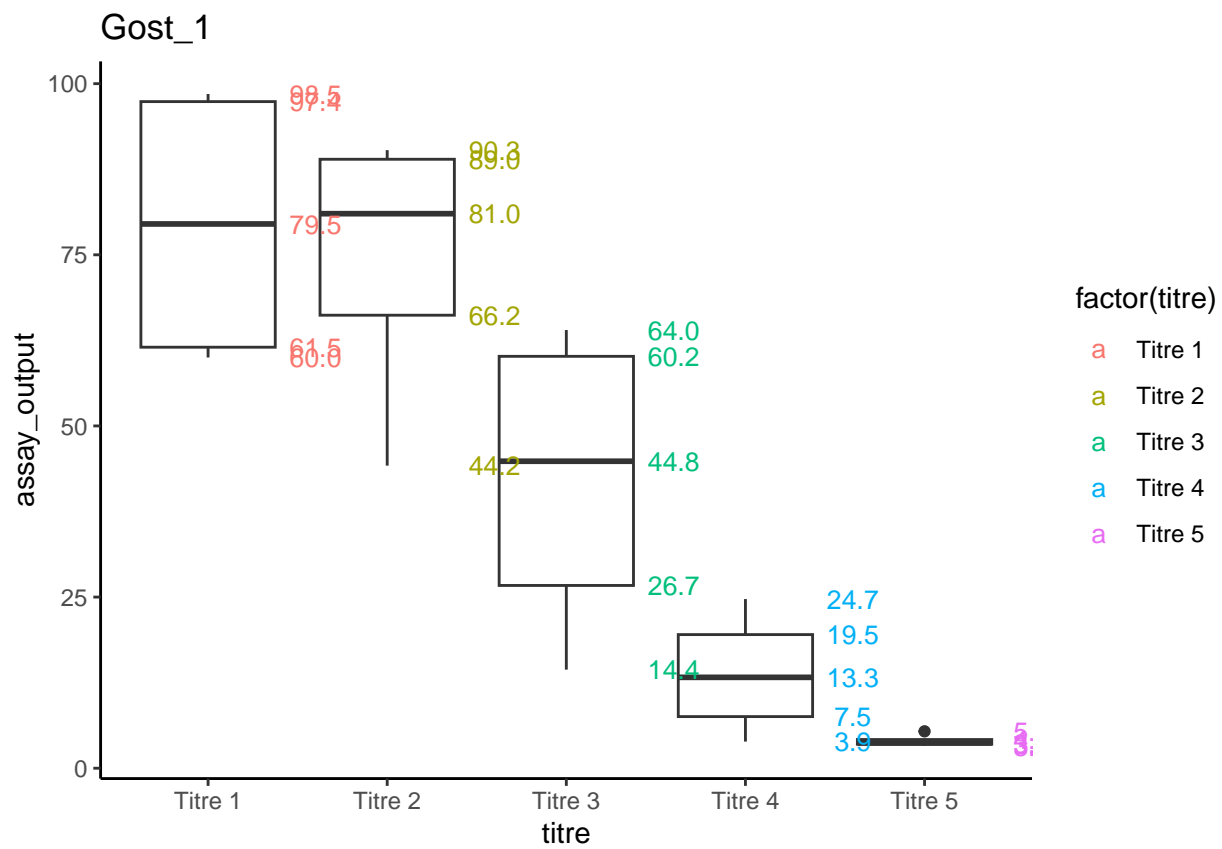
```
##
## [[36]]
```



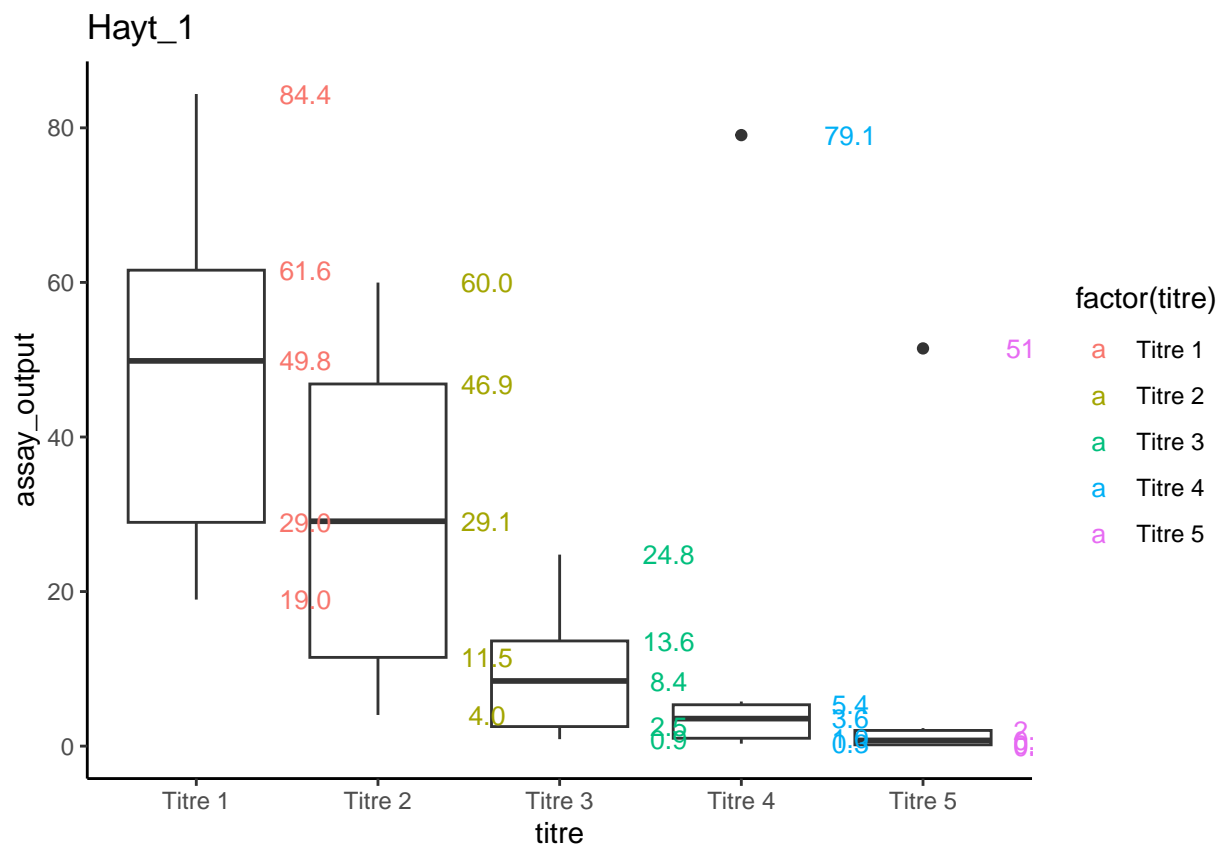
[[37]]



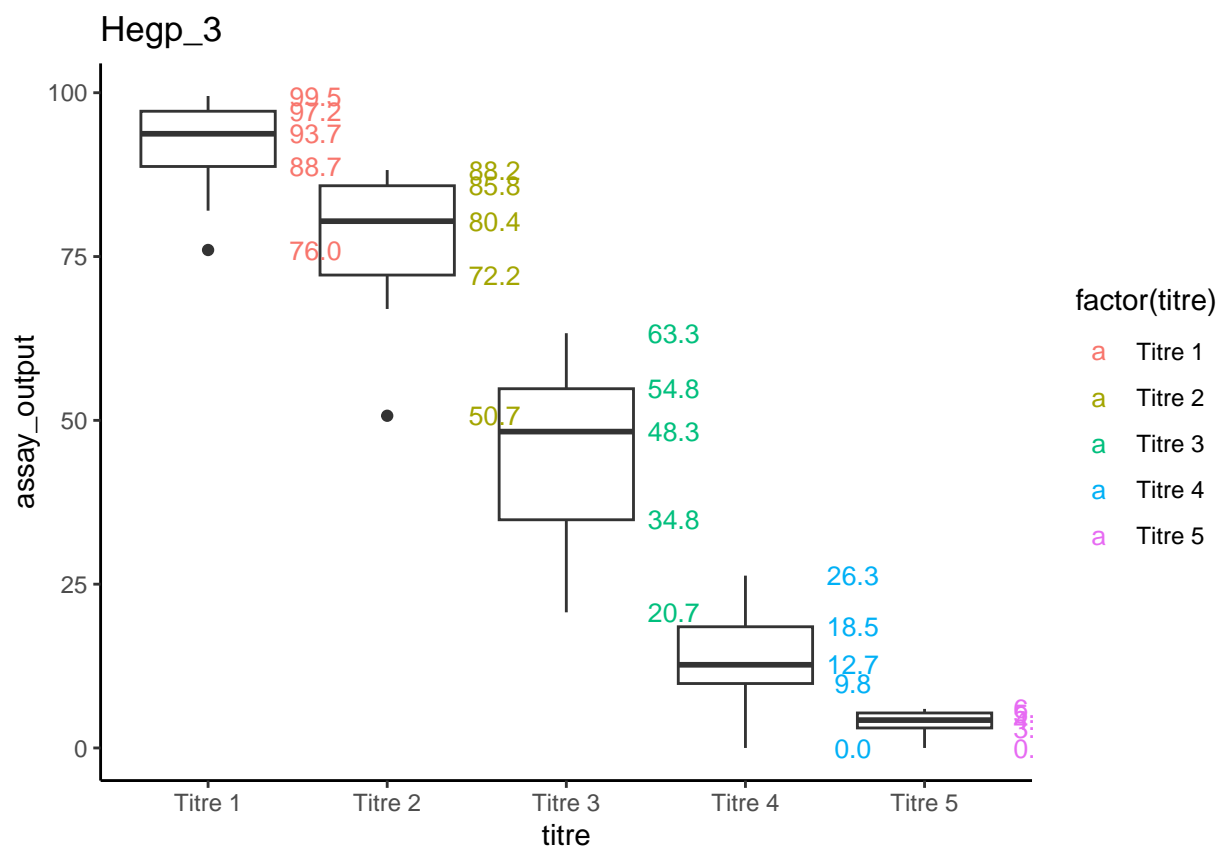
```
##
## [[38]]
```

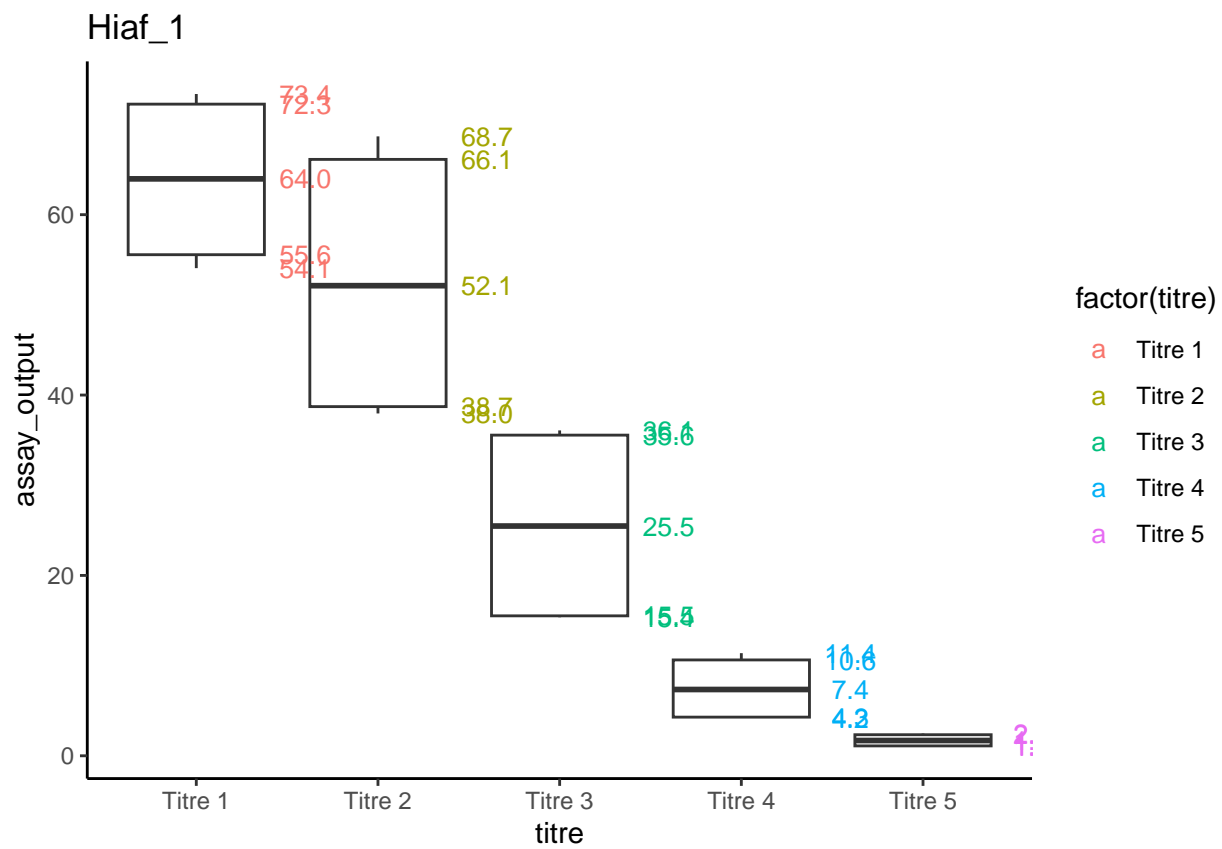
```
##
## [[39]]
```



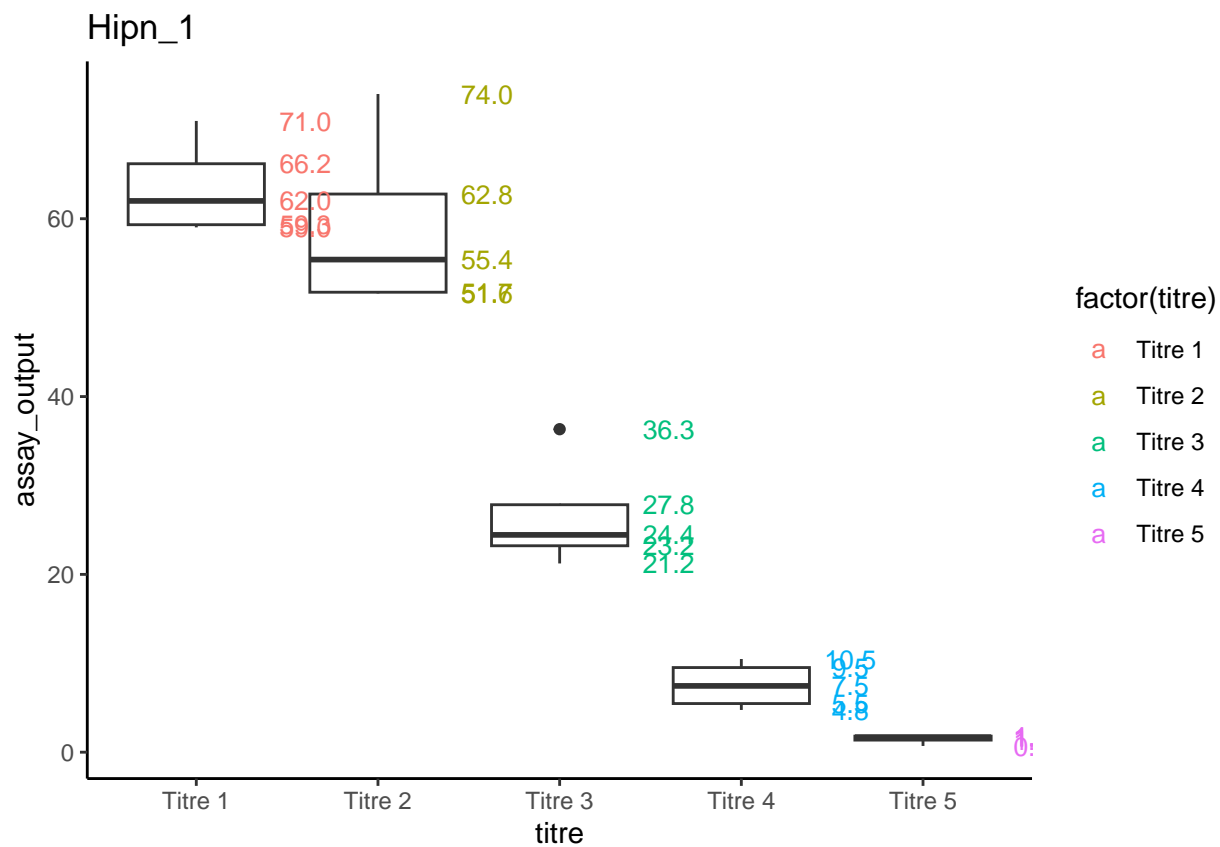
```
##
## [[40]]
```



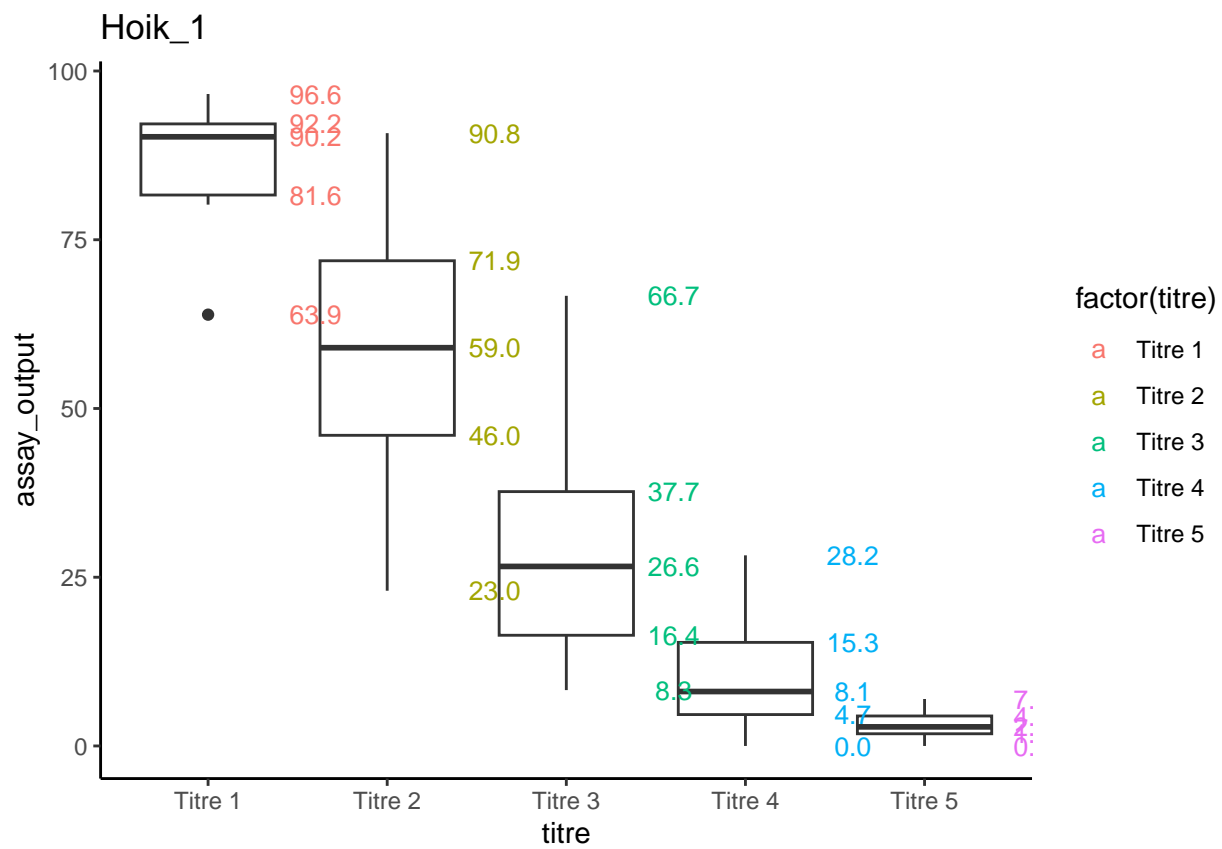
```
##
## [[41]]
```



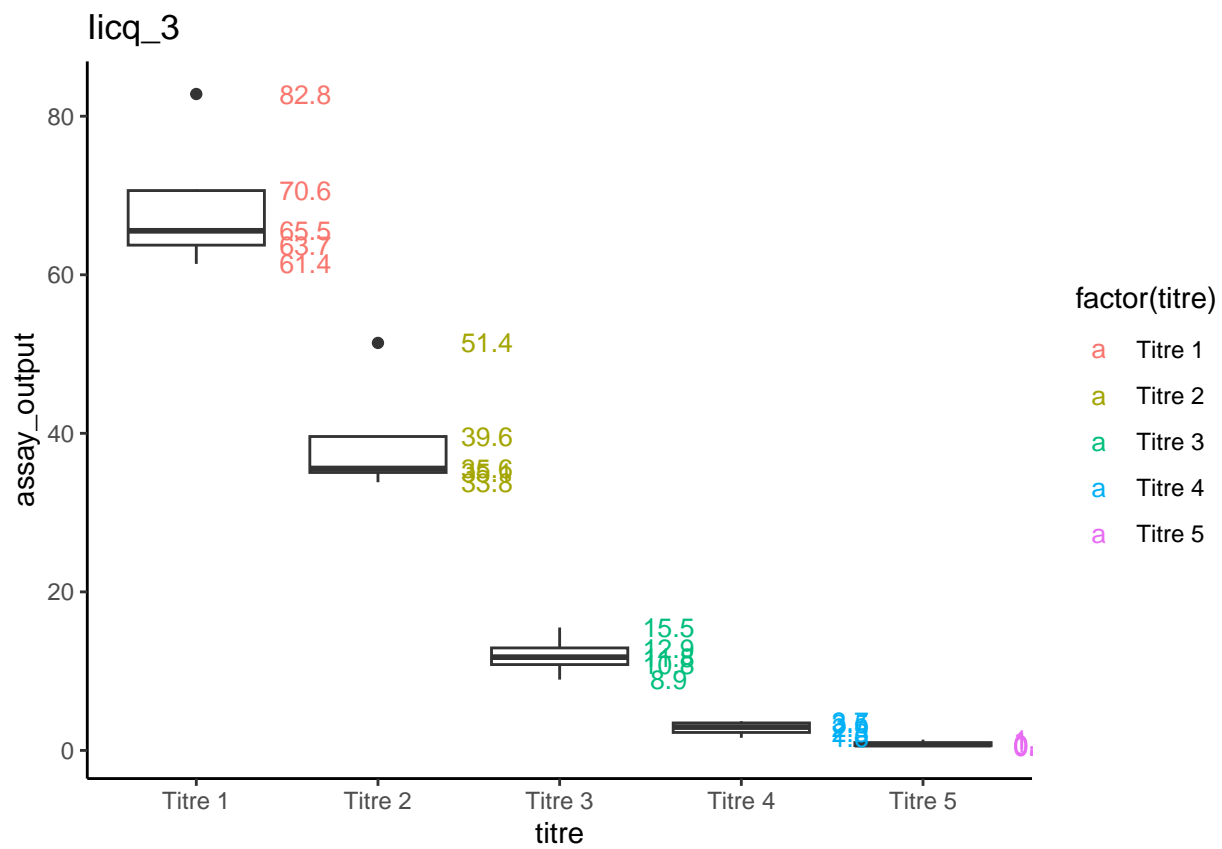
```
##
## [[42]]
```



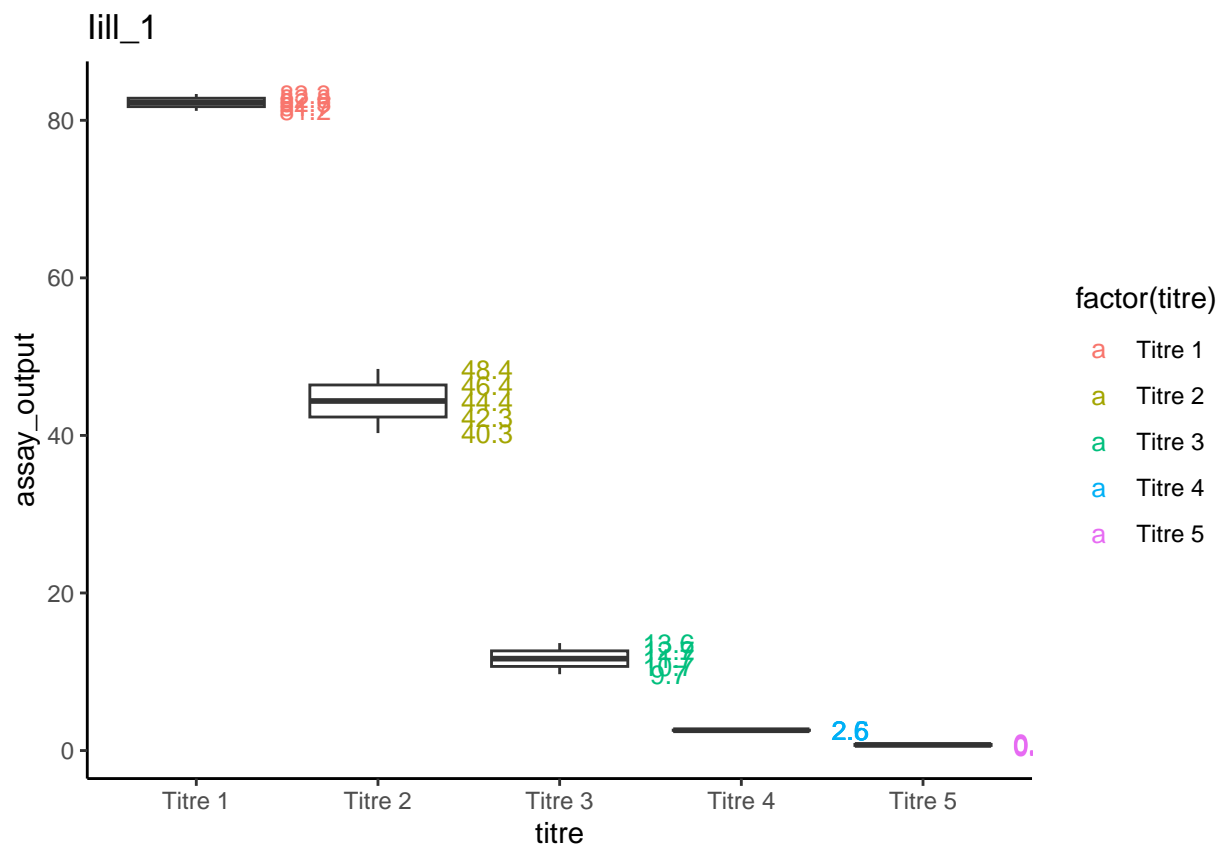
[[43]]



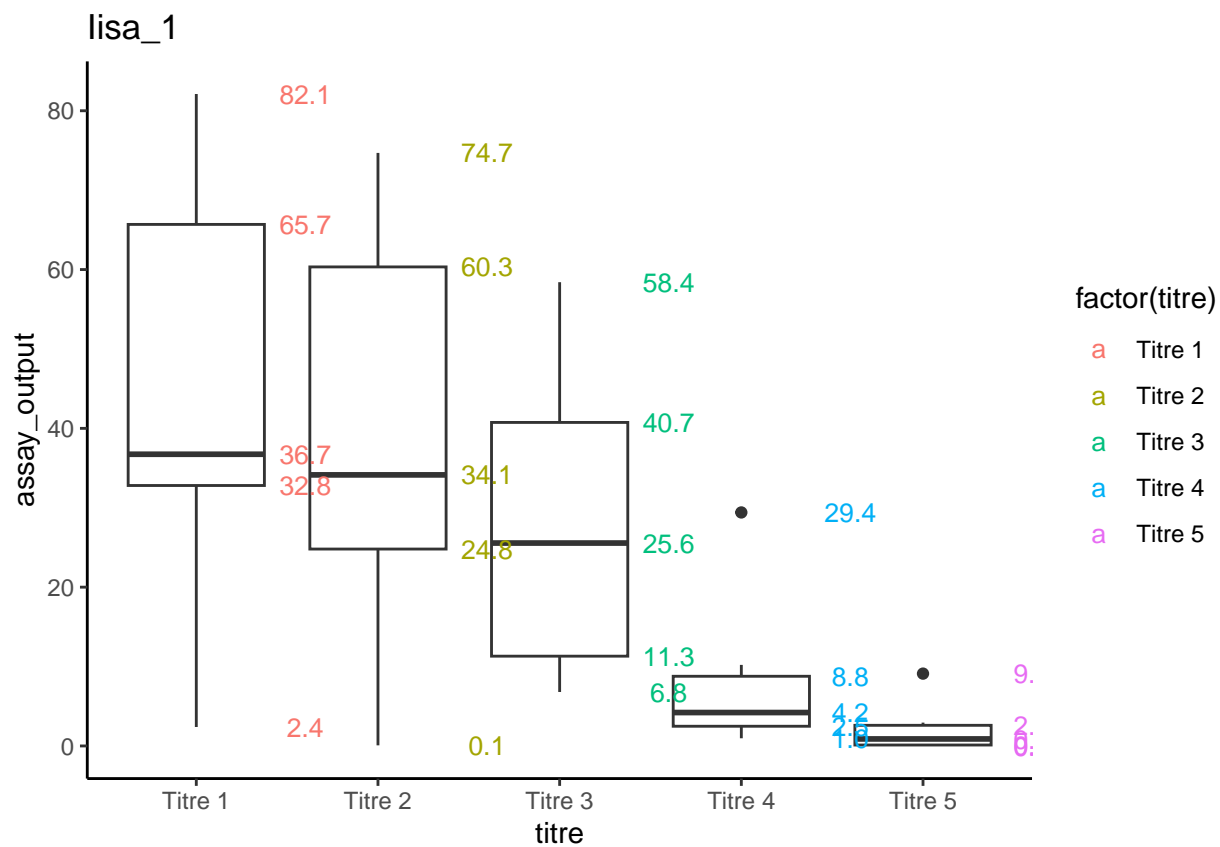
```
##
## [[44]]
```



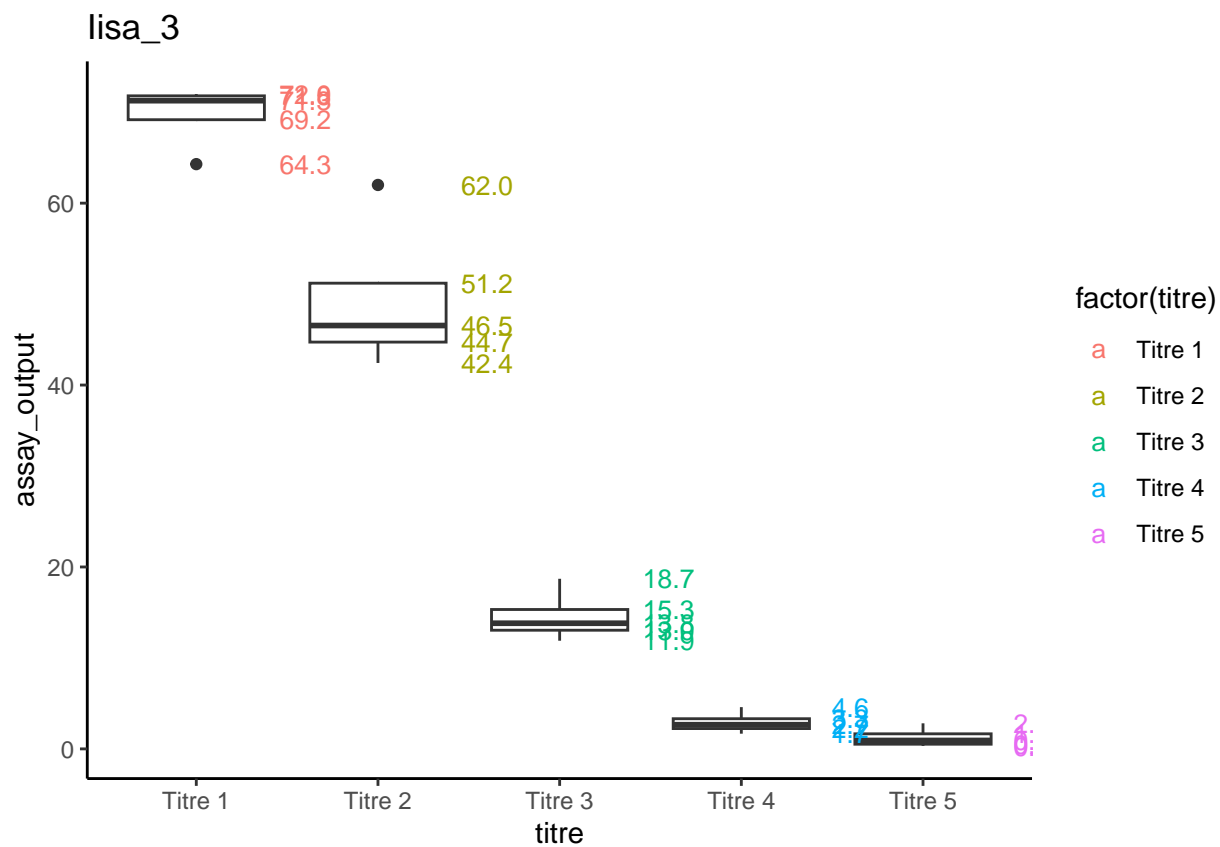
[[45]]



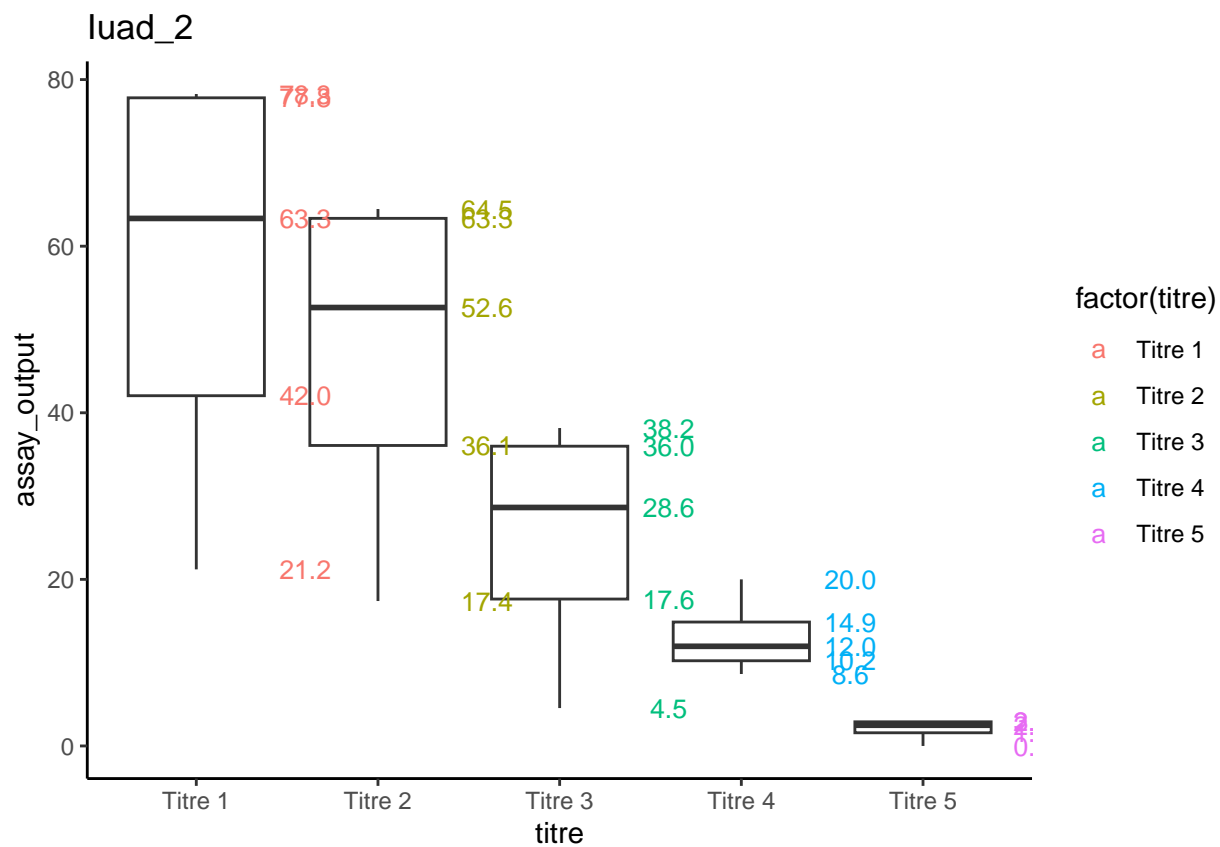
[[46]]



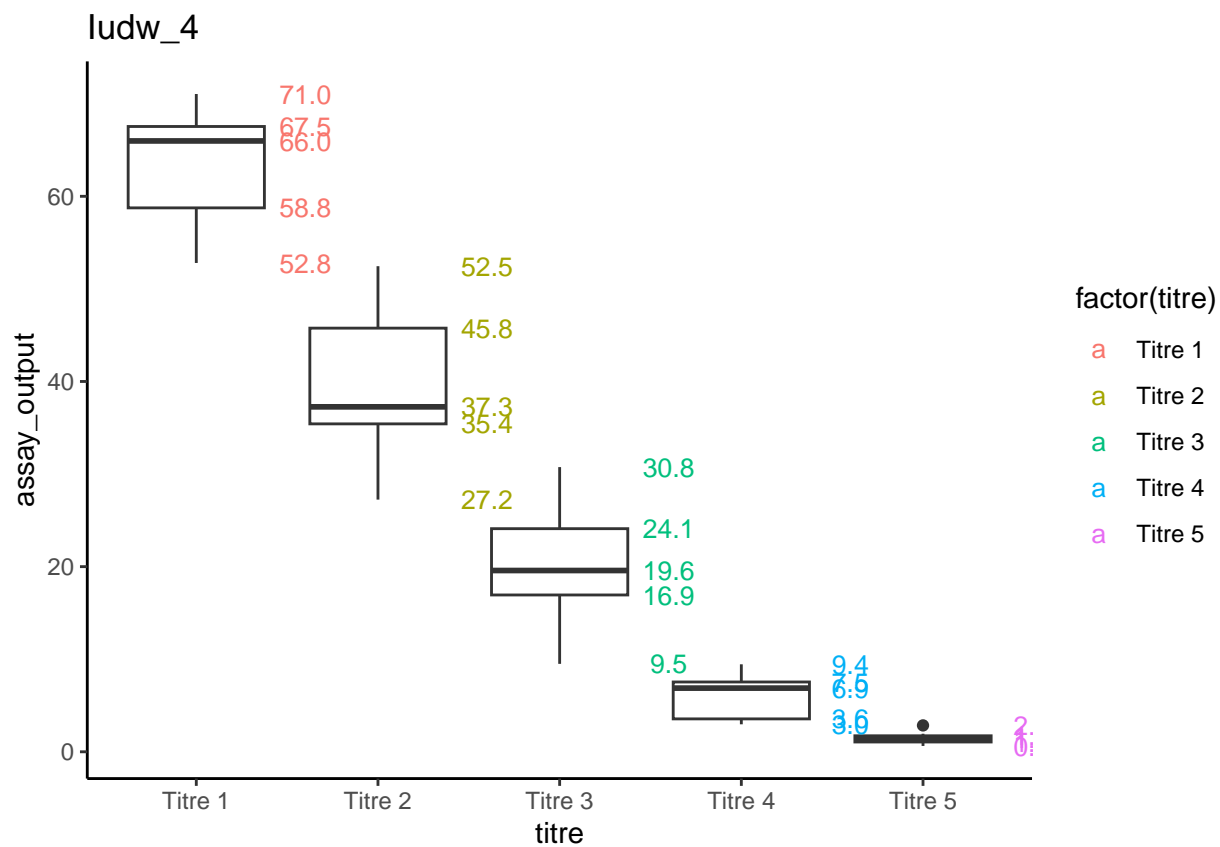
[[47]]



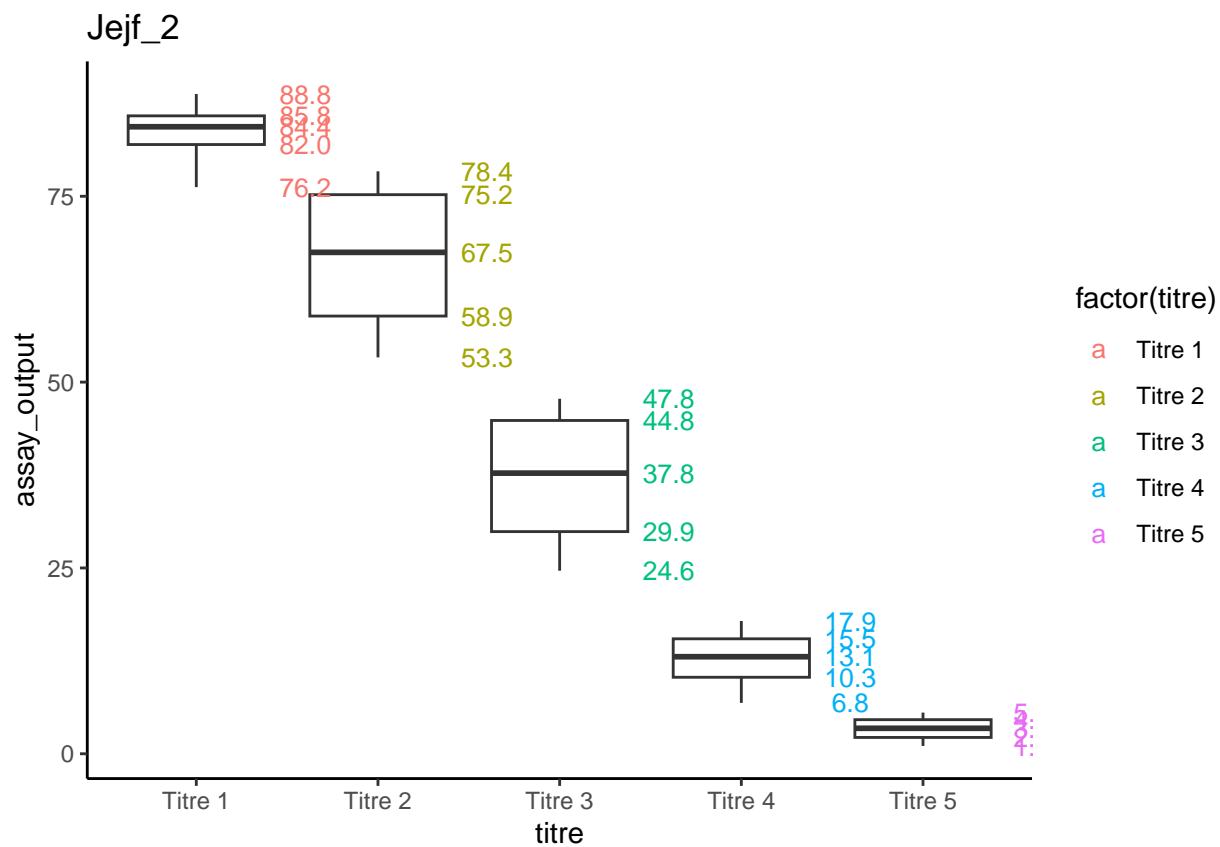
```
##
## [[48]]
```



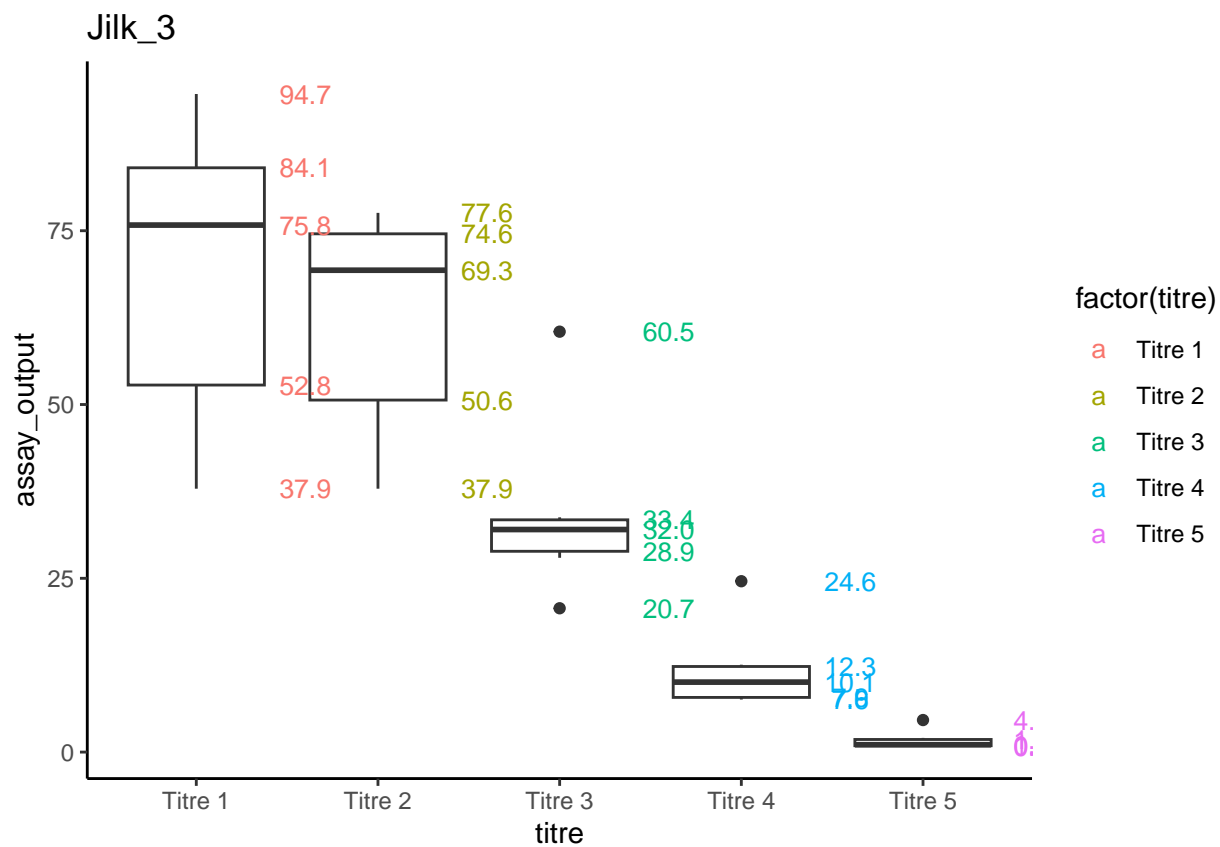
[[49]]



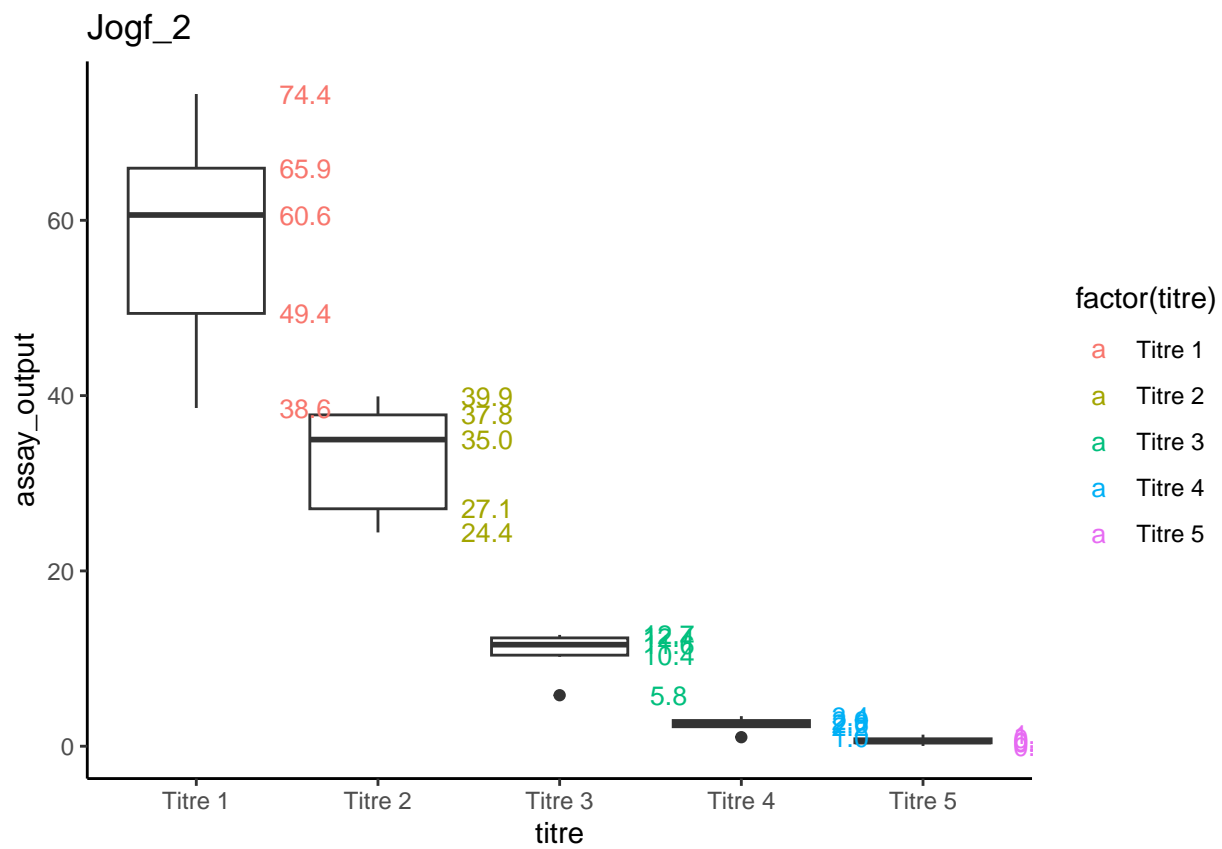
```
##
## [[50]]
```



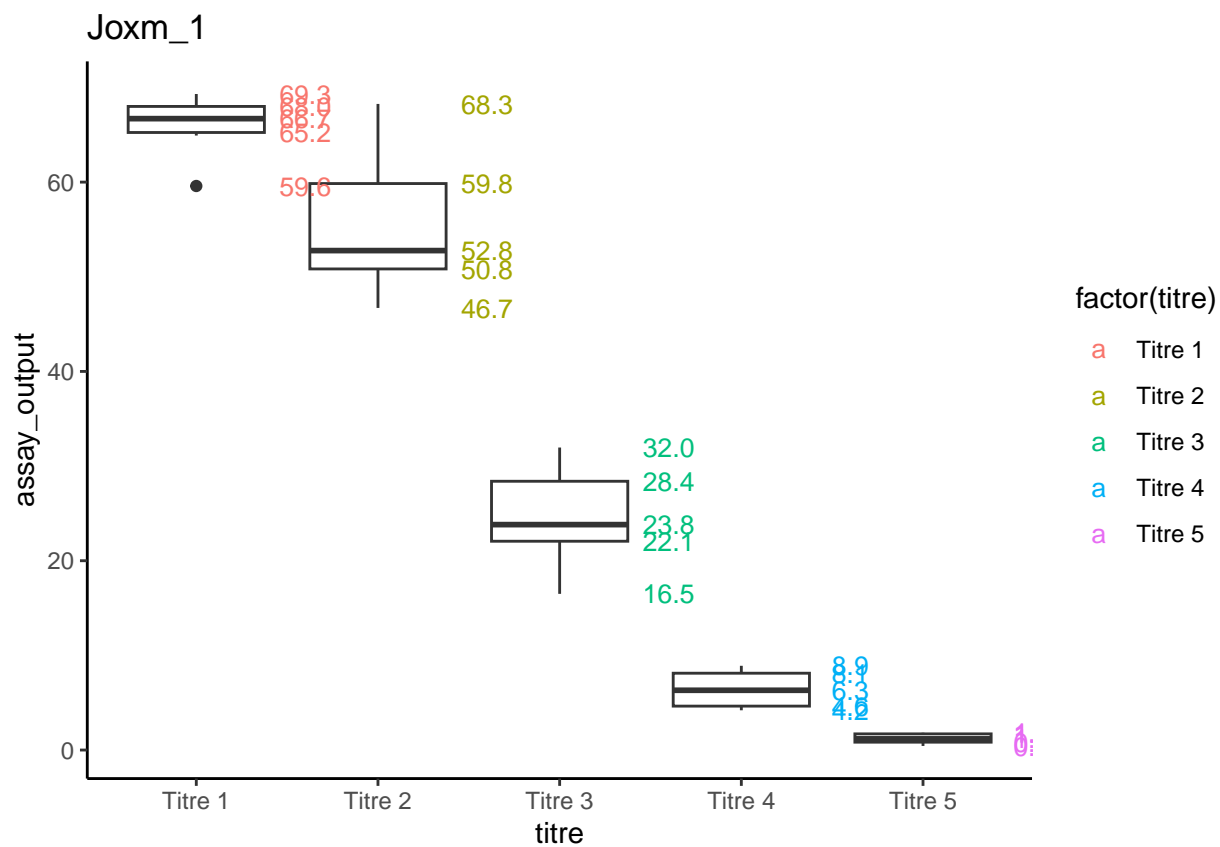
```
##
## [[51]]
```



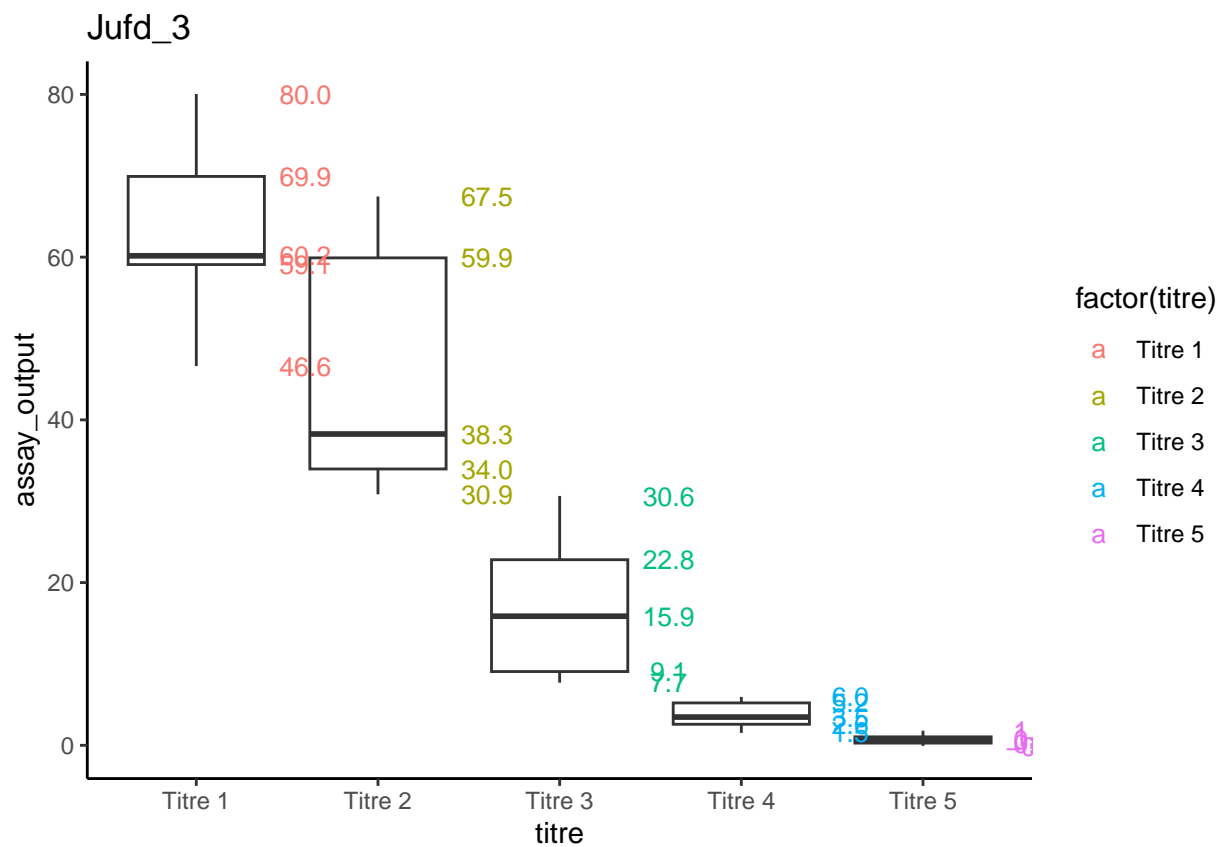
```
##
## [[52]]
```



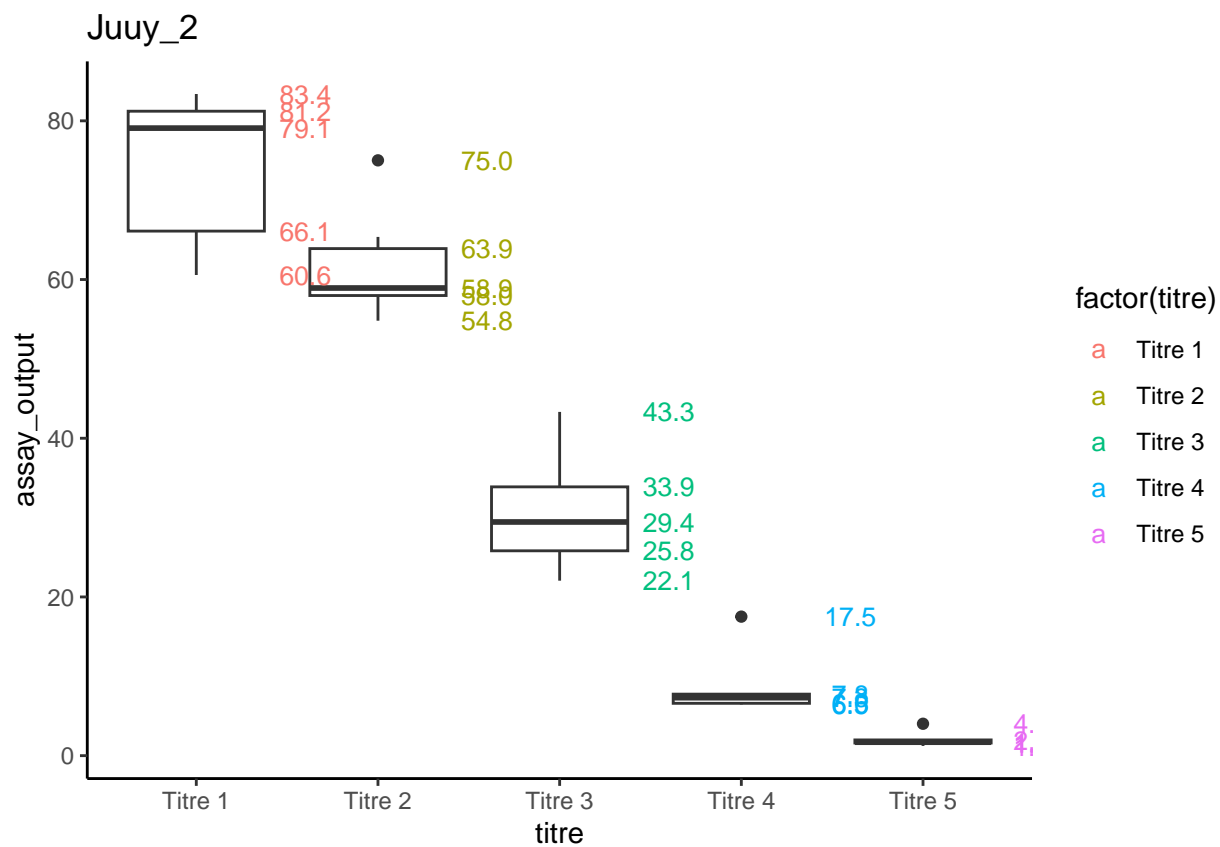
```
##
## [[53]]
```



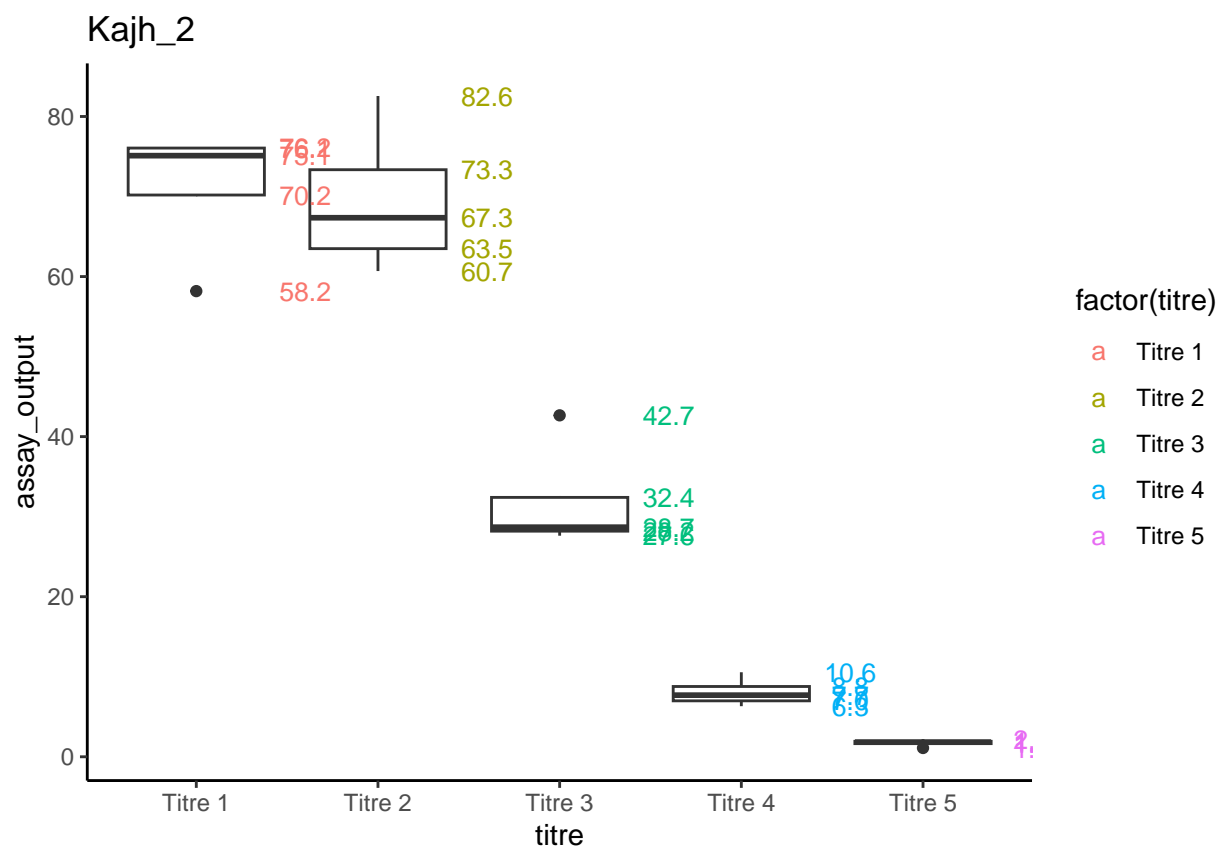
```
##
## [[54]]
```

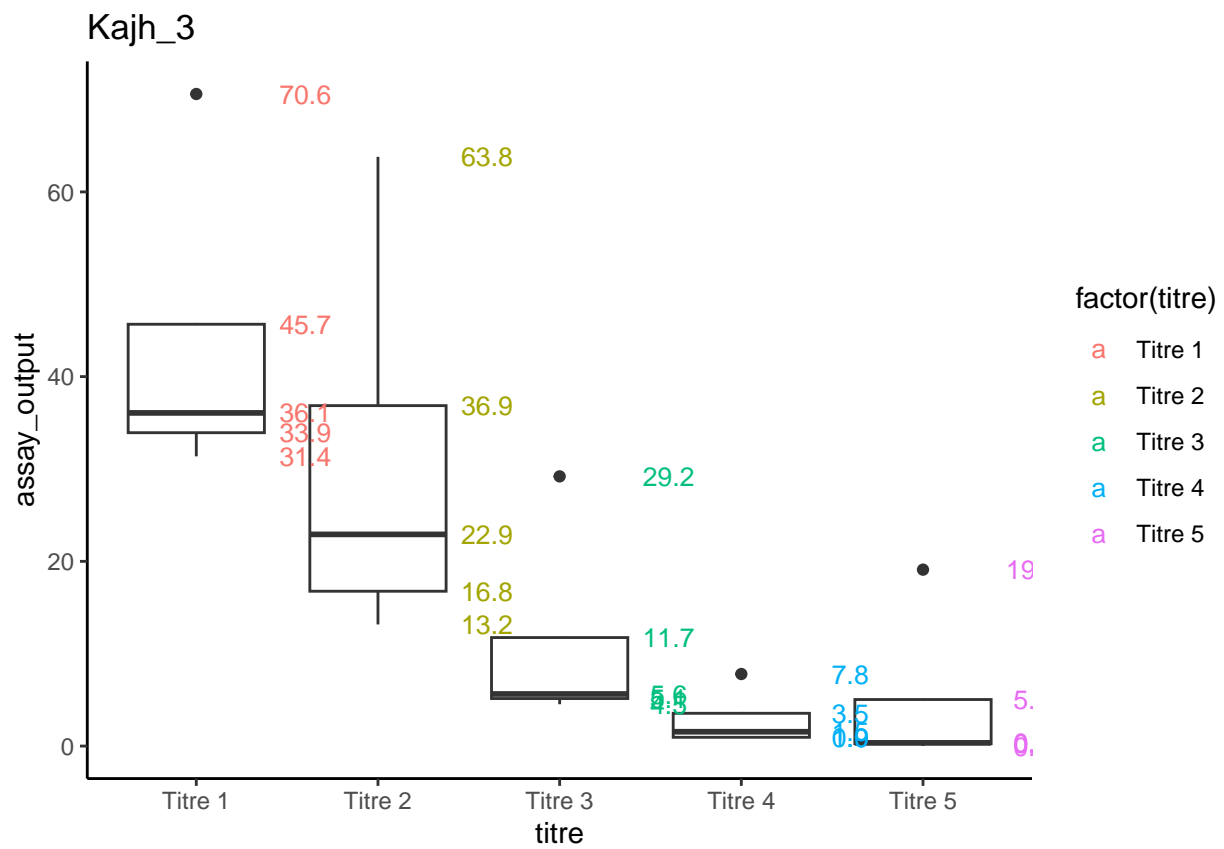
[[55]]



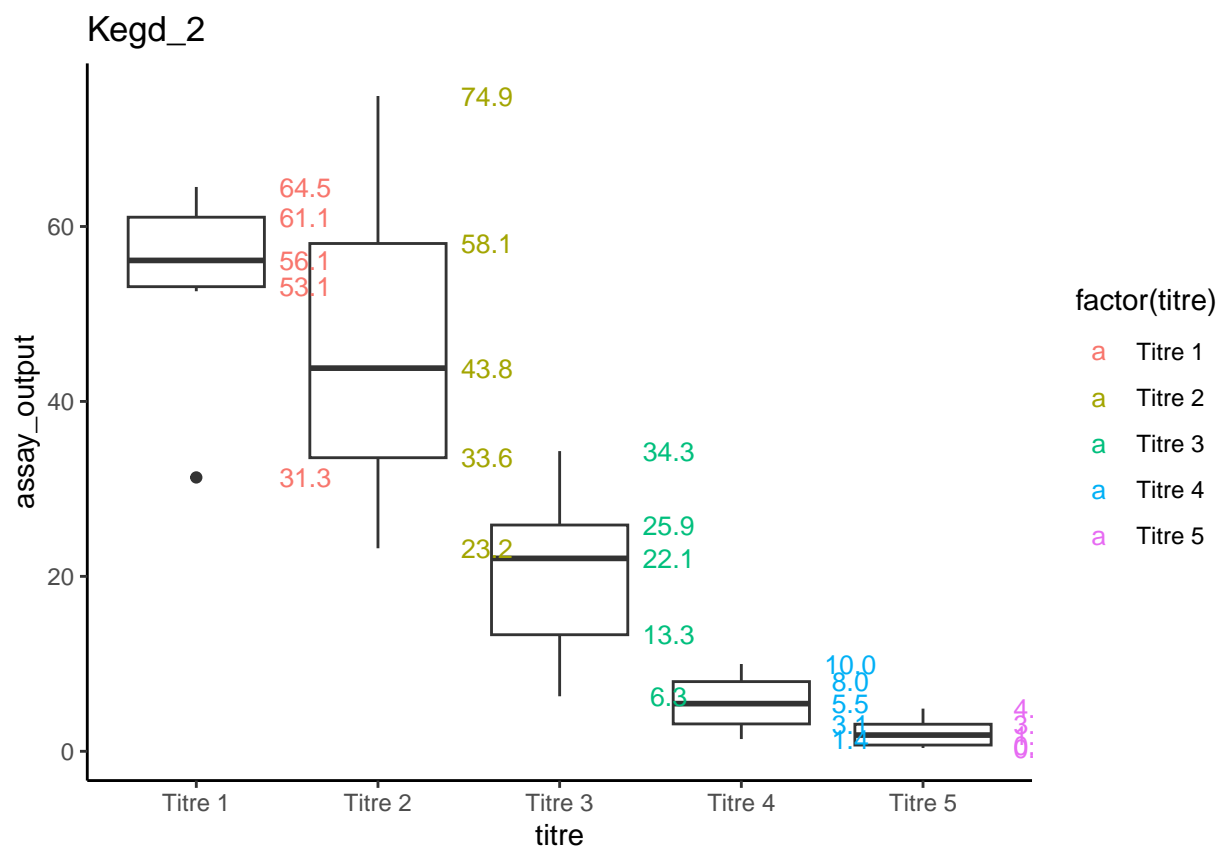
```
##
## [[56]]
```



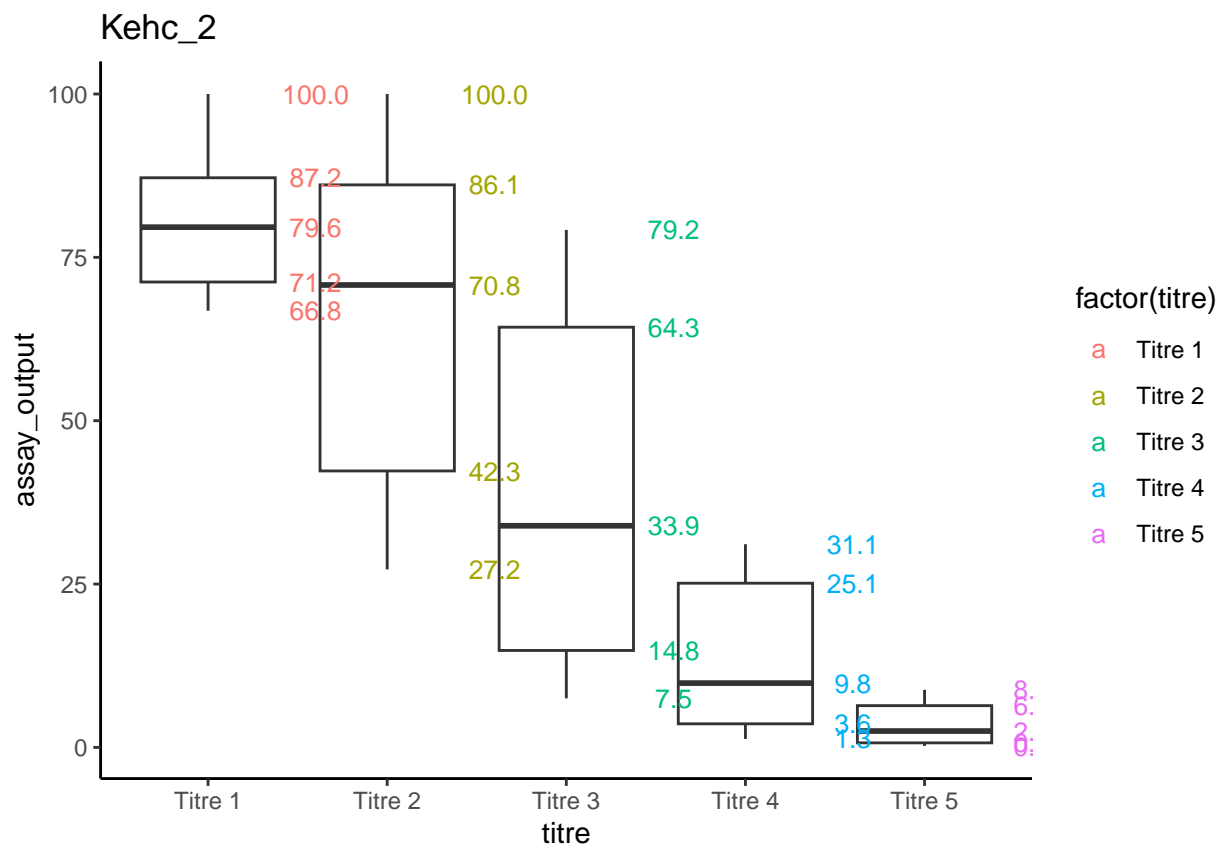
[[57]]



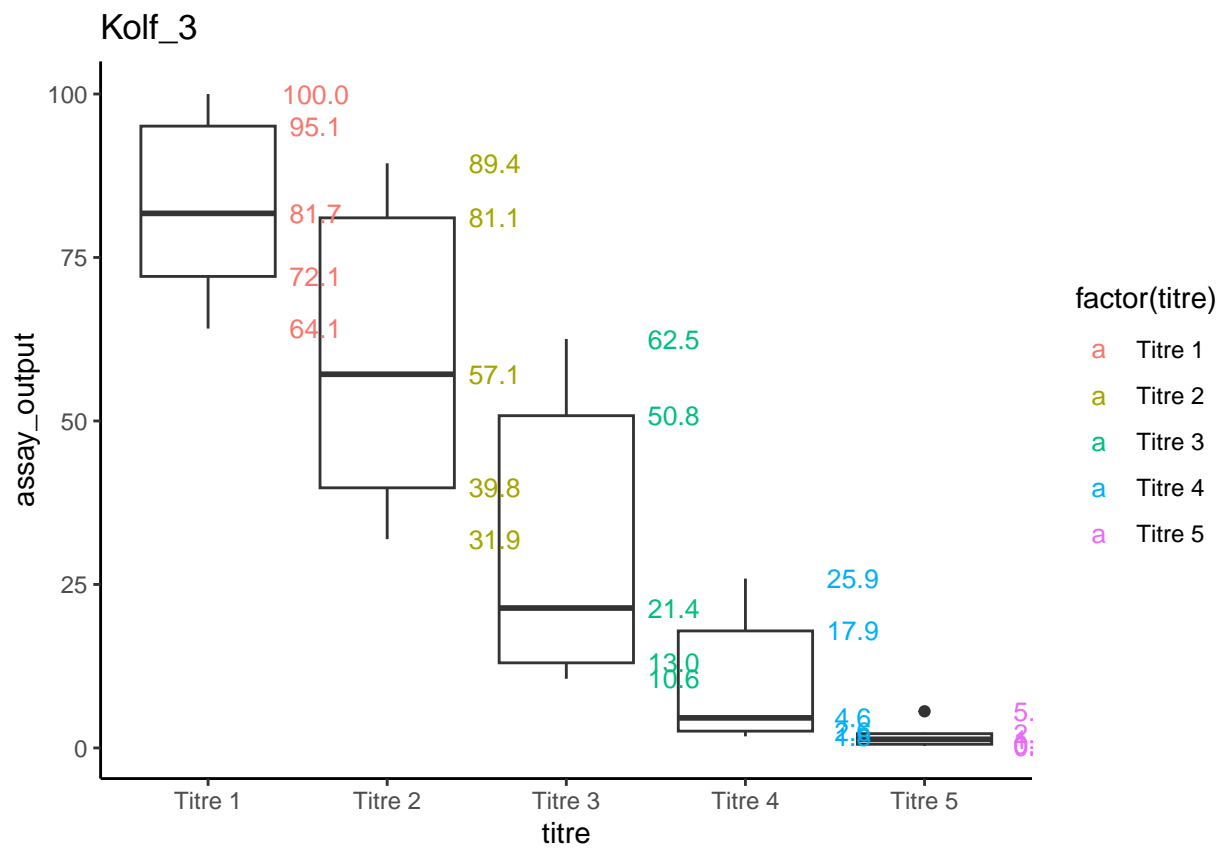
```
##
## [[58]]
```



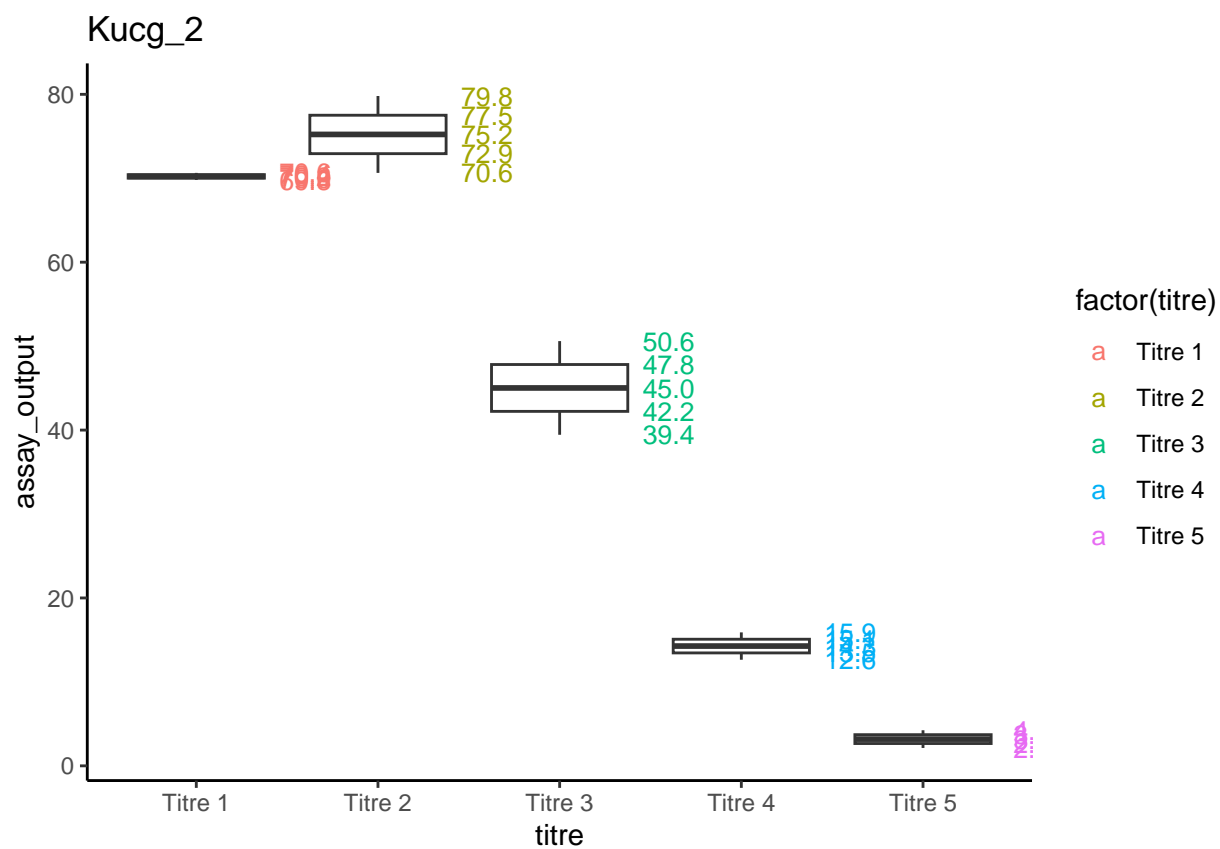
[[59]]



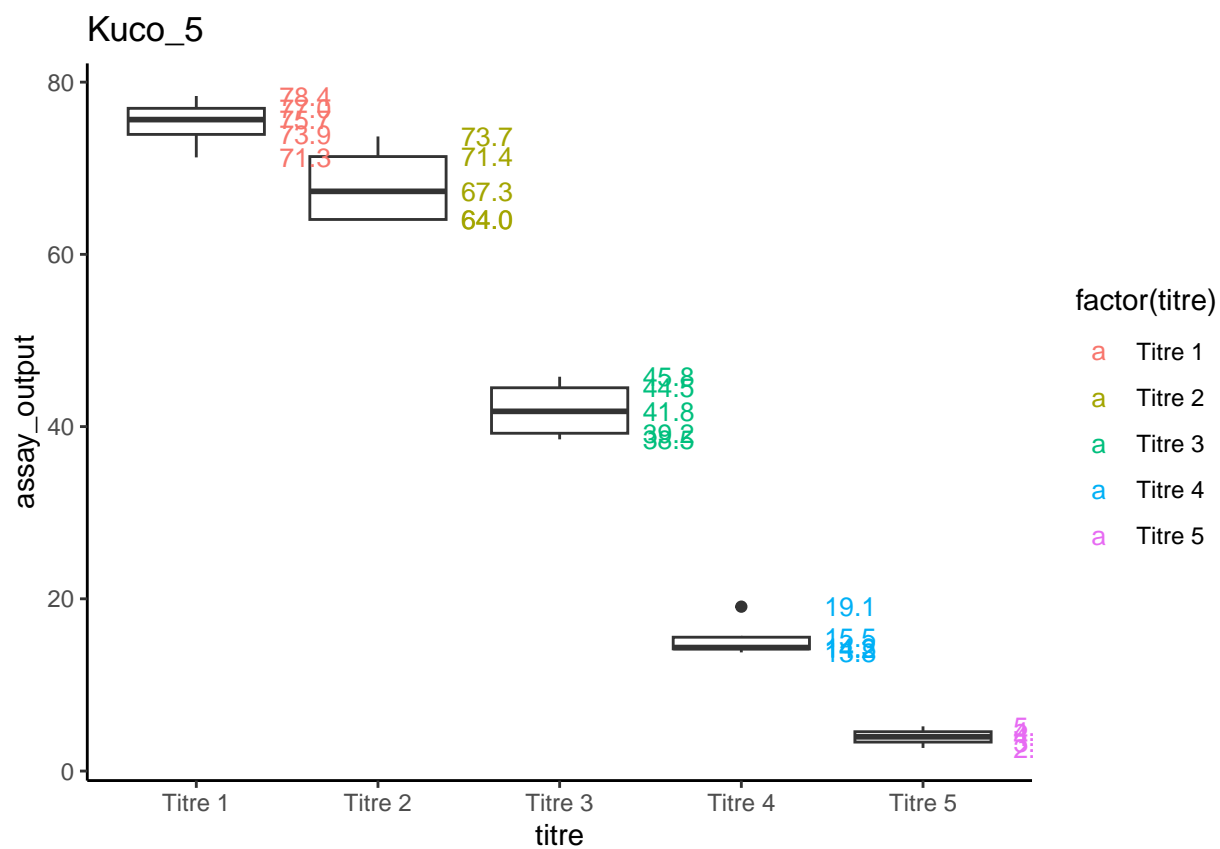
```
##
## [[60]]
```



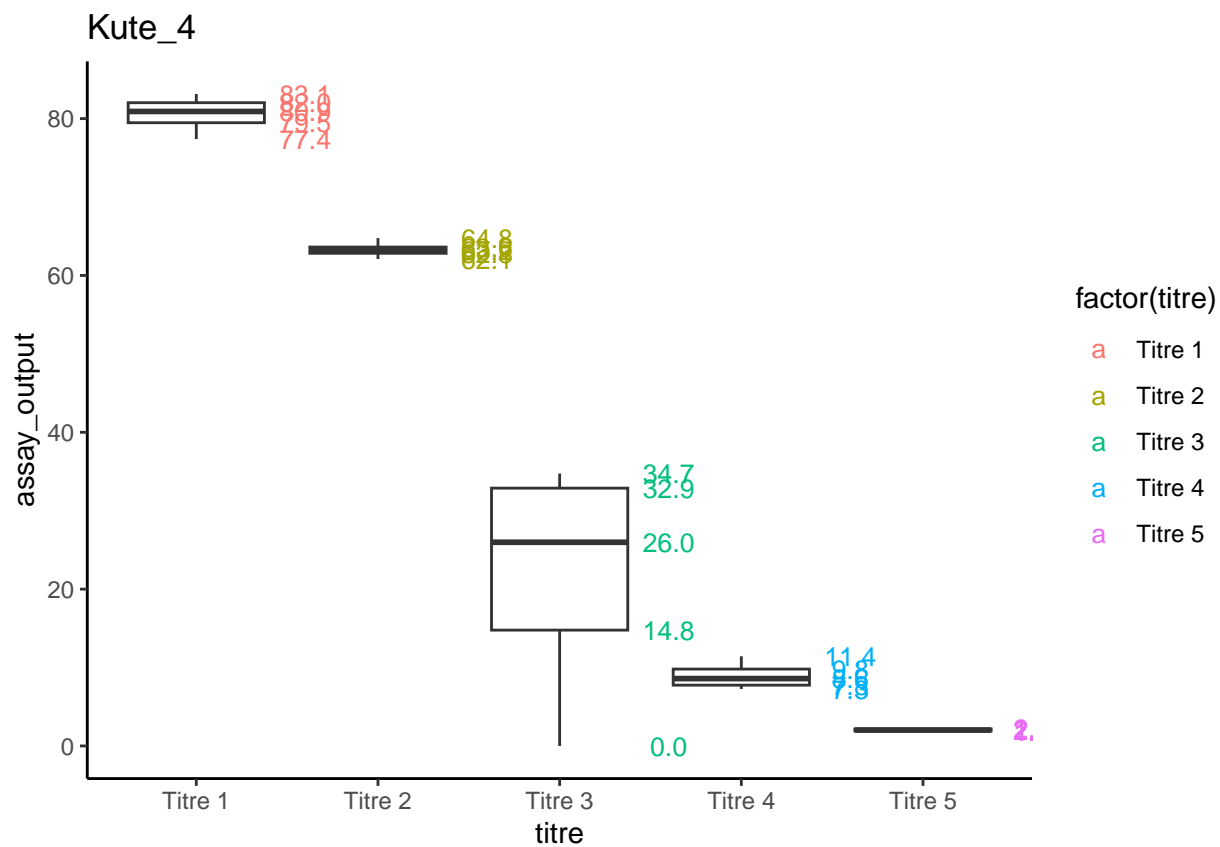
```
##
## [[61]]
```



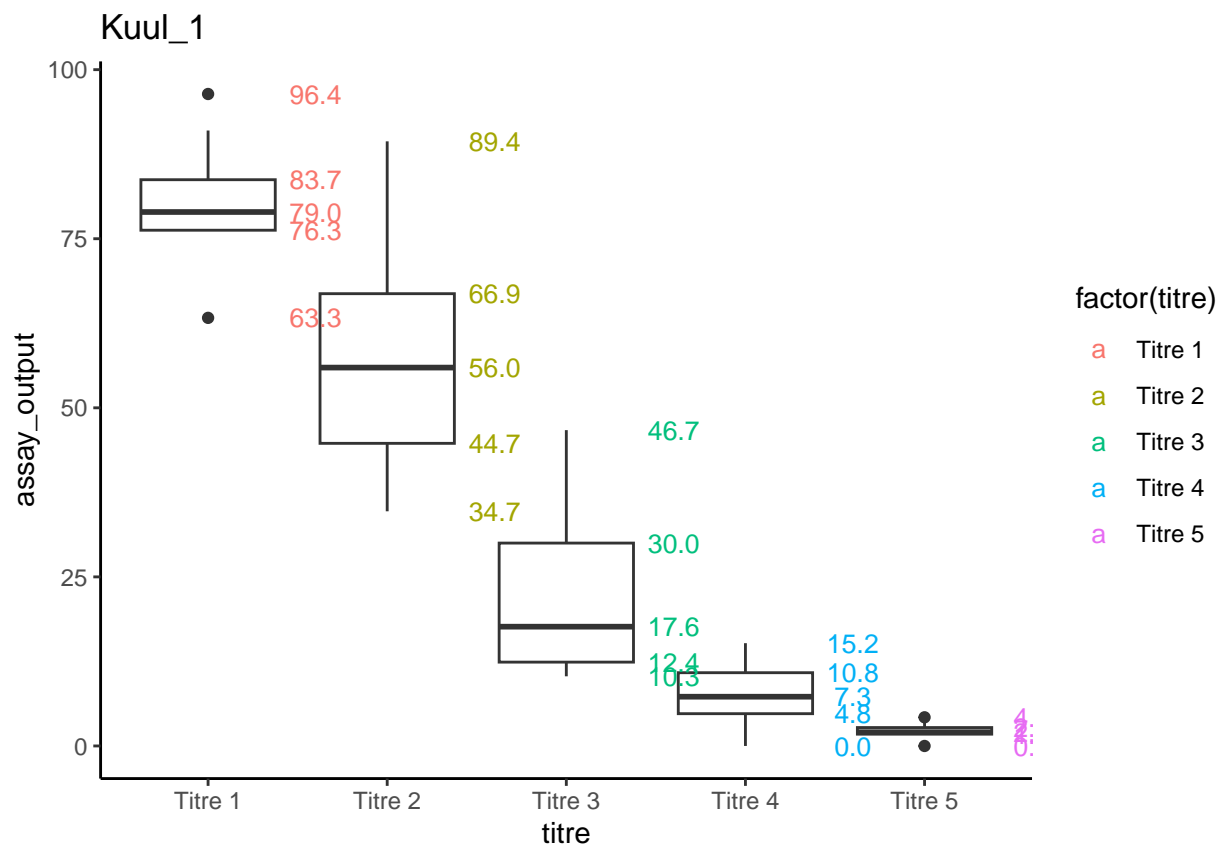
[[62]]



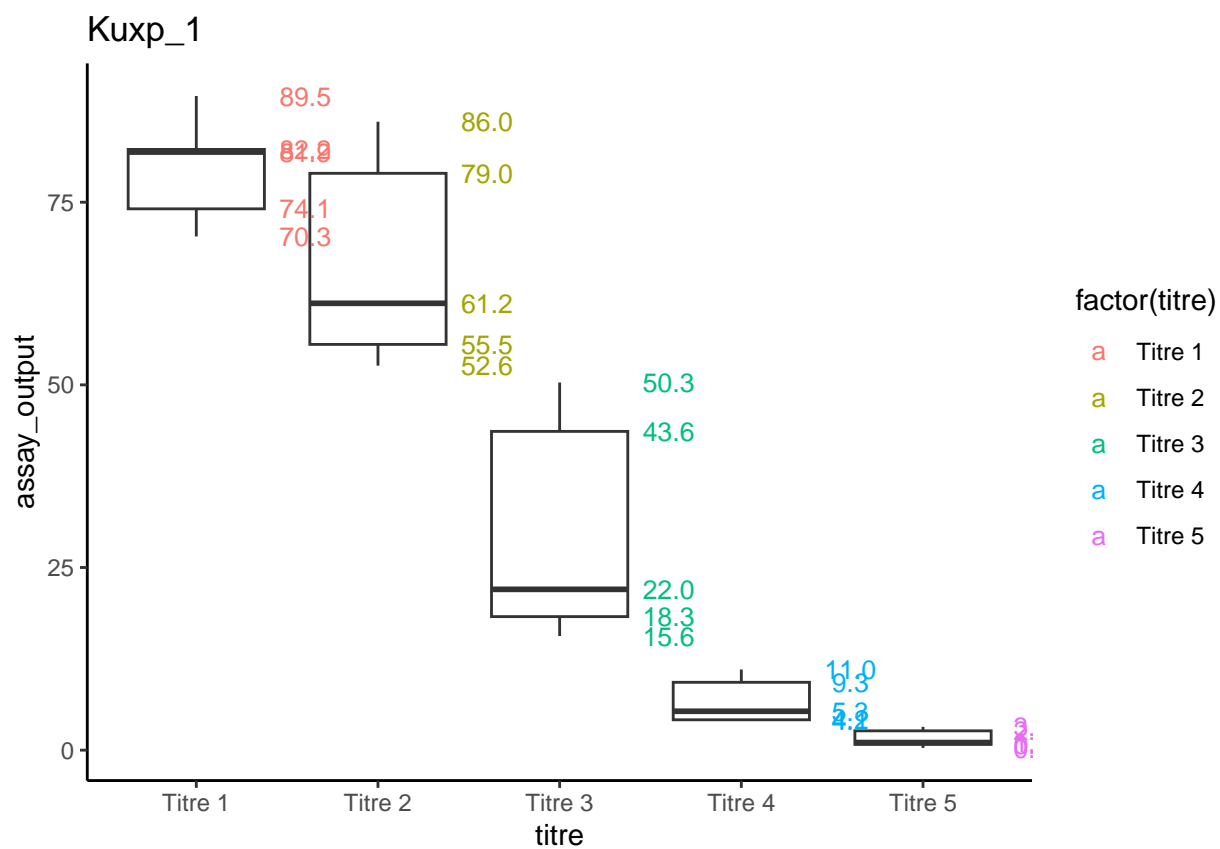
[[63]]



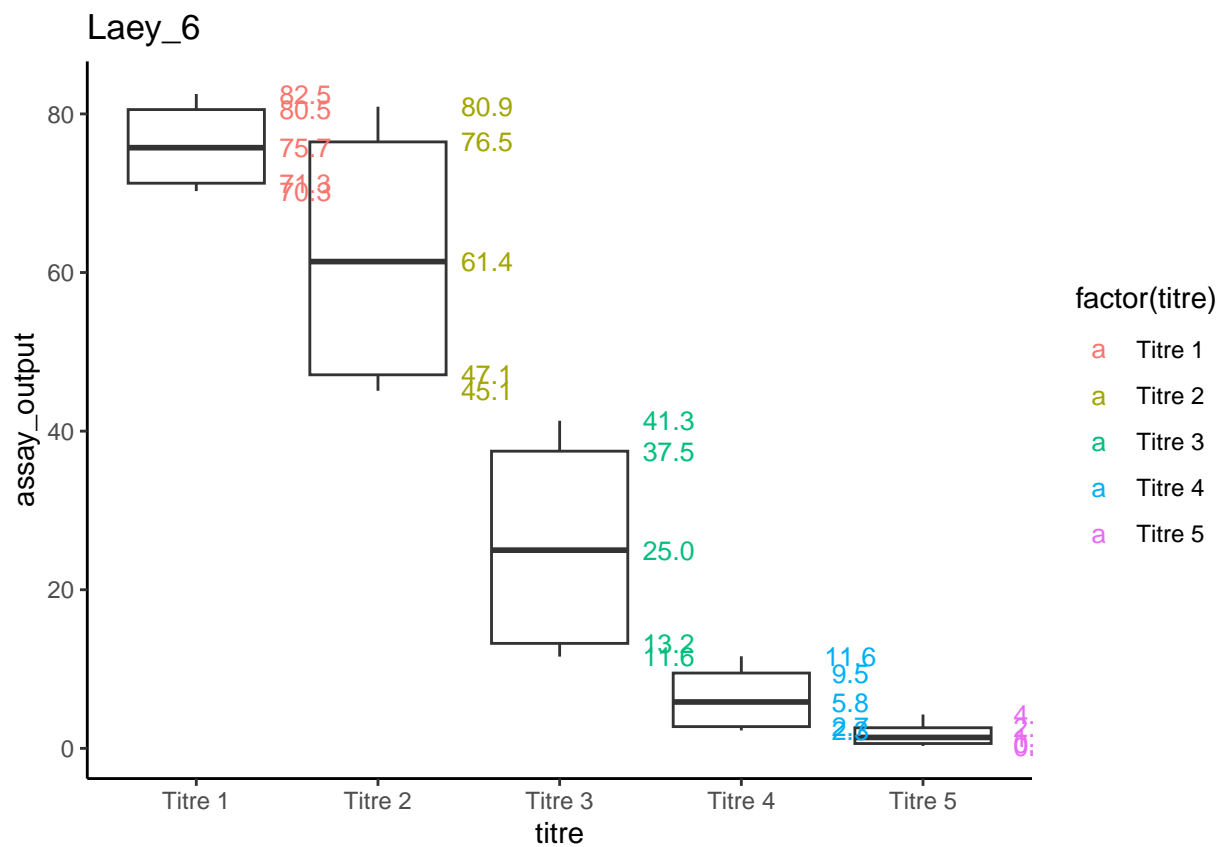
[[64]]



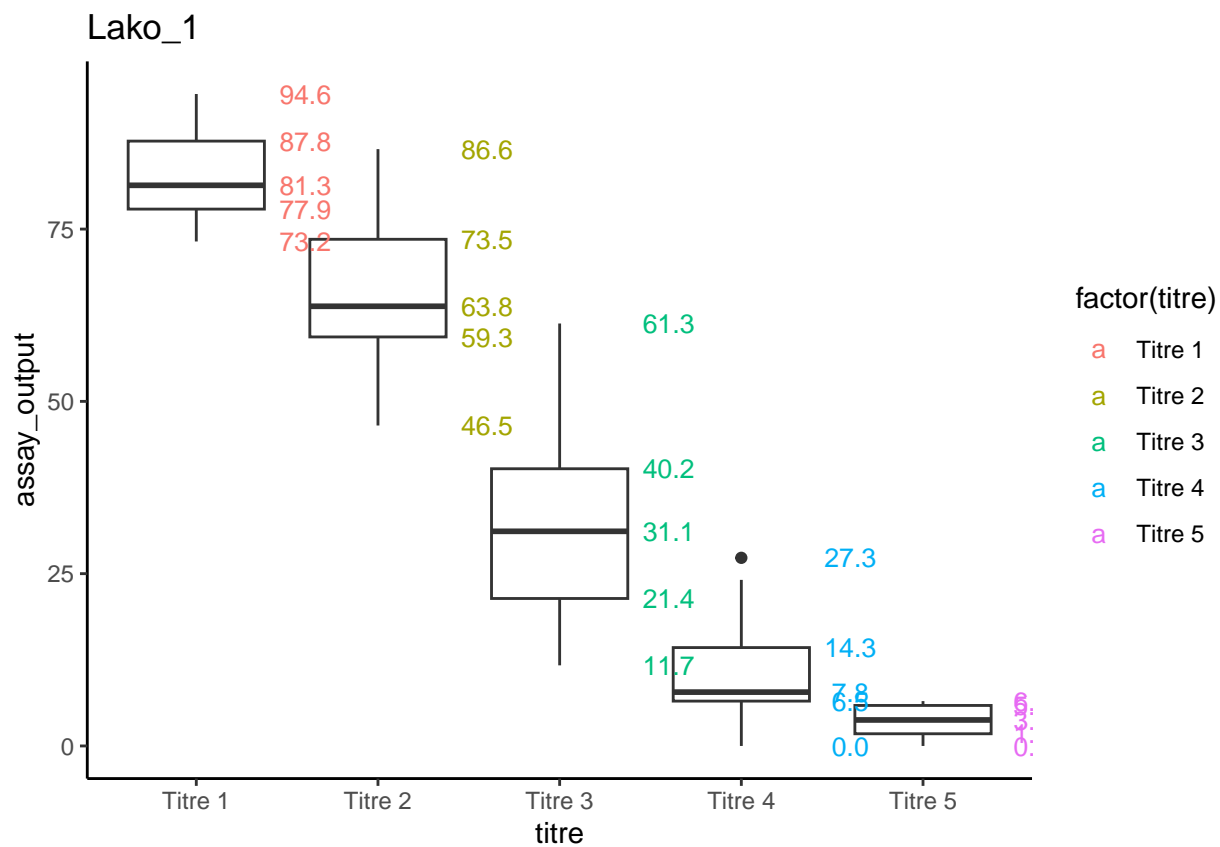
```
##
## [[65]]
```



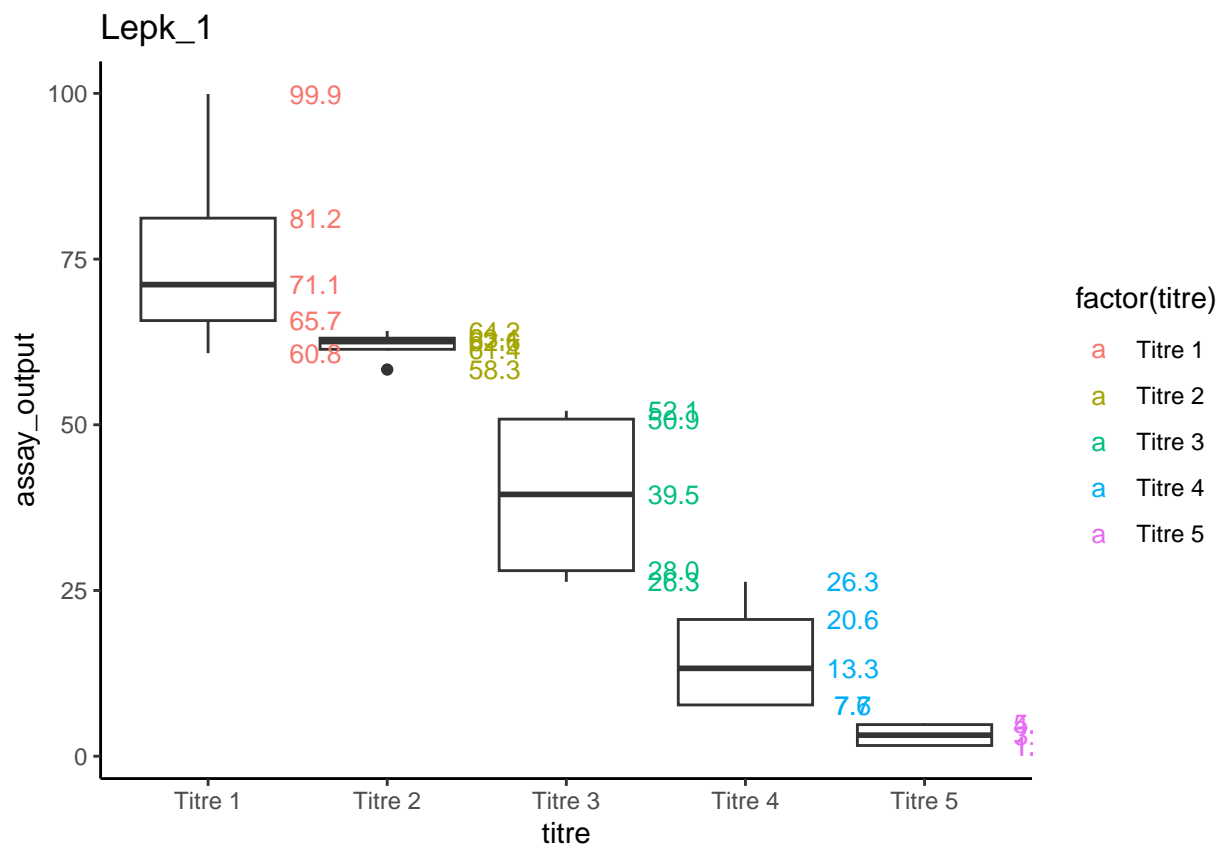
```
##
## [[66]]
```



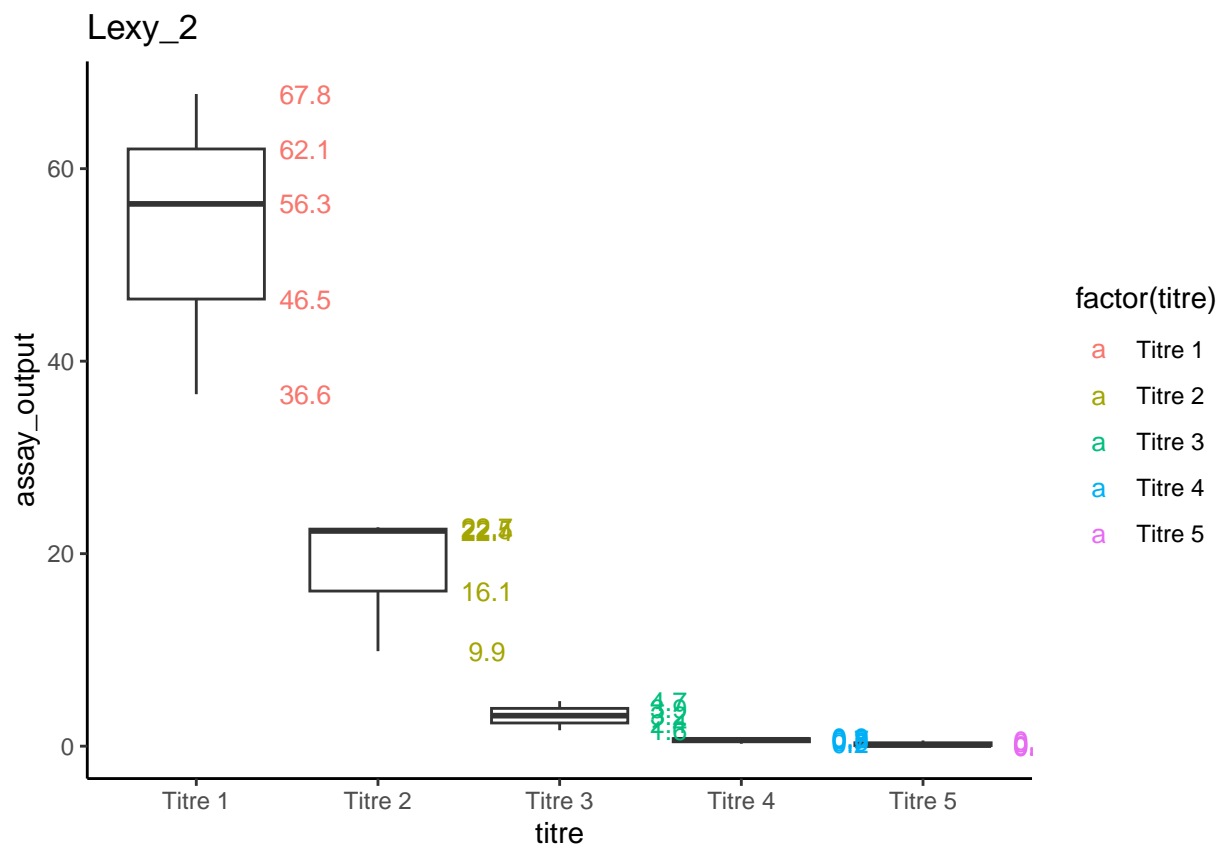
[[67]]



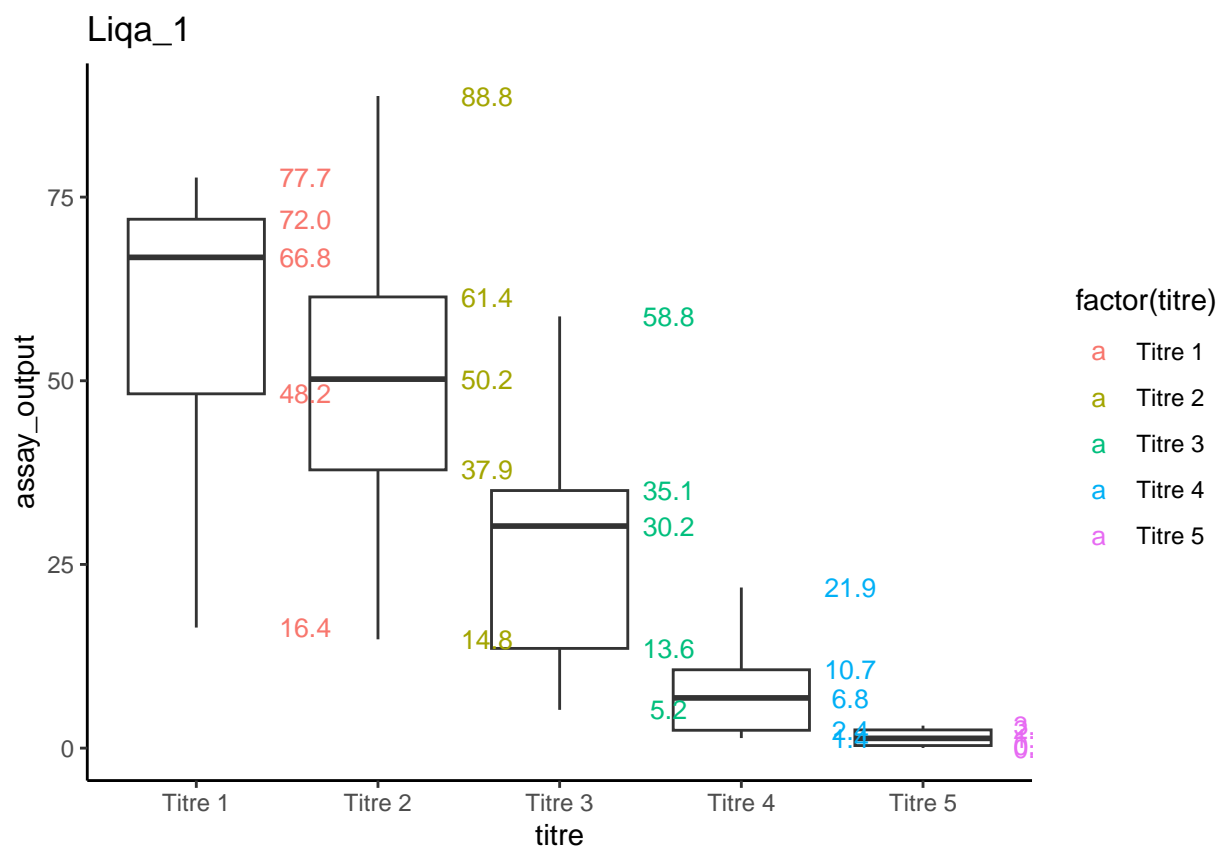
```
##
## [[68]]
```



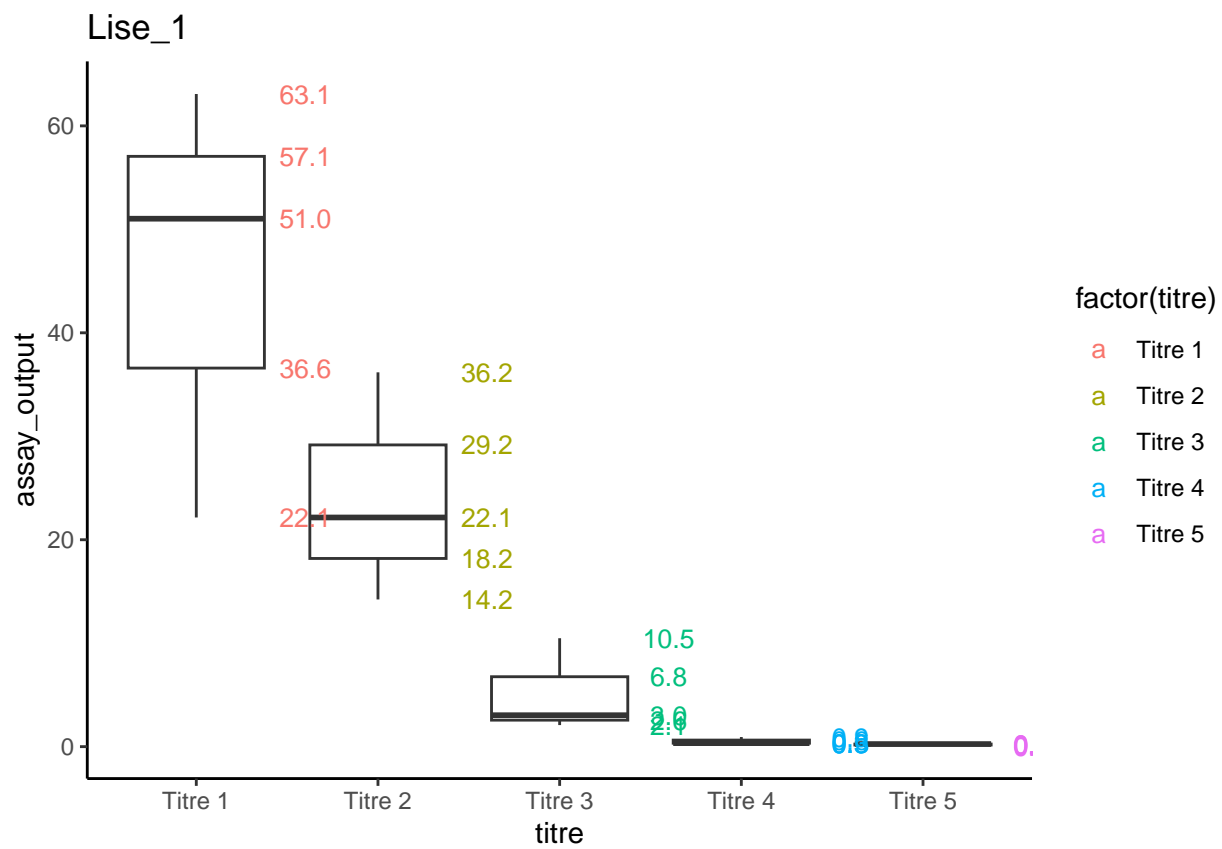
[[69]]



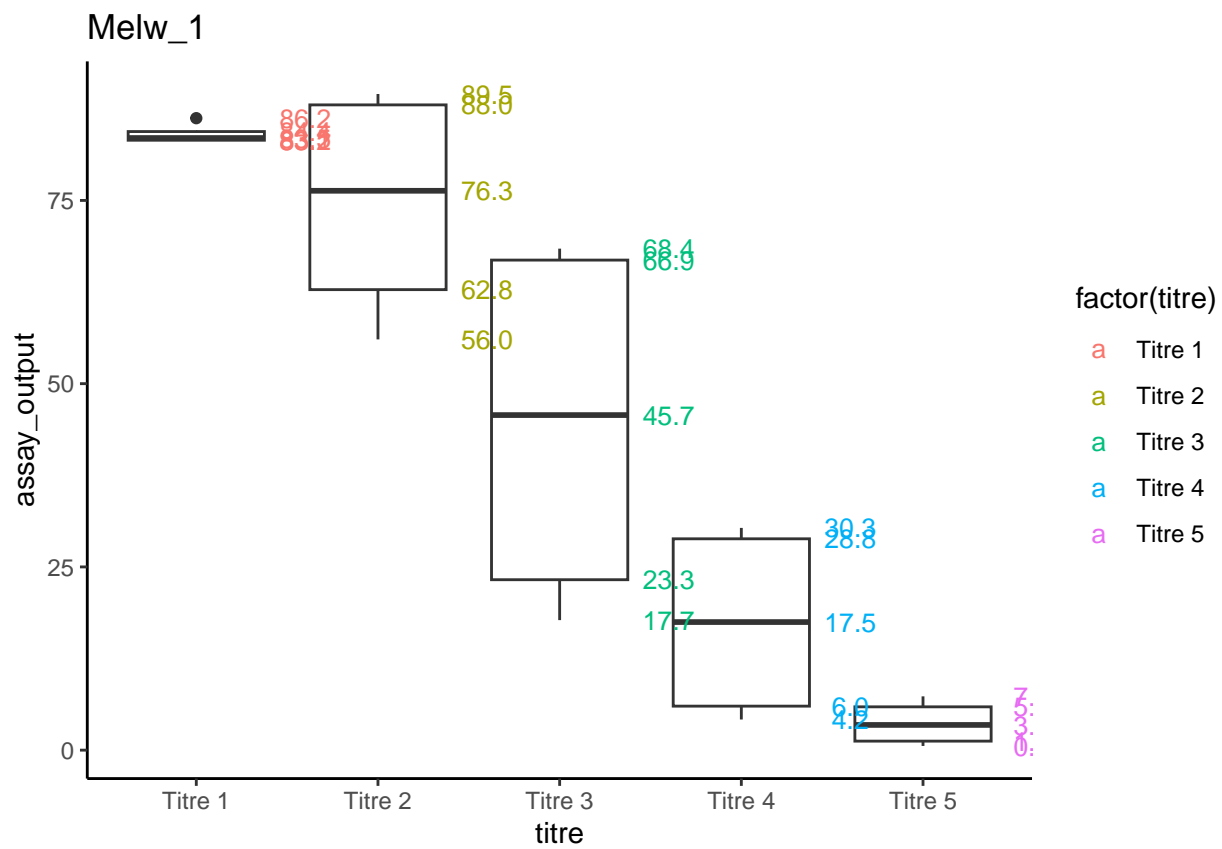
```
##
## [[70]]
```

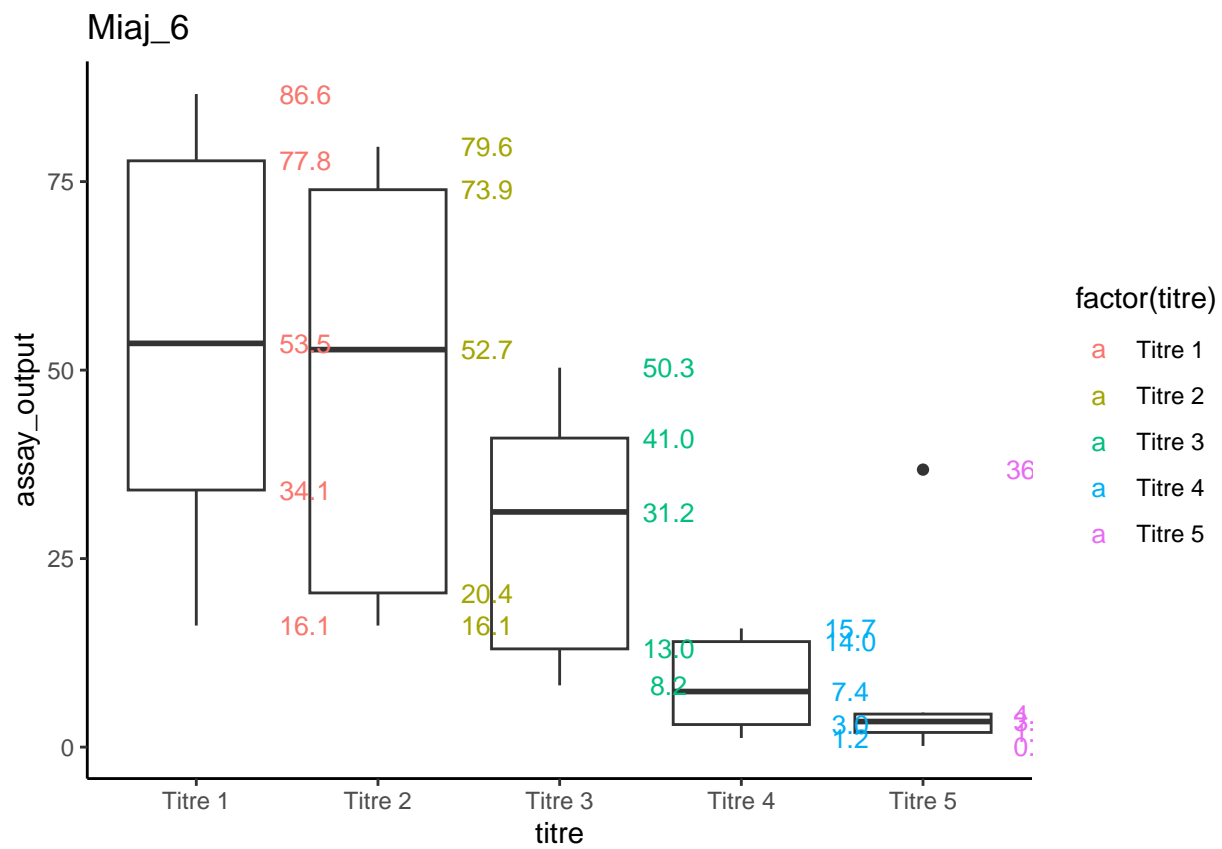
```
##  
## [[71]]
```



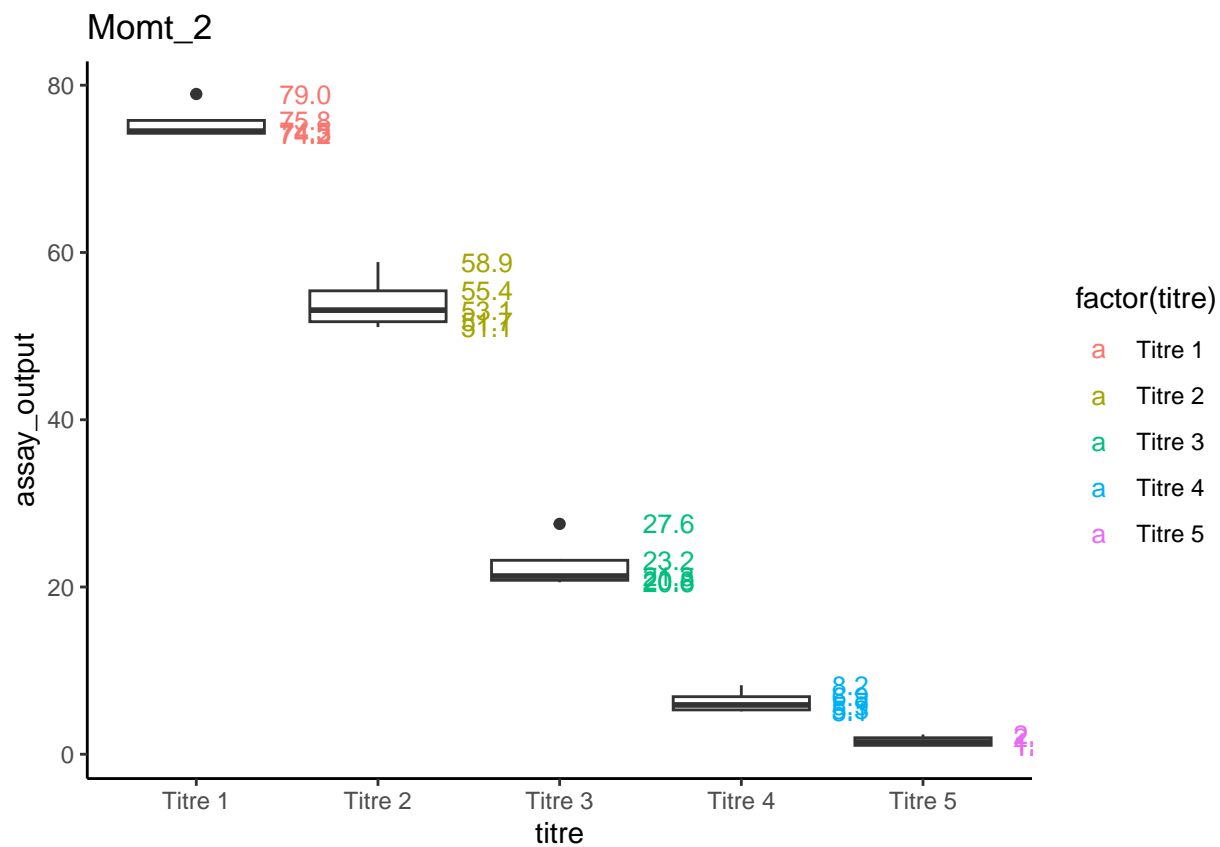
```
##
## [[72]]
```



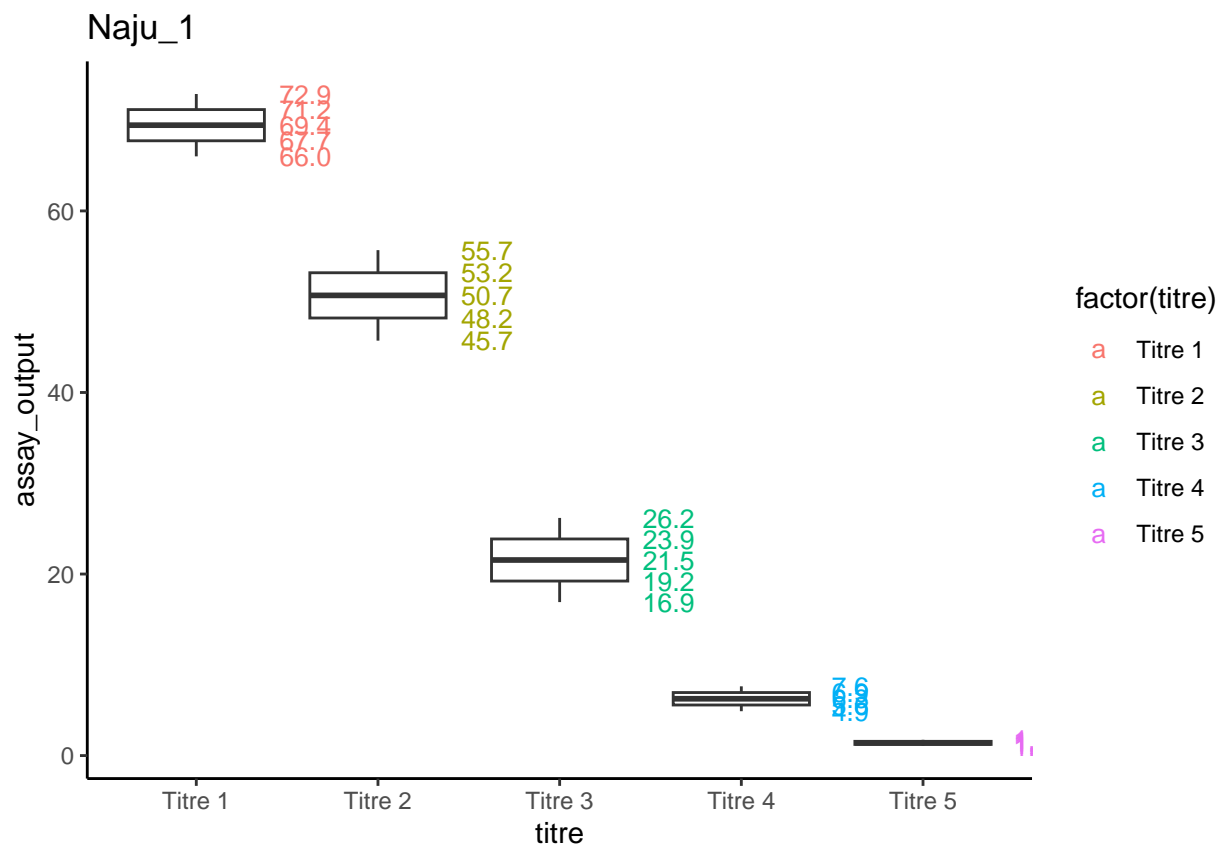
```
##
## [[73]]
```



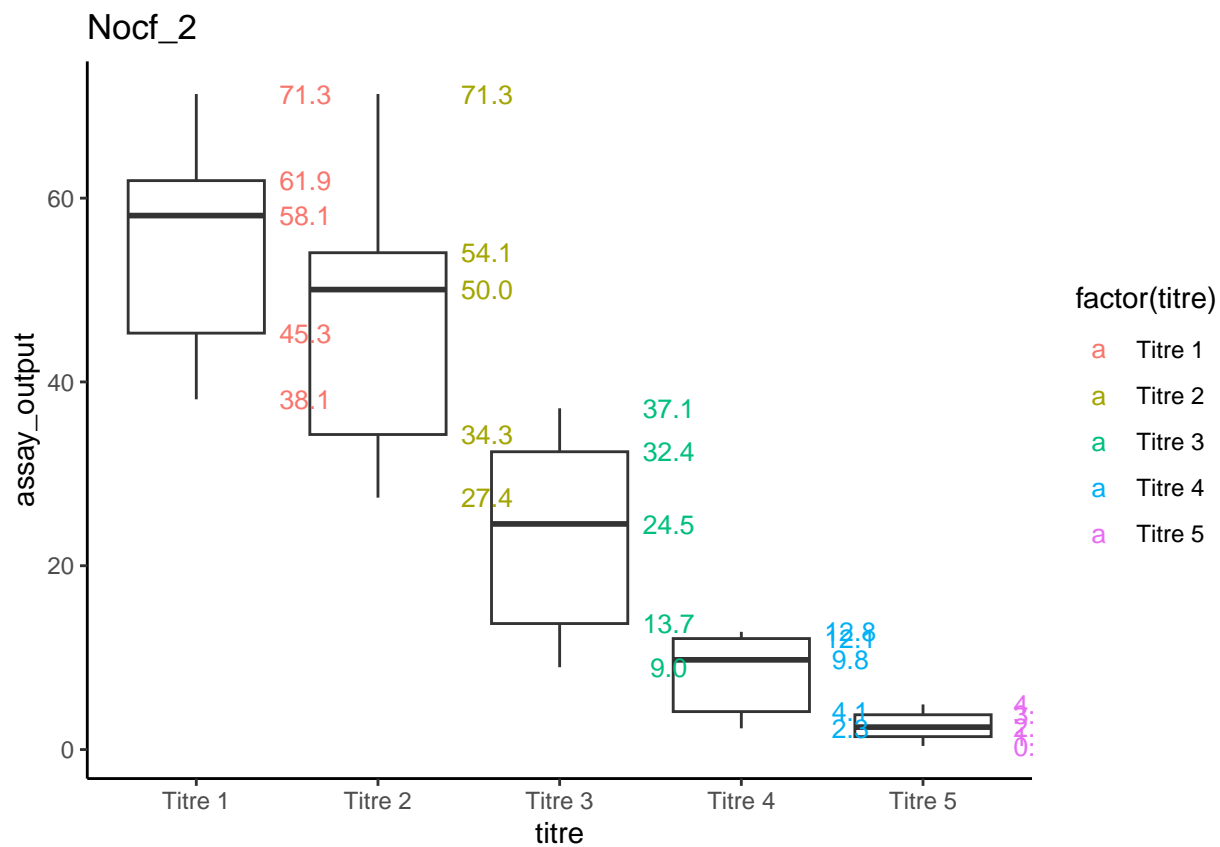
[[74]]



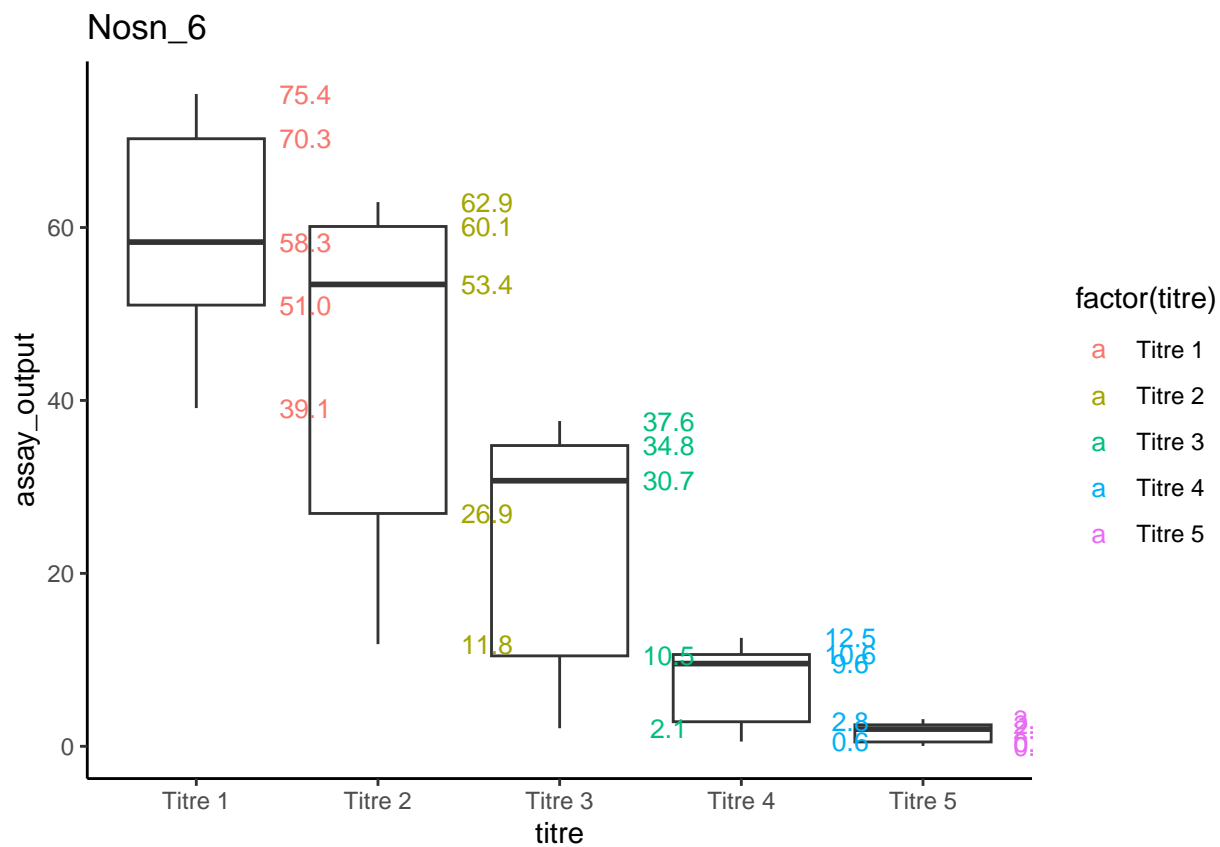
[[75]]



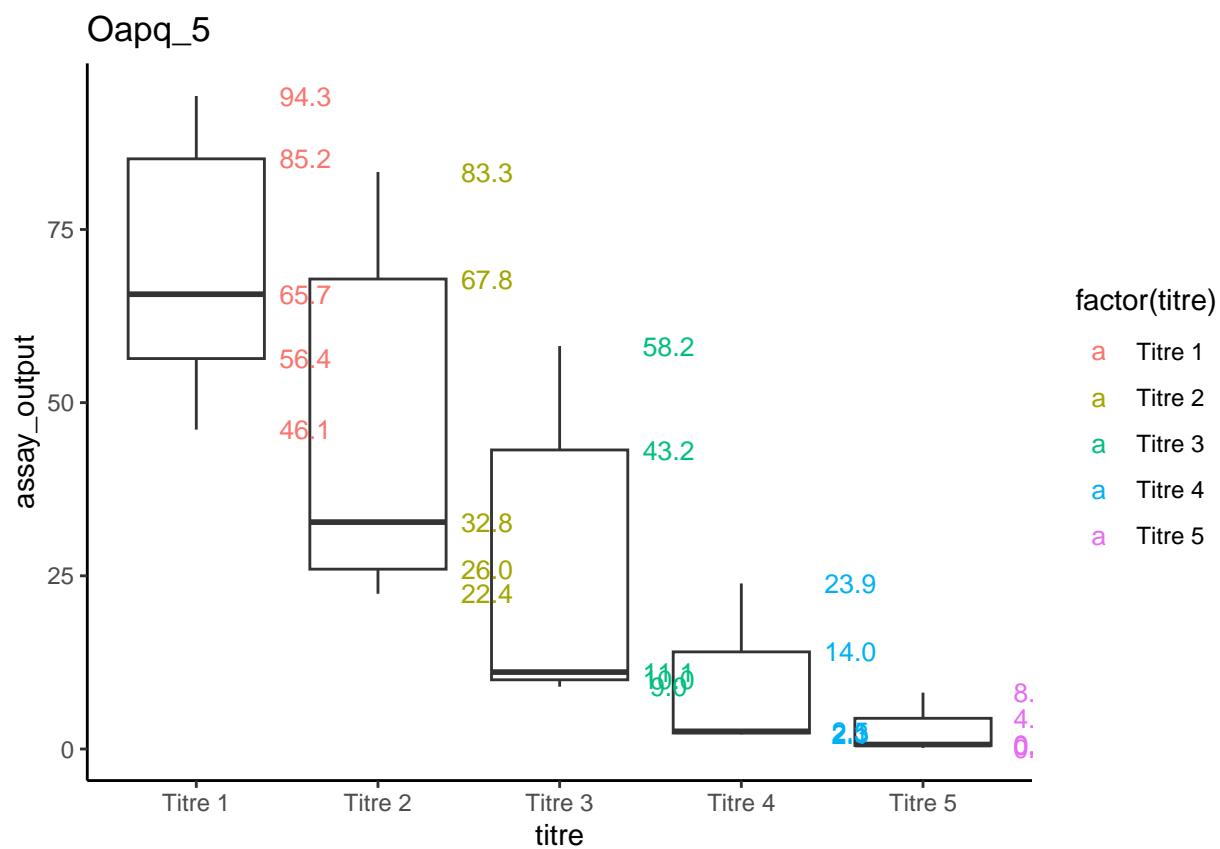
```
##
## [[76]]
```



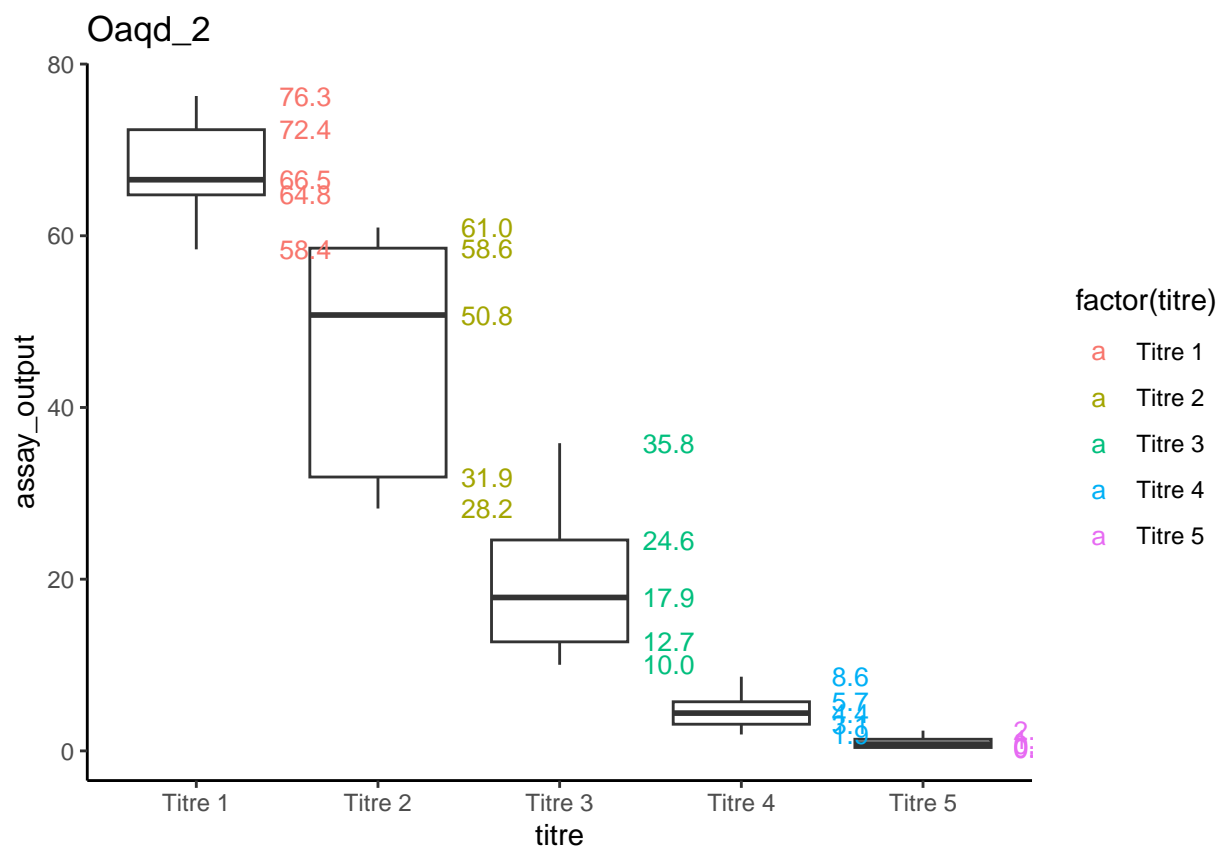
```
##
## [[77]]
```



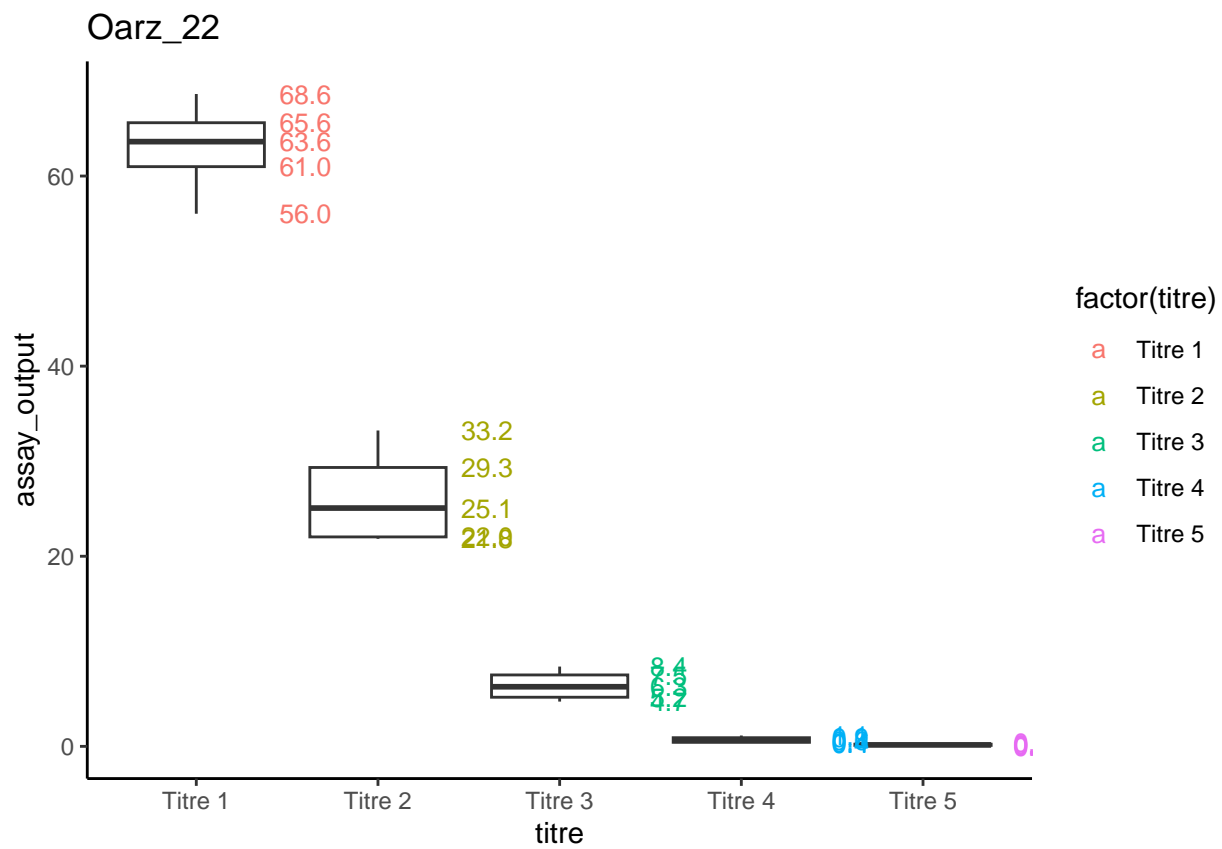
[[78]]



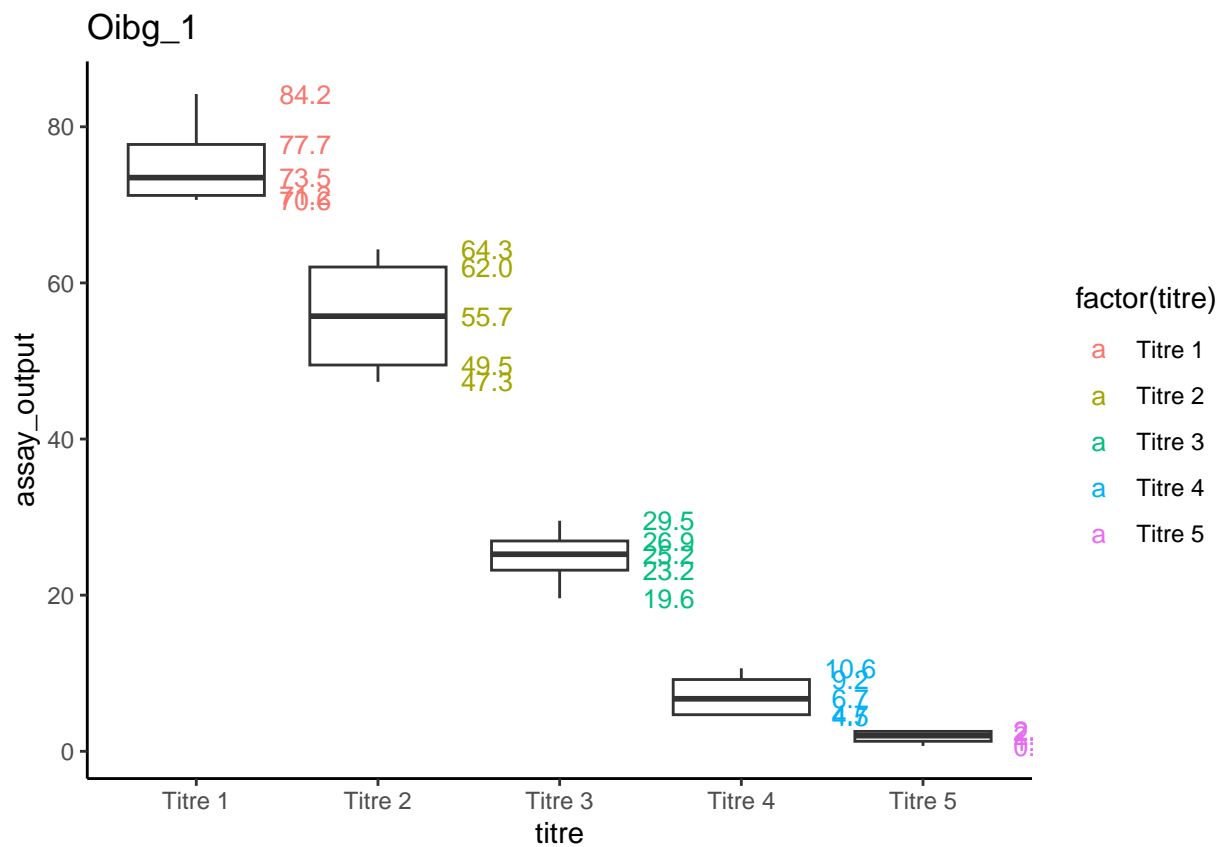
[[79]]



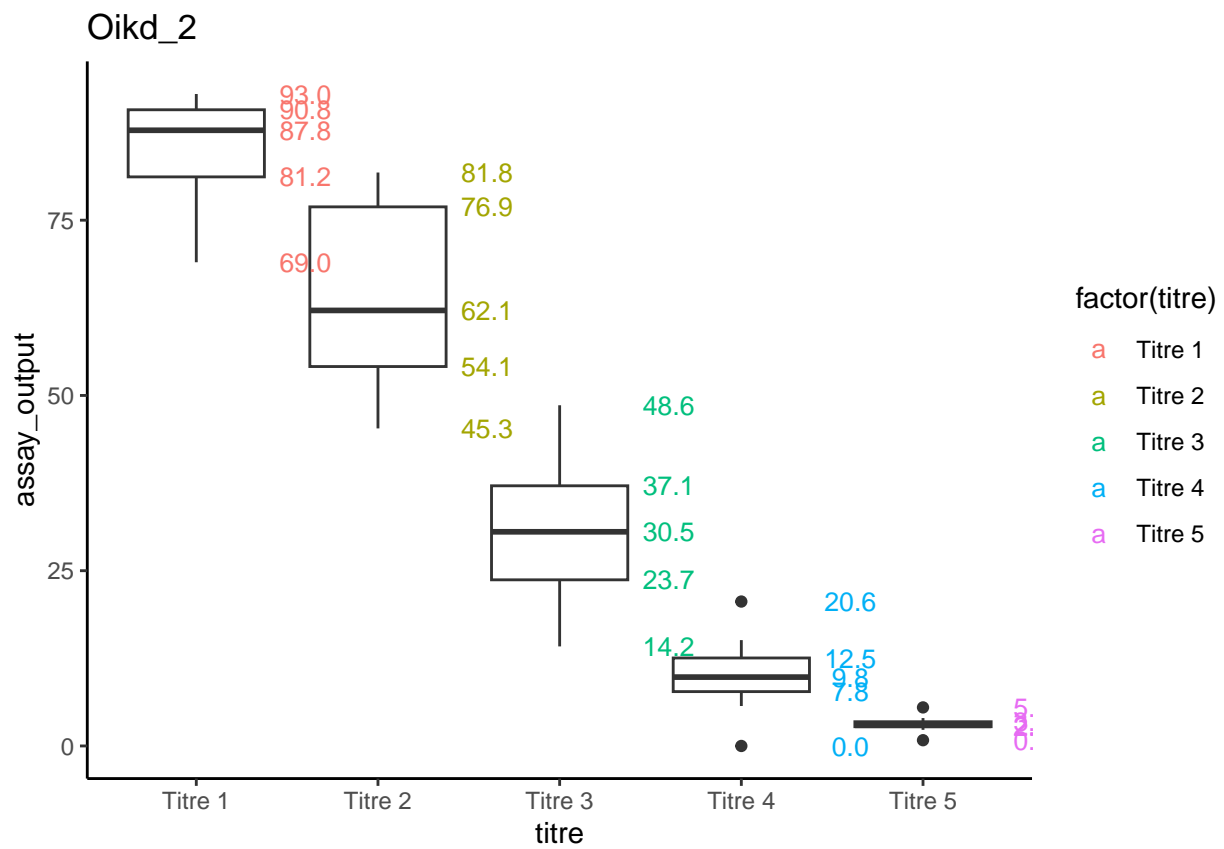
[[80]]



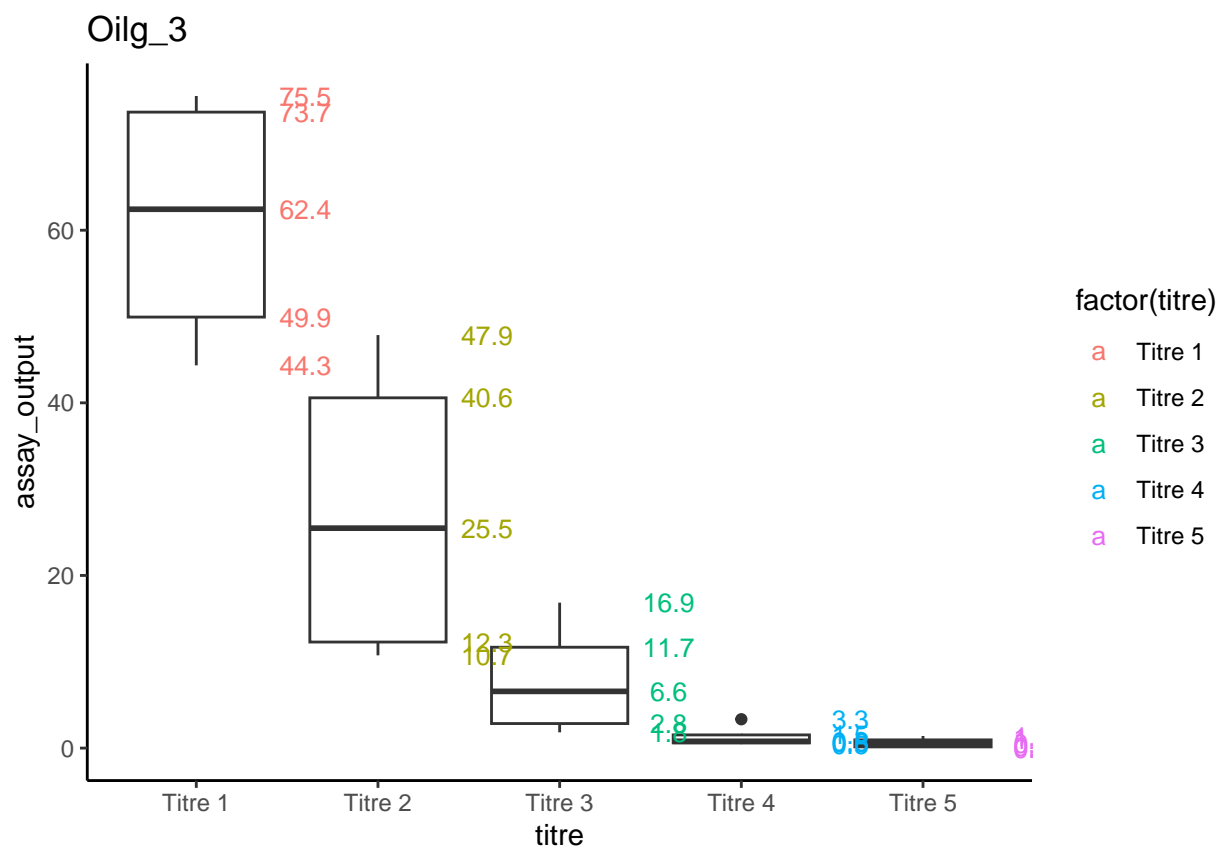
```
##
## [[81]]
```



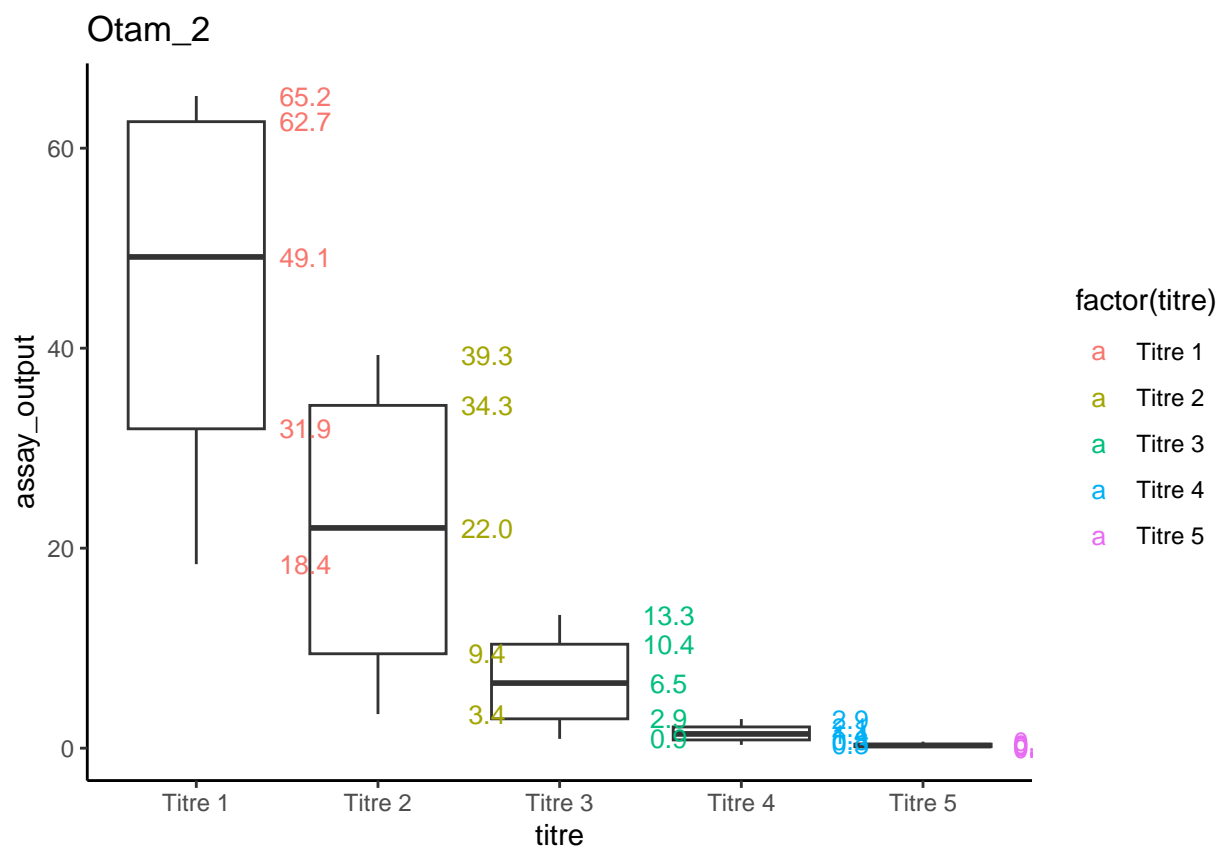
[[82]]



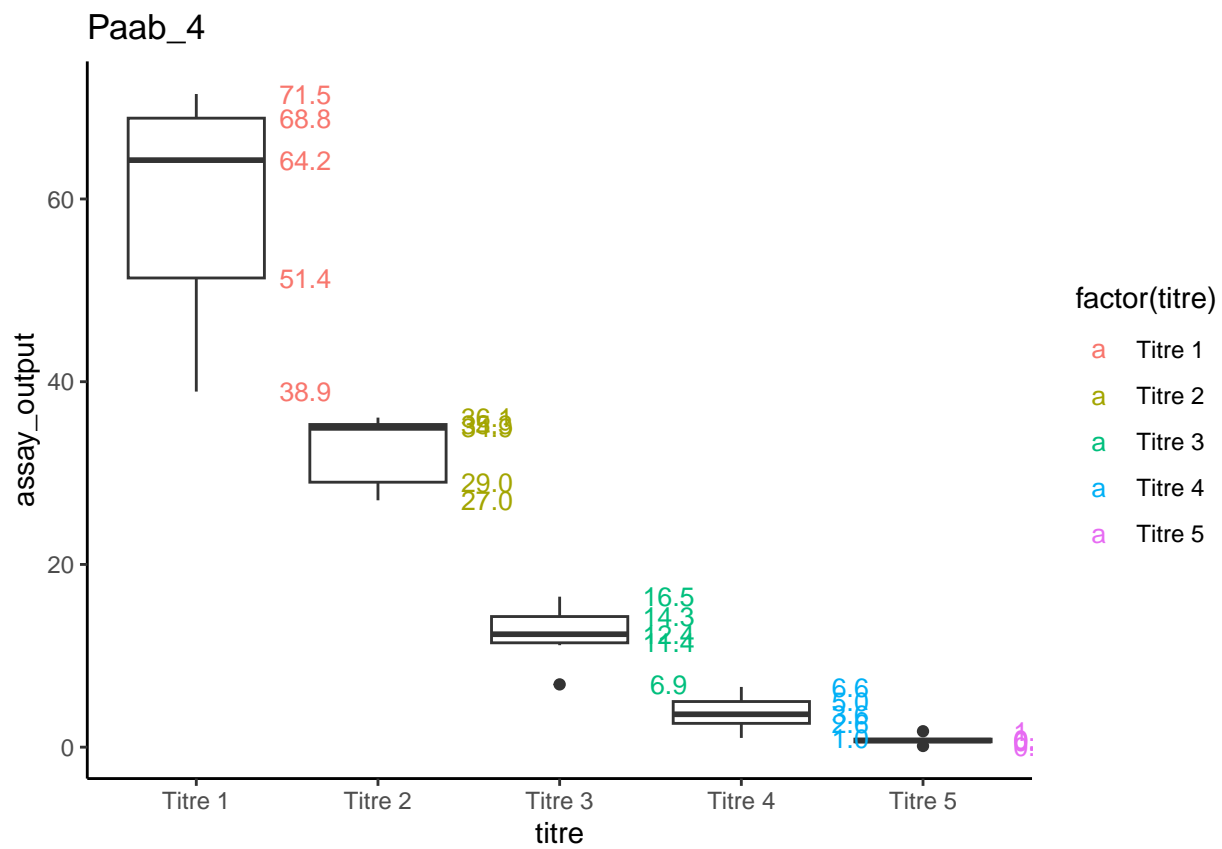
```
##
## [[83]]
```



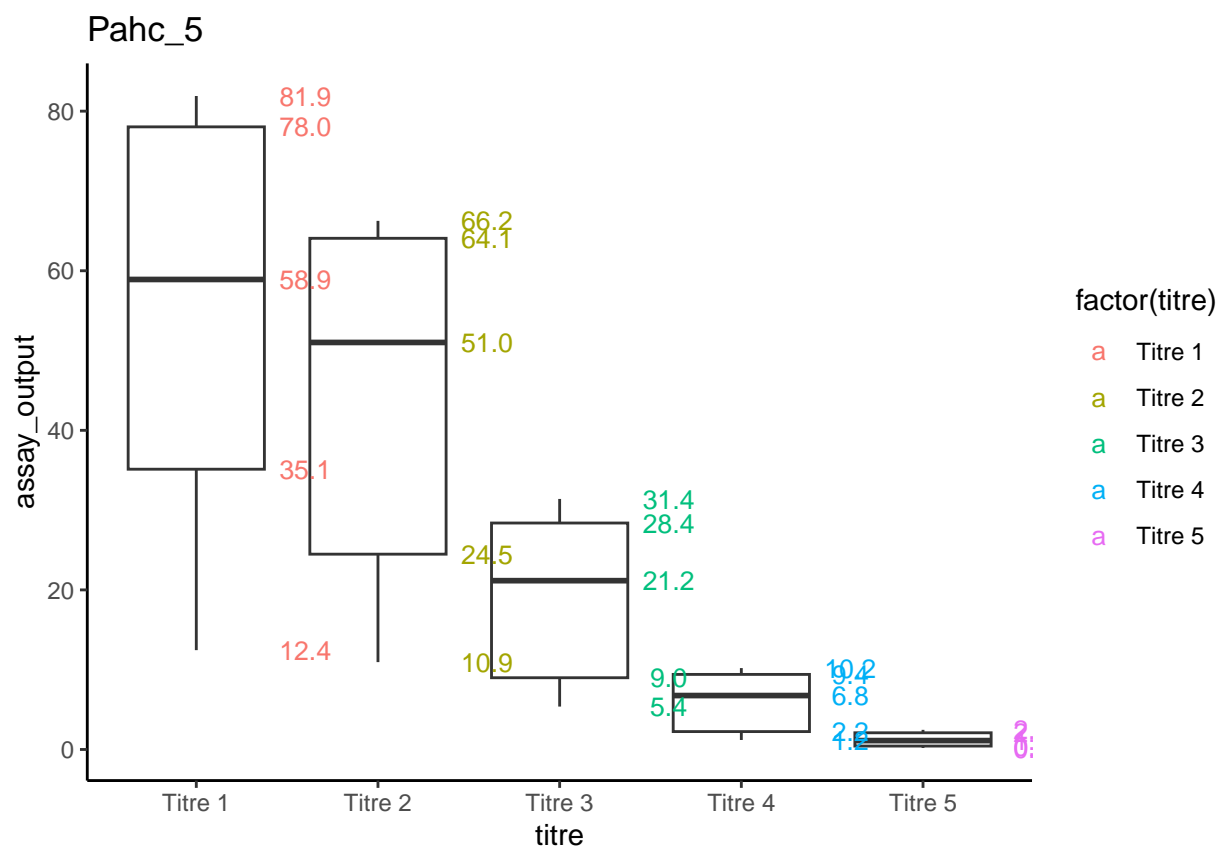
```
##
## [[84]]
```



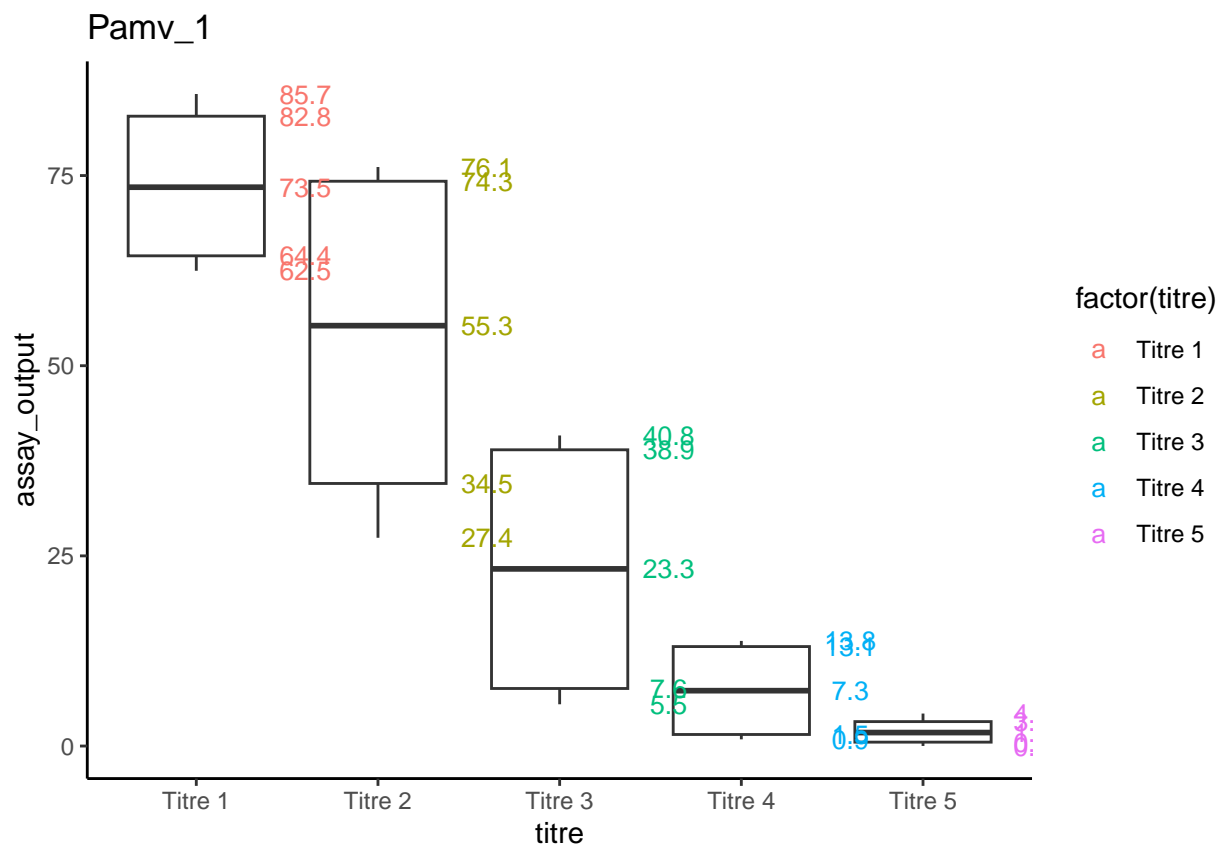
```
##
## [[85]]
```



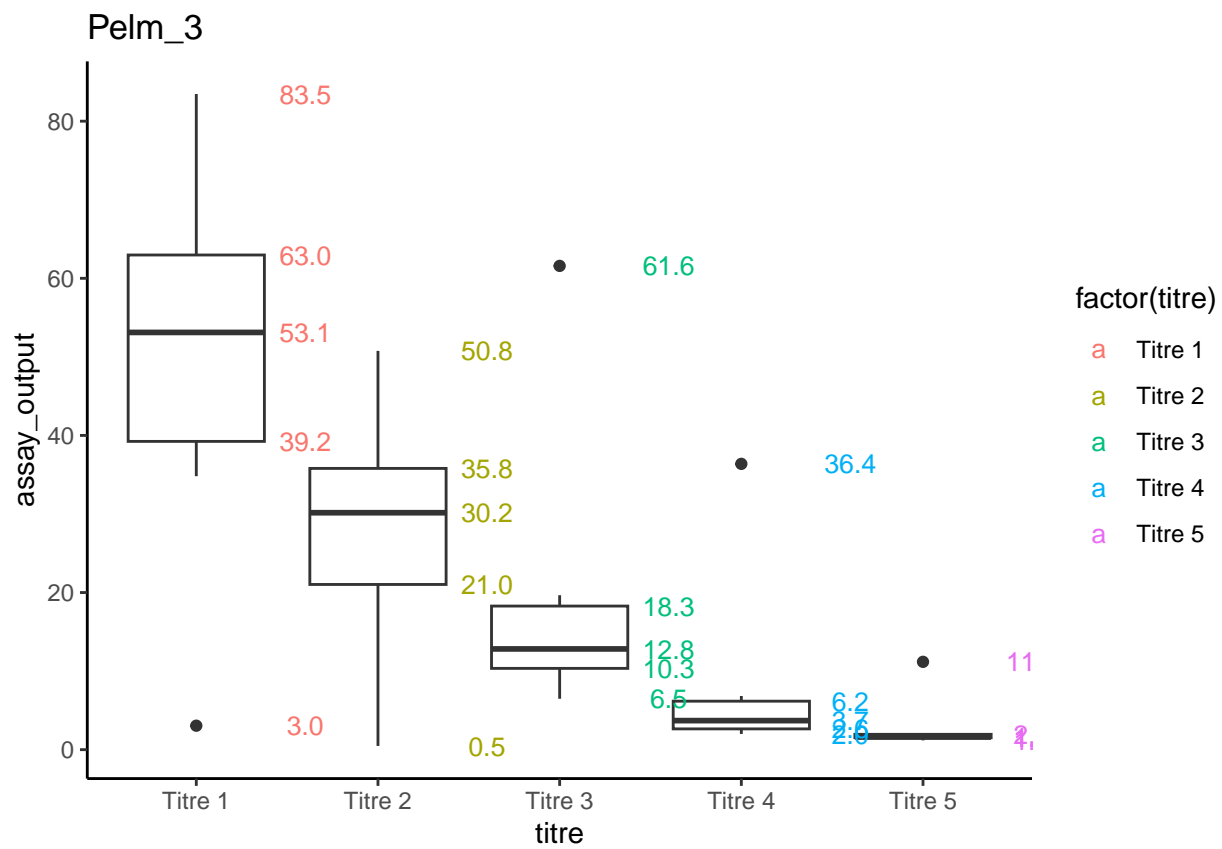
[[86]]



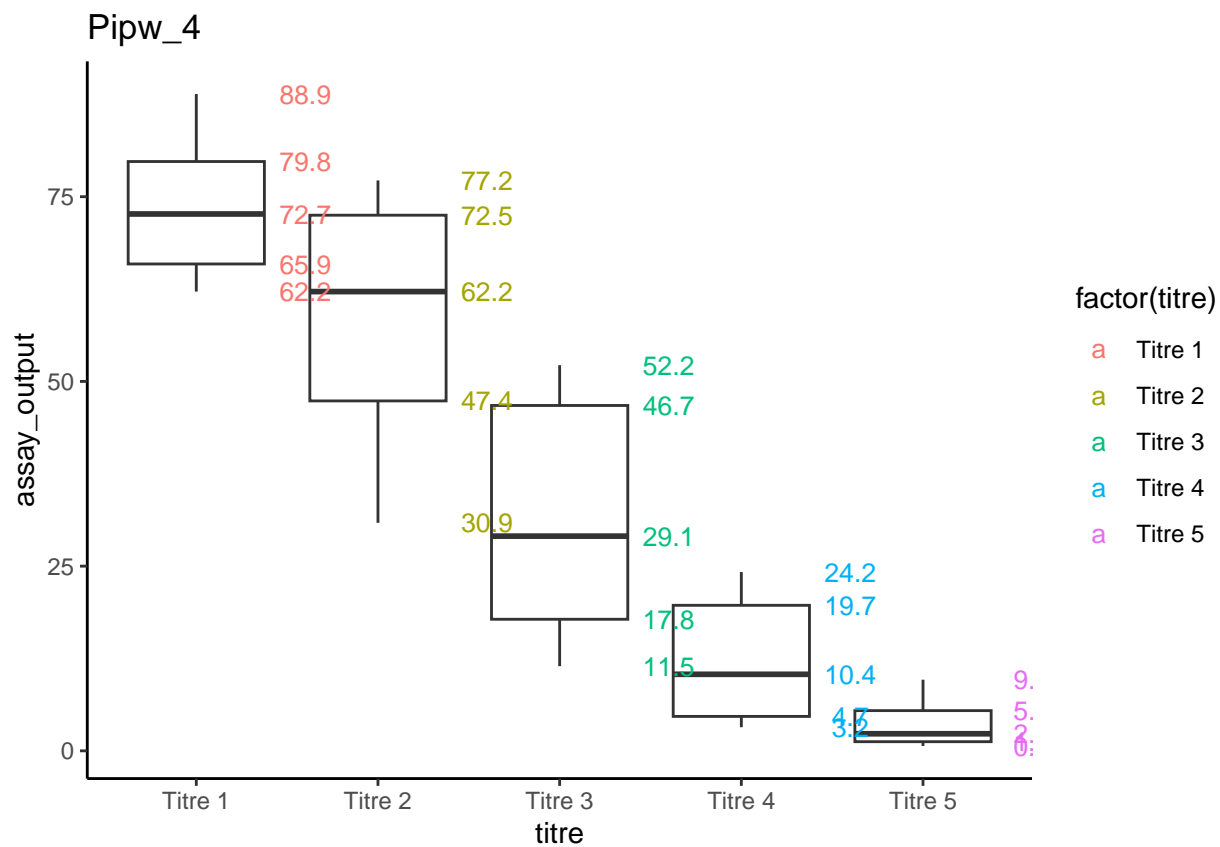
```
##
## [[87]]
```



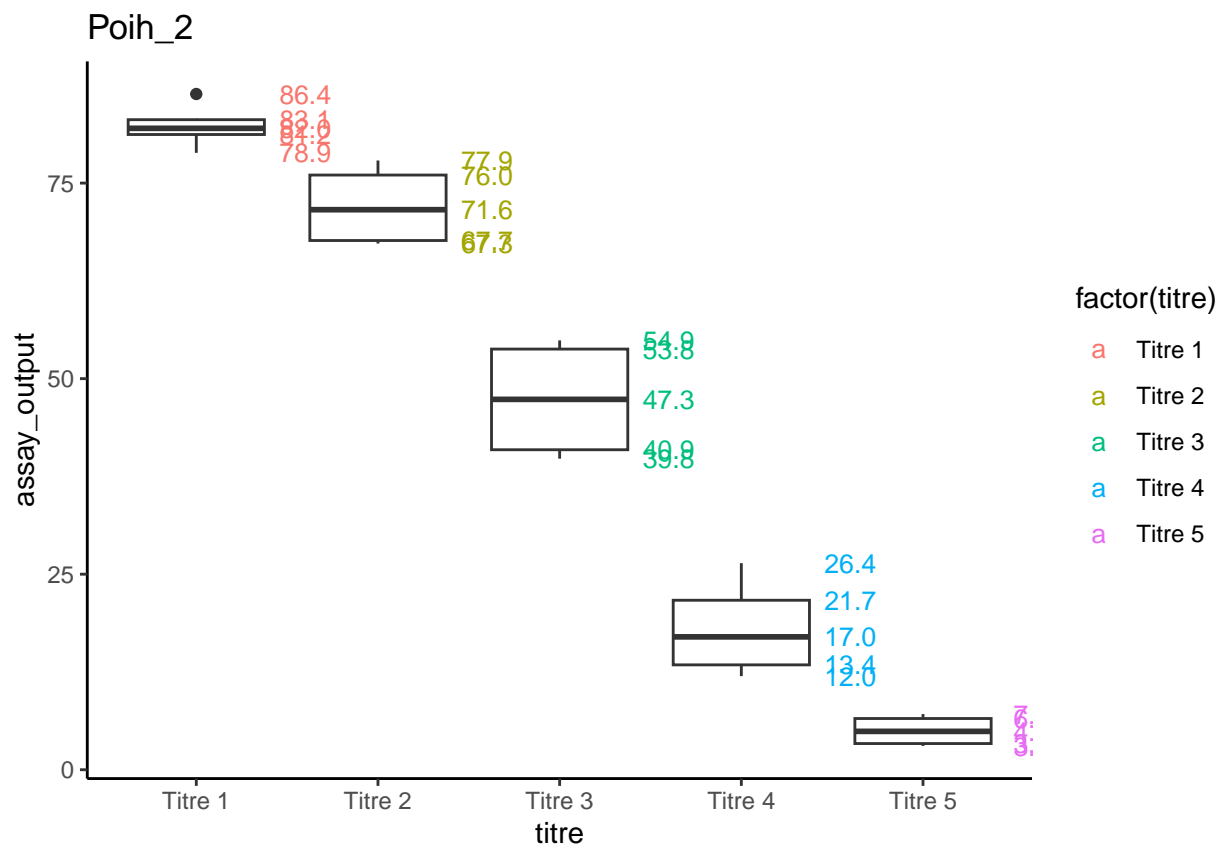
```
##
## [[88]]
```



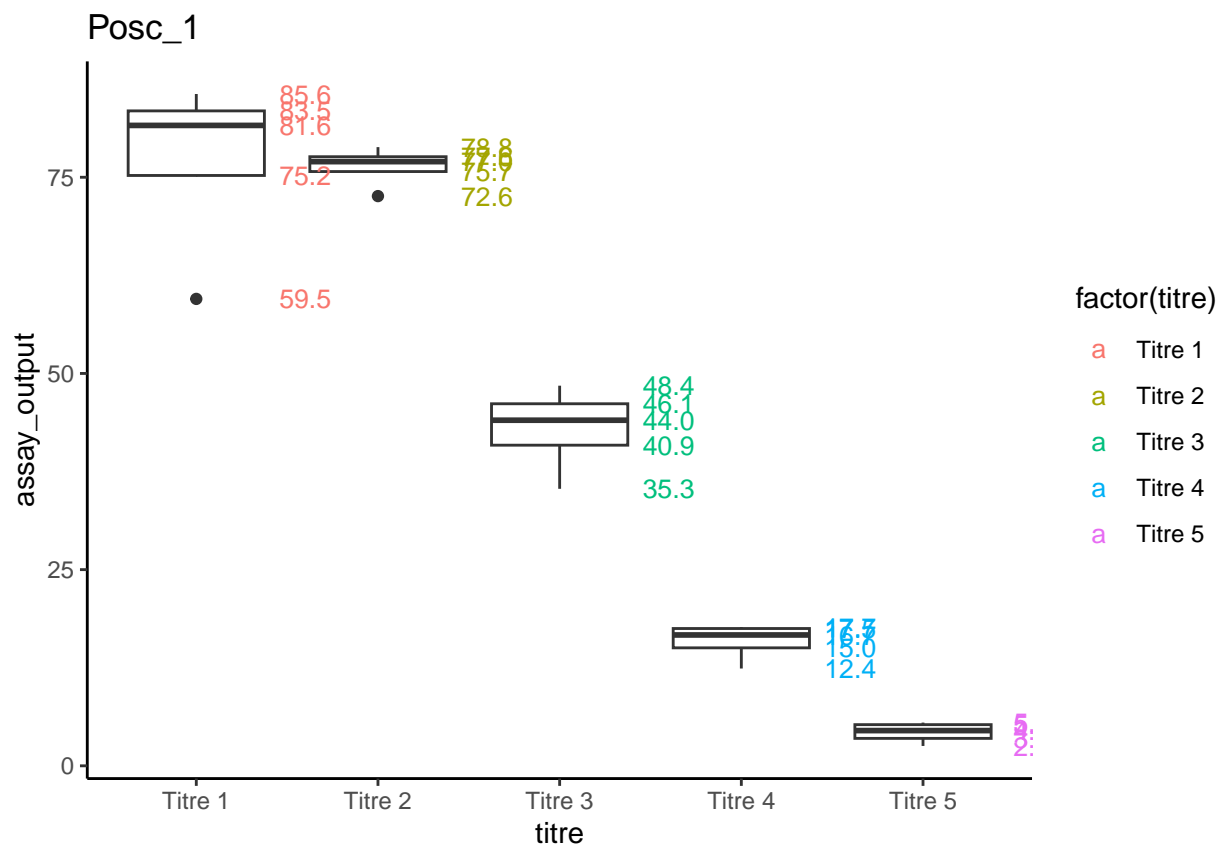
[[89]]



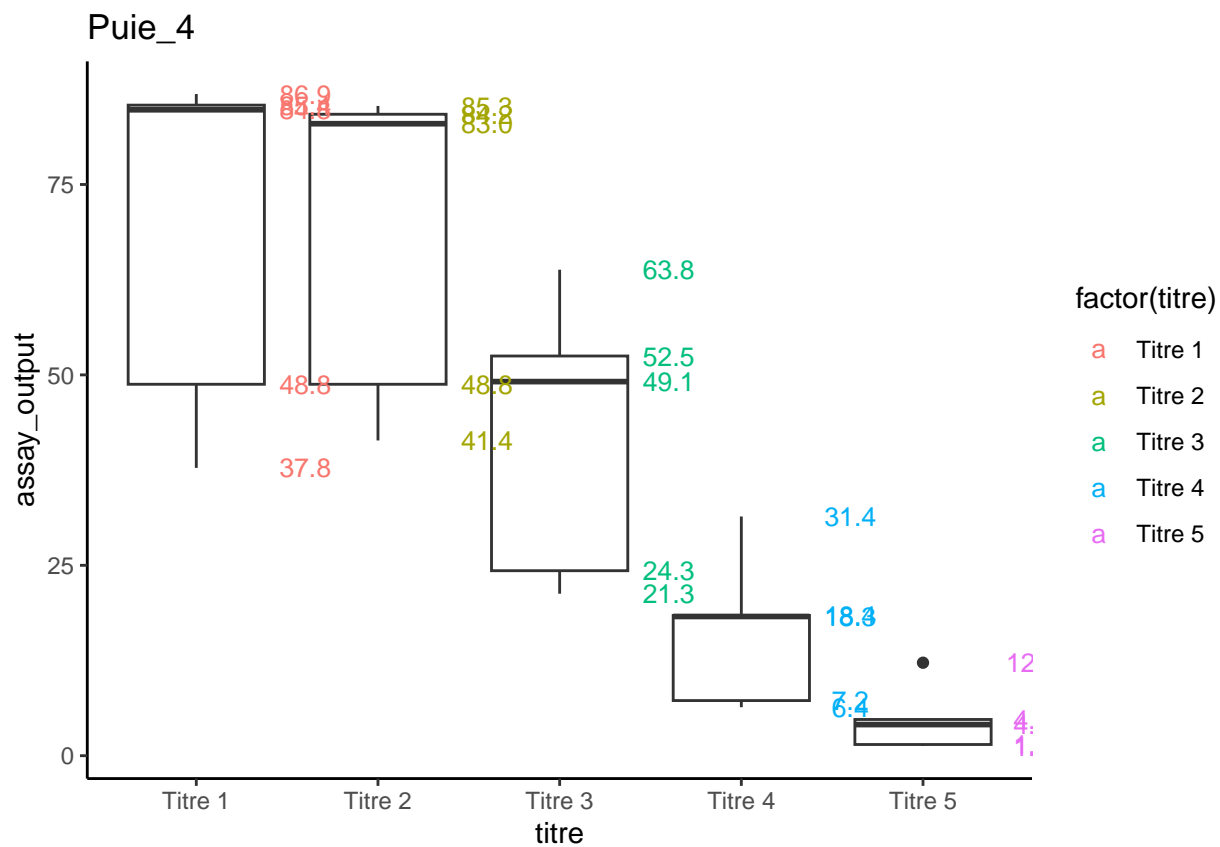
```
##
## [[90]]
```



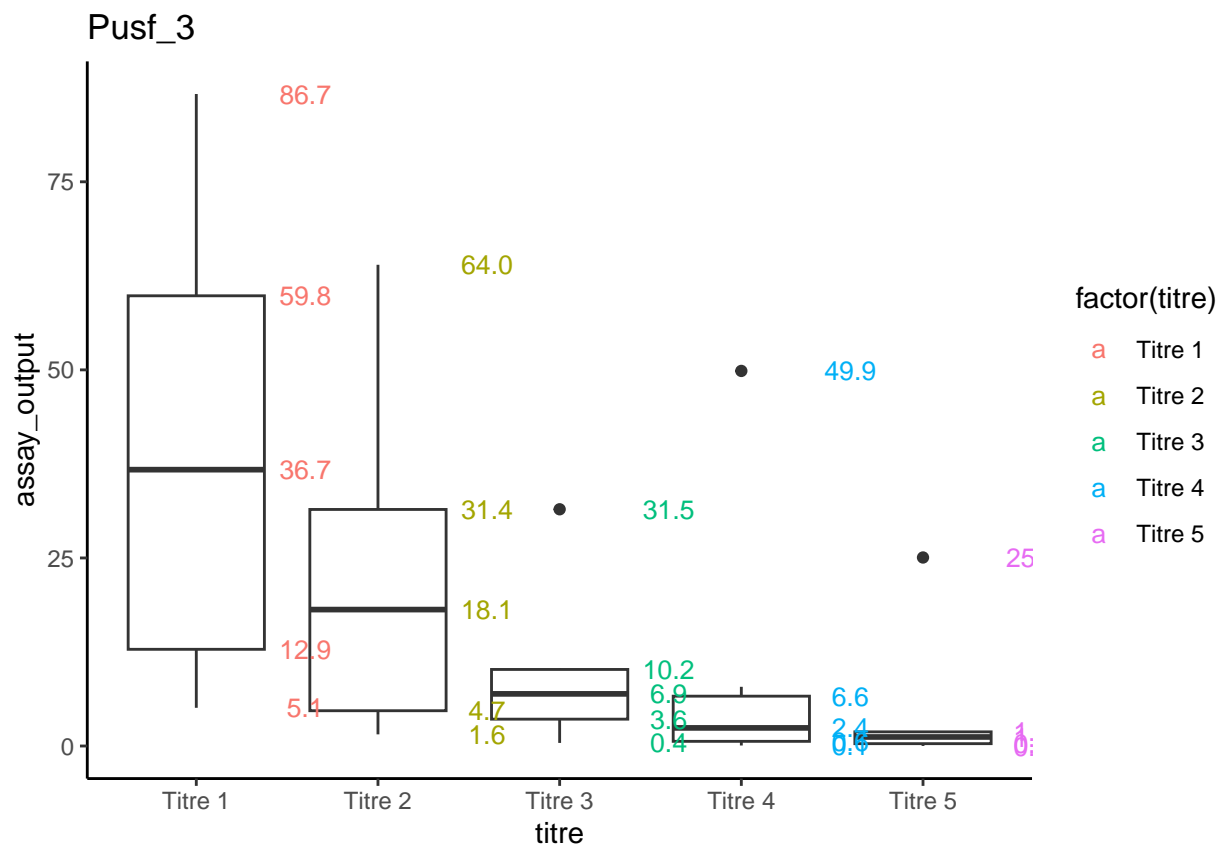
[[91]]



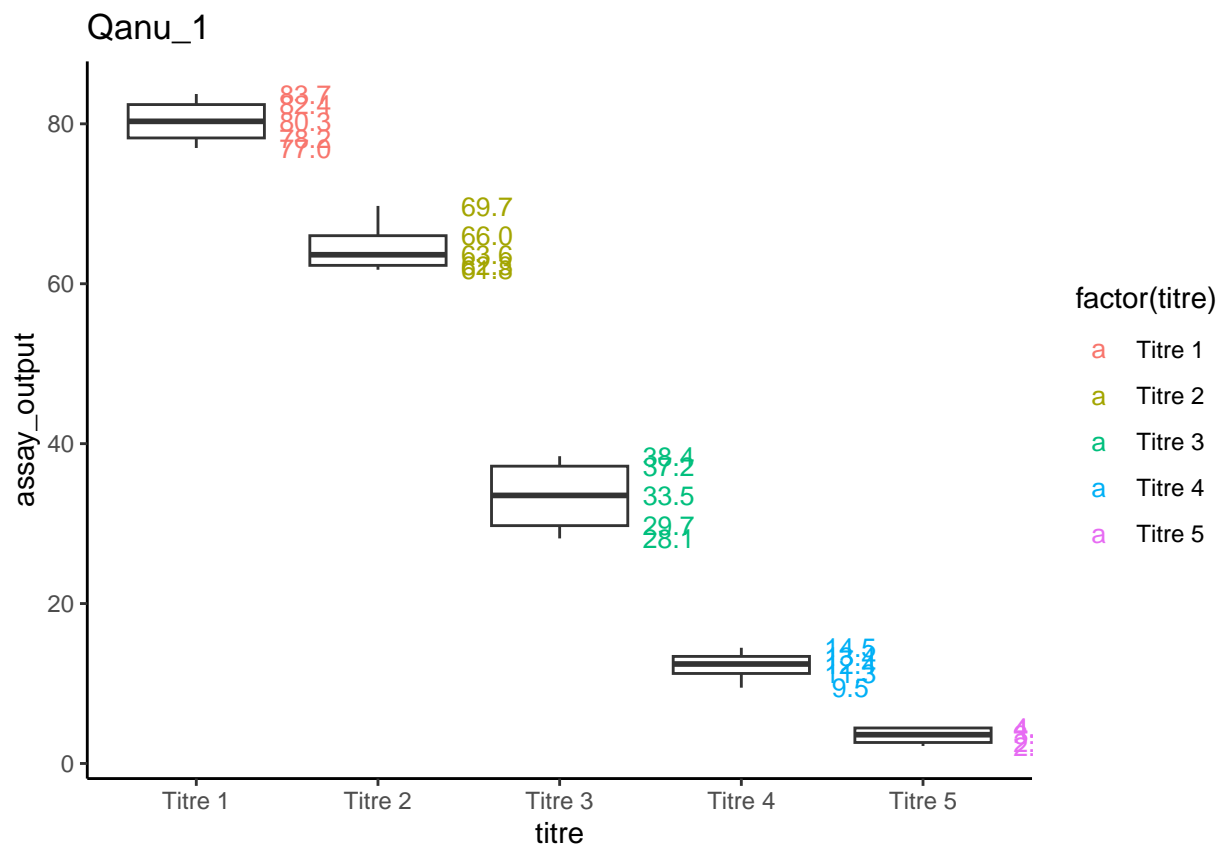
[[92]]



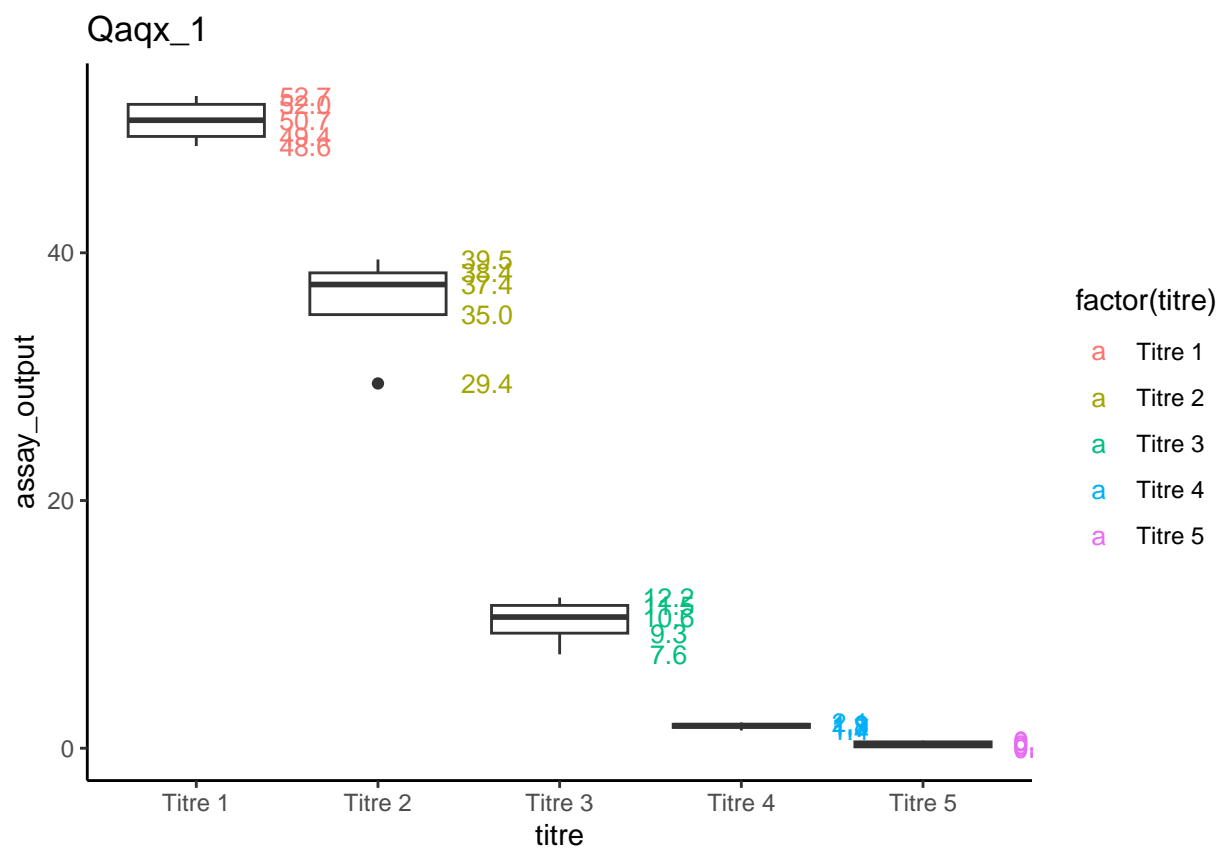
[[93]]



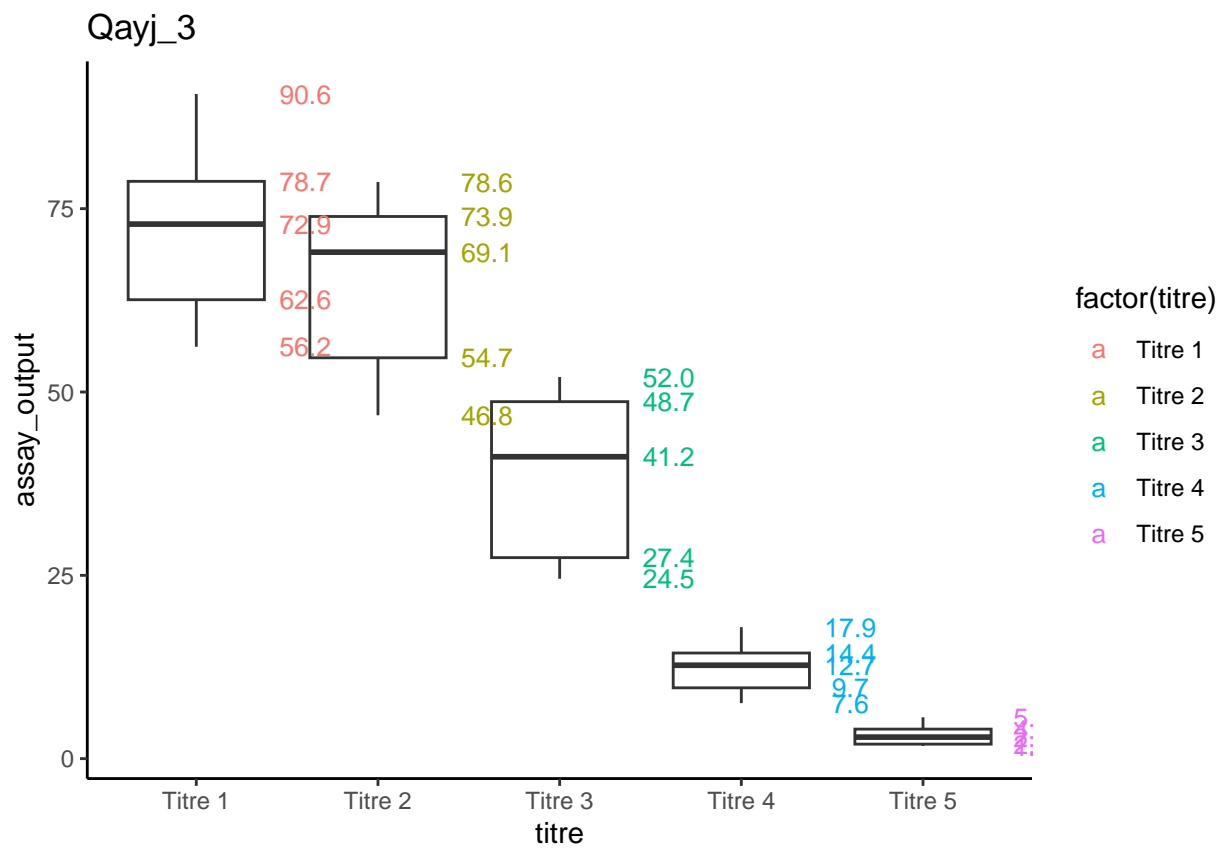
[[94]]



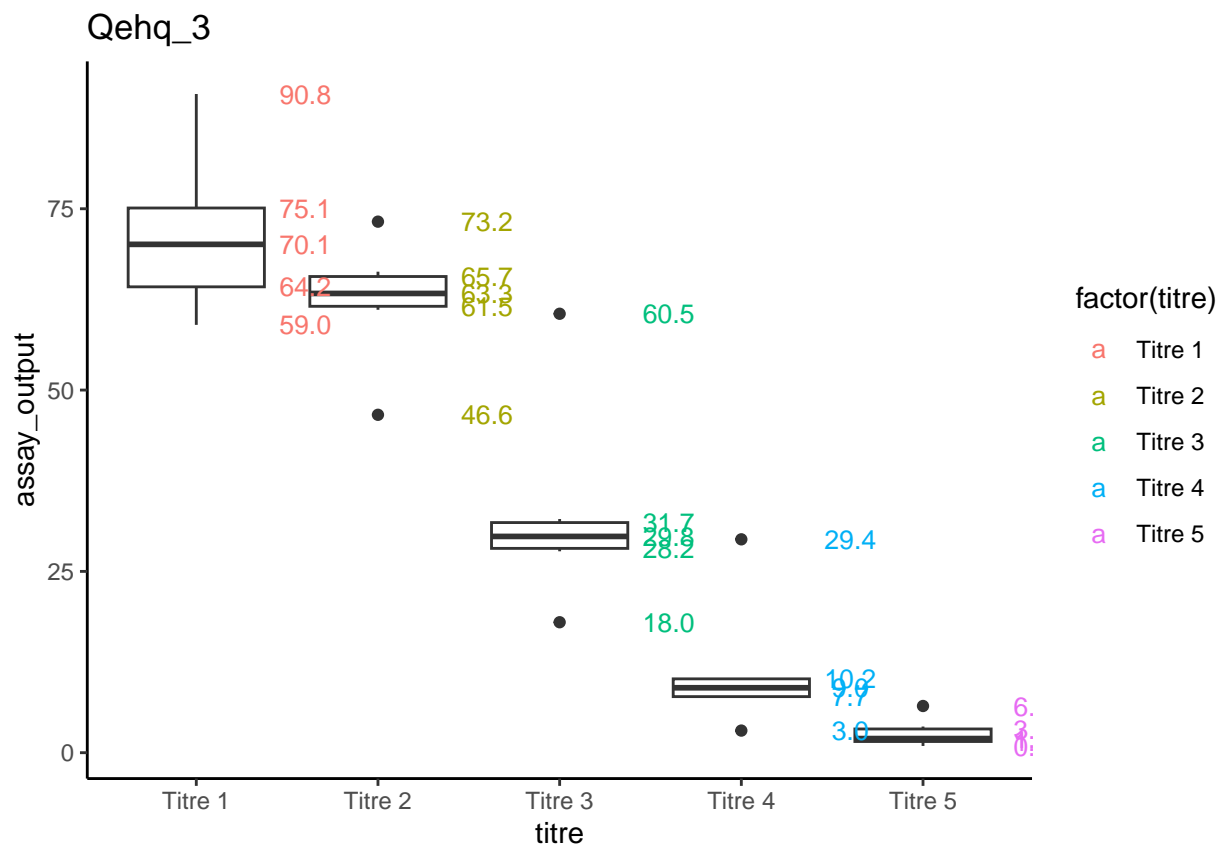
[[95]]



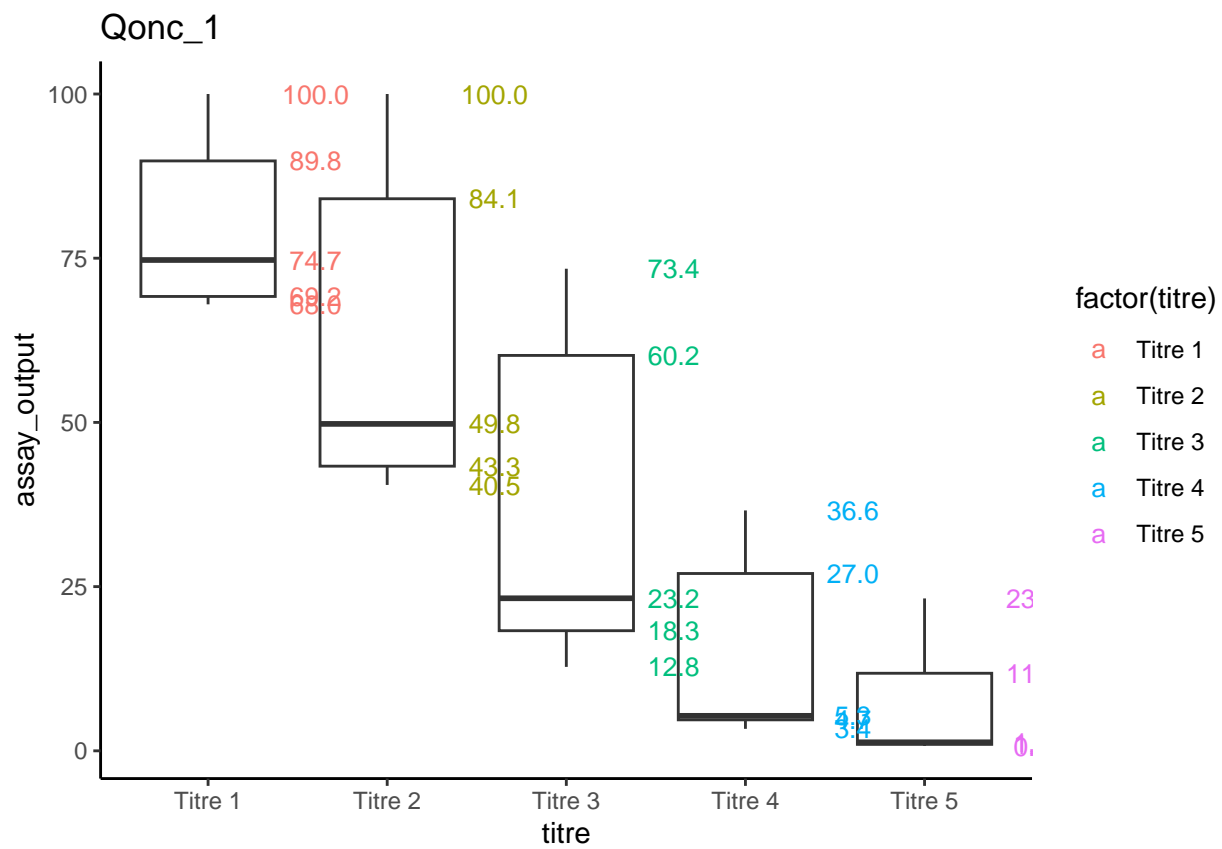
[[96]]



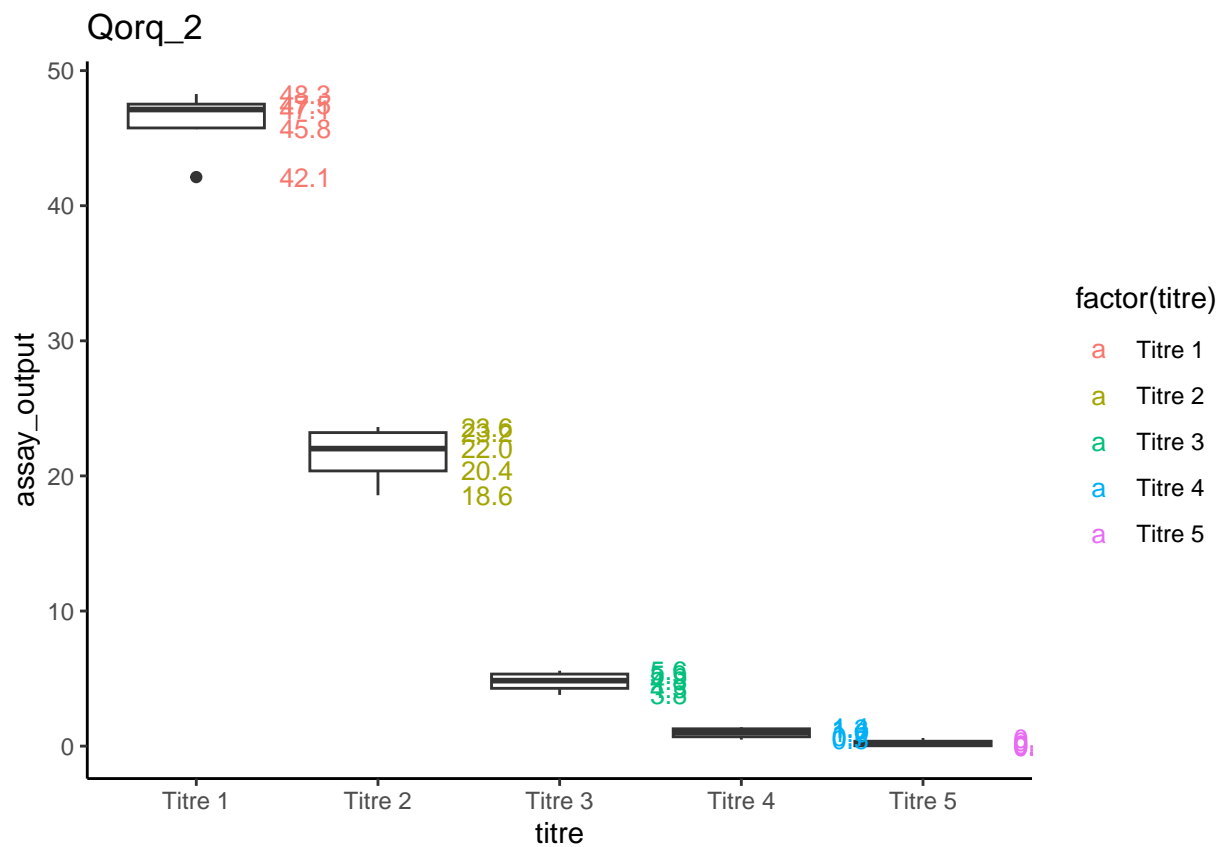
[[97]]



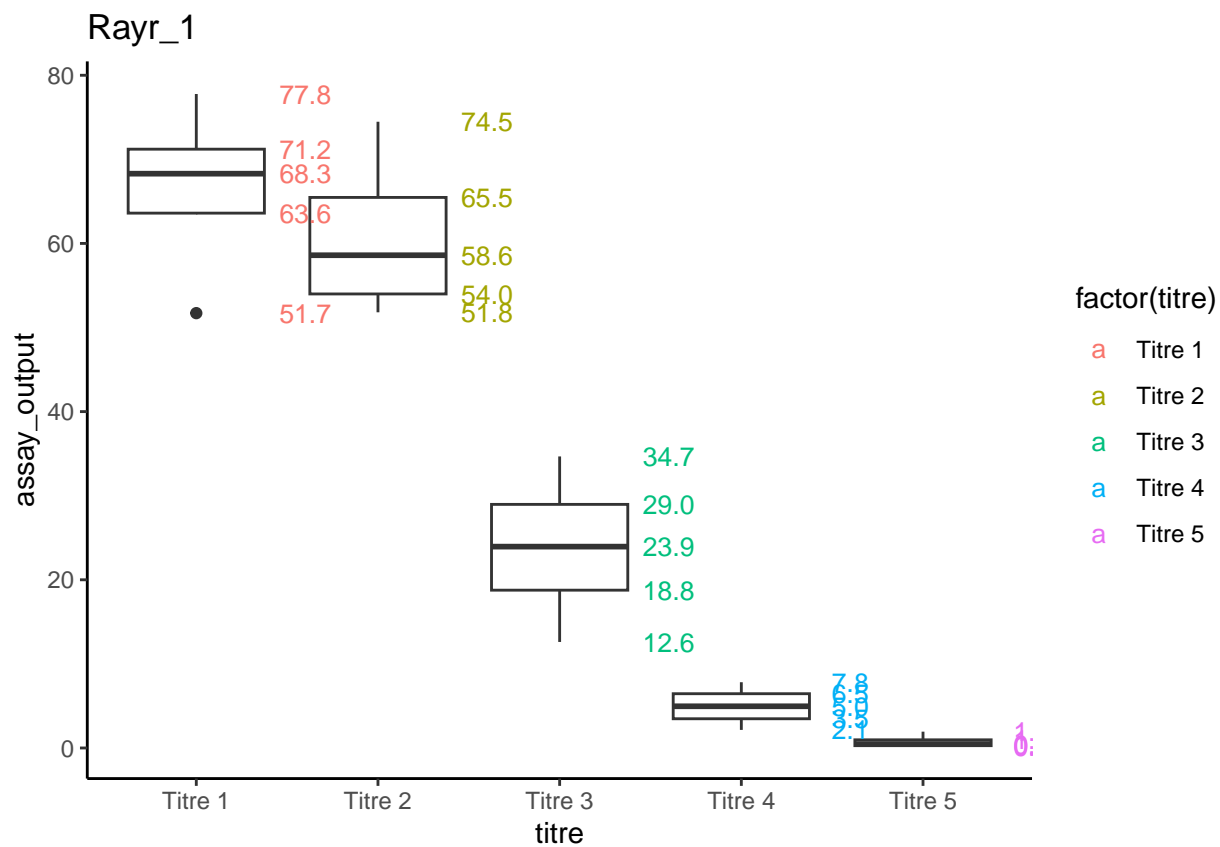
[[98]]



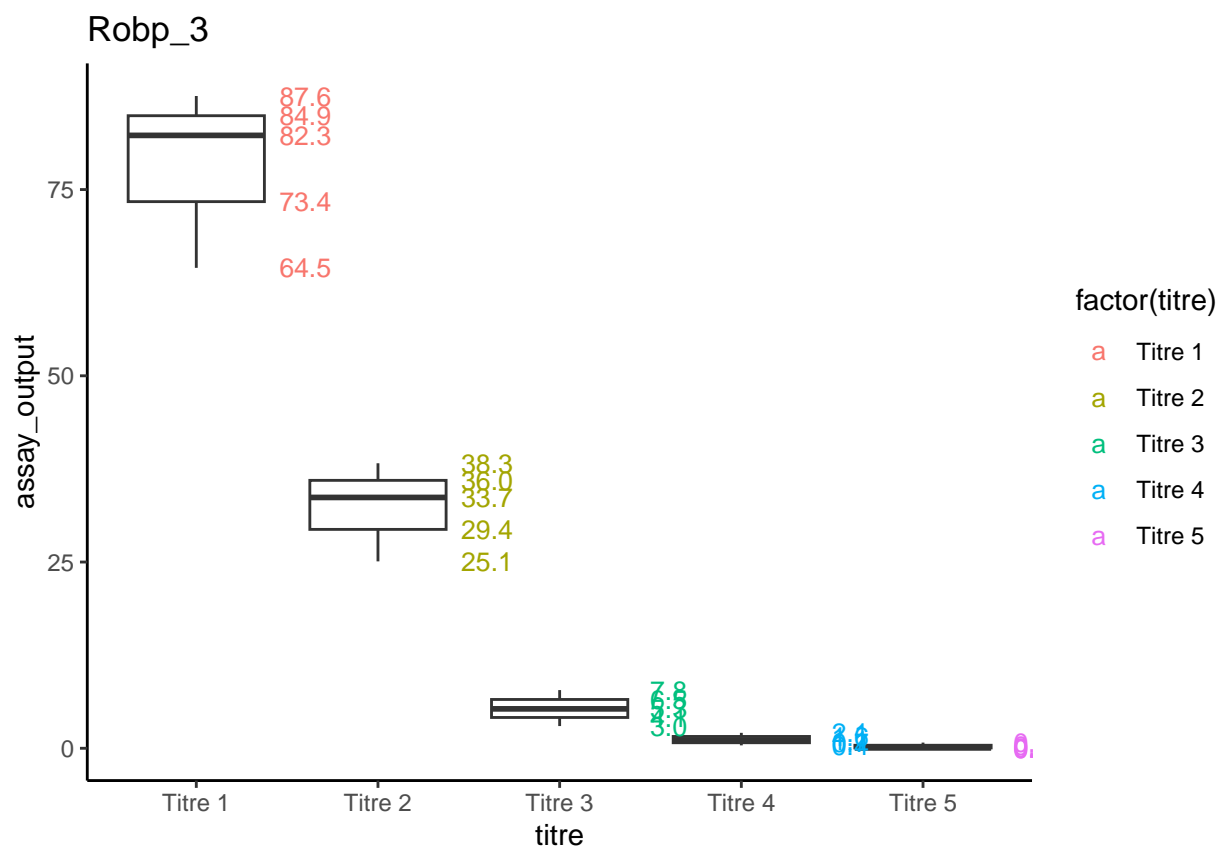
[[99]]



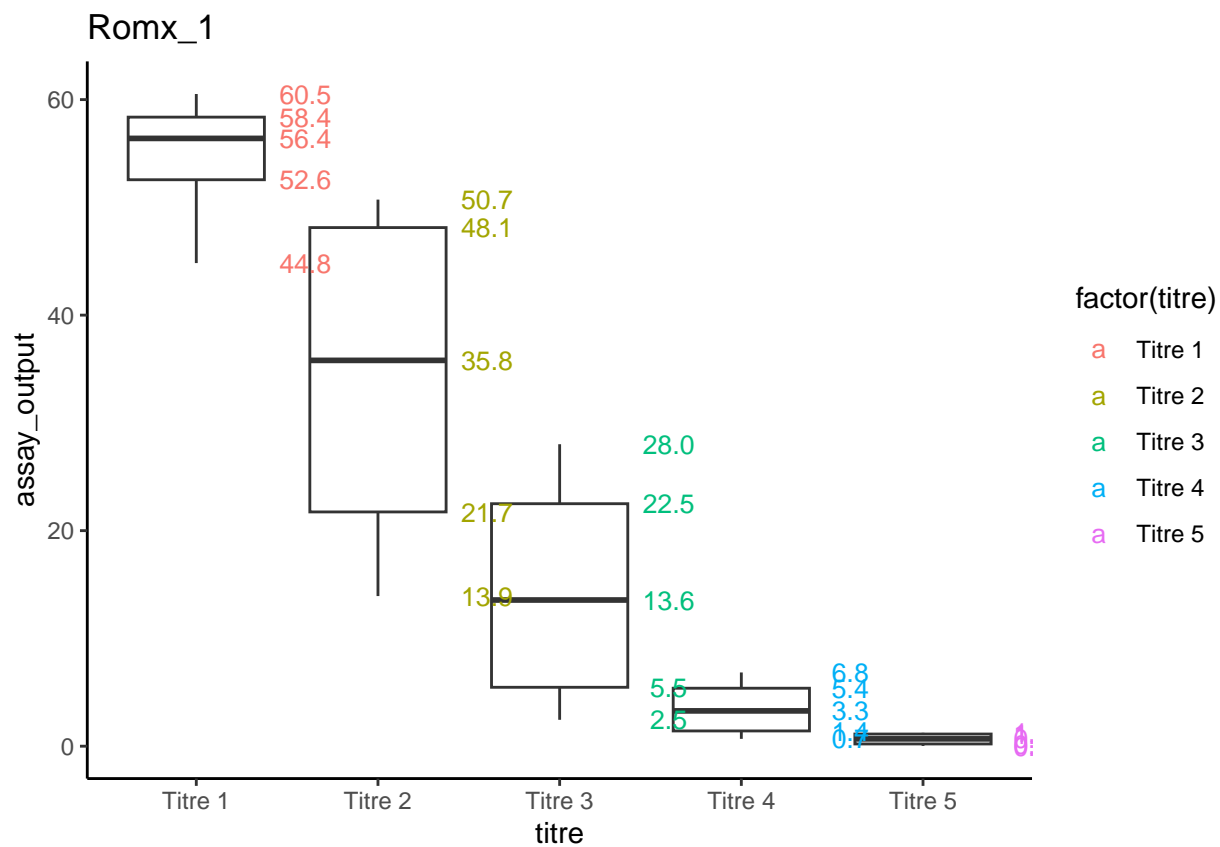
```
##
## [[100]]
```



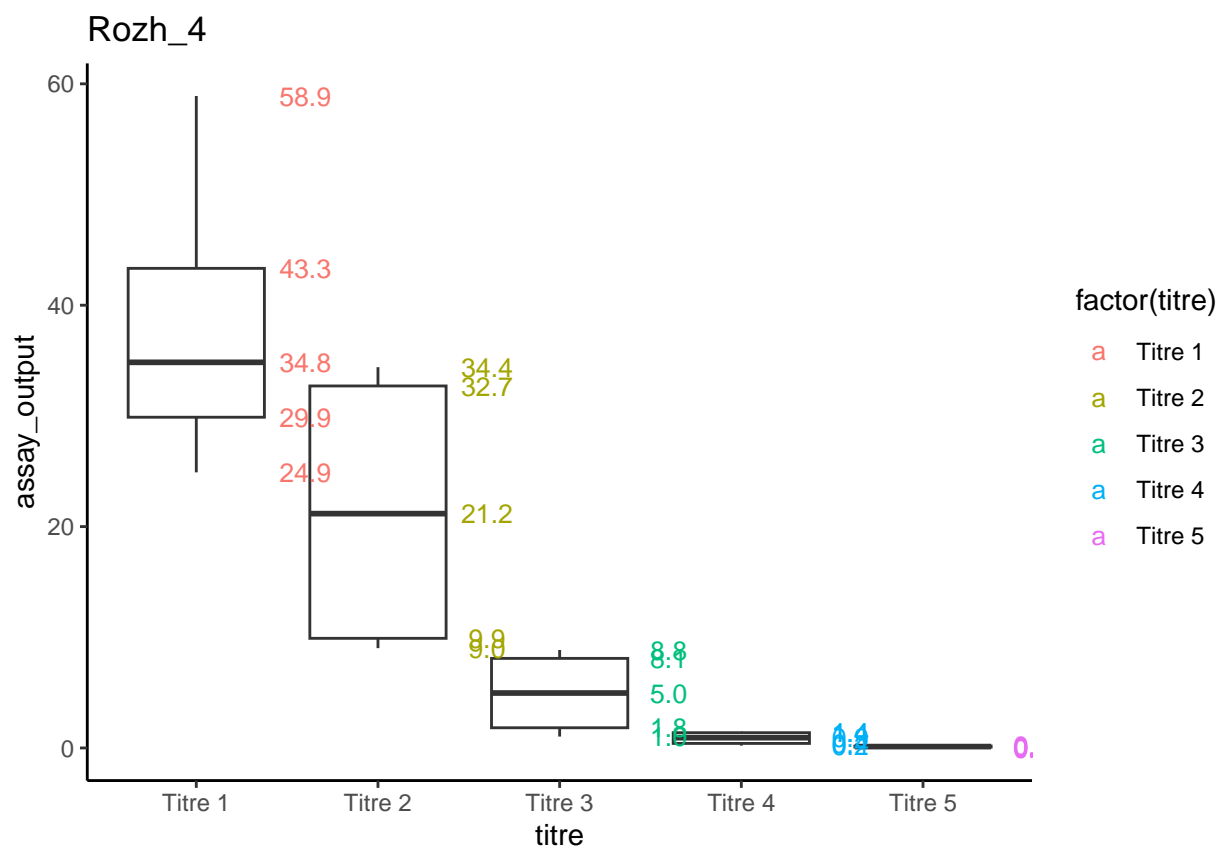
```
##
## [[101]]
```



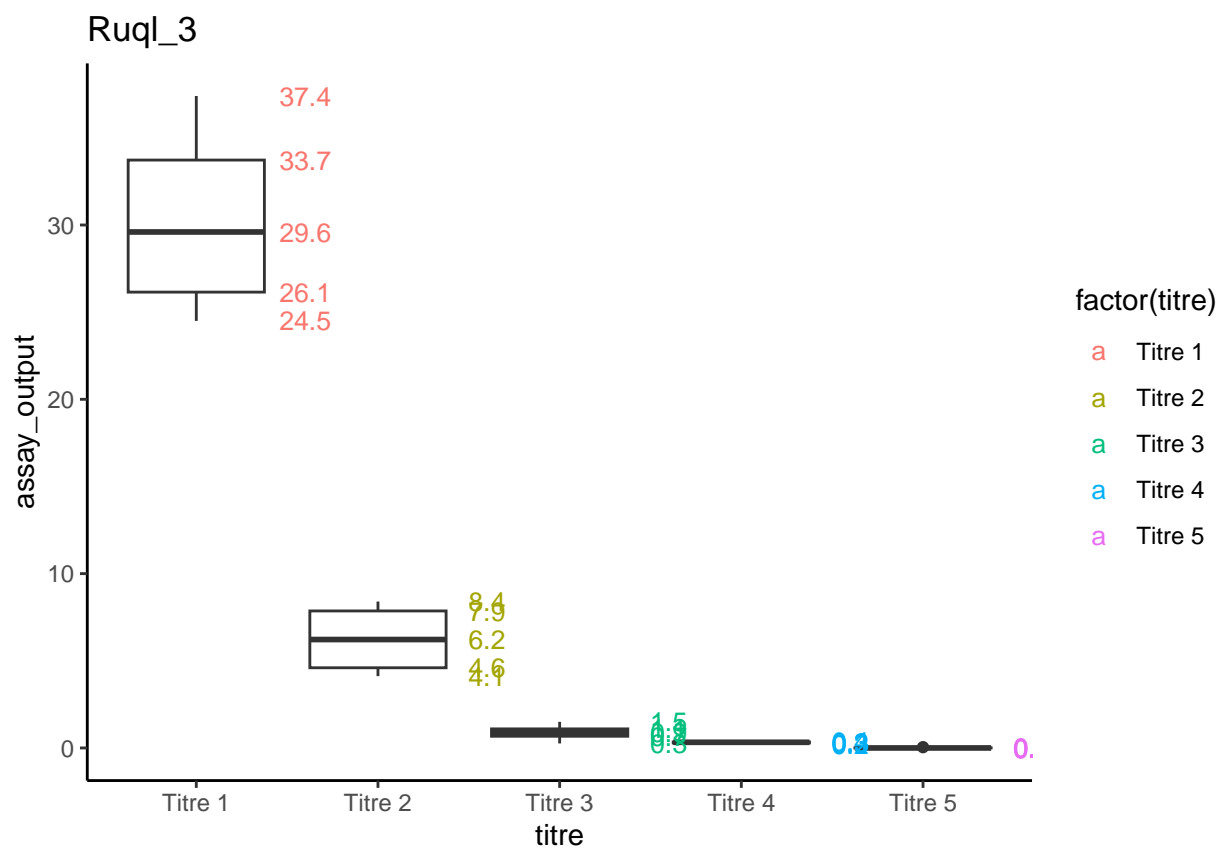
```
##
## [[102]]
```

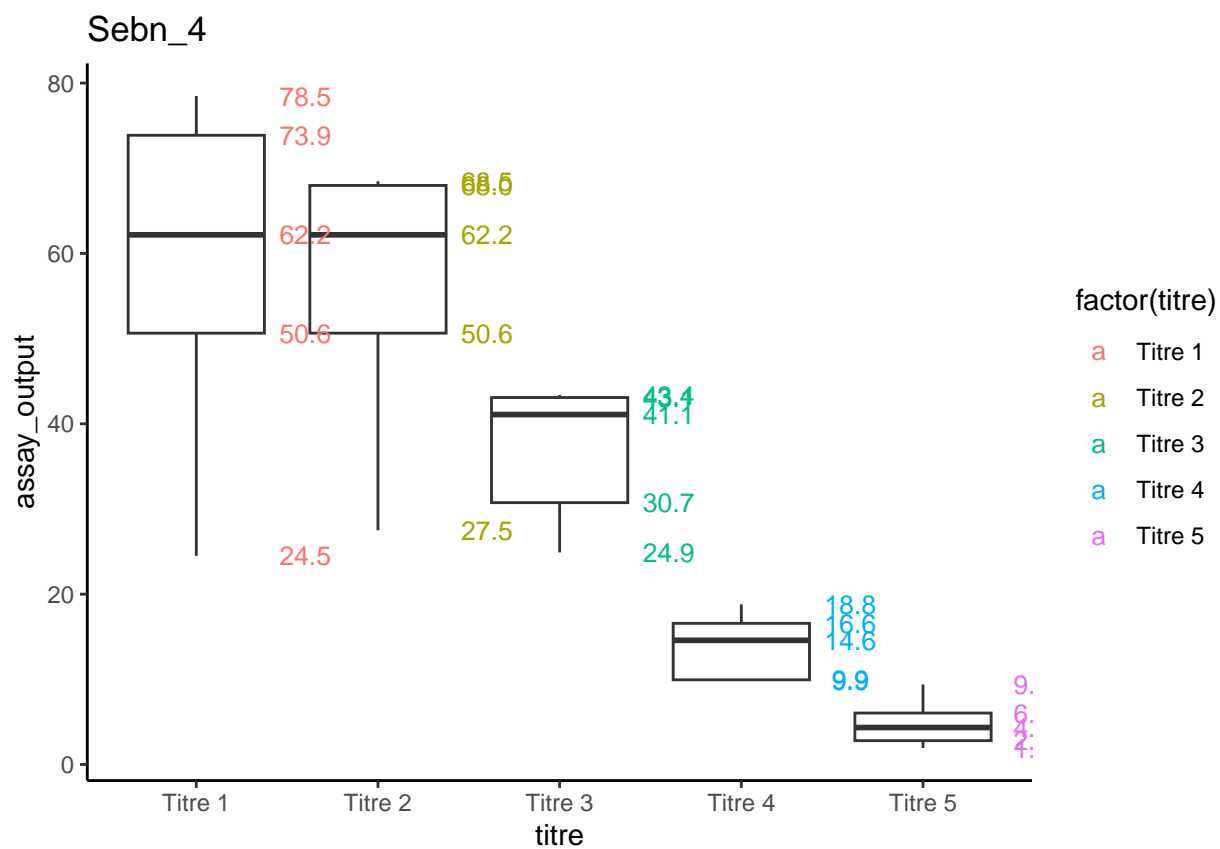
```
##
## [[103]]
```



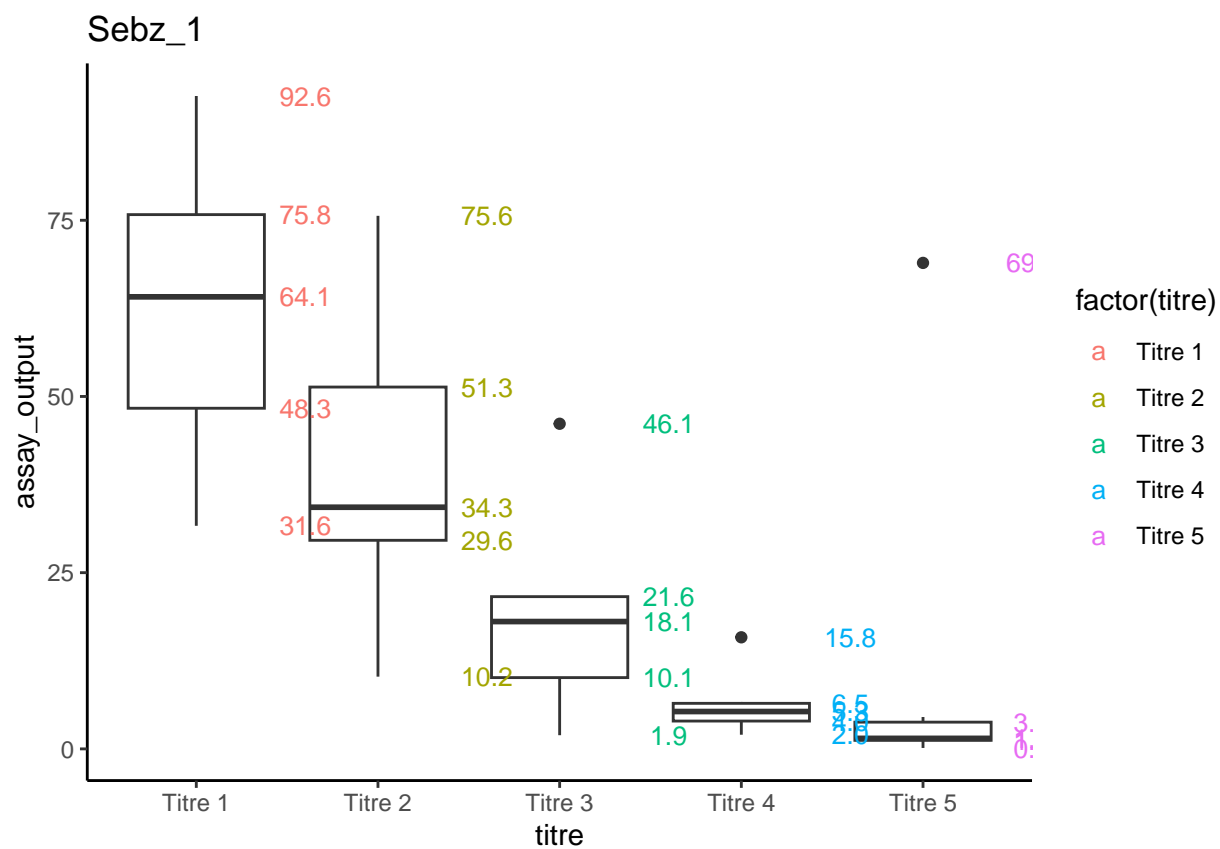
```
##
## [[104]]
```



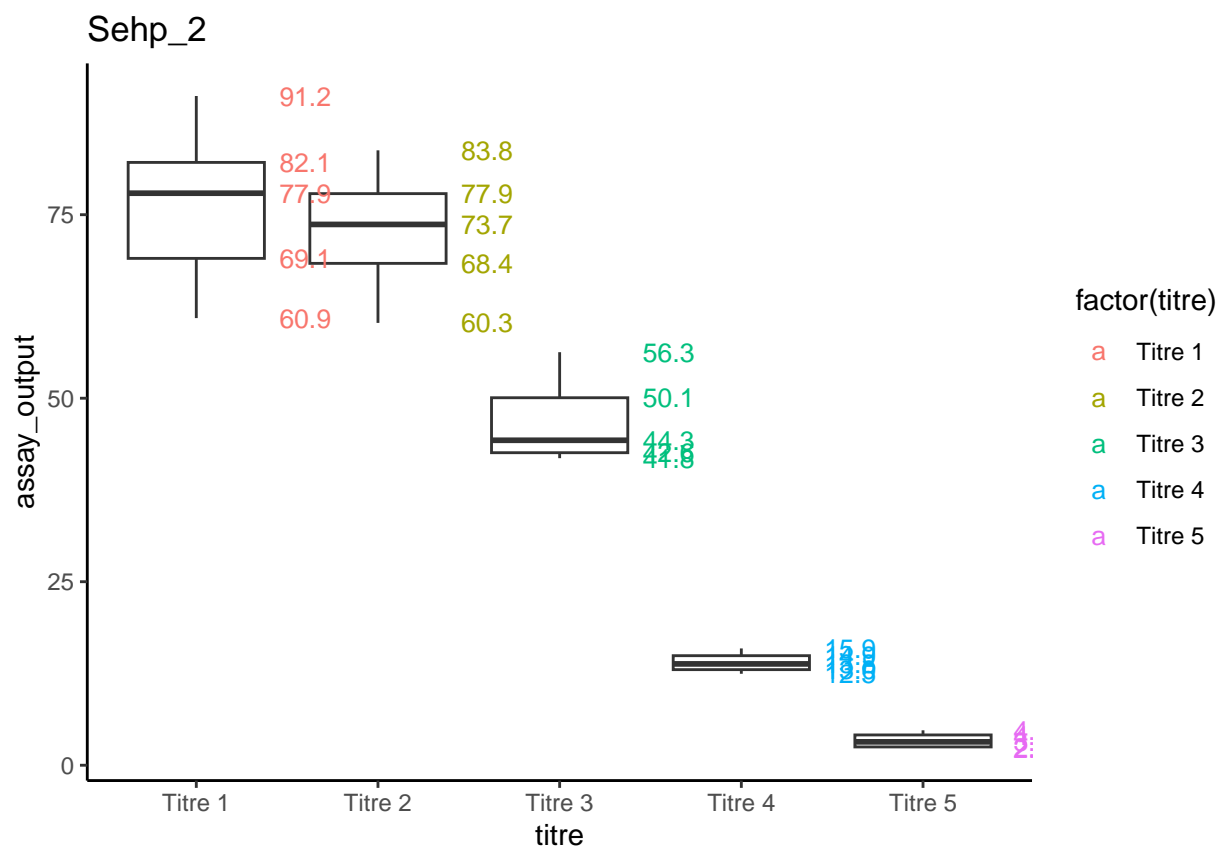
```
##
## [[105]]
```



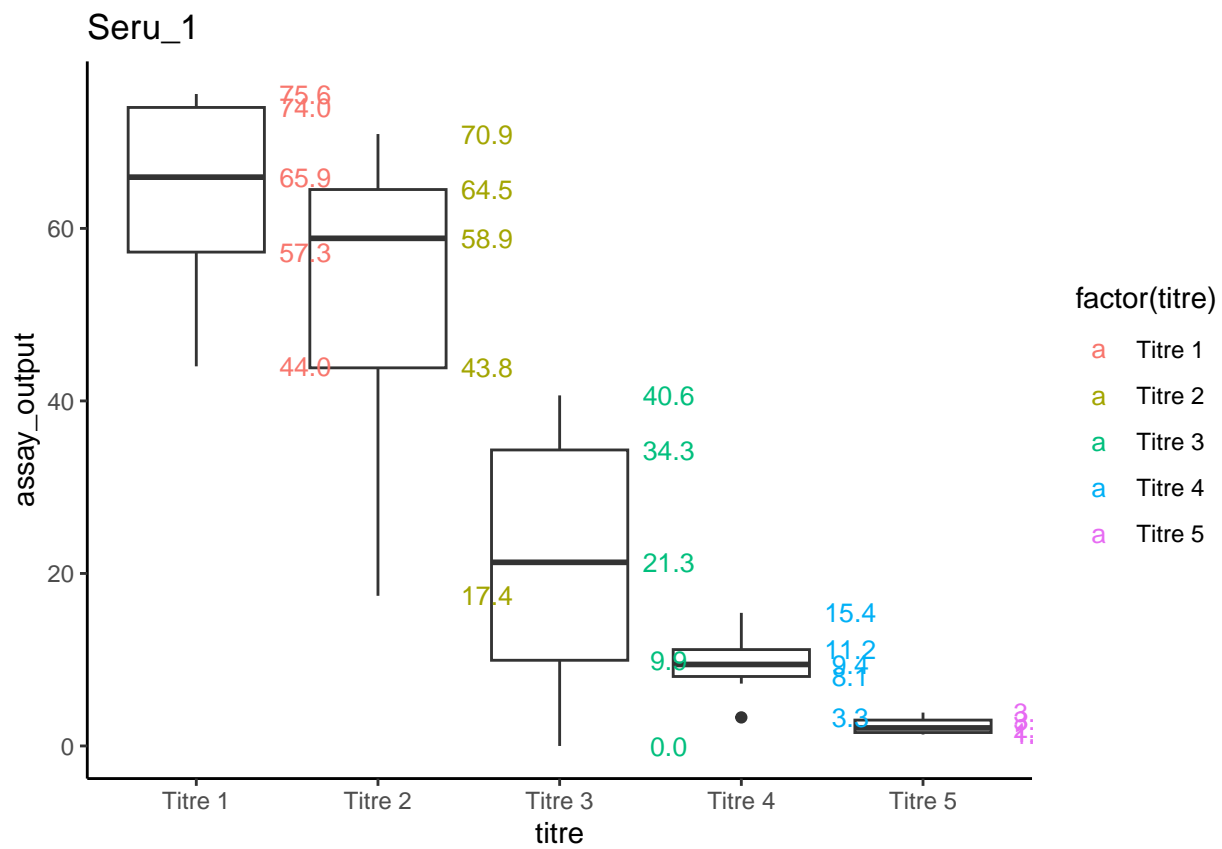
```
##  
## [[106]]
```



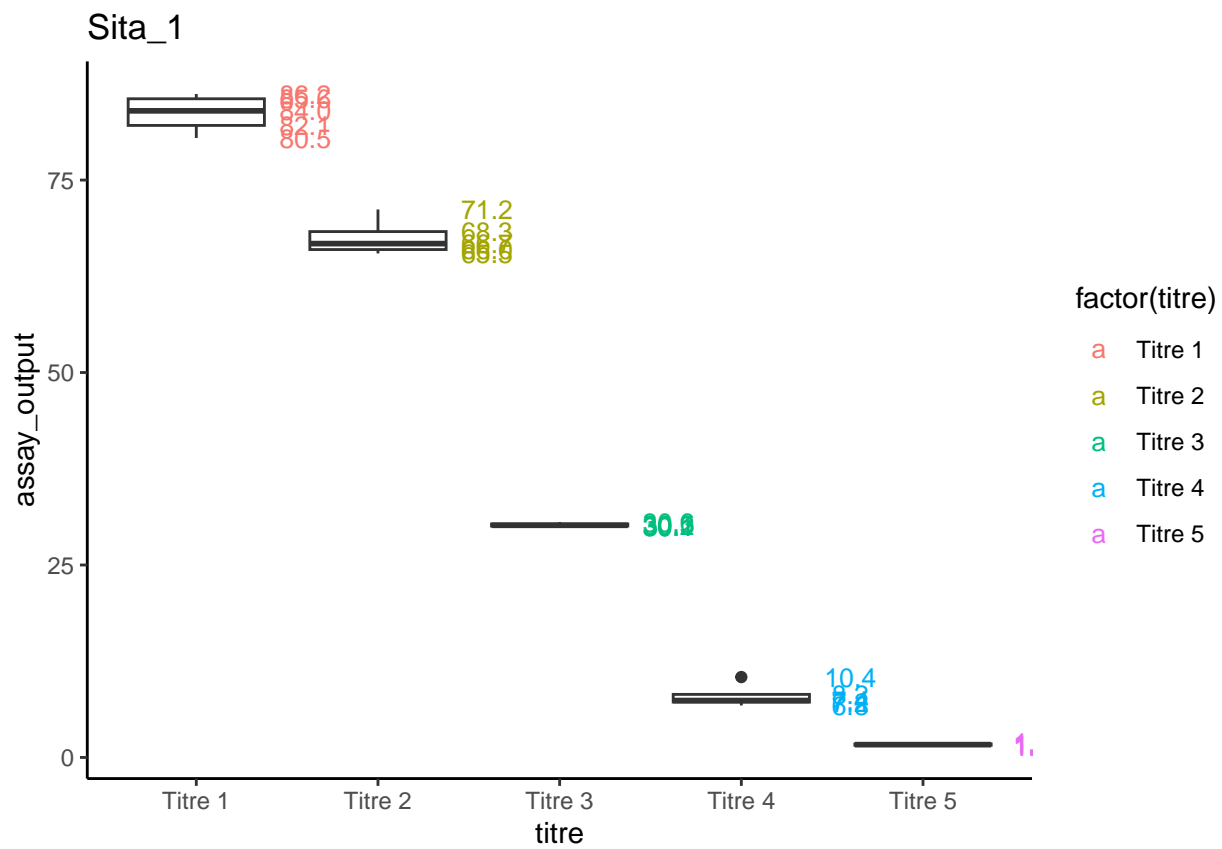
```
##
## [[107]]
```



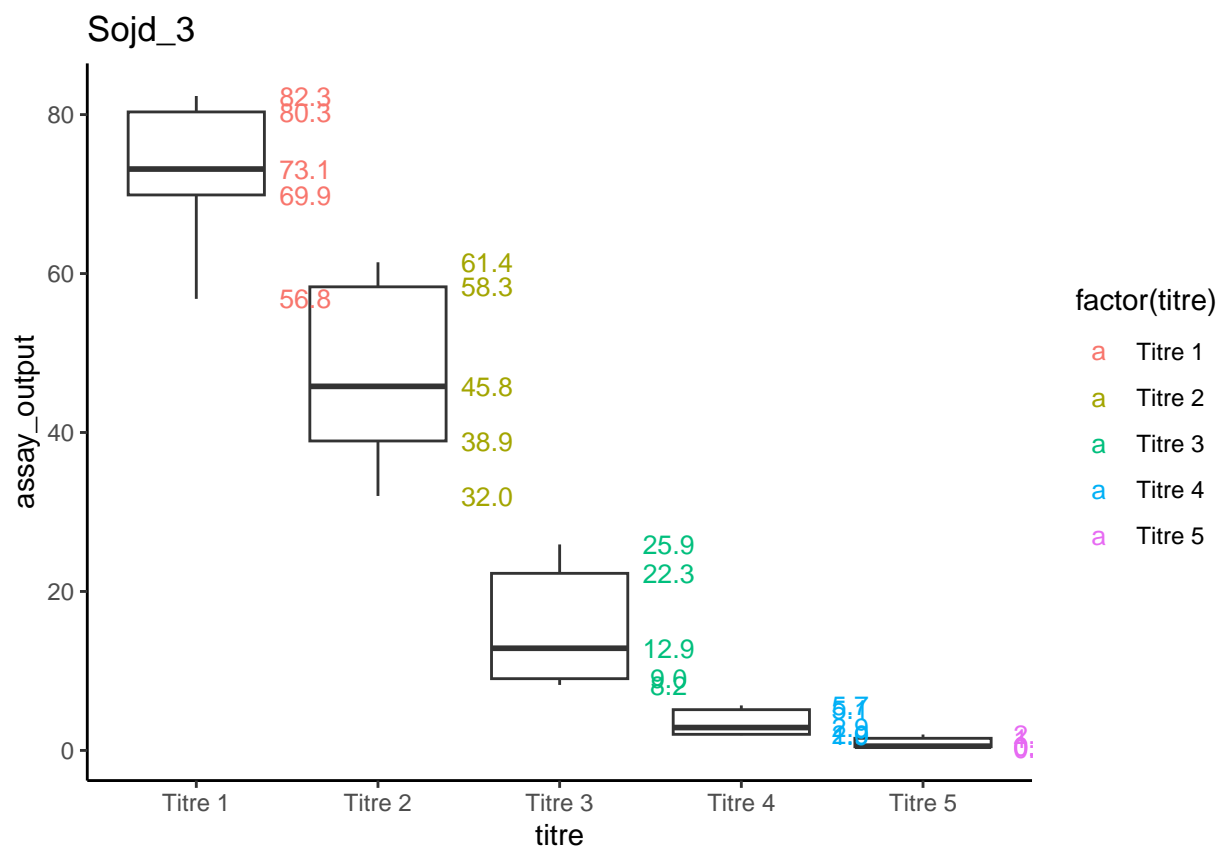
```
##
## [[108]]
```



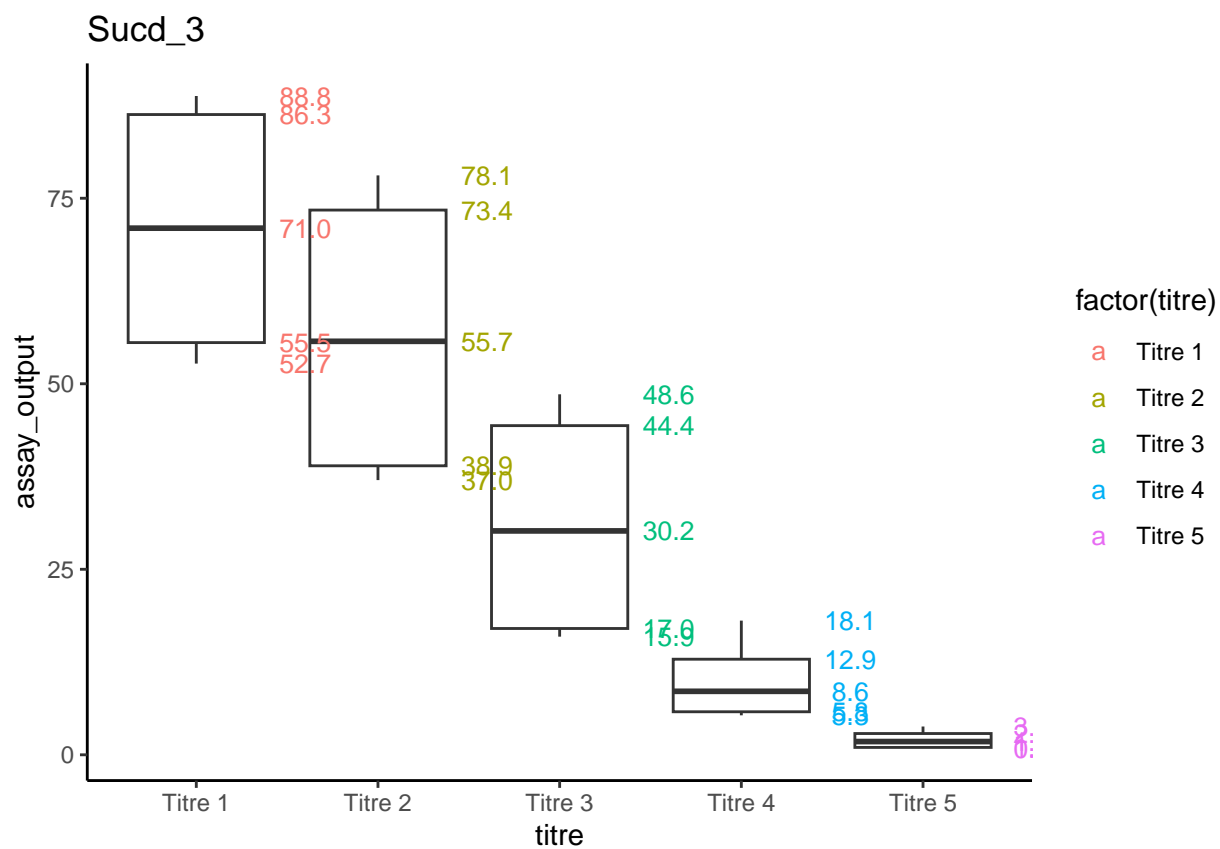
[[109]]



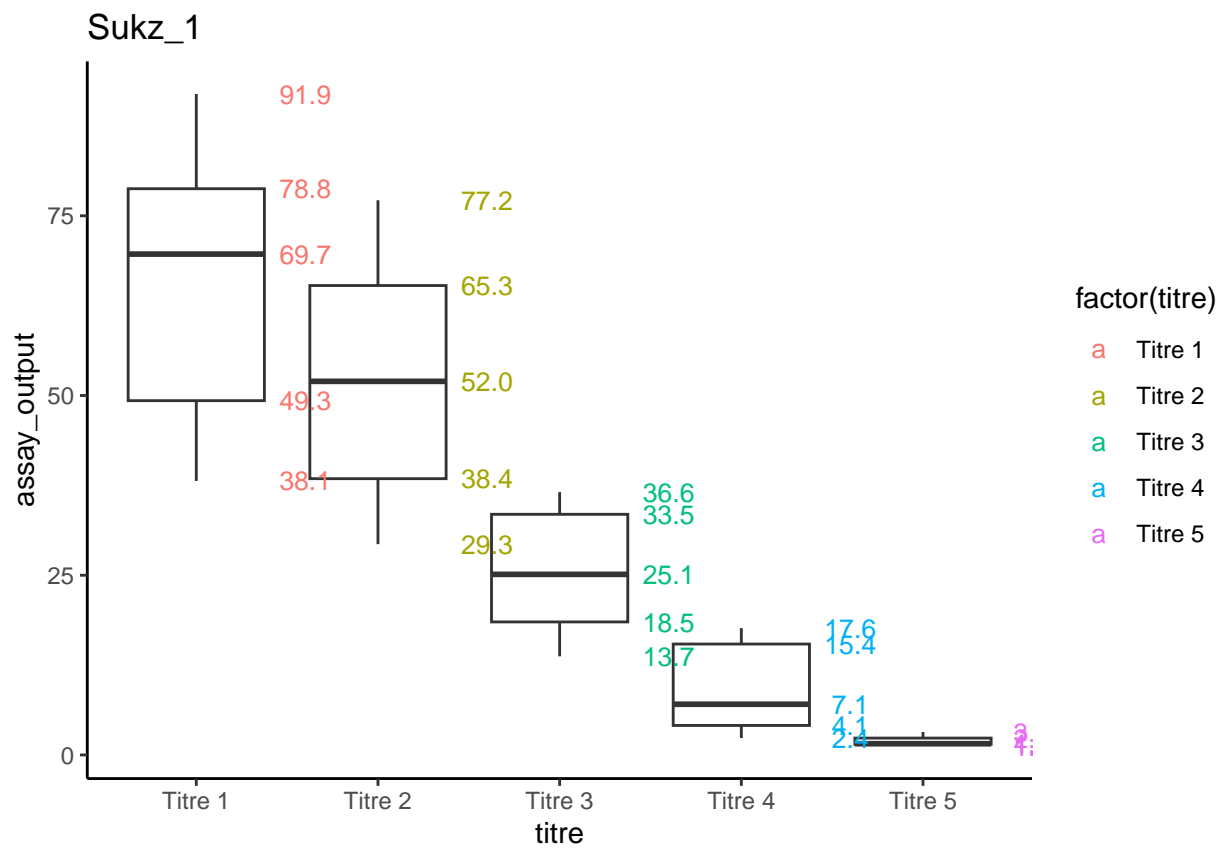
[[110]]



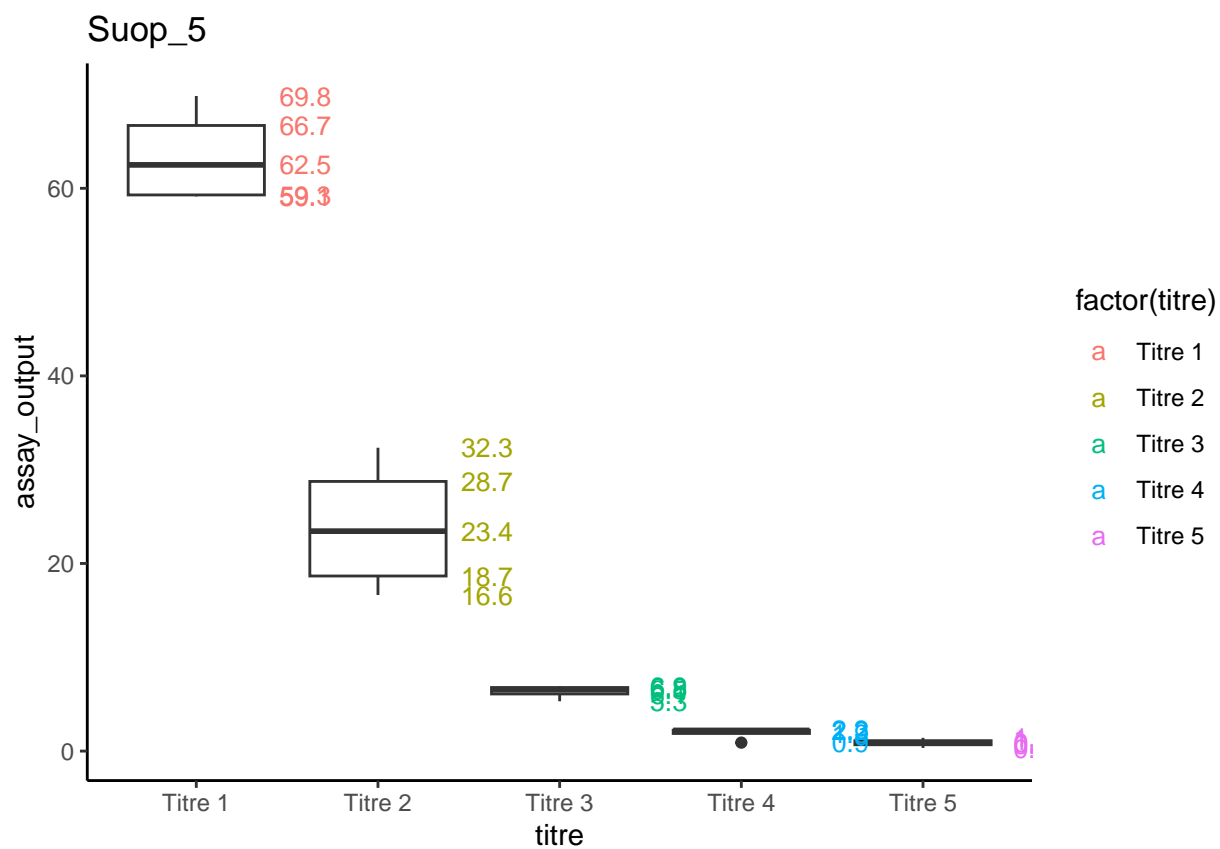
```
##
## [[111]]
```



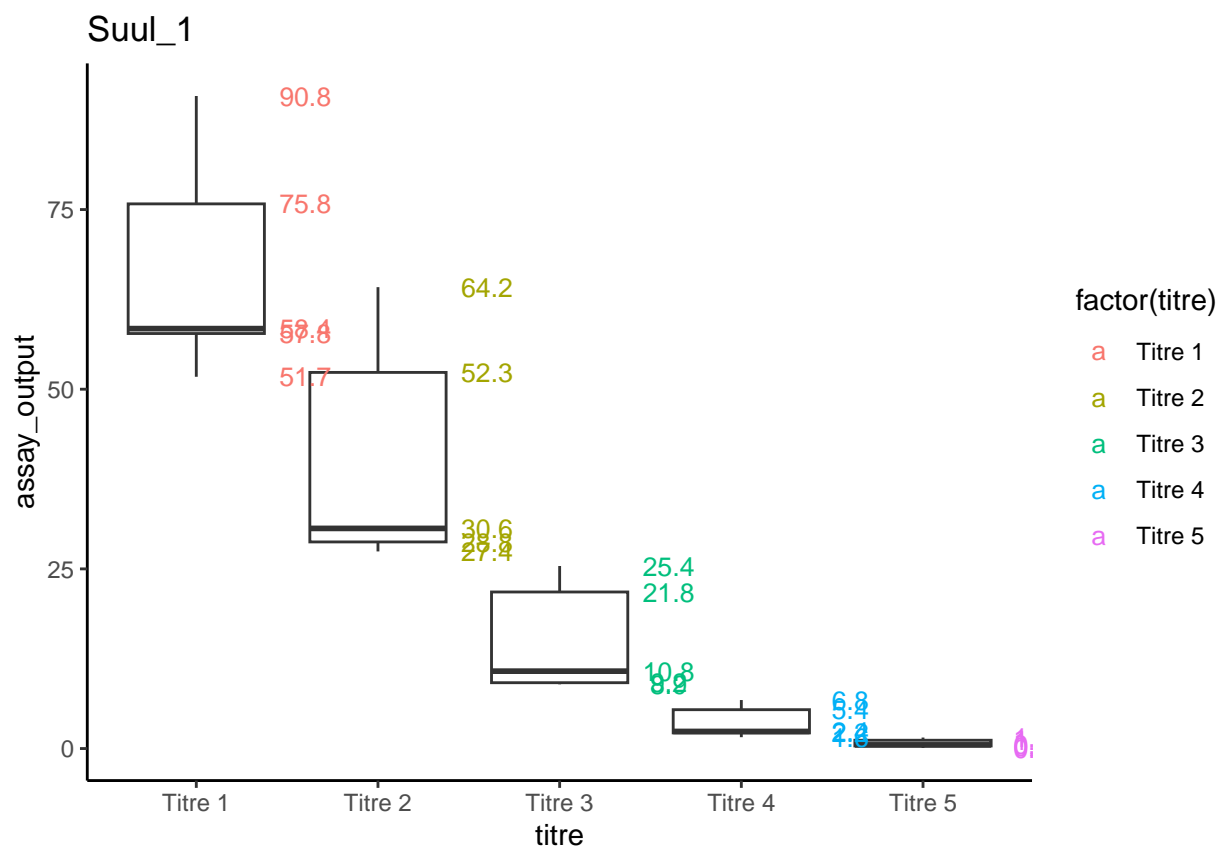
```
##
## [[112]]
```



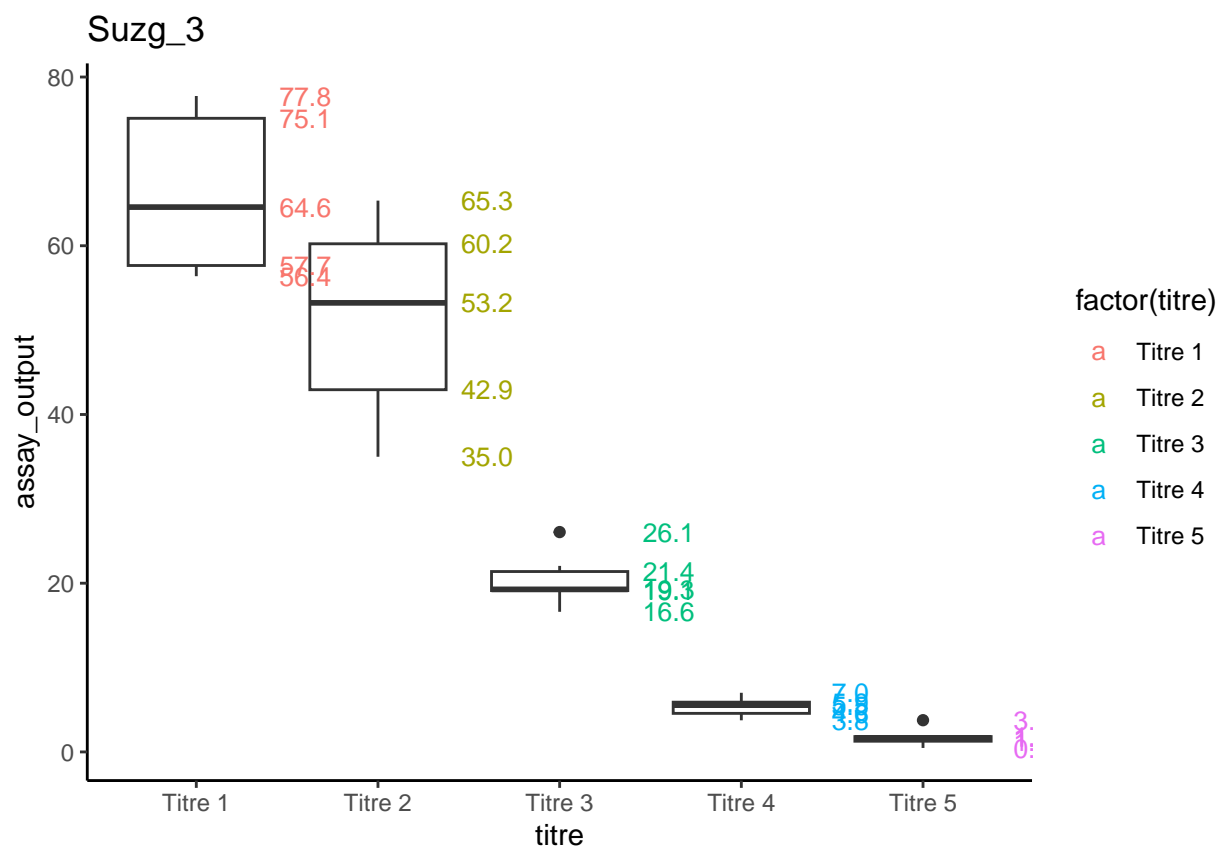
```
##
## [[113]]
```



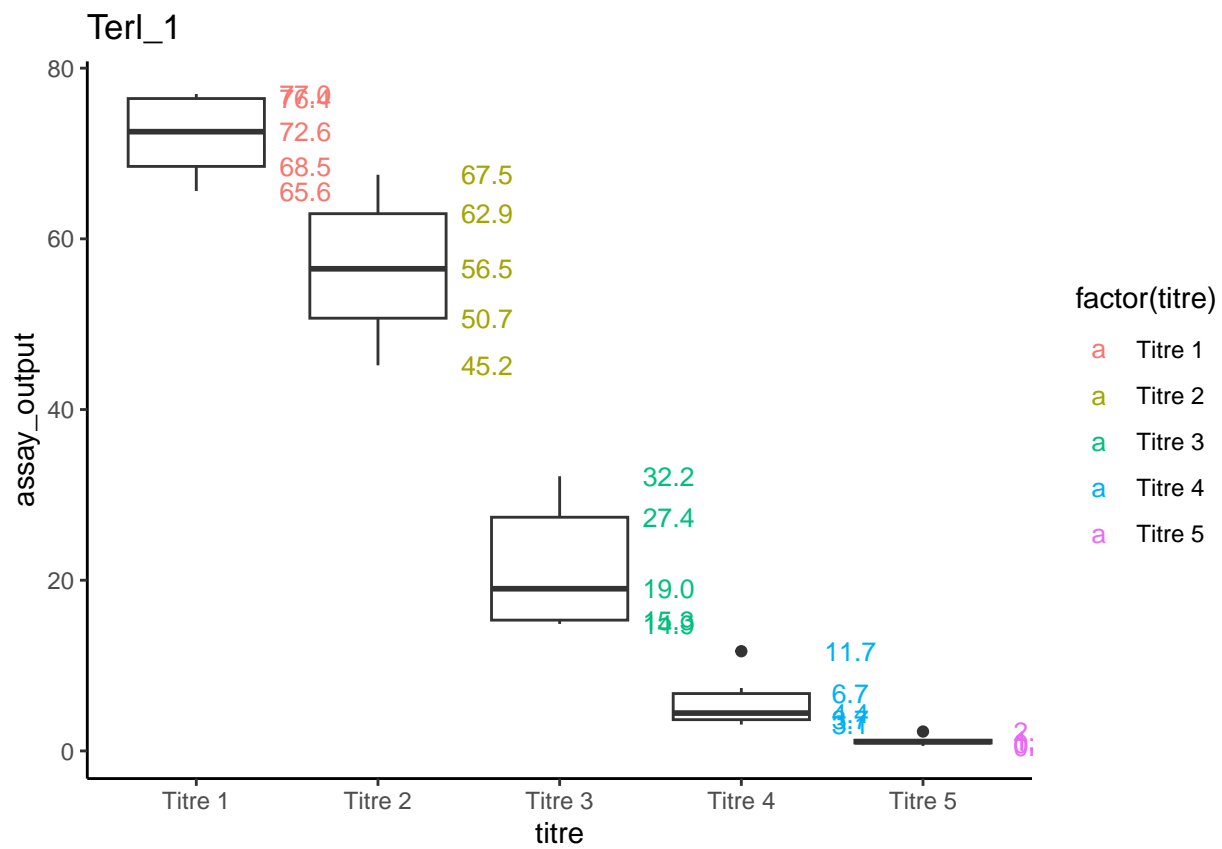
```
##
## [[114]]
```



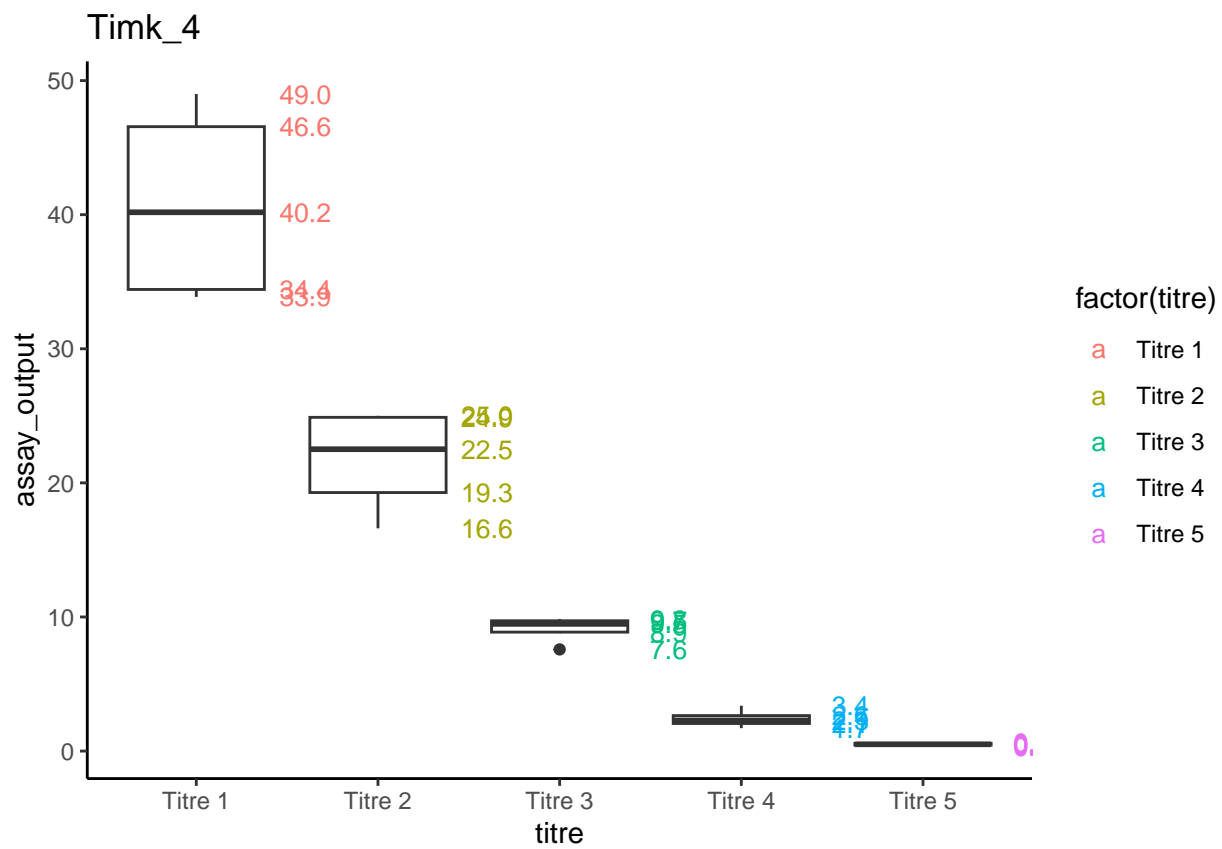
```
##
## [[115]]
```



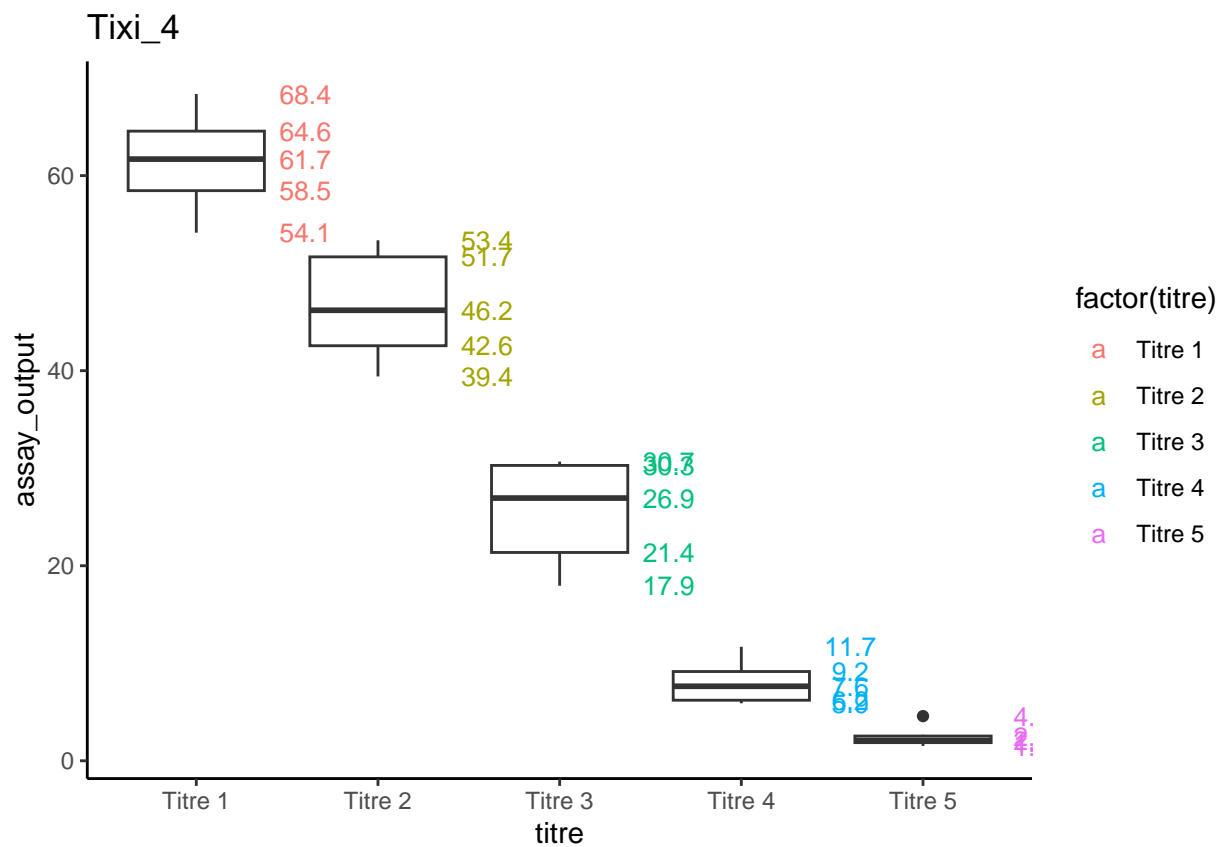
```
##
## [[116]]
```



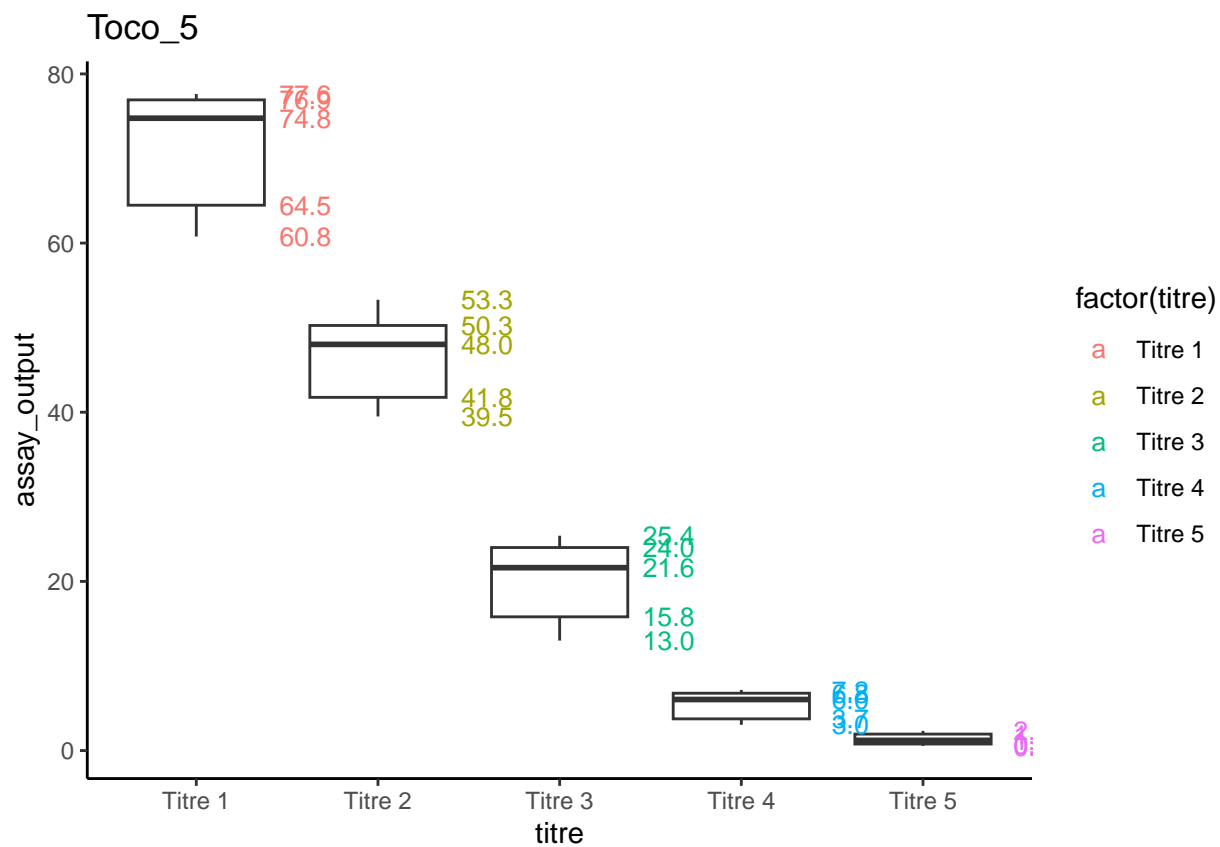
```
##
## [[117]]
```



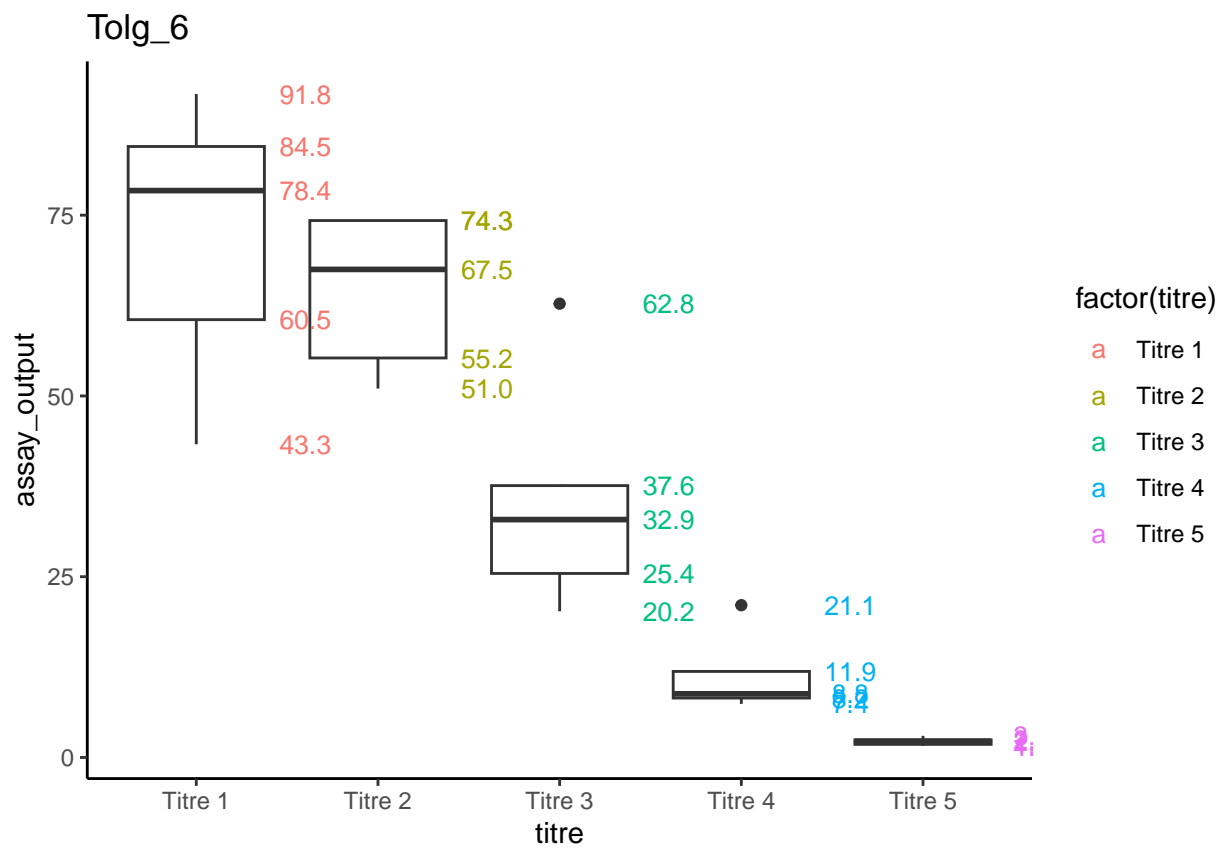
```
##
## [[118]]
```

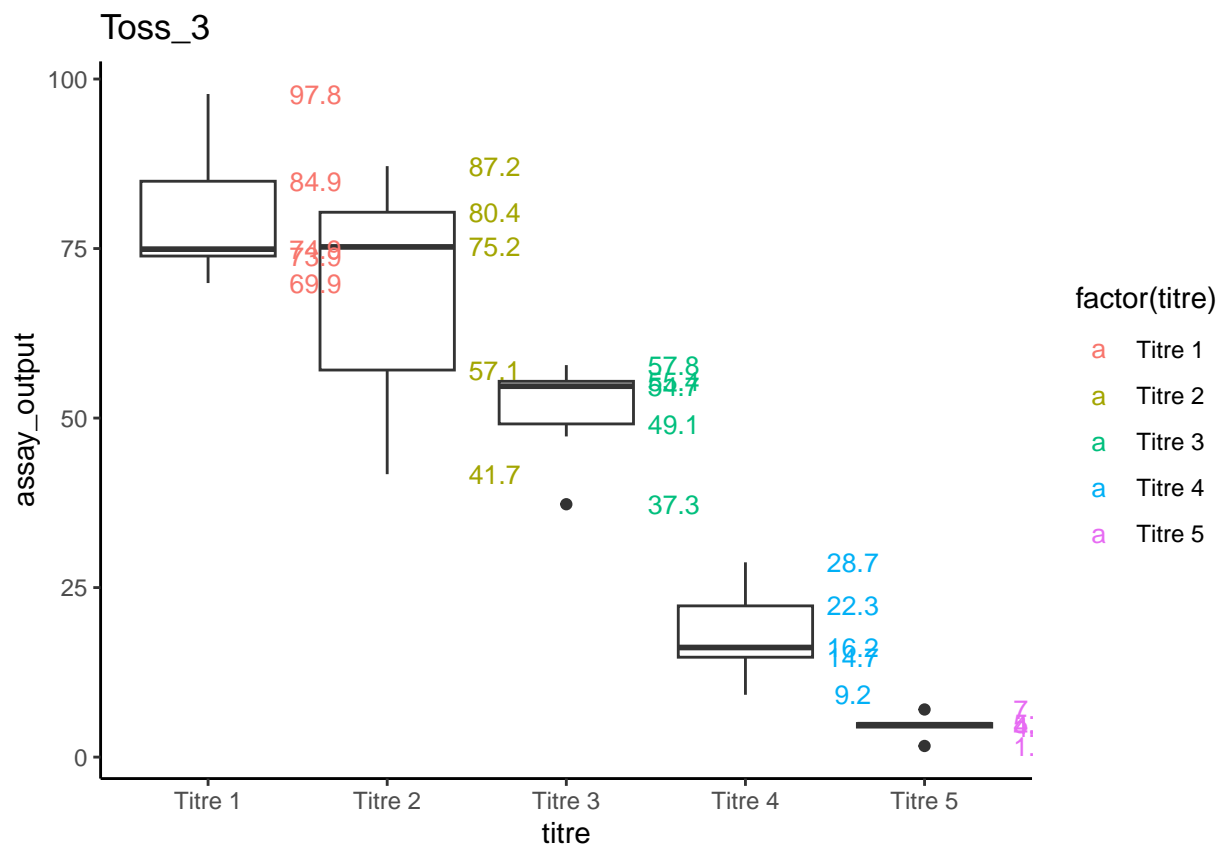
```
##
## [[119]]
```



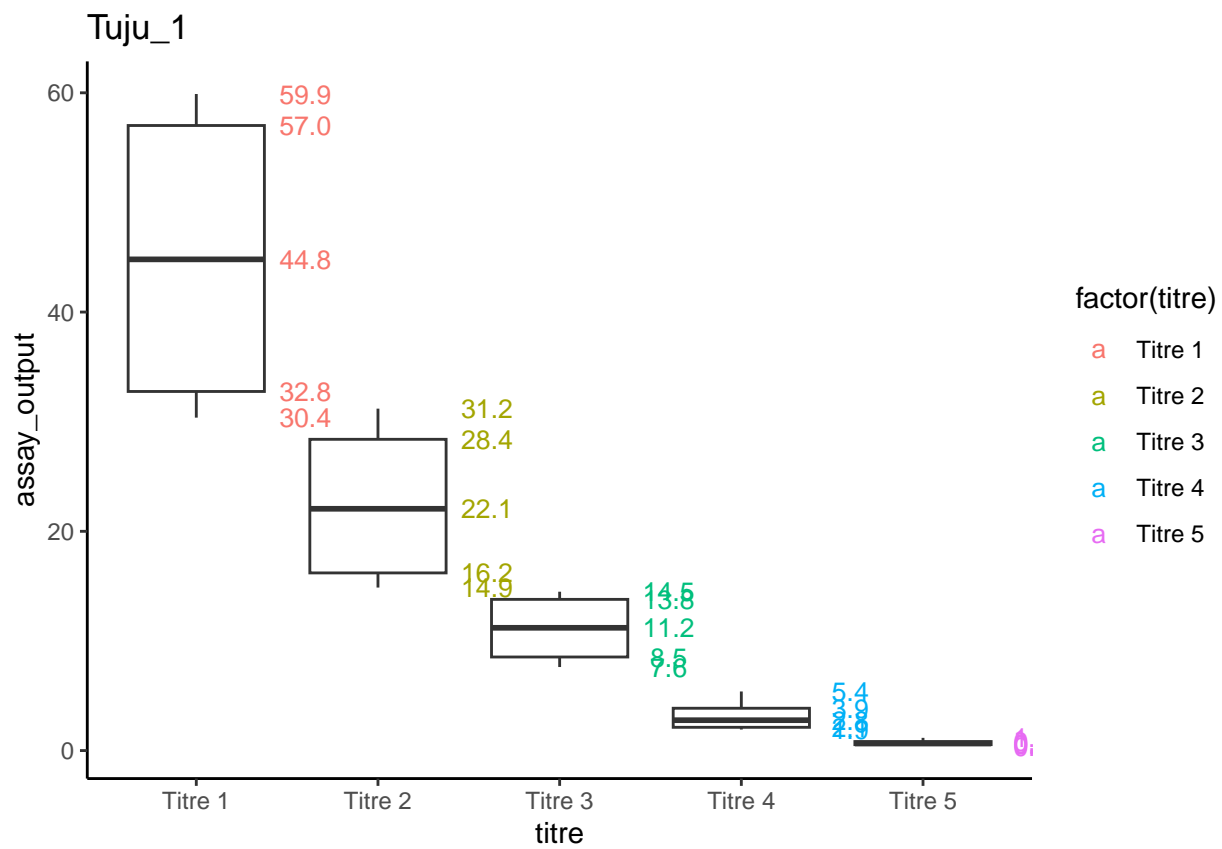
```
##
## [[120]]
```



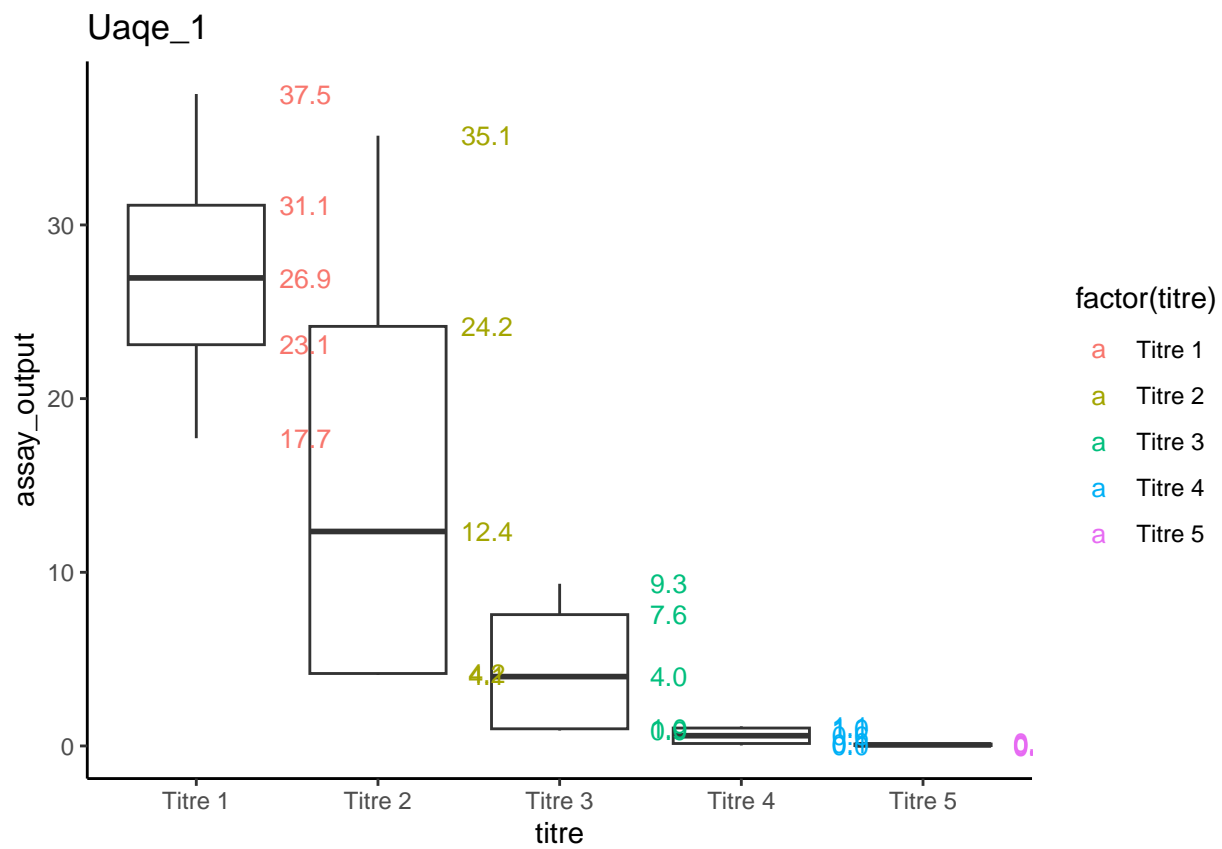
```
##
## [[121]]
```



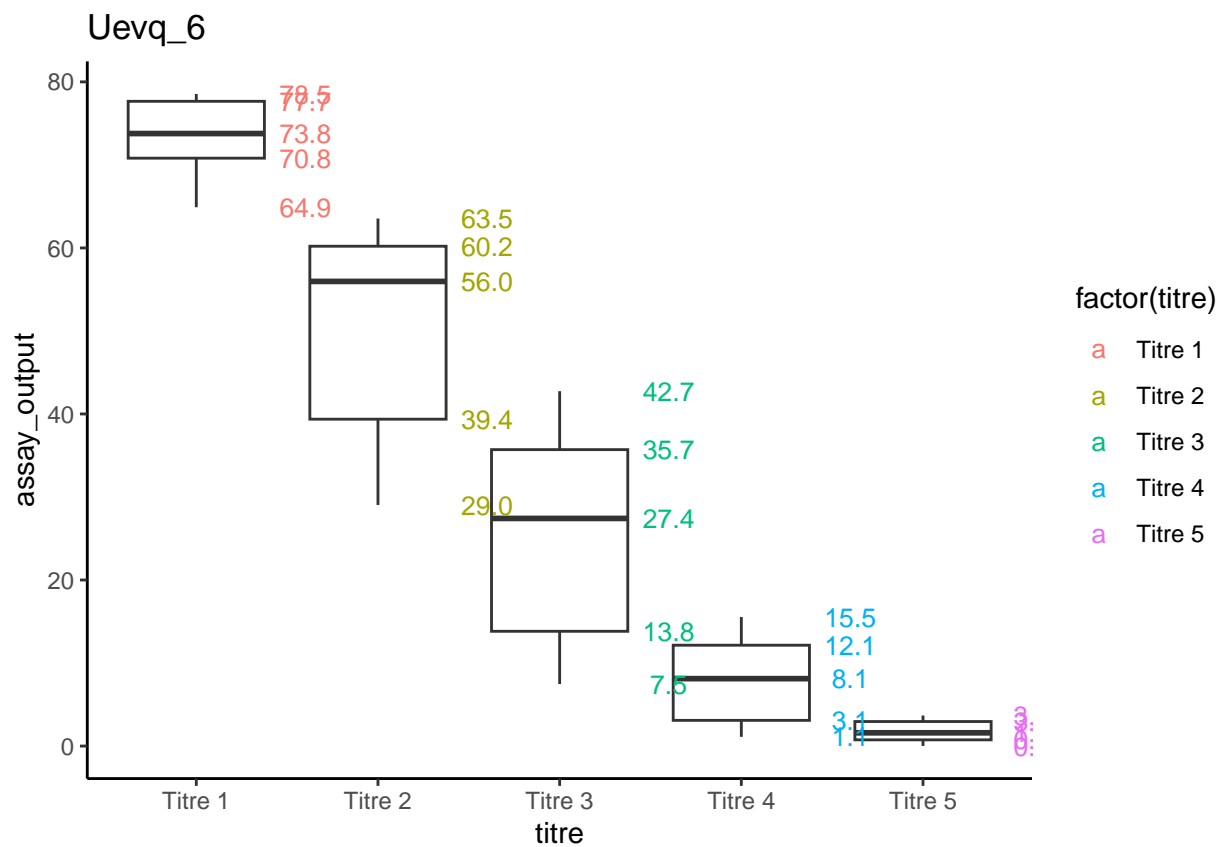
[[122]]



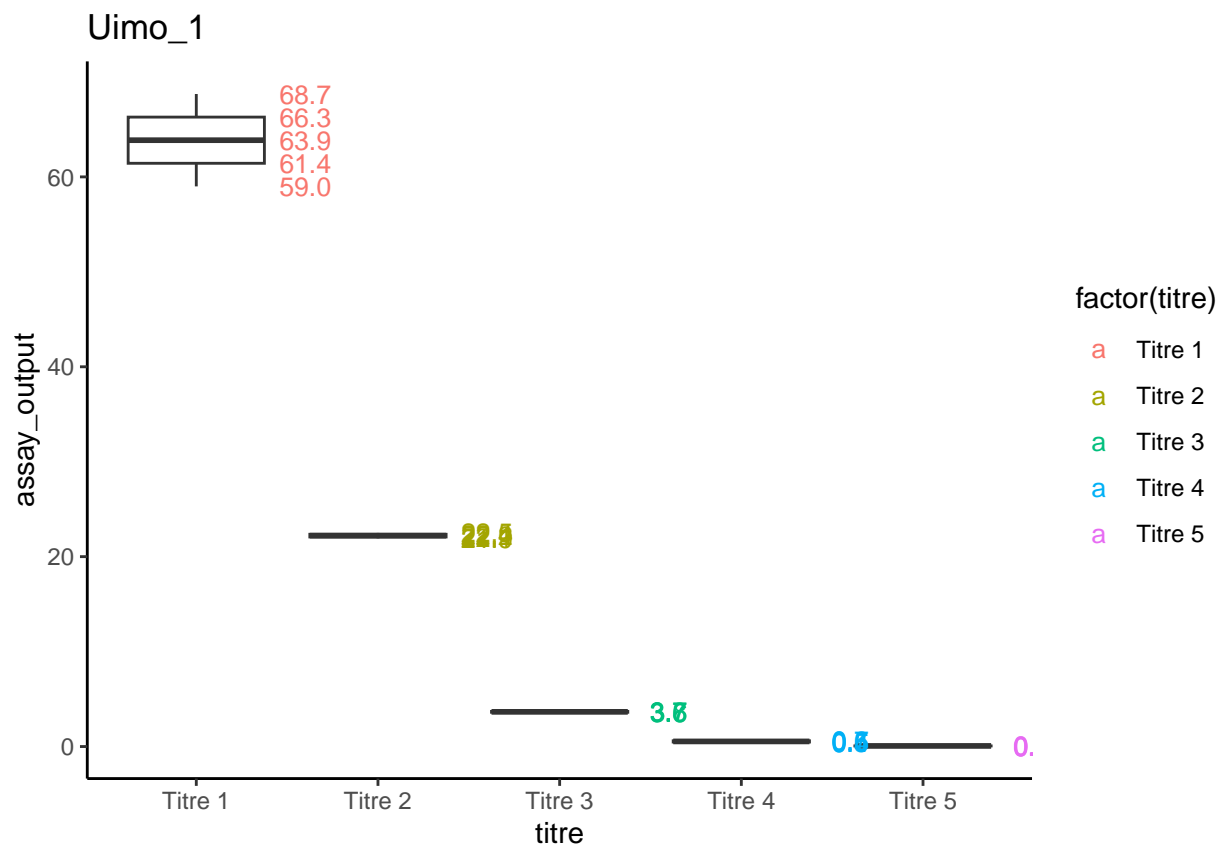
```
##
## [[123]]
```



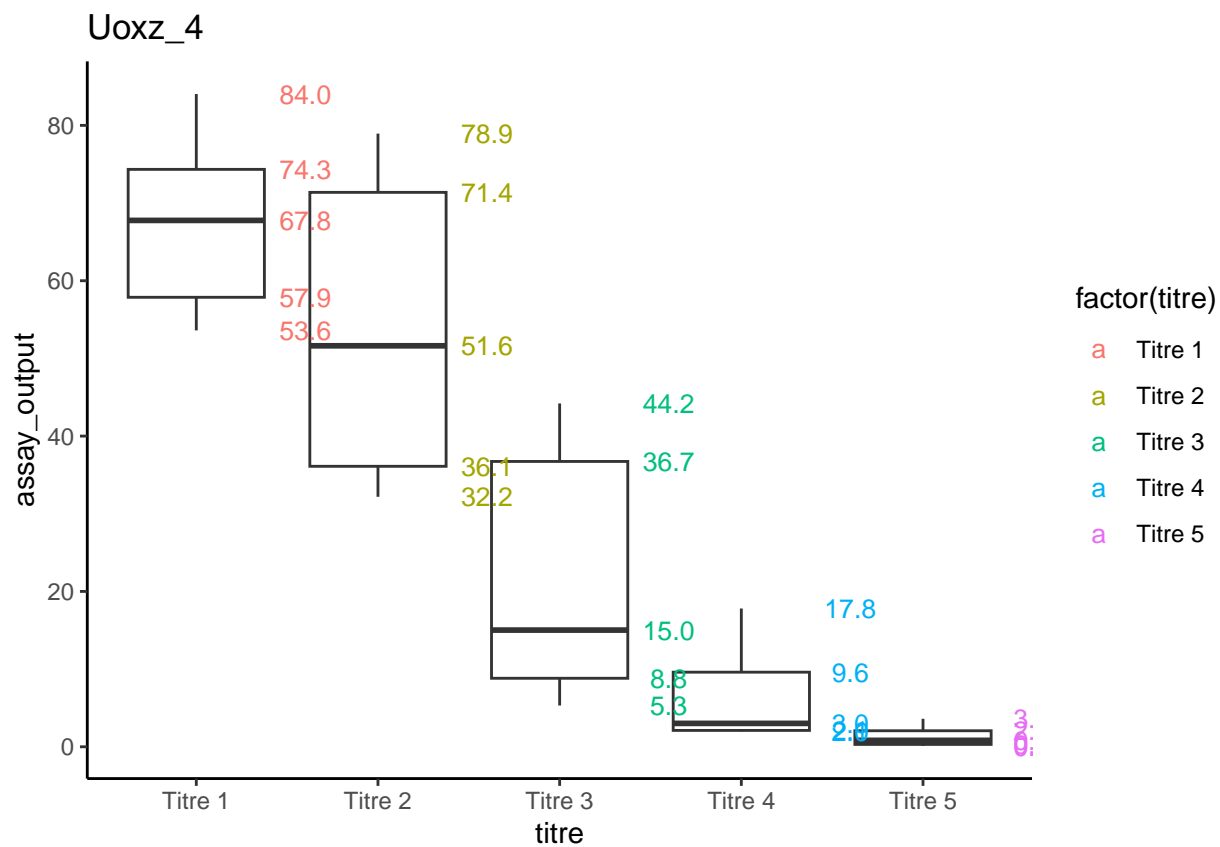
```
##
## [[124]]
```



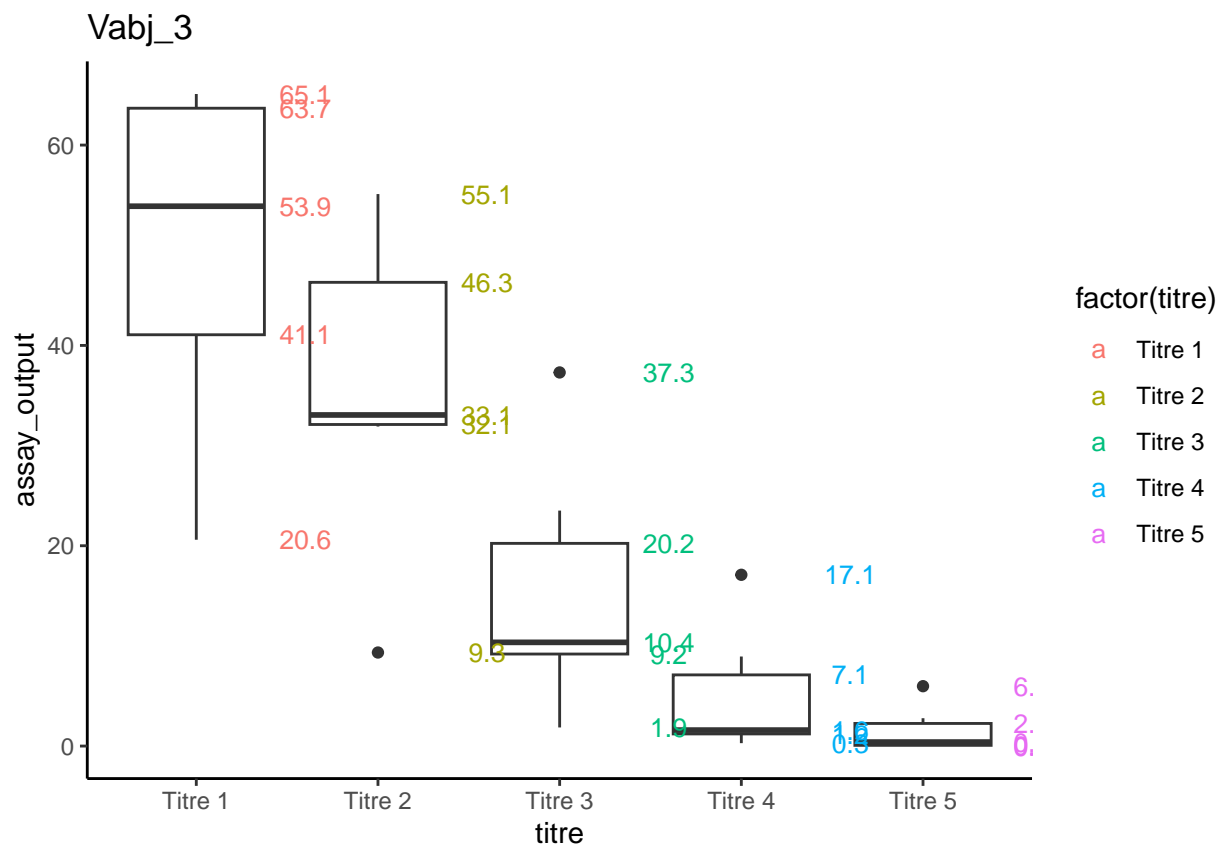
```
##
## [[125]]
```



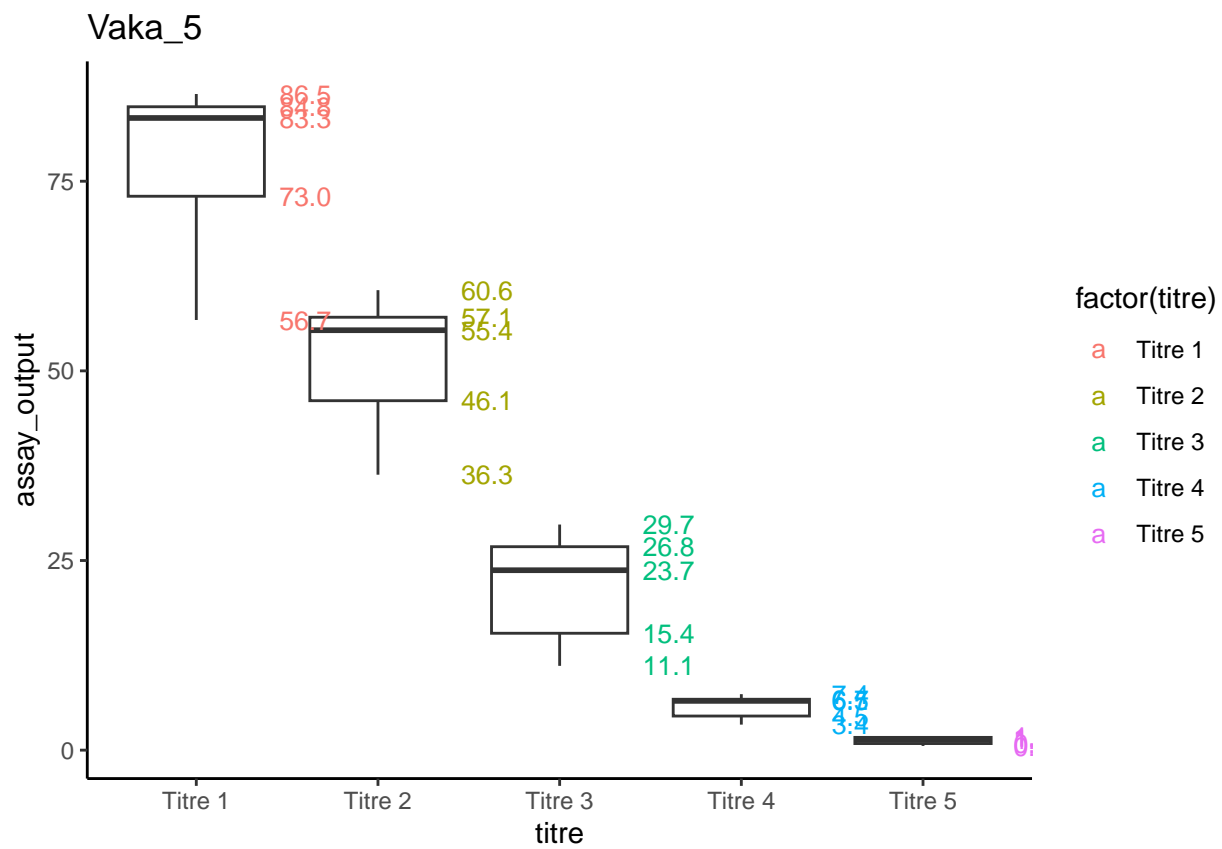
[[126]]



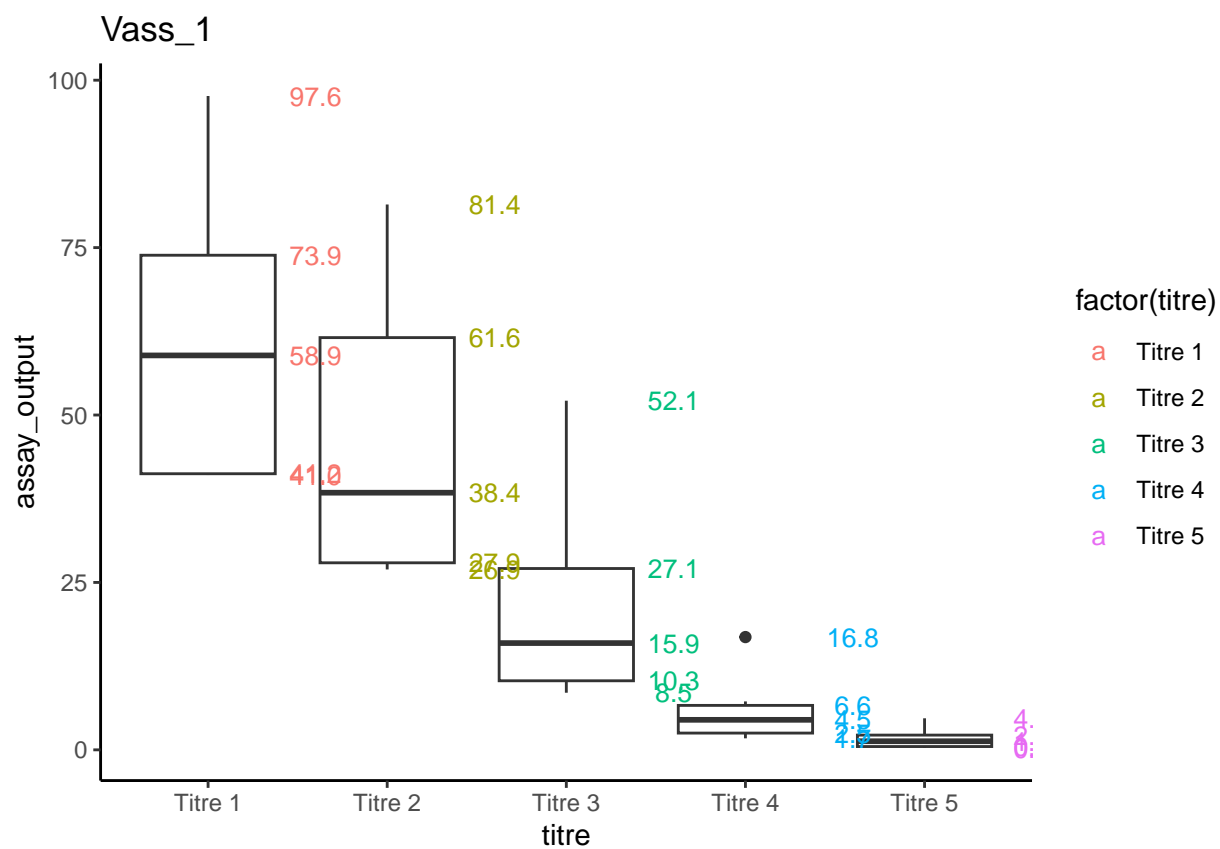
```
##
## [[127]]
```



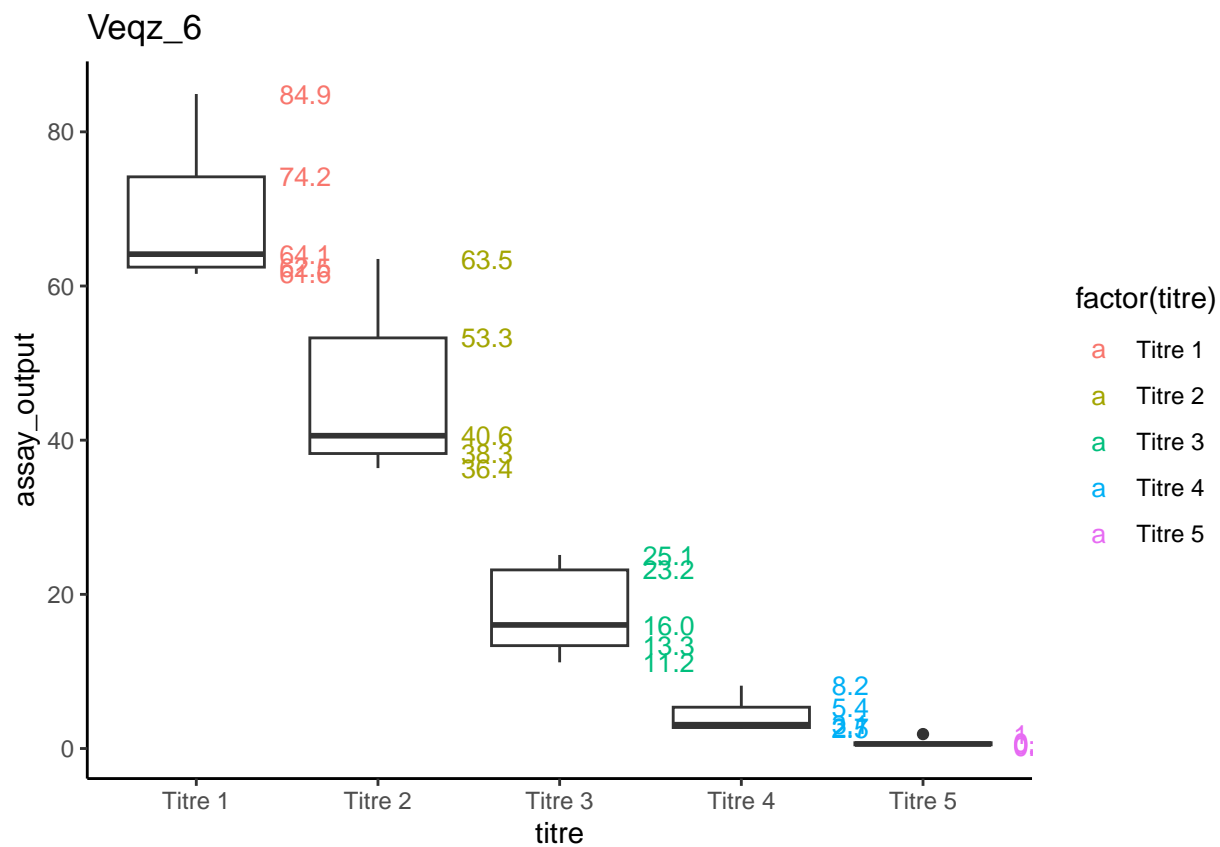
```
##
## [[128]]
```



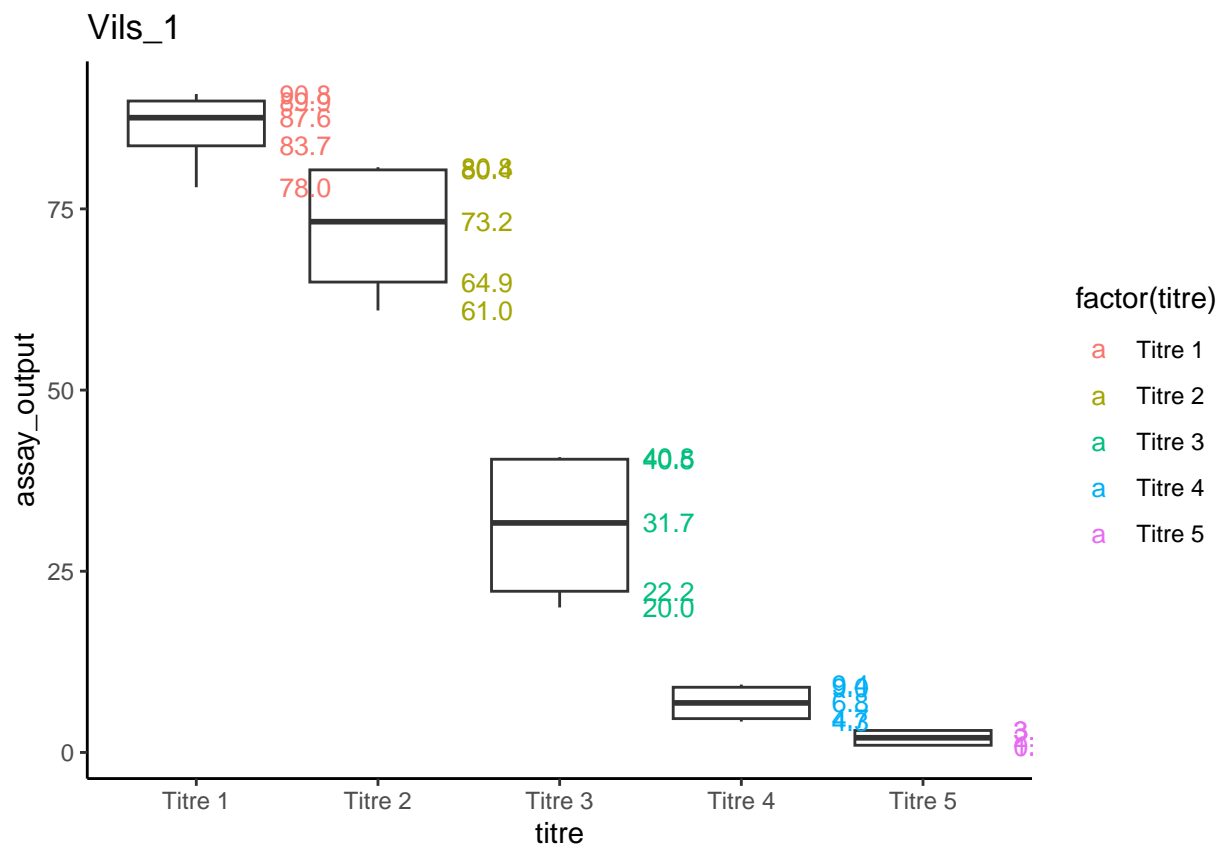
```
##
## [[129]]
```



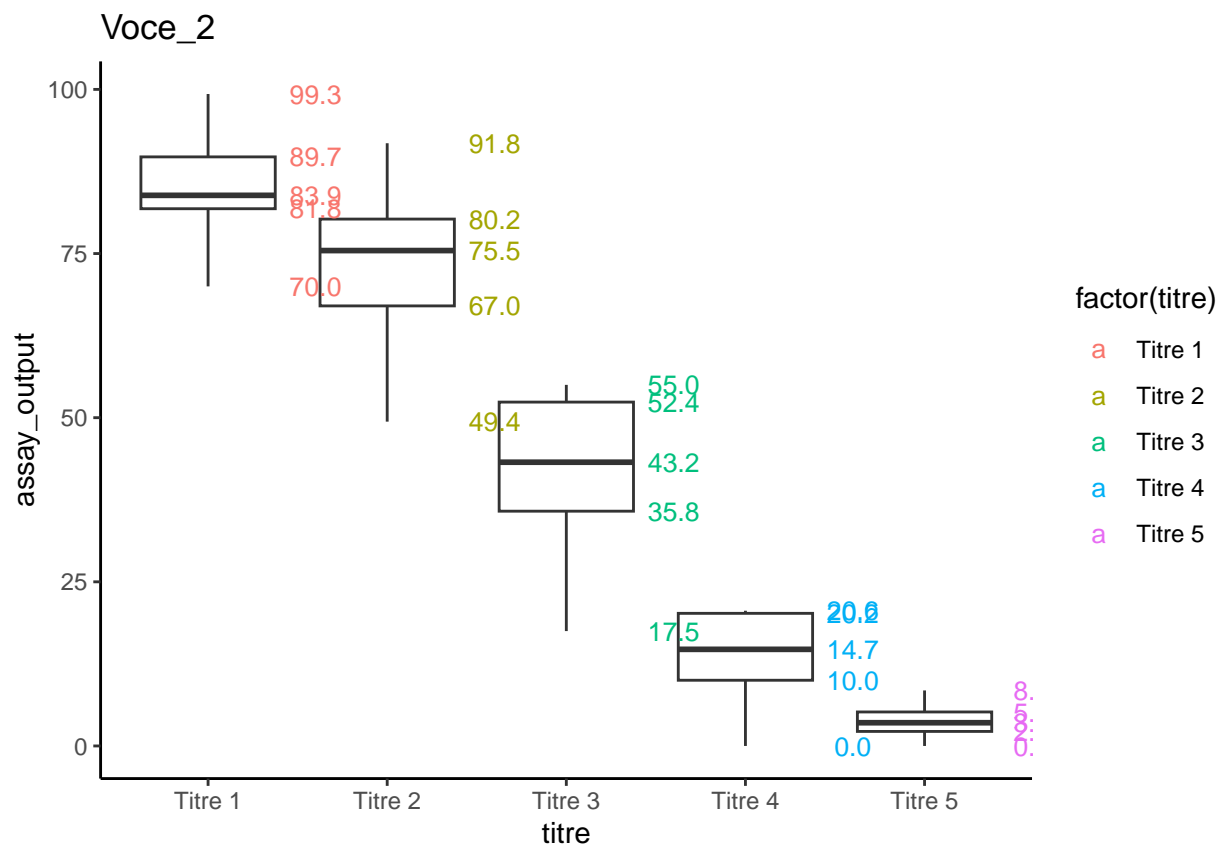
```
##
## [[130]]
```



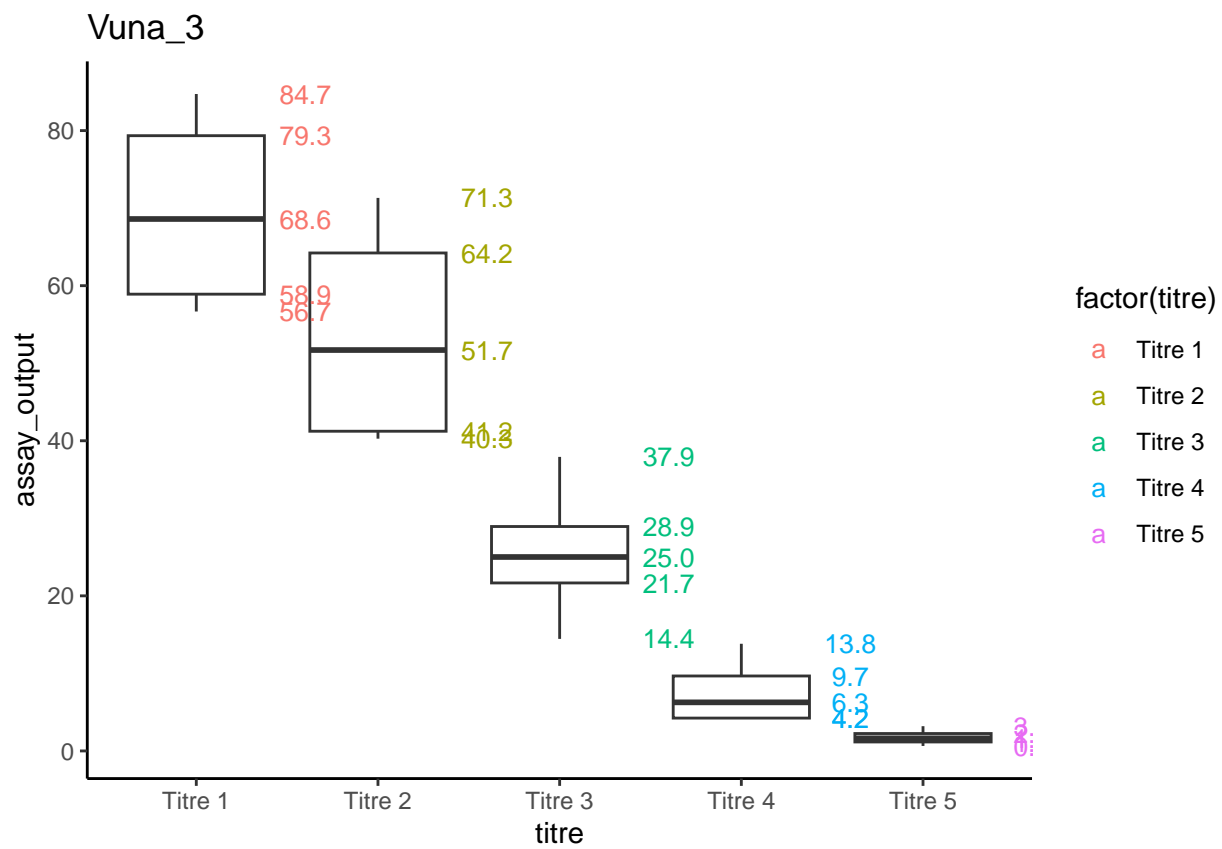
[[131]]



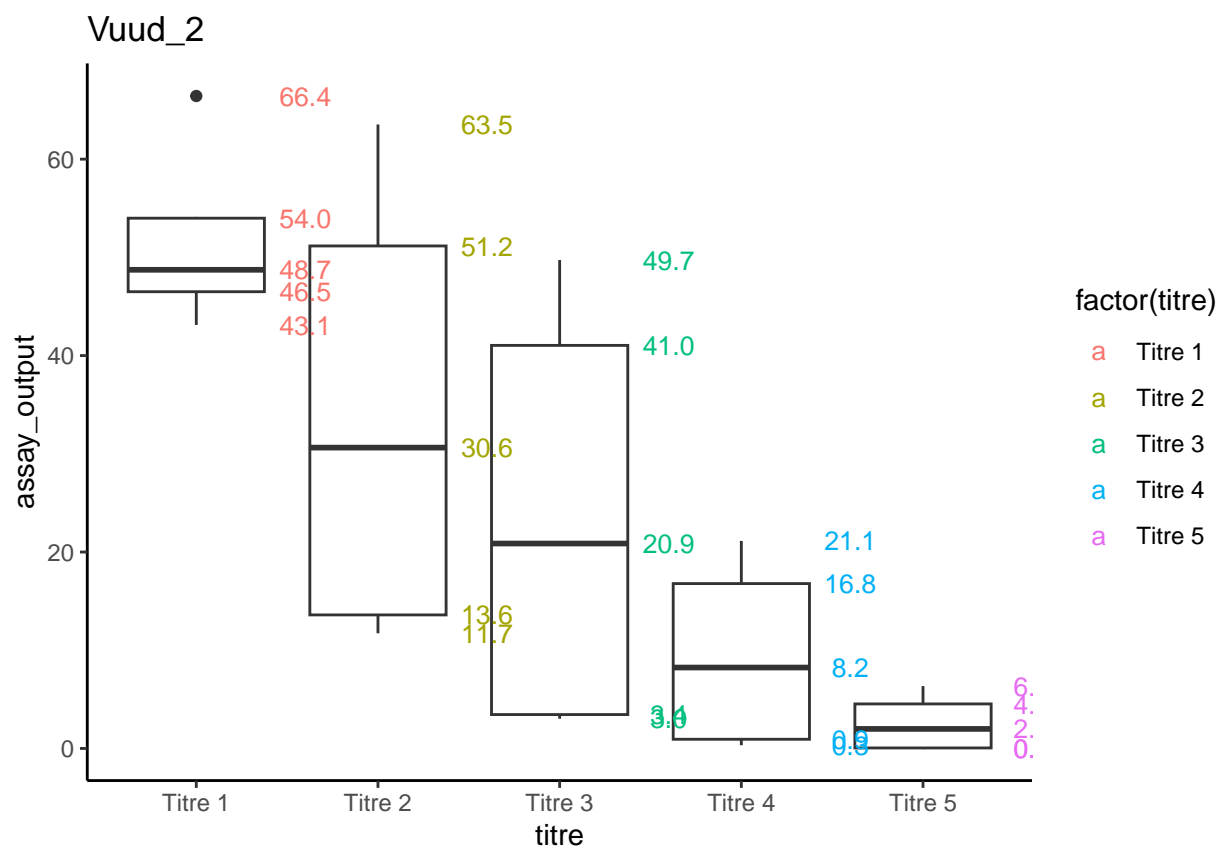
```
##
## [[132]]
```



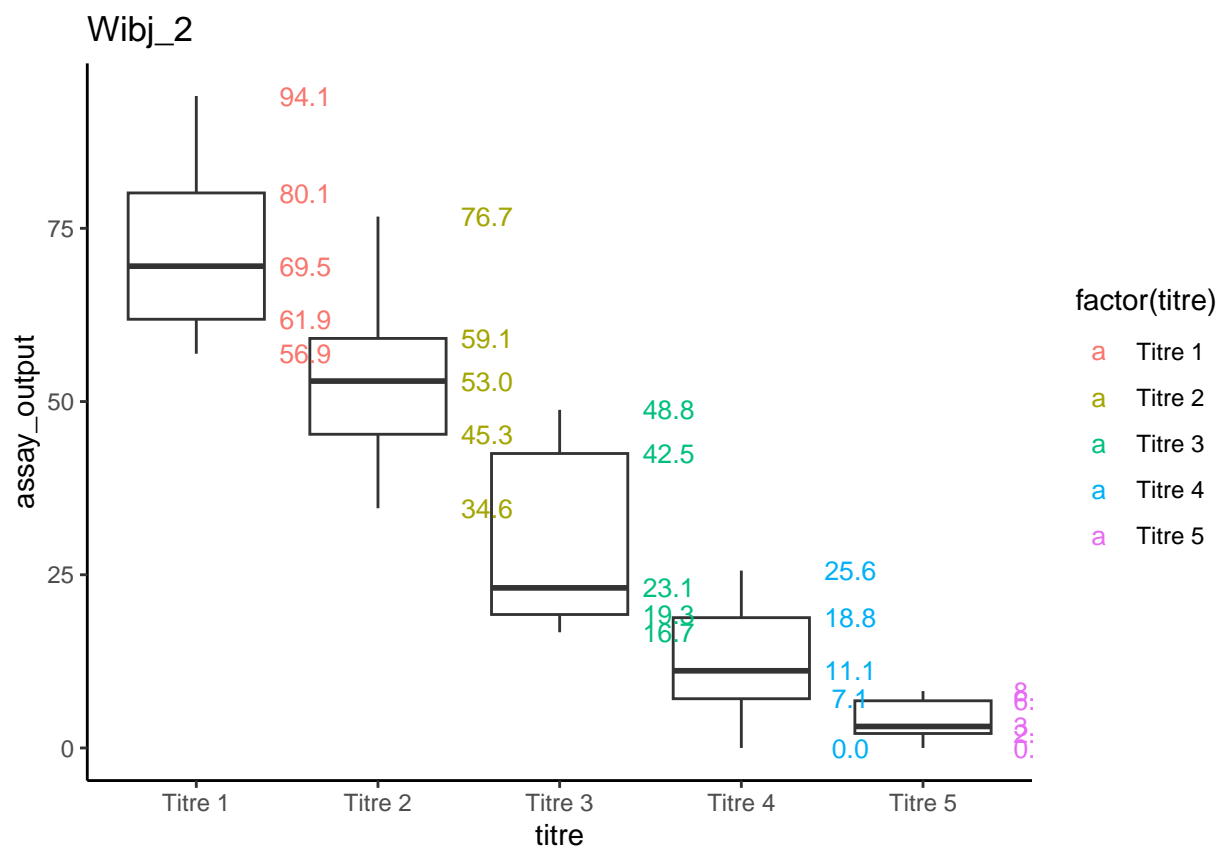
```
##
## [[133]]
```



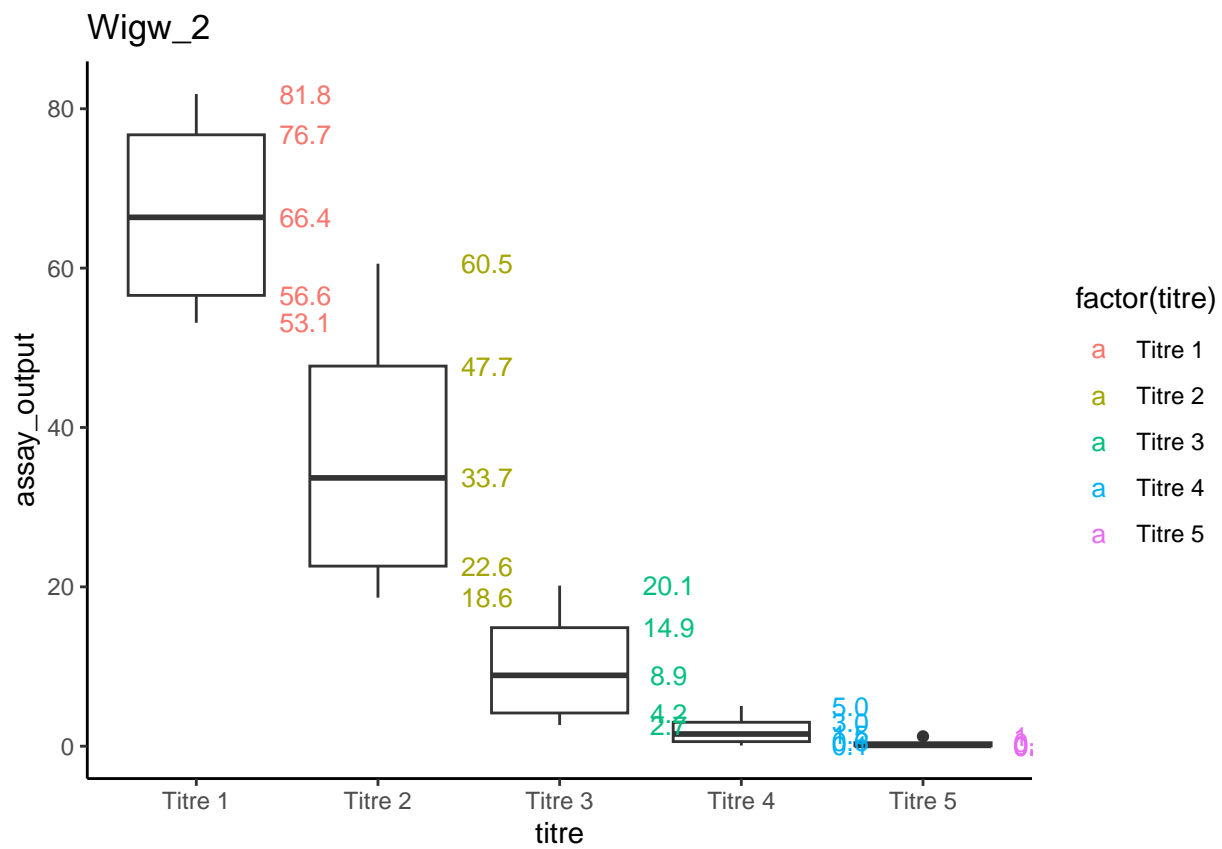
```
##
## [[134]]
```

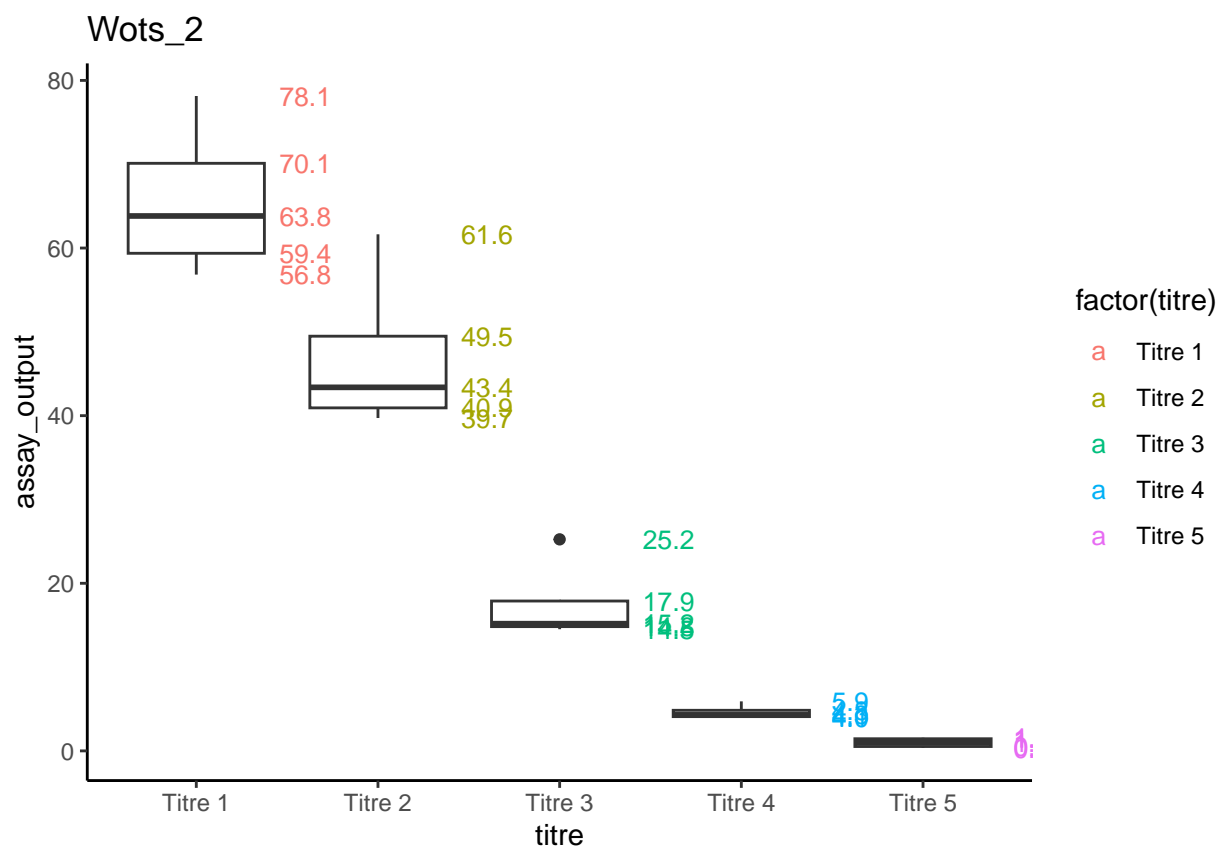
```
##
## [[135]]
```



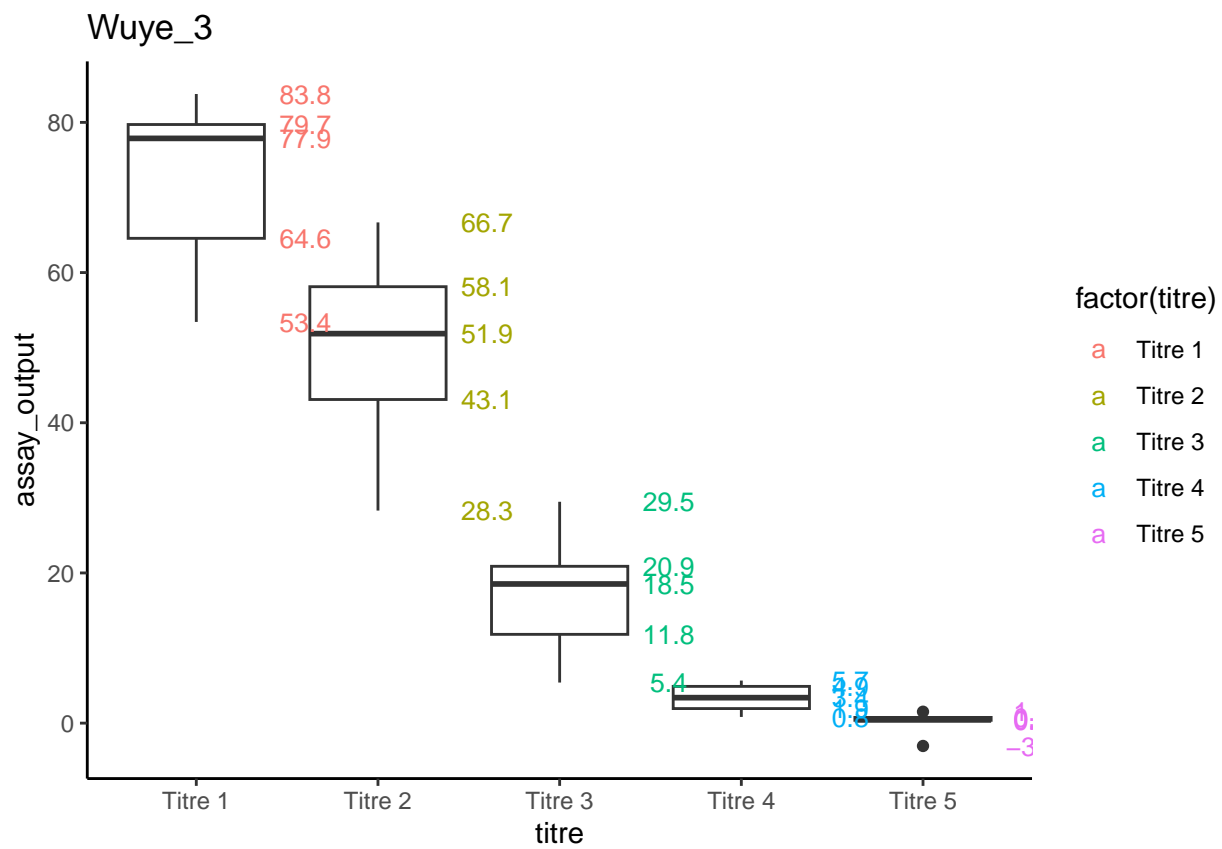
```
##
## [[136]]
```



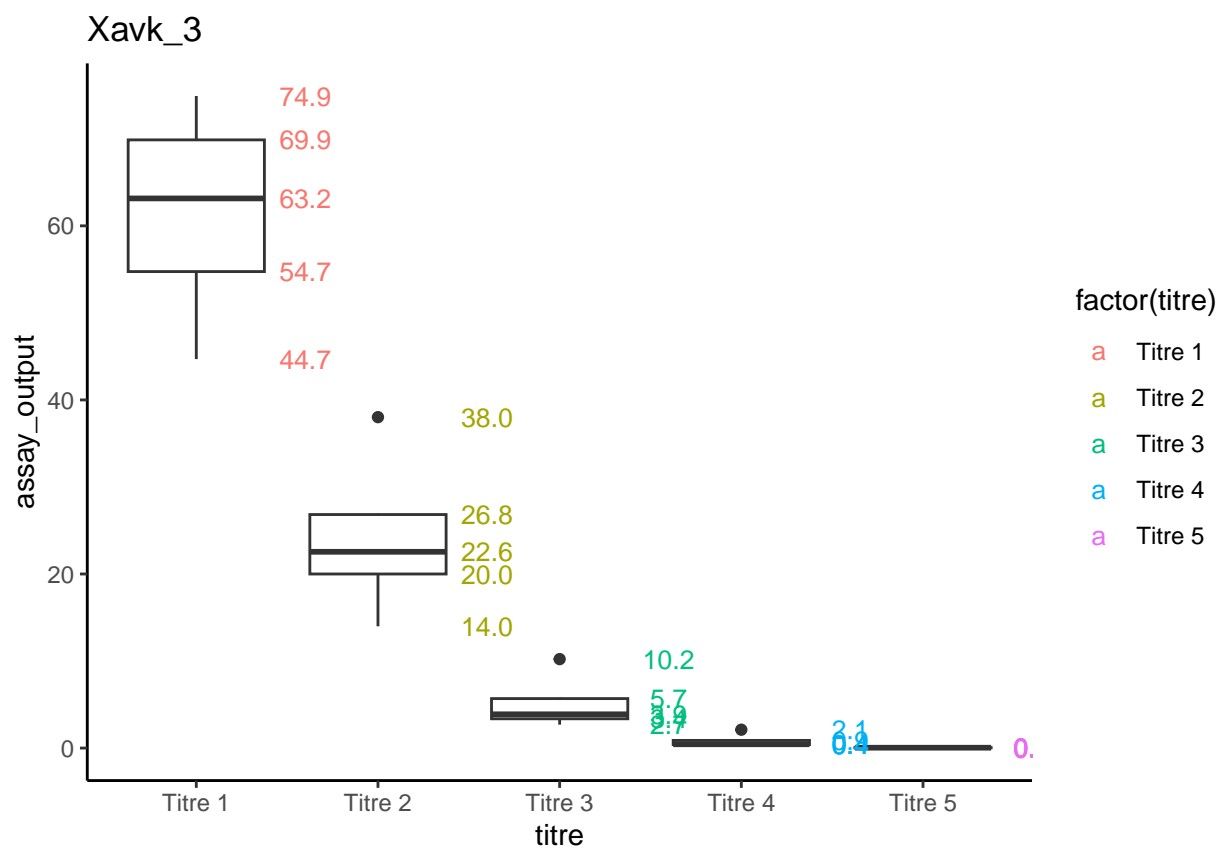
```
##
## [[137]]
```



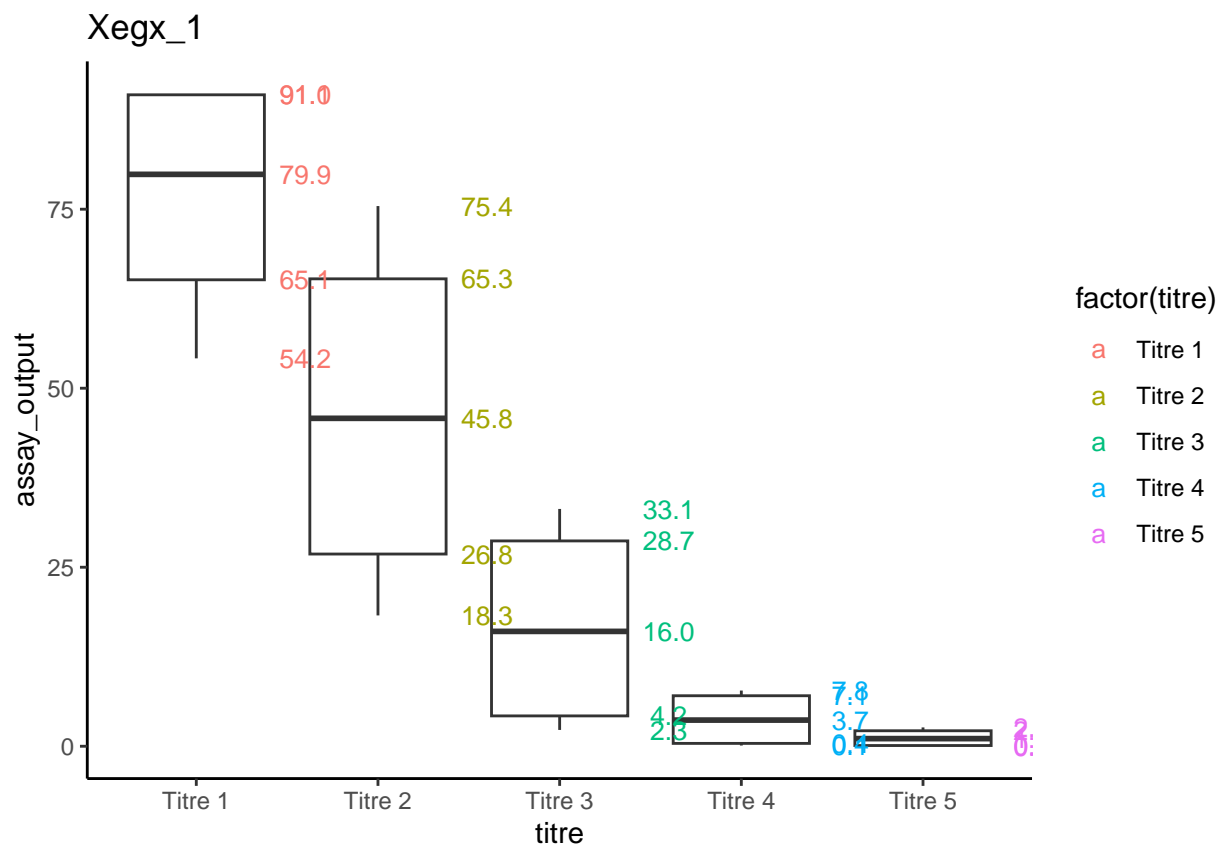
```
##
## [[138]]
```



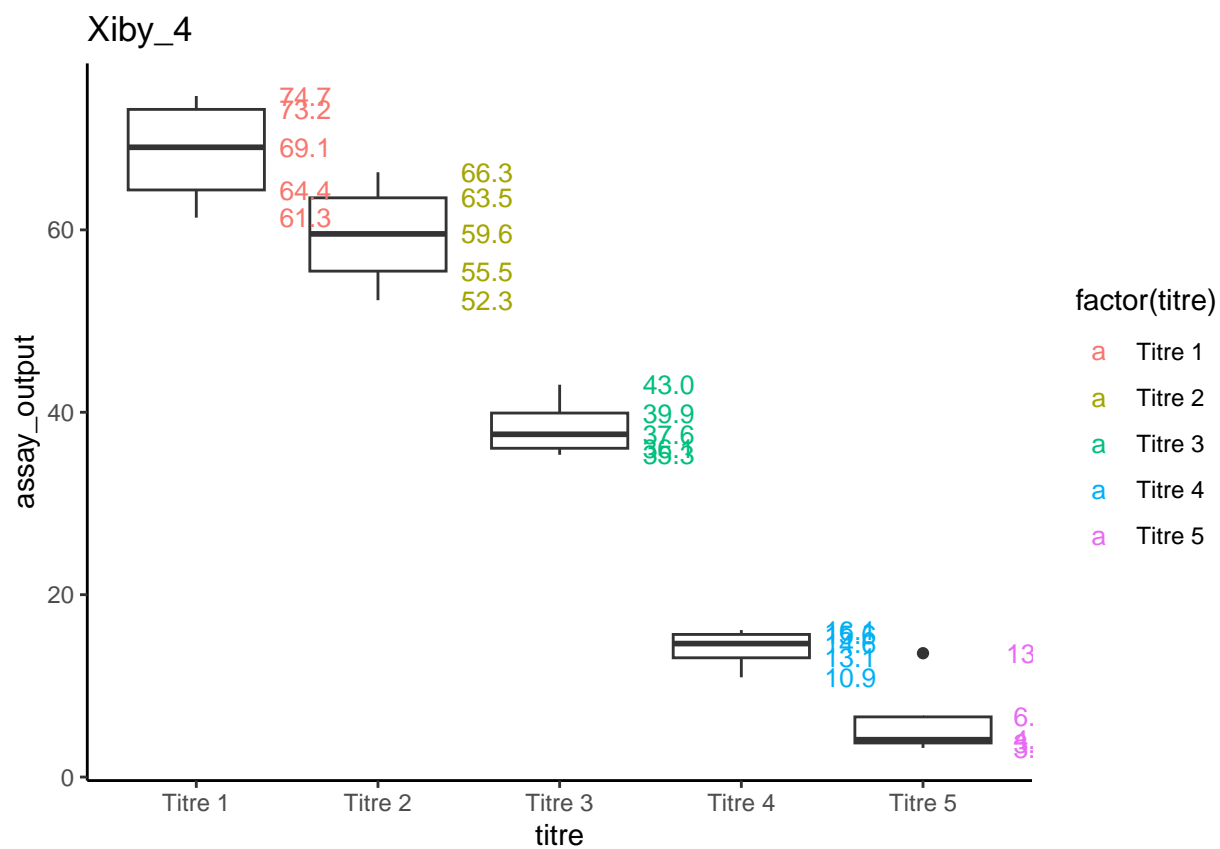
```
##
## [[139]]
```



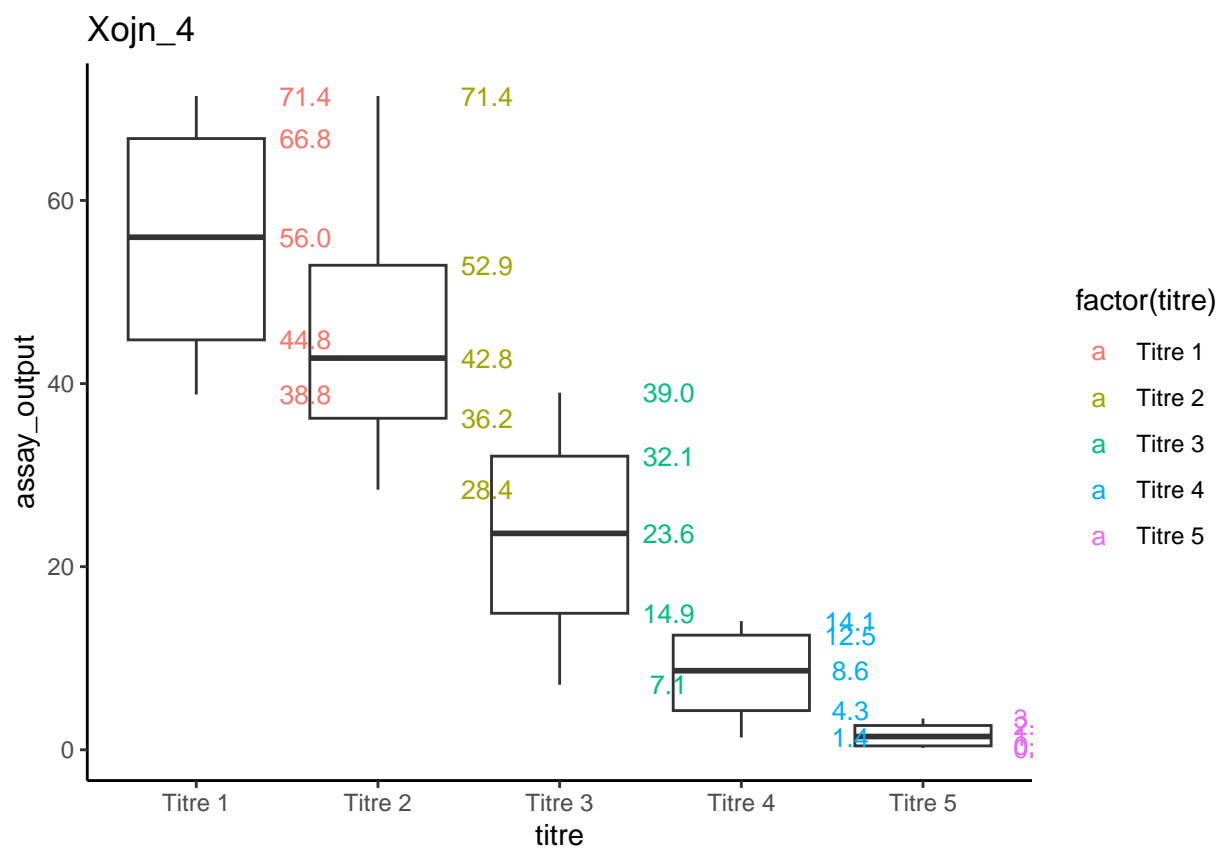
```
##
## [[140]]
```



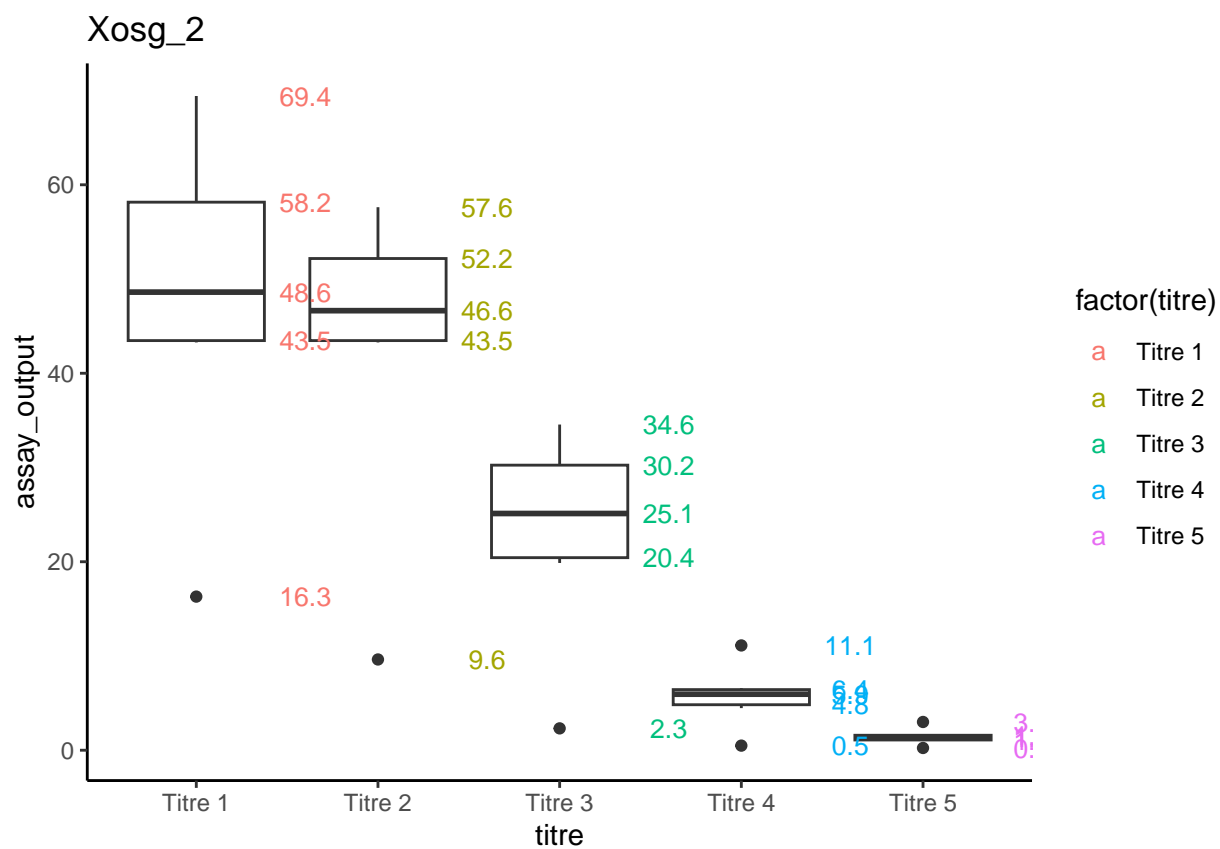
```
##
## [[141]]
```



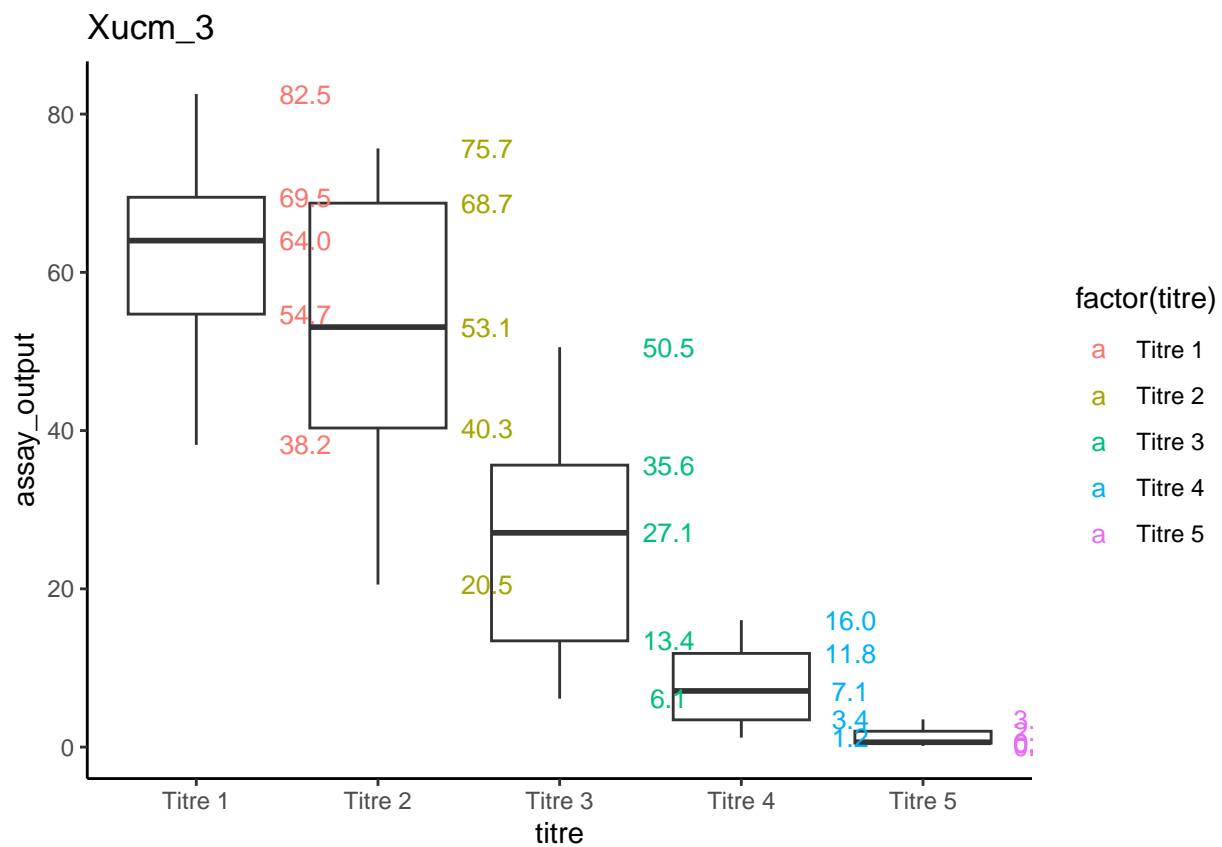
```
##
## [[142]]
```

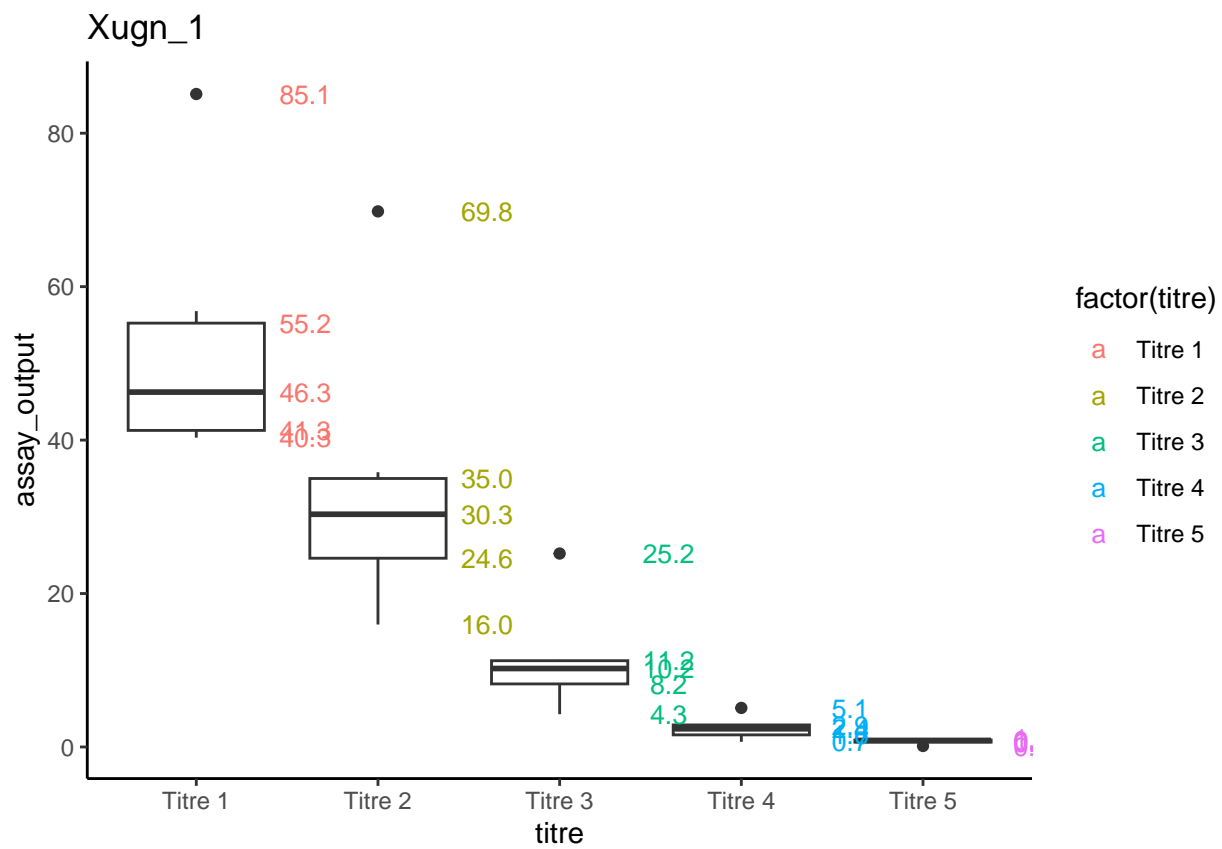
```
##
## [[143]]
```



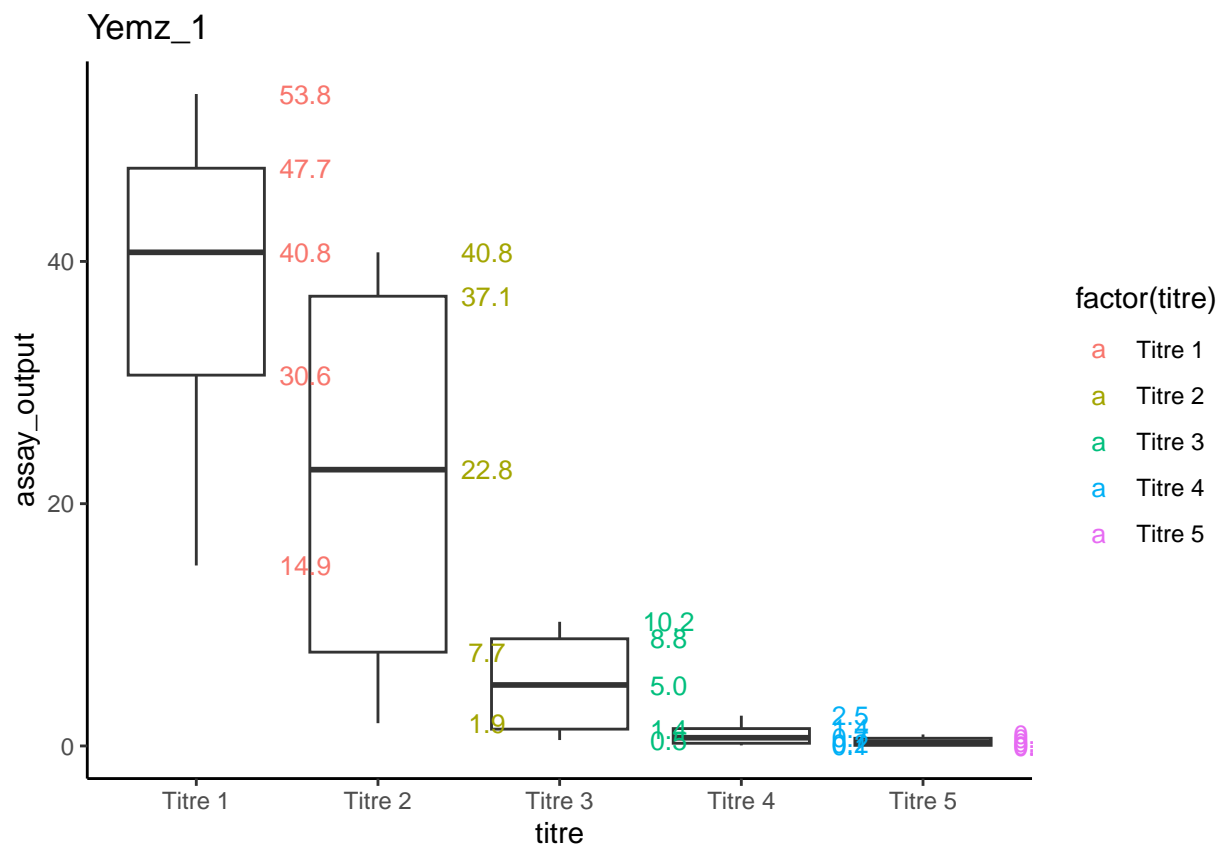
```
##
## [[144]]
```



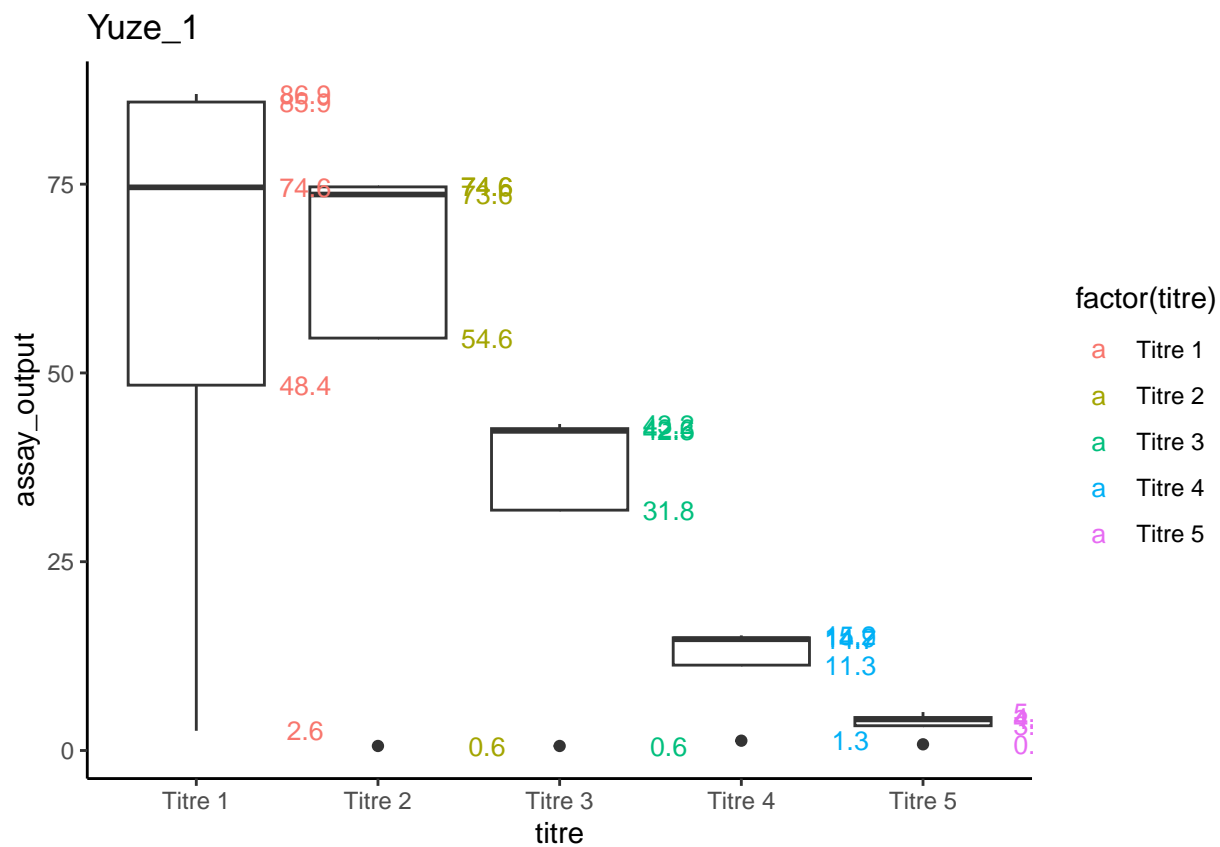
```
##
## [[145]]
```



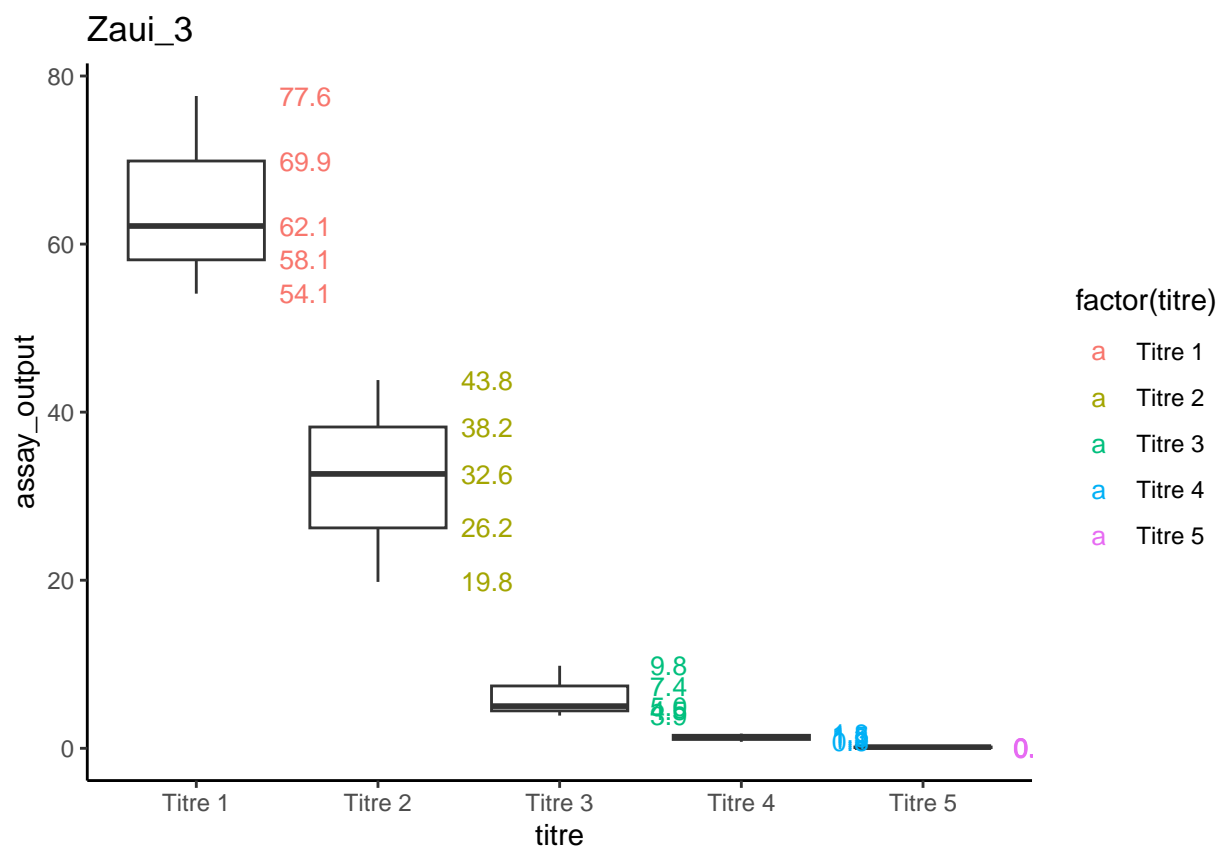
```
##
## [[146]]
```



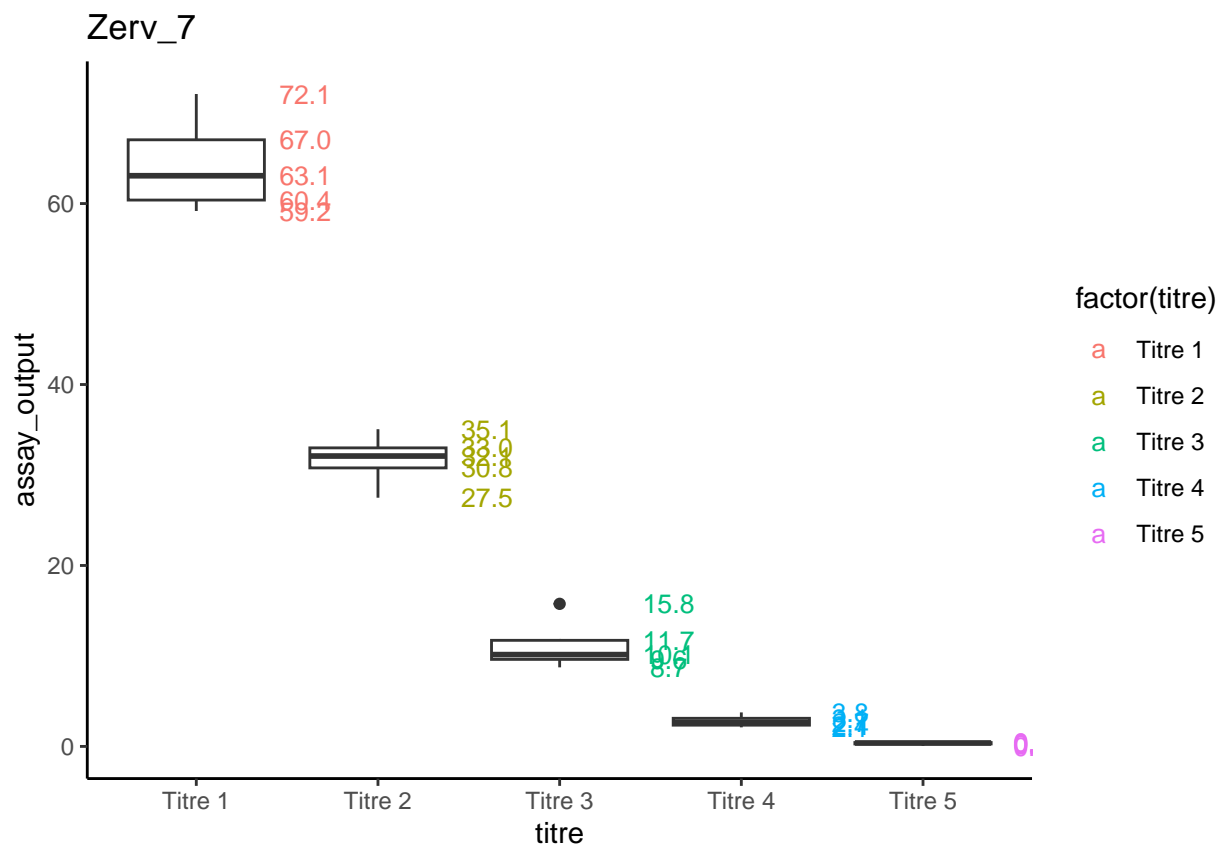
[[147]]



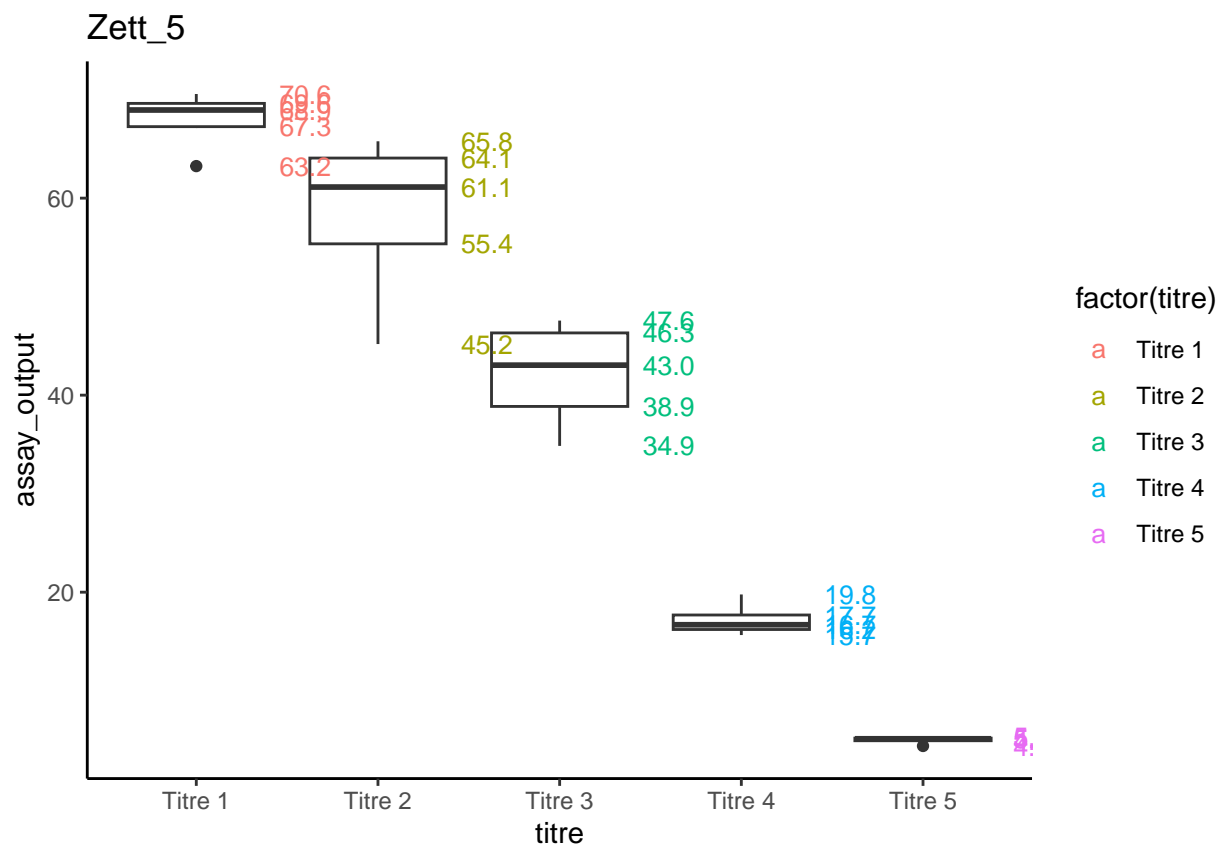
```
##
## [[148]]
```



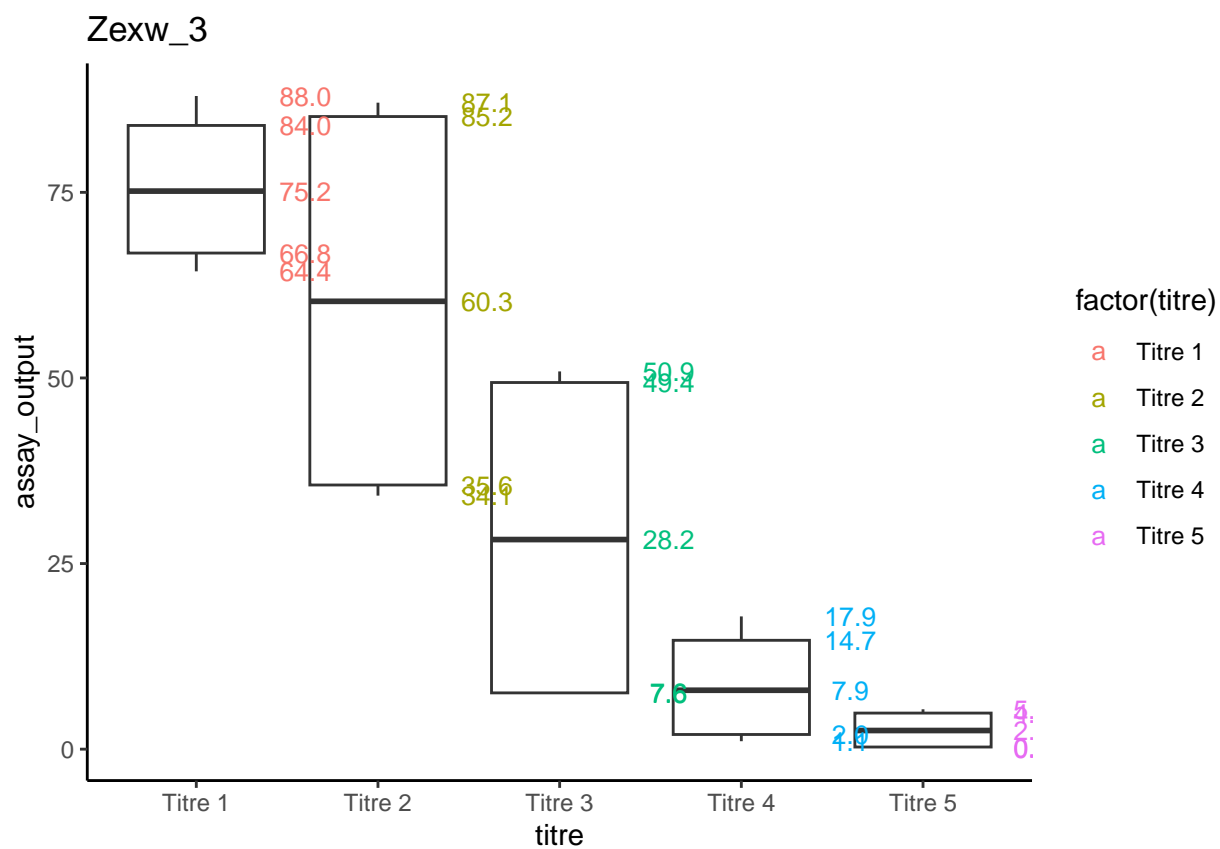
```
##
## [[149]]
```



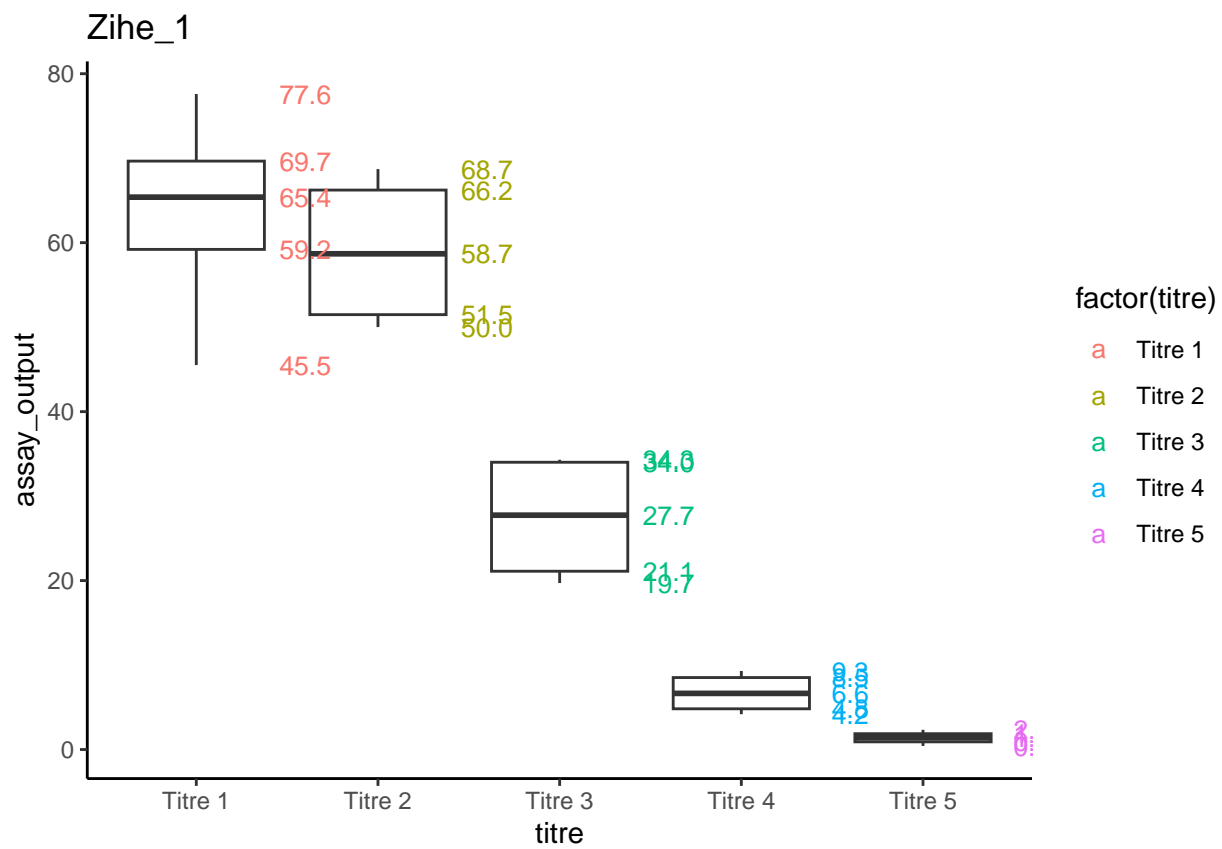
```
##
## [[150]]
```

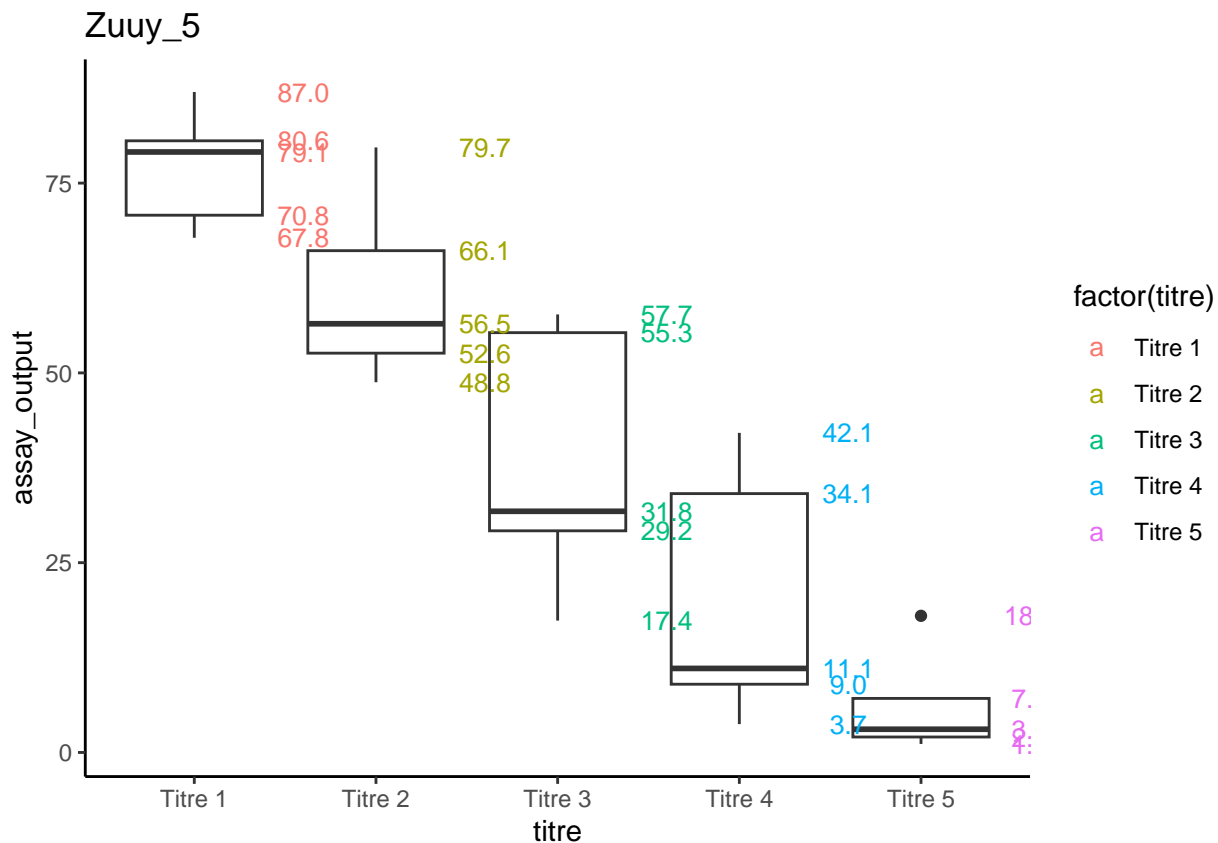
```
##
## [[151]]
```



```
##
## [[152]]
```



```
##
## [[153]]
```



```
#names(box_plot_vector)=vector_data_per_cell_line$cell_line
```

dealing with outliers

Now i want to get rid of outliers, however i can't just get rid of all outlier values because they might actually be biologically relevant, have to check with other values in the replicates or screen,

creating outlier function which determines if the value is outlier by $(1.5 \times IQR + UQ)$ or $1.5 \times IQR - LQ$

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}
```

this is a function that applies outlier per titre (checks if value is a outlier compared to other values in other screens)

```
outlier_function=function(i){
  i %>%
  group_by(titre) %>%
  mutate(outlier = ifelse(is_outlier(assay_output), assay_output, as.numeric(NA)))
}
```

use this to remove outlying observations. per screen (per cell) there are 5 titres (with 2 replicates), if at least 3 of these points are outliers that set of points for that replicate is removed (whole observation may be outlier)

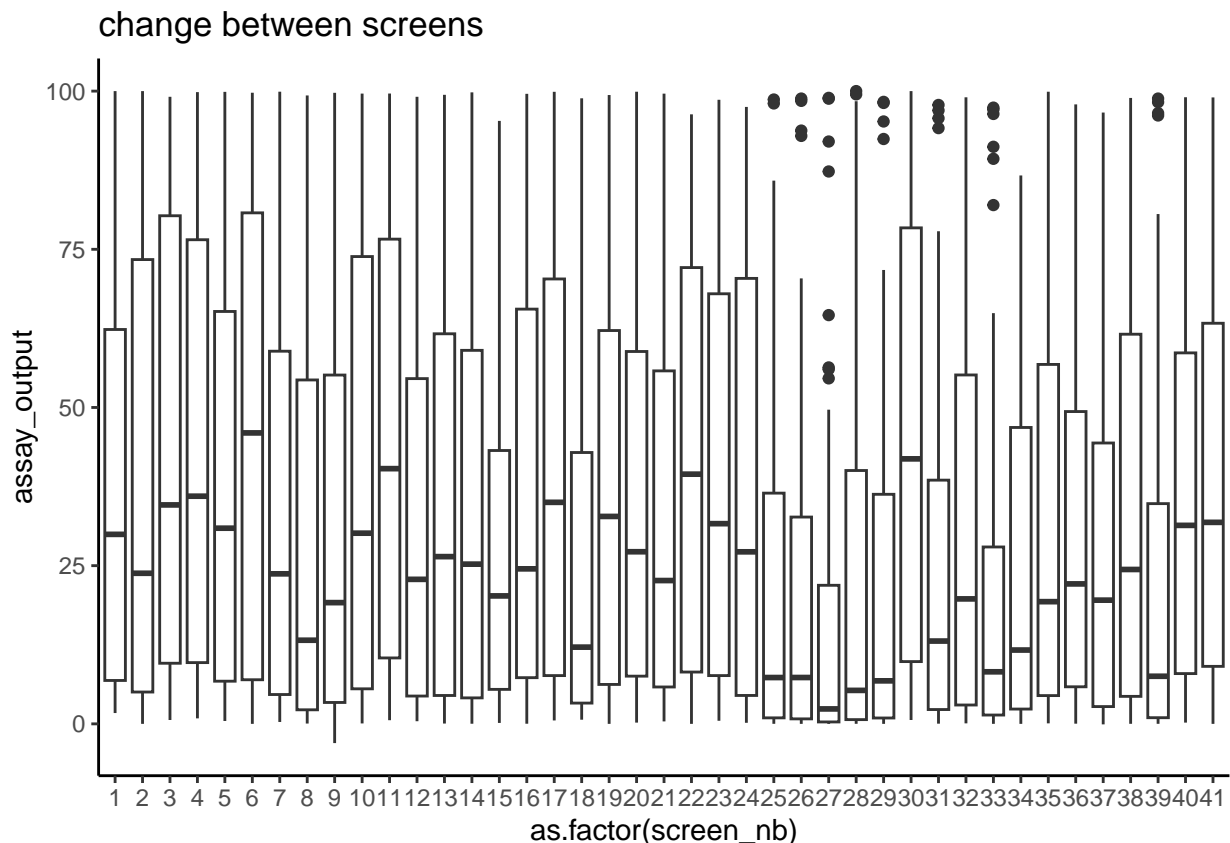
```
all_outliers=purrr::map(vector_data_per_cell_line$data,outlier_function)
HIV1_vector_data=bind_rows(all_outliers)%>%group_by(cell_line,screen_nb,replicate)%>%filter(((sum(is.na
```

normalising between the screens

In order to get rid of any batch effects from the screening process, we want to find the zscore to normalise within the screen, this makes it more comparable between screens

plot the boxplot for each screen, before normalisation

```
ggplot(data = HIV1_vector_data, aes(y=assay_output, x=as.factor(screen_nb), group = screen_nb,)) +  
  geom_boxplot() +  
  ggtitle('change between screens') +  
  theme_classic()
```

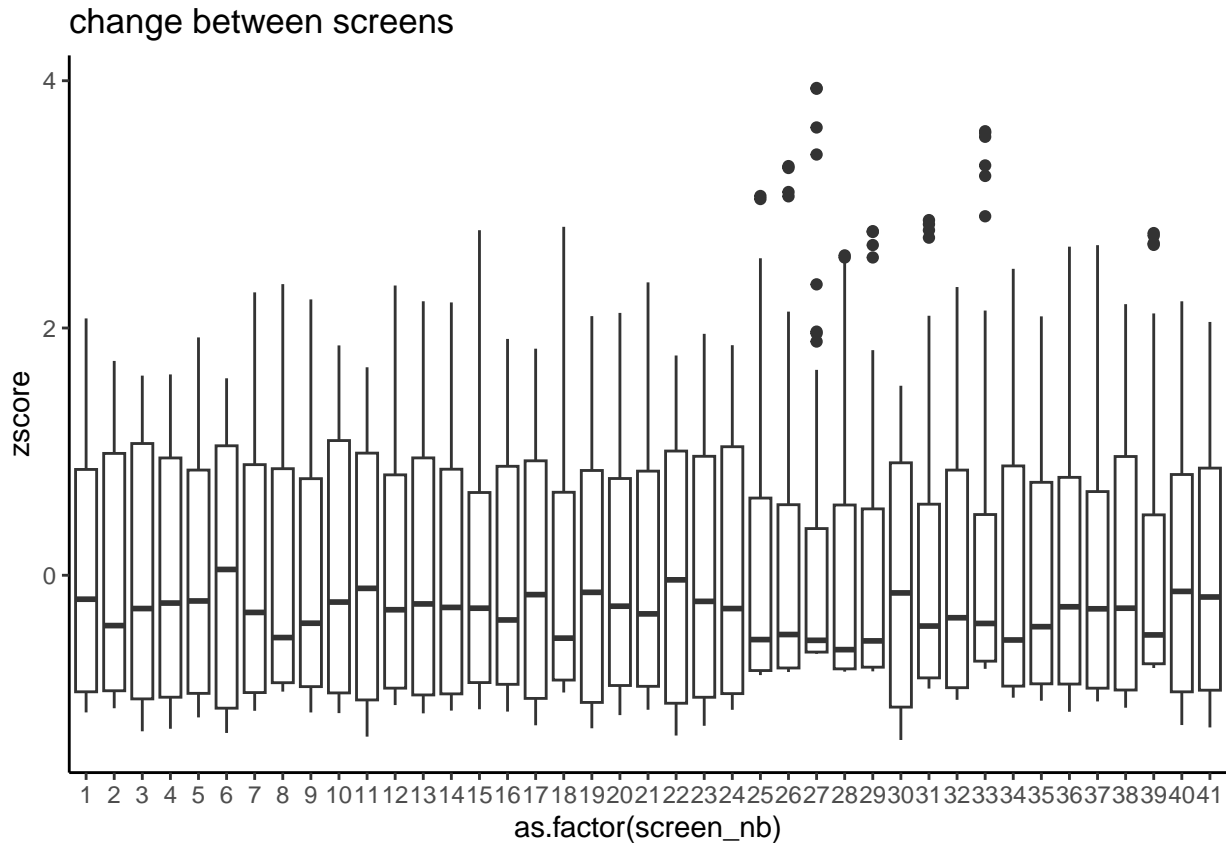


okay going to apply zscore to all values to standardise between screens to try and get rid of any batch effects to allow proper comparison

```
#HIV_vector_group_data = HIV1_vector_data%>%summarise(mean_output=mean(assay_output),sd_output=sd(assay_output))  
HIV1_vector_data=HIV1_vector_data%>%  
  group_by(screen_nb)%>%  
  mutate(zscore=(((assay_output-mean(assay_output))/sd(assay_output))))%>%  
  ungroup()
```

making box plot for after normalisation

```
ggplot(data = HIV1_vector_data, aes(y=zscore, x=as.factor(screen_nb), group = screen_nb,)) +  
  geom_boxplot() +  
  ggtitle('change between screens') +  
  theme_classic()
```



finding parameters

so now i have the normalised max mean percentage (in z score) A bit worried that by normalising it and getting rid of the difference it might affect results

finding now the area under the curve using AUC function

```
HIV1_vector_data=HIV1_vector_data%>%group_by(batch,cell_line,screen,screen_nb,replicate)%>%
  mutate(area_under_curve=AUC(infection_volume_ul,(zscore)))%>%
  ungroup()
```

updating the df list per cell line

```
vector_data_per_cell_line=HIV1_vector_data%>%nest_by(cell_line,.keep = T)
```

fitting the curves

##logarithmic

Here we use a logarithmic

$$a + b \times \log_2(x - c)$$

with 3 parameters, a- y offset b- slope and c- x offset, to help plot the graph in our analysis a higher a and b should correspond to more permissive/susceptible and a lower c would correspond to less permissive. However this depends on how the data is fitted, the fit may mean these values aren't directly correlated with those phenotype

```
logarithmic_func <- function(x, a, b,c) {
  a + b * log2(x-c)
}
```

this is the function for fitting. using a robust nlsLM function, fits the data using start points, having to choose good starting volumes, saves coefs to the database. There is also error handling using tryCatch- giving error message and instead adds NA

Plotting function, using ggplot, use logarithmic function to create predicted list of values, these values are then fed into geom line to be the fitted line. then add the parameters onto the graph averaged for each cell line

```
apply_plot_log=function(i){
i <- i %>%
  mutate(predicted = logarithmic_func(infection_volume_ul, i$a, i$b,i$c))

  # Create the plot
p=ggplot(data = i, aes(x = infection_volume_ul, y = zscore, group = interaction(screen_nb,replicate)))
  geom_point(aes(color=screen_nb),color = "blue", size = 2, alpha = 0.6) + # Actual data points
  geom_line(aes(y = predicted,color=screen_nb), color = "red", size = 1) + # Fitted line
  ggtitle(unique(i$cell_line)) + # Dynamic title
  labs(x = "Infection Volume (uL)", y = "Mean Z-Score") +
  theme_classic()

  params_text <- paste0(
    "Parameters:\n",
    "a = ", round(mean(i$a, na.rm = TRUE), 2), "\n",
    "b = ", round(mean(i$b, na.rm = TRUE), 2), "\n",
    "d = ", round(mean(i$c, na.rm = TRUE), 2), "\n"
  )

  p + annotate("text",
    x = Inf, y = -Inf,
    label = params_text,
    hjust = 1.1, vjust = -0.1,
    size = 3)
}
```

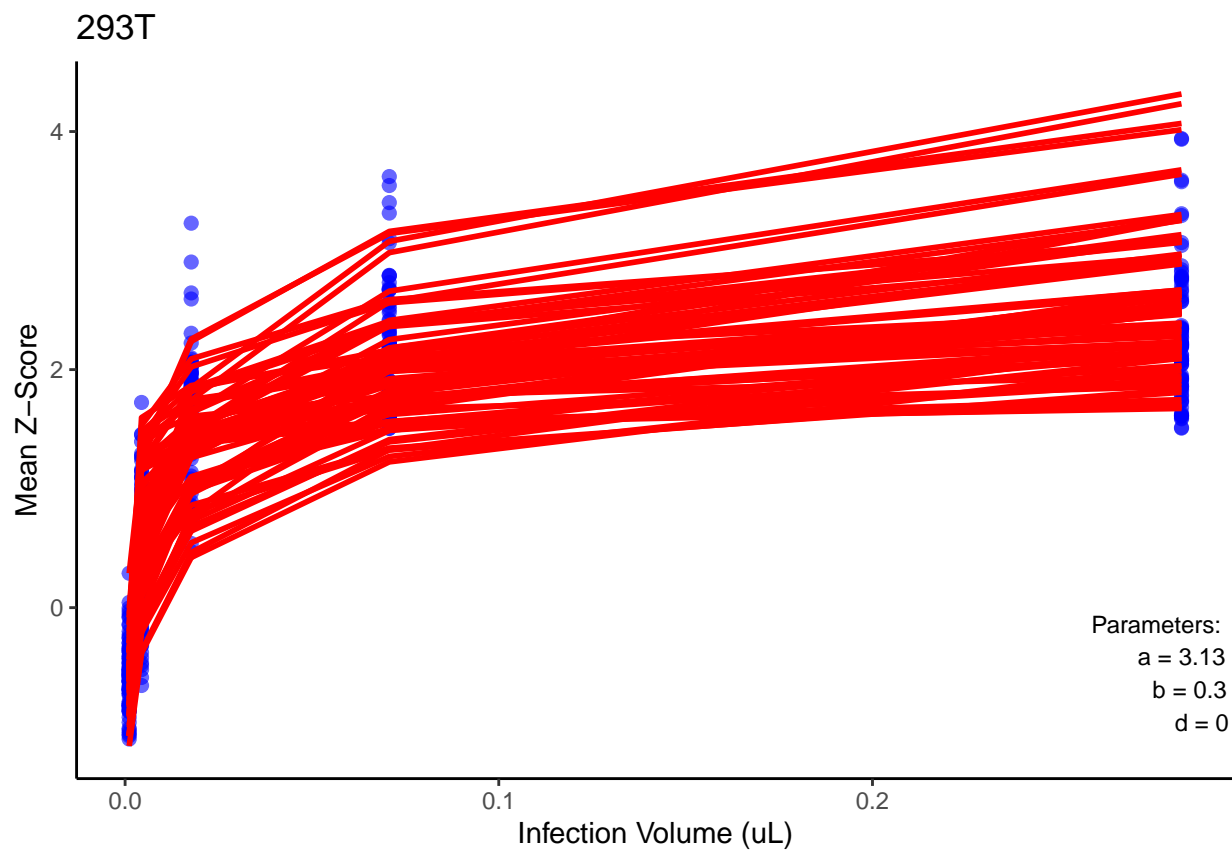
```
logarithmic_plot_vector=purrr::map(vector_data_per_cell_line$data,apply_plot_log)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

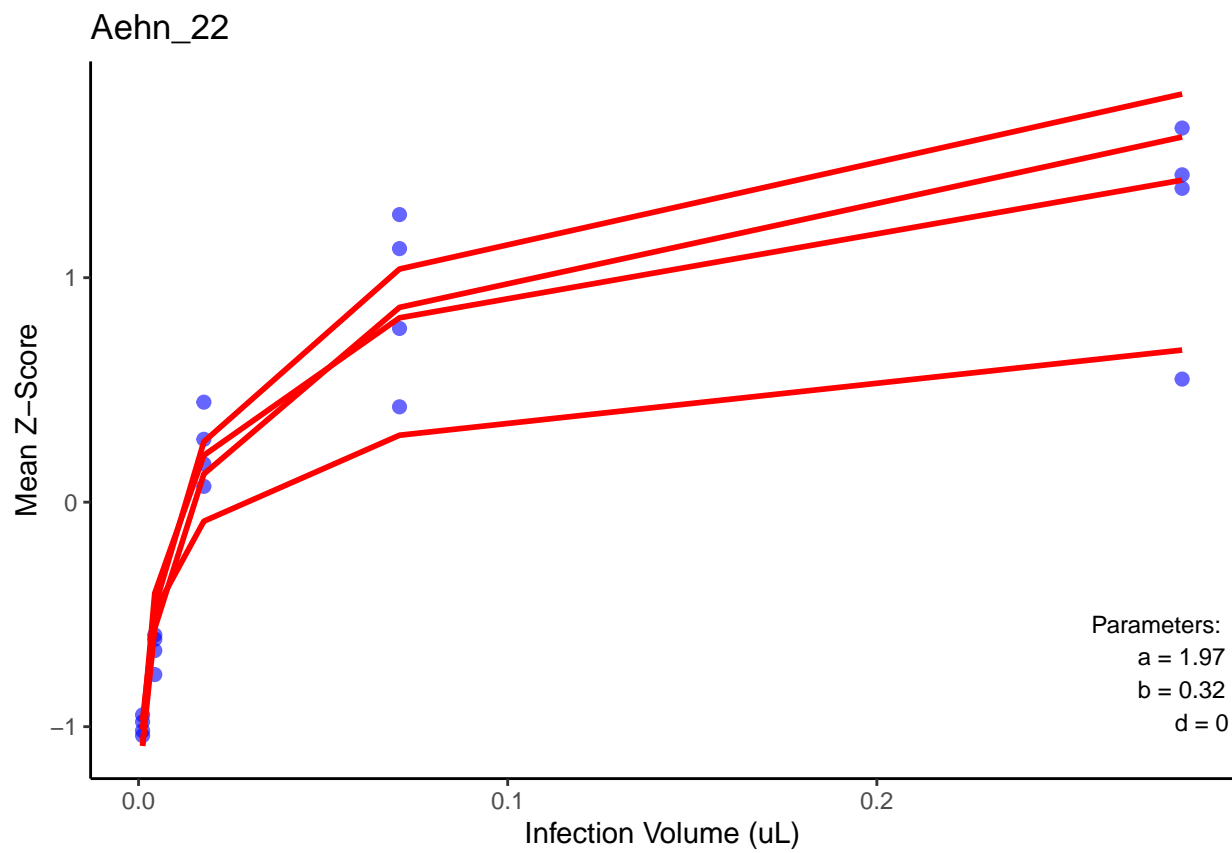
```
names(logarithmic_plot_vector)=vector_data_per_cell_line$cell_line
logarithmic_plot_vector
```

```
## $`293T`
```

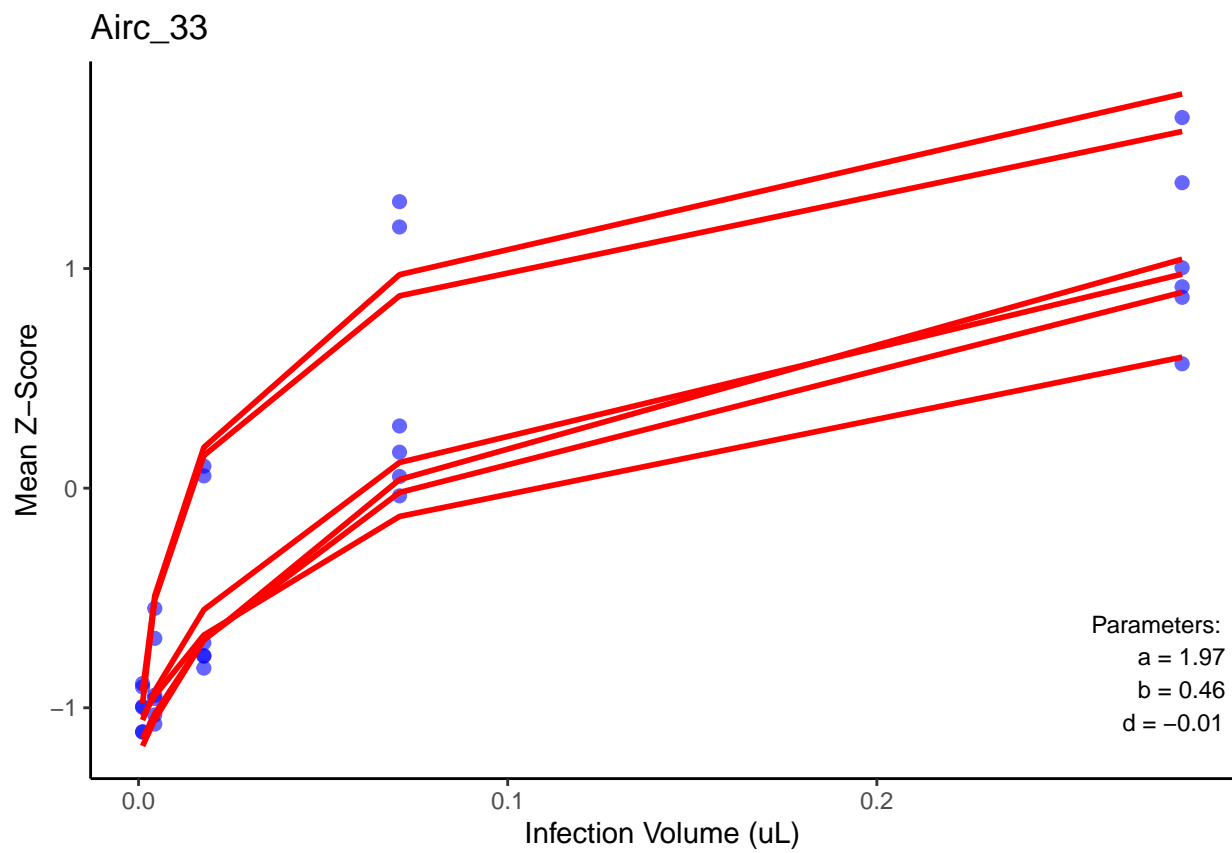
```
## Warning: Removed 15 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



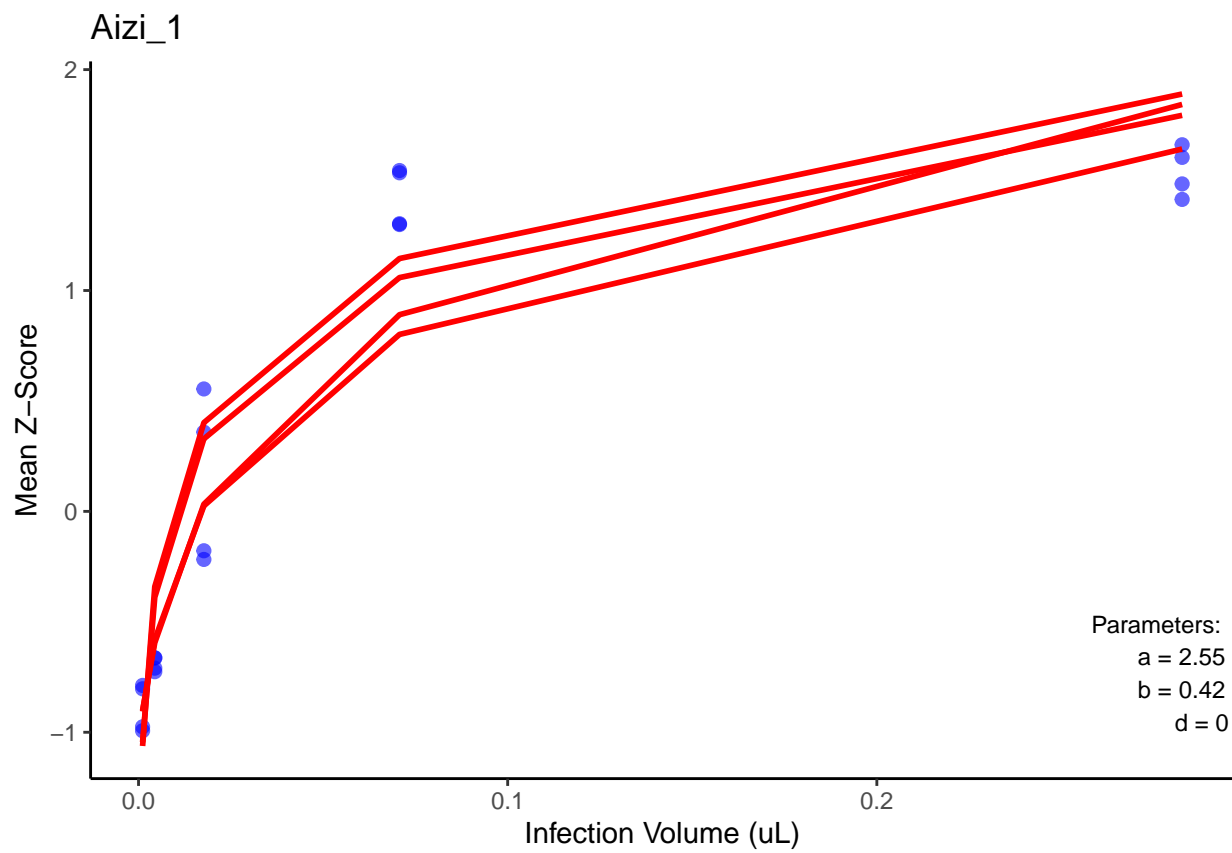
 ## \$Aehn_22



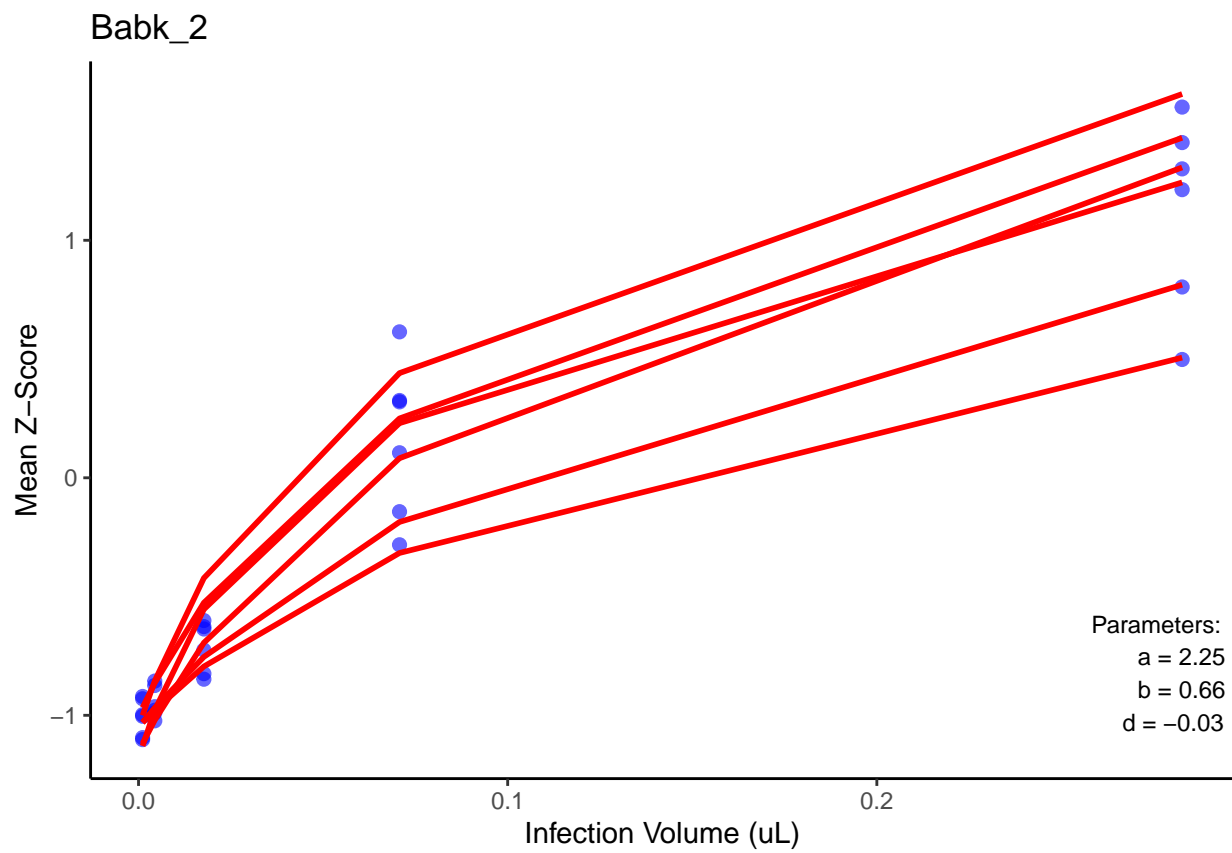
\$Airc_33



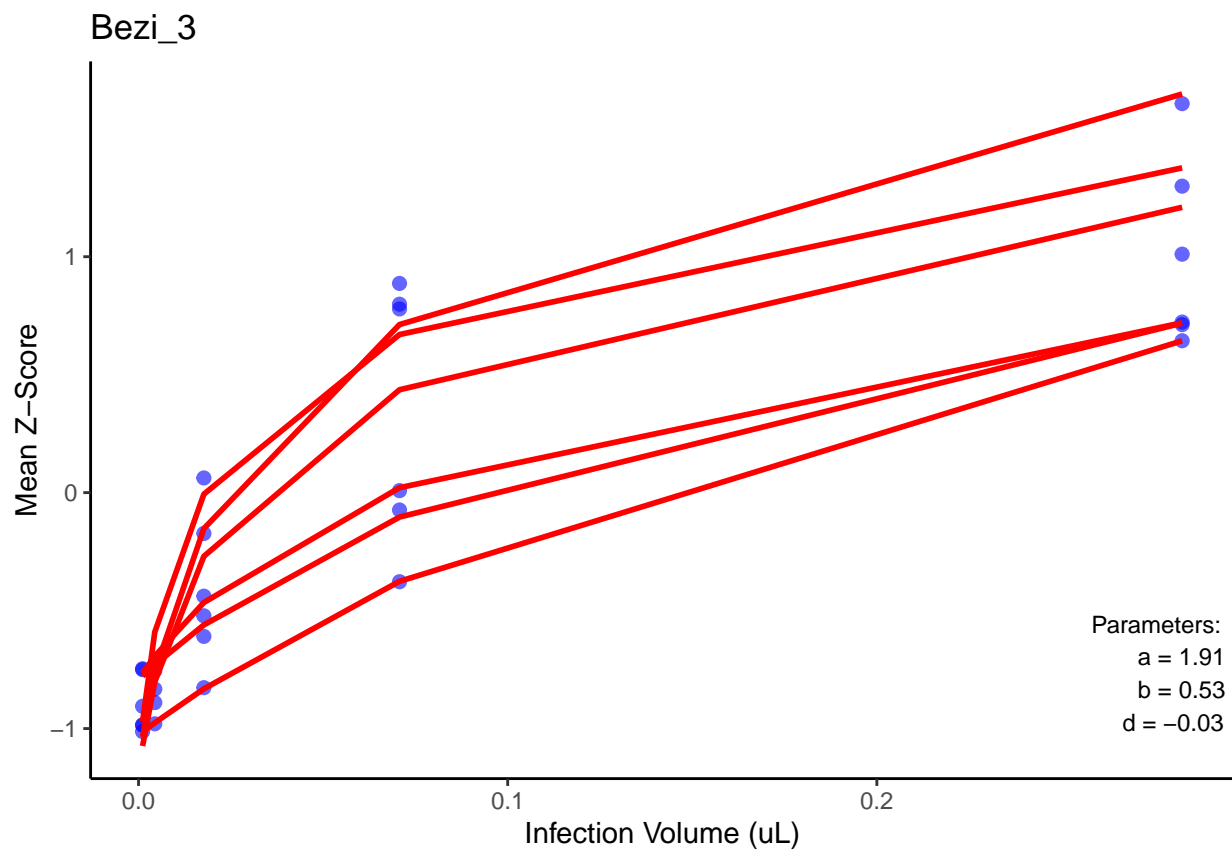
\$Aizi_1



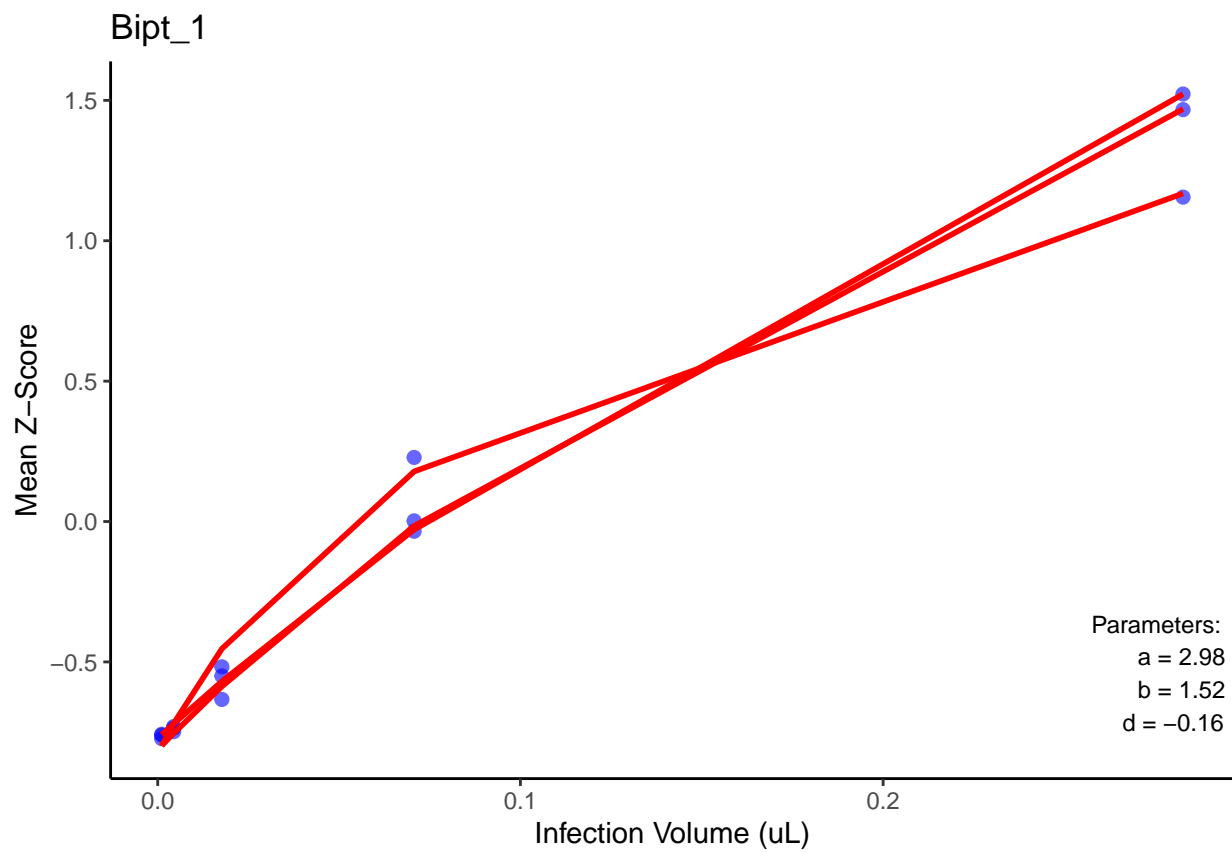
 ## \$Babk_2



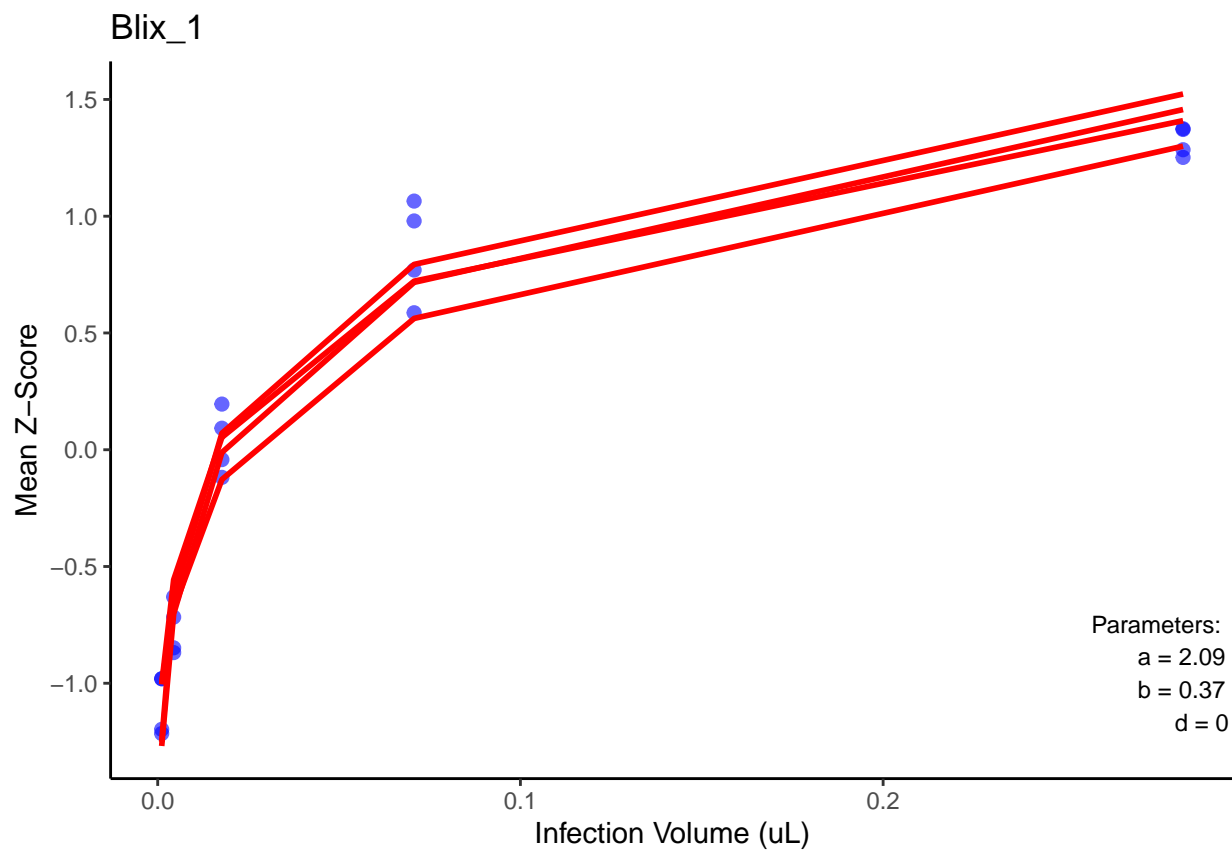
\$Bezi_3



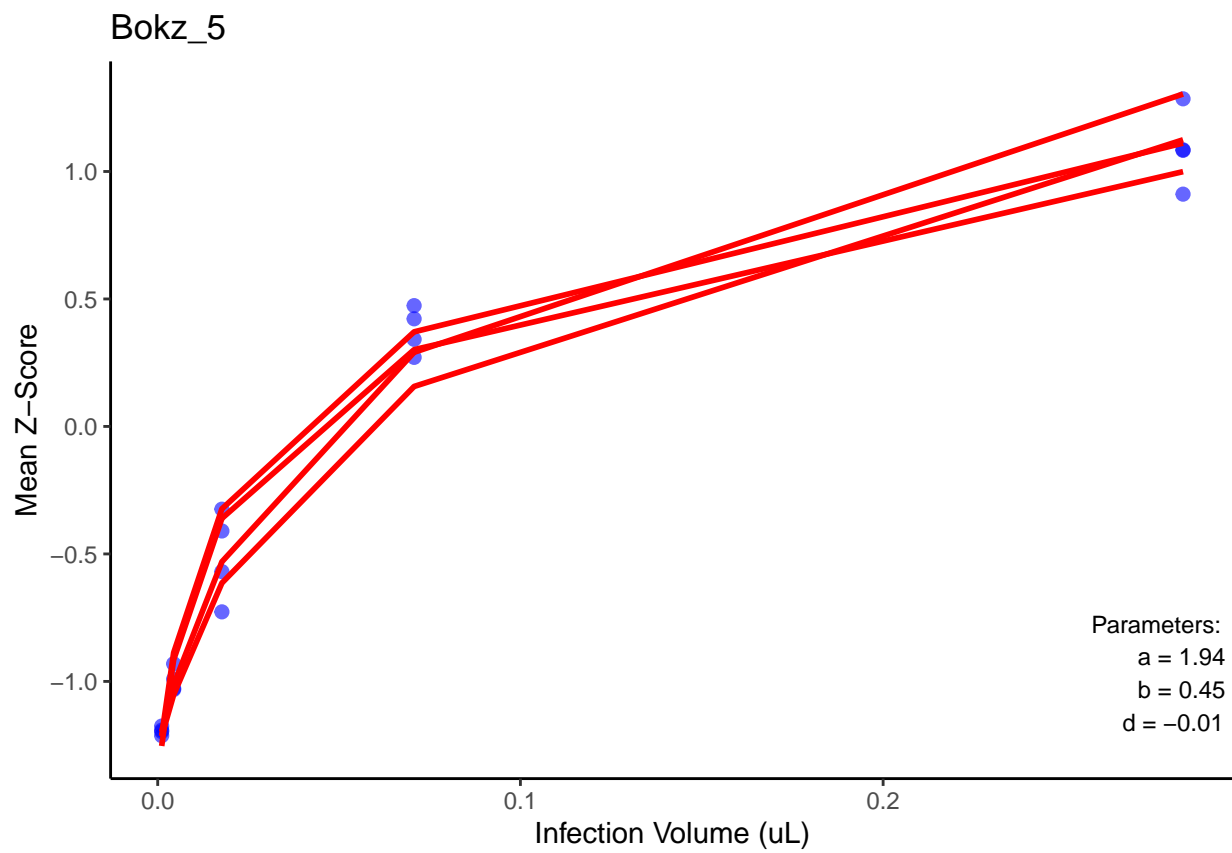
\$Bipt_1



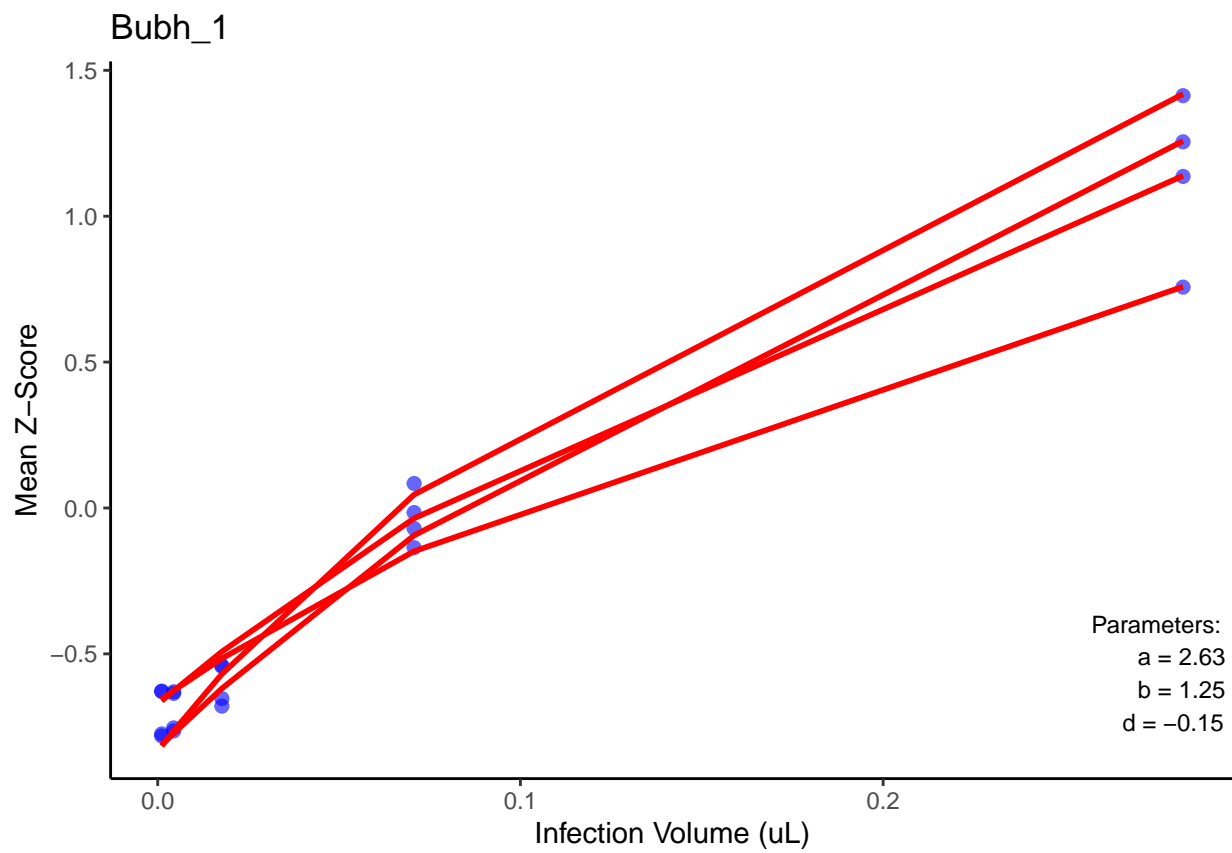
\$Blix_1



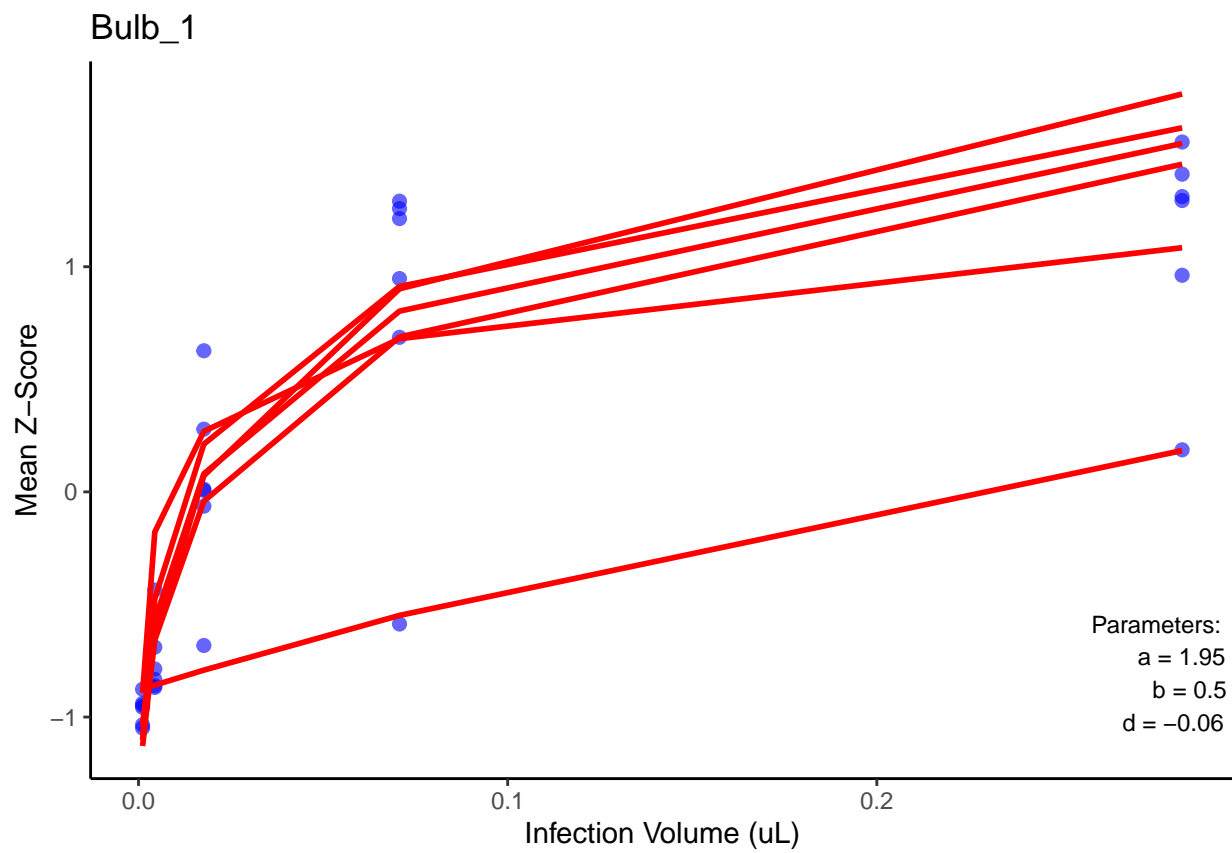
\$Bokz_5



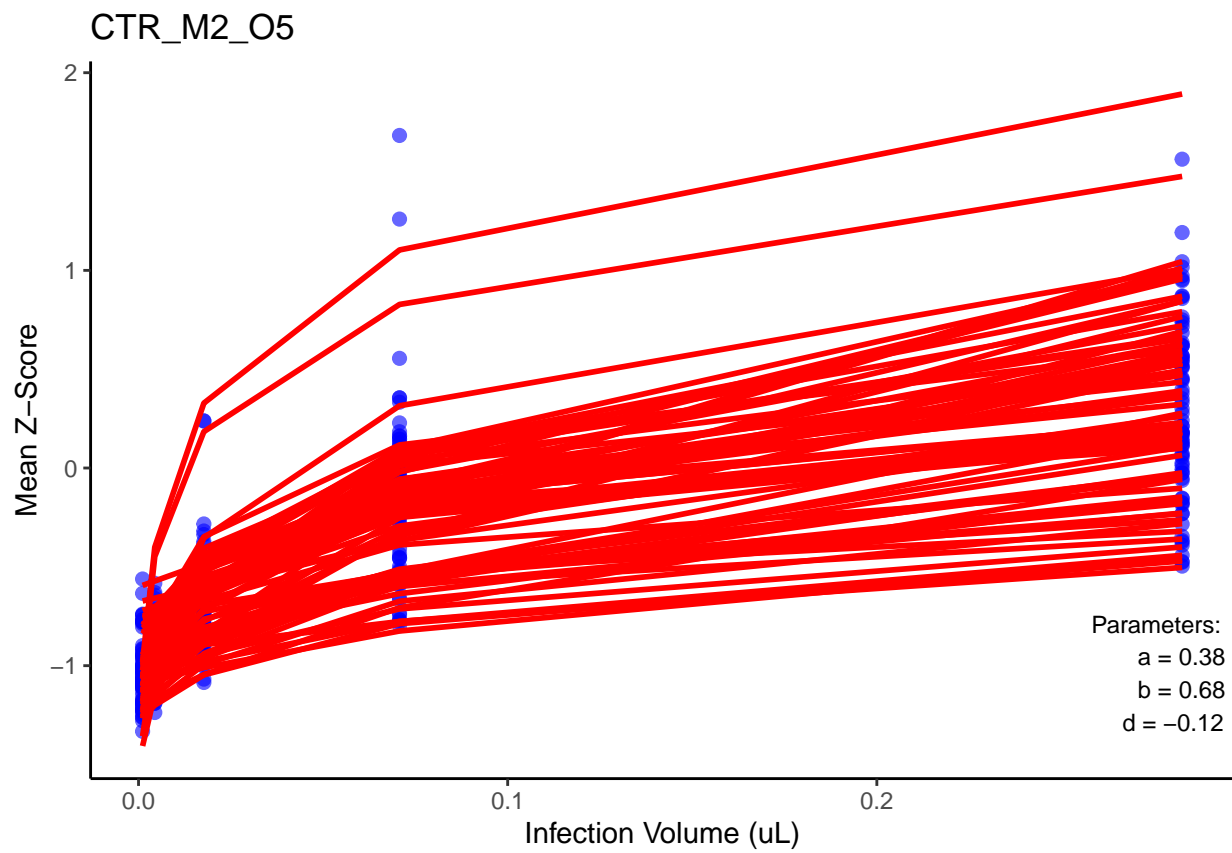
\$Bubh_1



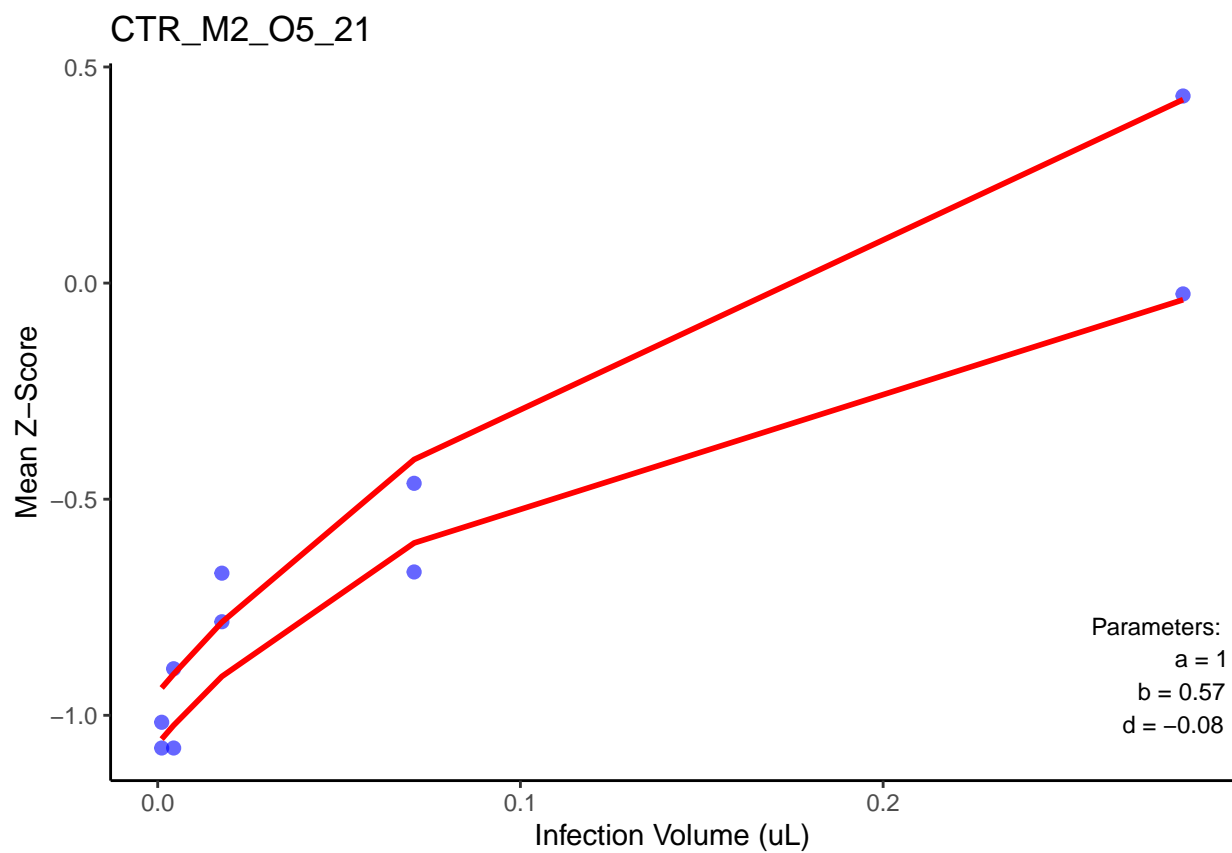
 ## \$Bulb_1



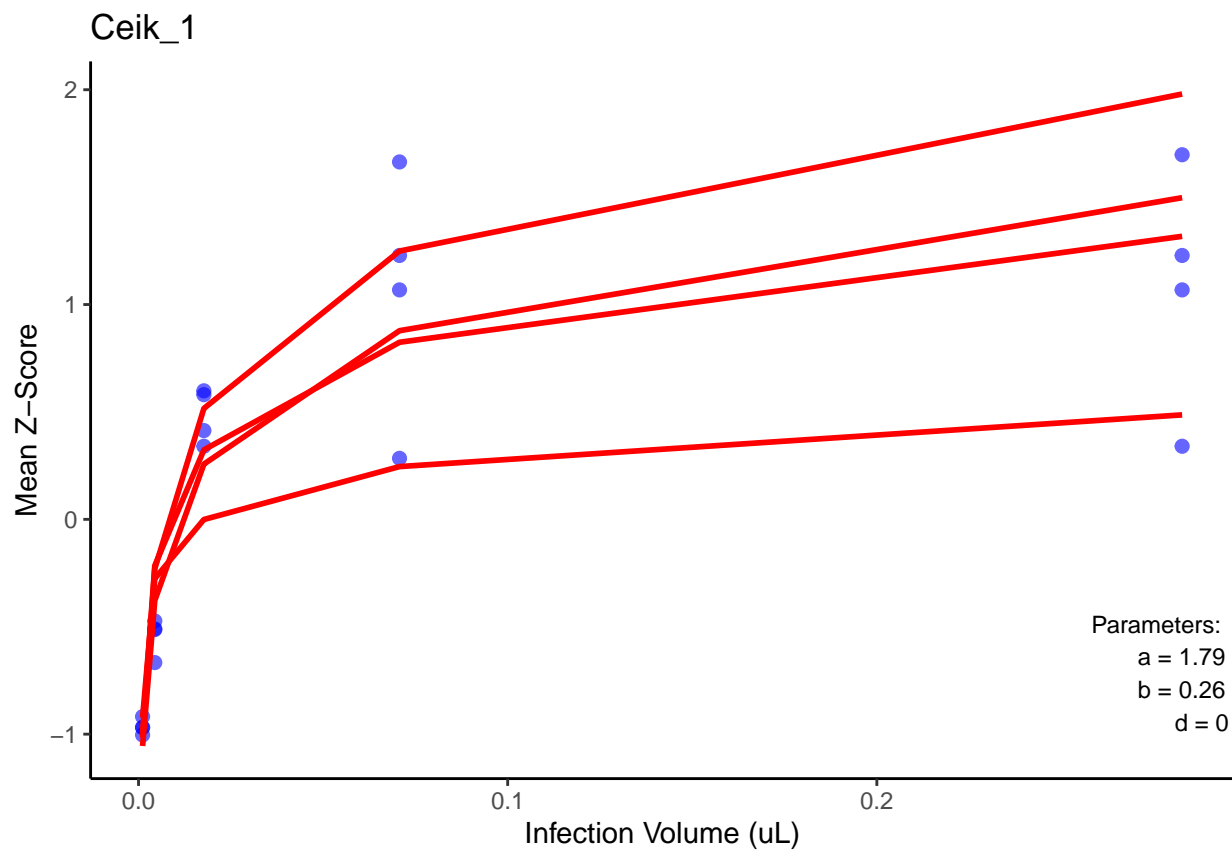
\$CTR_M2_05



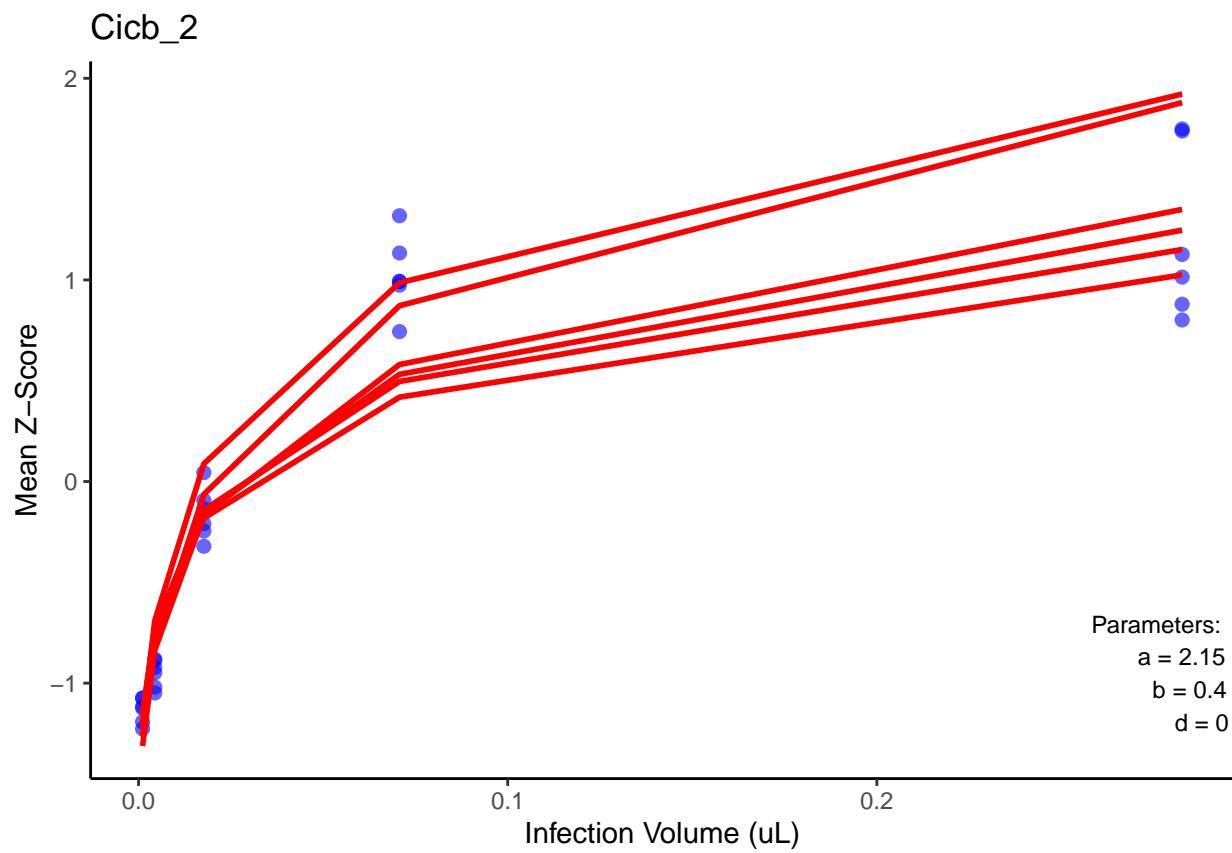
\$CTR_M2_O5_21



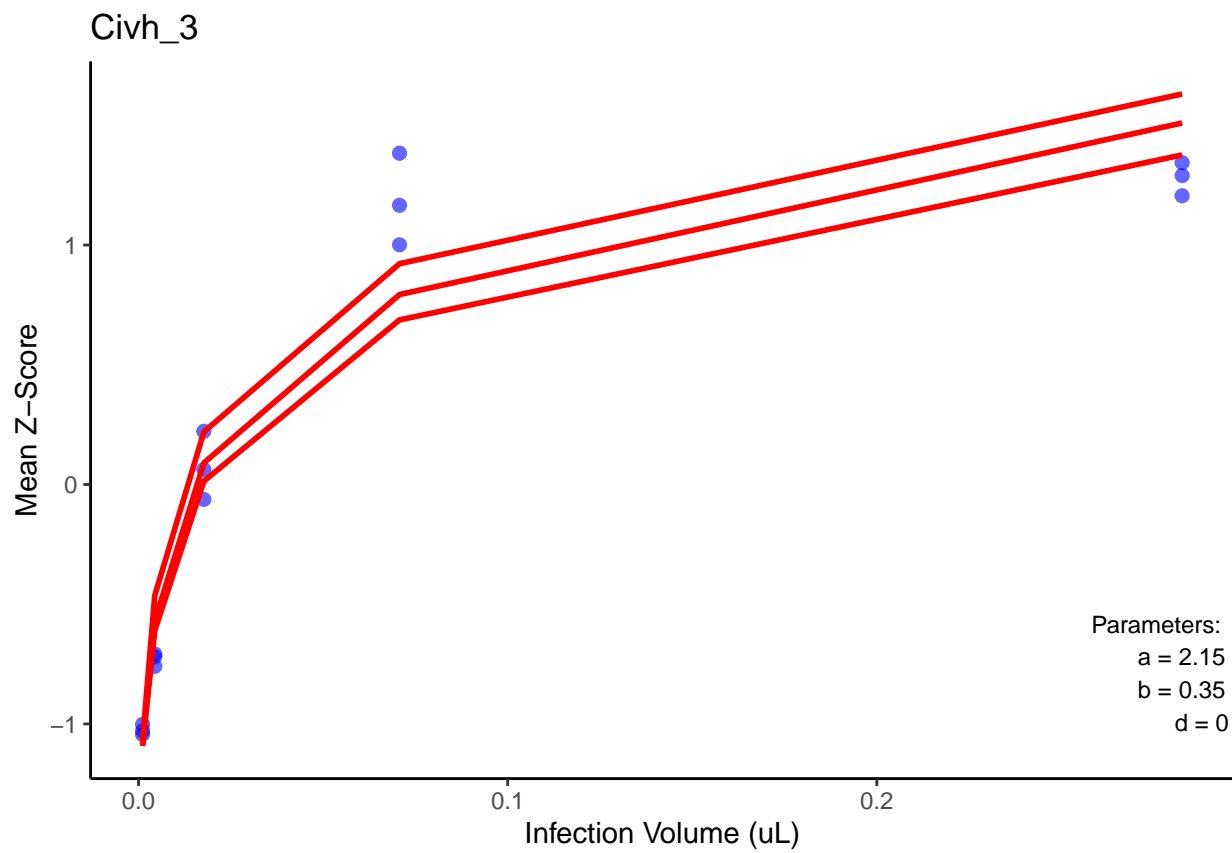
\$Ceik_1



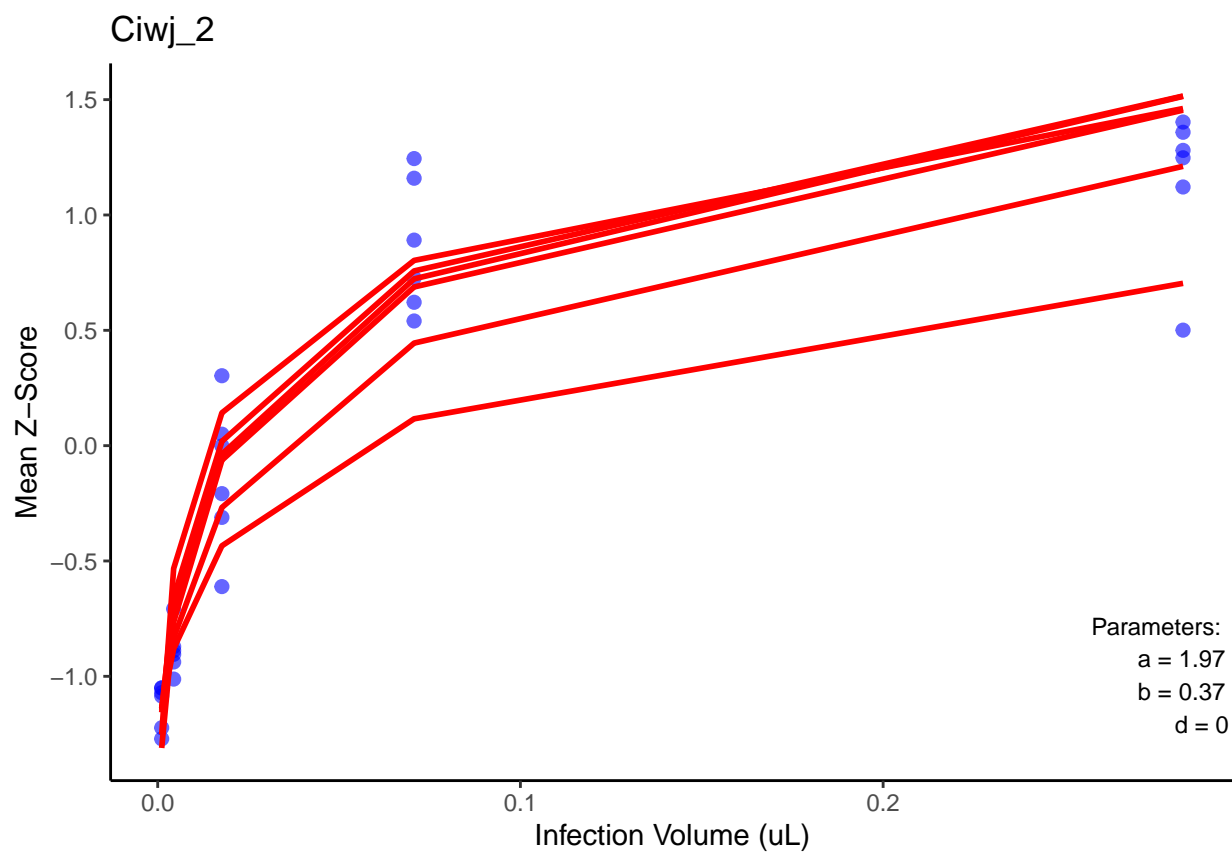
\$Cicb_2



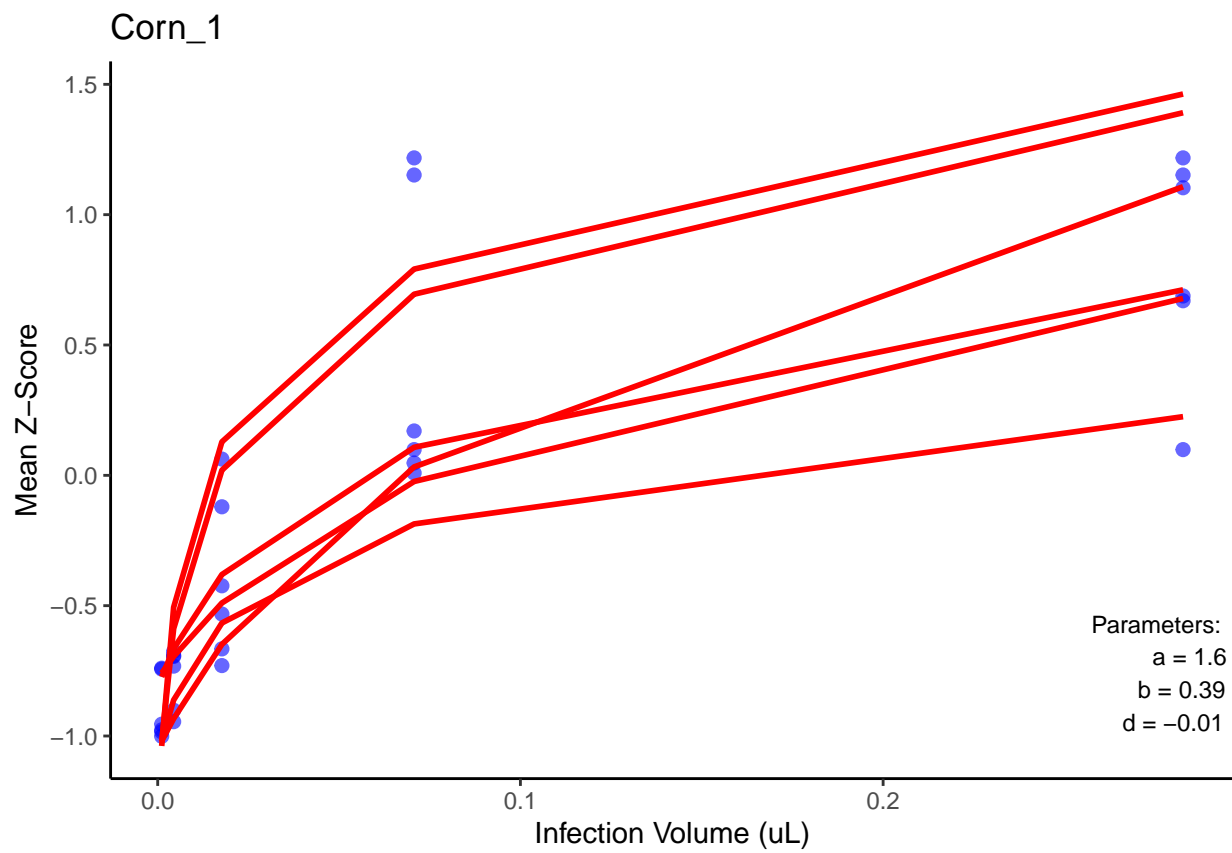
\$Civh_3



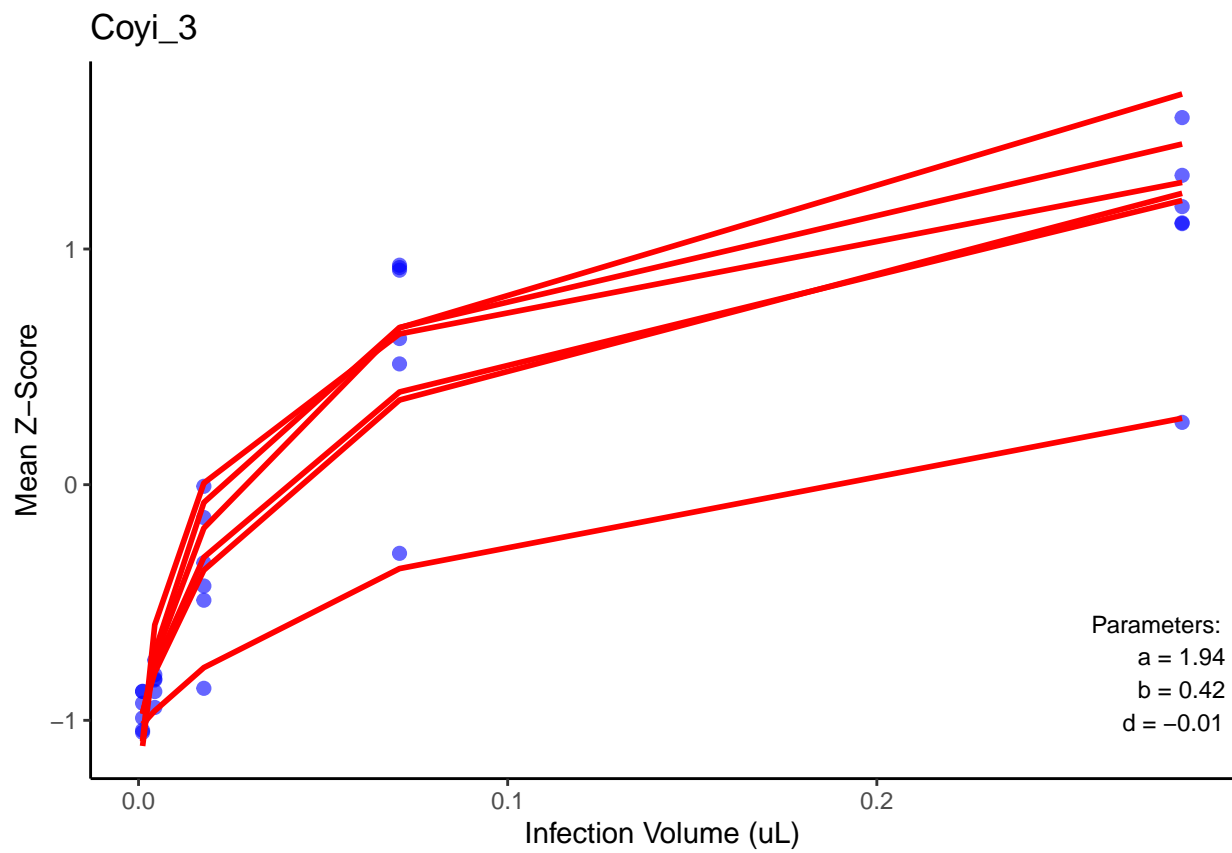
\$Ciwj_2



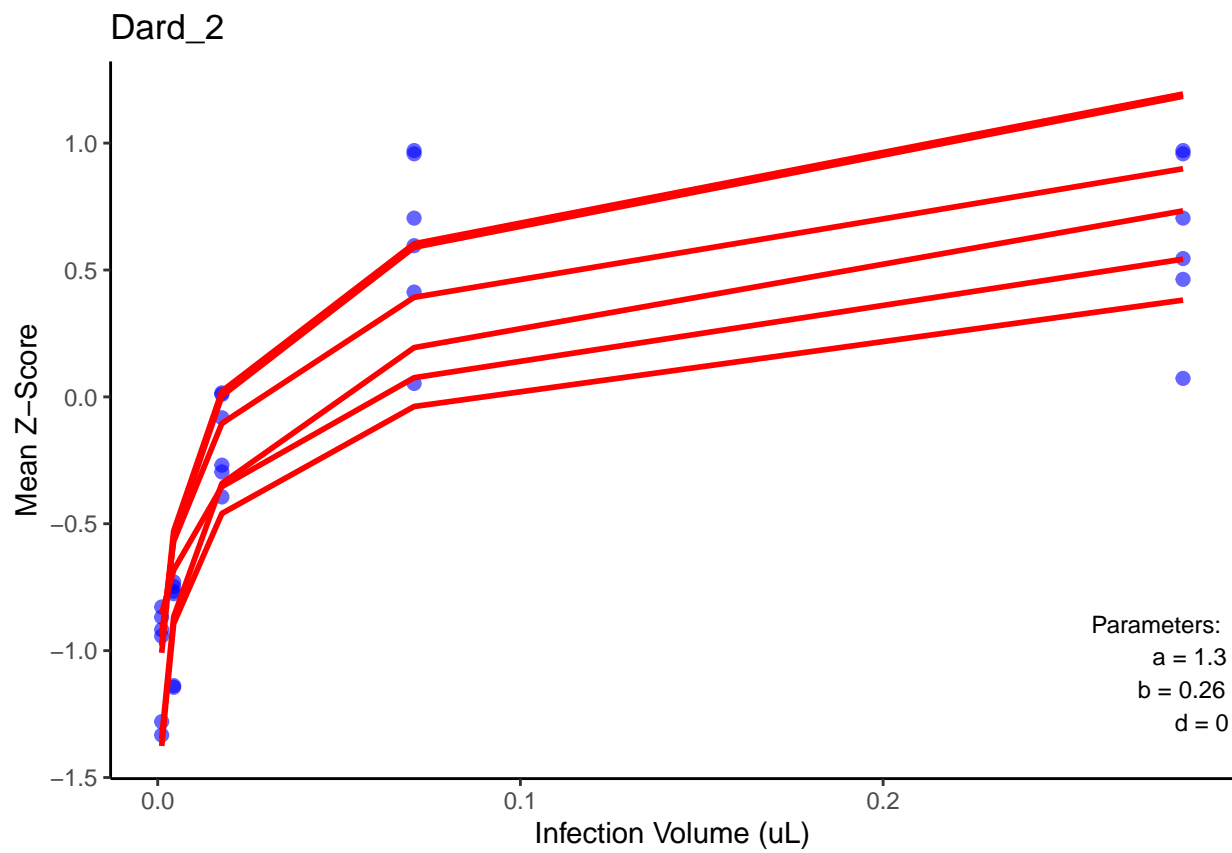
\$Corn_1



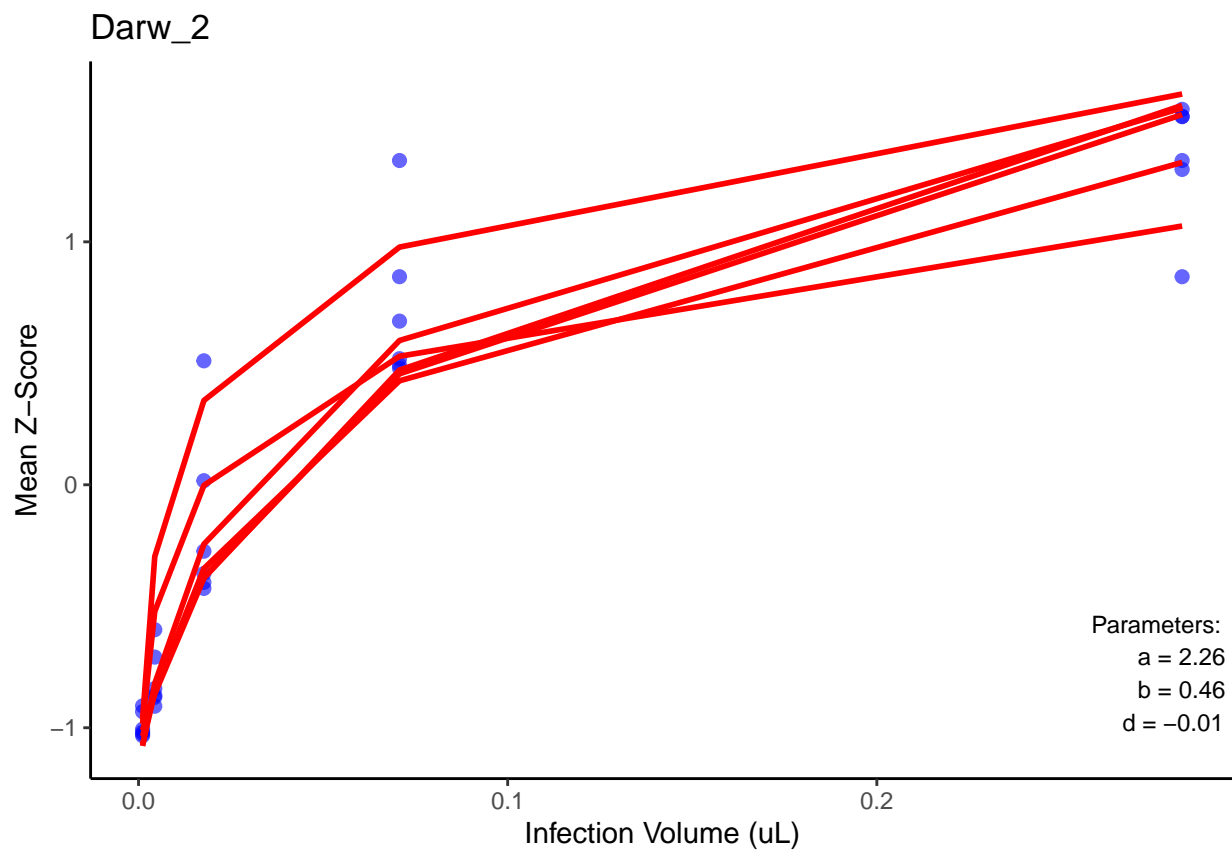
 ## \$Coyi_3



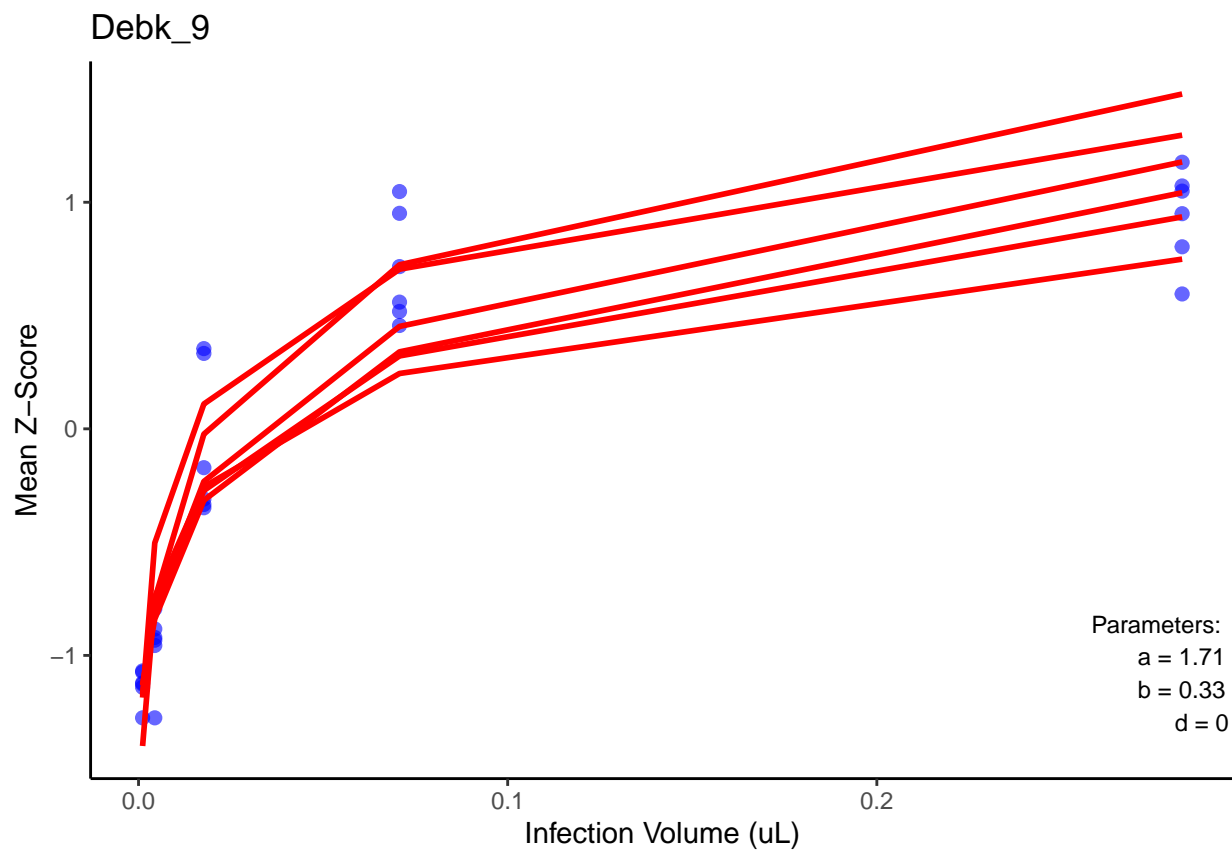
\$Dard_2



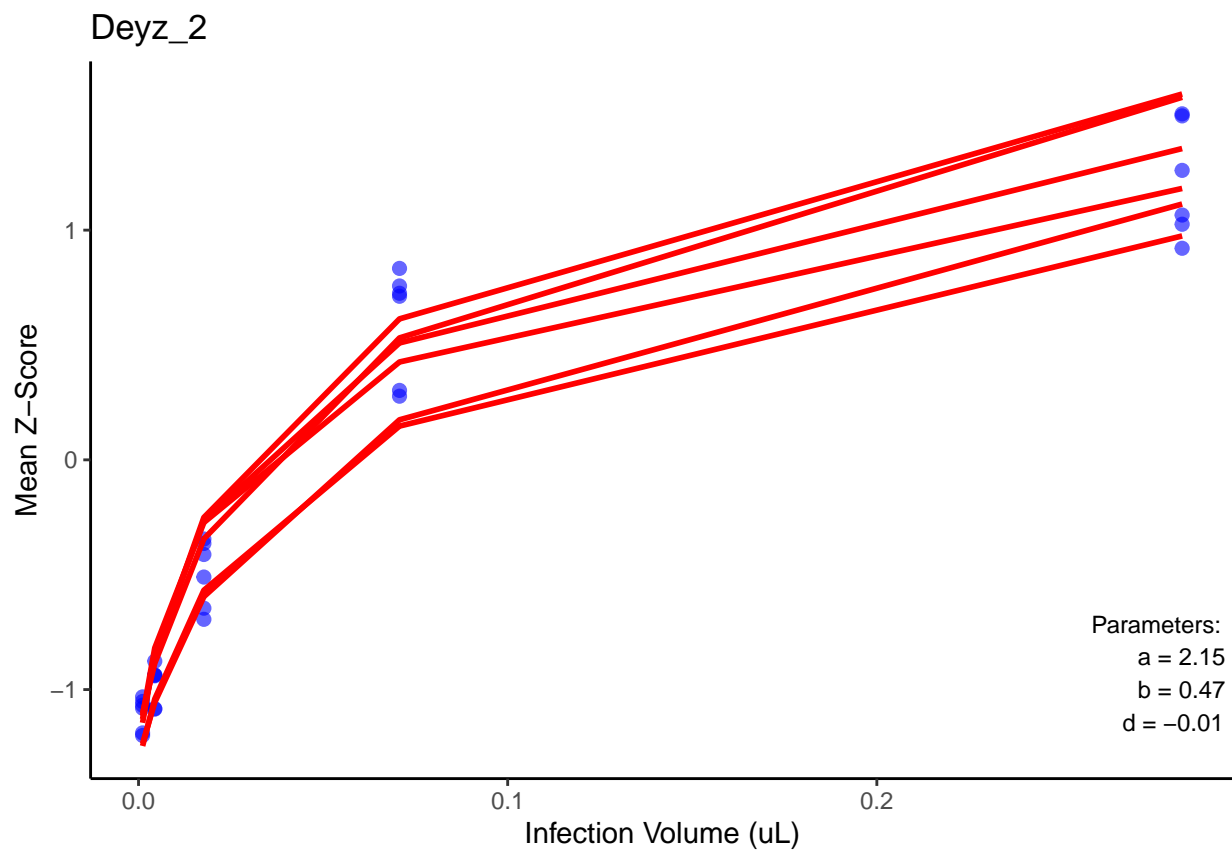
\$Darw_2



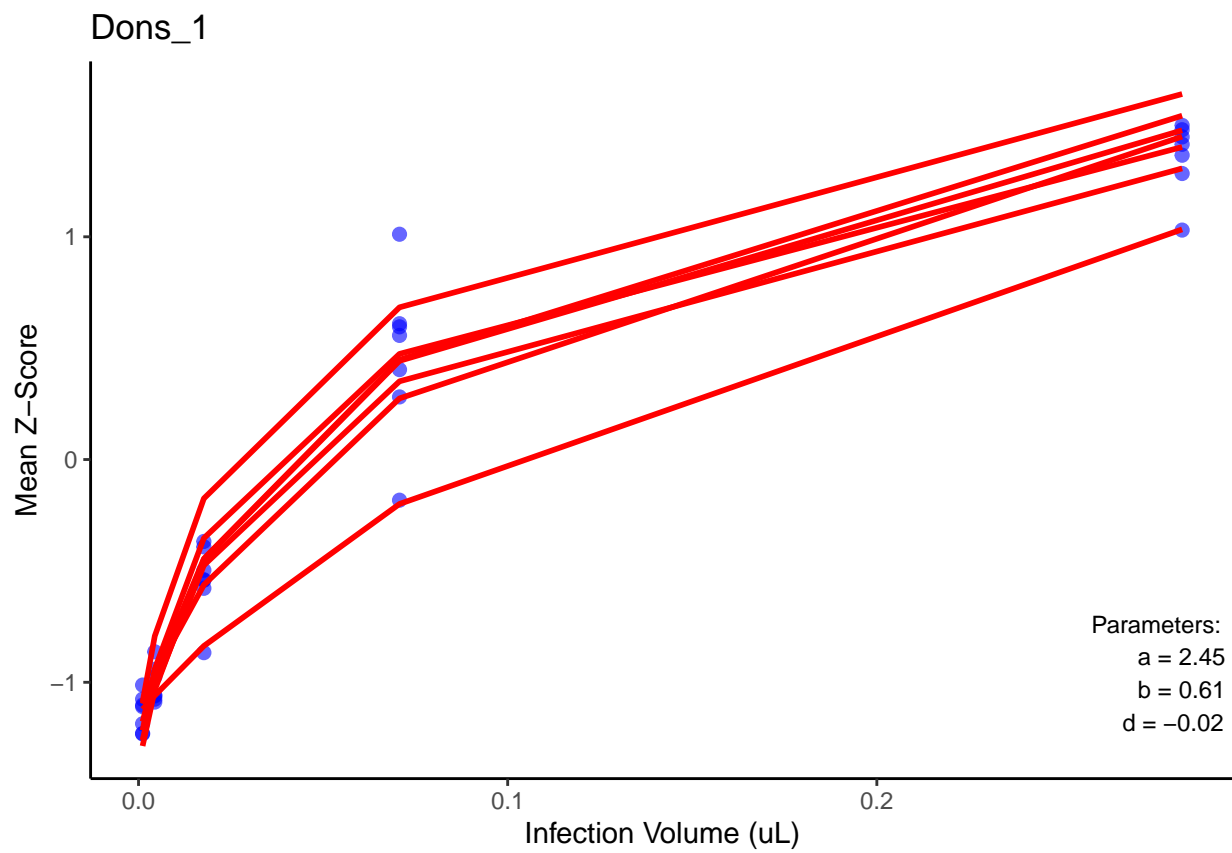
\$DebK_9



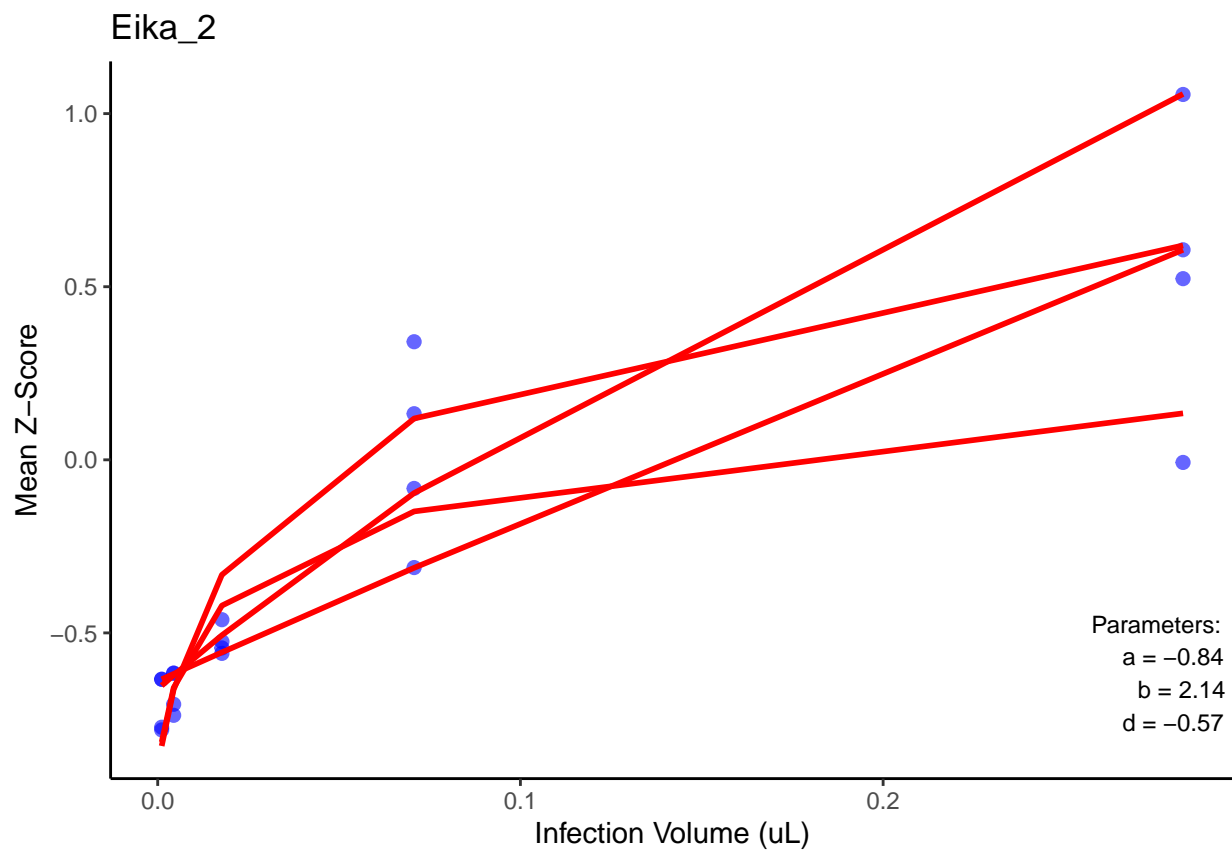
\$Deyz_2



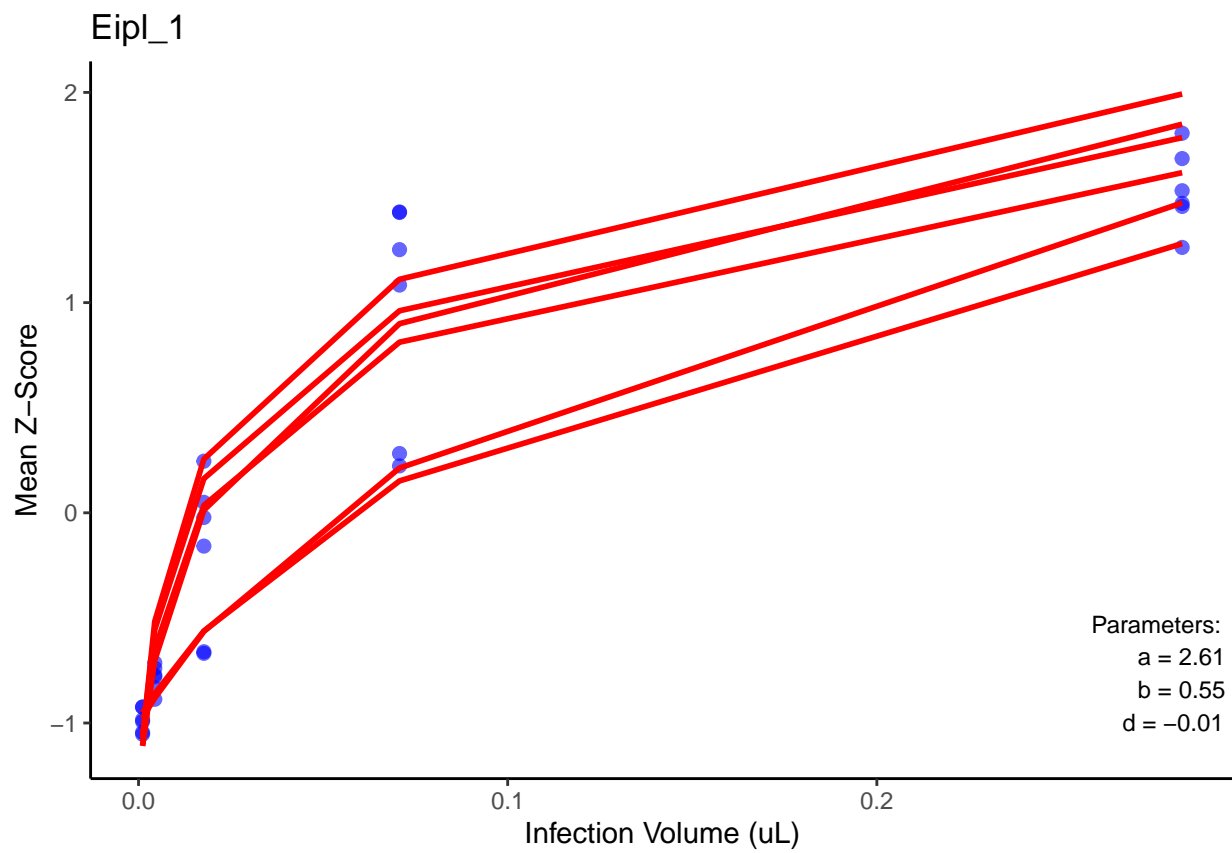
\$Dons_1



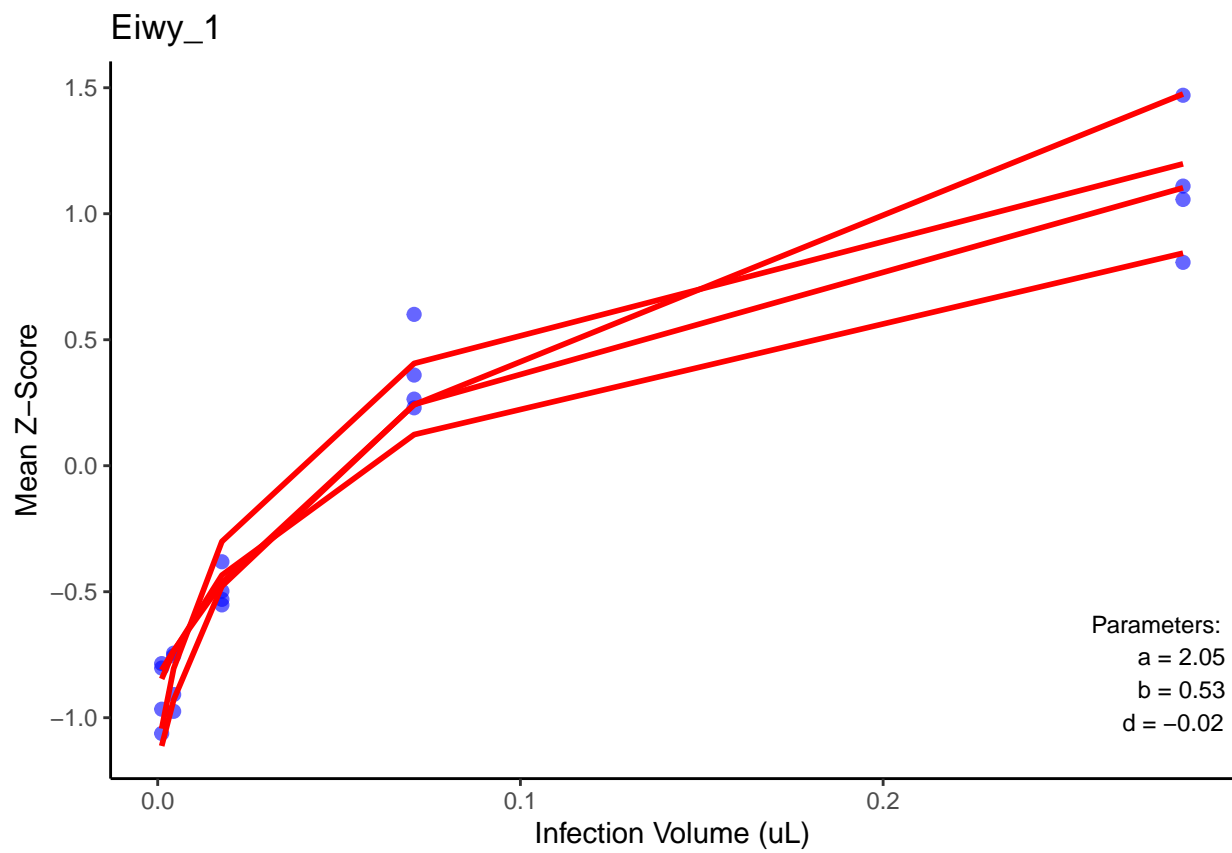
\$Eika_2



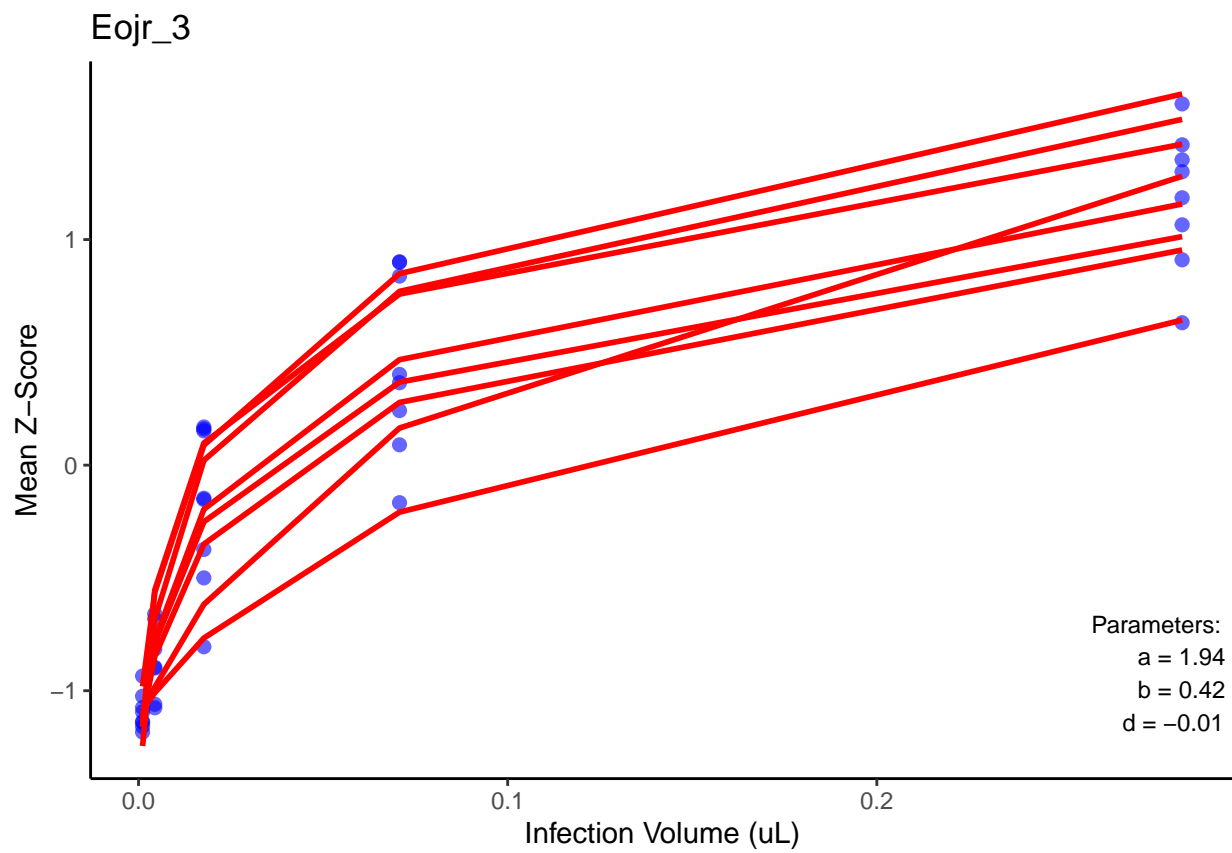
\$Eipl_1



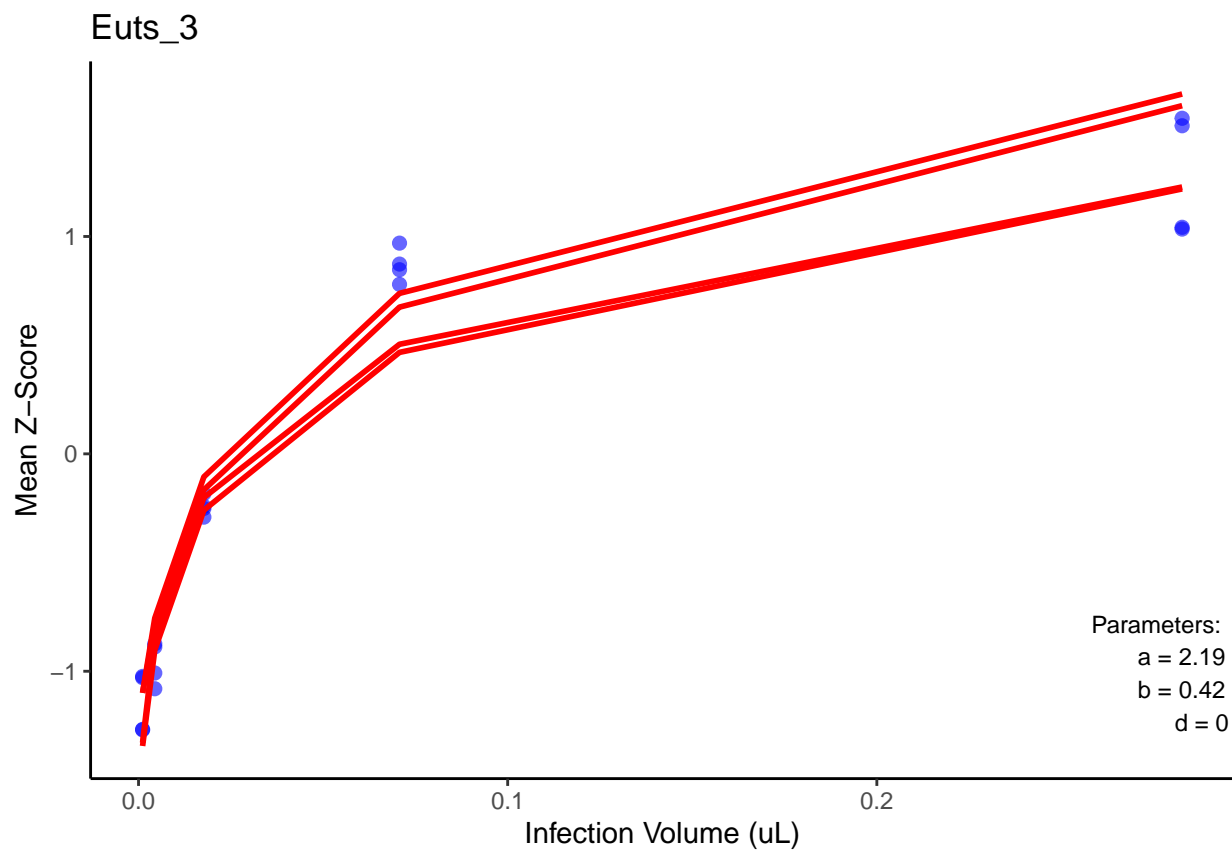
\$Eiwy_1



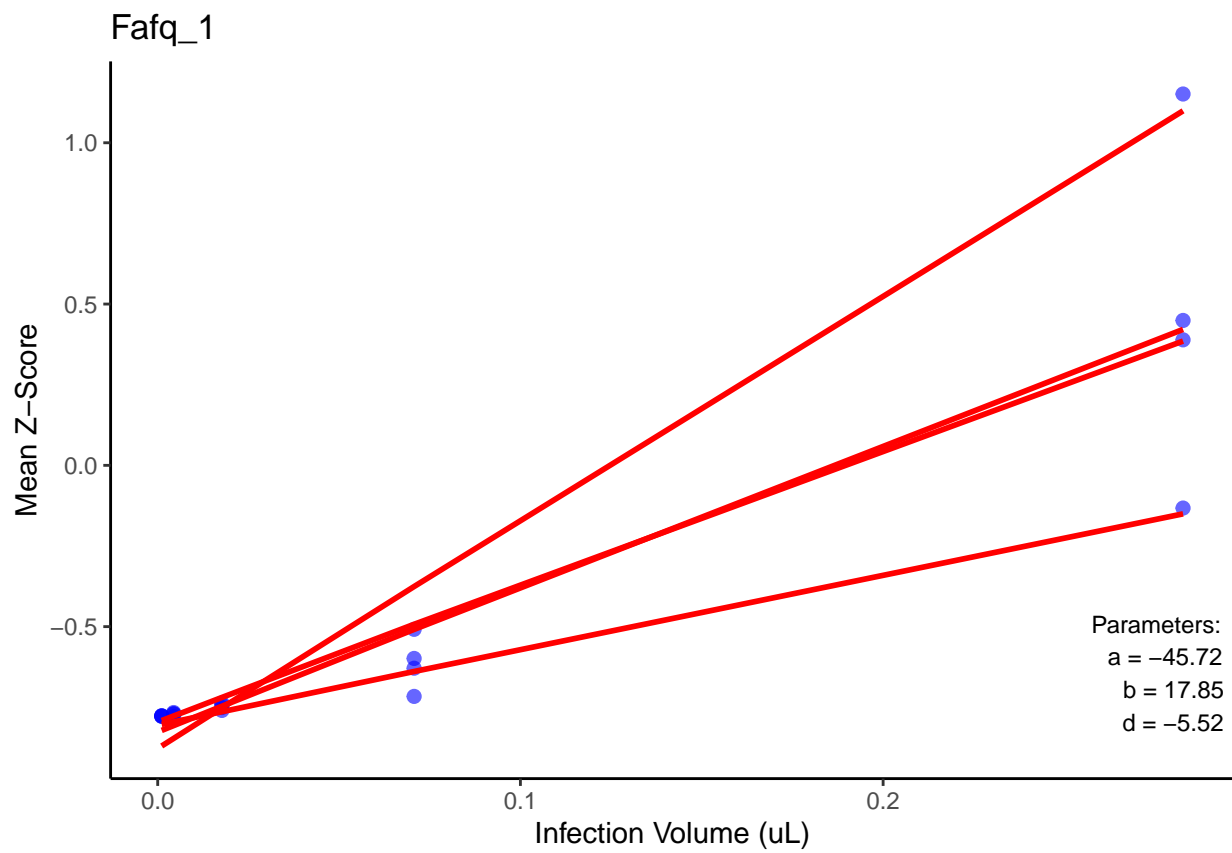
\$Eojr_3



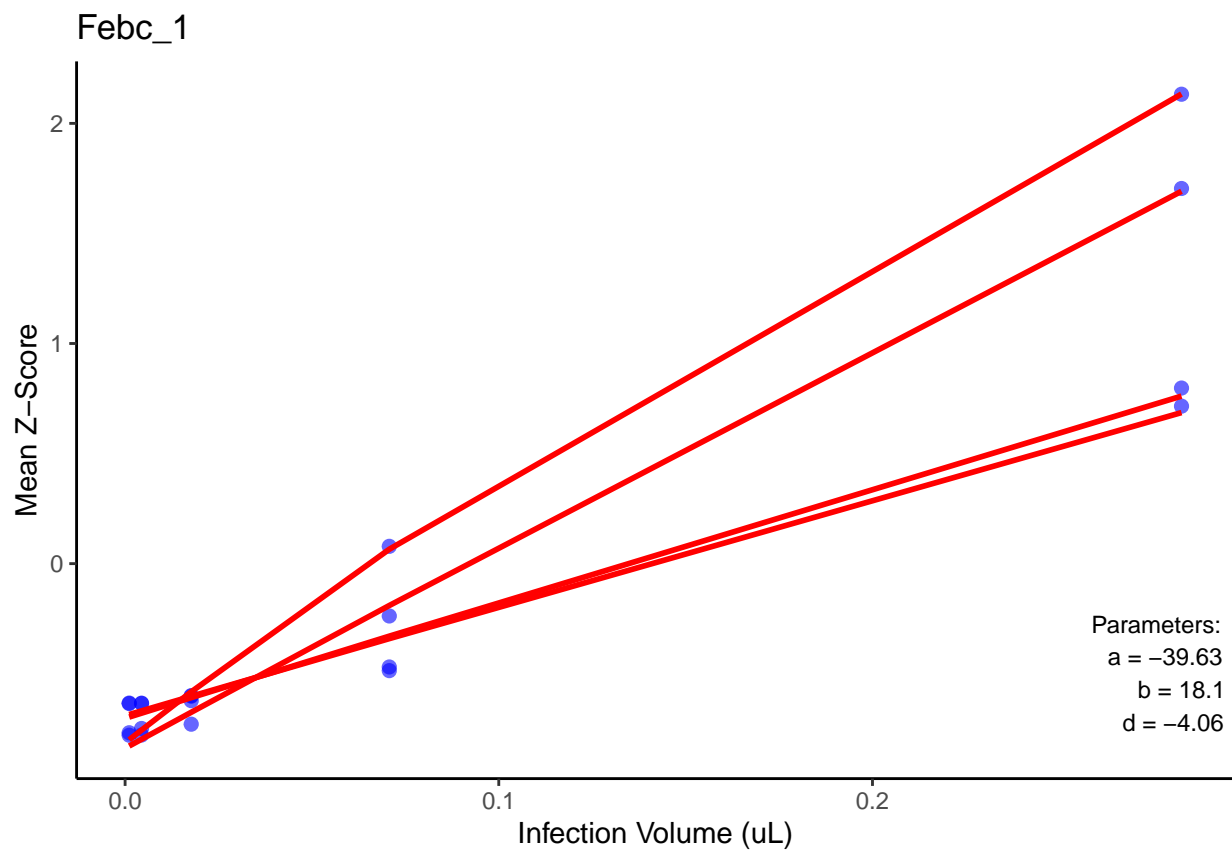
\$Euts_3



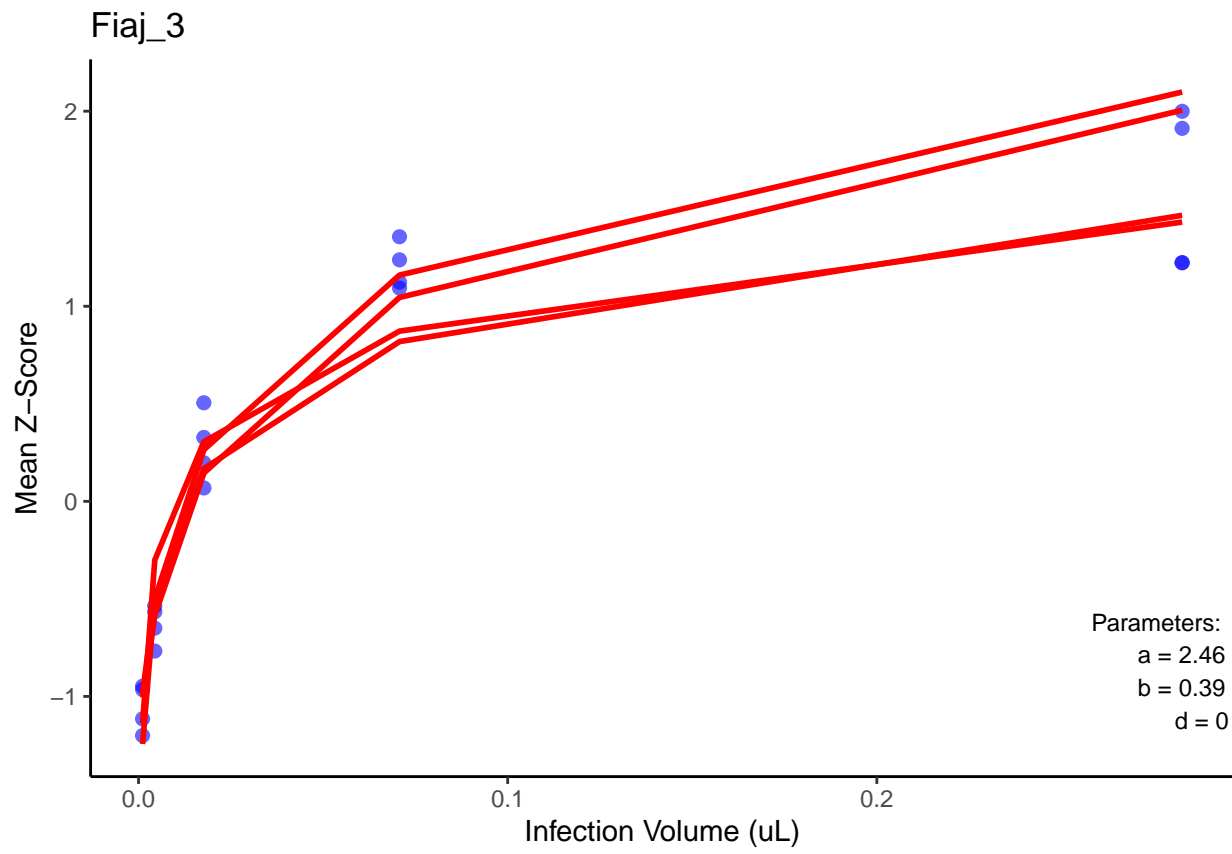
\$Fafq_1



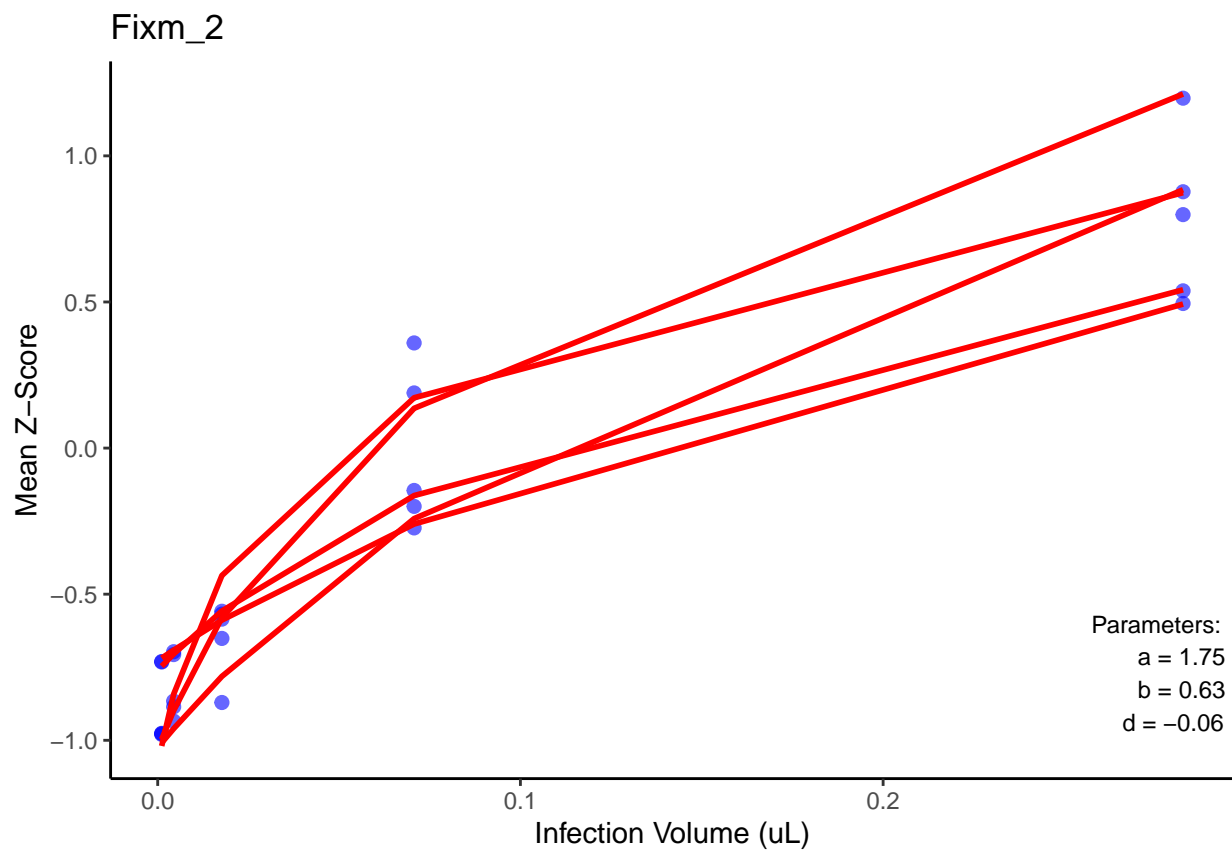
 ## \$Febc_1



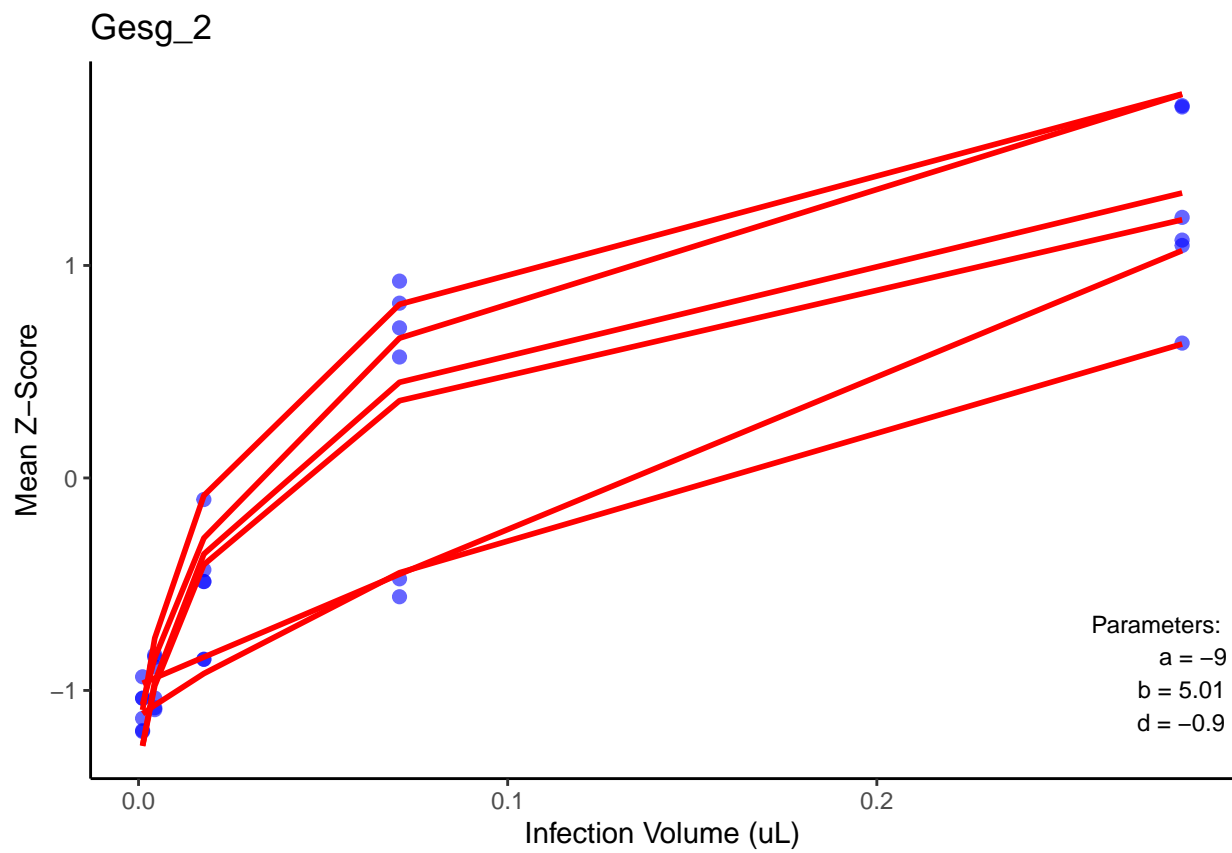
 ## \$Fiaj_3



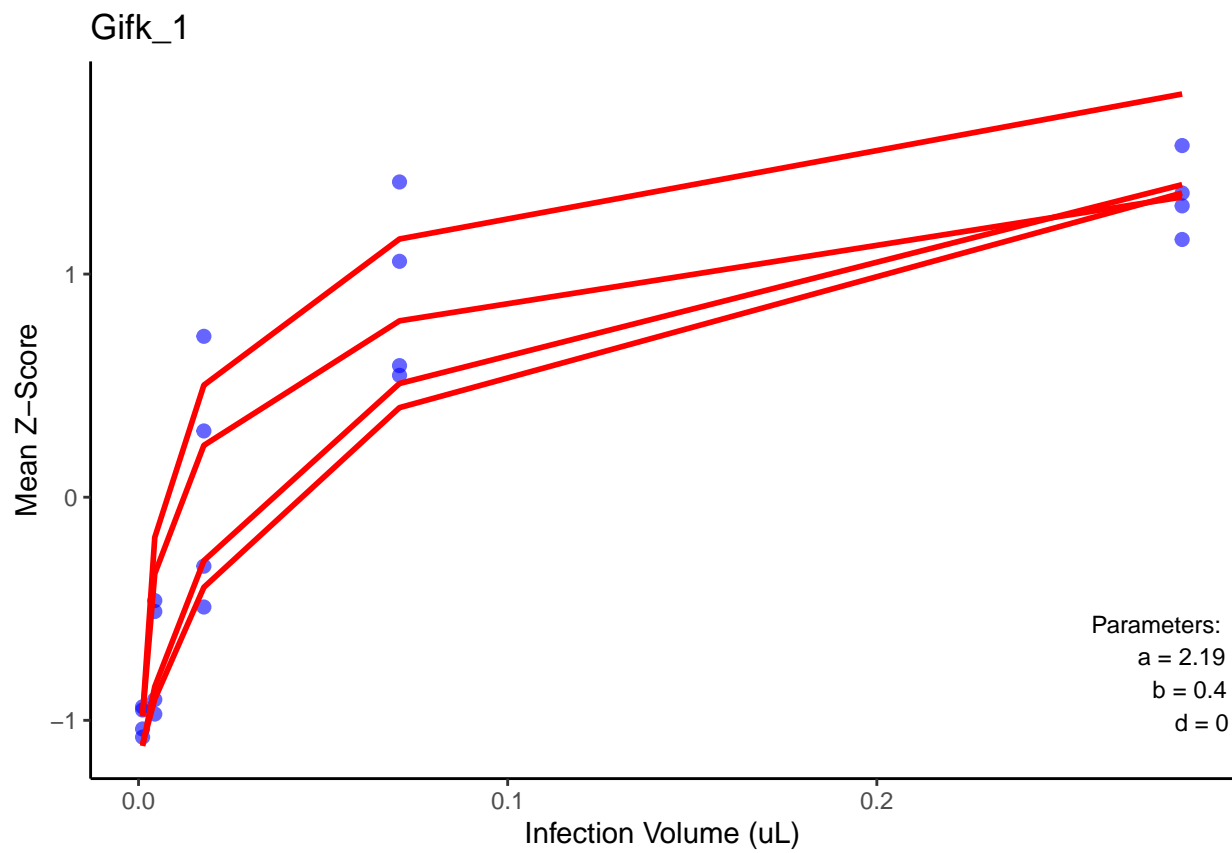
```
##
## $Fixm_2
```



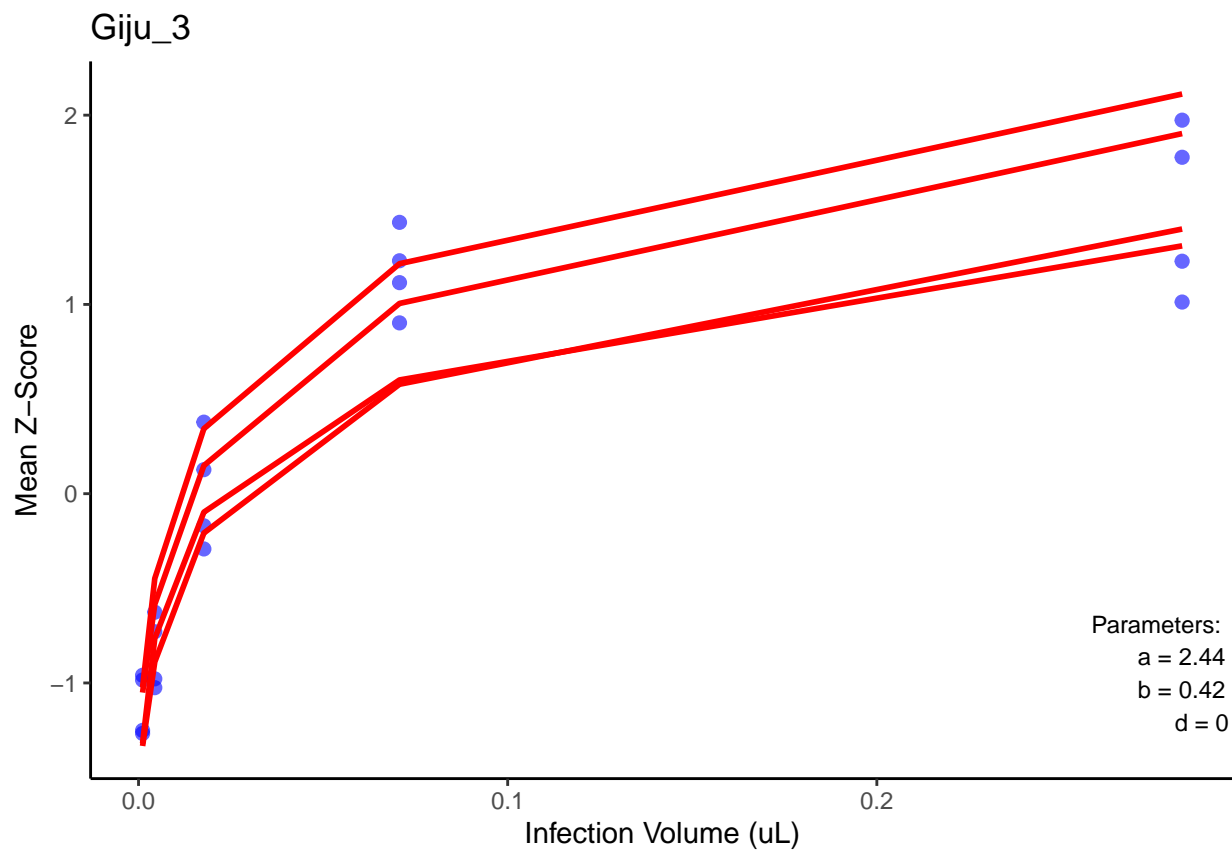
 ## \$Gesg_2



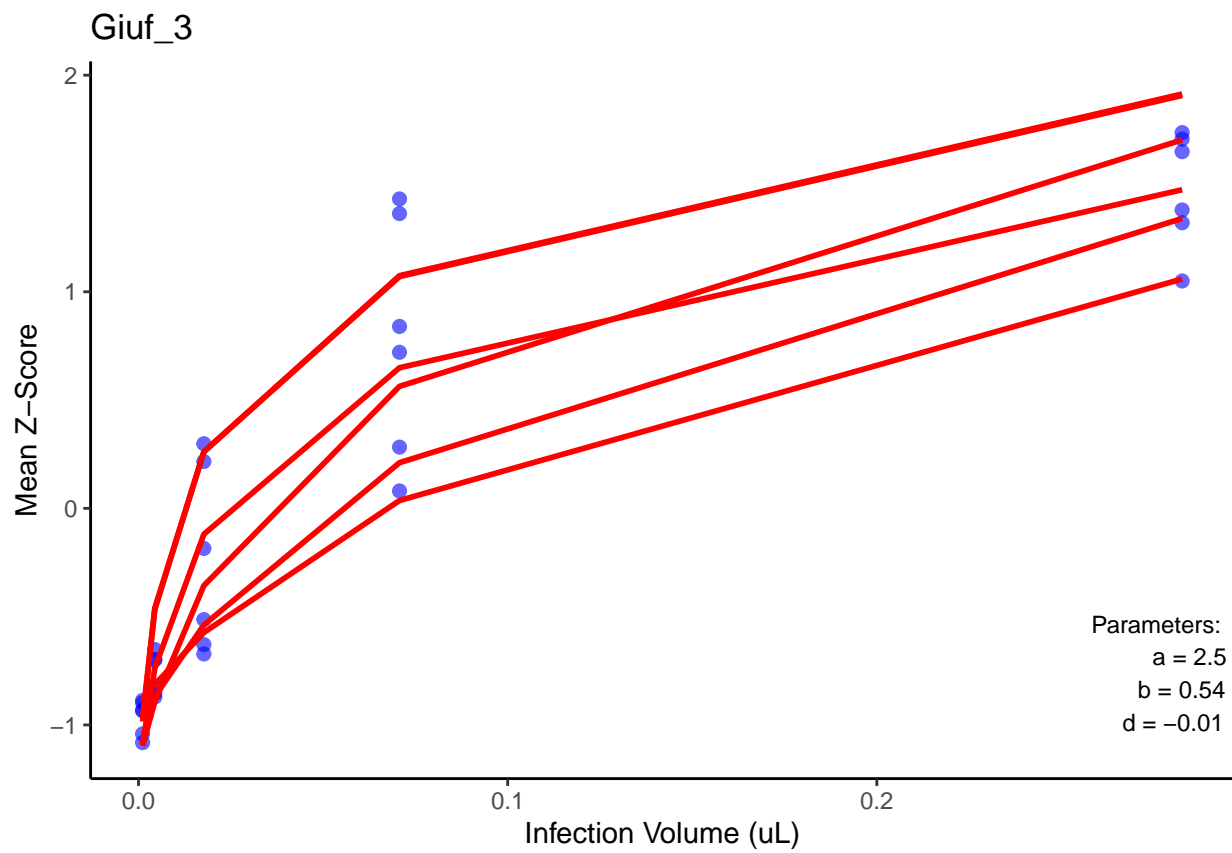
\$Gifk_1



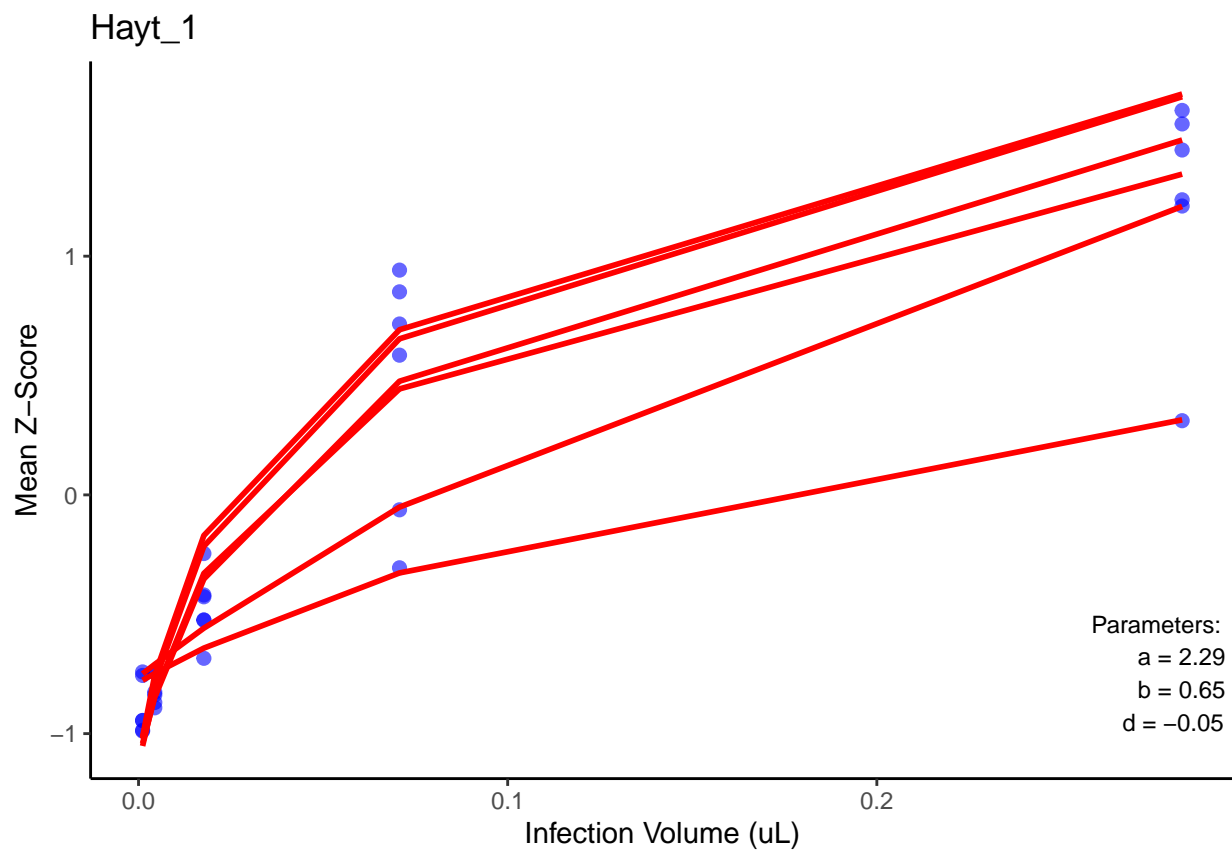
 ## \$Giju_3



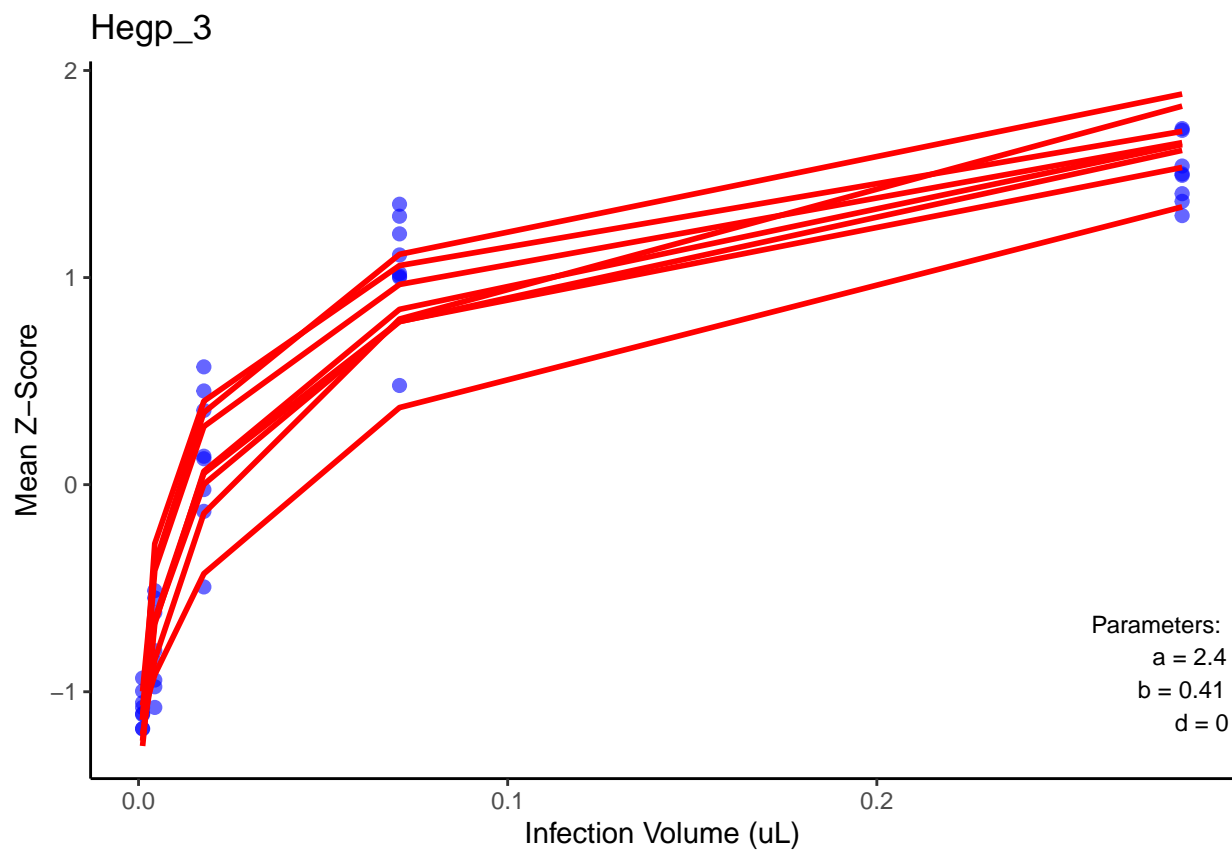
 ## \$Giuf_3



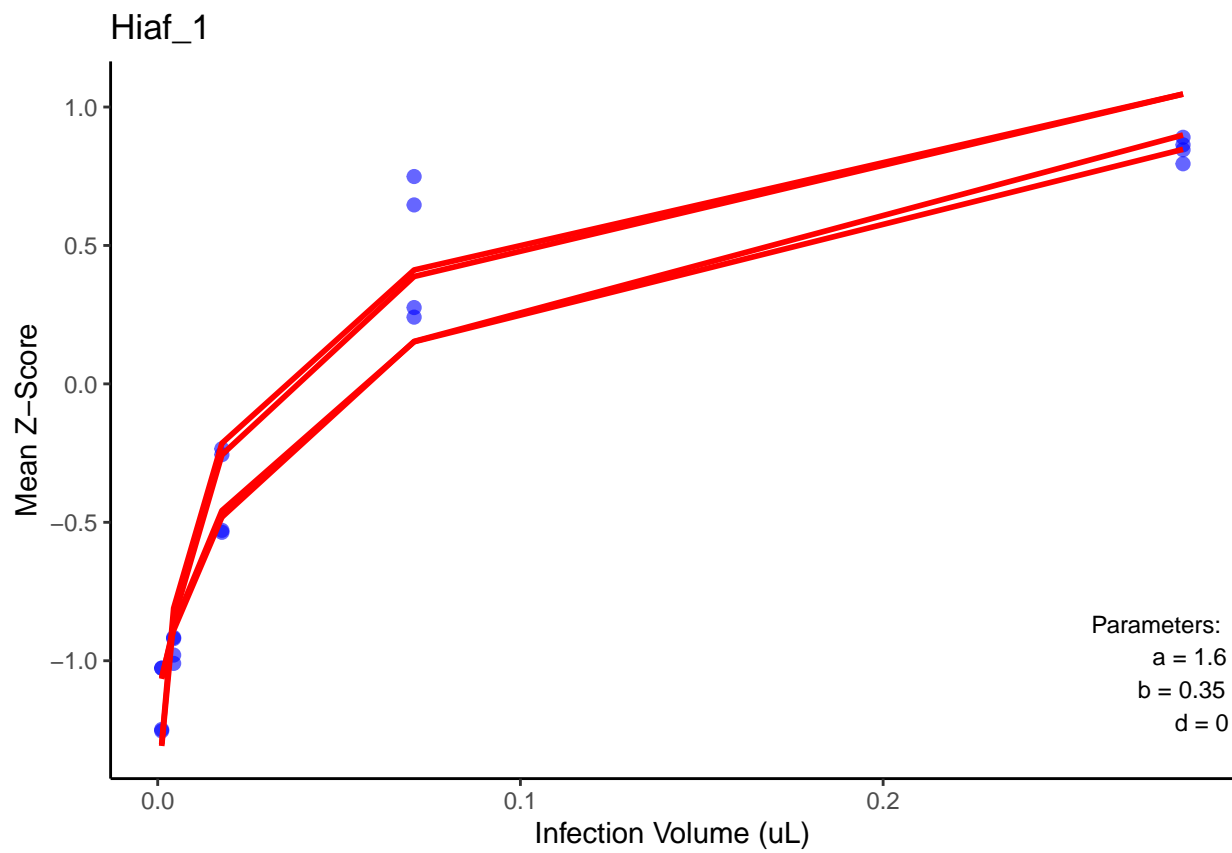
\$Gost_1



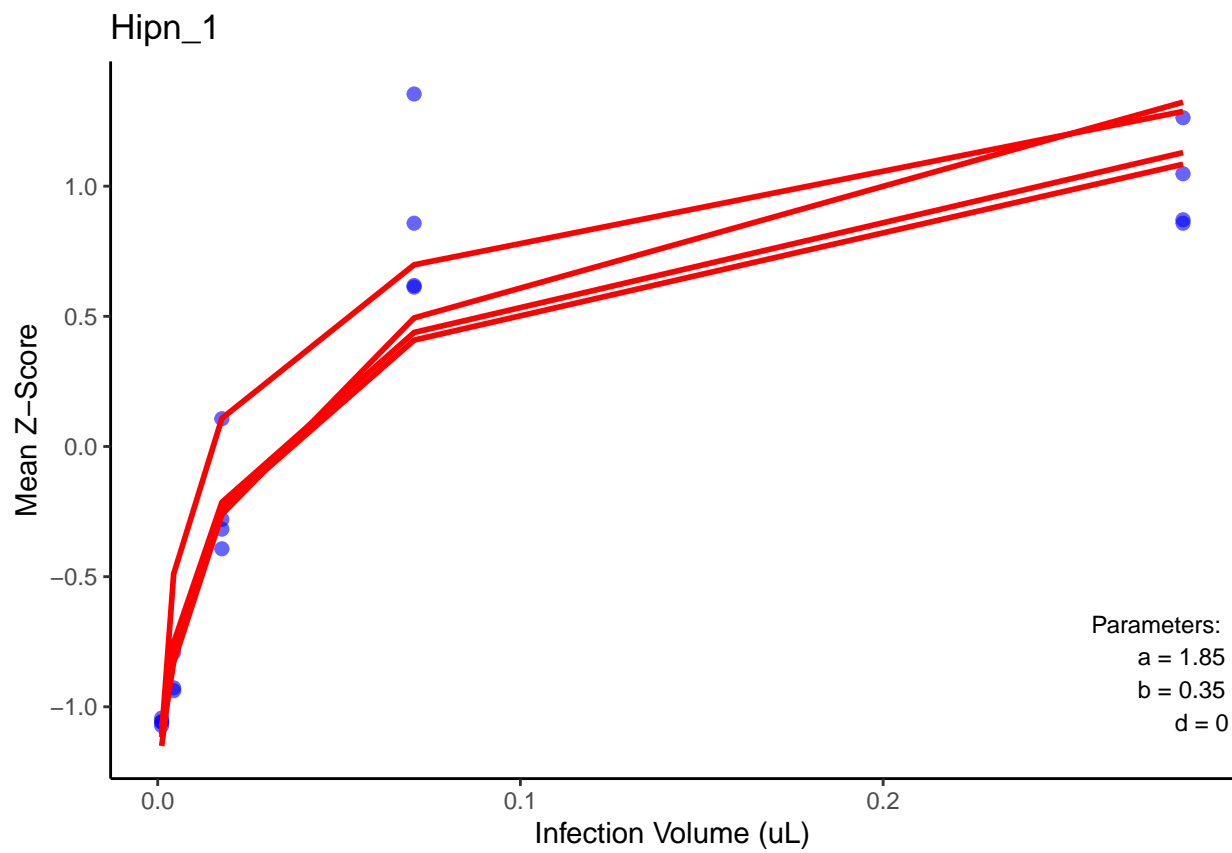
\$Hegp_3



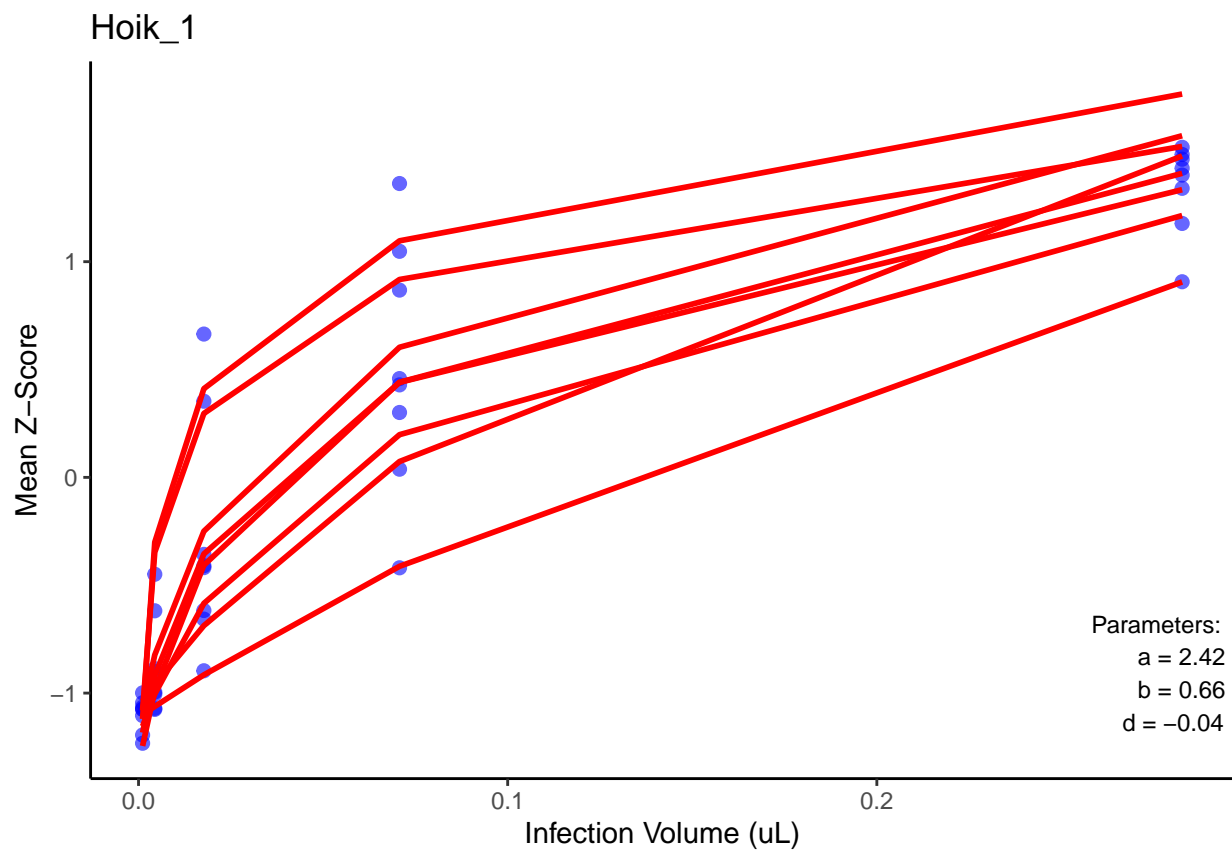
\$Hiaf_1



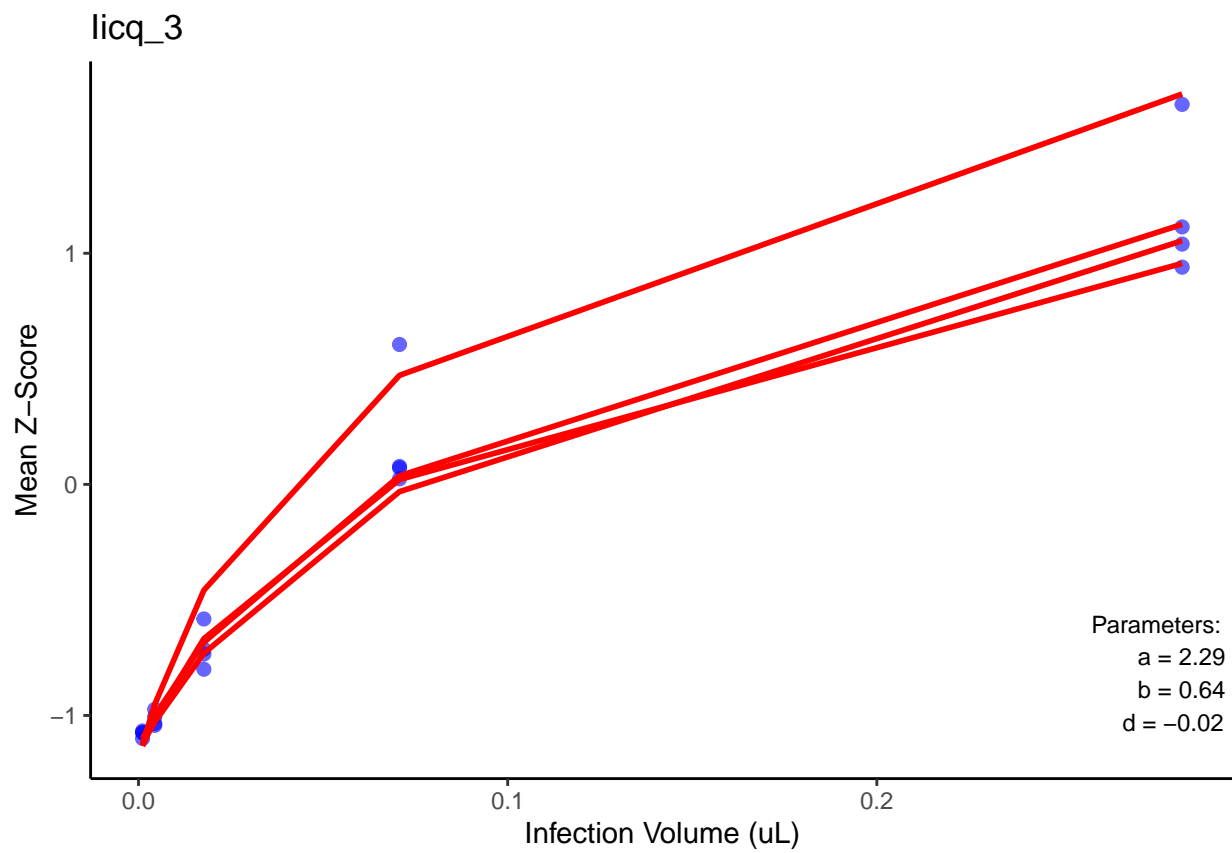
\$Hipn_1



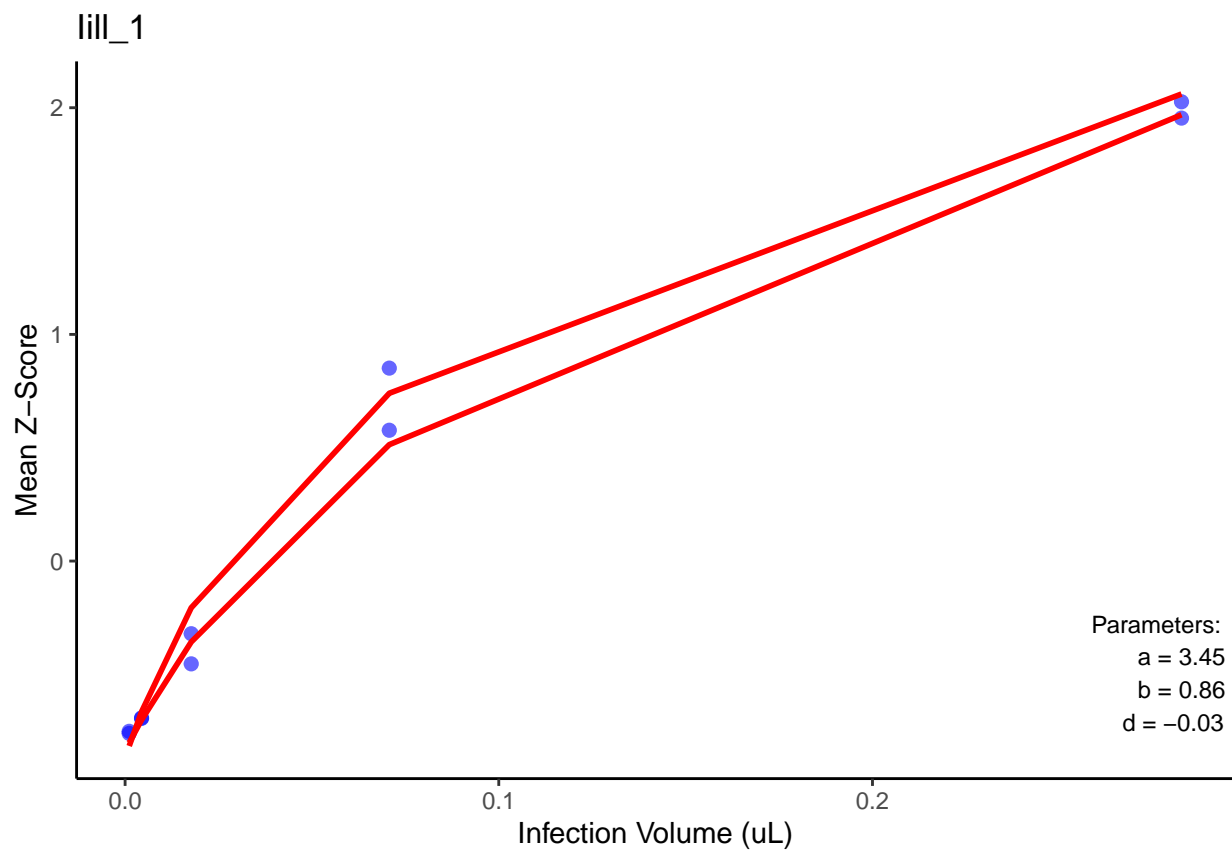
\$Hoik_1



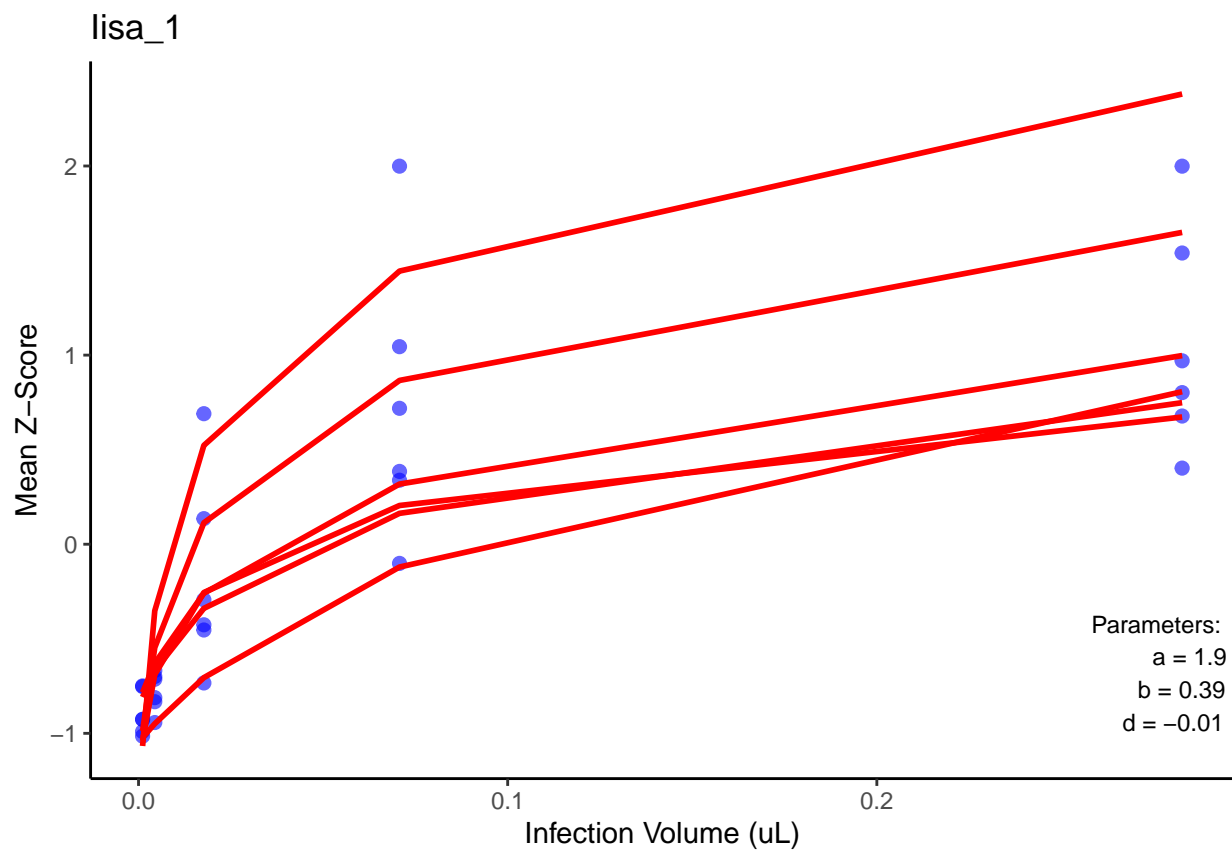
\$Iicq_3

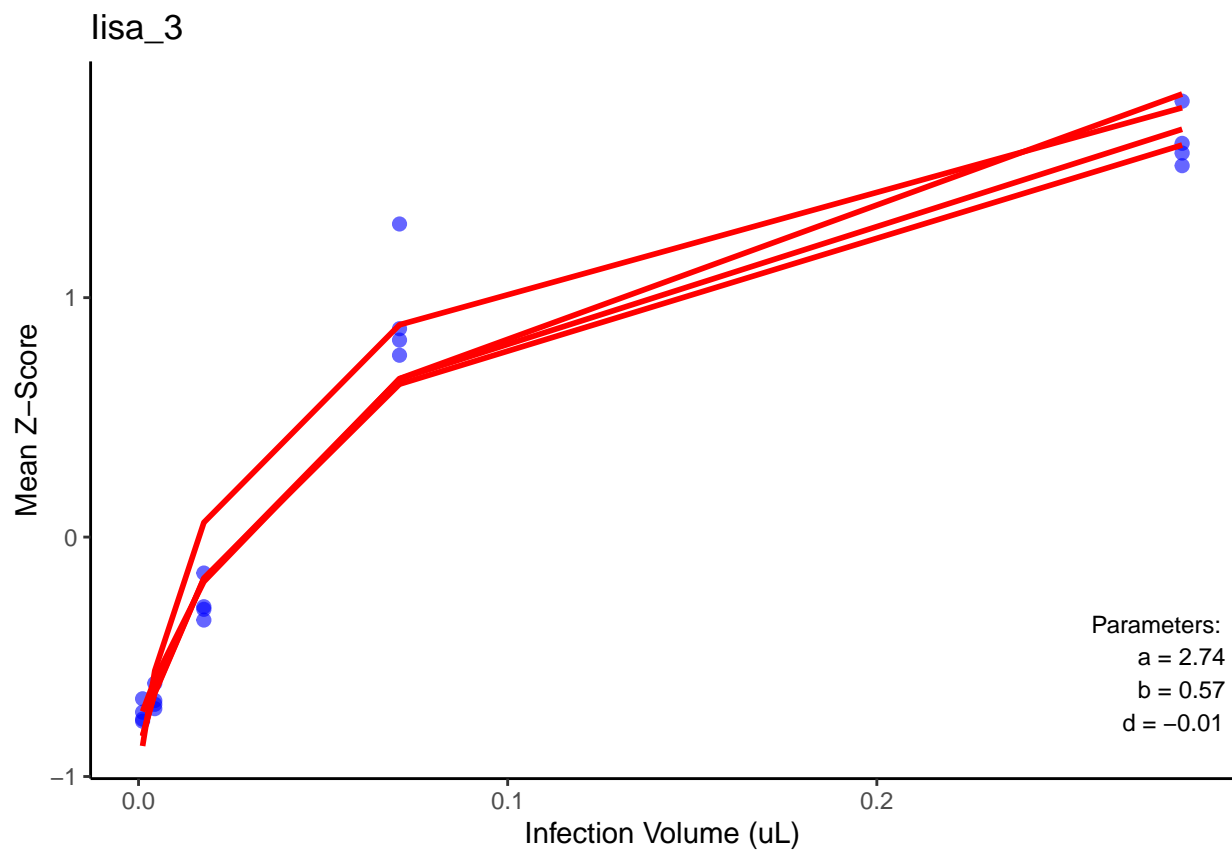


 ## \$Iill_1

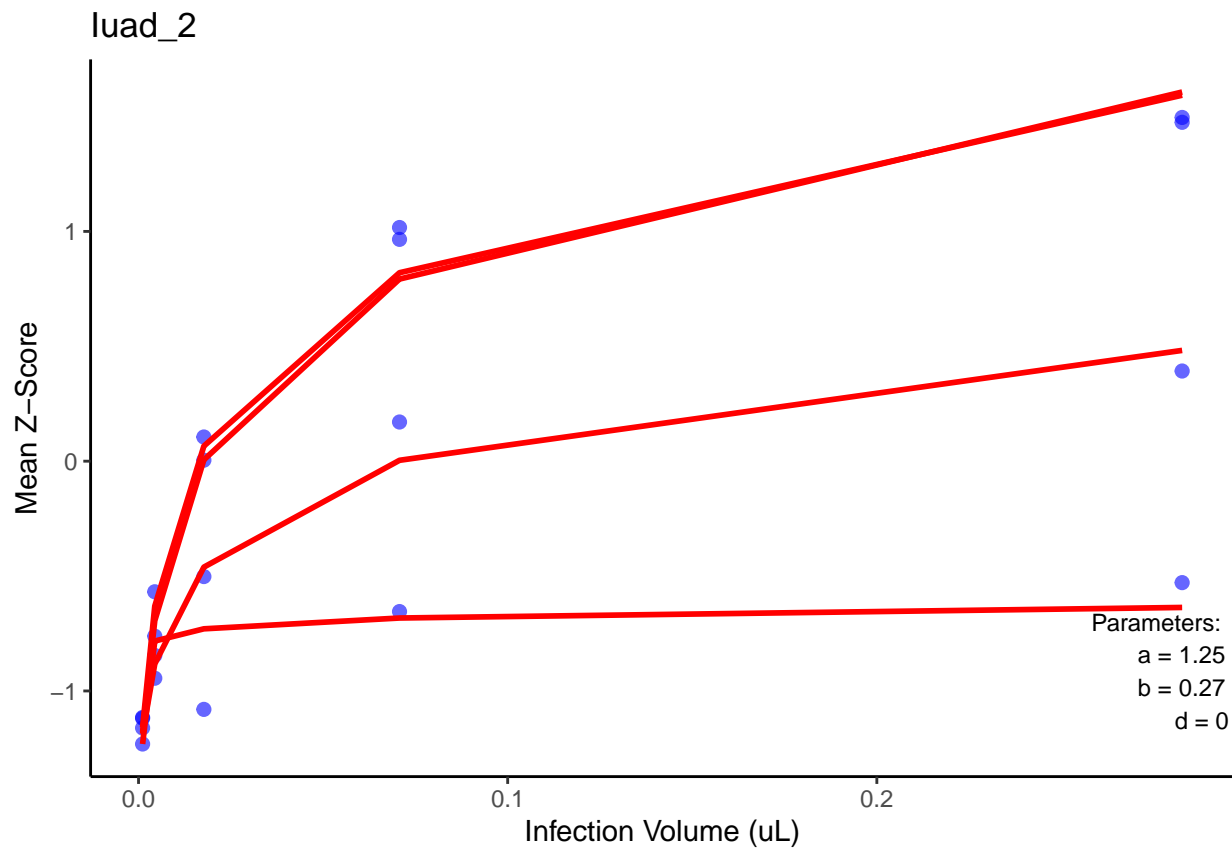


 ## \$Iisa_1

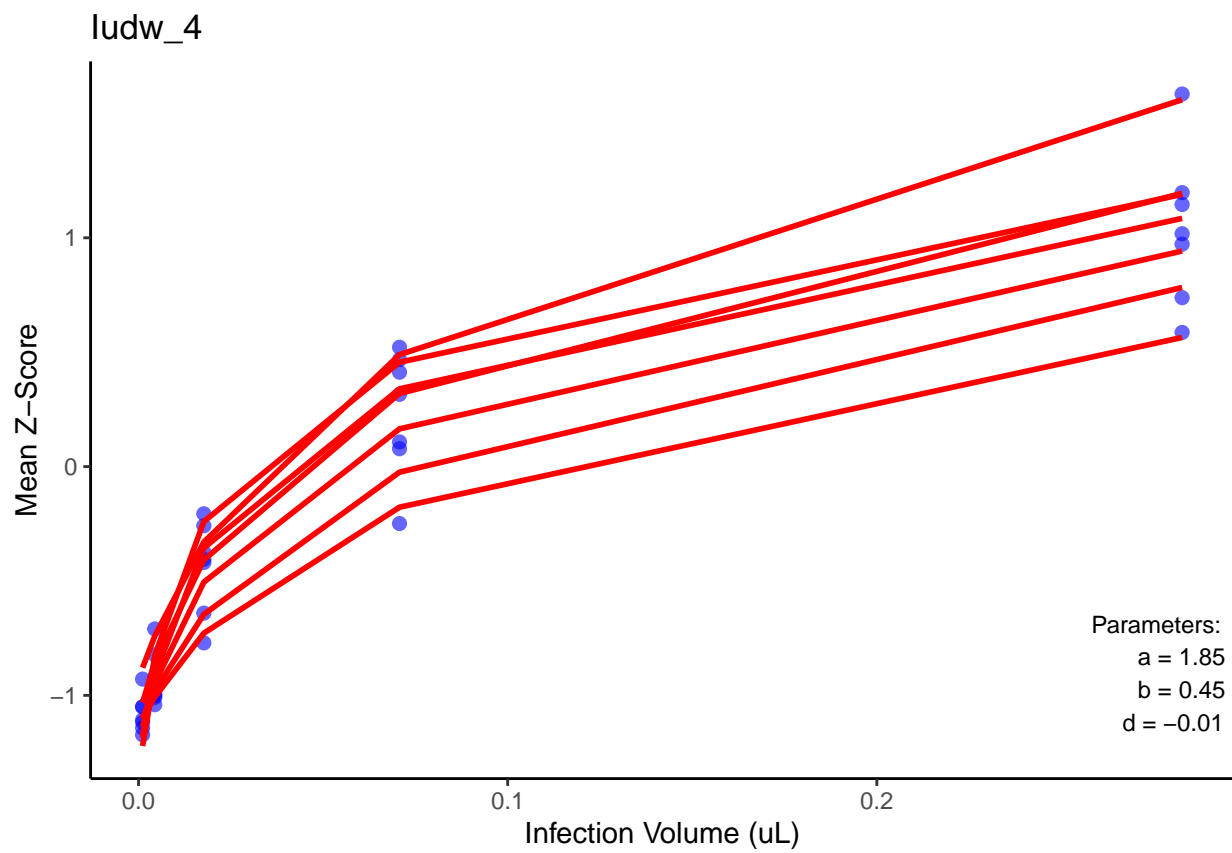




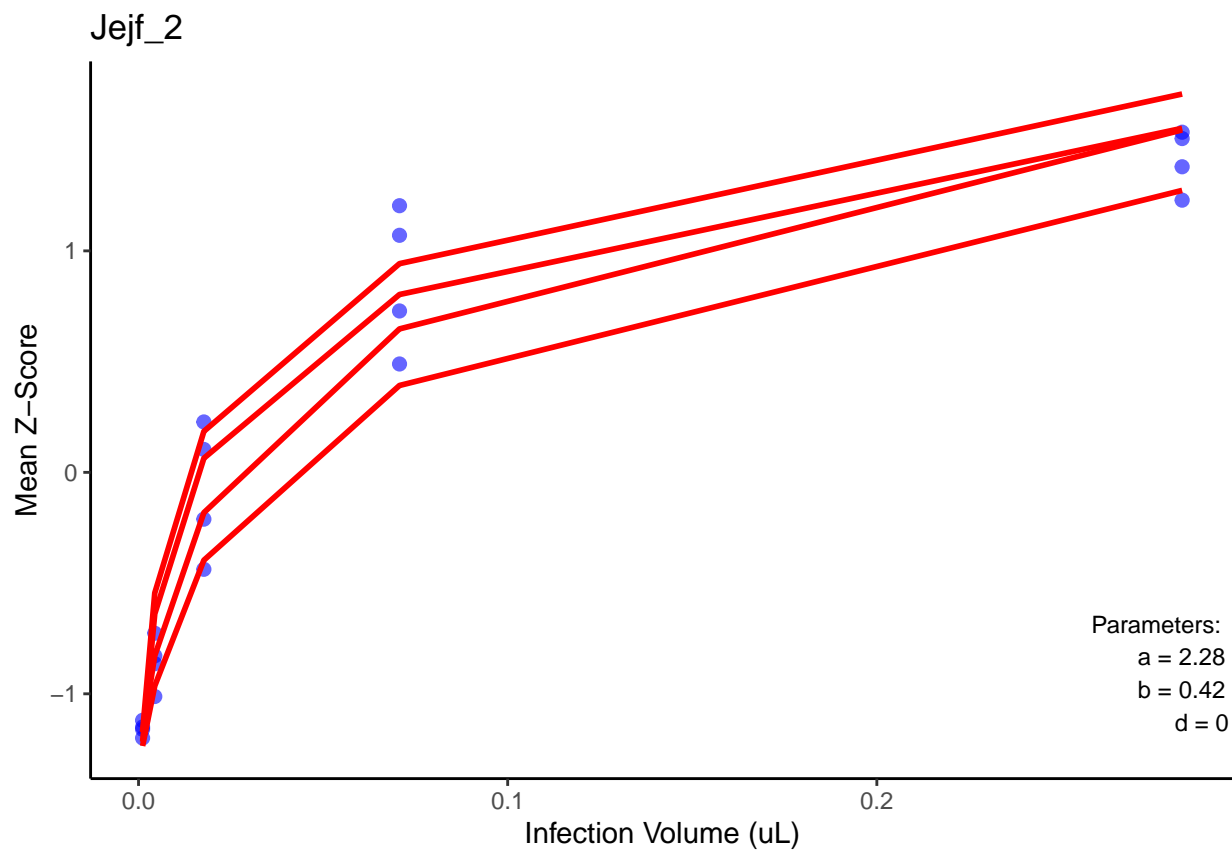
\$Iuad_2



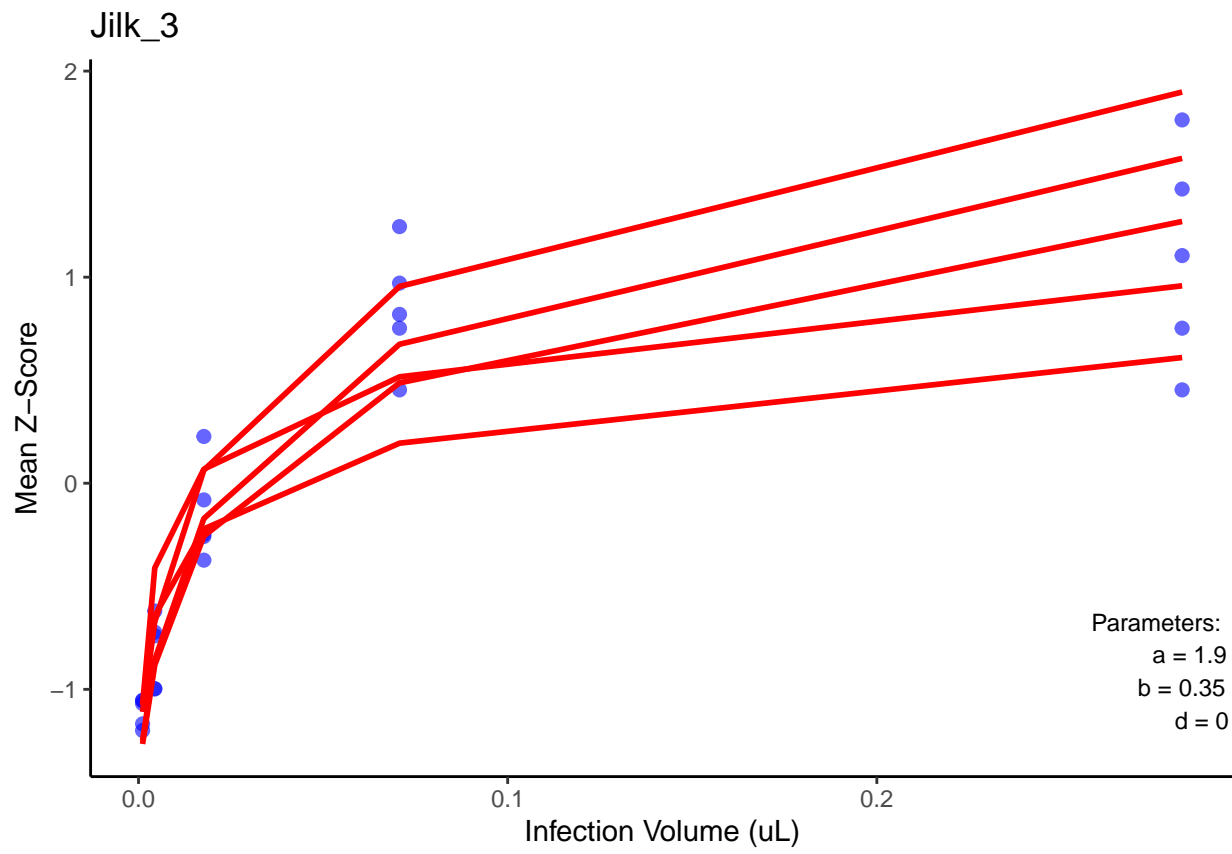
 ## \$Iudw_4



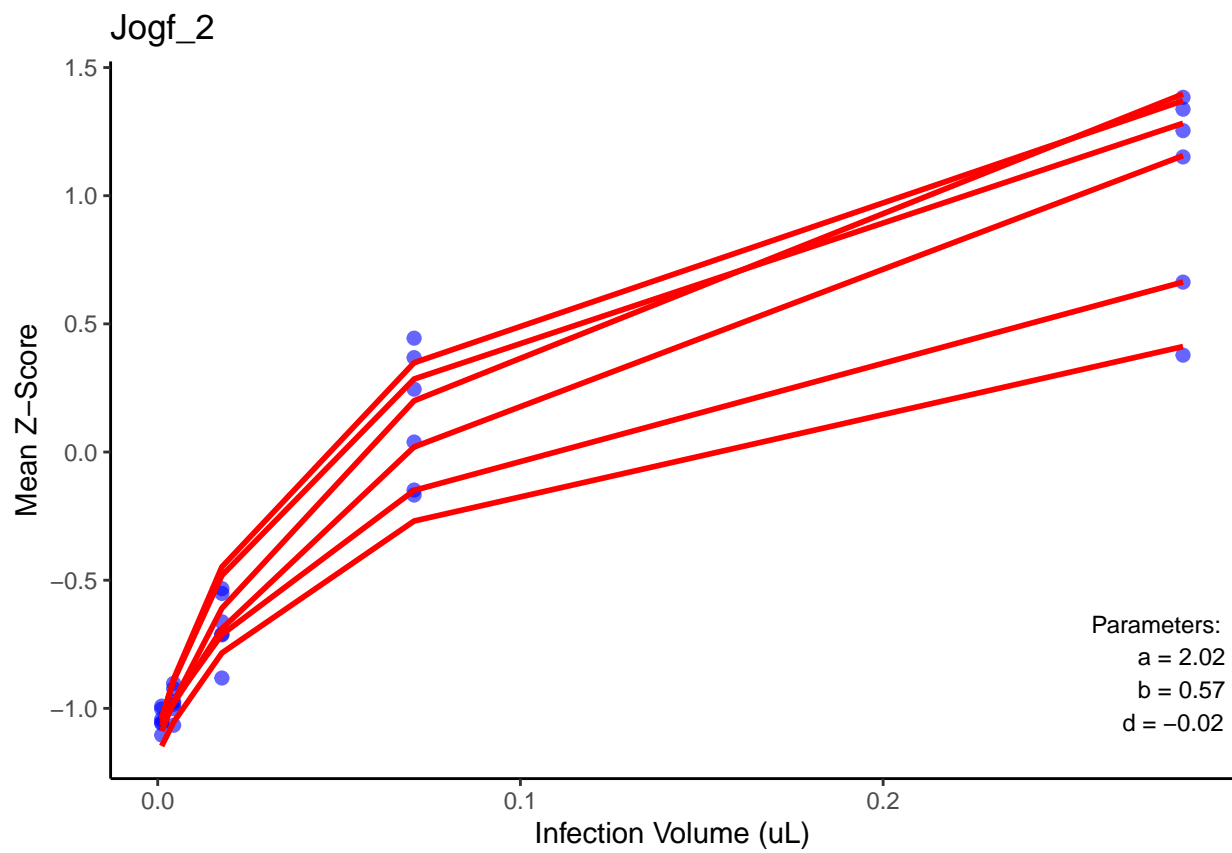
\$Jejf_2



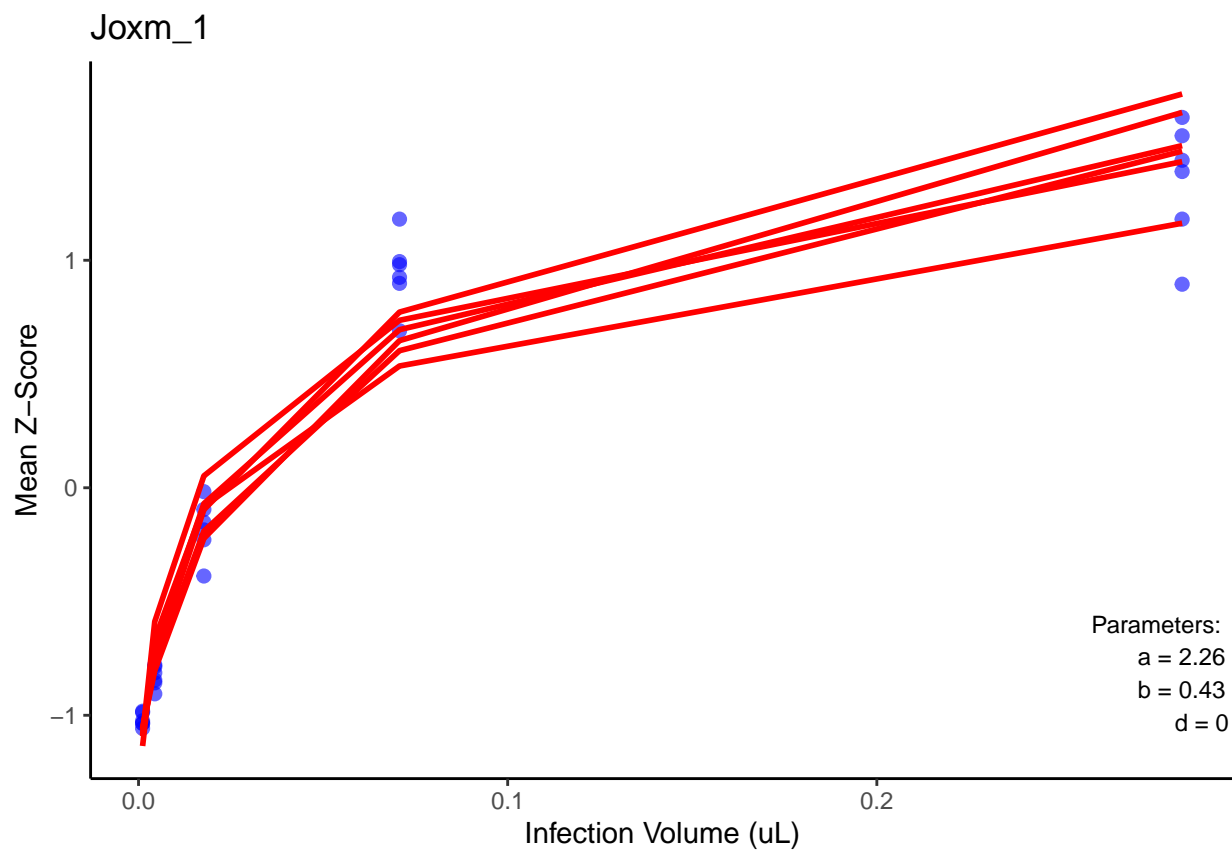
\$Jilk_3



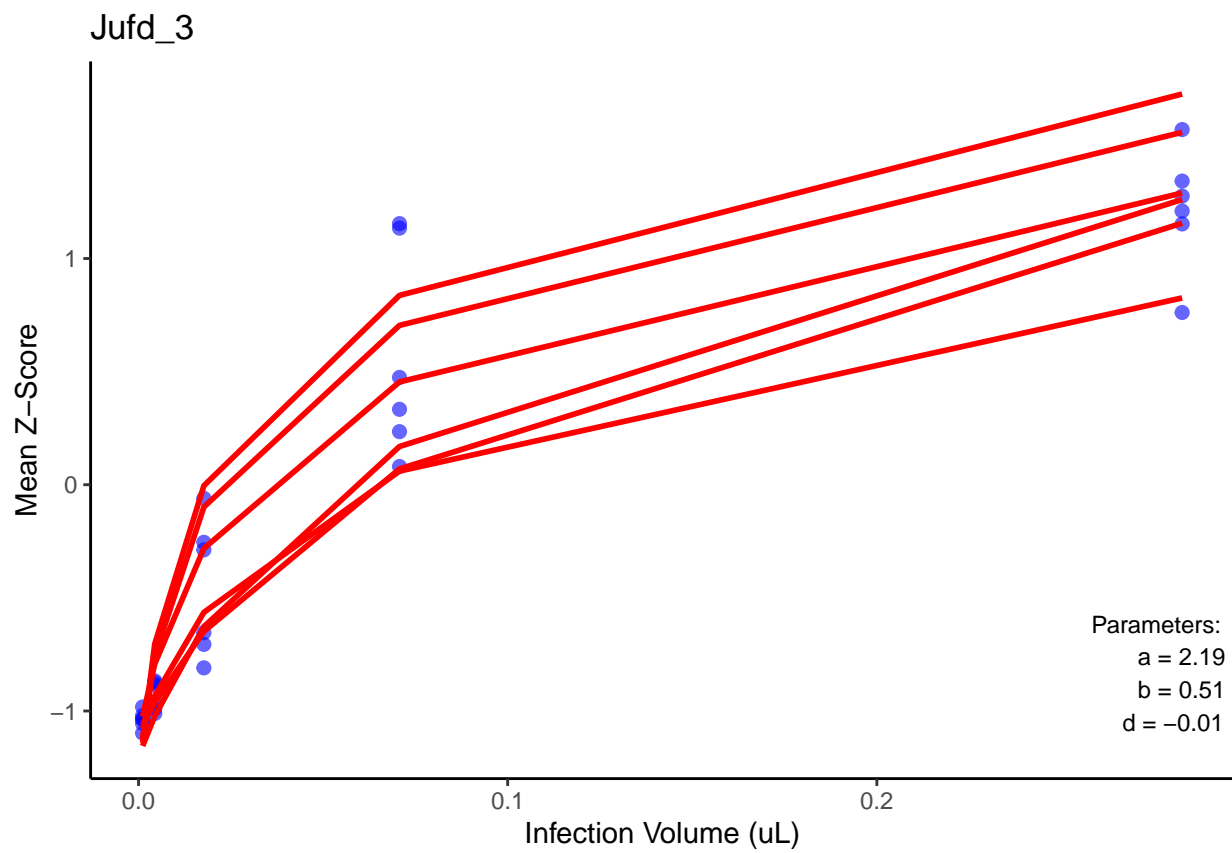
 ## \$Jogf_2



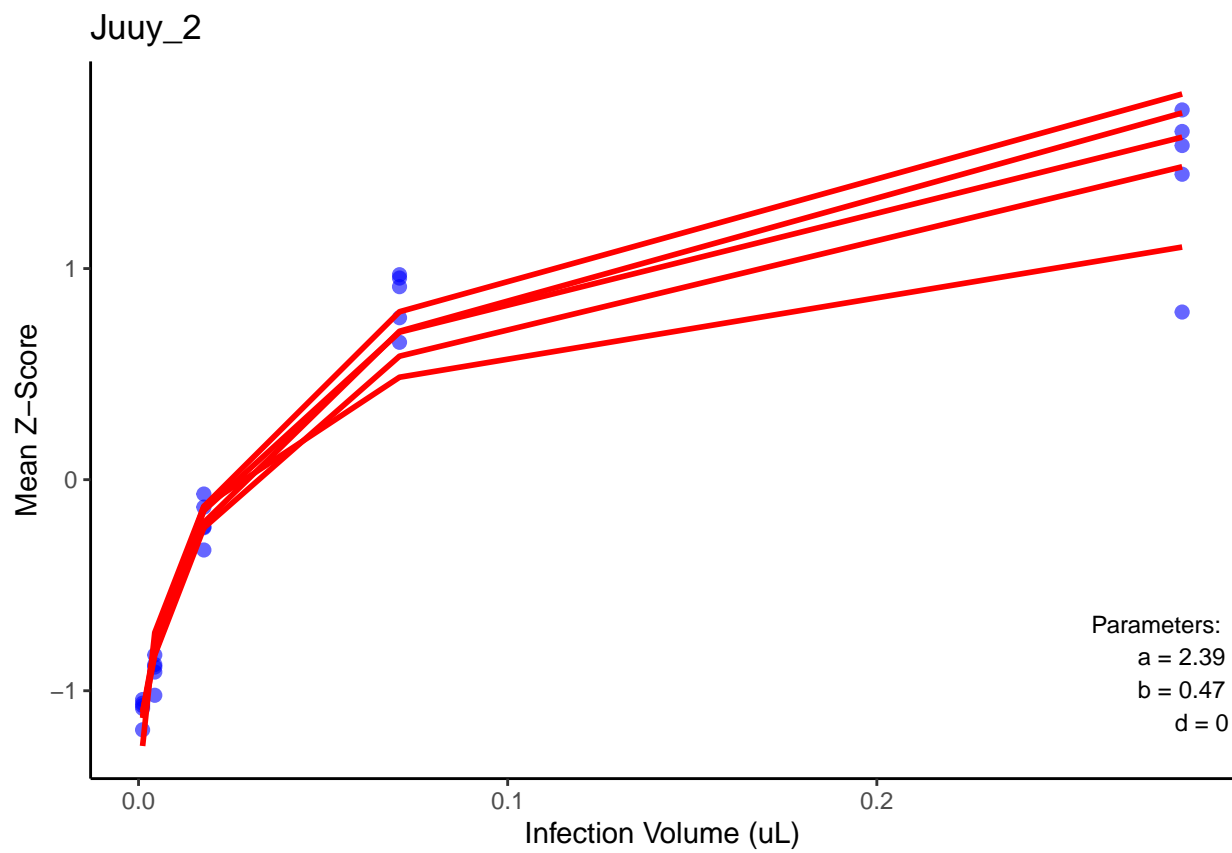
 ## \$Joxm_1



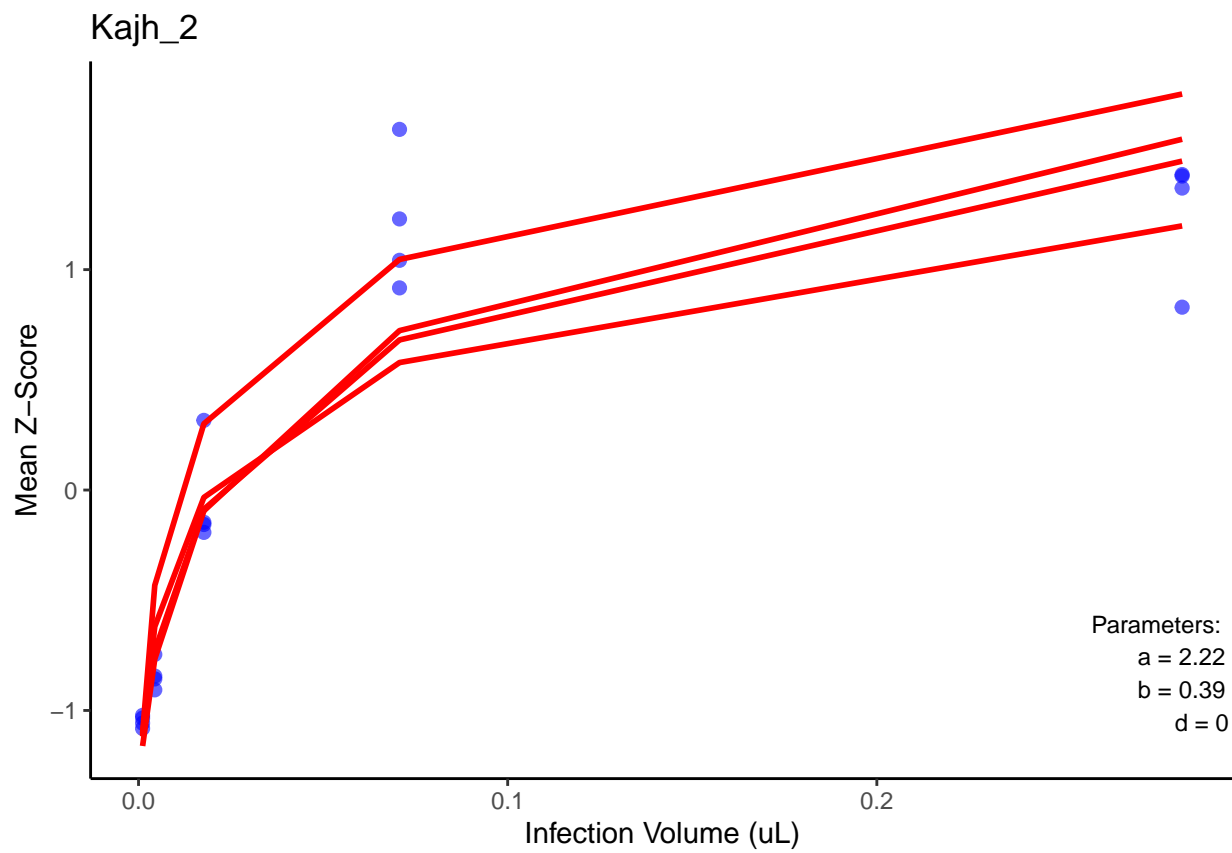
\$Jufd_3



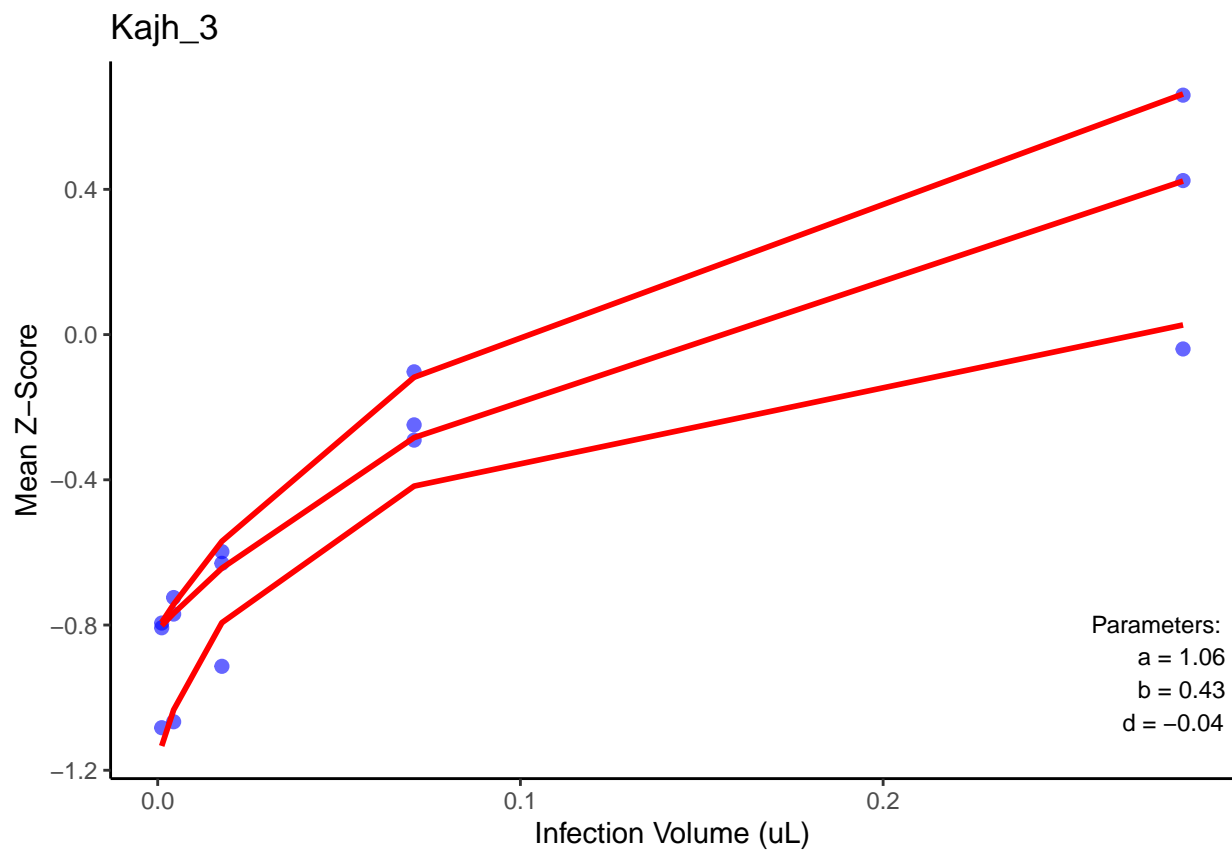
\$Juuy_2



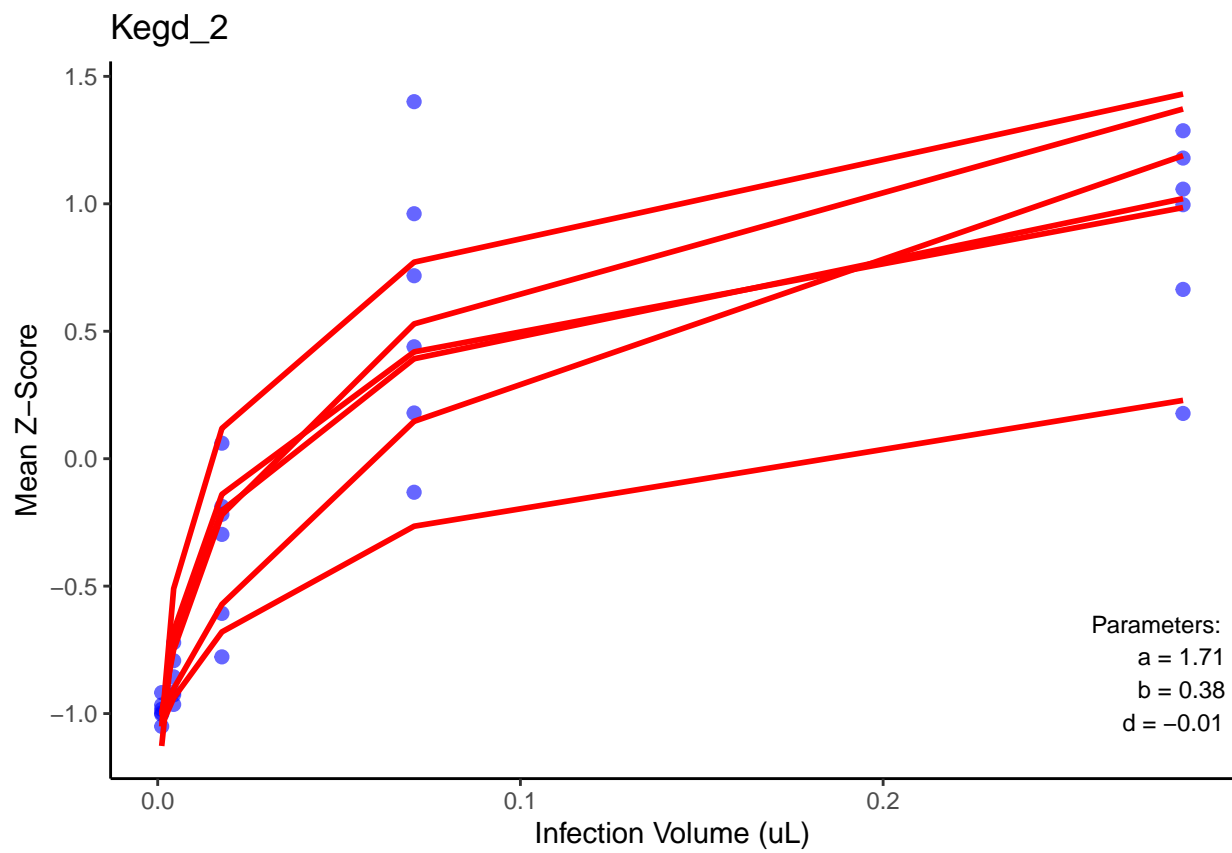
\$Kajh_2

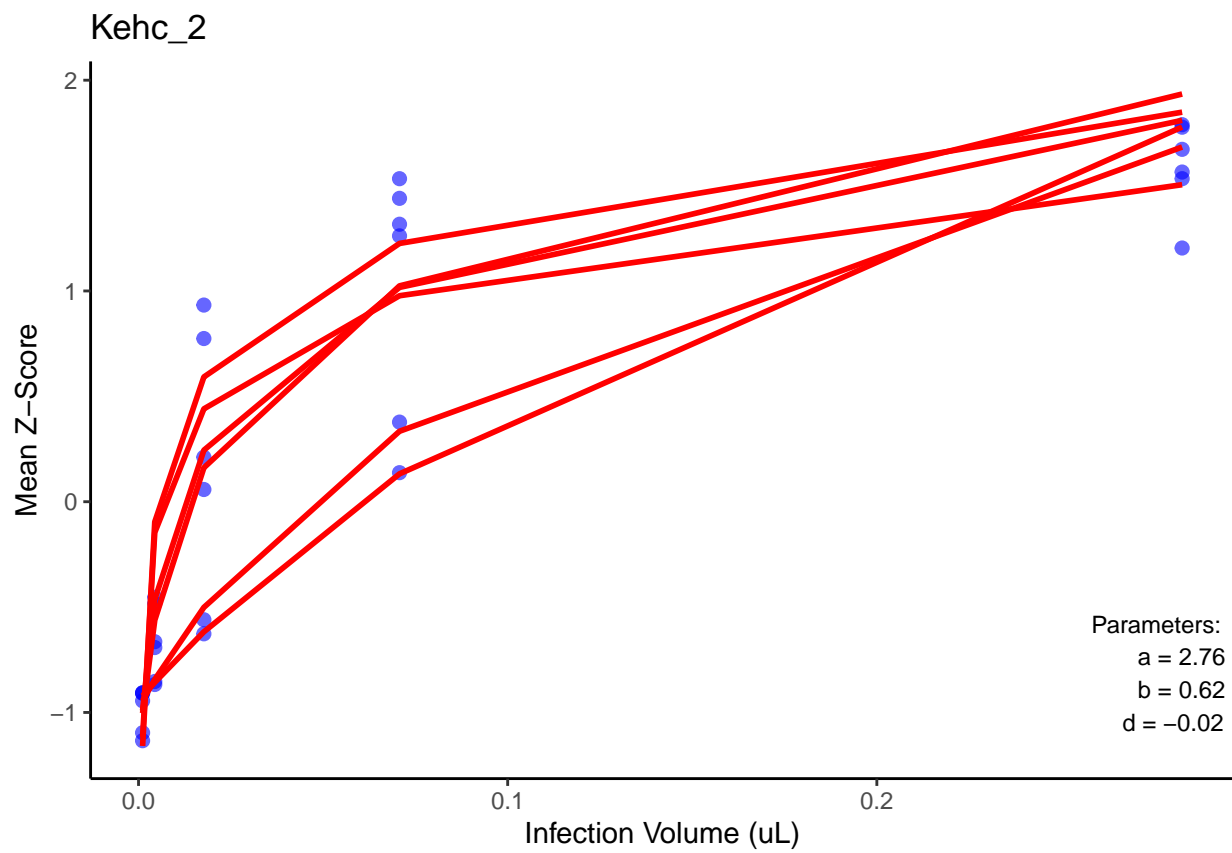


 ## \$Kajh_3

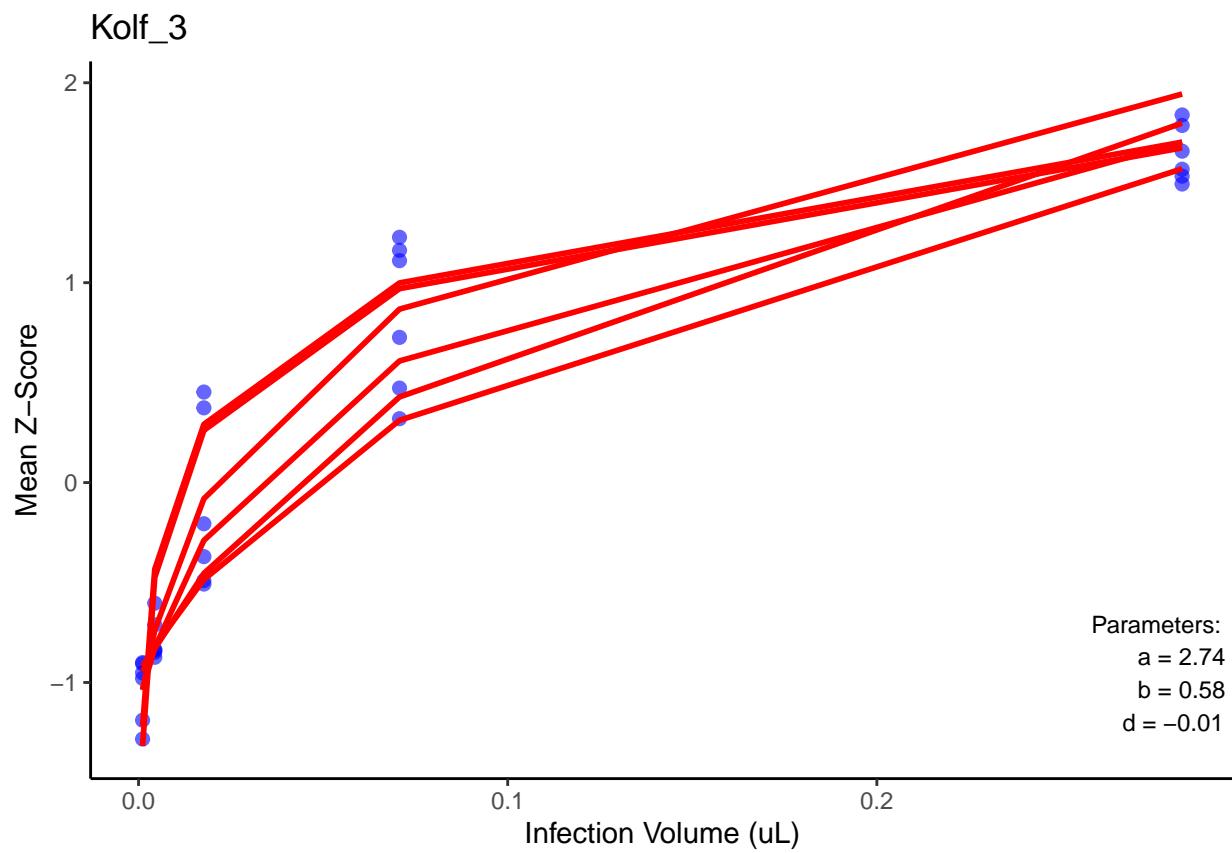


 ## \$Kegd_2

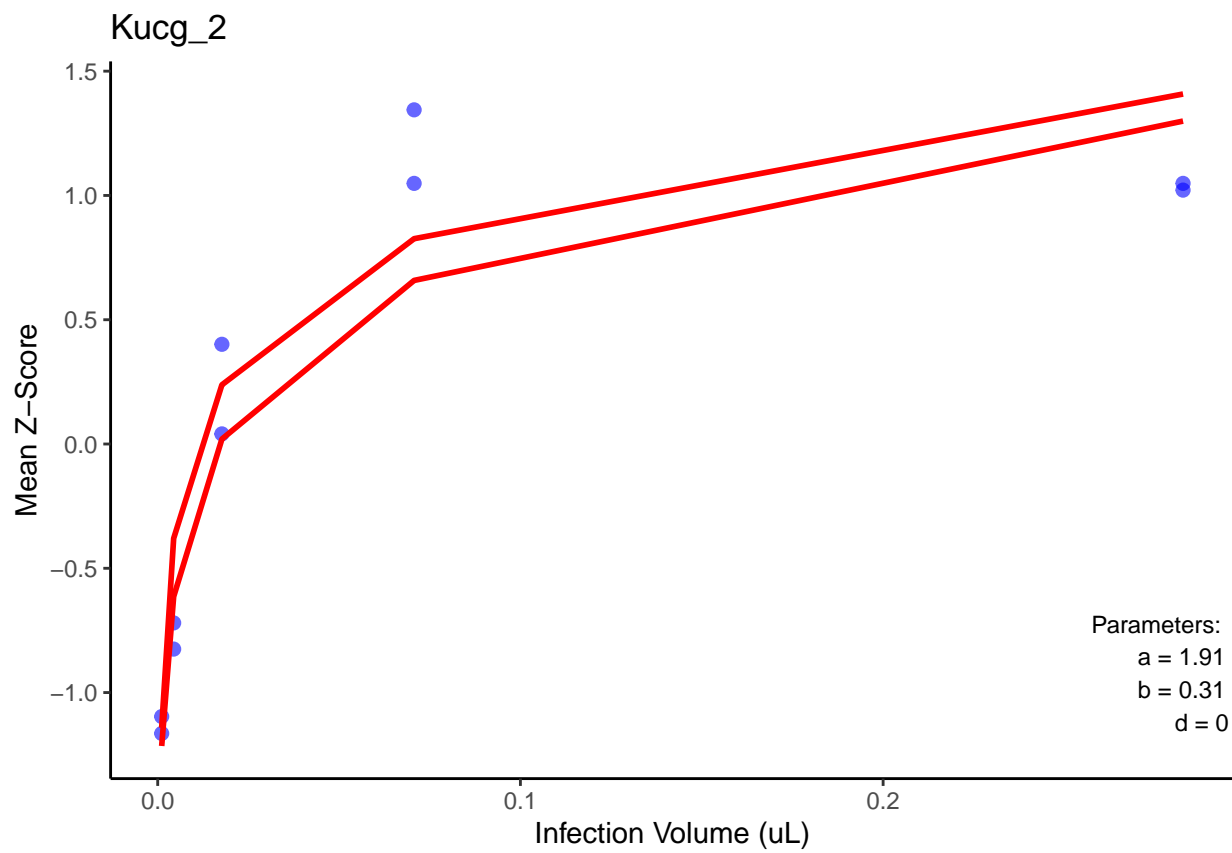




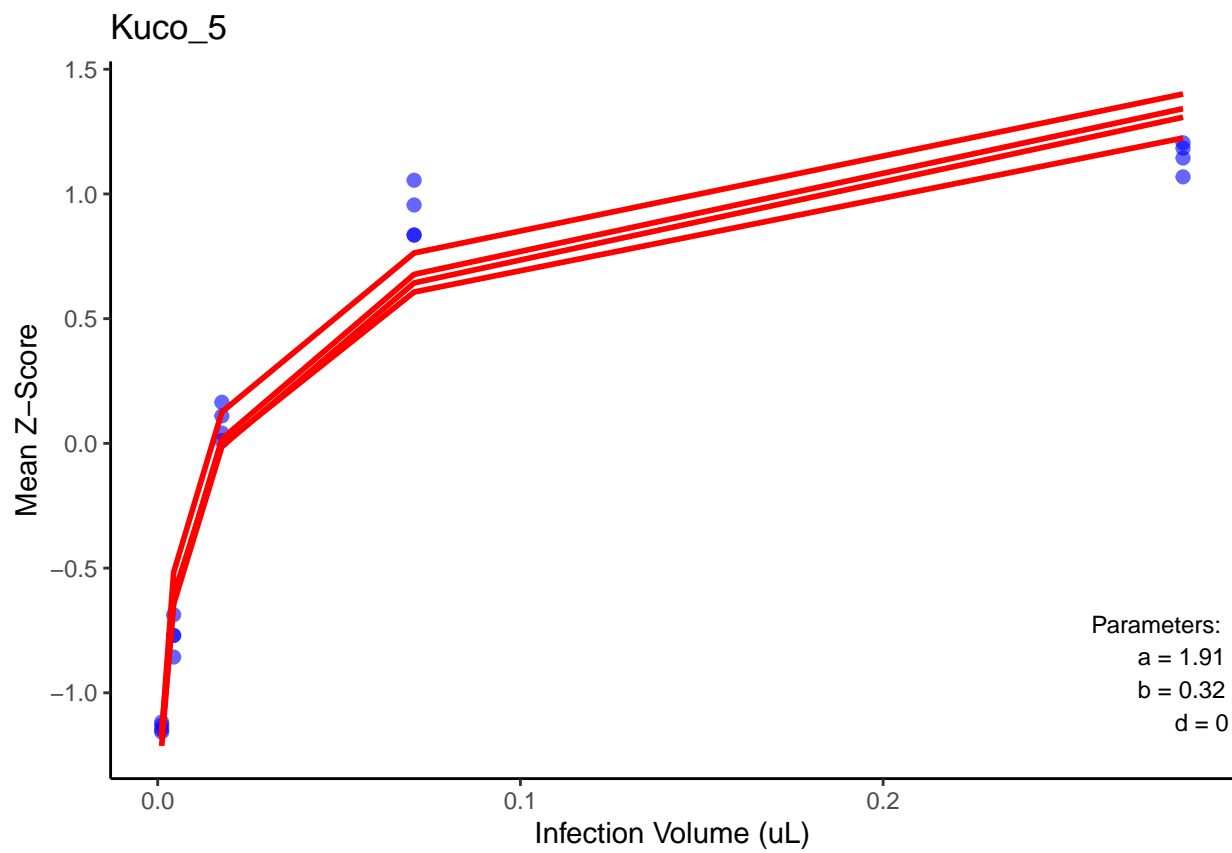
\$Kolf_3



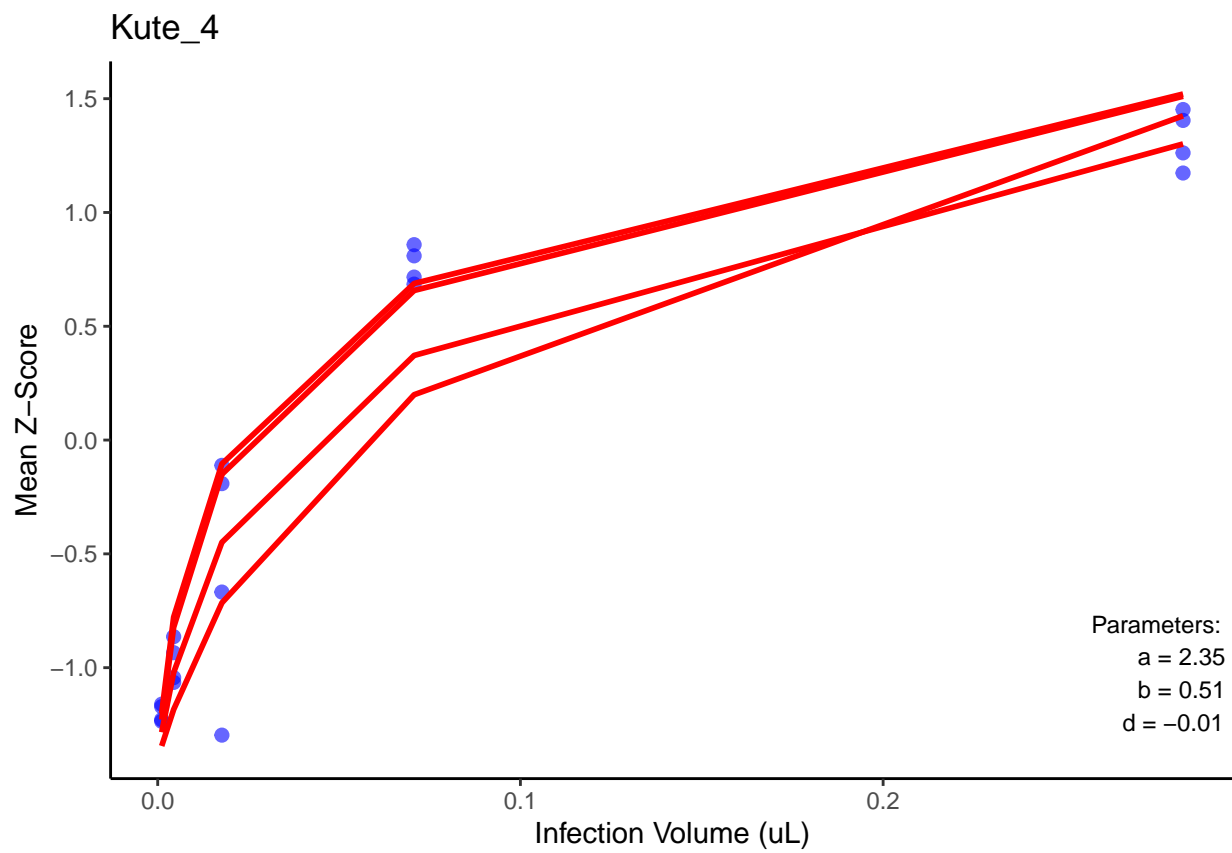
\$Kucg_2



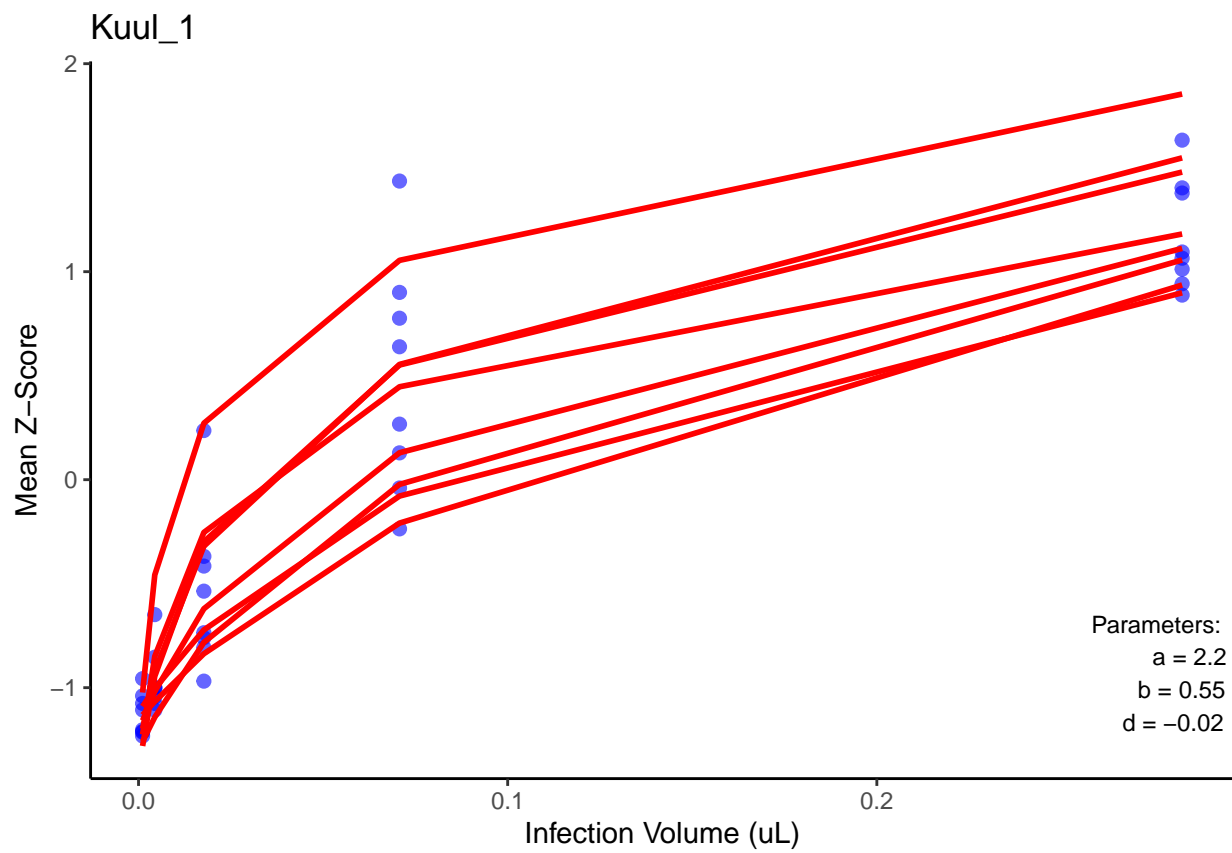
\$Kuco_5



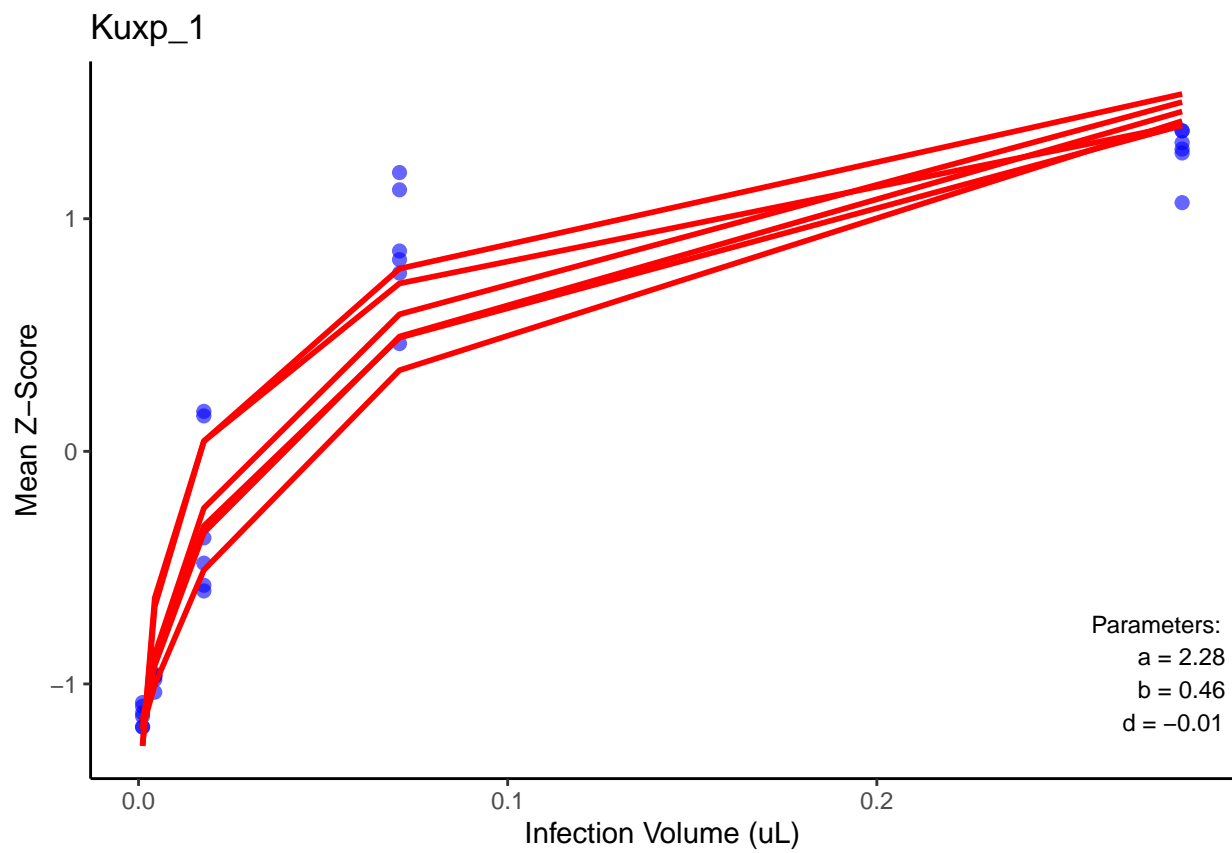
\$Kute_4



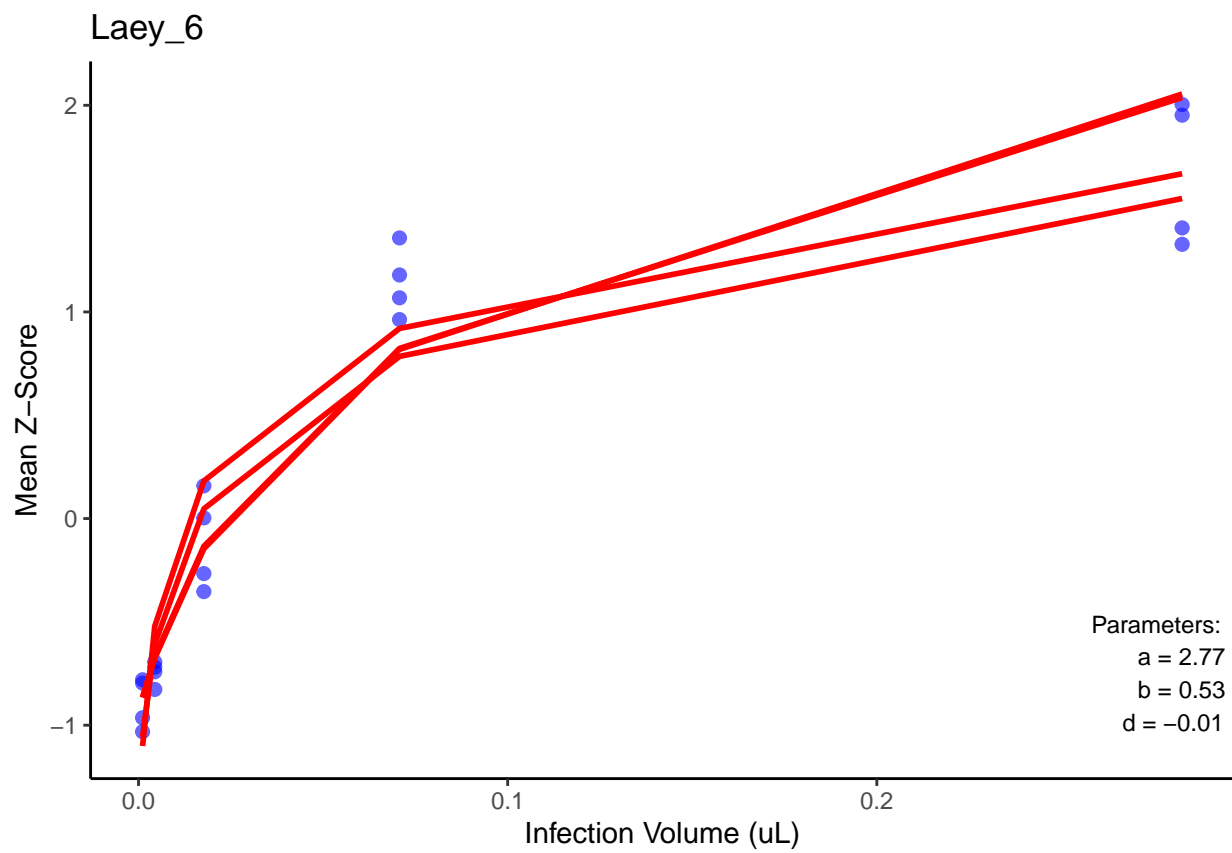
\$Kuul_1



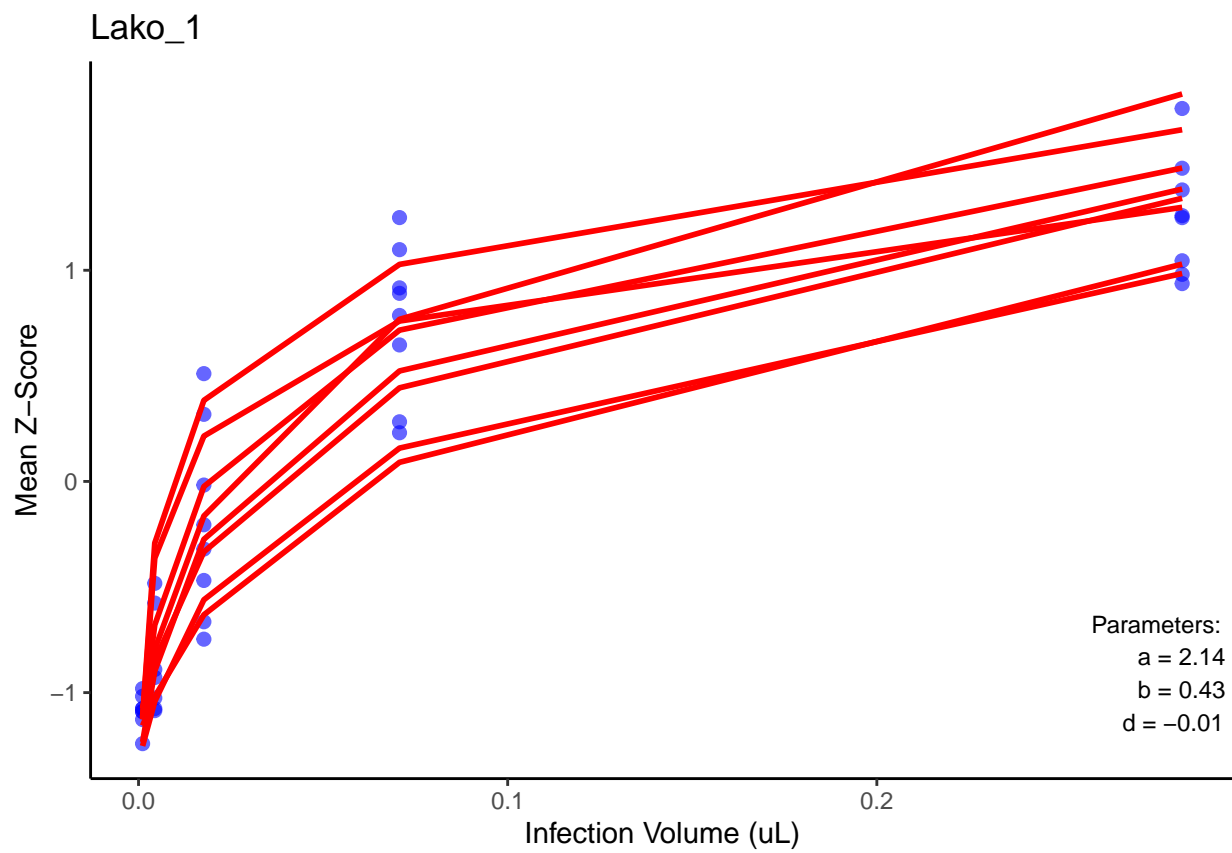
 ## \$Kuxp_1



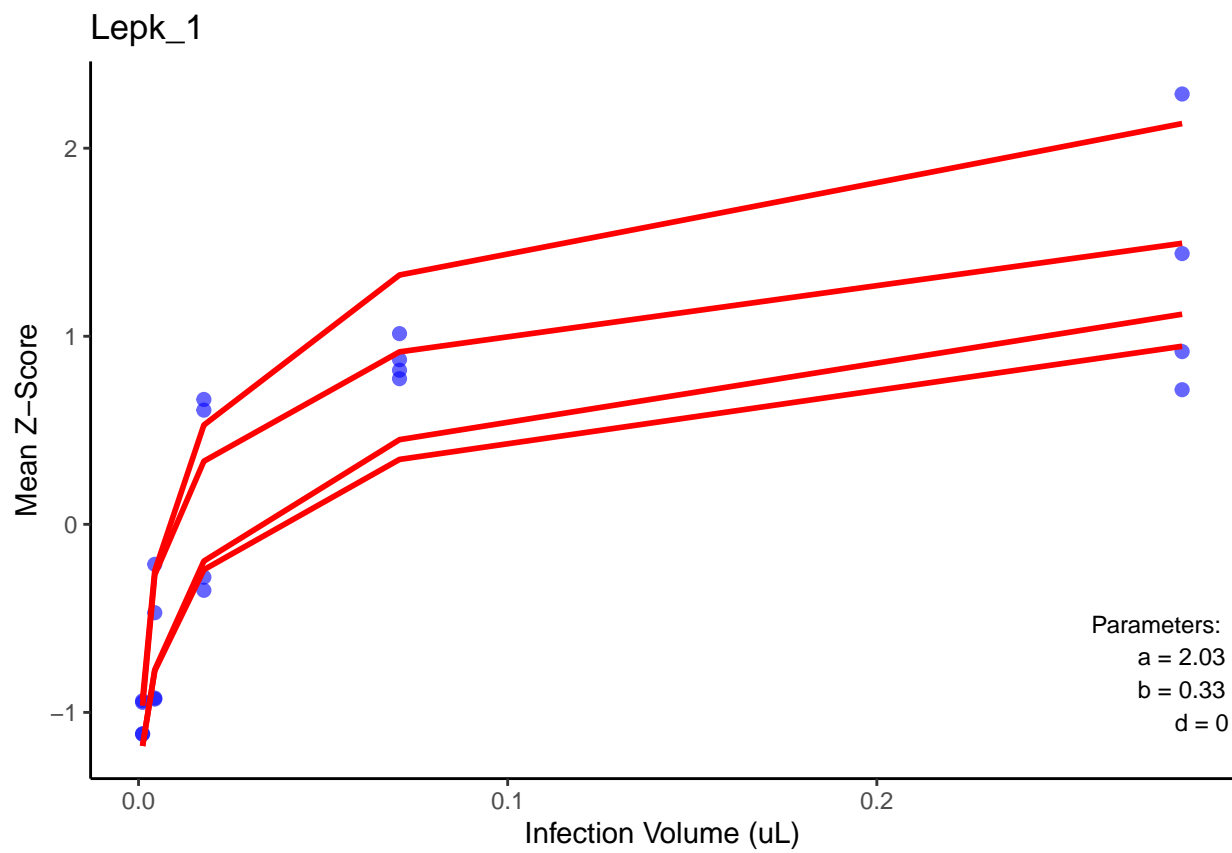
\$Laey_6



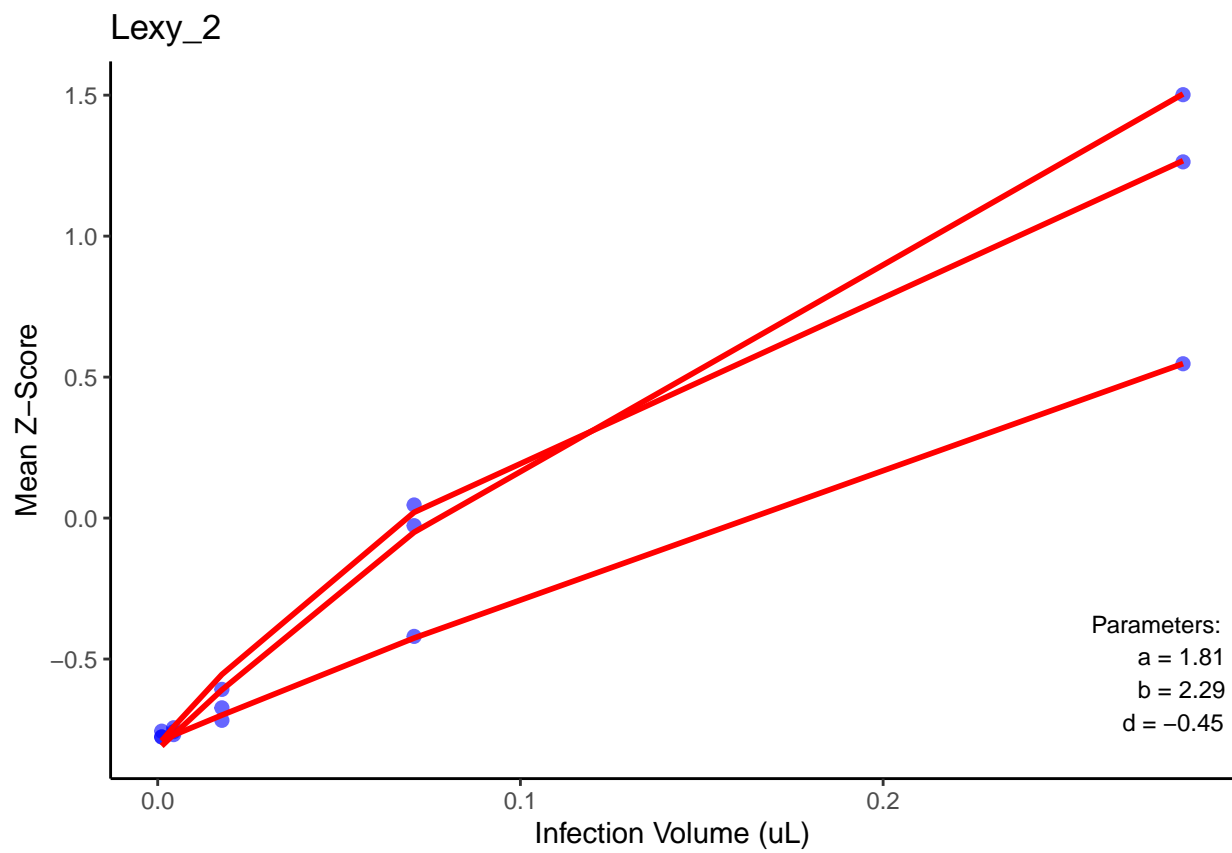
 ## \$Lako_1



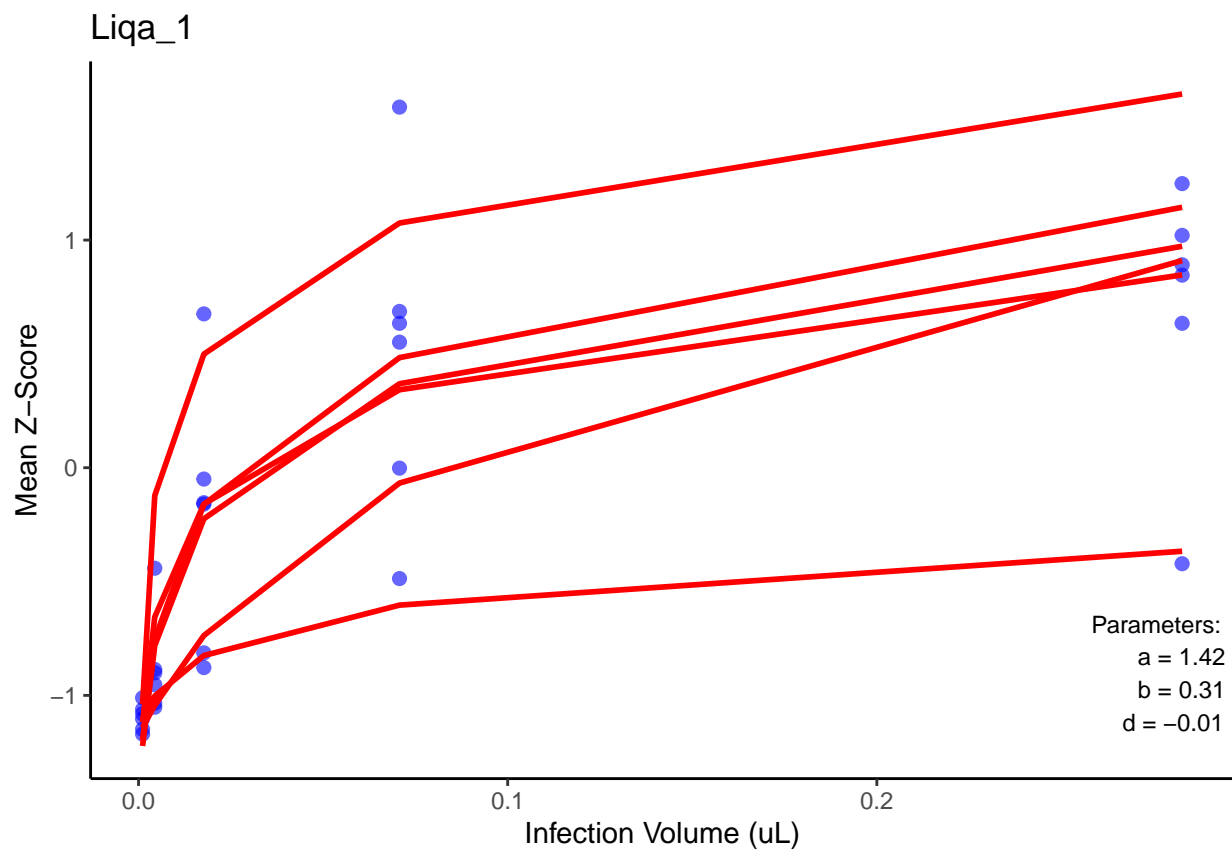
\$Lepk_1



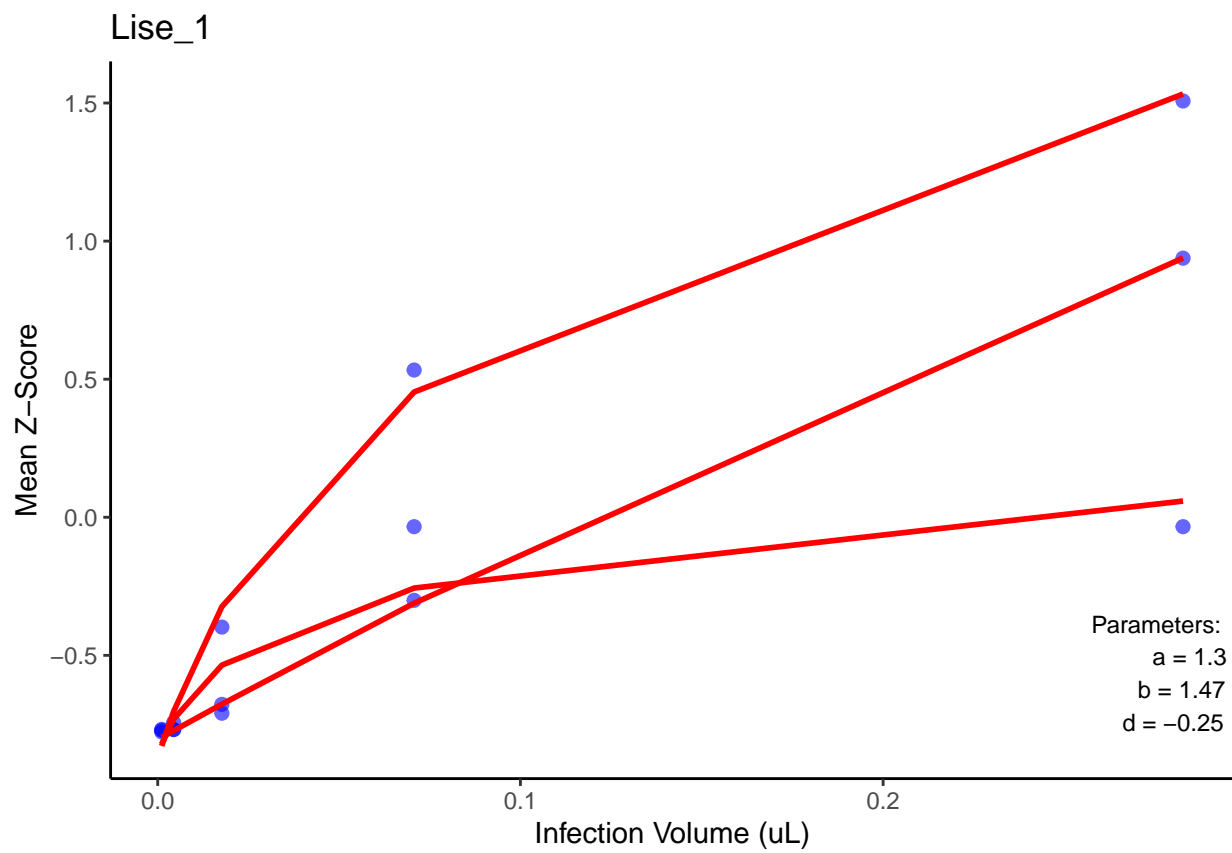
\$Lexy_2



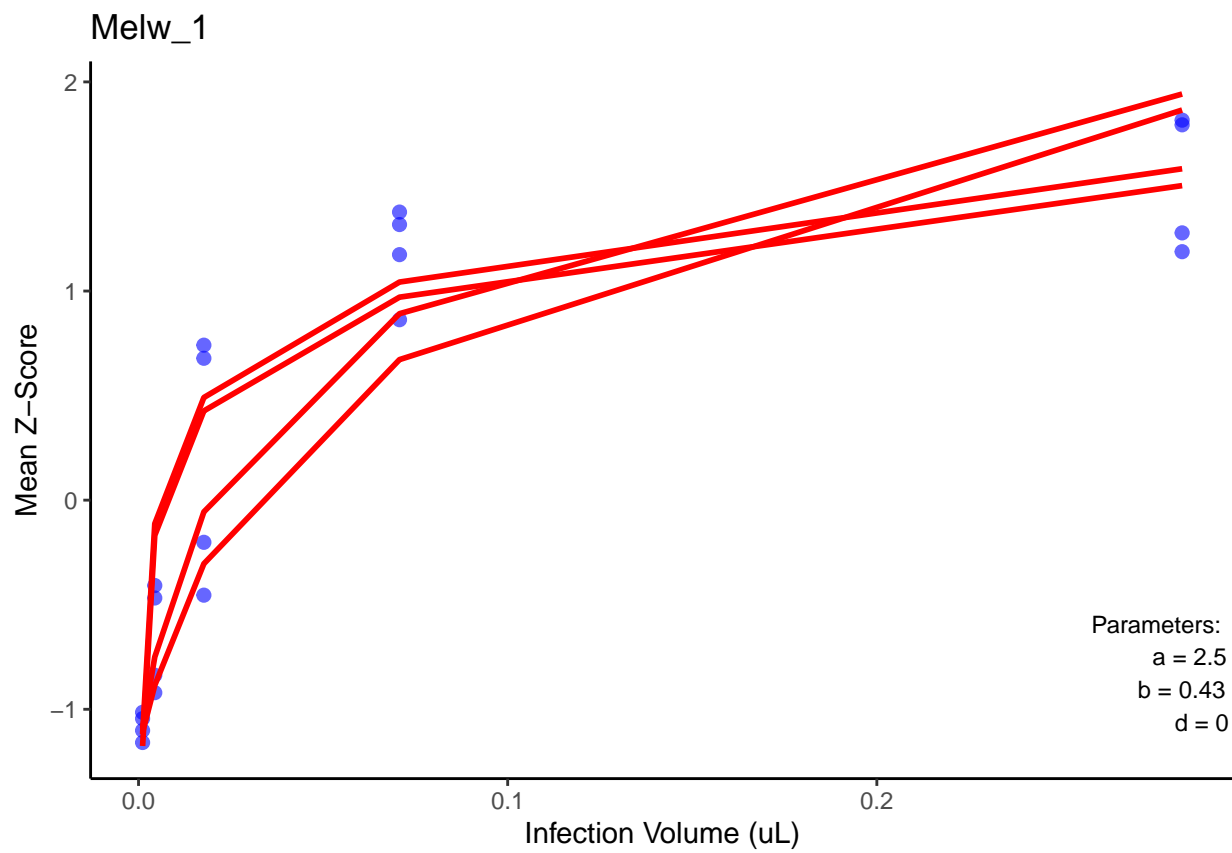
\$Liqa_1



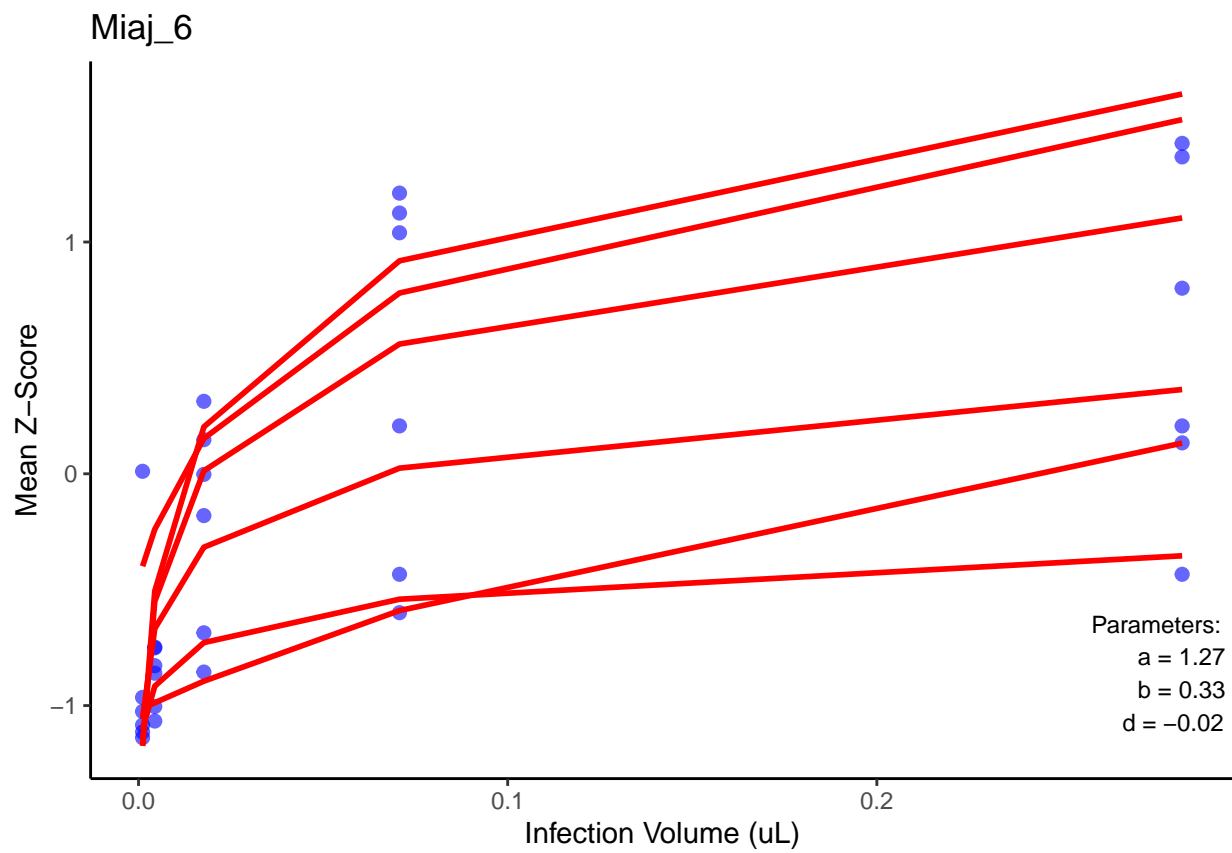
\$Lise_1



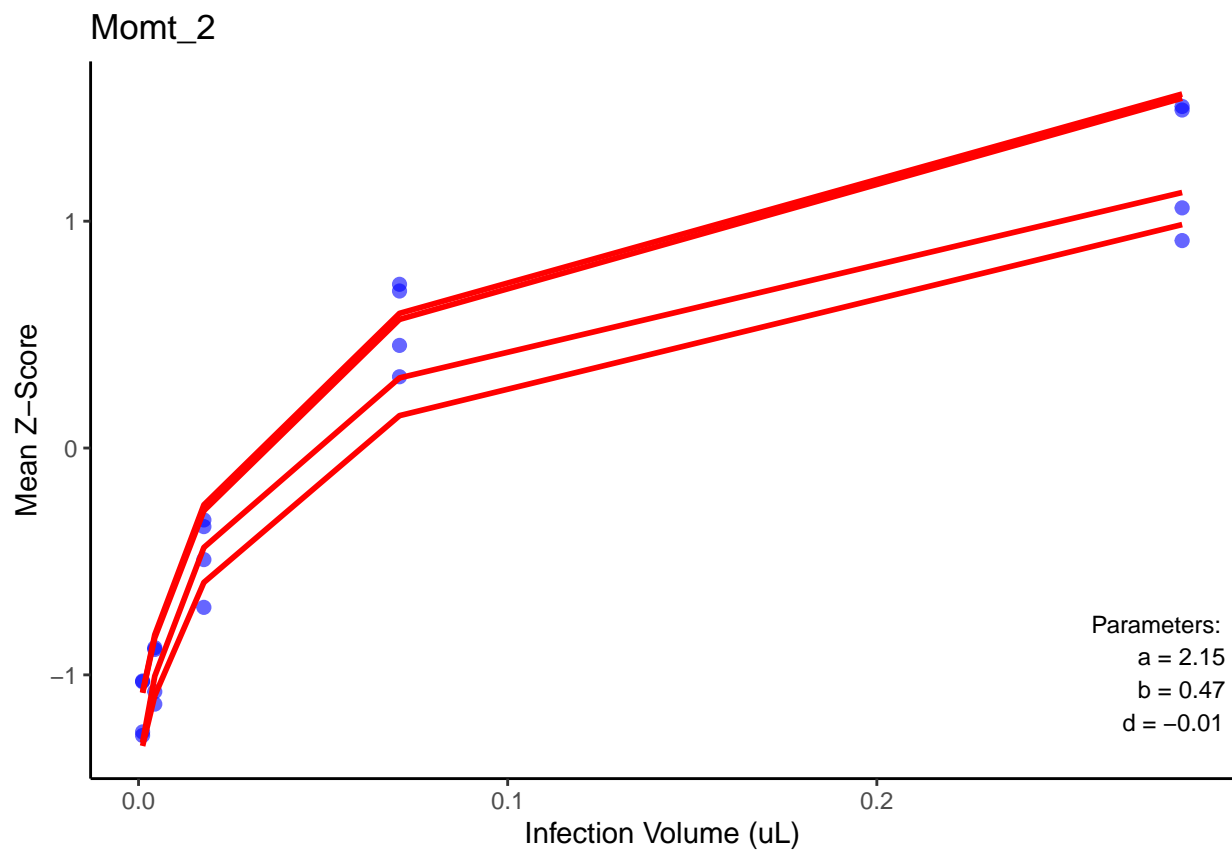
 ## \$Melw_1



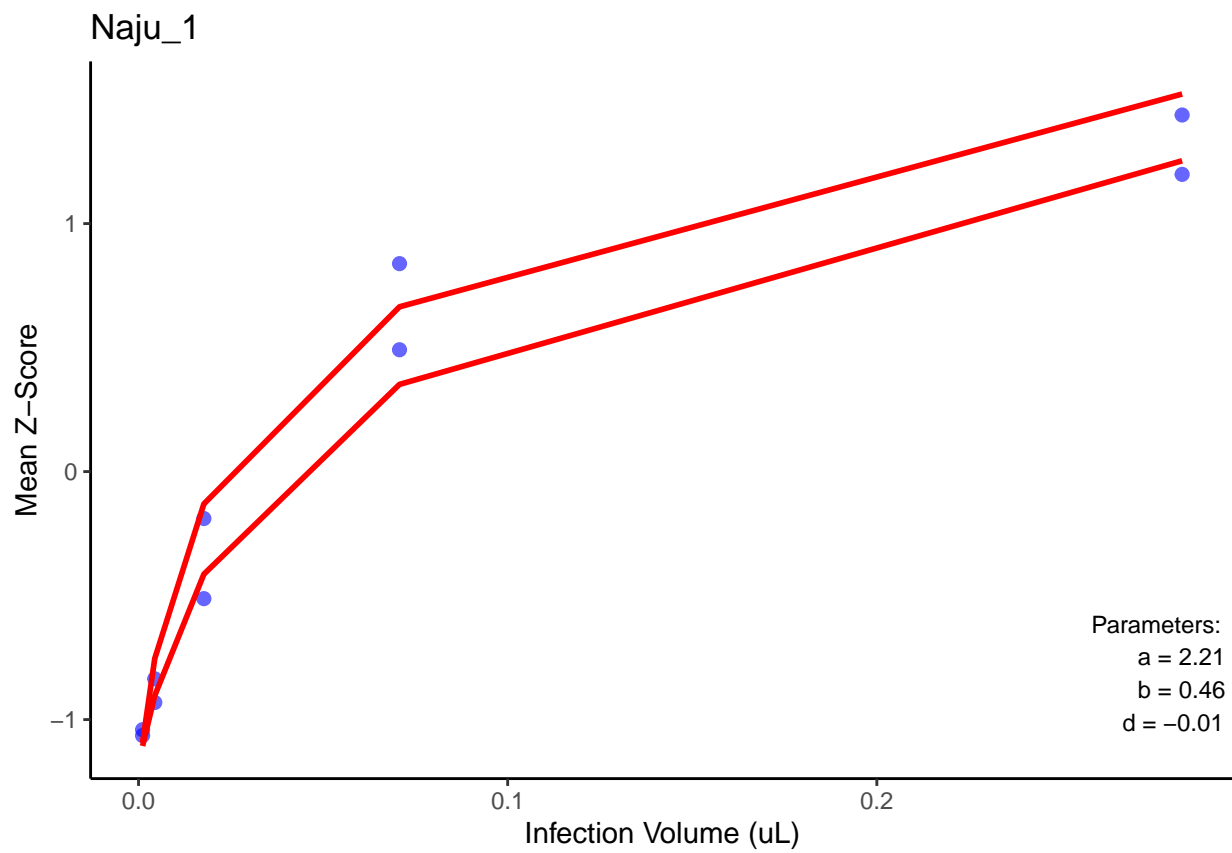
 ## \$Miaj_6



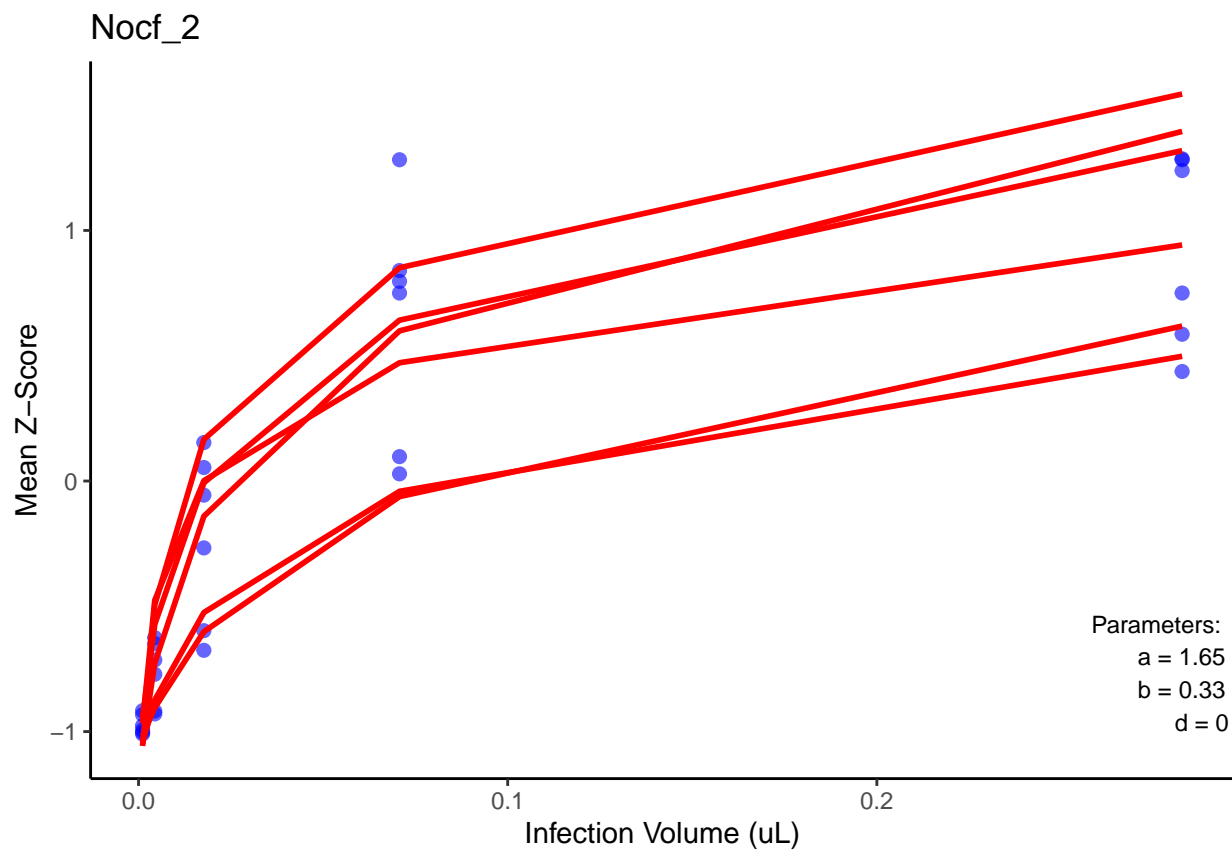
 ## \$Momt_2



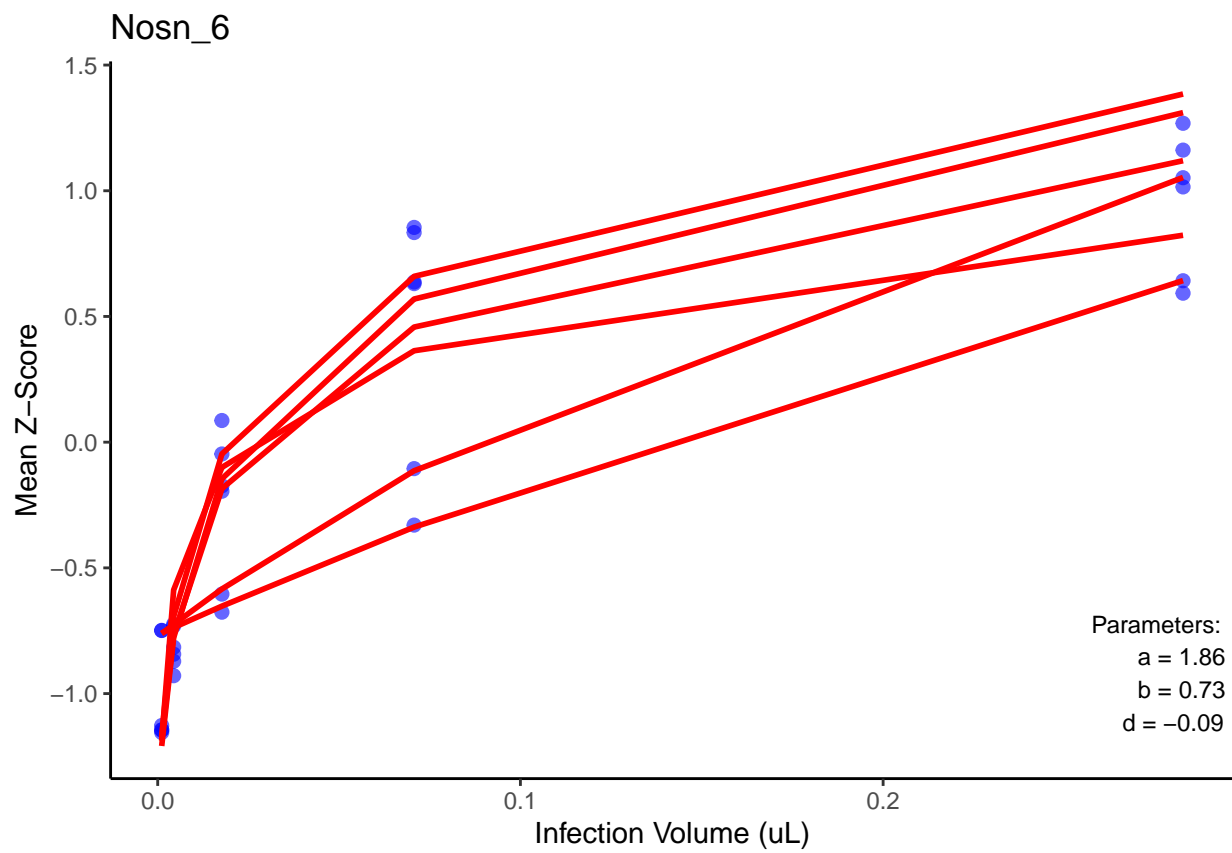
\$Naju_1



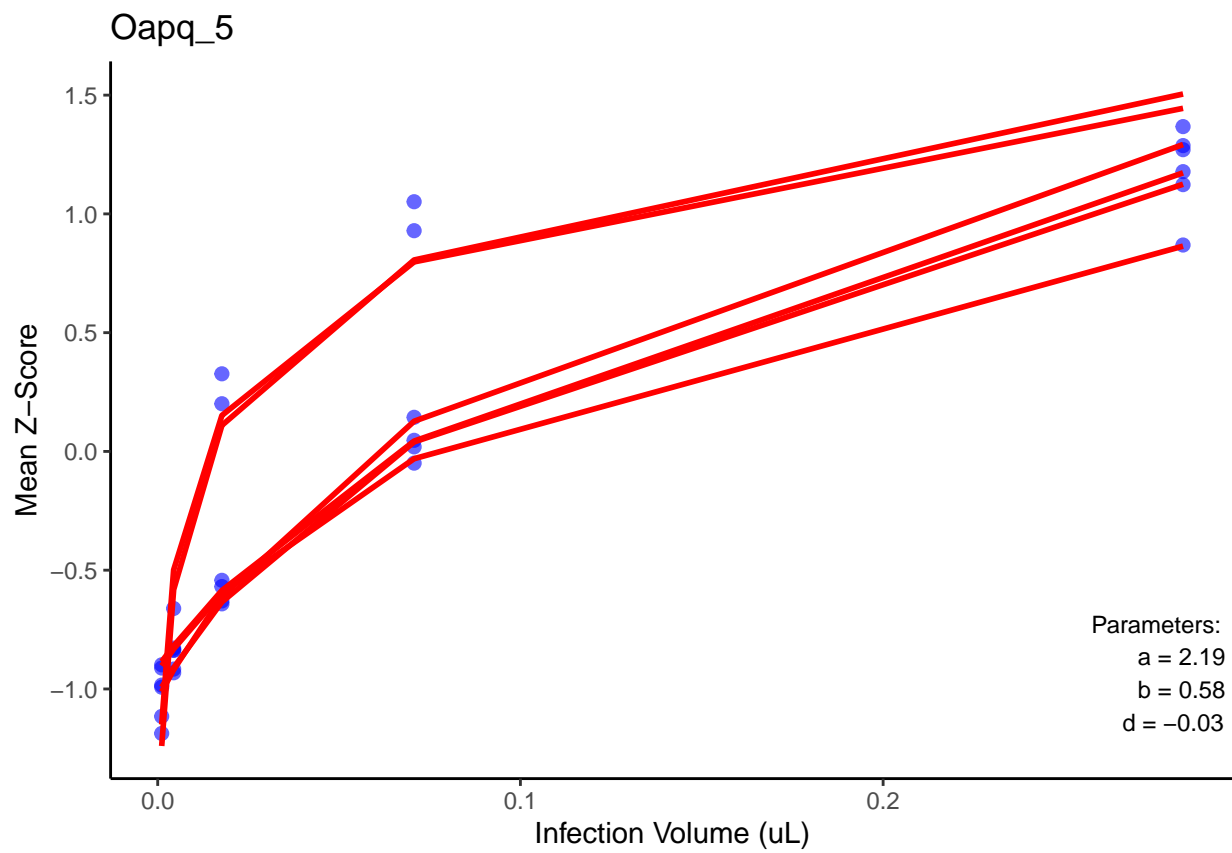
\$Nocf_2



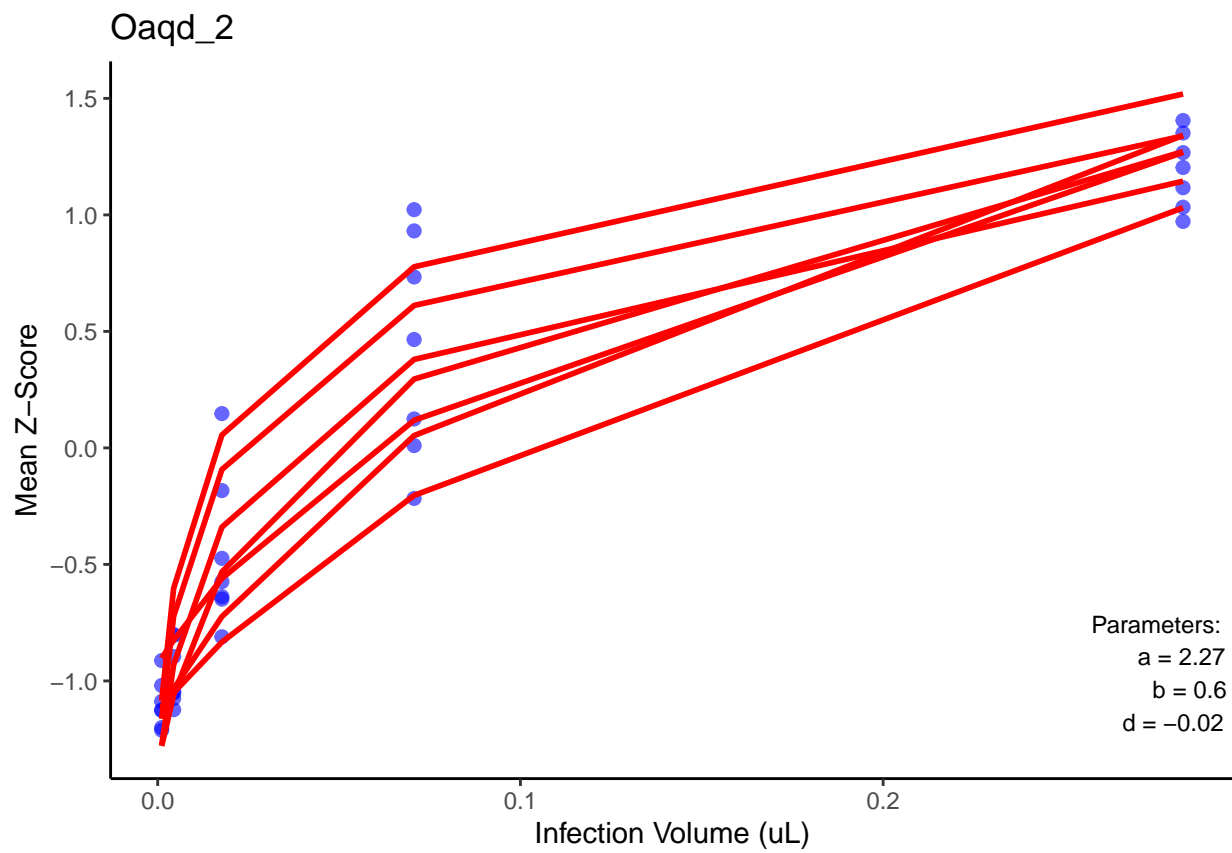
 ## \$Nosn_6



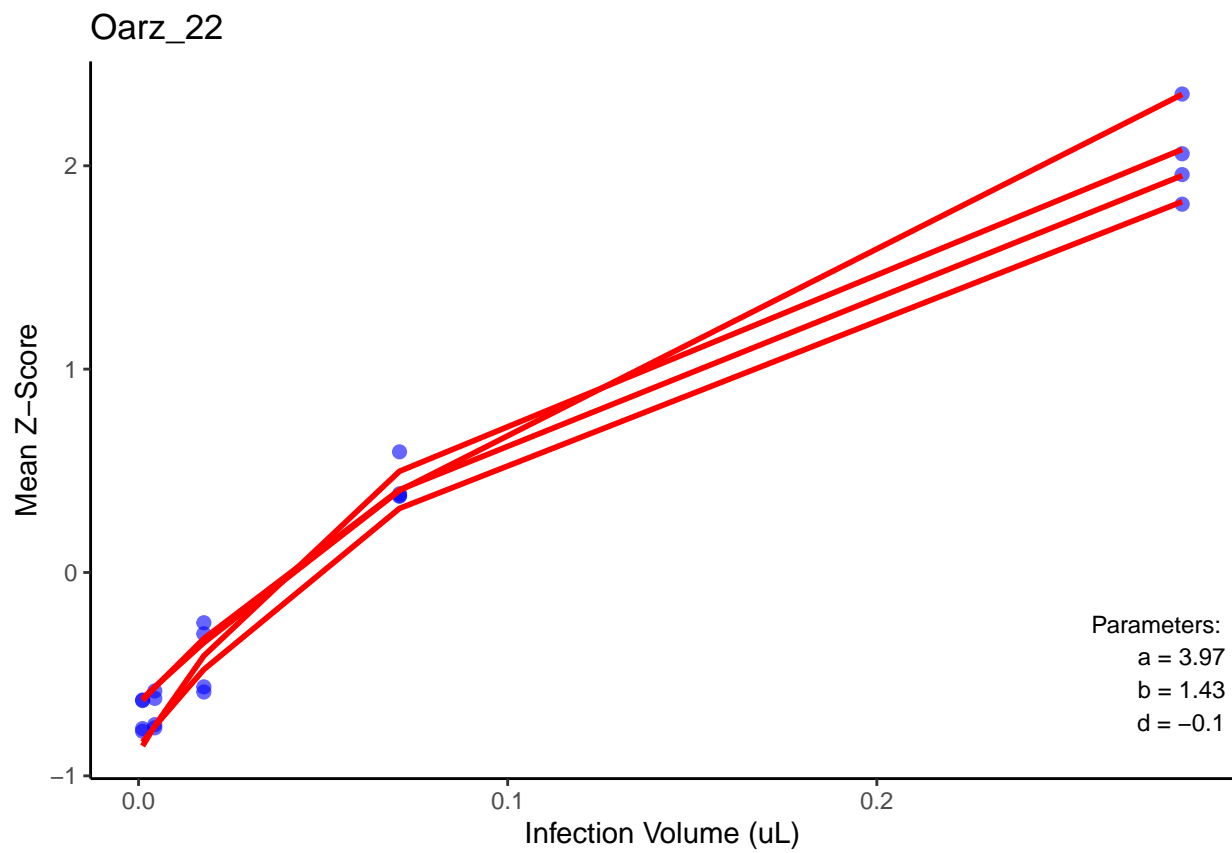
 ## σ_{apq_5}



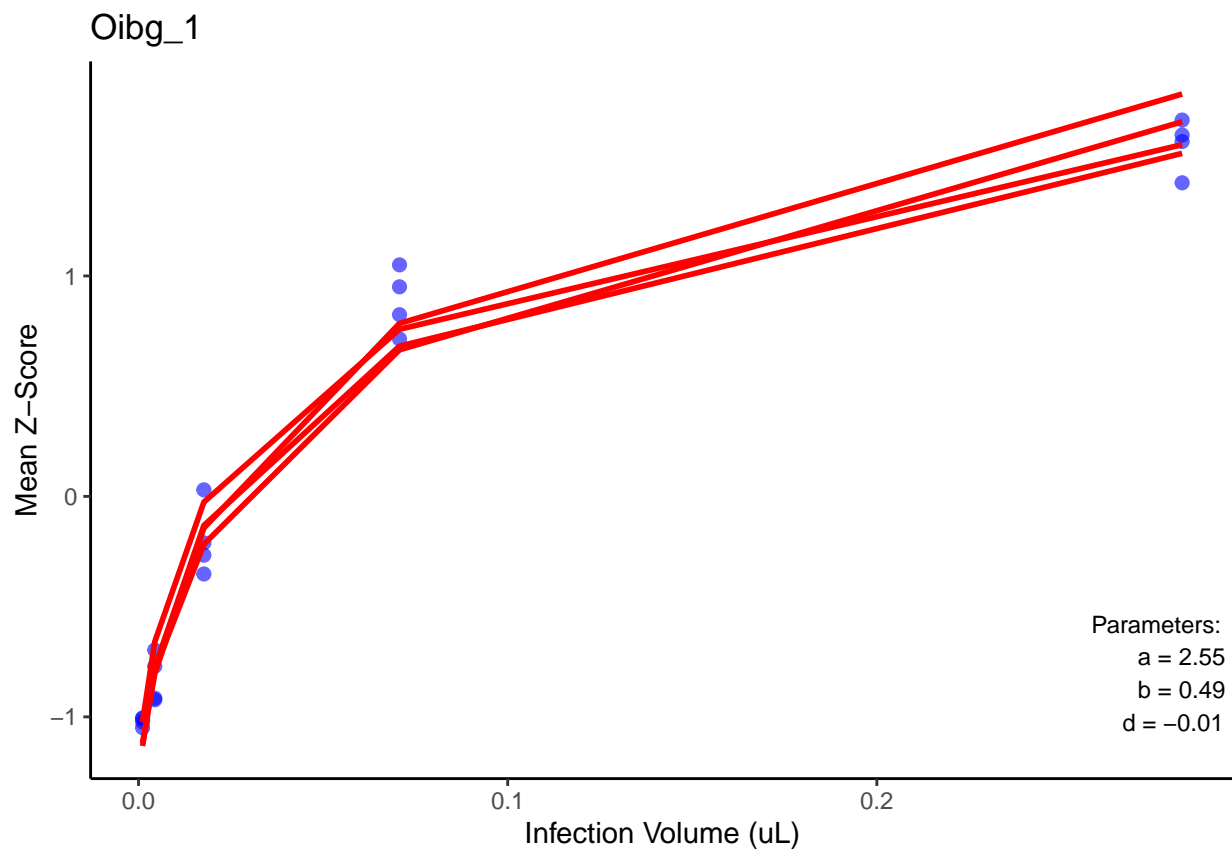
 ## \$0aqd_2



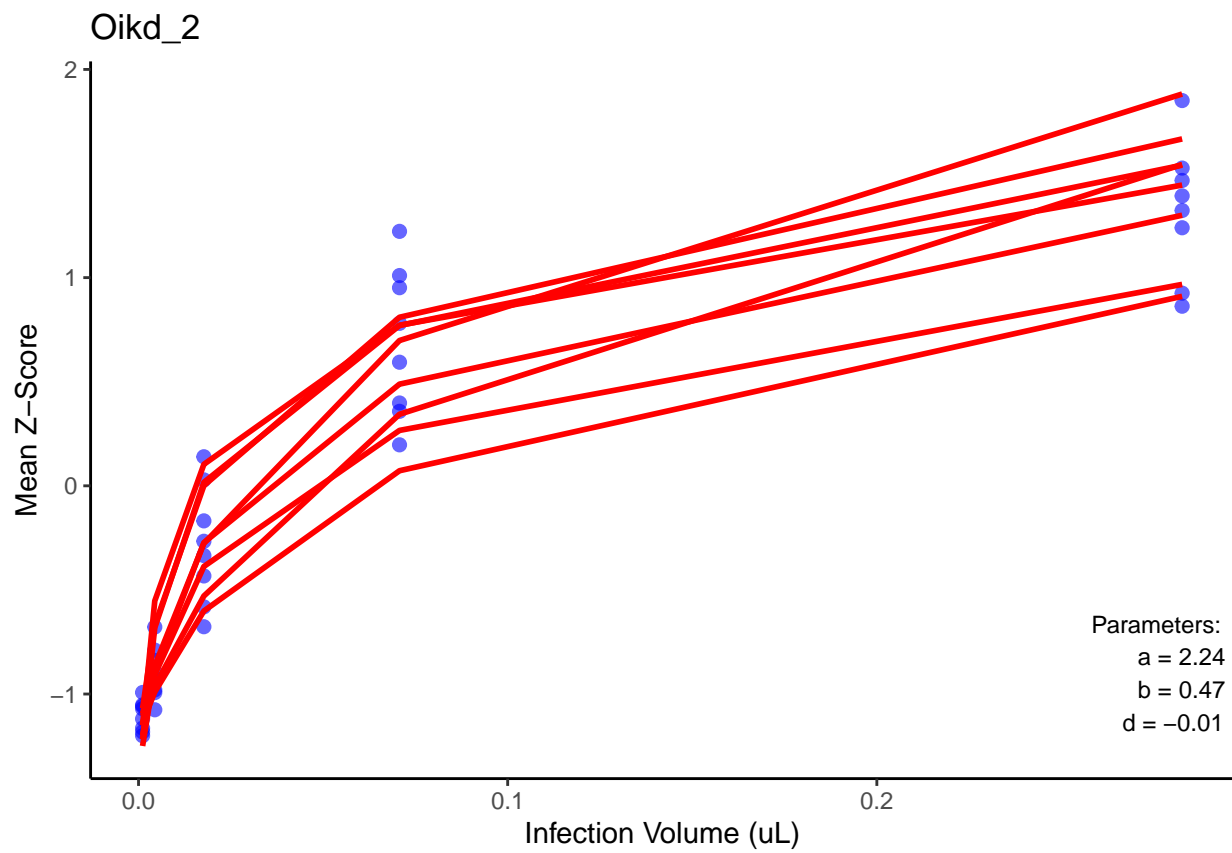
\$0arz_22



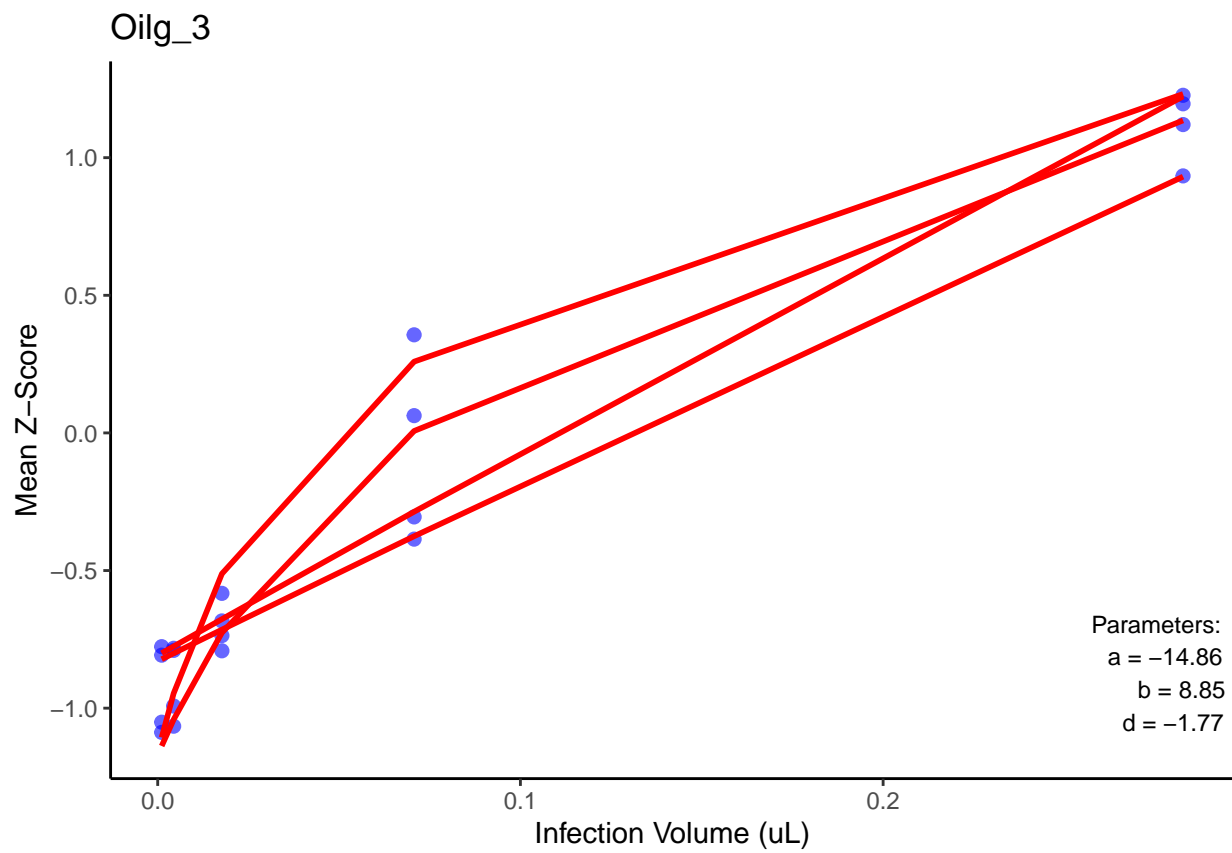
\$0ibg_1



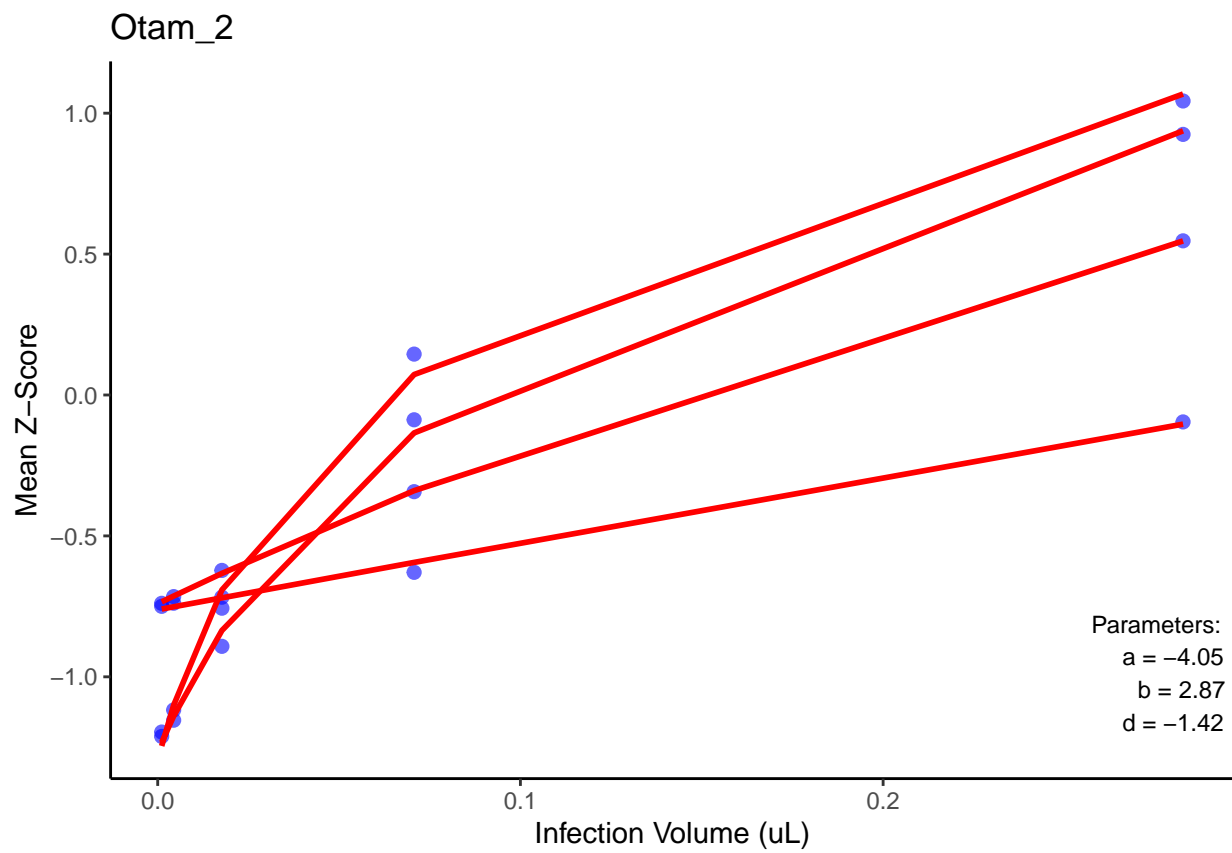
\$Oikd_2



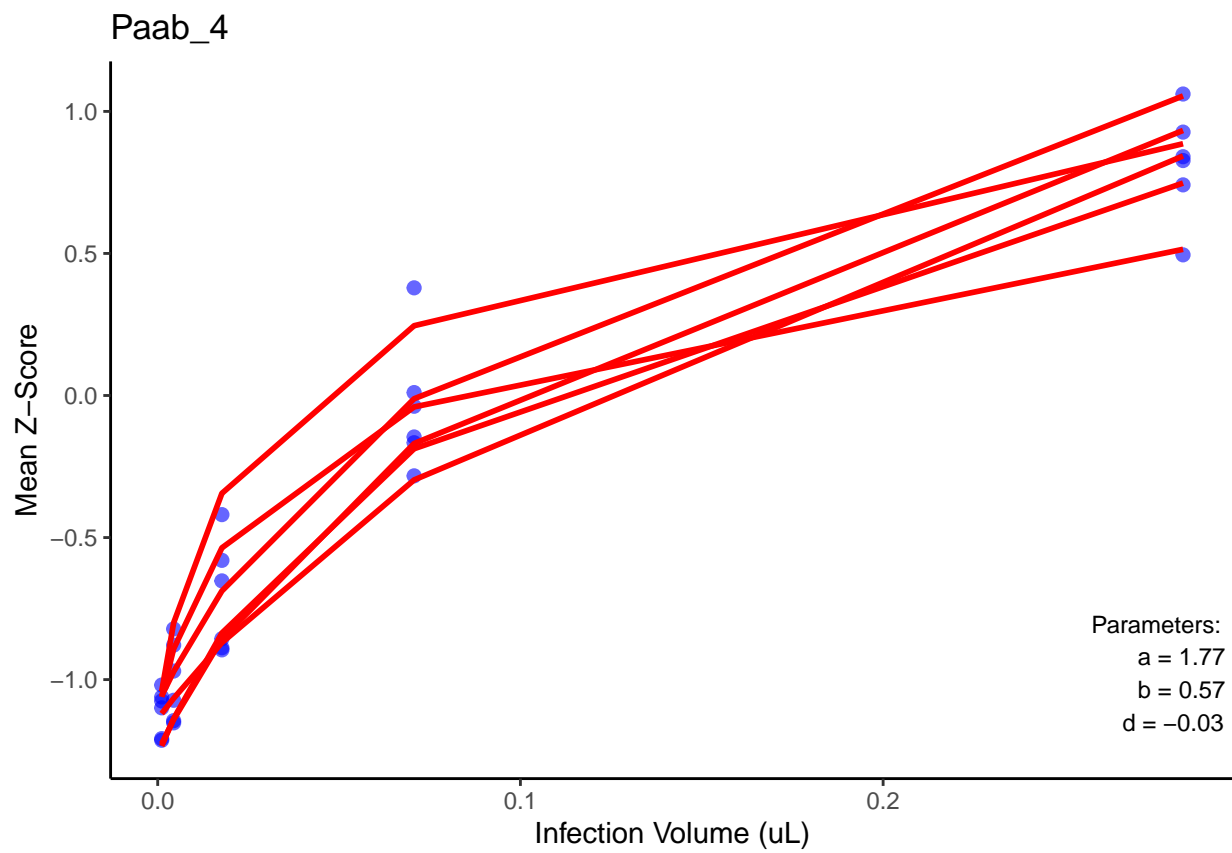
\$Oilg_3



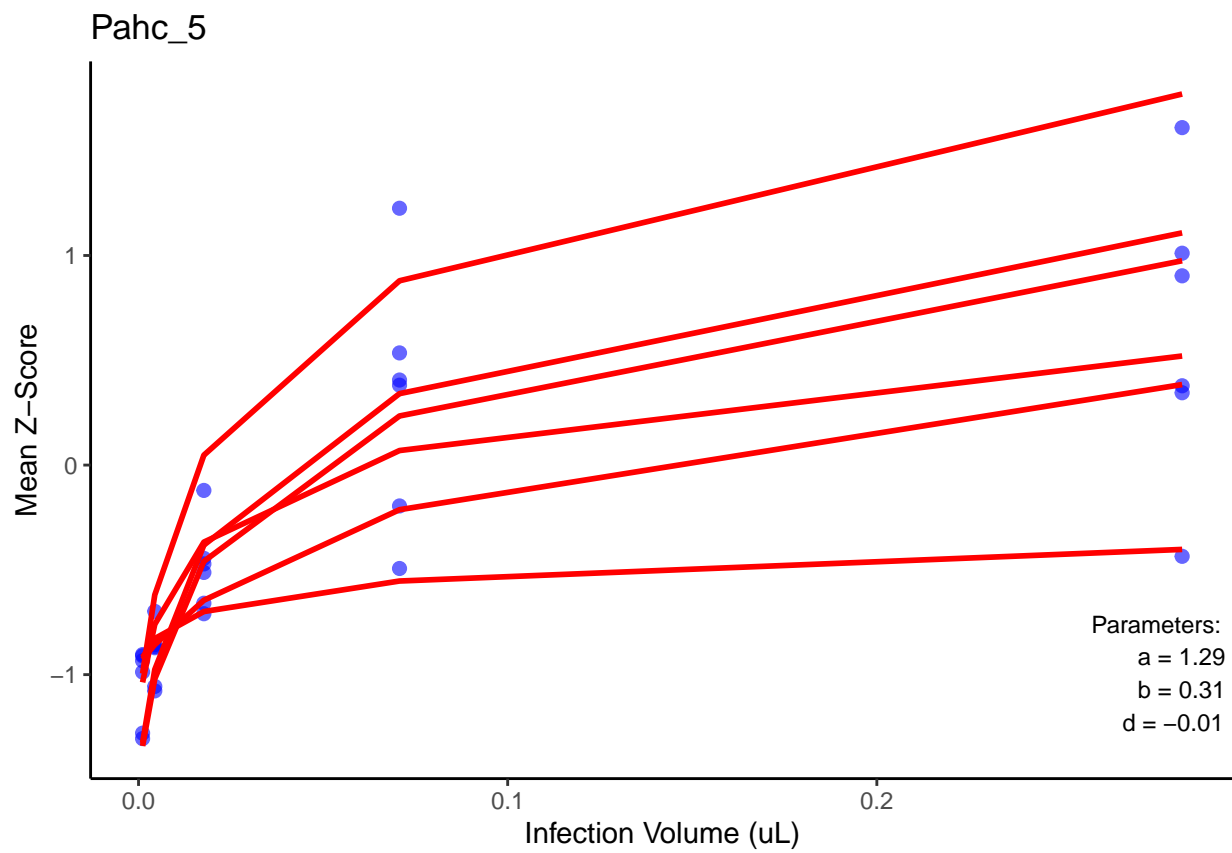
\$0tam_2



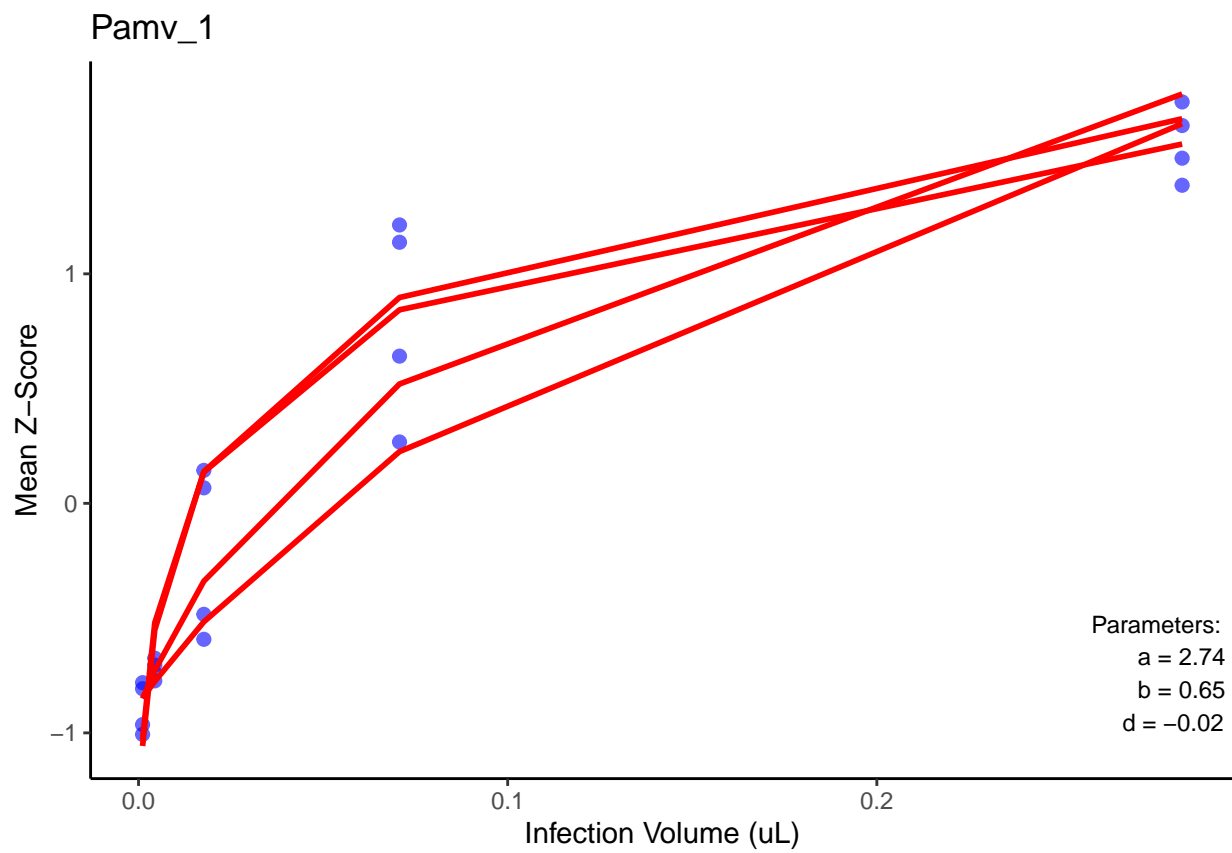
\$Paab_4



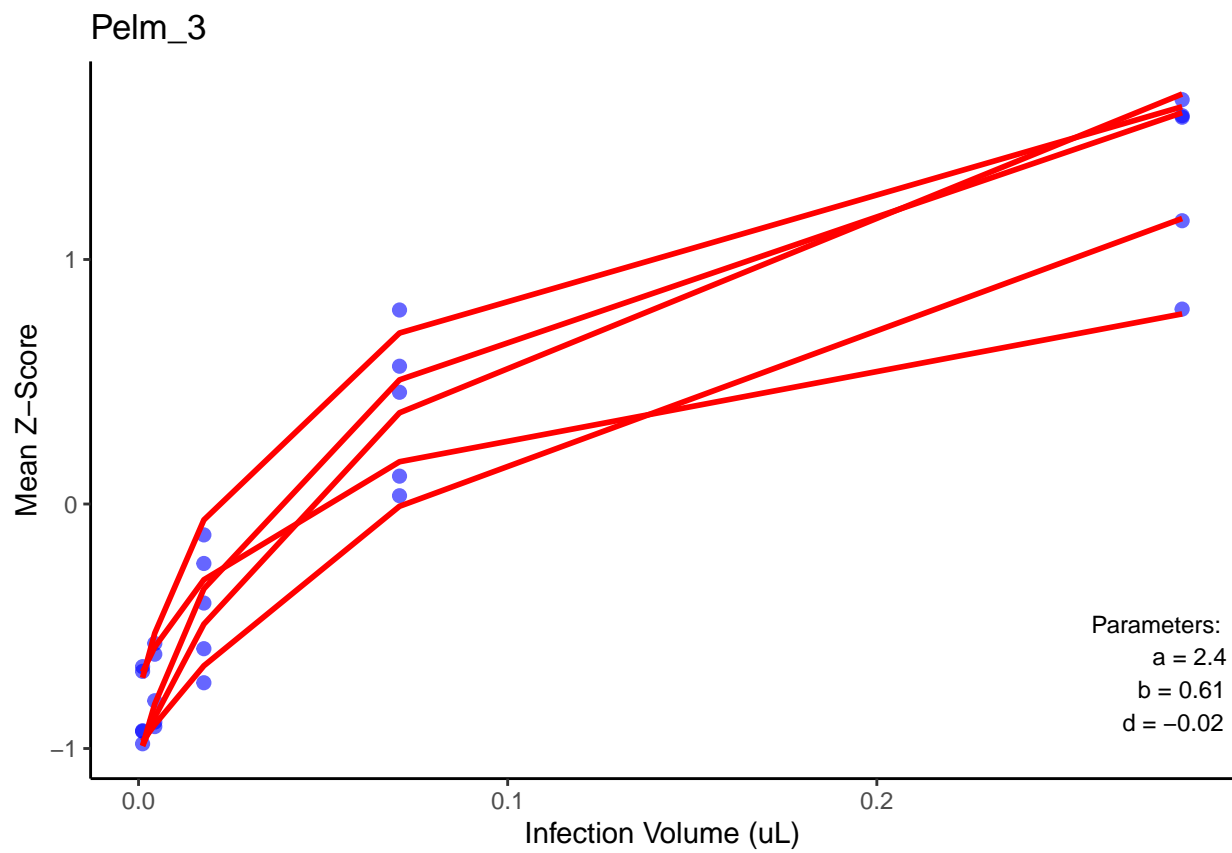
\$Pahc_5



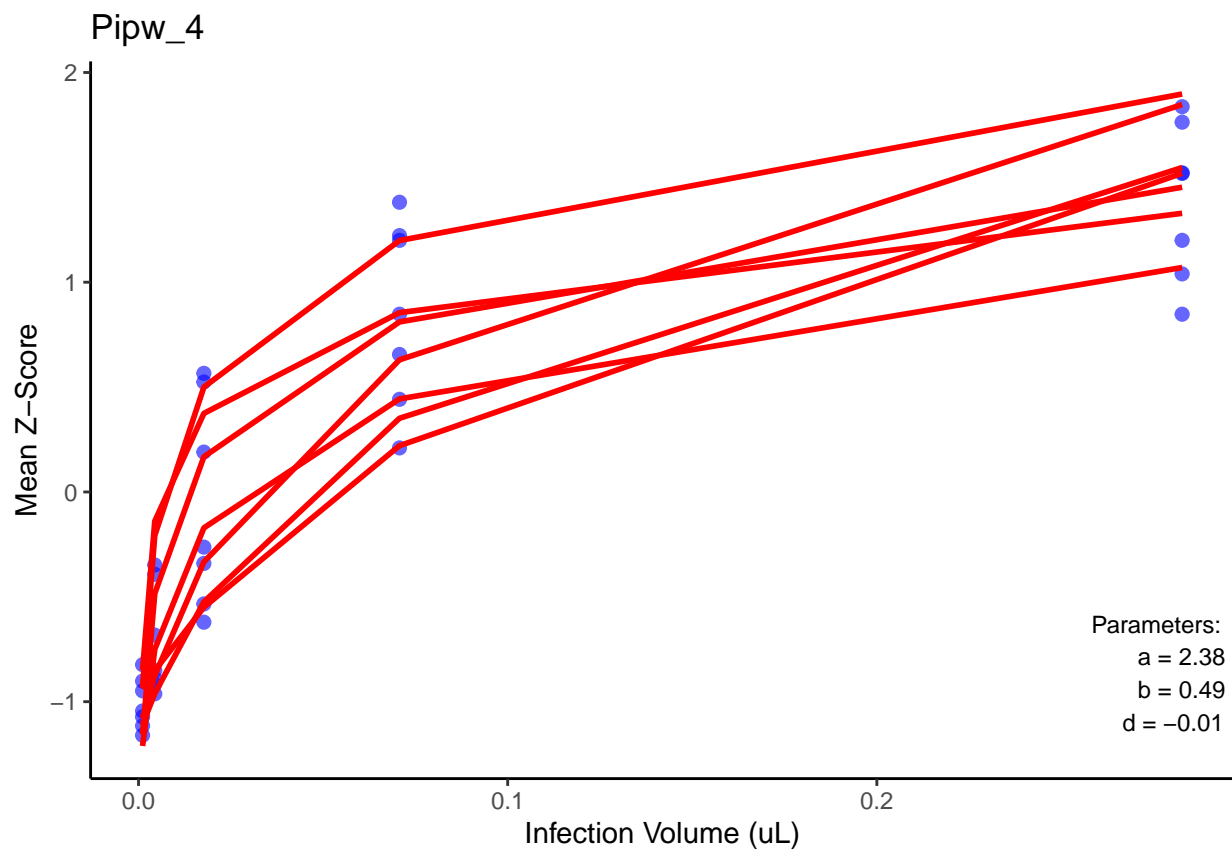
\$Pamv_1



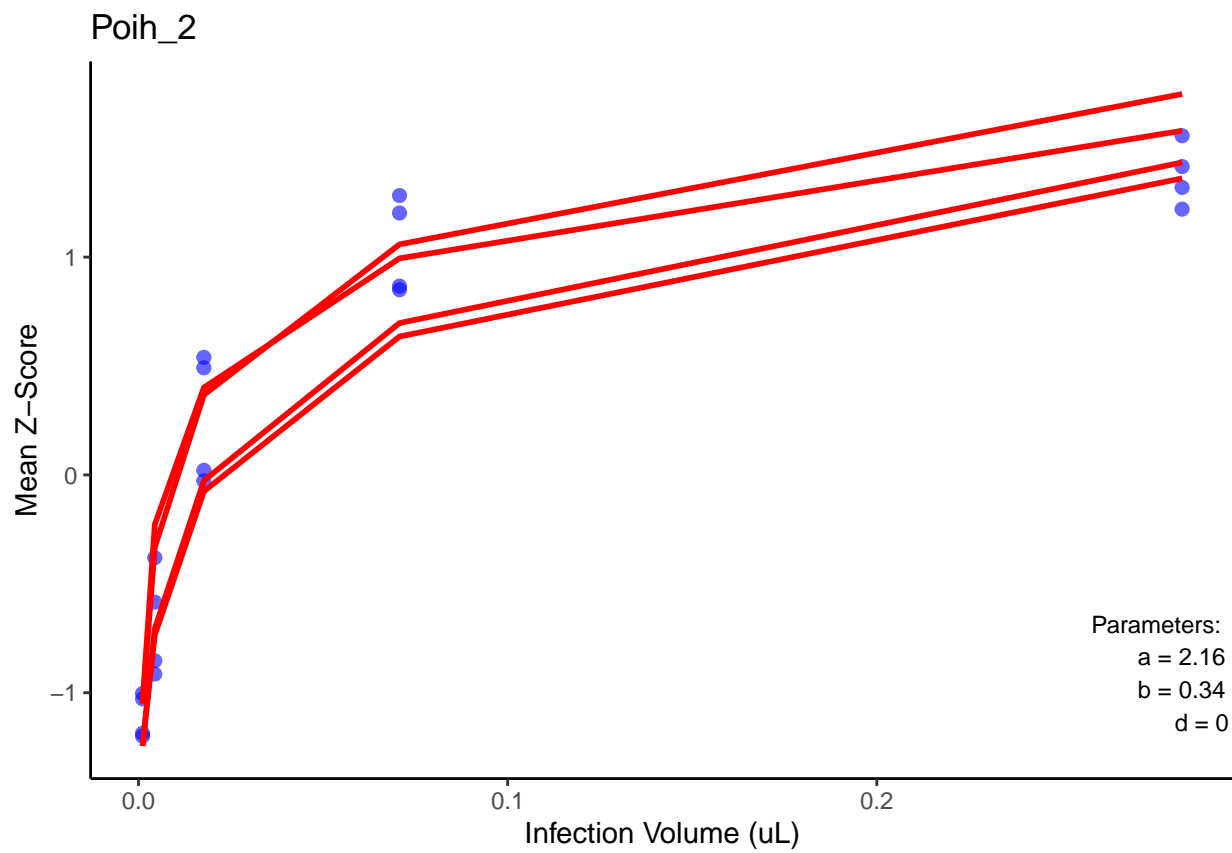
 ## \$Pelm_3



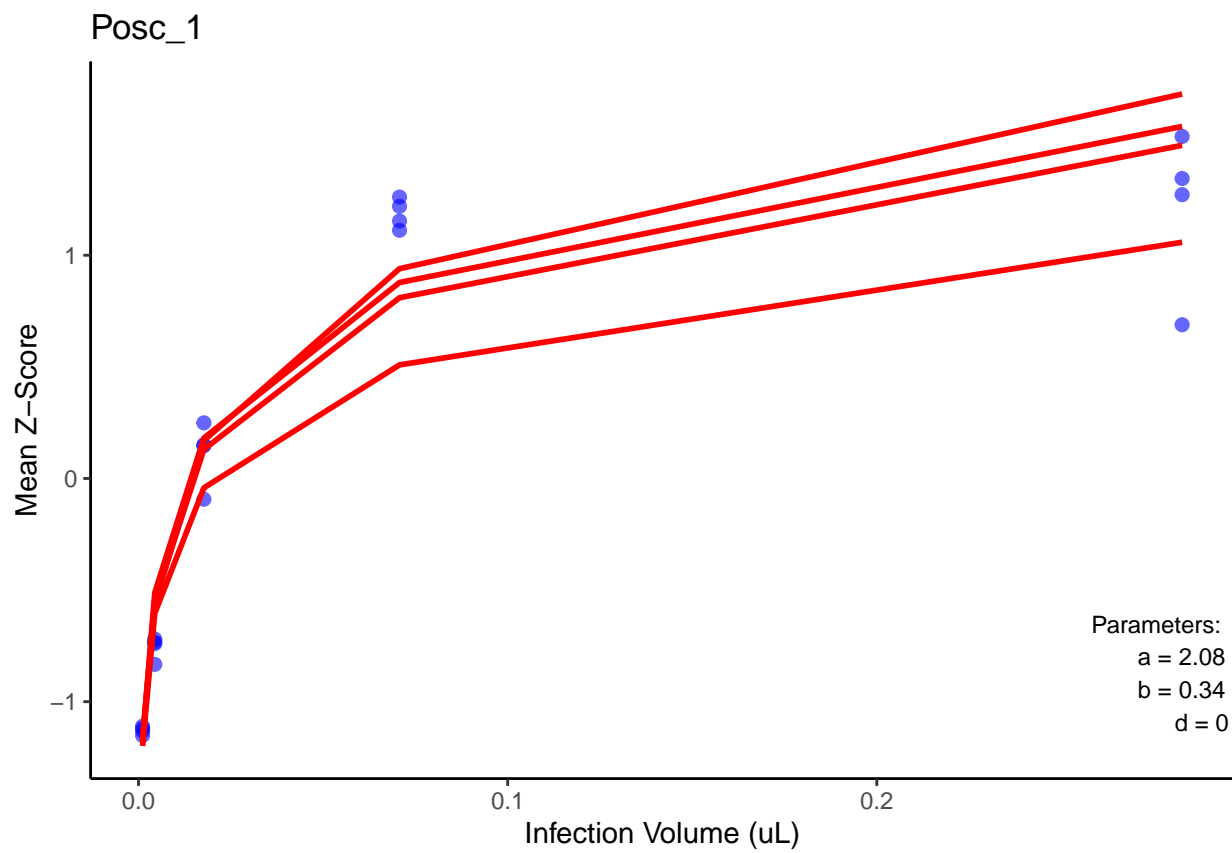
\$Pipw_4



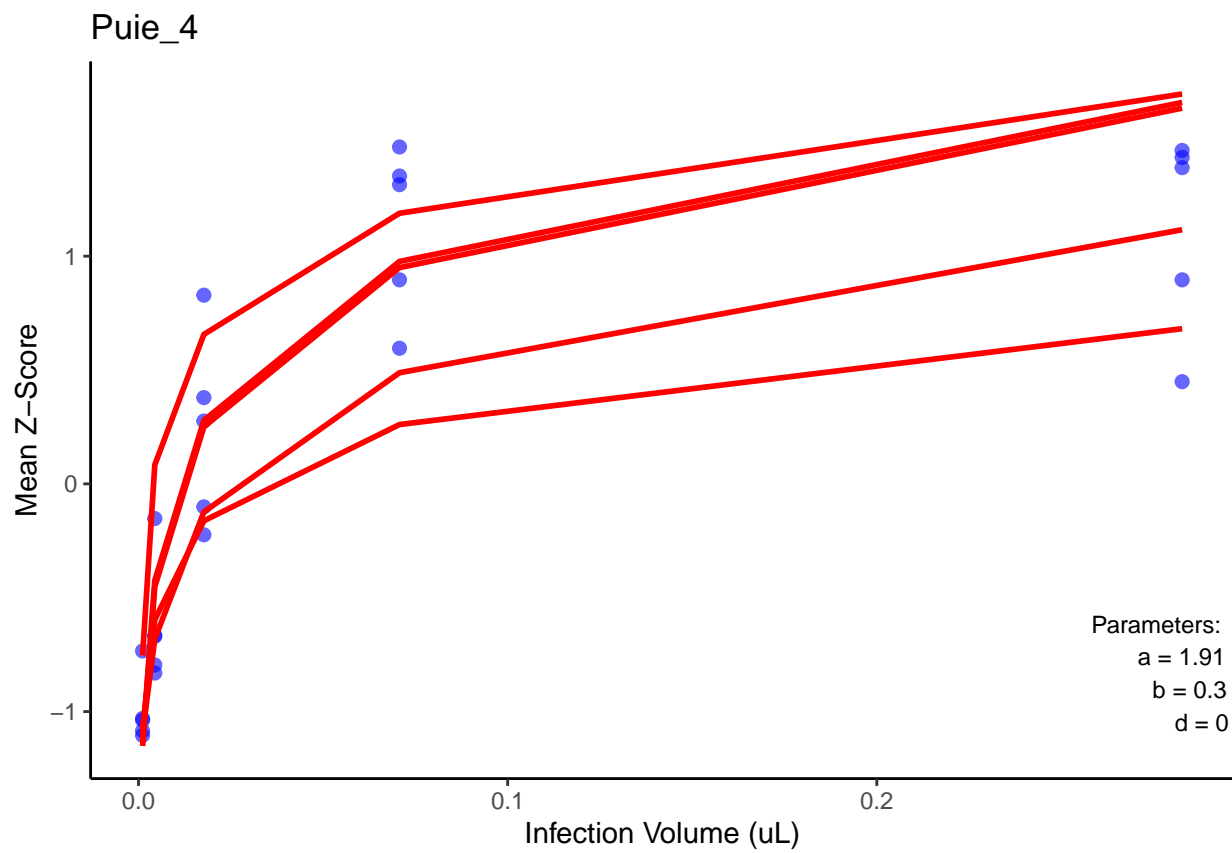
\$Poih_2



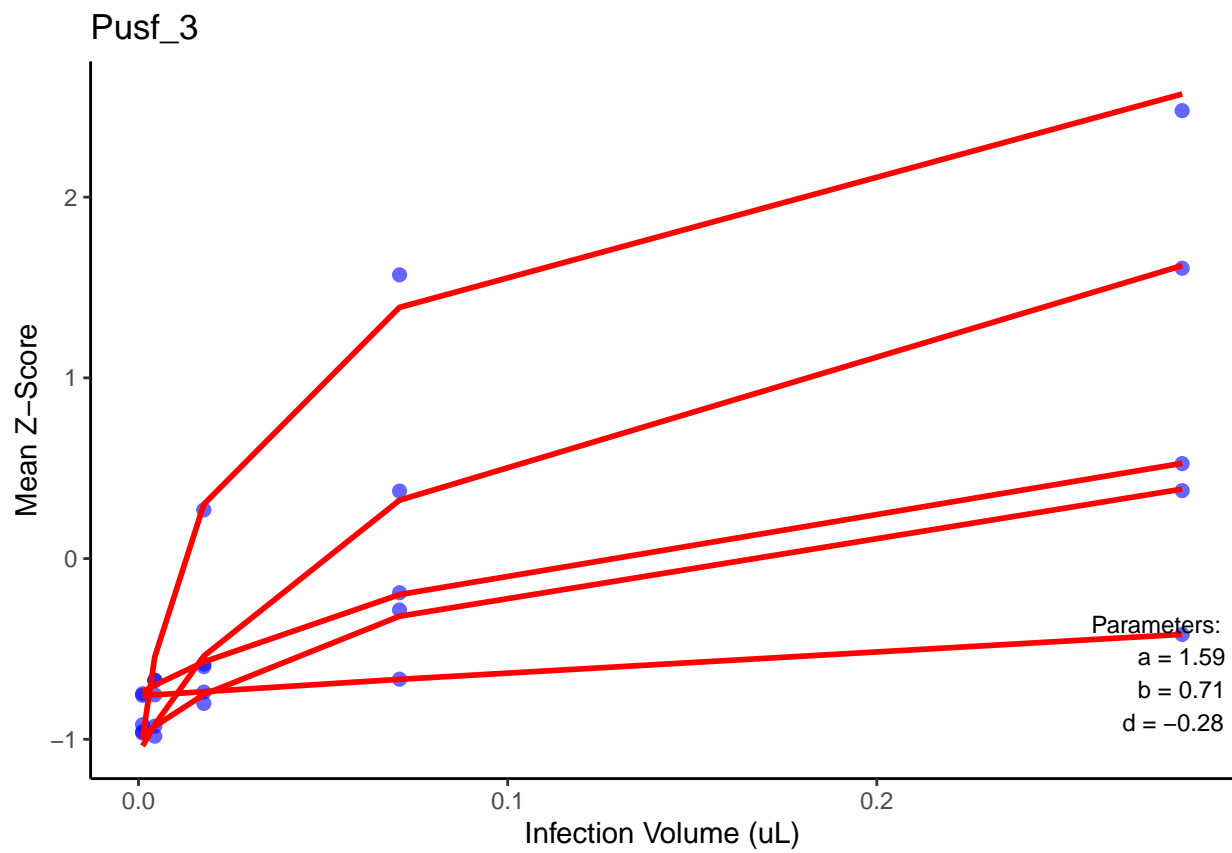
\$Posc_1



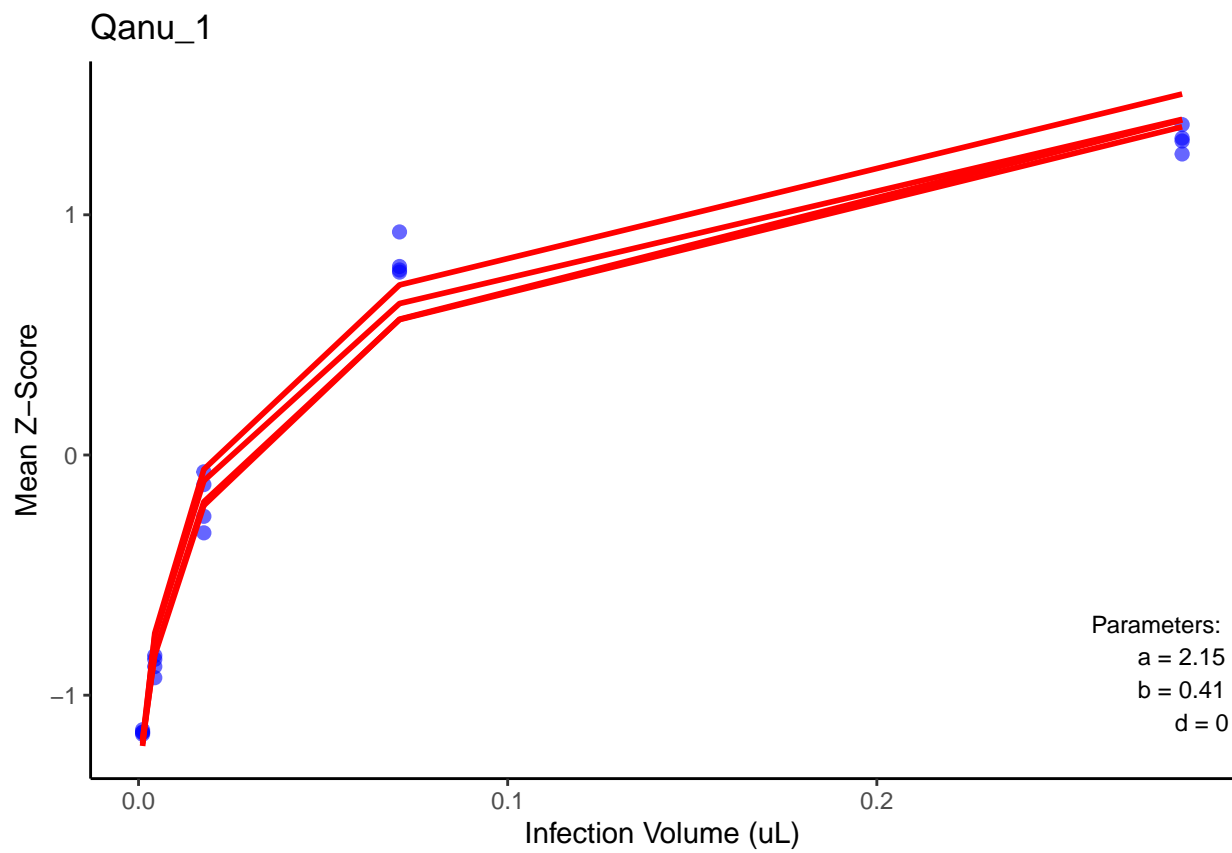
\$Puie_4



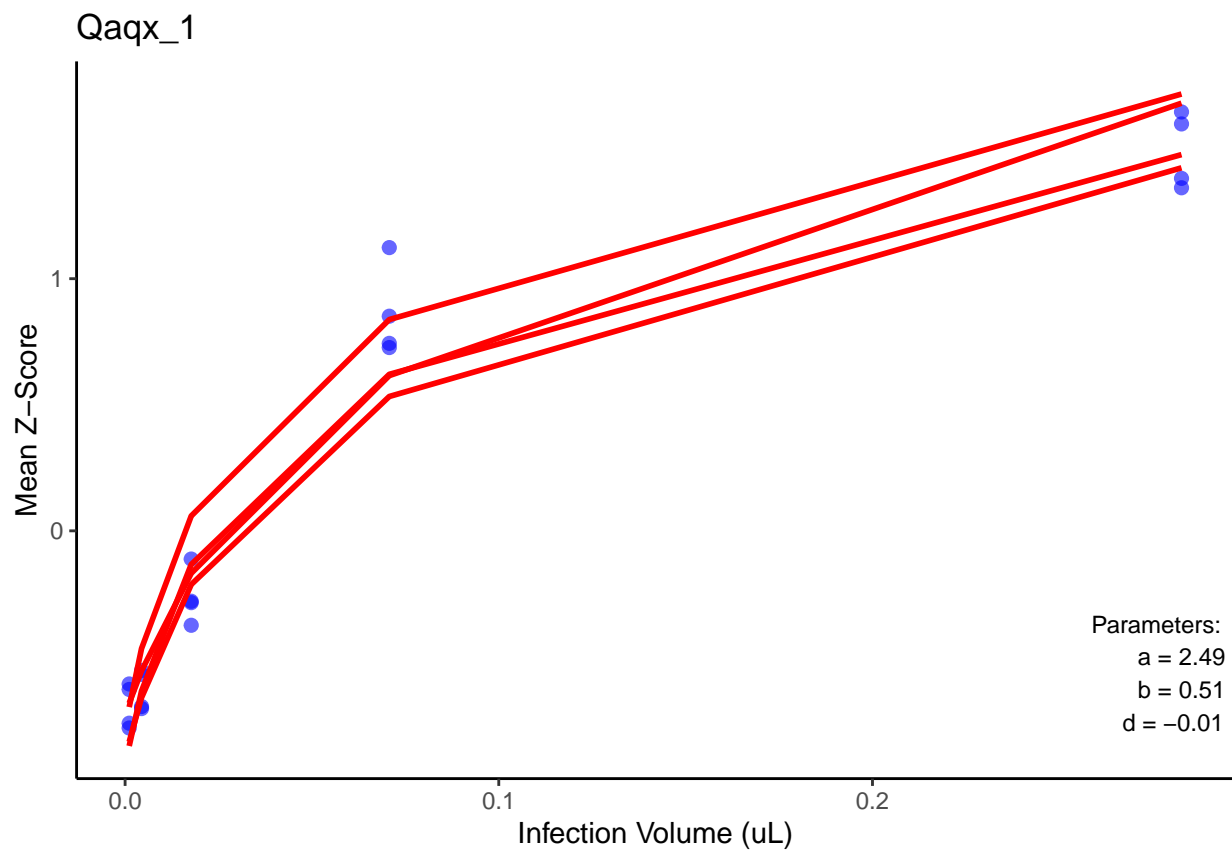
\$Pusf_3



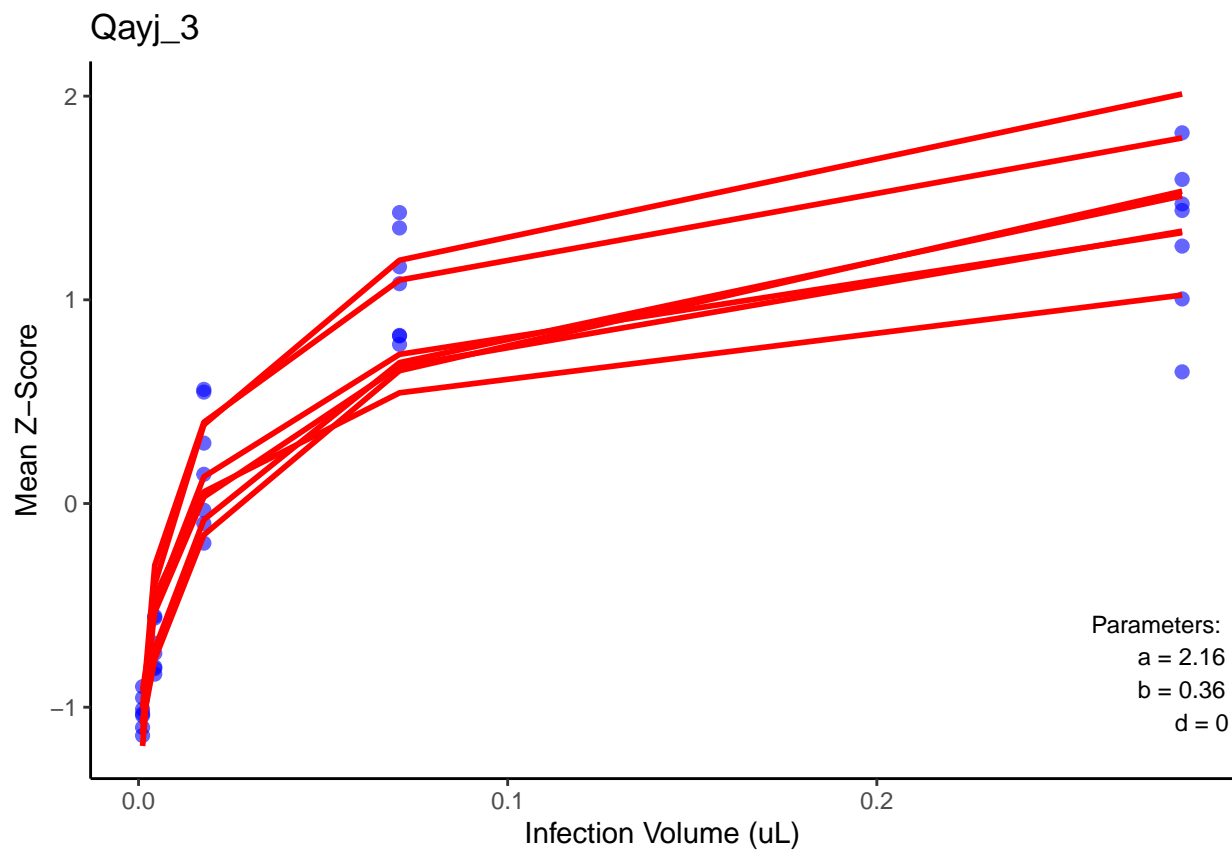
 ## \$Qanu_1



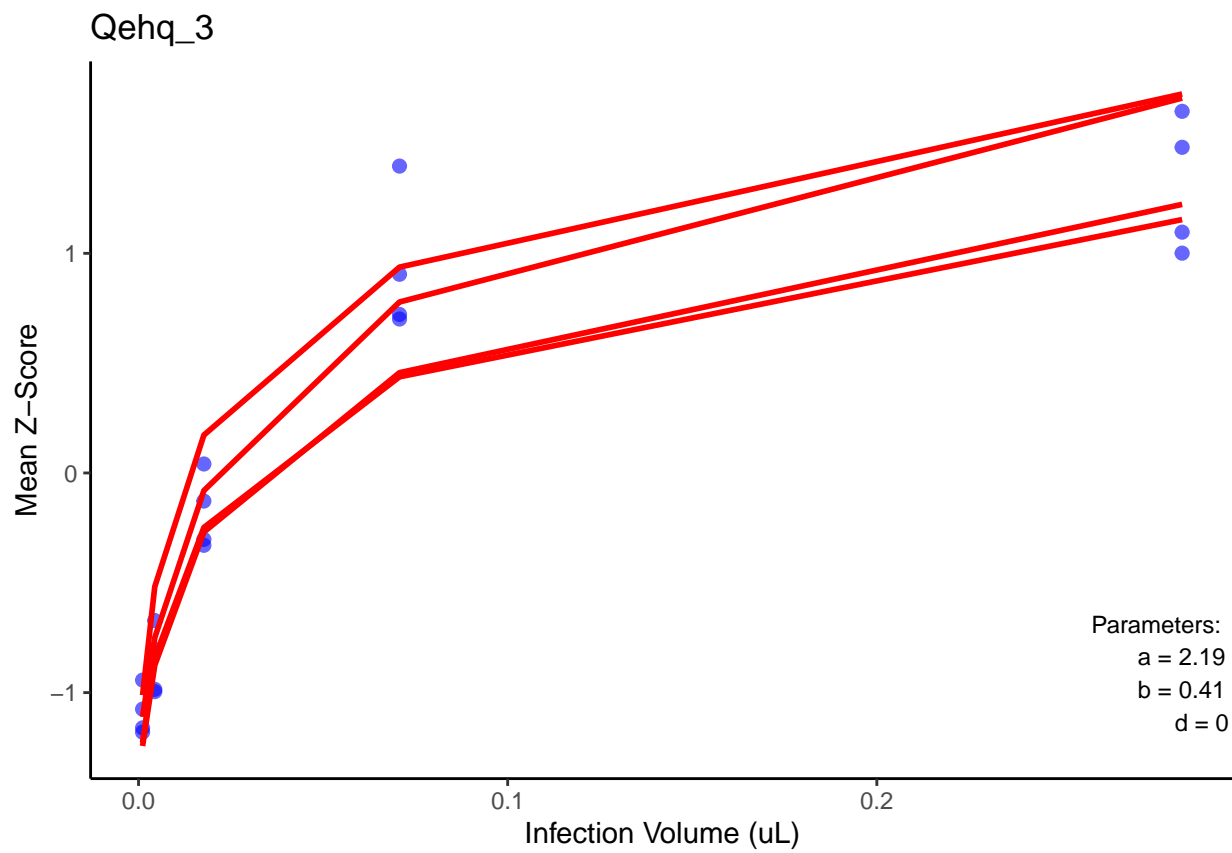
\$Qaqx_1



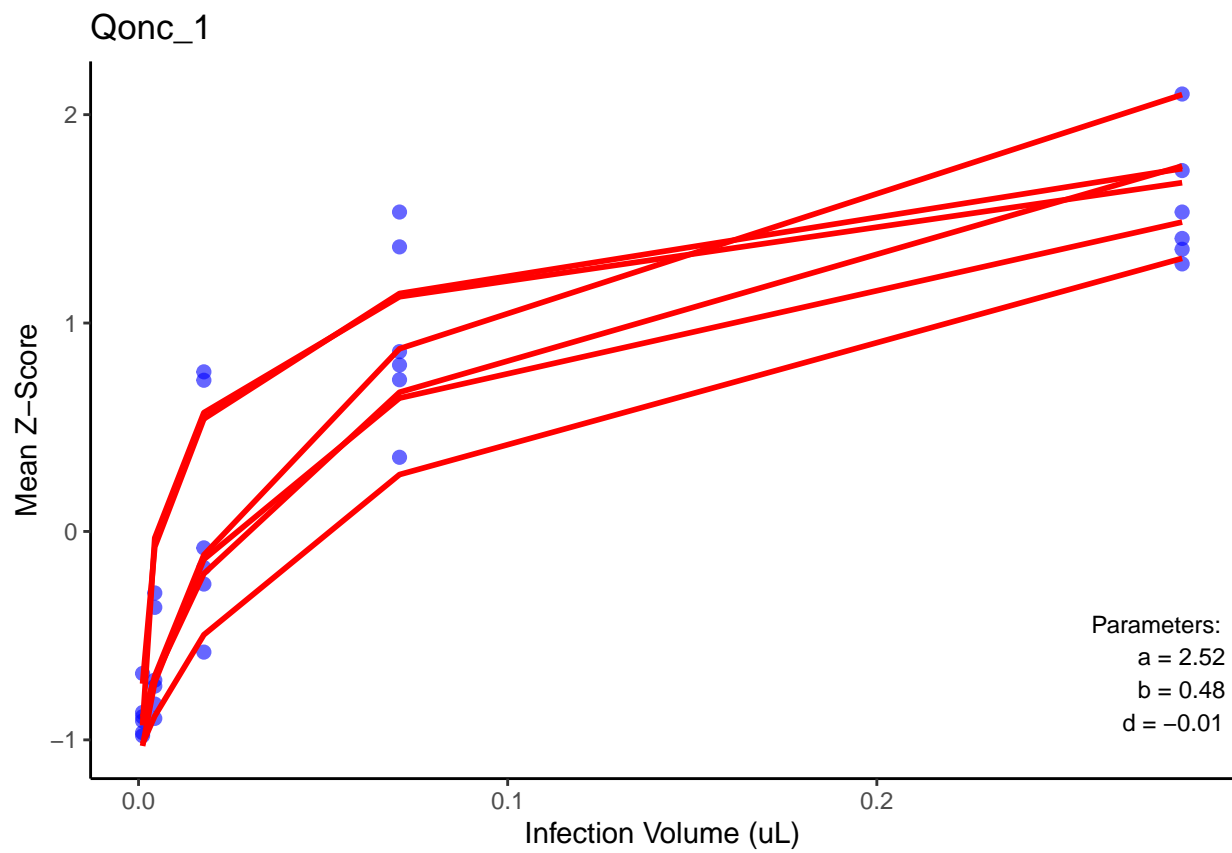
 ## \$Qayj_3



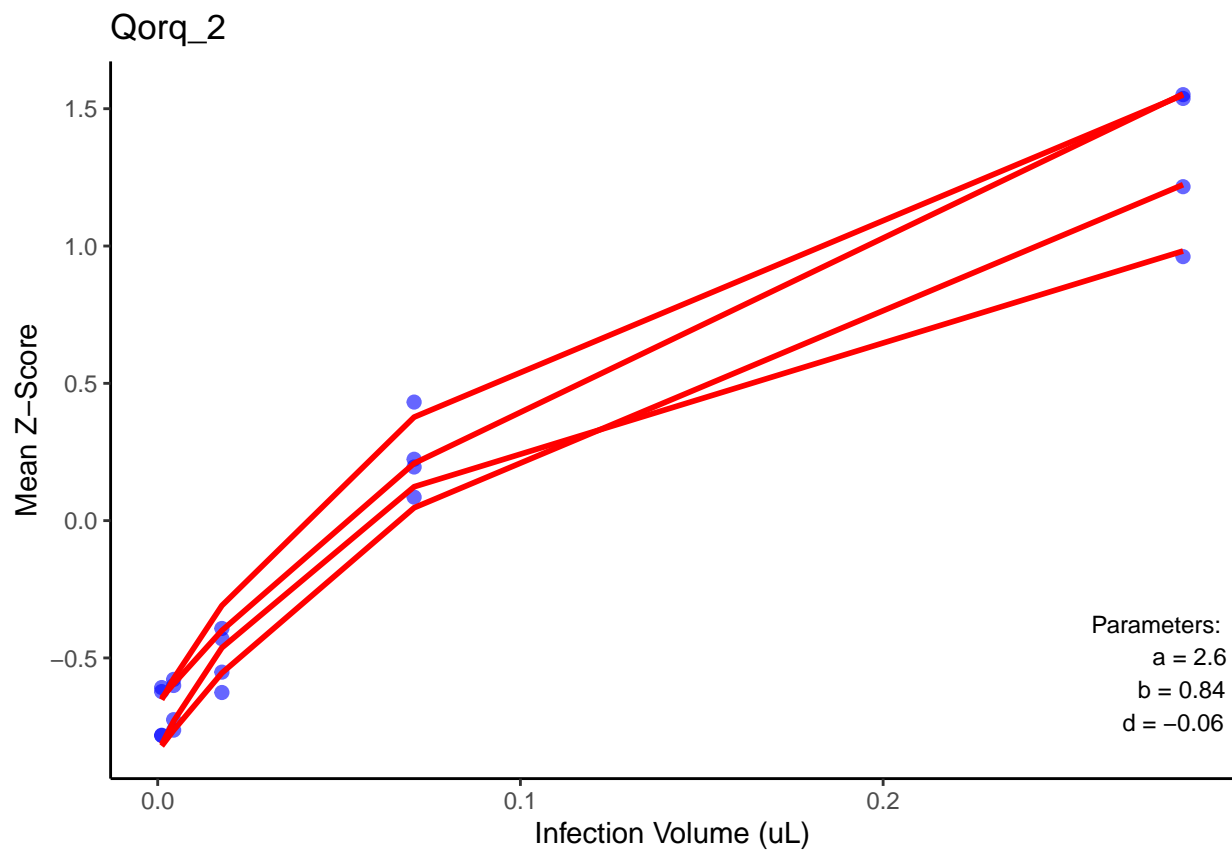
\$Qehq_3



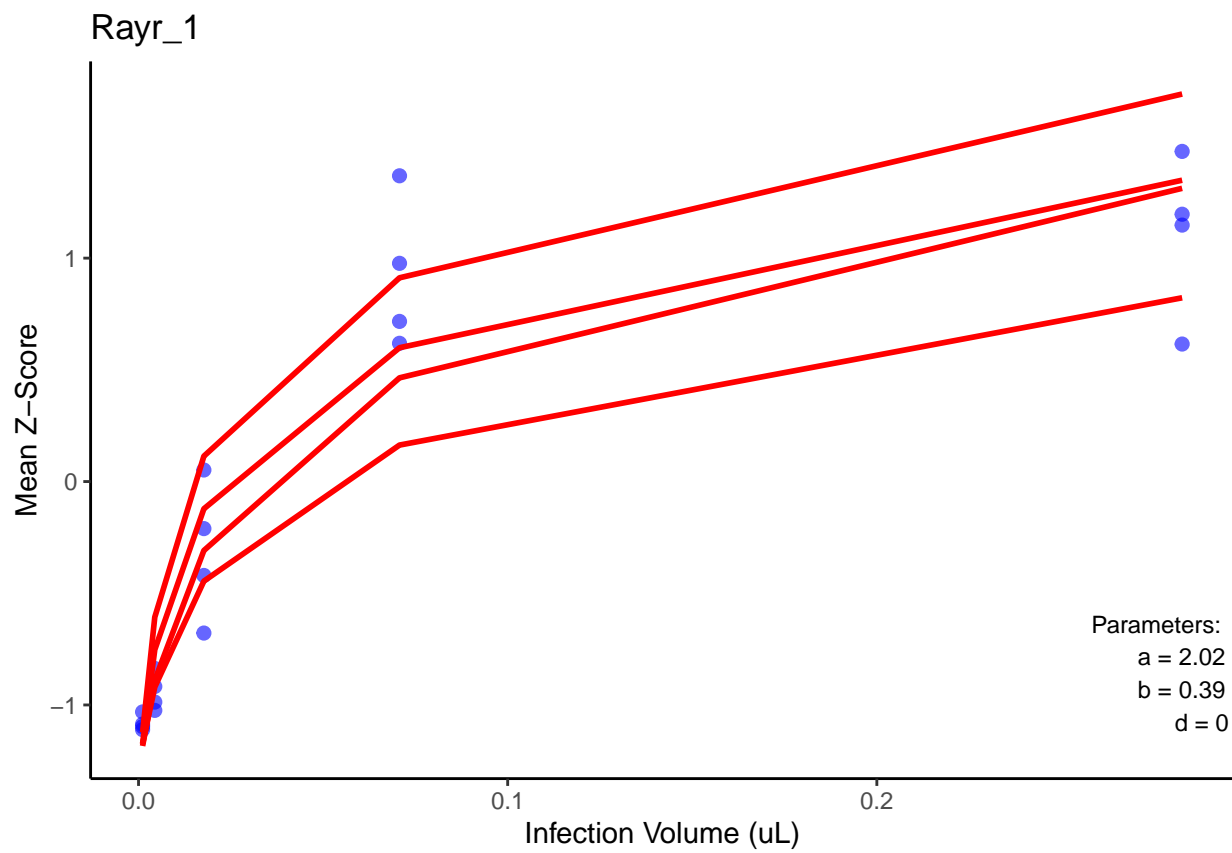
\$Qonc_1



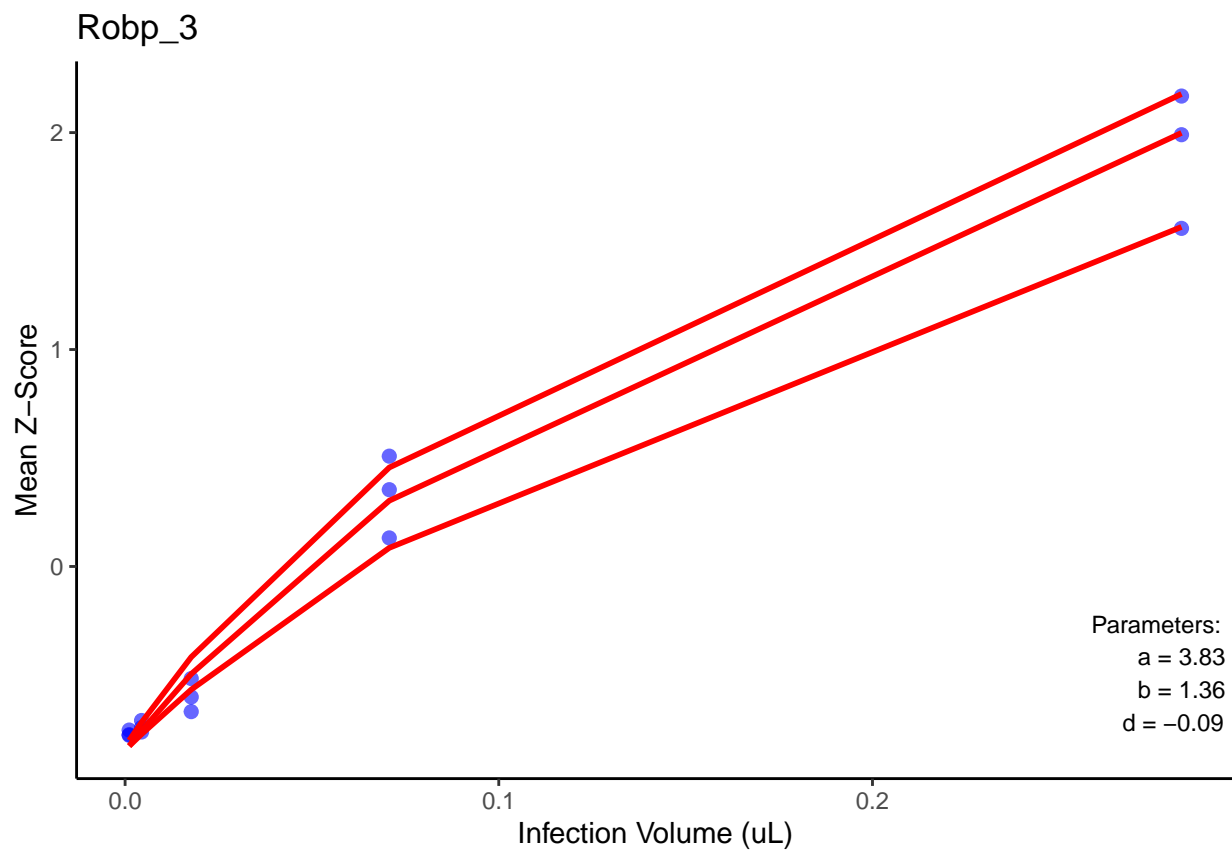
 ## \$Qorq_2



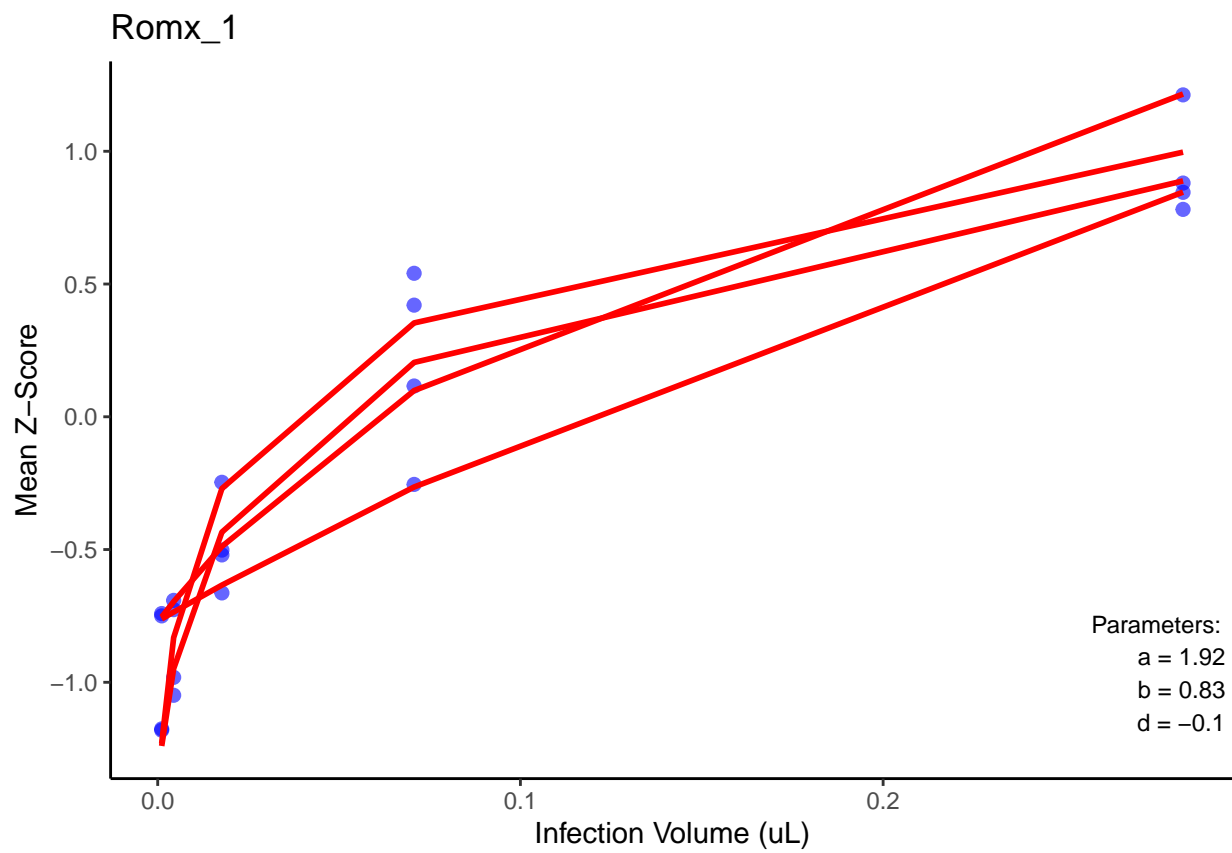
\$Rayr_1



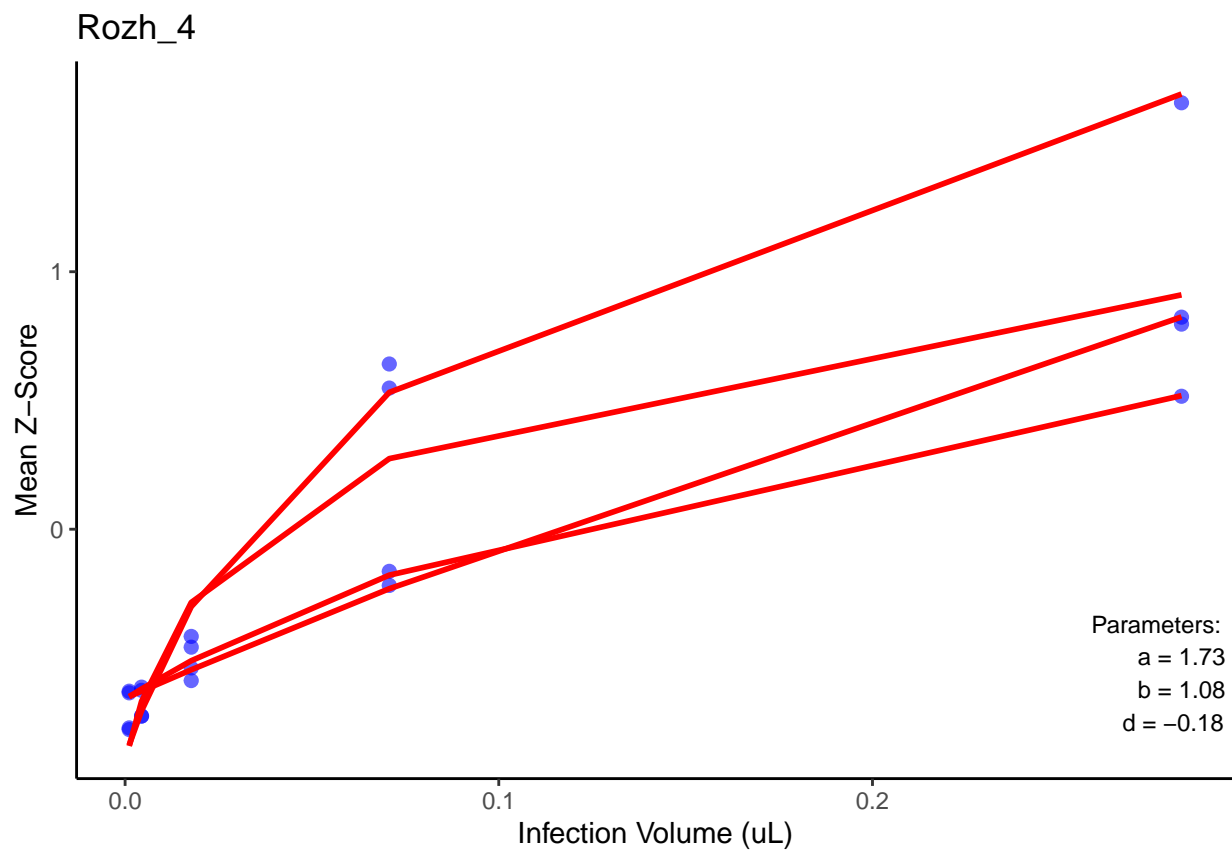
 ## \$Robp_3



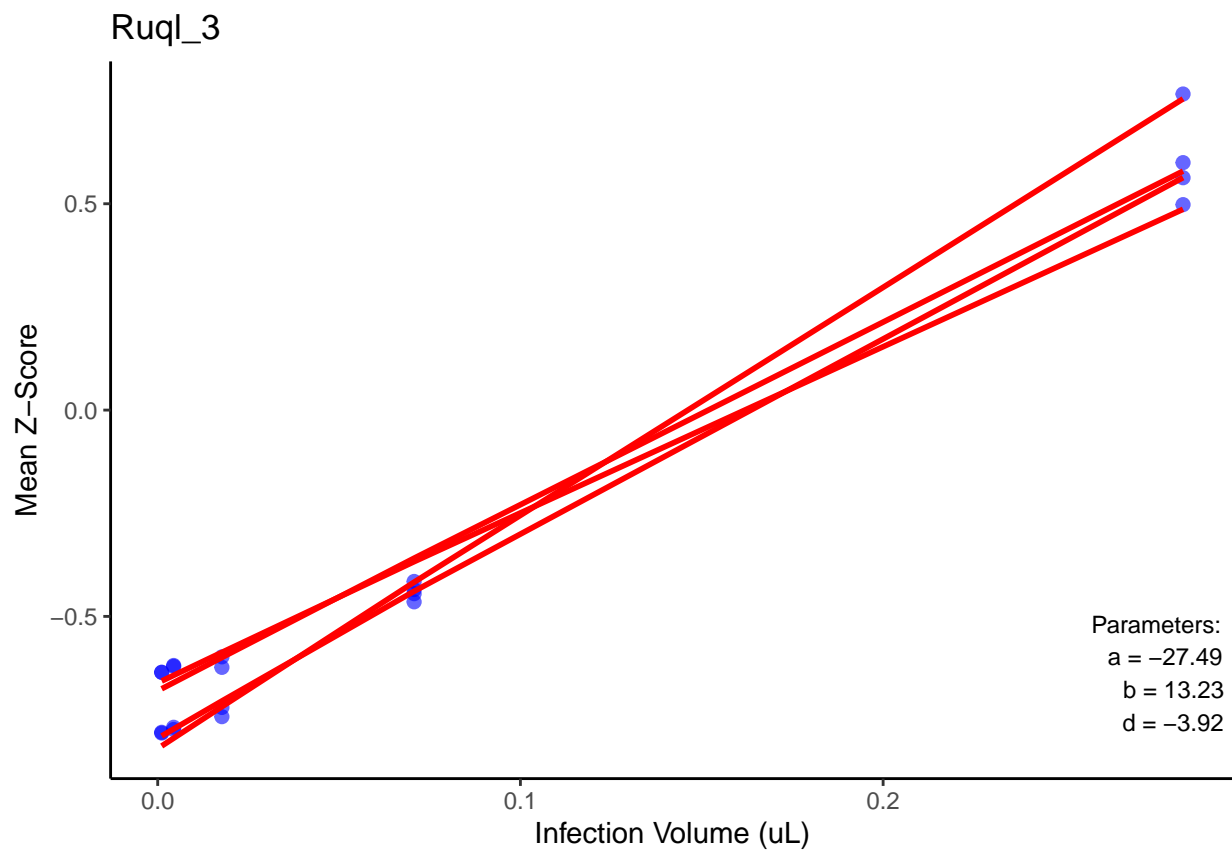
\$Romx_1



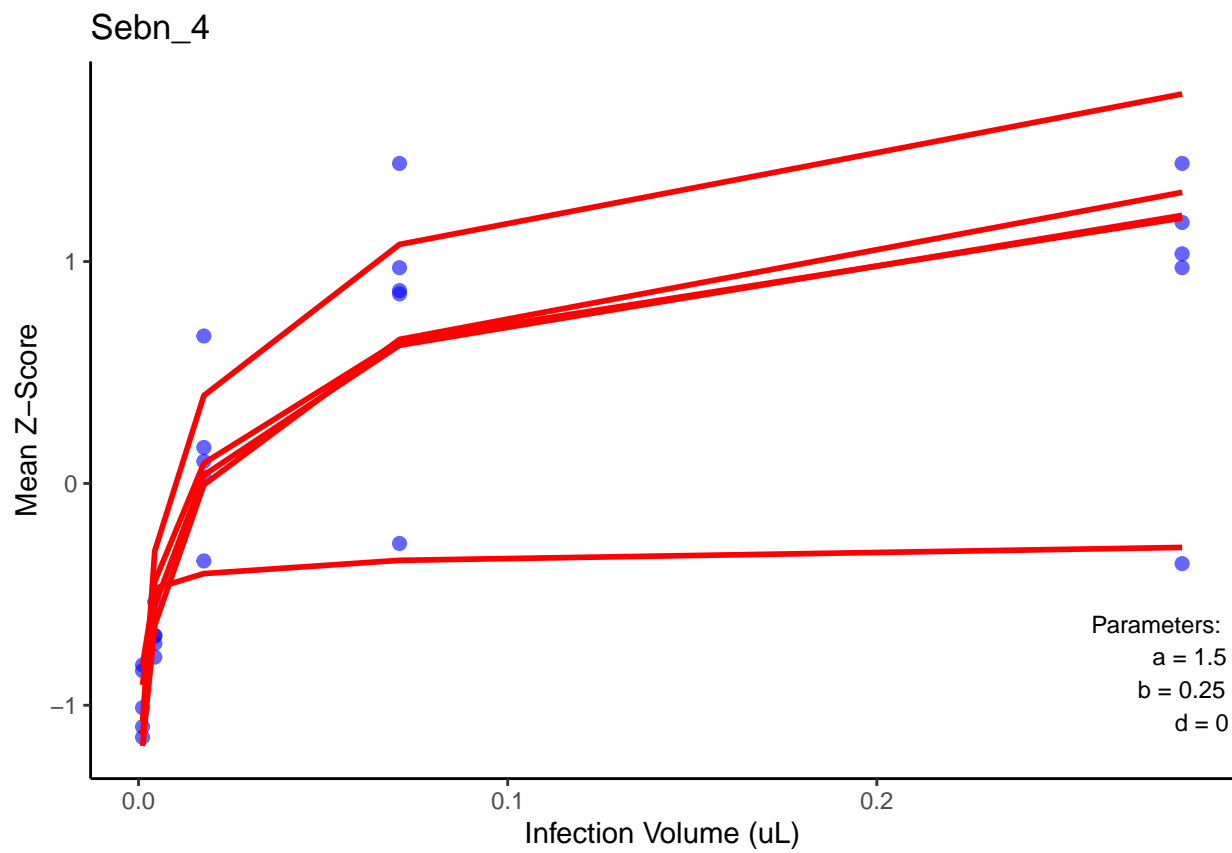
\$Rozh_4



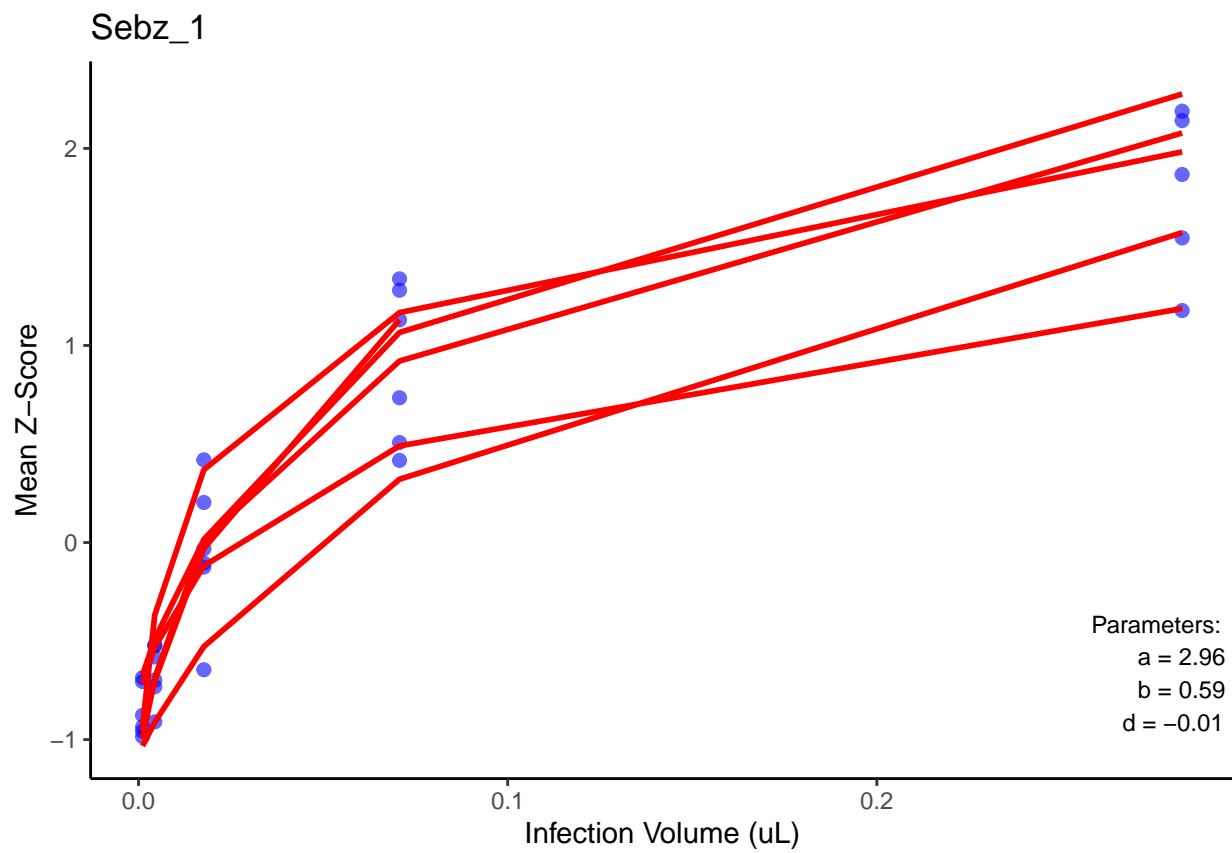
\$Ruql_3



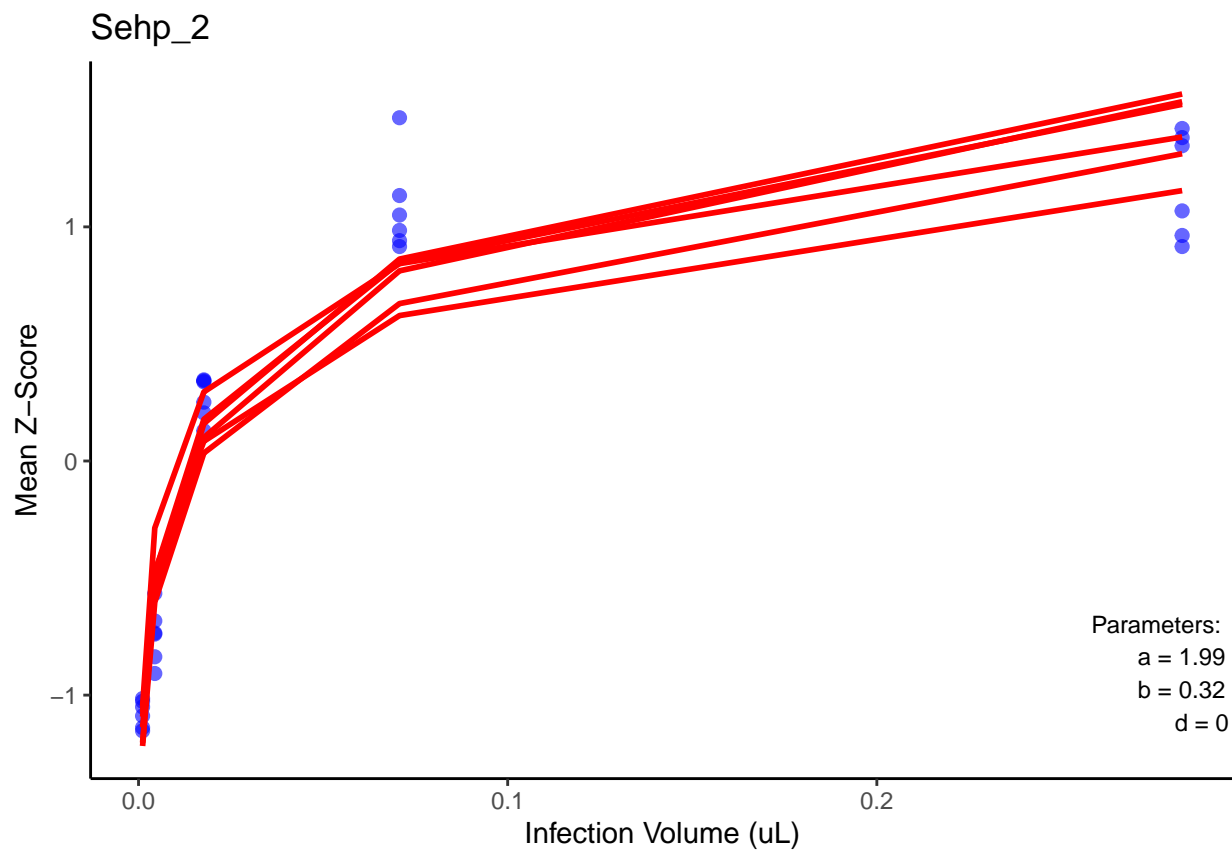
\$Sebn_4



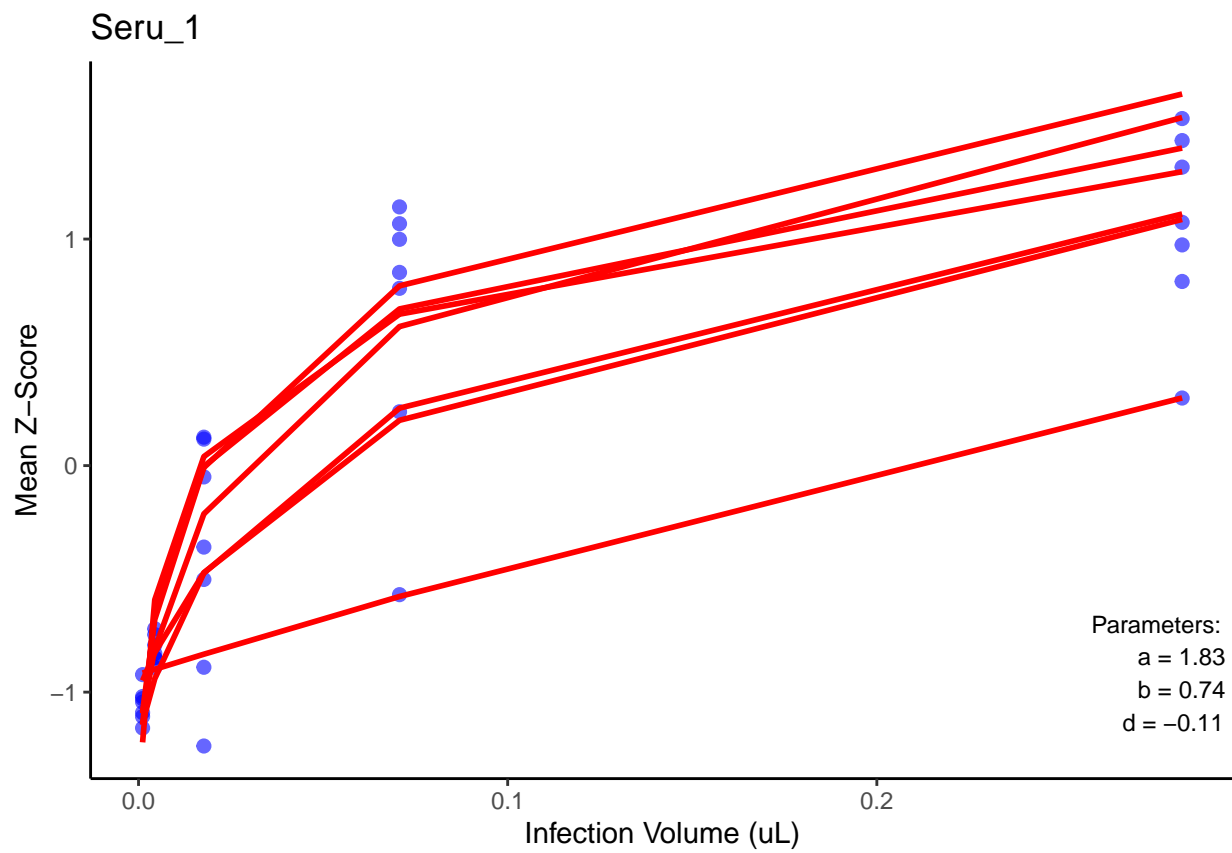
\$Sebz_1



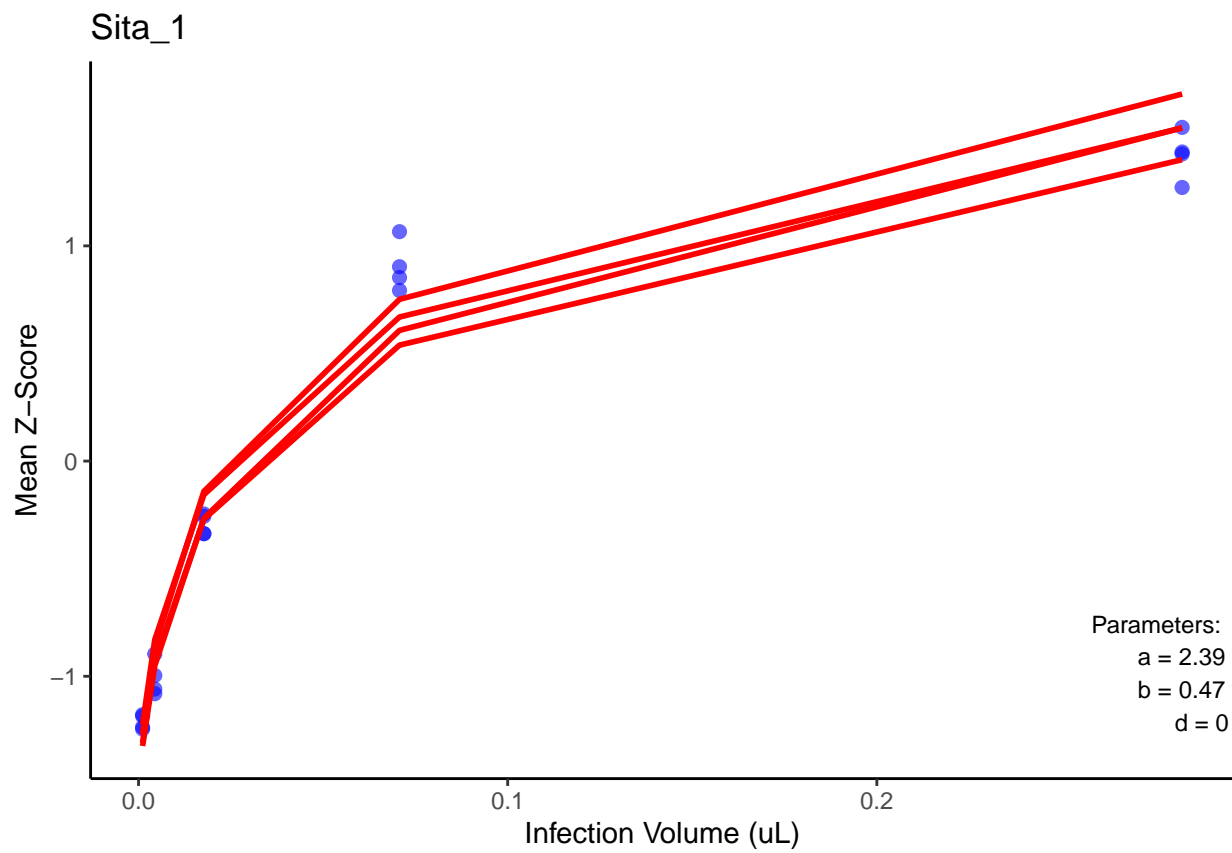
\$Sehp_2



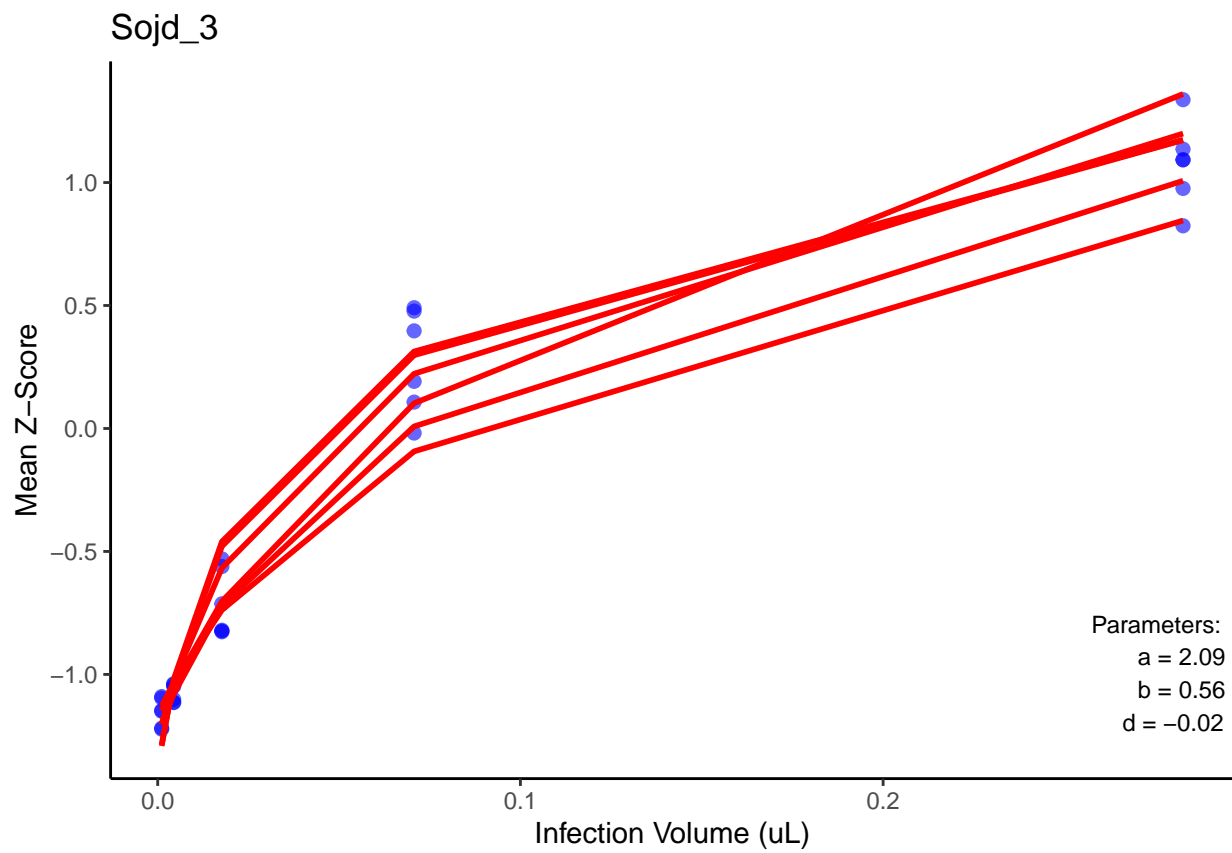
\$Seru_1



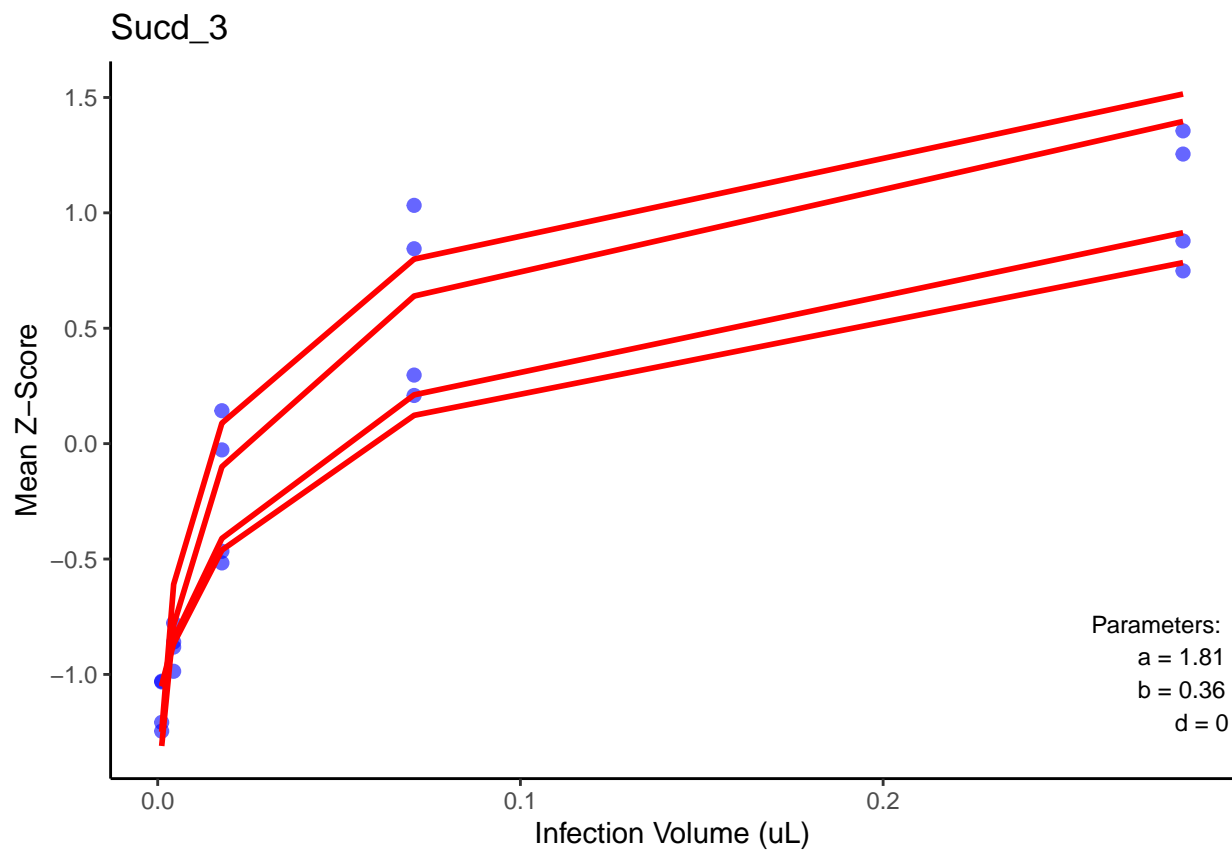
 ## \$Sita_1



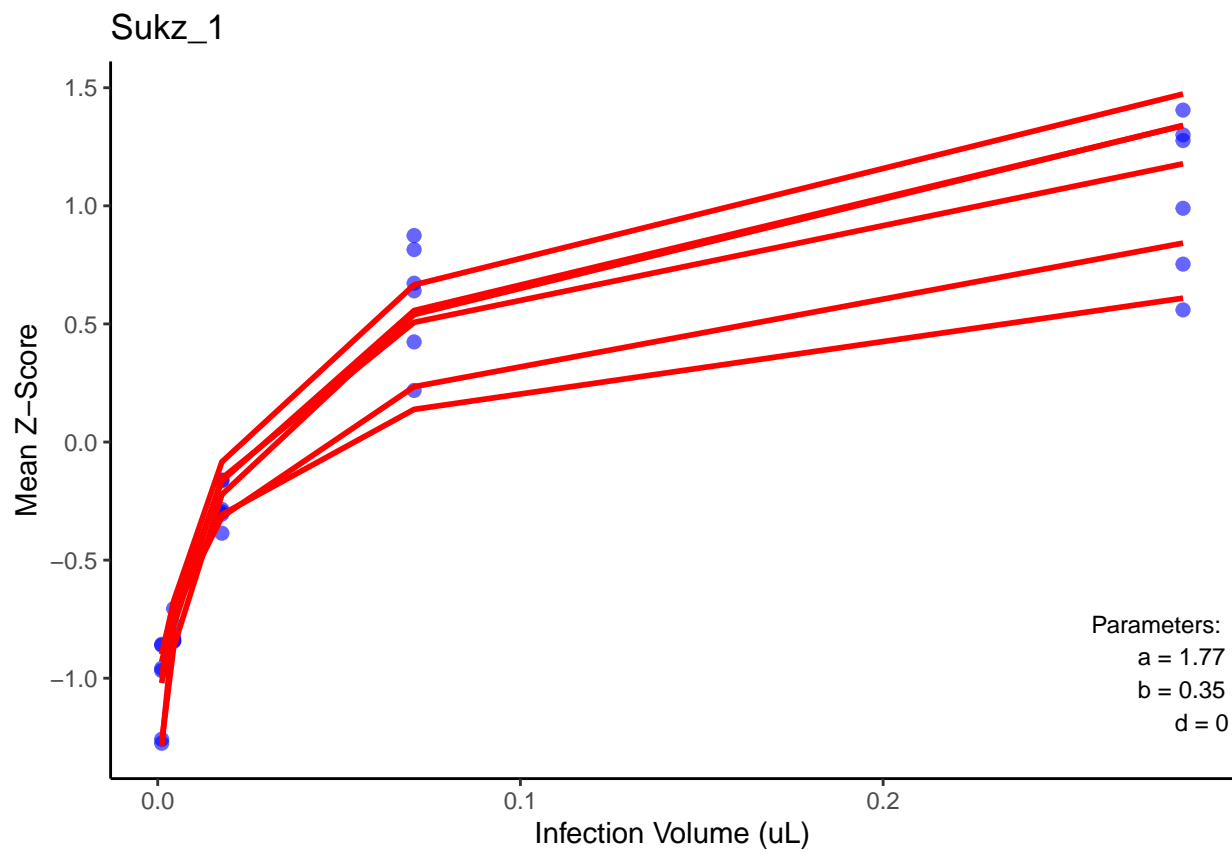
\$Sojd_3



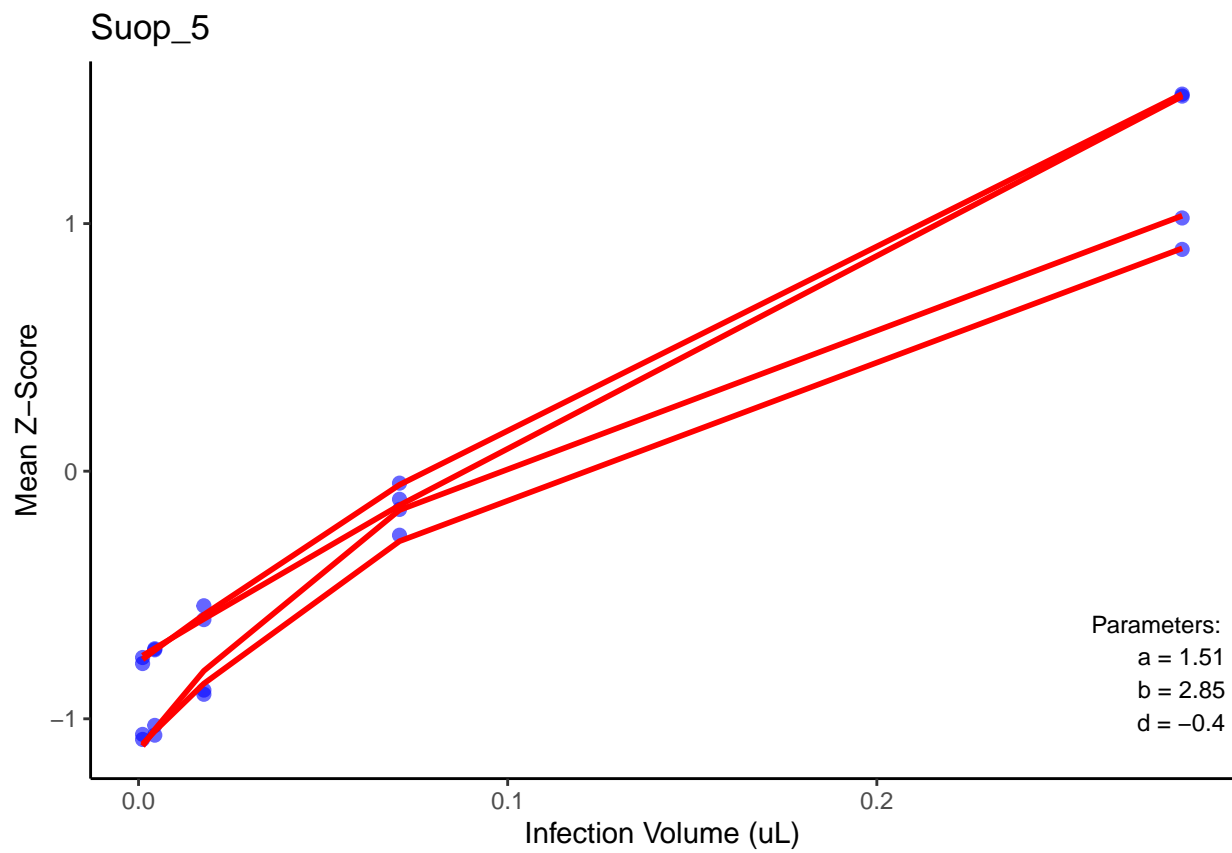
\$Sucd_3



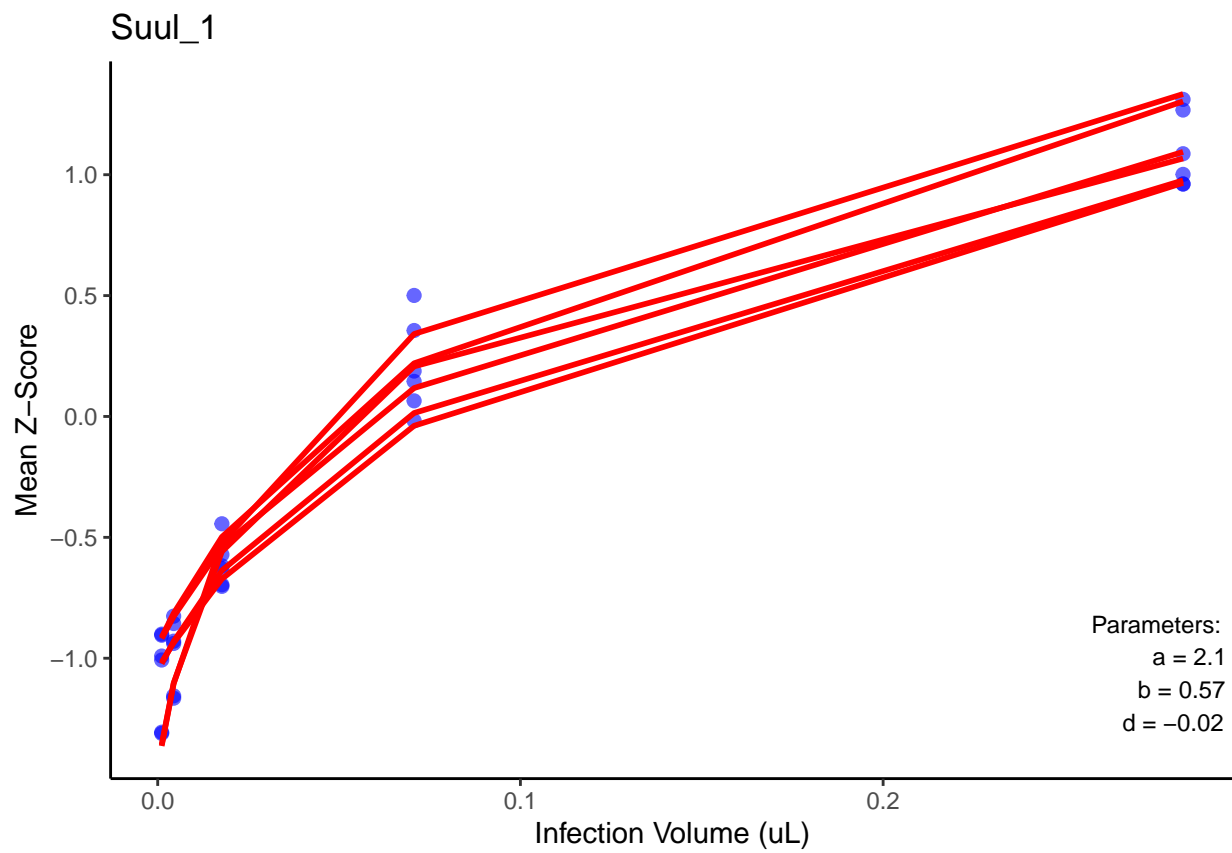
\$Sukz_1



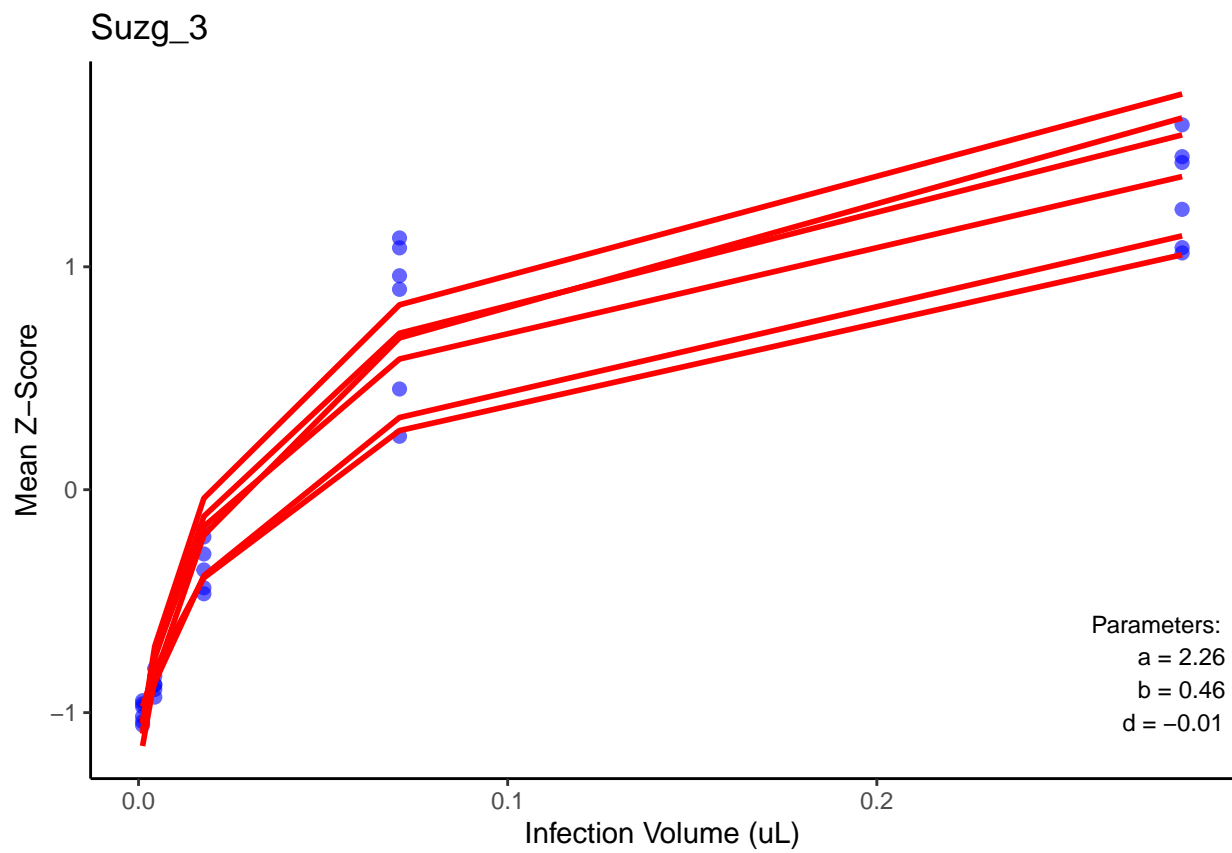
 ## \$Suop_5



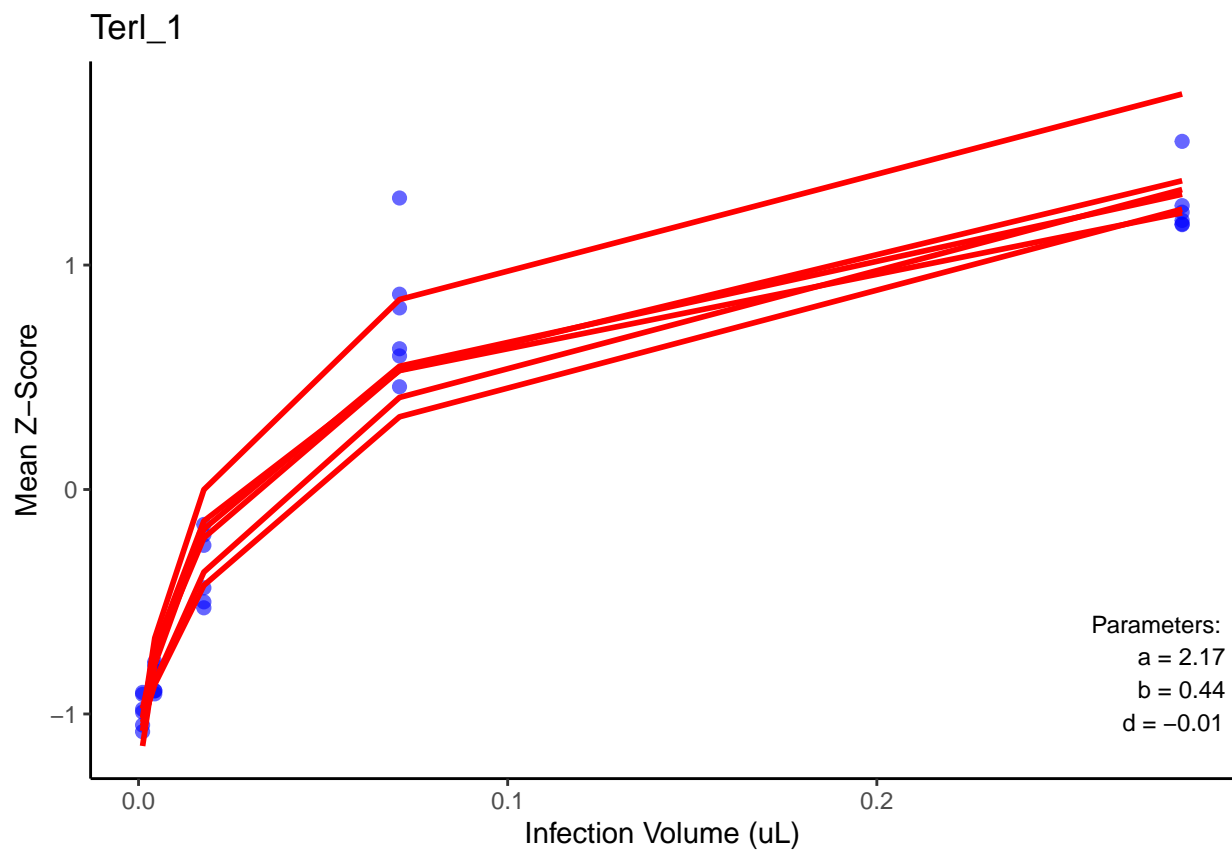
\$Suul_1



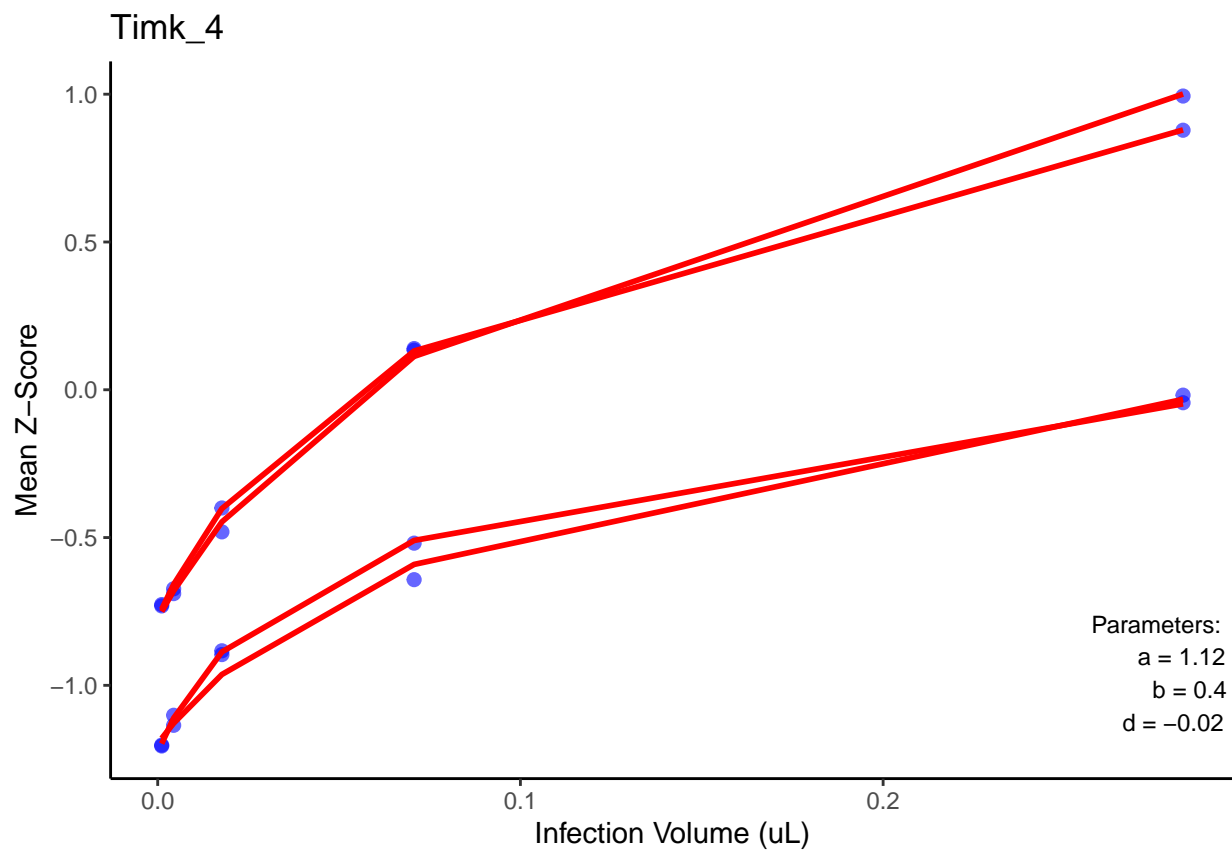
\$Suzg_3



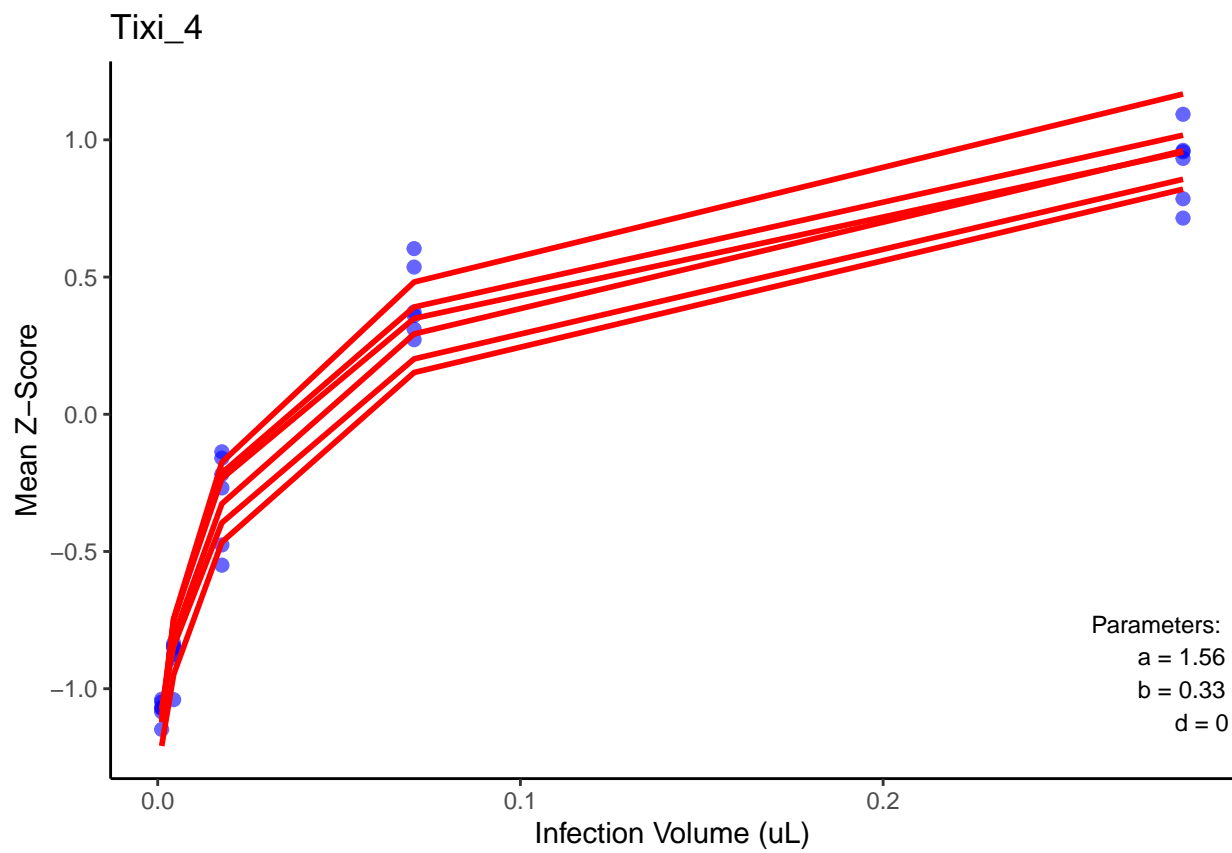
\$Terl_1



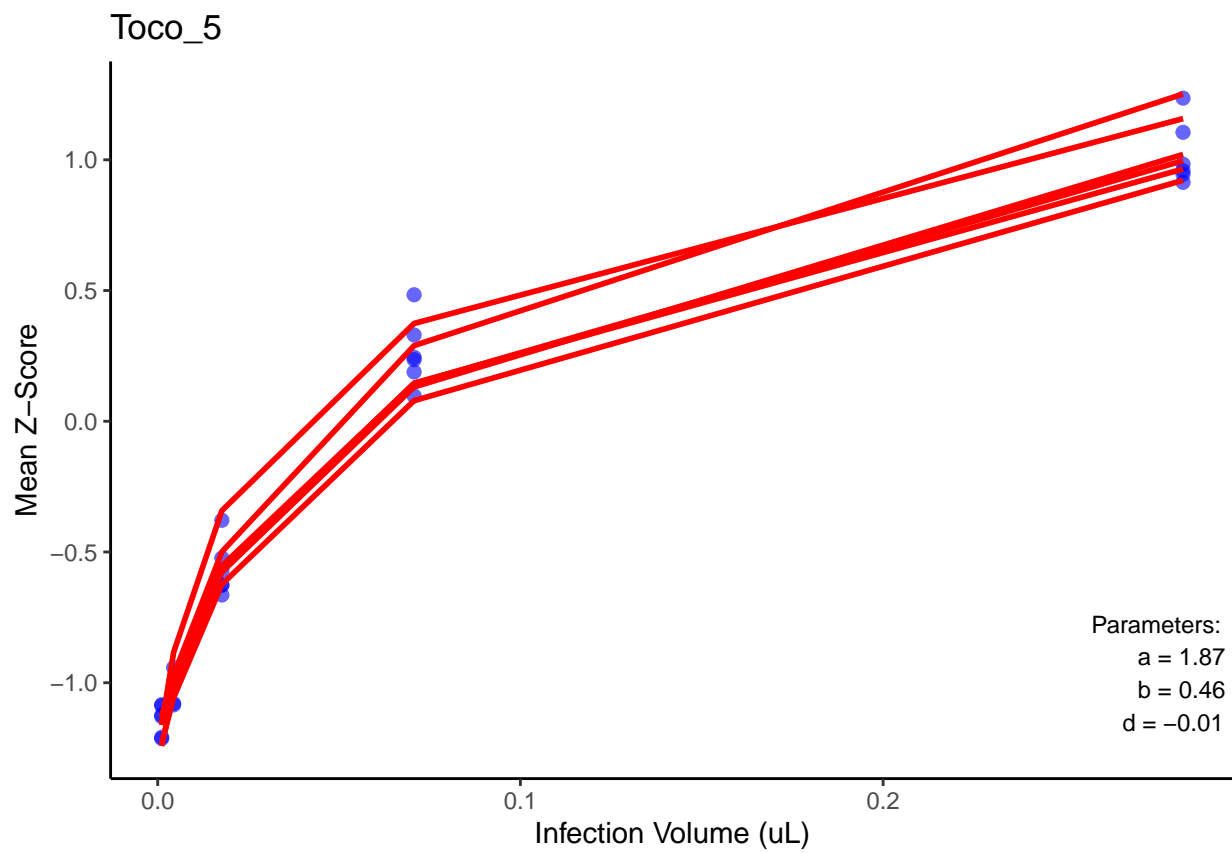
\$Timk_4



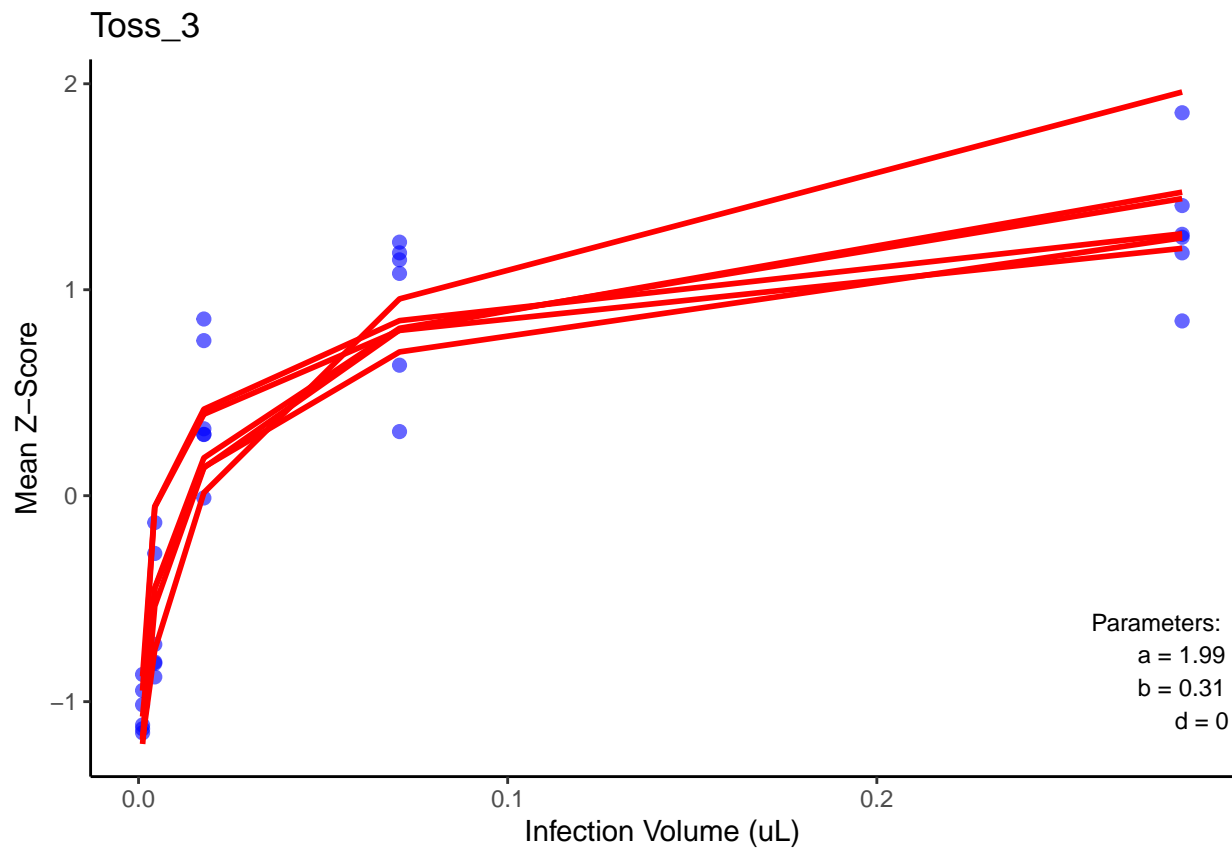
\$Tixi_4



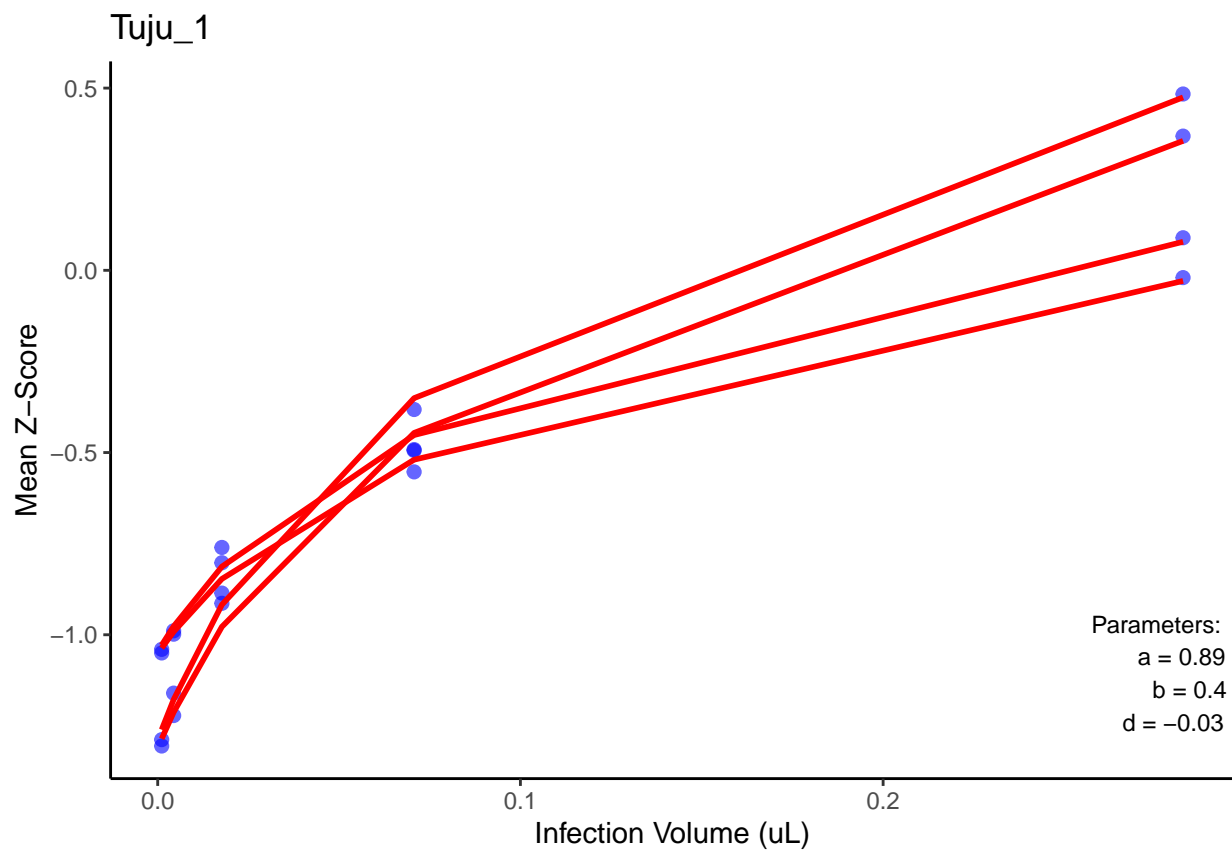
 ## \$Toco_5



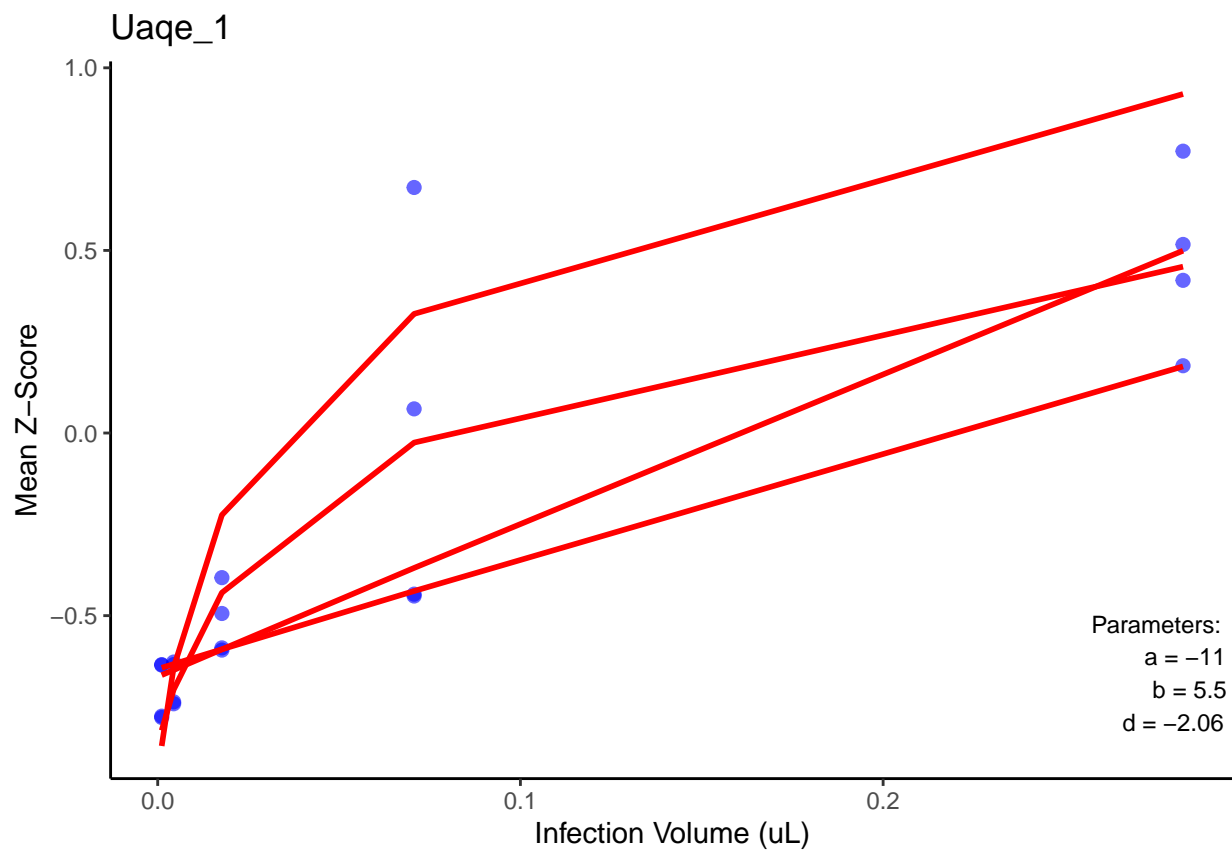
 ## \$Tolg_6



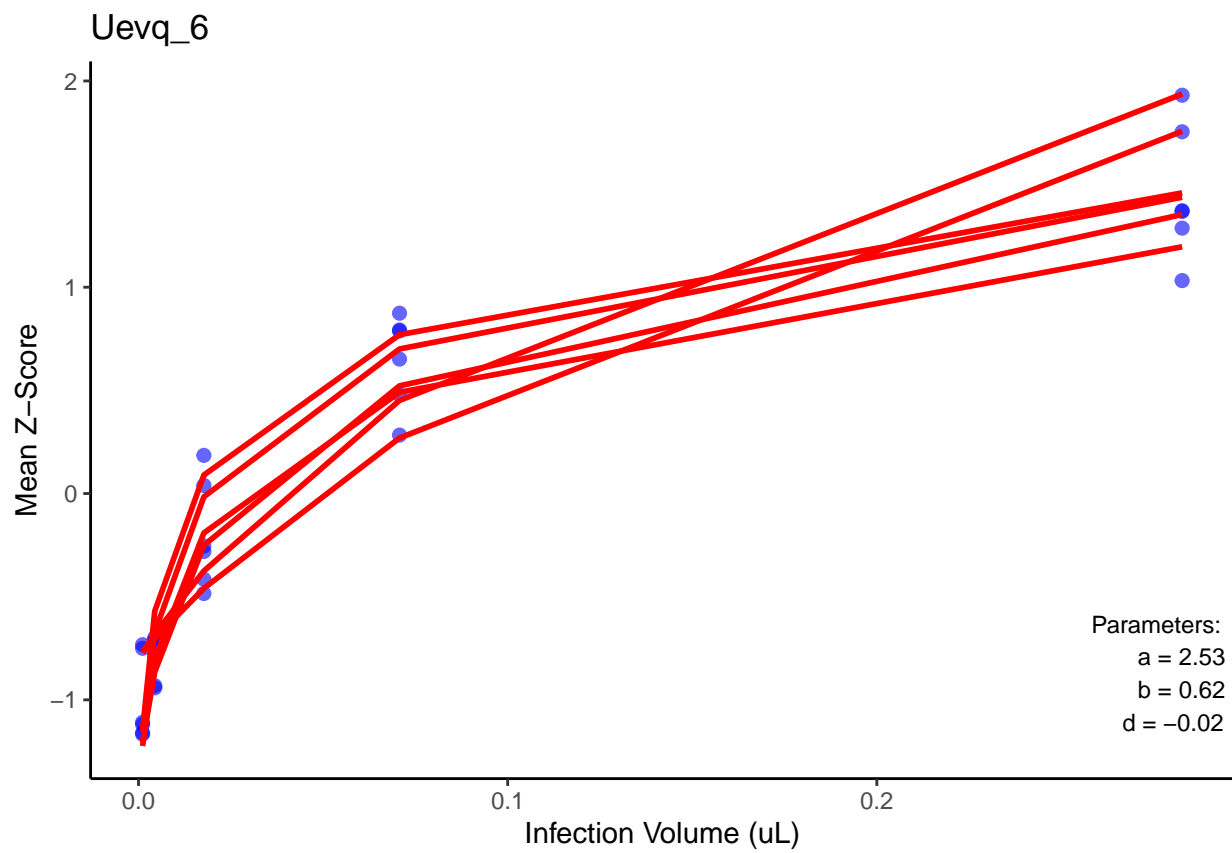
\$Tuju_1



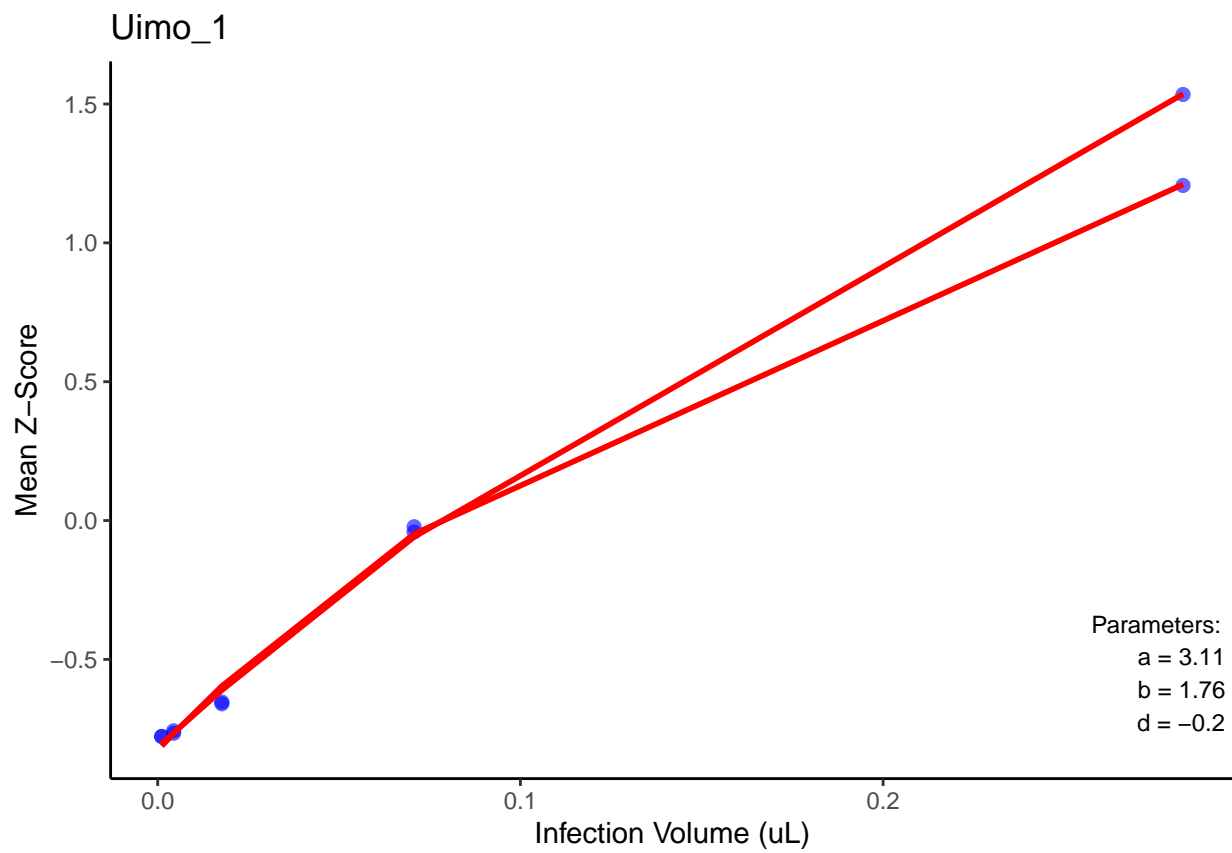
 ## \$Uaqe_1



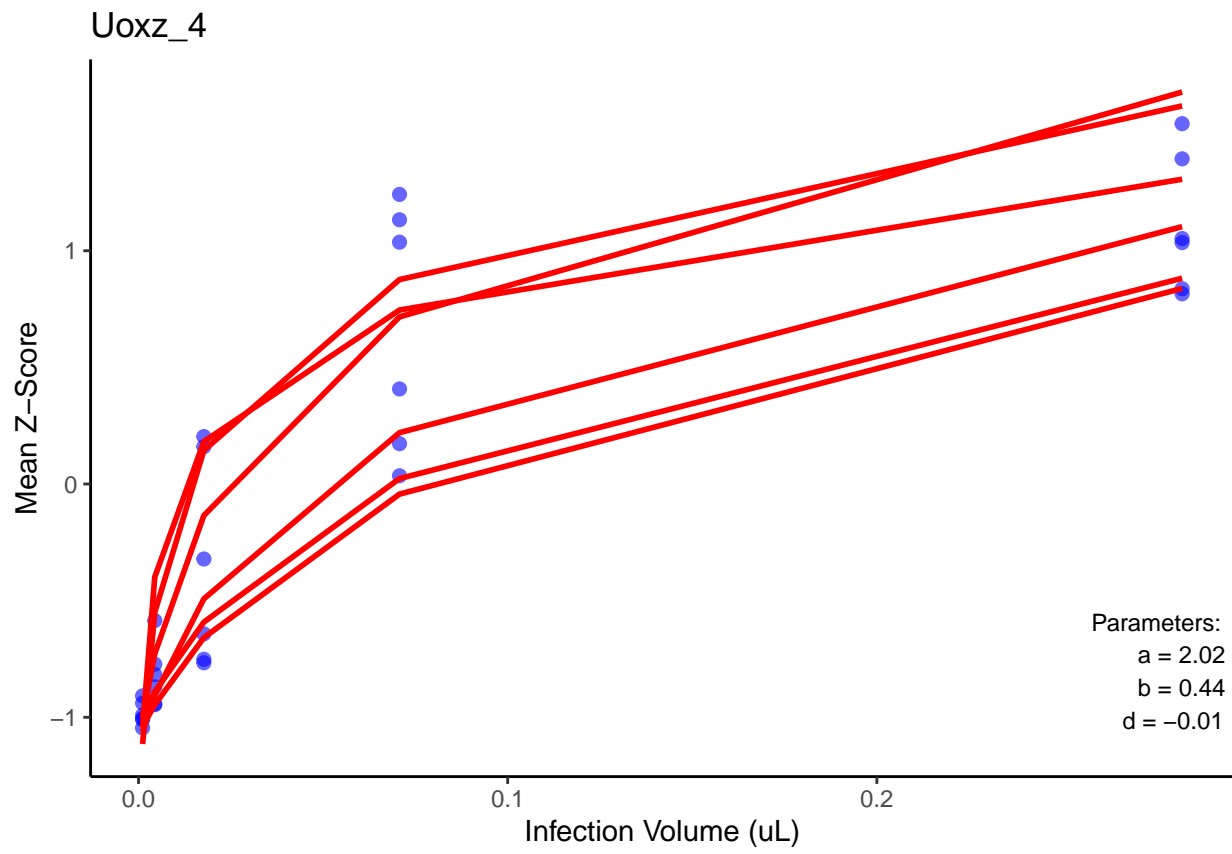
 ## \$Uevq_6



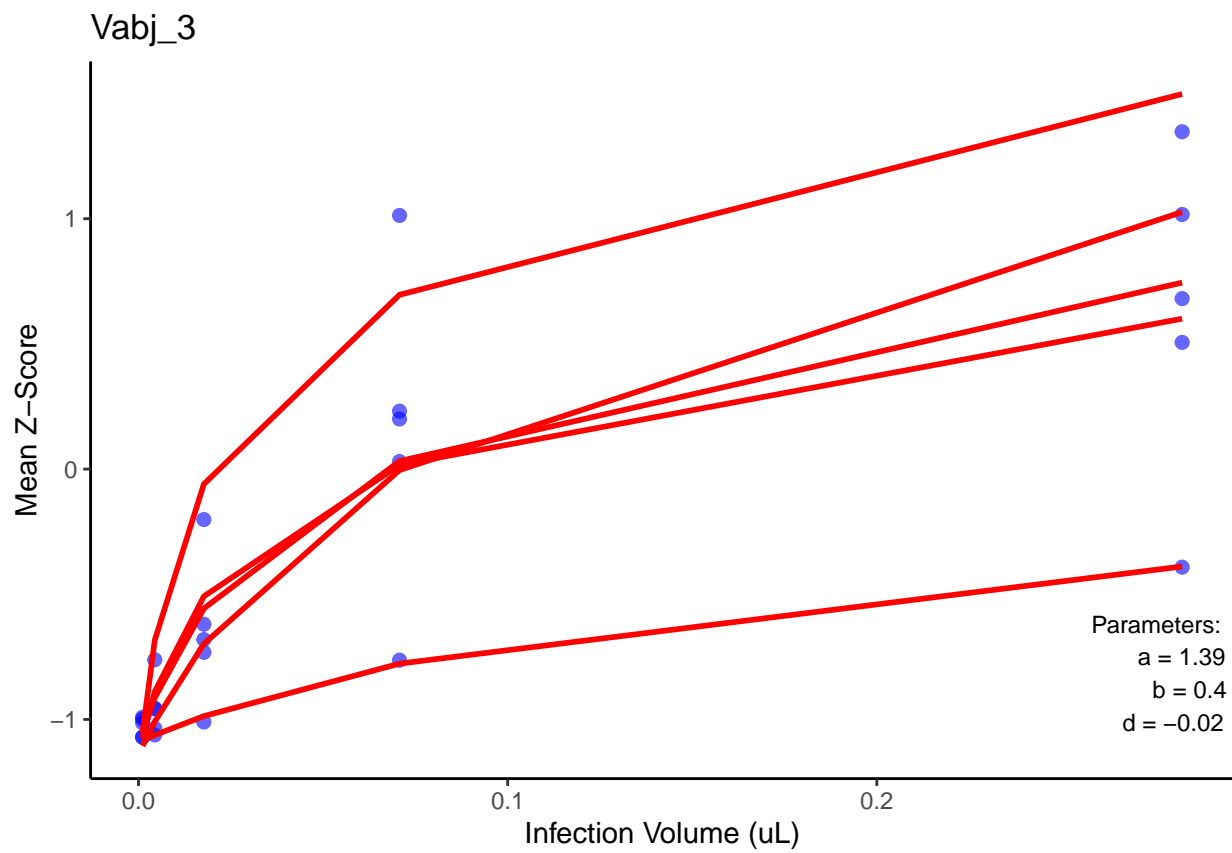
 ## \$Uimo_1



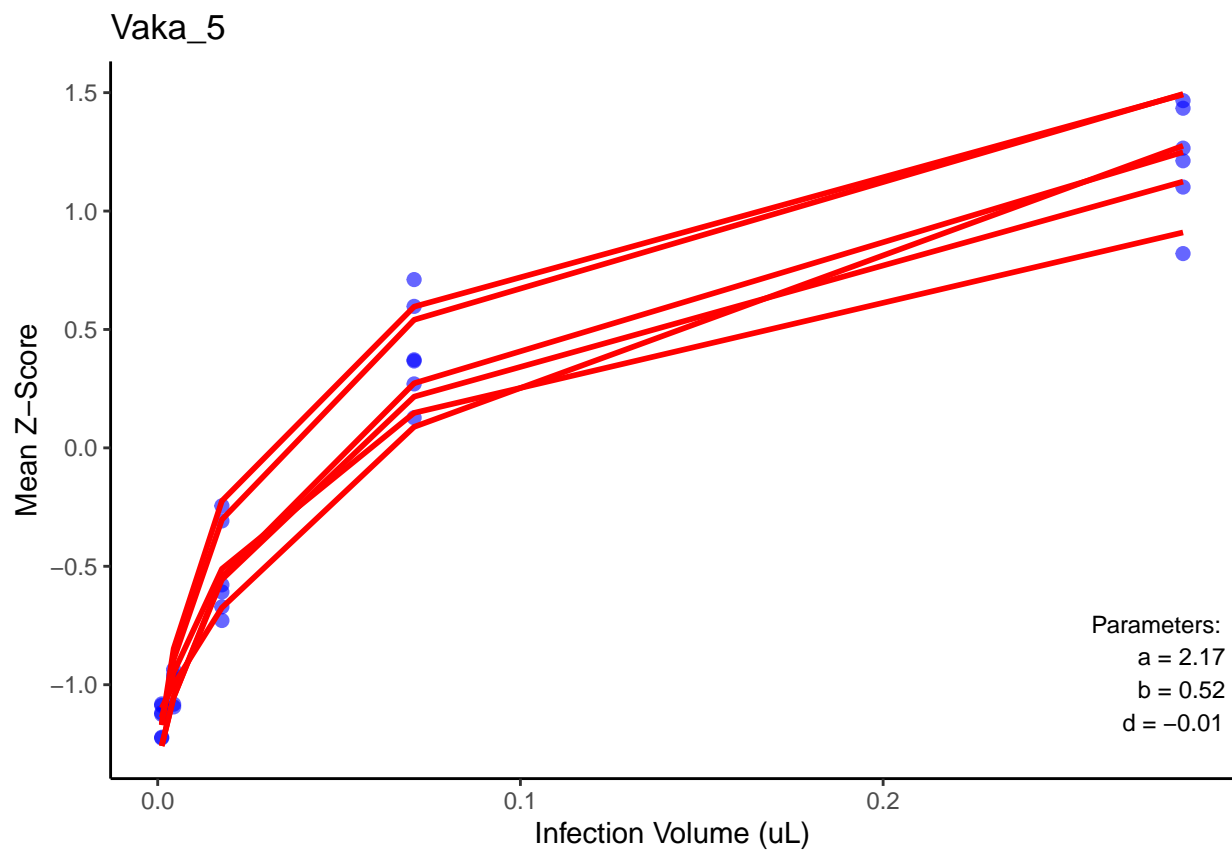
\$Uoxz_4



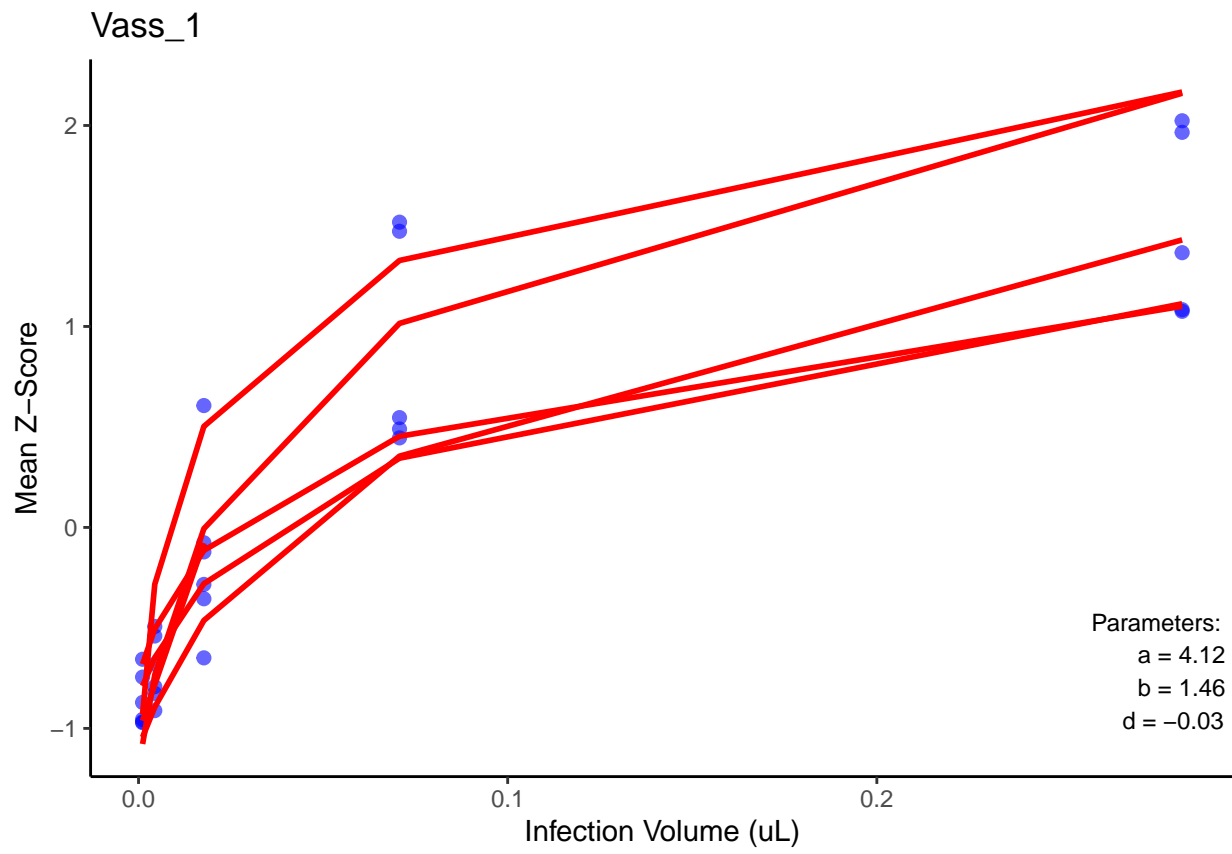
\$Vabj_3



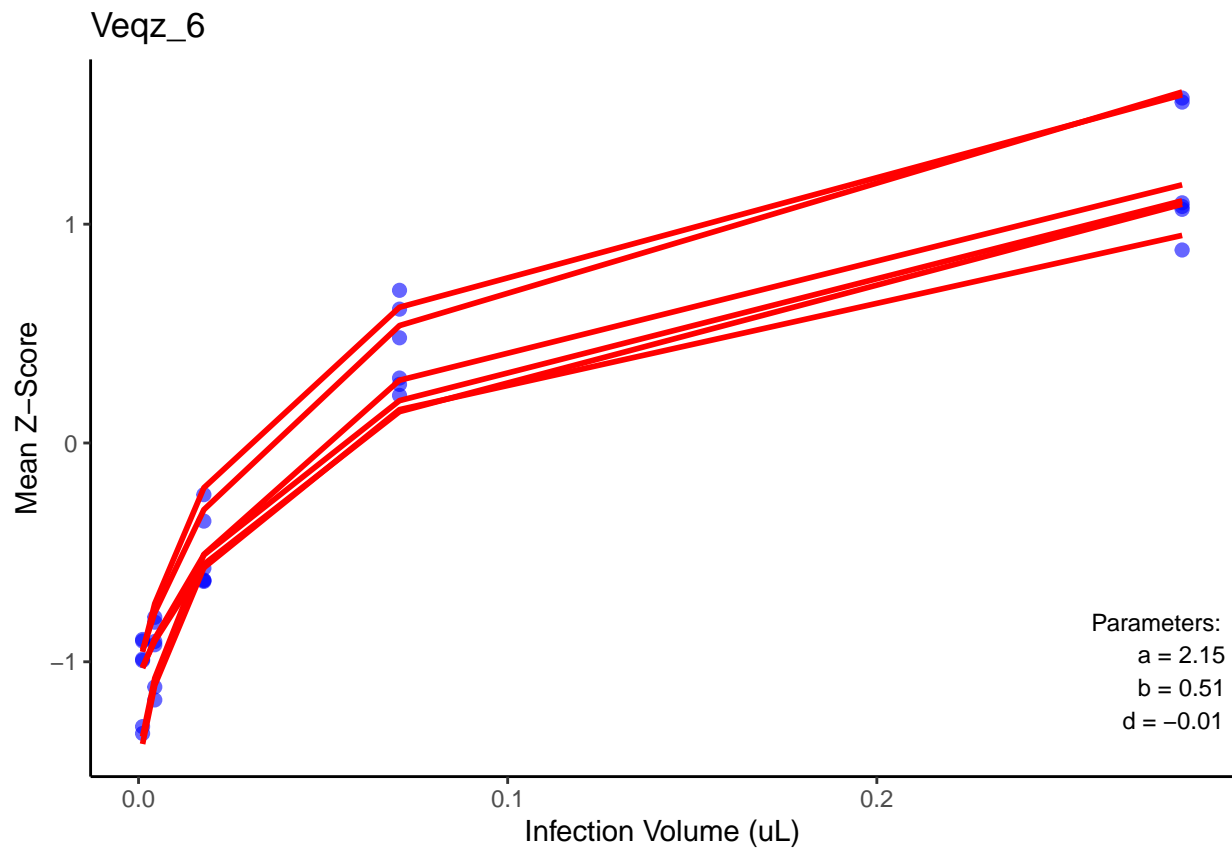
 ## \$Vaka_5



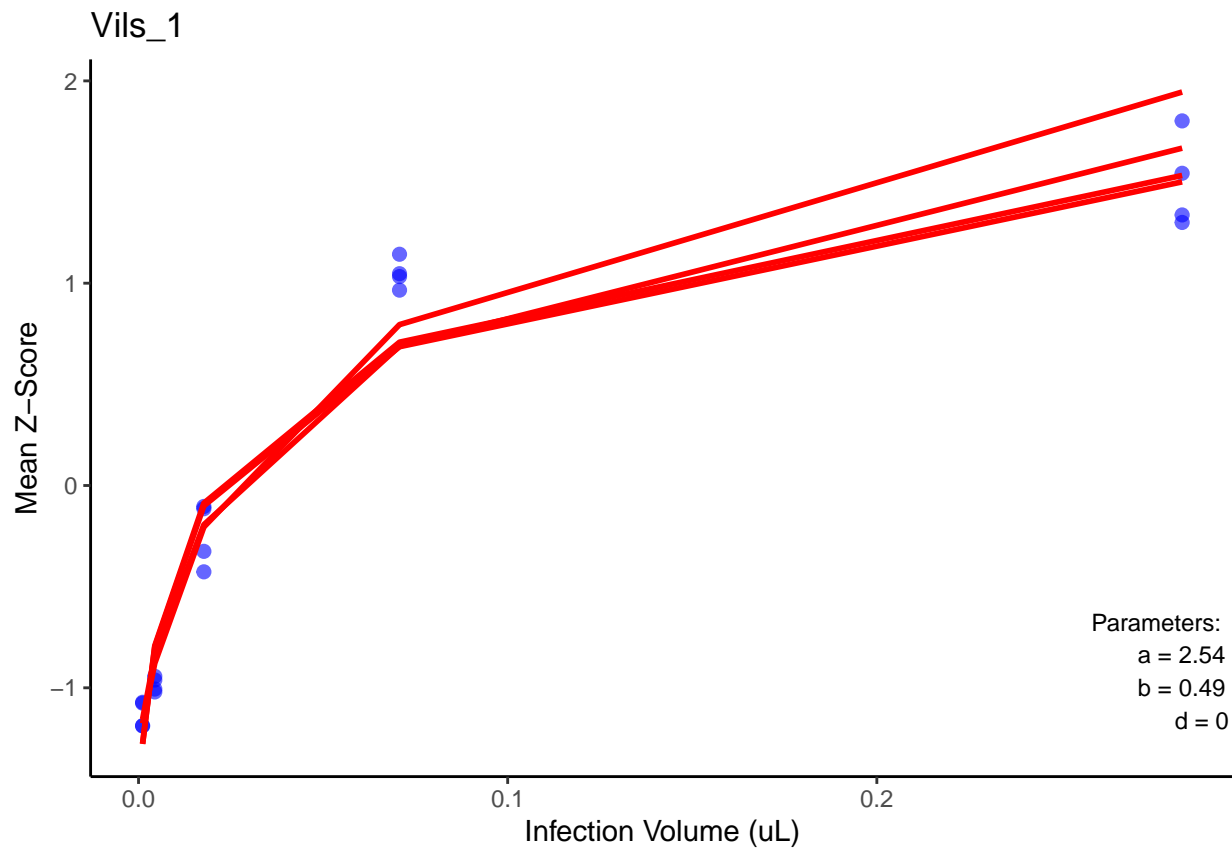
 ## \$Vass_1



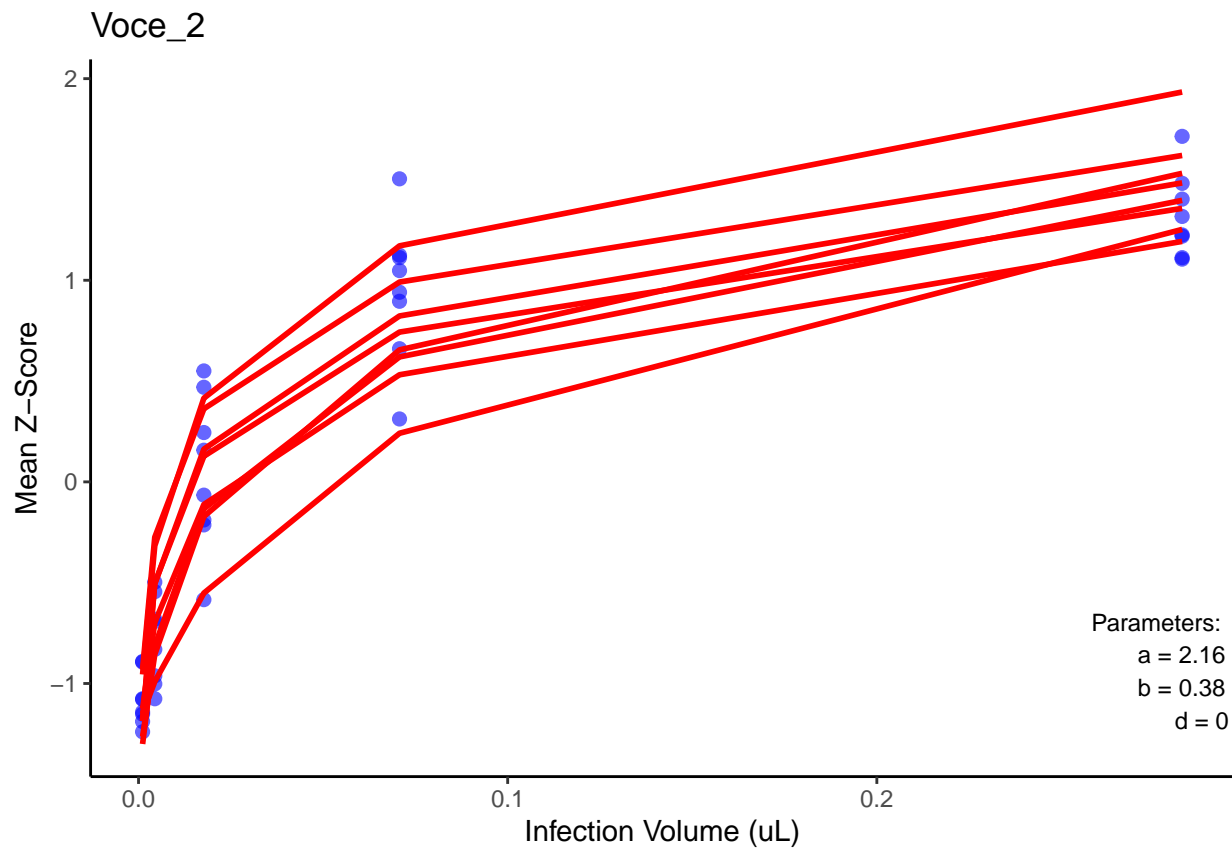
 ## \$Veqz_6



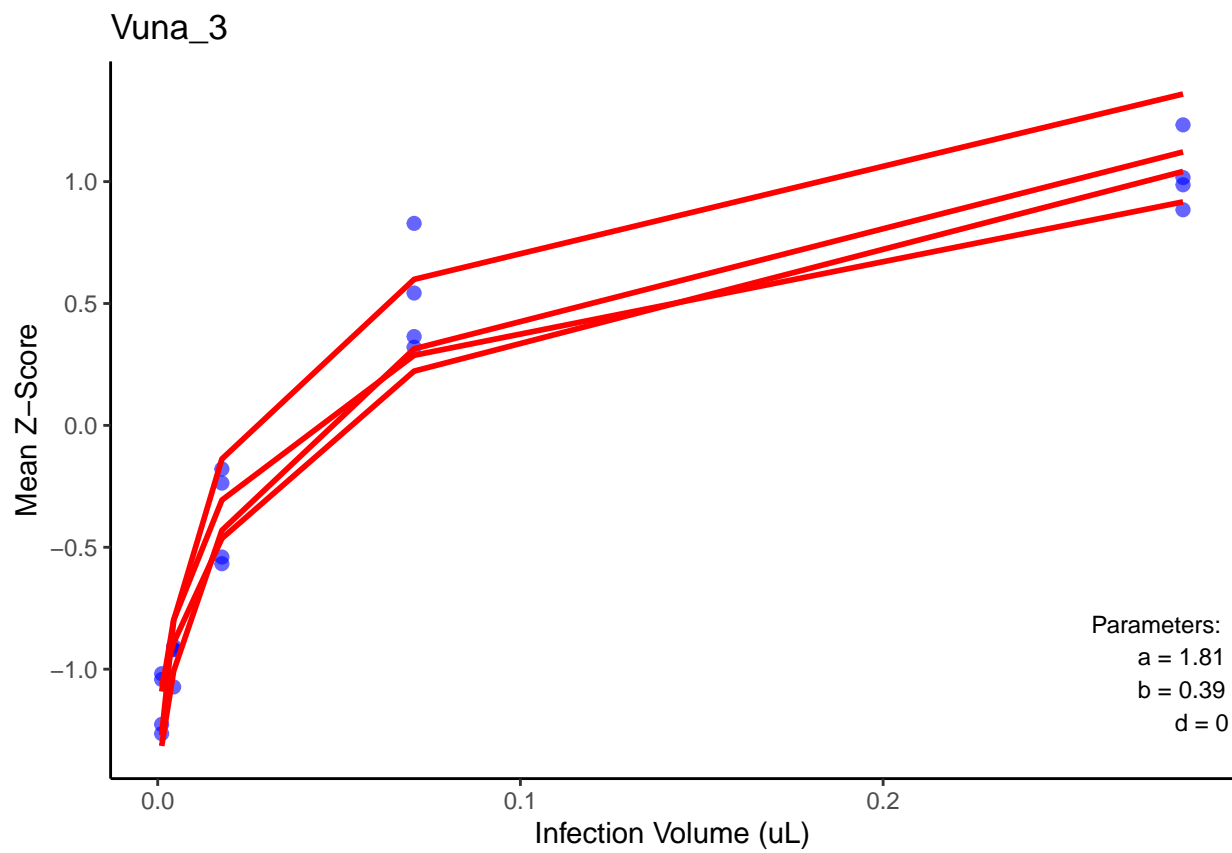
 ## \$Vils_1



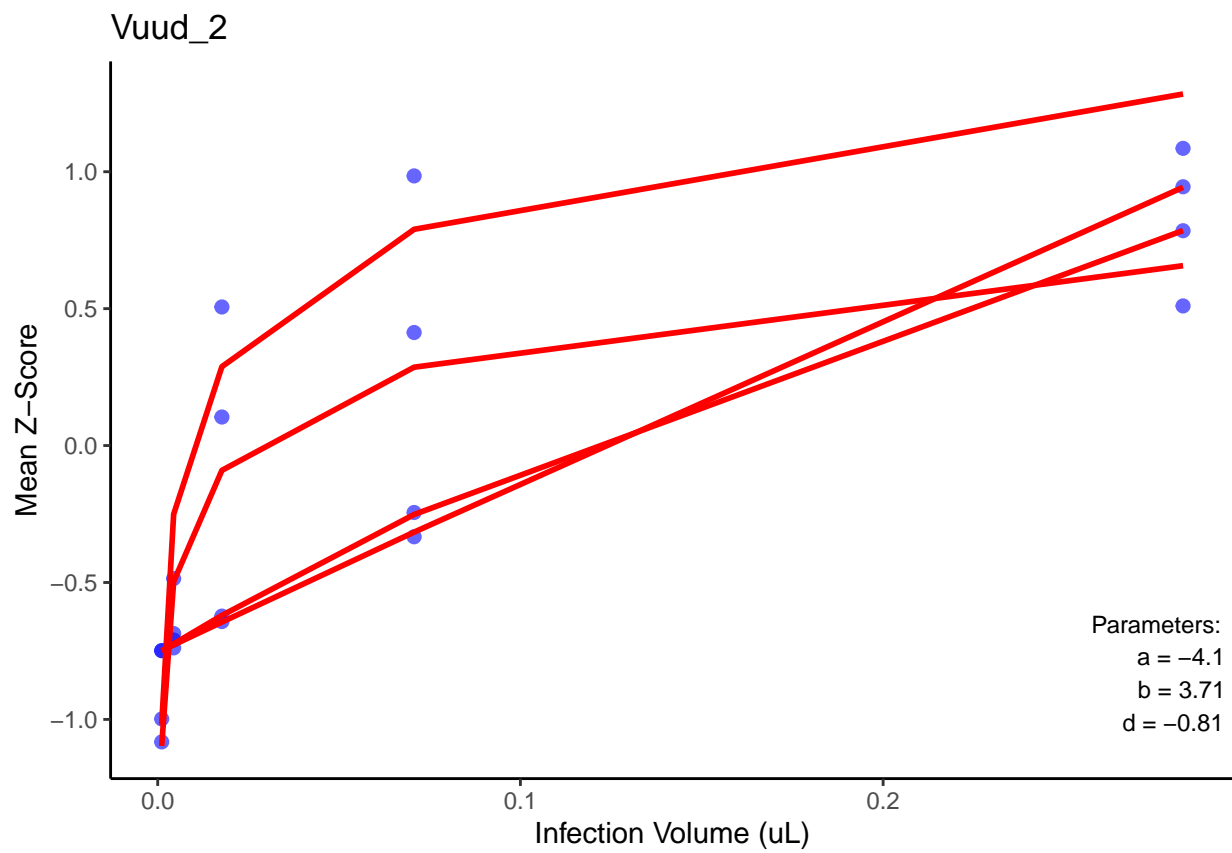
\$Voce_2



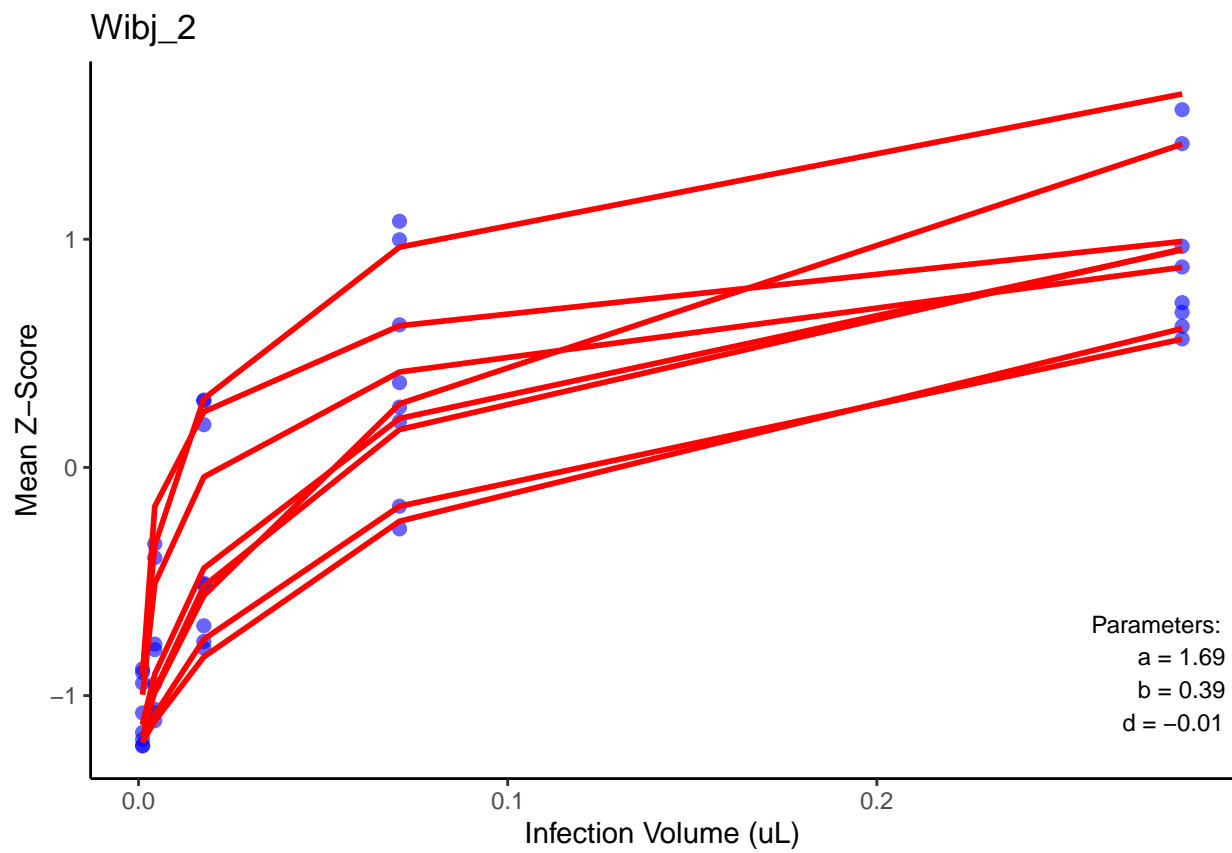
\$Vuna_3



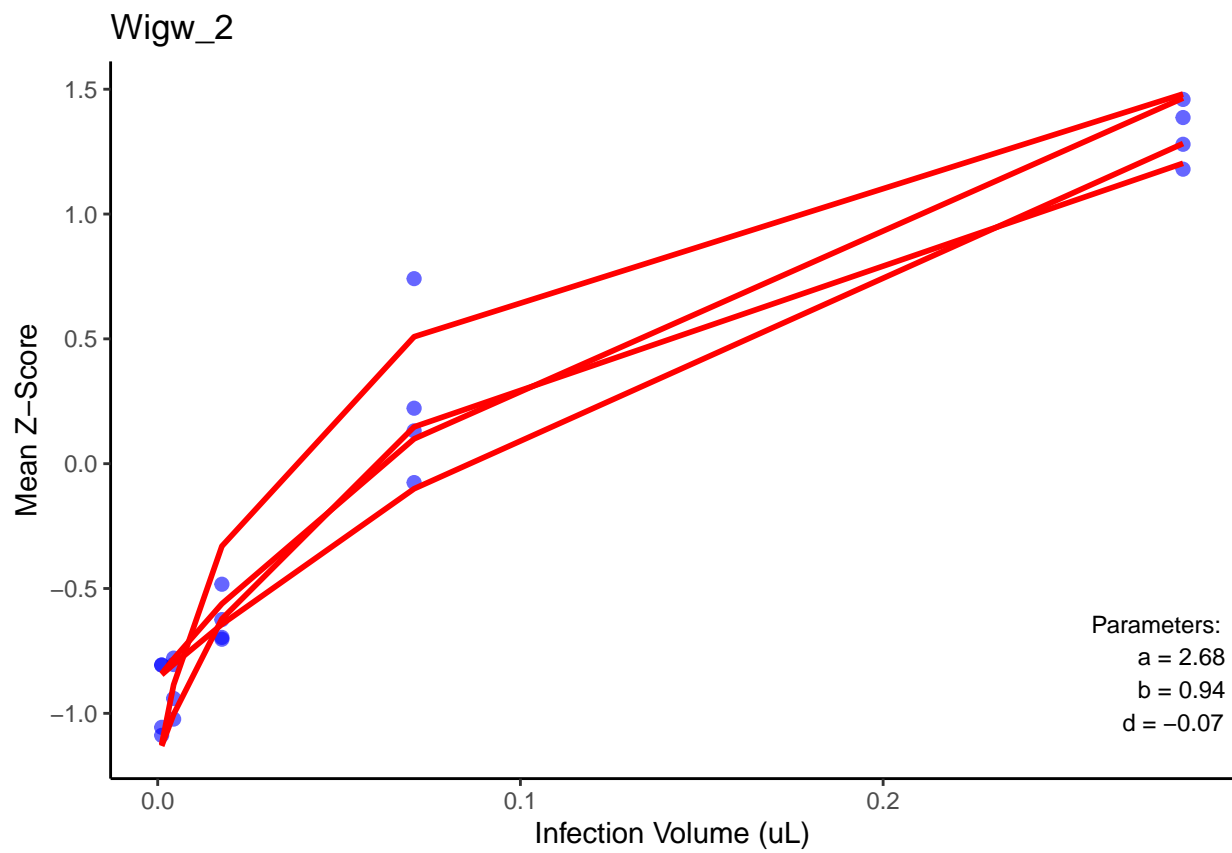
 ## \$Vuud_2



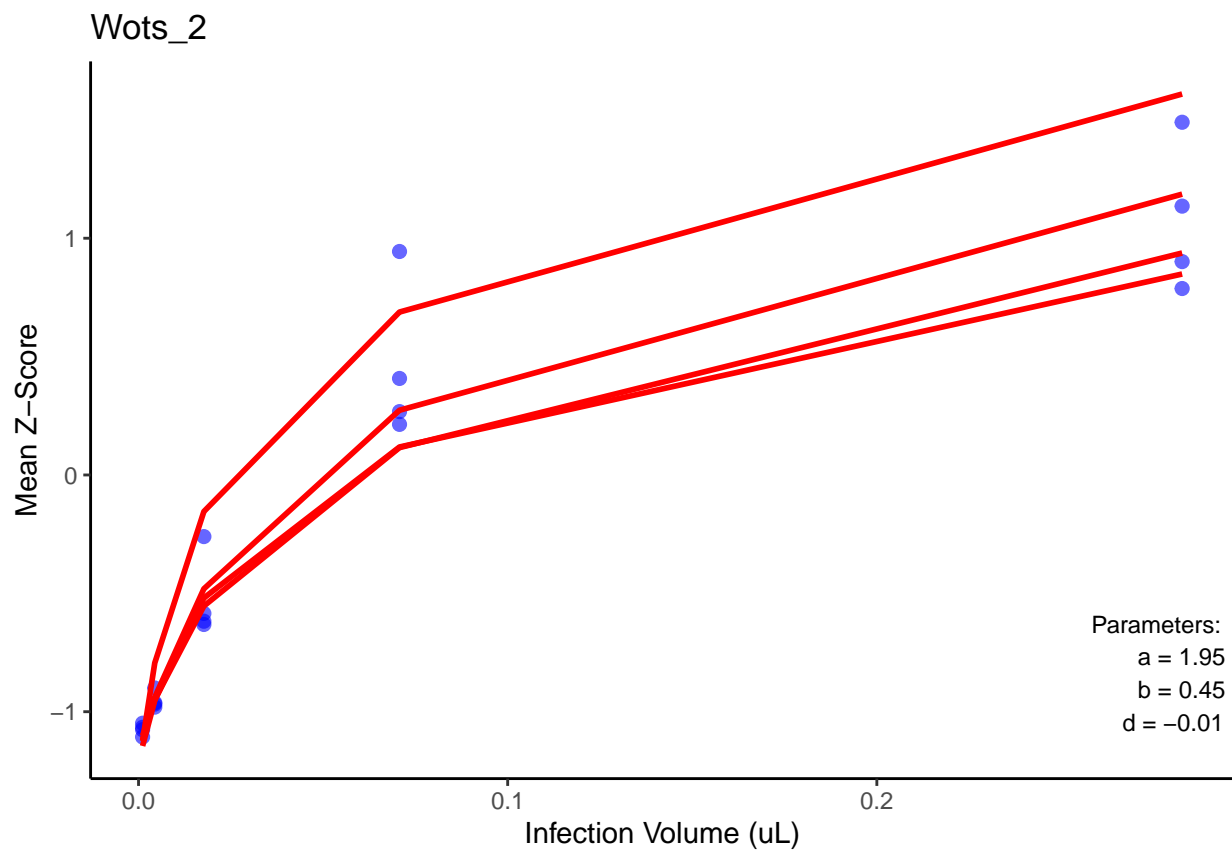
 ## \$Wibj_2



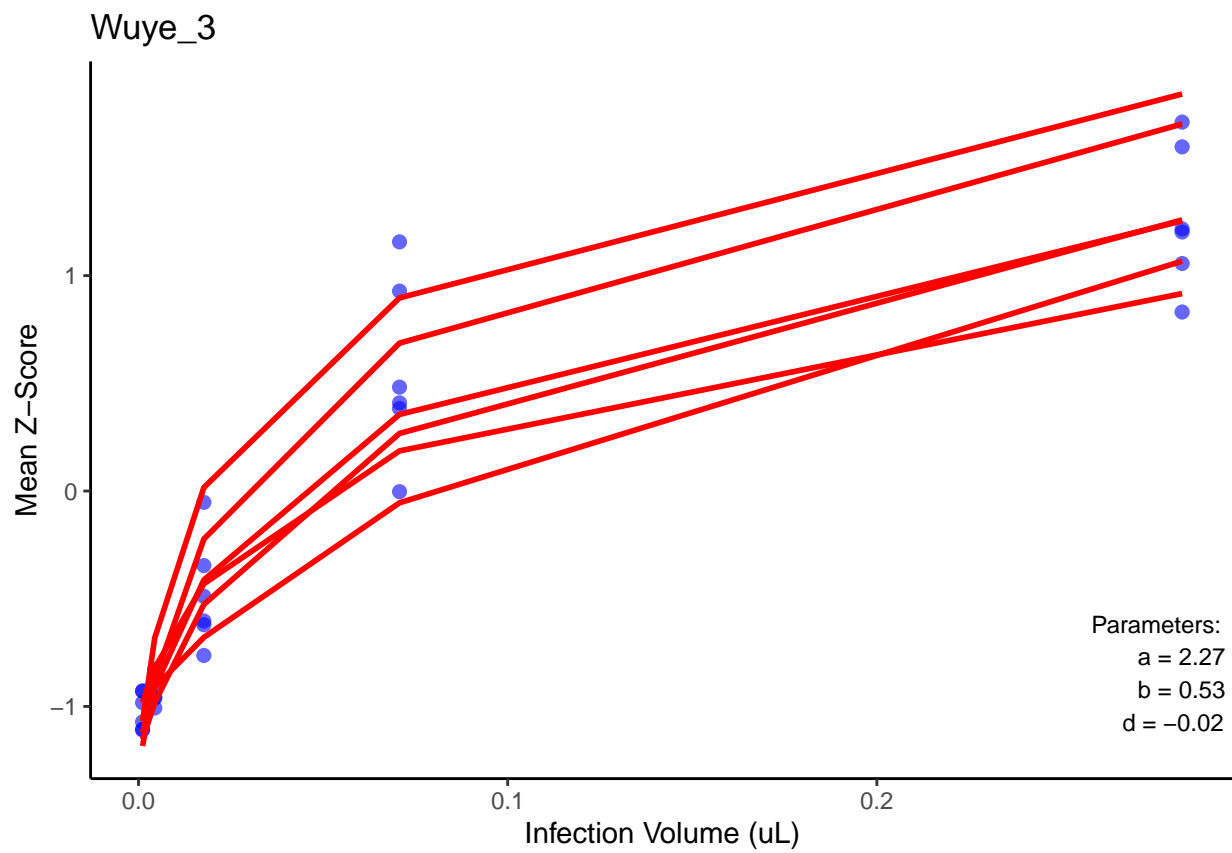
\$Wigw_2



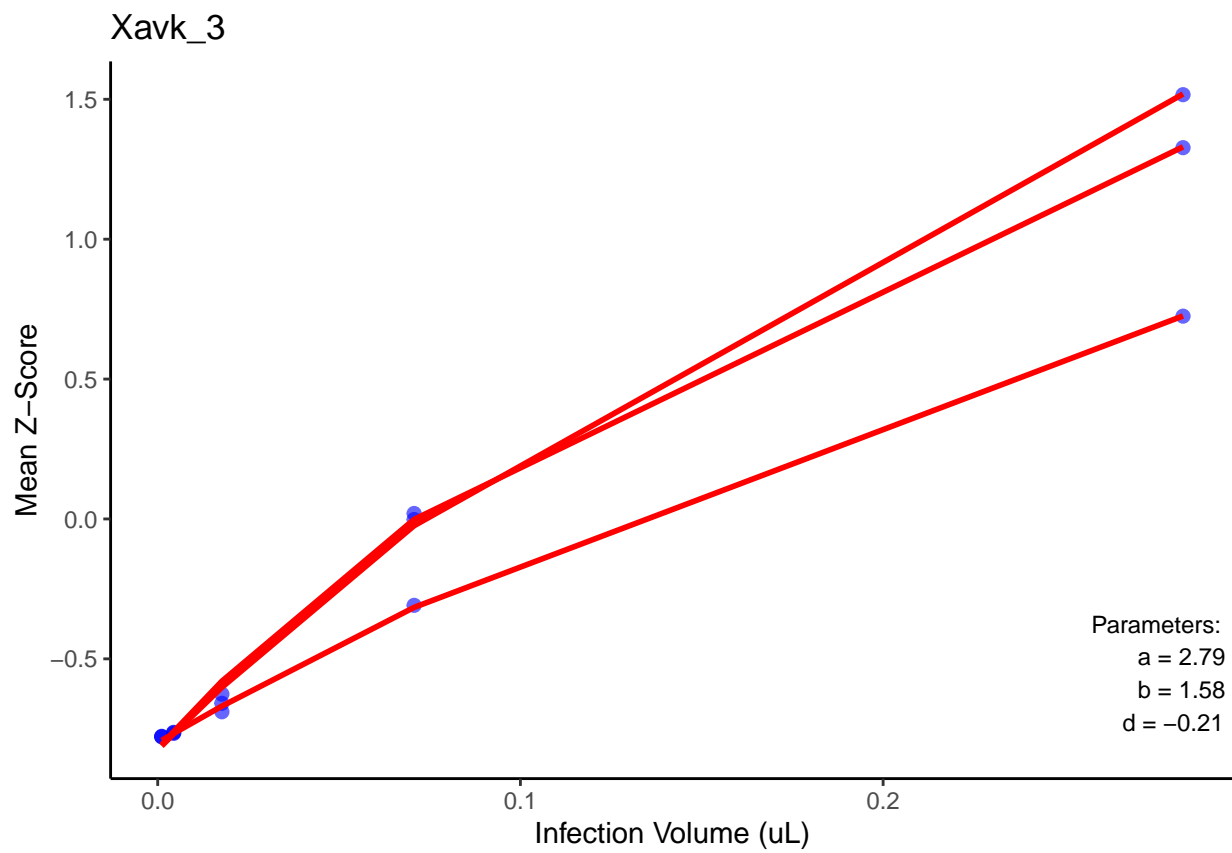
\$Wots_2



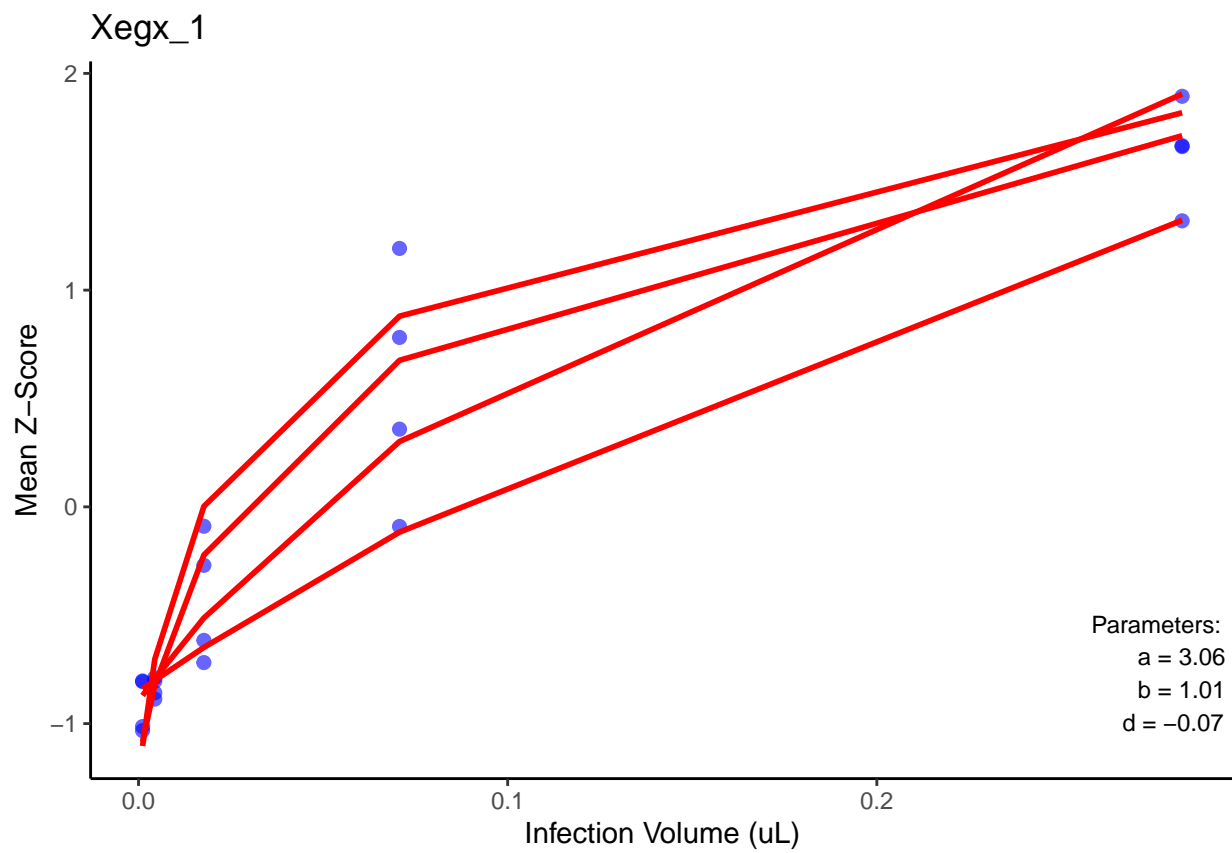
\$Wuye_3



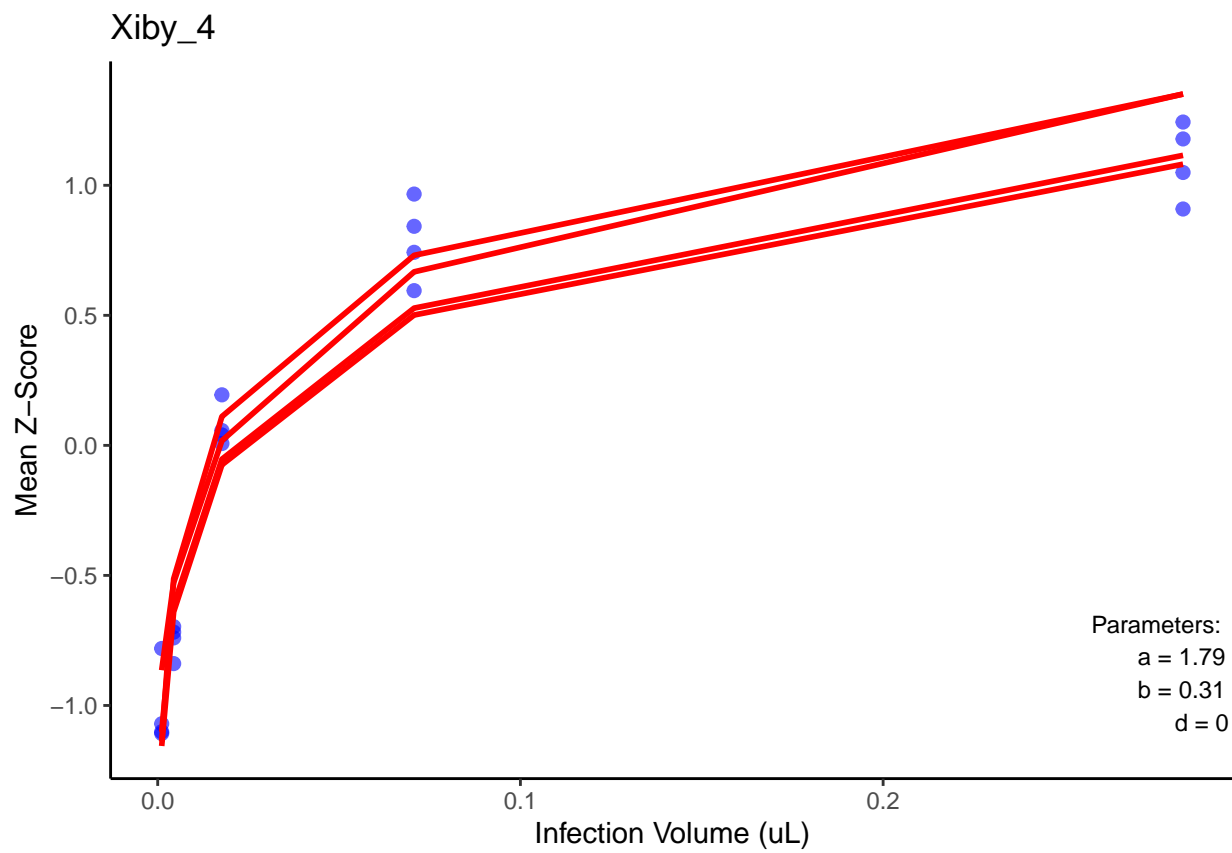
\$Xavk_3



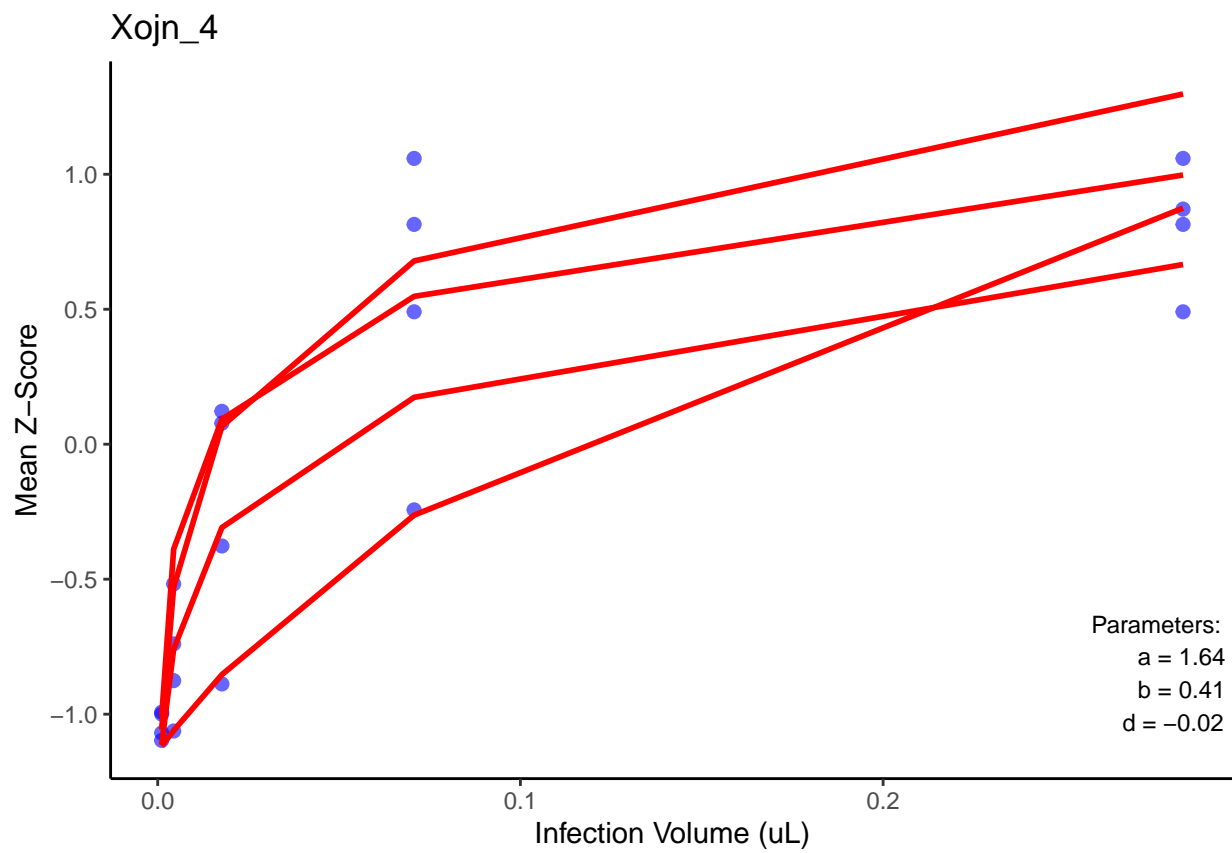
\$Xegx_1



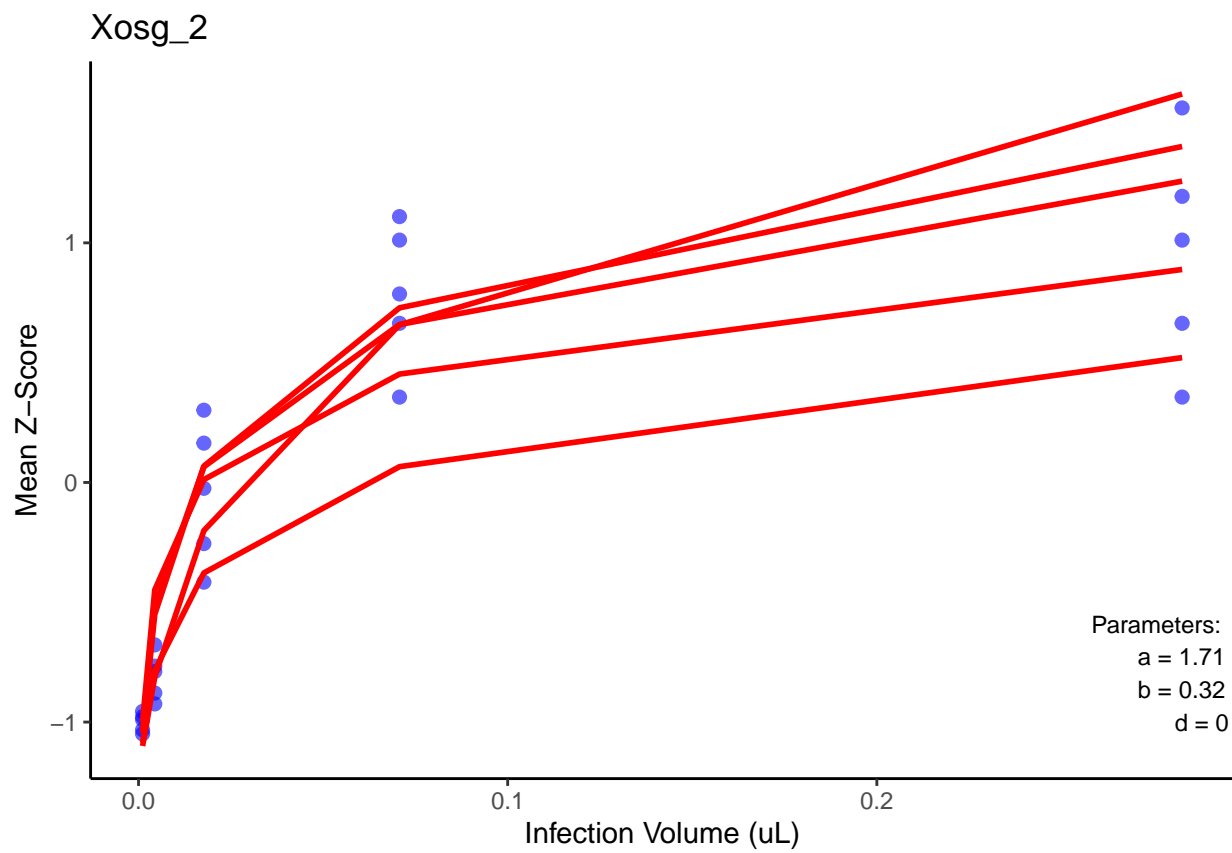
\$Xiby_4



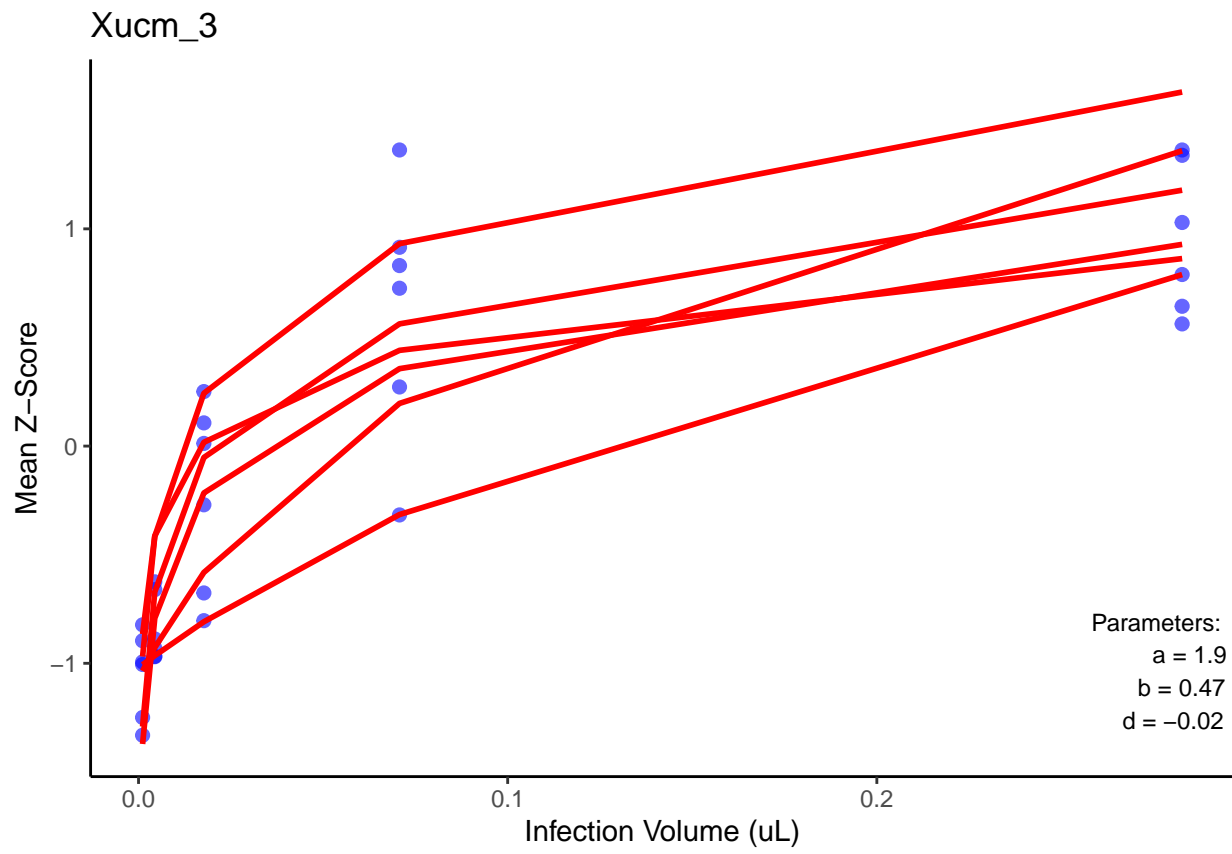
\$Xojn_4



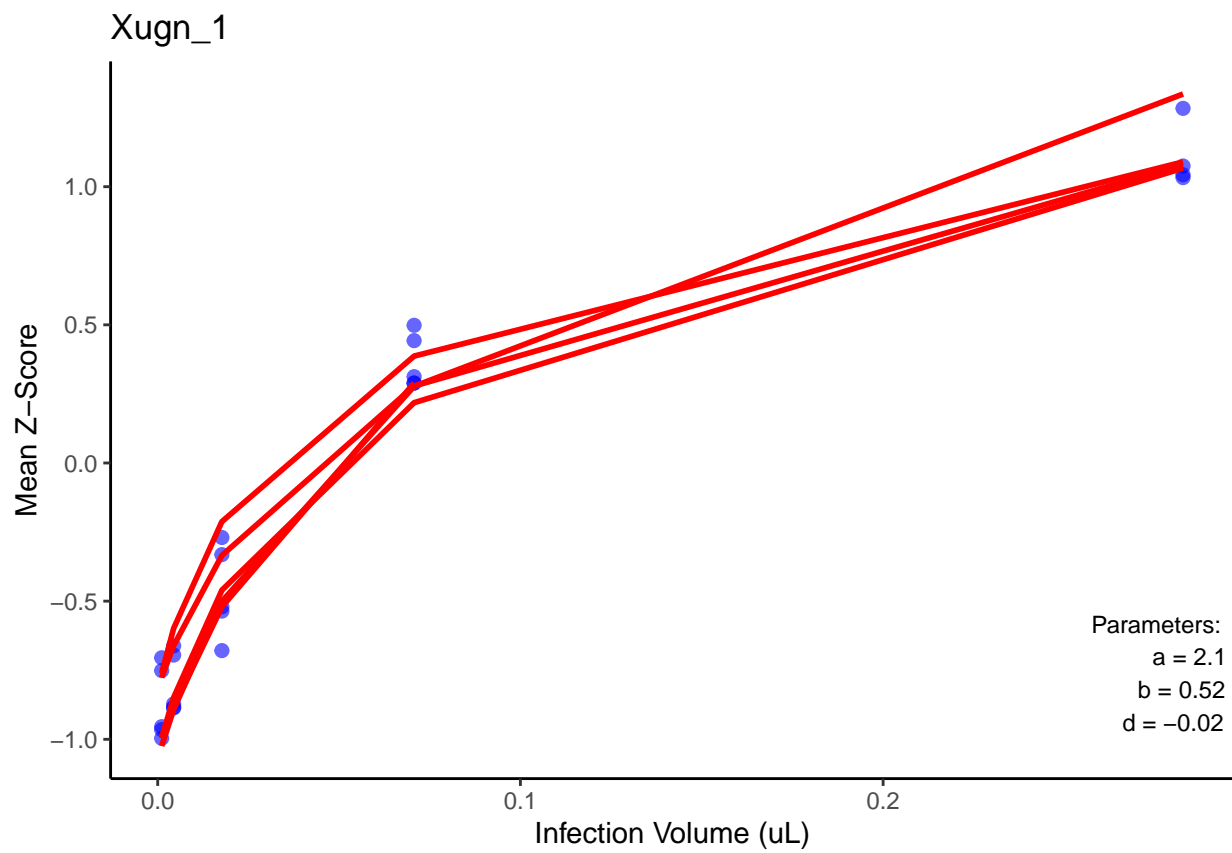
\$Xosg_2



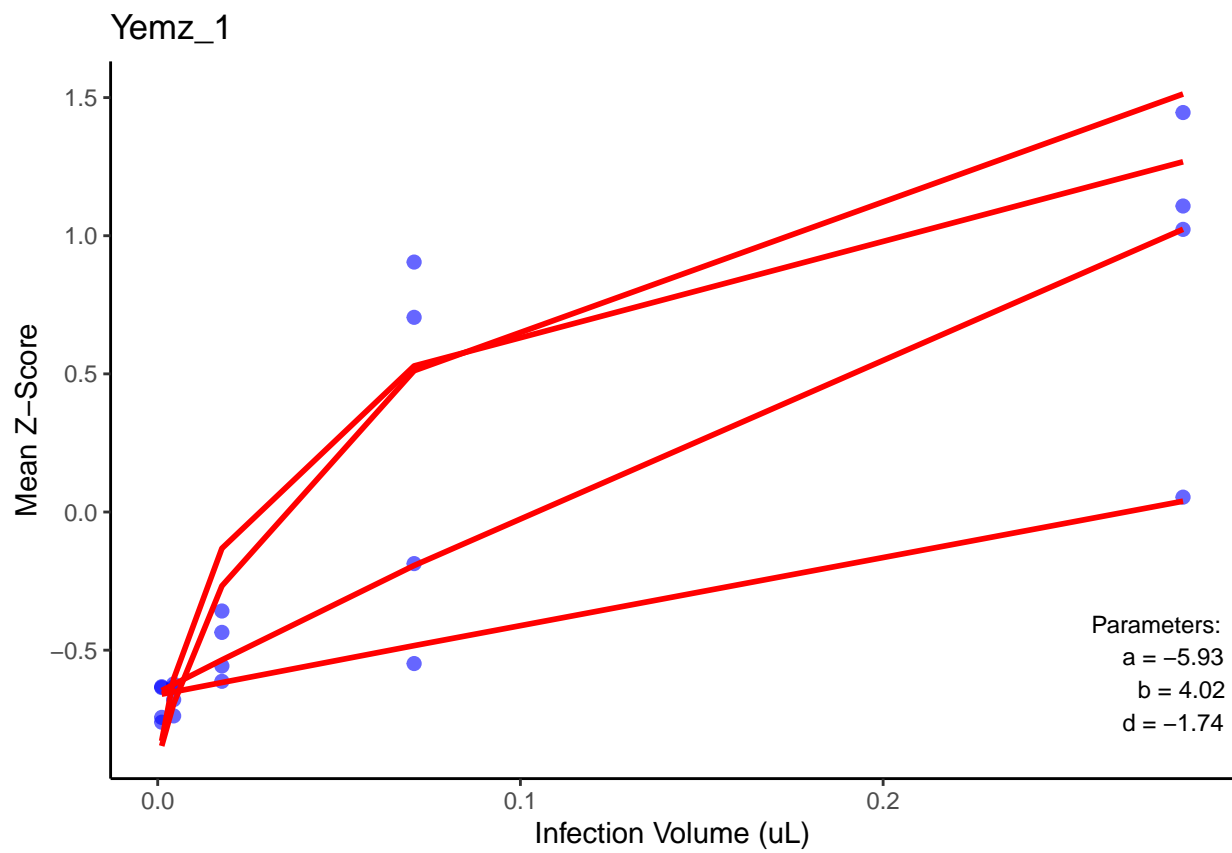
 ## \$Xucm_3



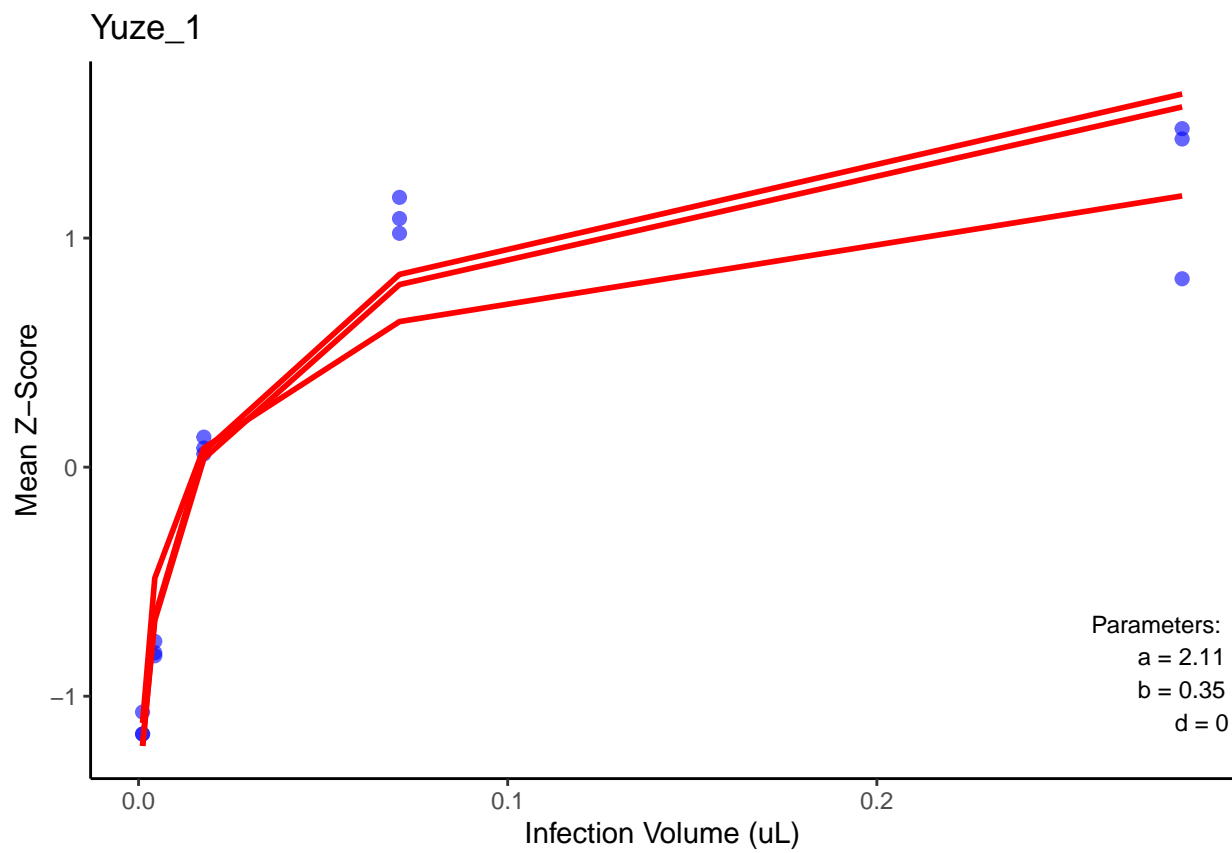
 ## \$Xugn_1



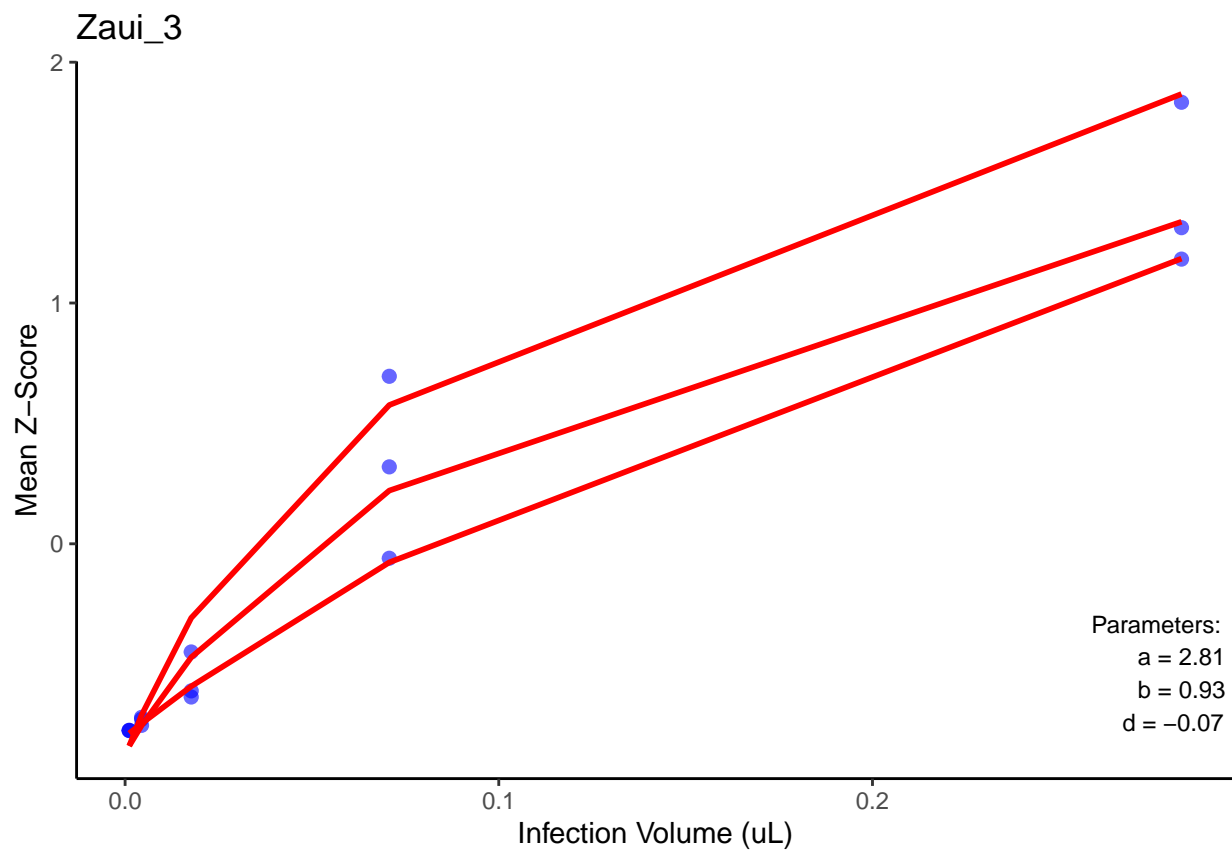
\$Yemz_1



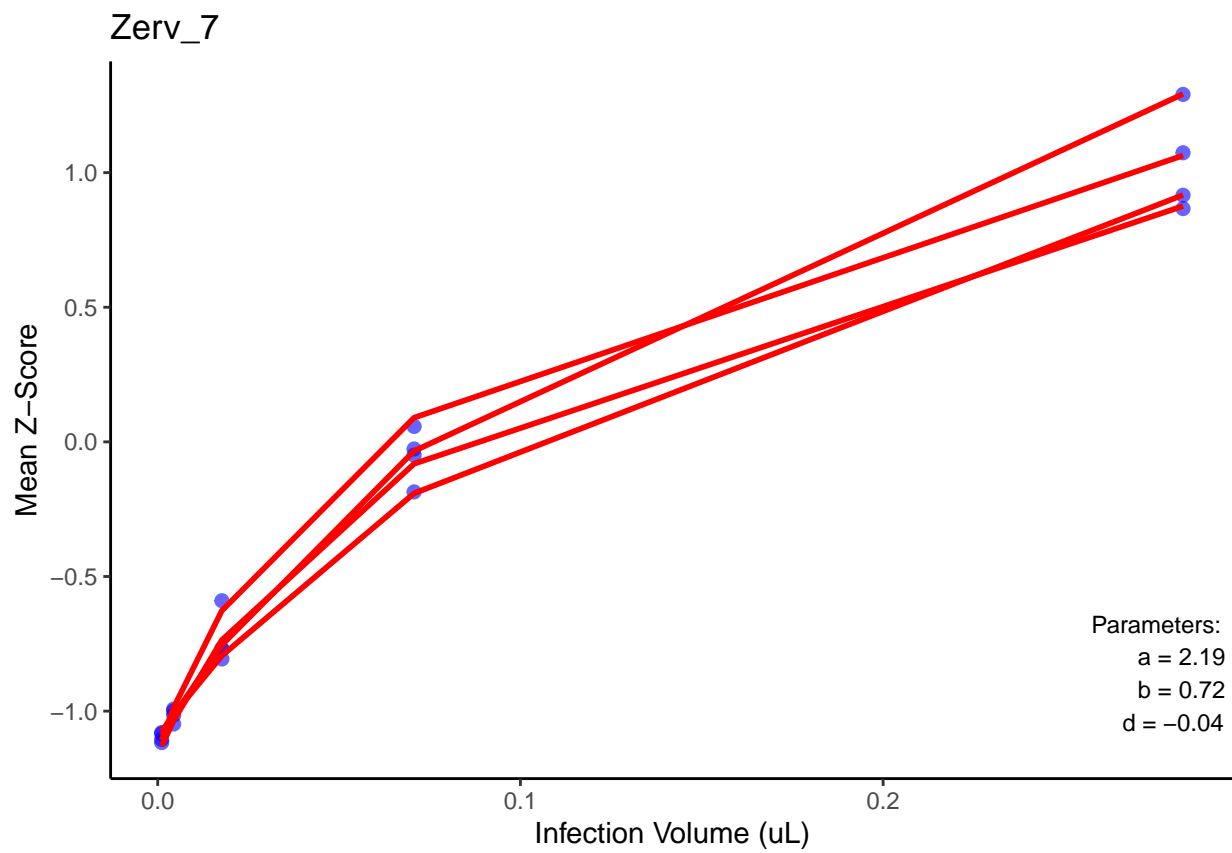
\$Yuze_1



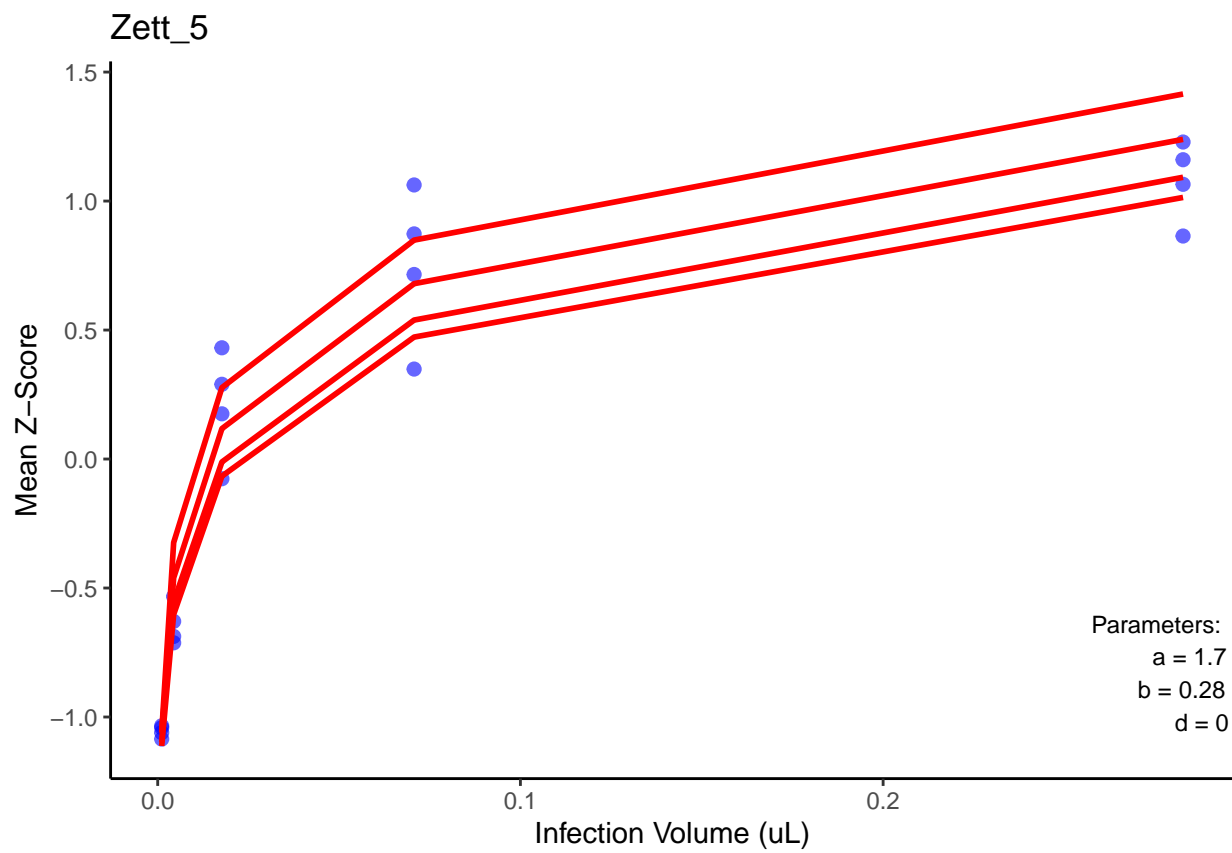
\$Zaui_3



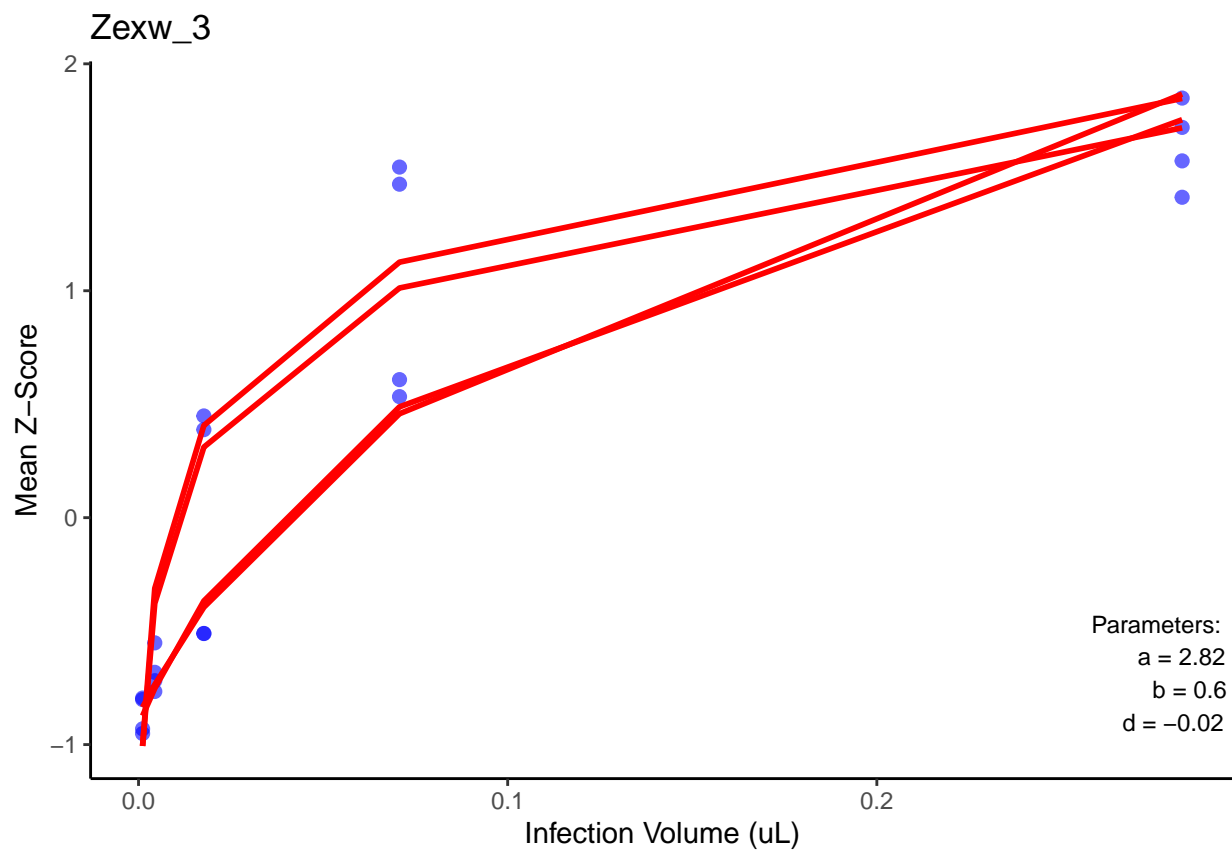
\$Zerv_7



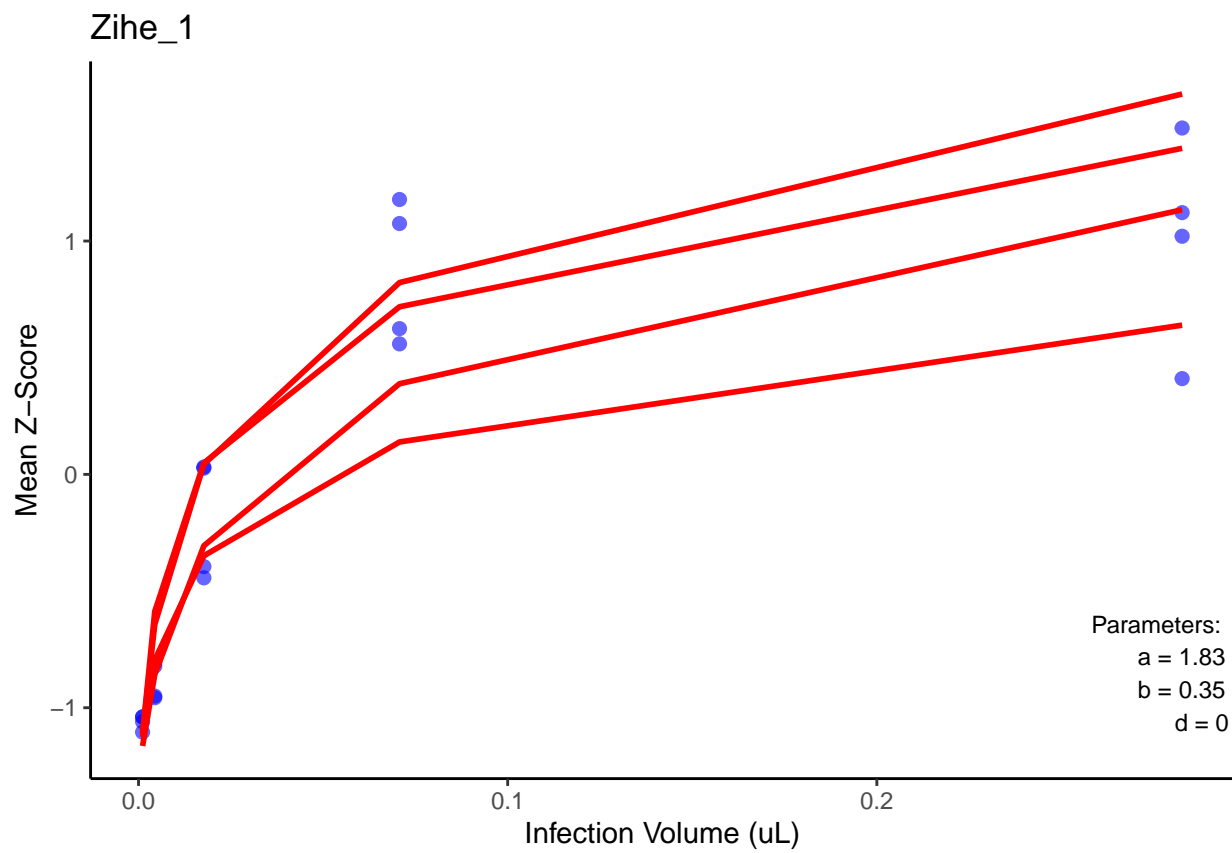
\$Zett_5



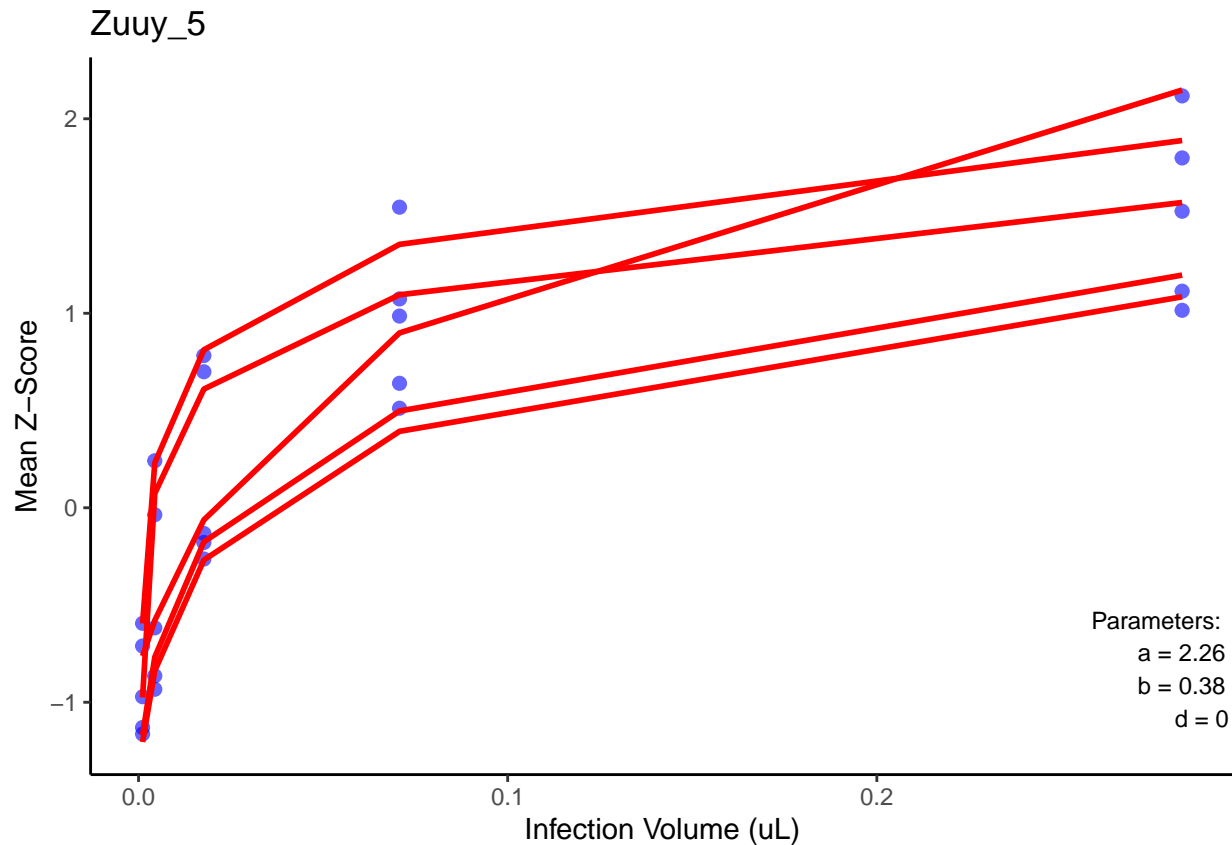
\$Zexw_3



\$Zihe_1



 ## \$Zuuy_5



logistic

Decided to use to use a 4 parameter logistic,

$$c + \frac{d - c}{1 + e^{b(x - e)}}$$

c - min,d-max, b-slope, e- x offset a higher c and d and b should correspond to an increase in permissiveness, an increase in e should correspond with lower permissiveness , however the fitting of the graph may not fit completely accurately, e.g parameters such as b and c may be extra large/small to allow a more accurate fit compared to what the values actually represent. So said parameters may have to be reevaluated later in the analysis

I use Drc (dose response package) as they provide a robust 4 parameter fitting with drm function, i then take the coefficients out of the model and store them, per replicate

```
# Correct 4PL formula (parentheses fix)
logistic_func <- function(x, b, d, e,c){
  c+ ((d-c)/(1 + exp(b*(x - e))))# Fixed denominator placement
}

# Robust fitting function
logistic_fit <- function(data) {
  fitted_data <- data %>%
    group_by(screen_nb,replicate) %>% # Verify these columns exist in your data
    group_modify(~ {
      tryCatch({ # Fit with drm, explicitly stating NO log transformation
        model <- drm(zscore ~ infection_volume_ul, fct = L.4(), data = .x)
        coefs <- coef(model)
      }
```



```

      .x %>%
        mutate(
          logis_b = coefs["b:(Intercept)"],
          logis_e = coefs["e:(Intercept)"],
          logis_c = coefs["c:(Intercept)"],
          logis_d = coefs["d:(Intercept)"]
        )
    }, error = function(e) {
      warning(paste("drc error for screen_nb:", .x$screen_nb[1], ":", e$message))
      .x %>%
        mutate(logis_b = NA, logis_e = NA, logis_d = NA)
    })
  }) %>%
  ungroup()
  return(fitted_data)
}

# Apply to cell line data
vector_data_per_cell_line <- purrr::map(vector_data_per_cell_line$data, logistic_fit) %>% bind_rows() %>% ne

```

similar to plot logarithmic, i create predicted then use ggplot to plot the line over the actual points to assess the fit, and use the parameters to see if the values i get for each parameter make biological sense

```

apply_plot_logistic <- function(i) {
  # Generate predictions using row-wise parameters
  plot_data <- i %>%
    mutate(
      predicted = logistic_func(
        x = infection_volume_ul,
        b = logis_b,
        c = logis_c,
        d = logis_d,
        e = logis_e
      )
    )

  # Create base plot
  p <- ggplot(plot_data, aes(x = infection_volume_ul, y = zscore)) +
    # Actual data points
    geom_point(
      aes(color = factor(screen_nb)), # Color by screen_nb
      size = 2,
      alpha = 0.6,
      show.legend = T
    ) +
    # Fitted curves
    geom_line(
      aes(y = predicted, group = interaction(screen_nb, replicate)),
      color = "red",
      linewidth = 1
    ) +
    # Labels and theme
    labs(
      title = unique(plot_data$cell_line),

```

```

    x = "Infection Volume (microL)",
    y = "Mean Z-Score"
  ) +
  theme_classic(base_size = 12)

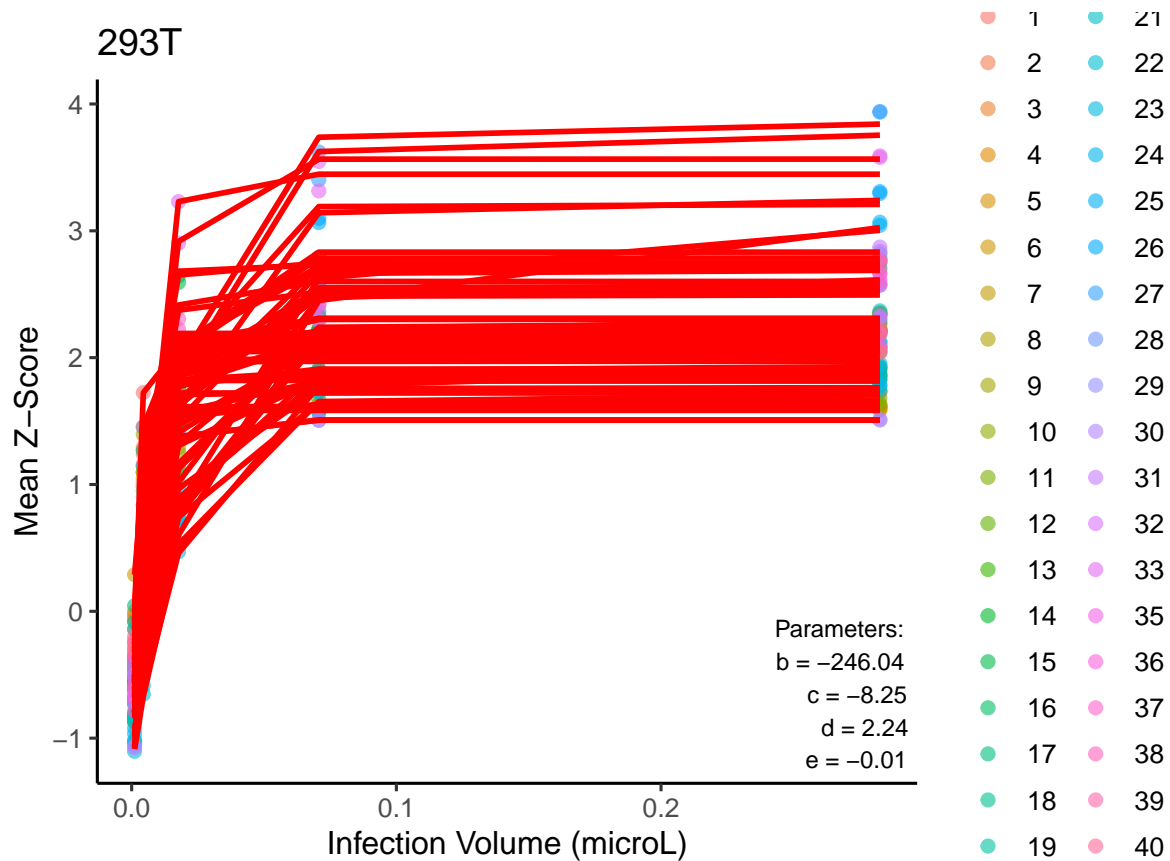
# Add regression parameters annotation
params_text <- paste0(
  "Parameters:\n",
  "b = ", round(mean(plot_data$logis_b, na.rm = TRUE), 2), "\n",
  "c = ", round(mean(plot_data$logis_c, na.rm = TRUE), 2), "\n",
  "d = ", round(mean(plot_data$logis_d, na.rm = TRUE), 2), "\n",
  "e = ", round(mean(plot_data$logis_e, na.rm = TRUE), 2)
)

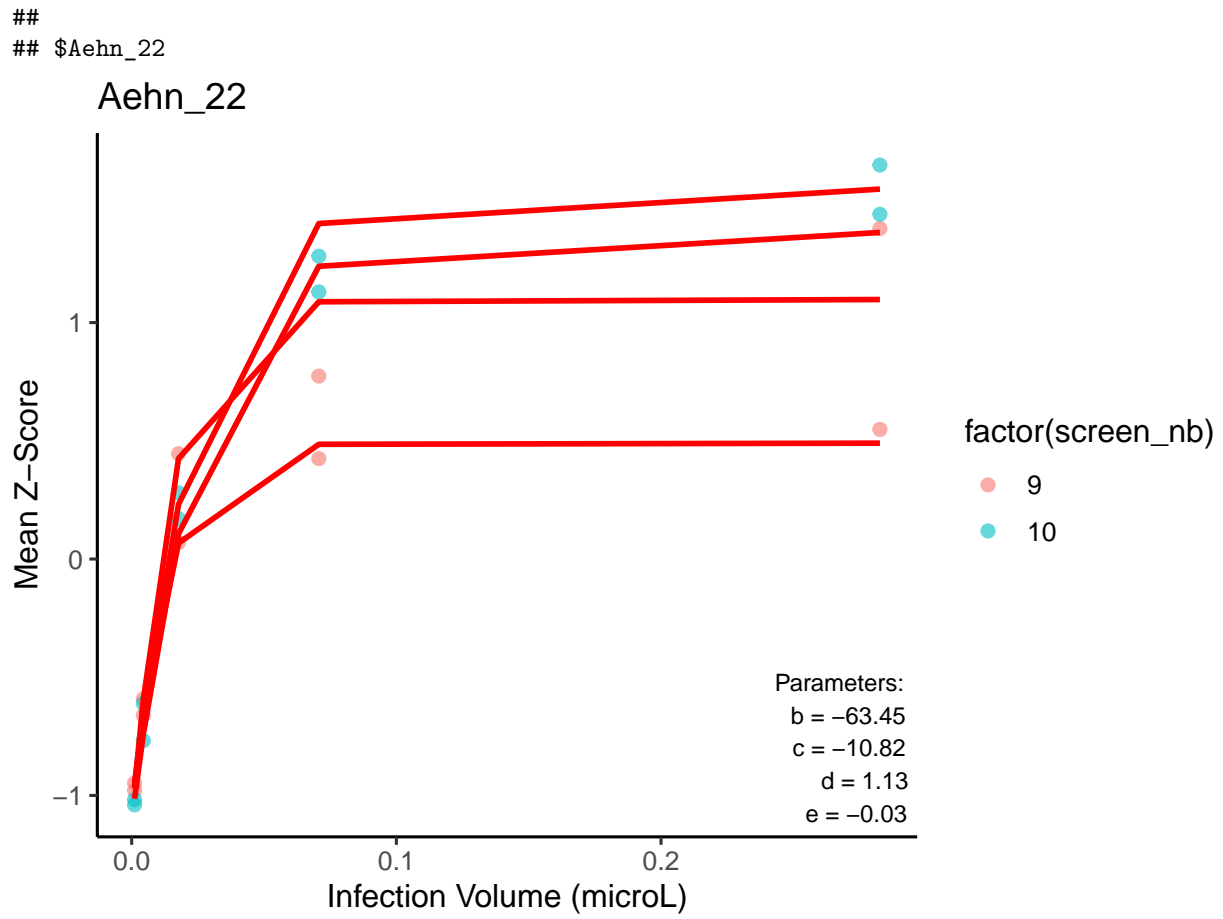
p + annotate("text",
  x = Inf, y = -Inf,
  label = params_text,
  hjust = 1.1, vjust = -0.1,
  size = 3)
}

logistic_plot_vector=purrr::map(vector_data_per_cell_line$data,apply_plot_logistic)
names(logistic_plot_vector)=vector_data_per_cell_line$cell_line
logistic_plot_vector

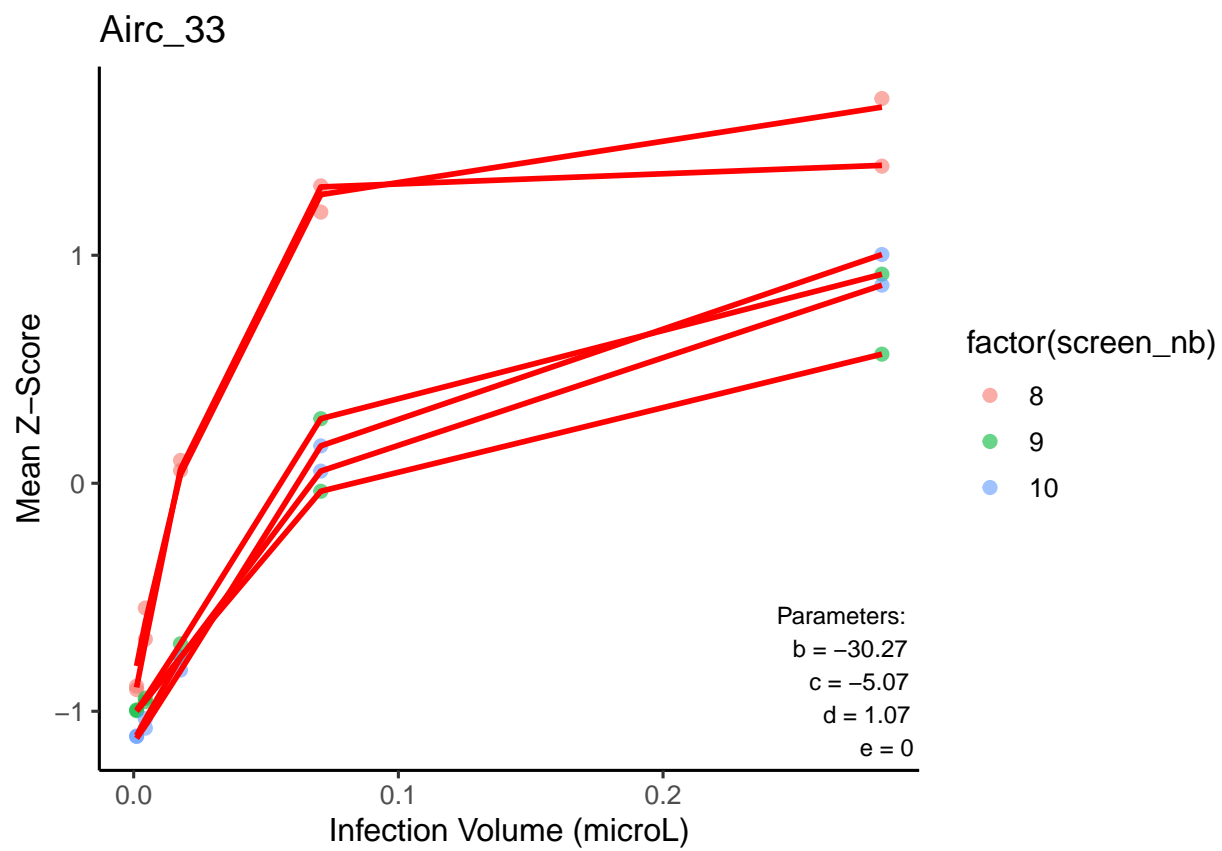
```

\$`293T`

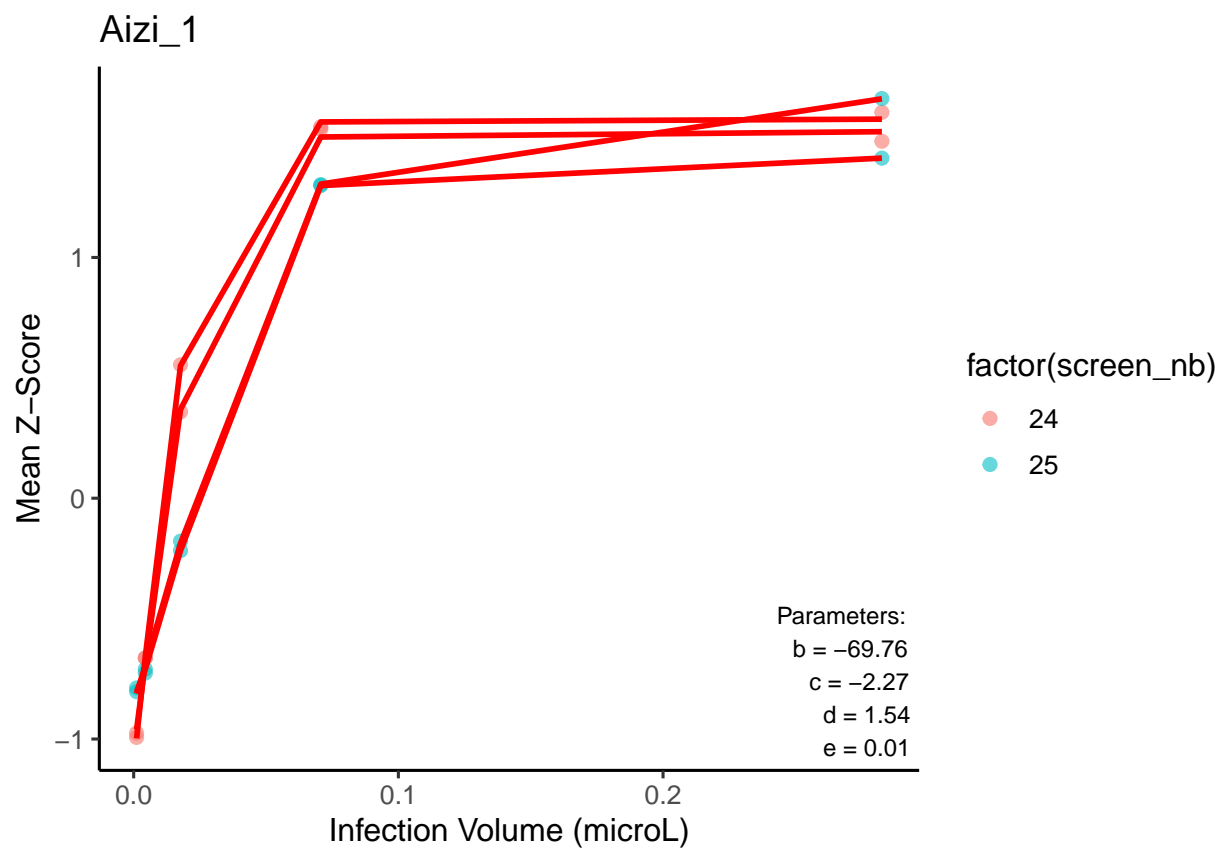




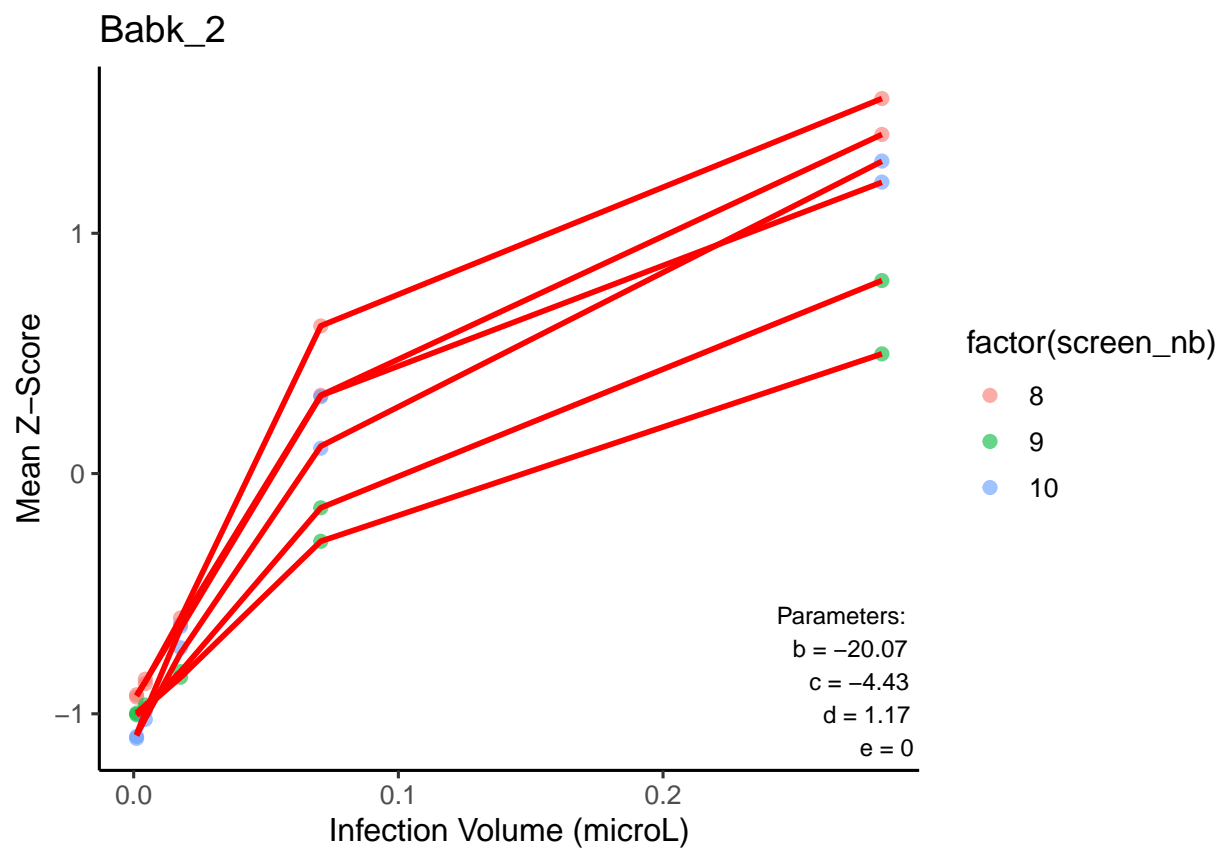
```
##
## $Airc_33
```



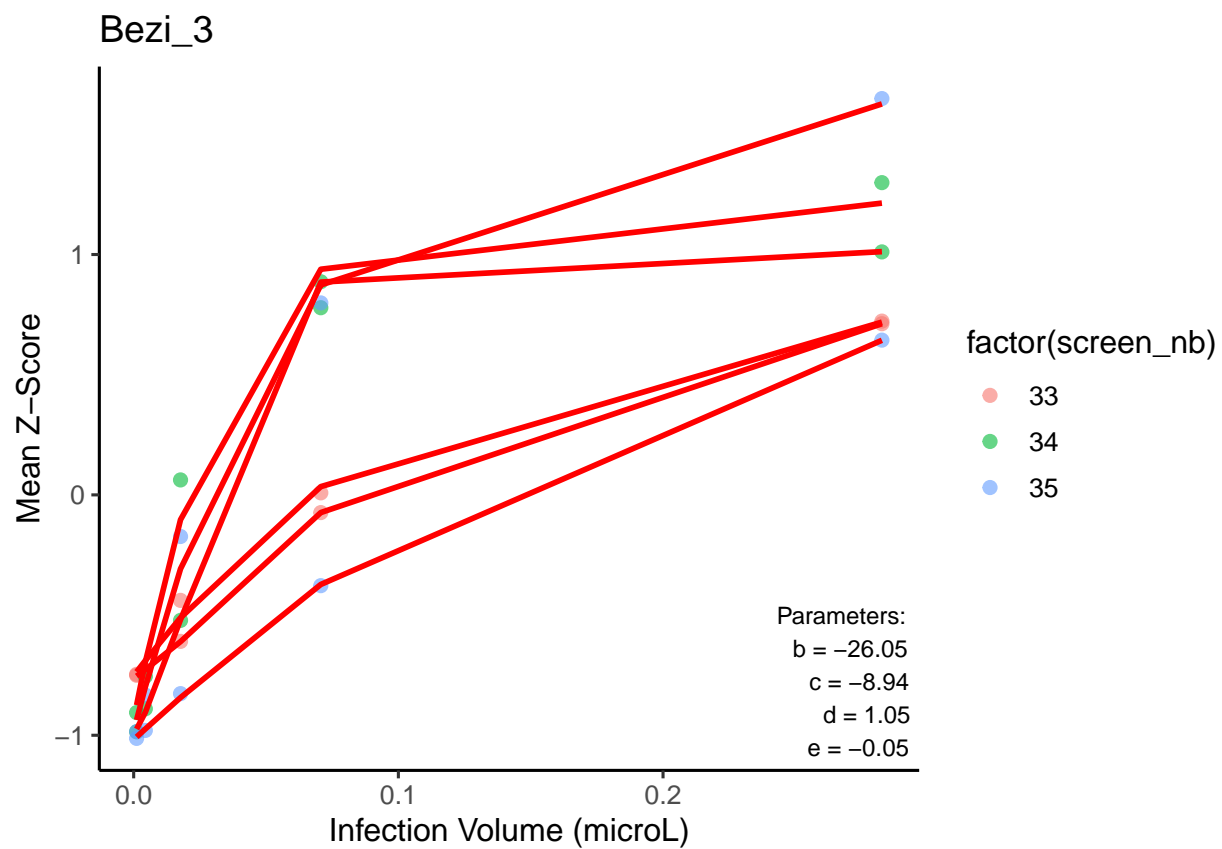
```
##  
## $Aizi_1
```



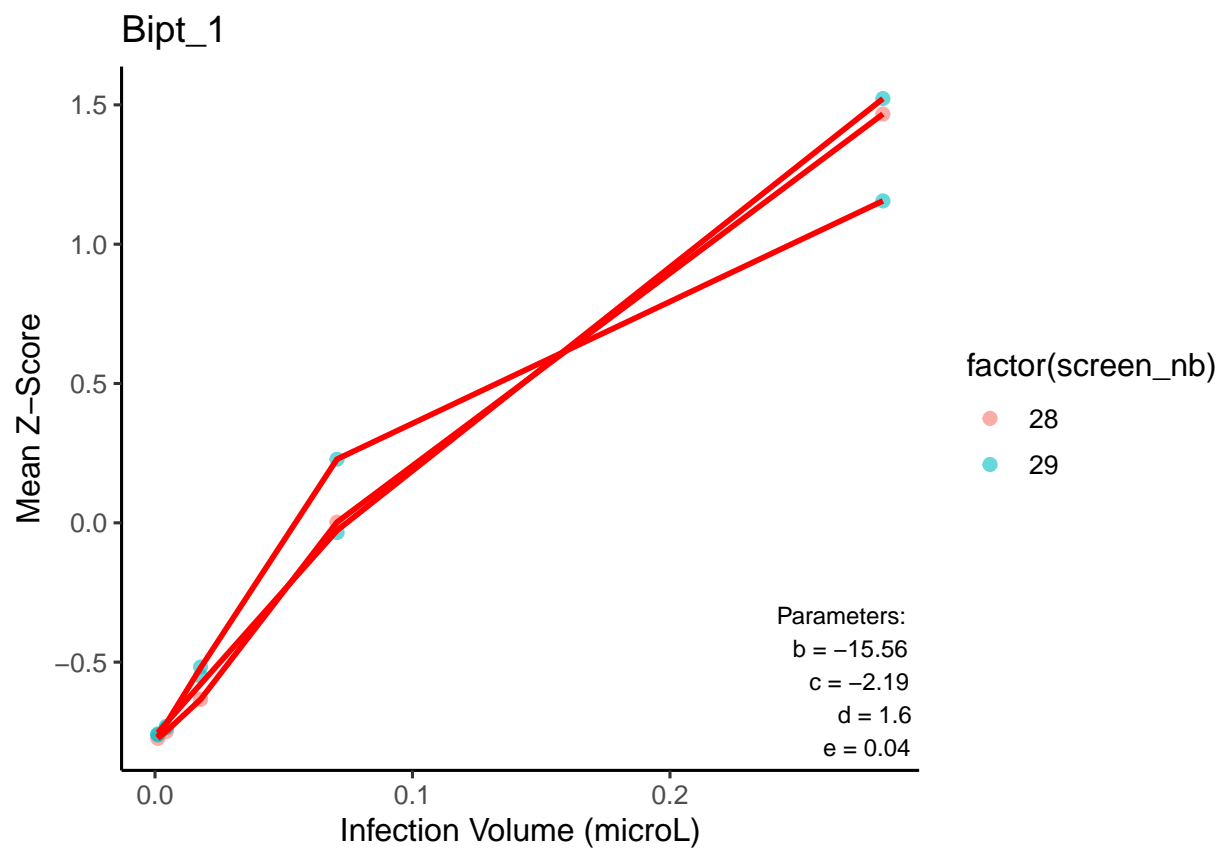
\$Babk_2



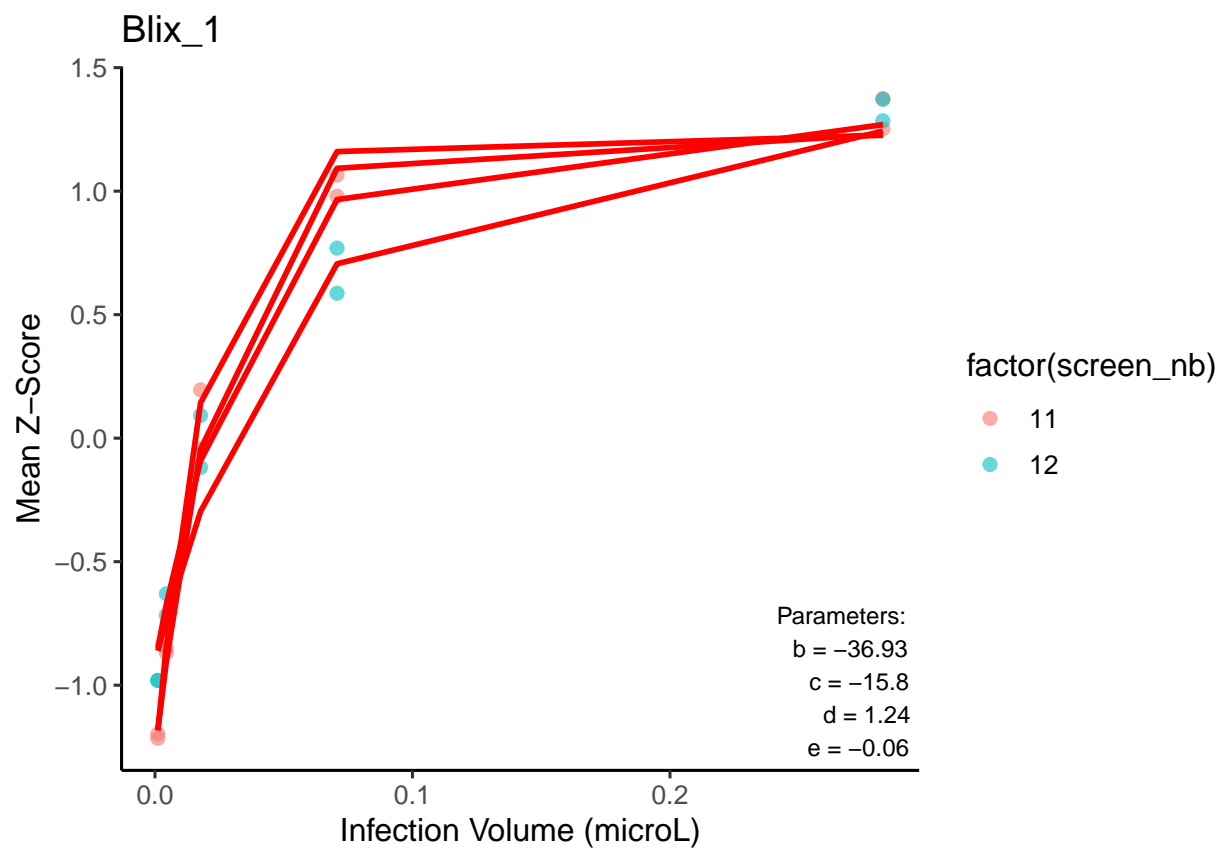
\$Bezi_3



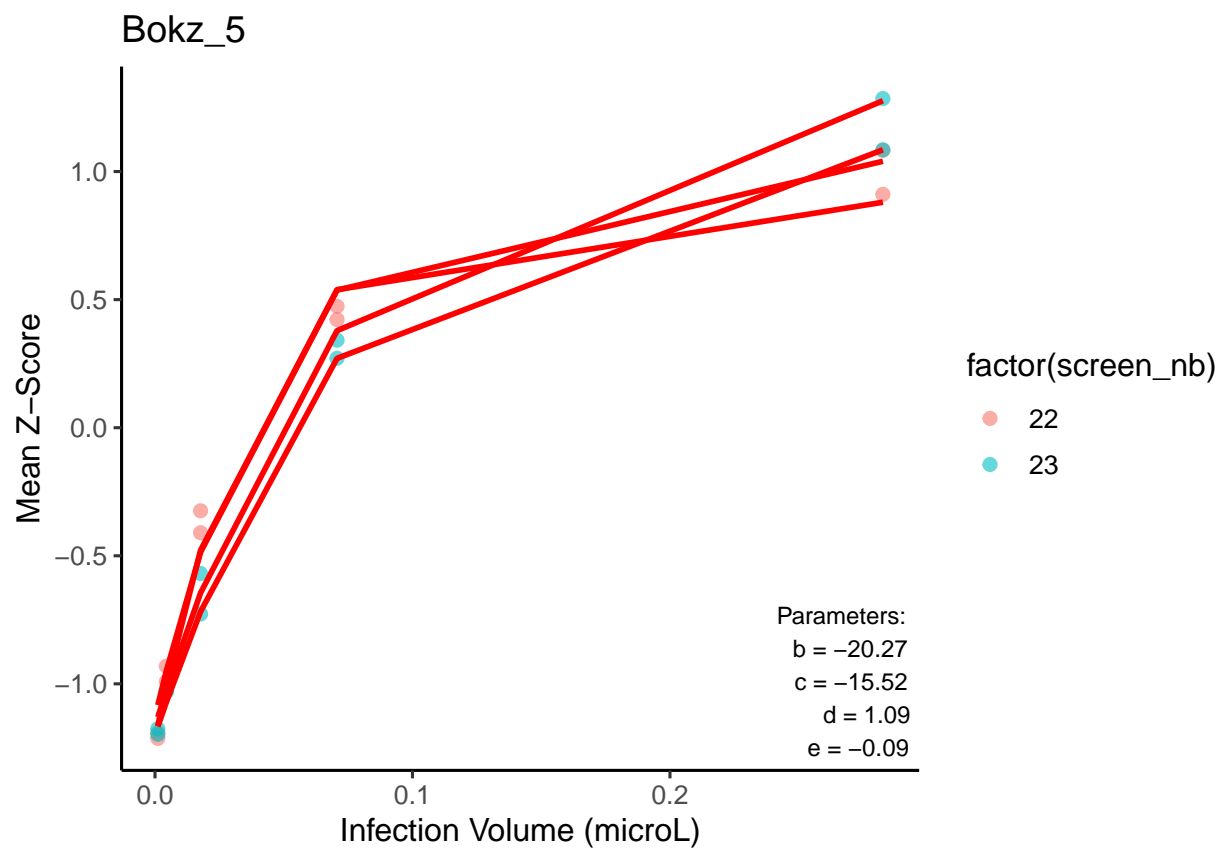
\$Bipt_1



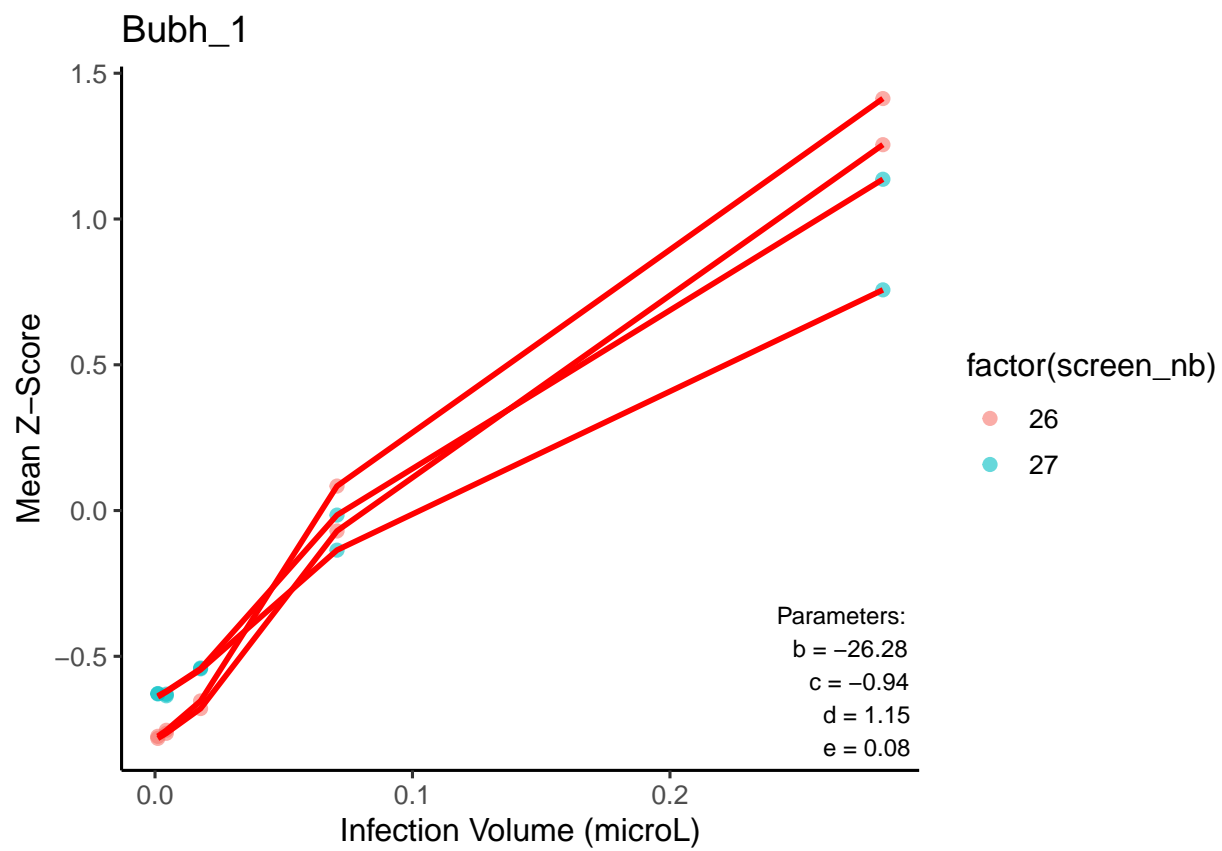
```
##  
## $Blix_1
```

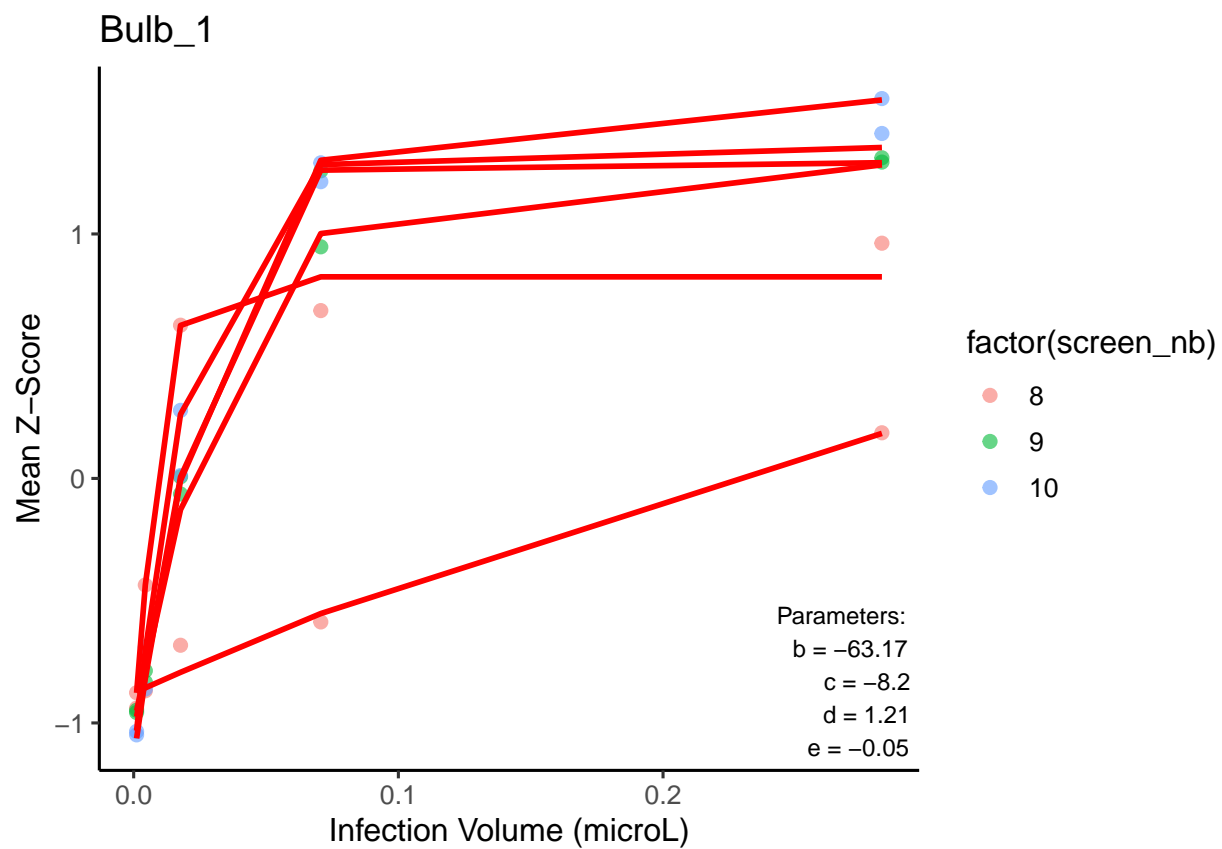
\$Bokz_5



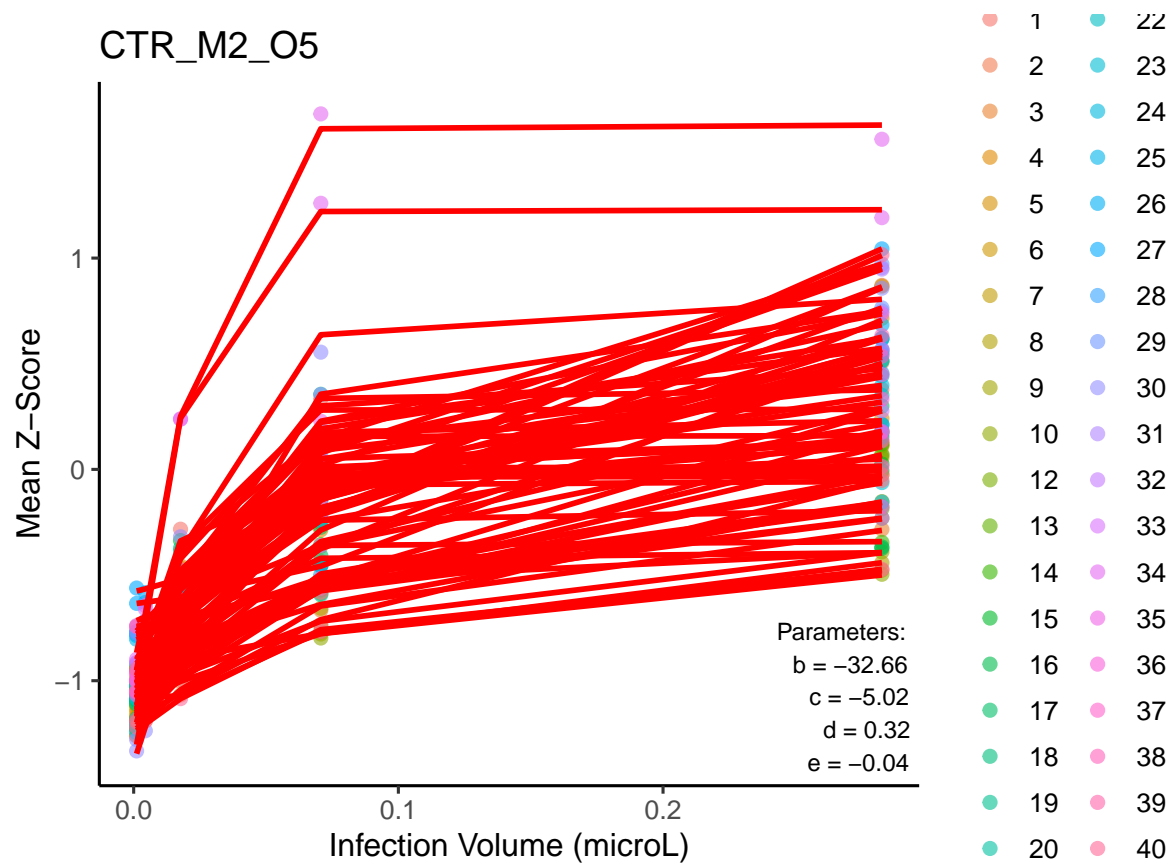
\$Bubh_1



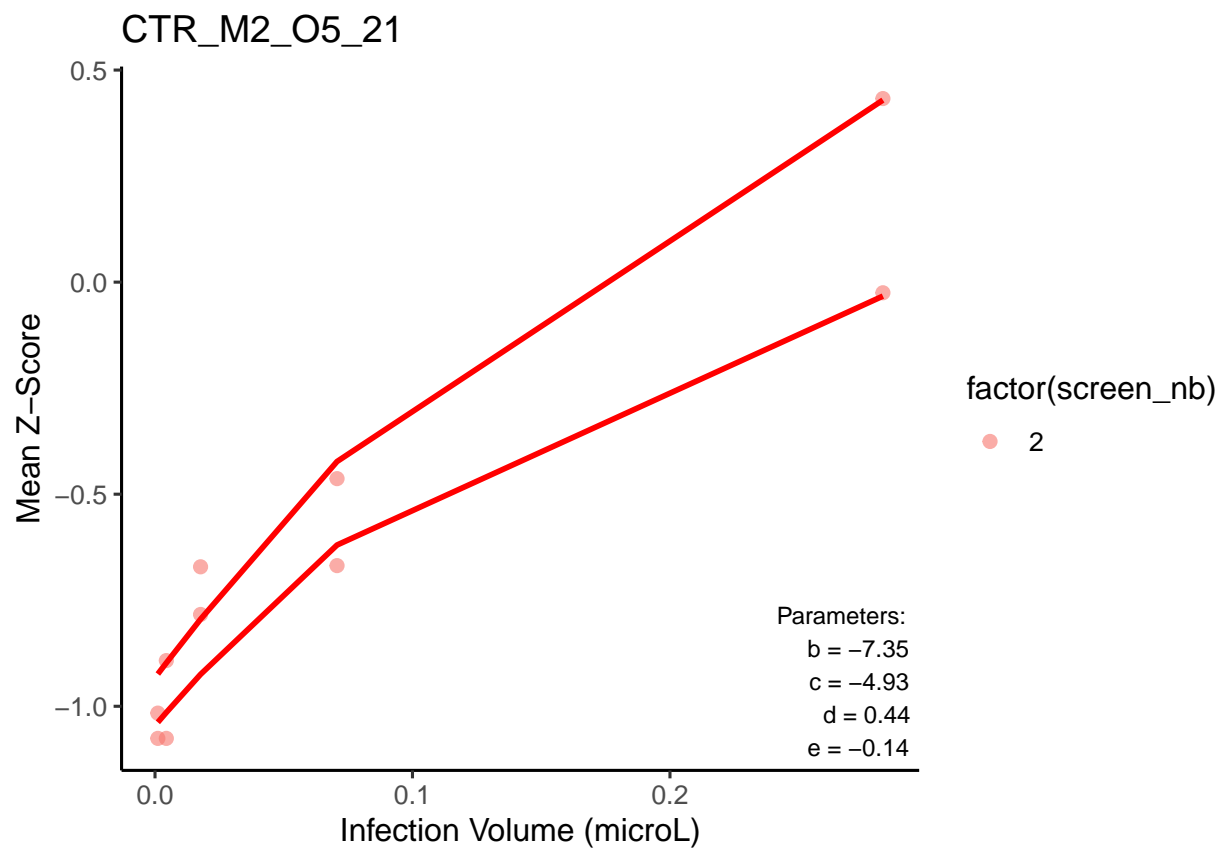
```
##  
## $Bulb_1
```



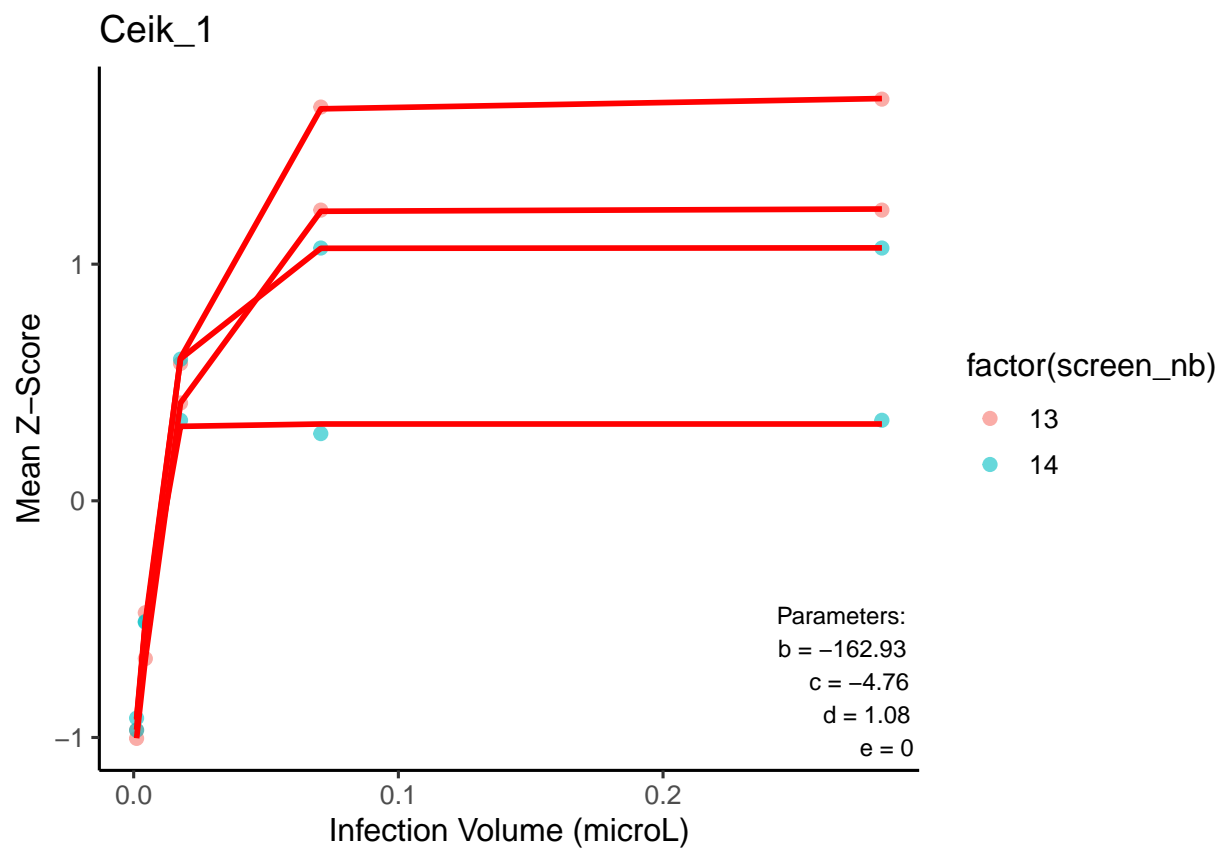
 ## \$CTR_M2_05



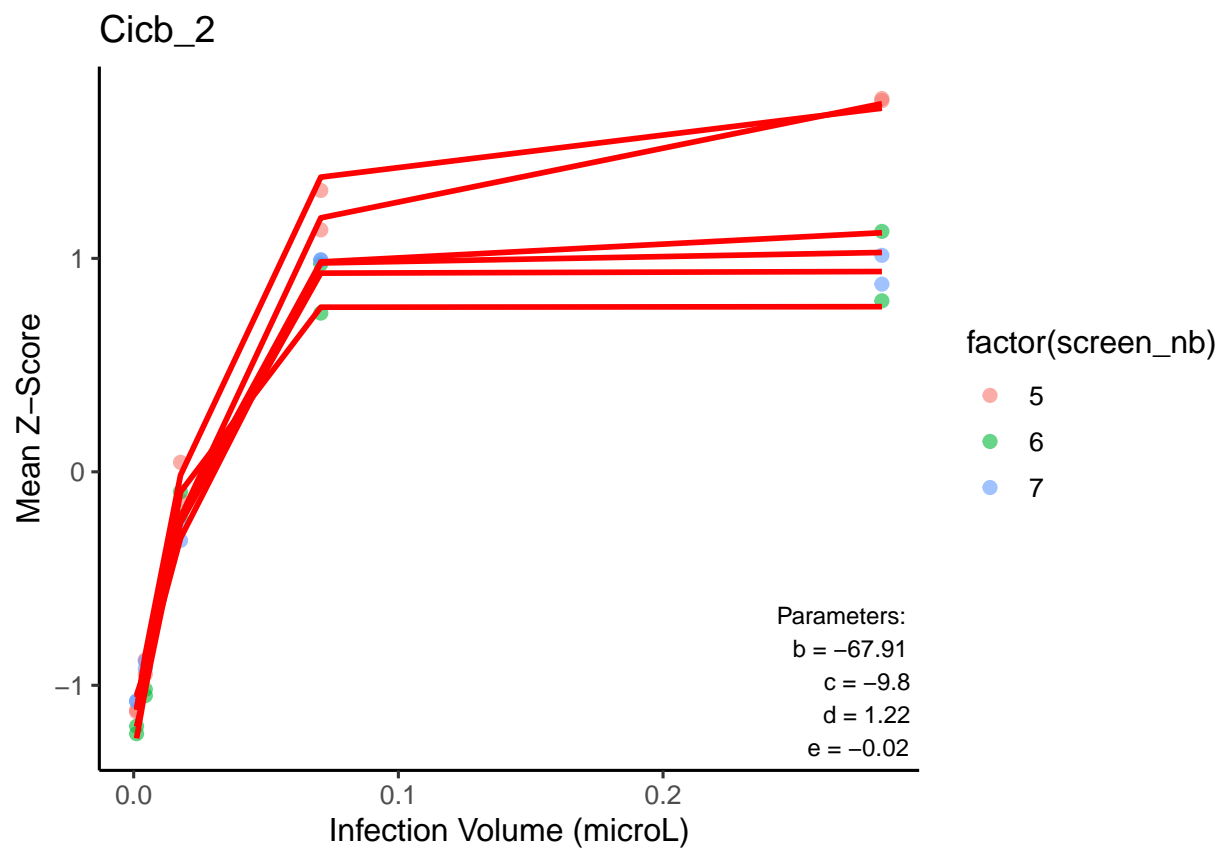
 ## \$CTR_M2_O5_21



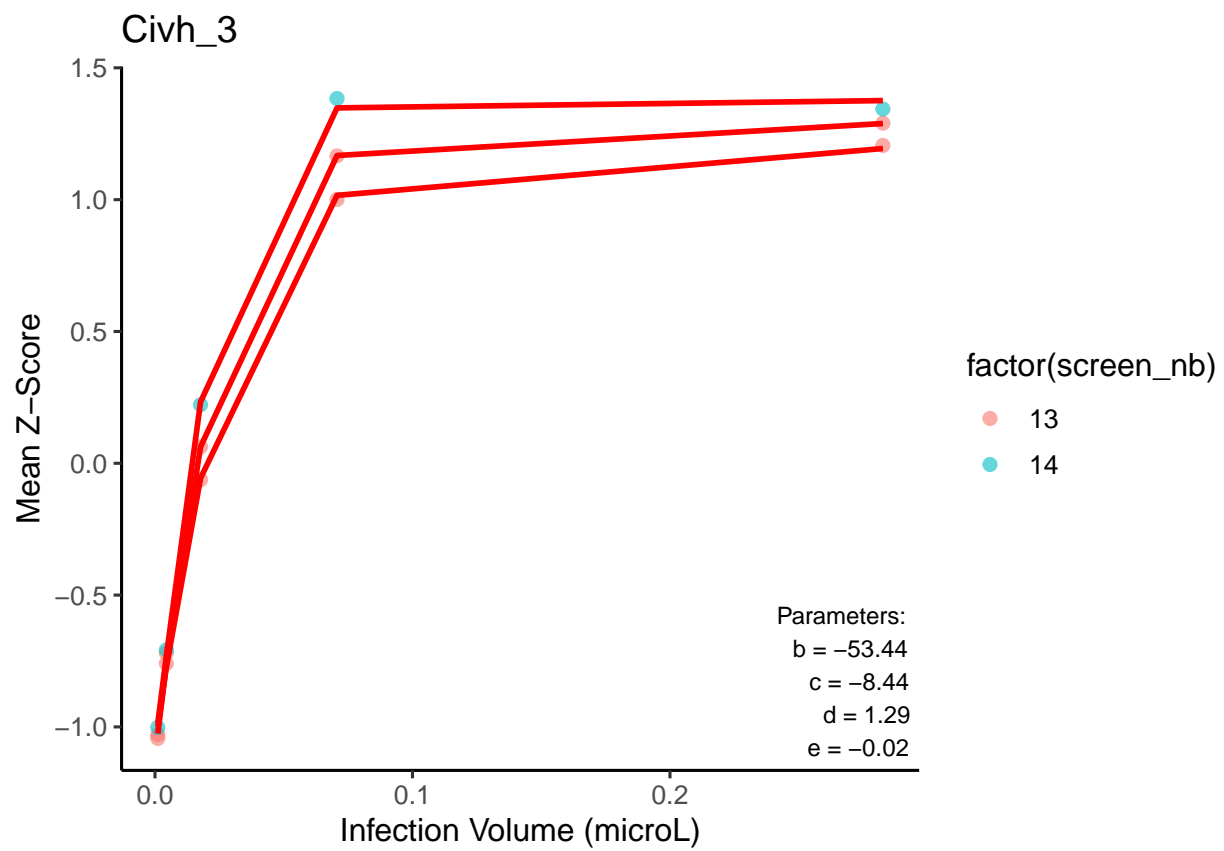
\$Ceik_1



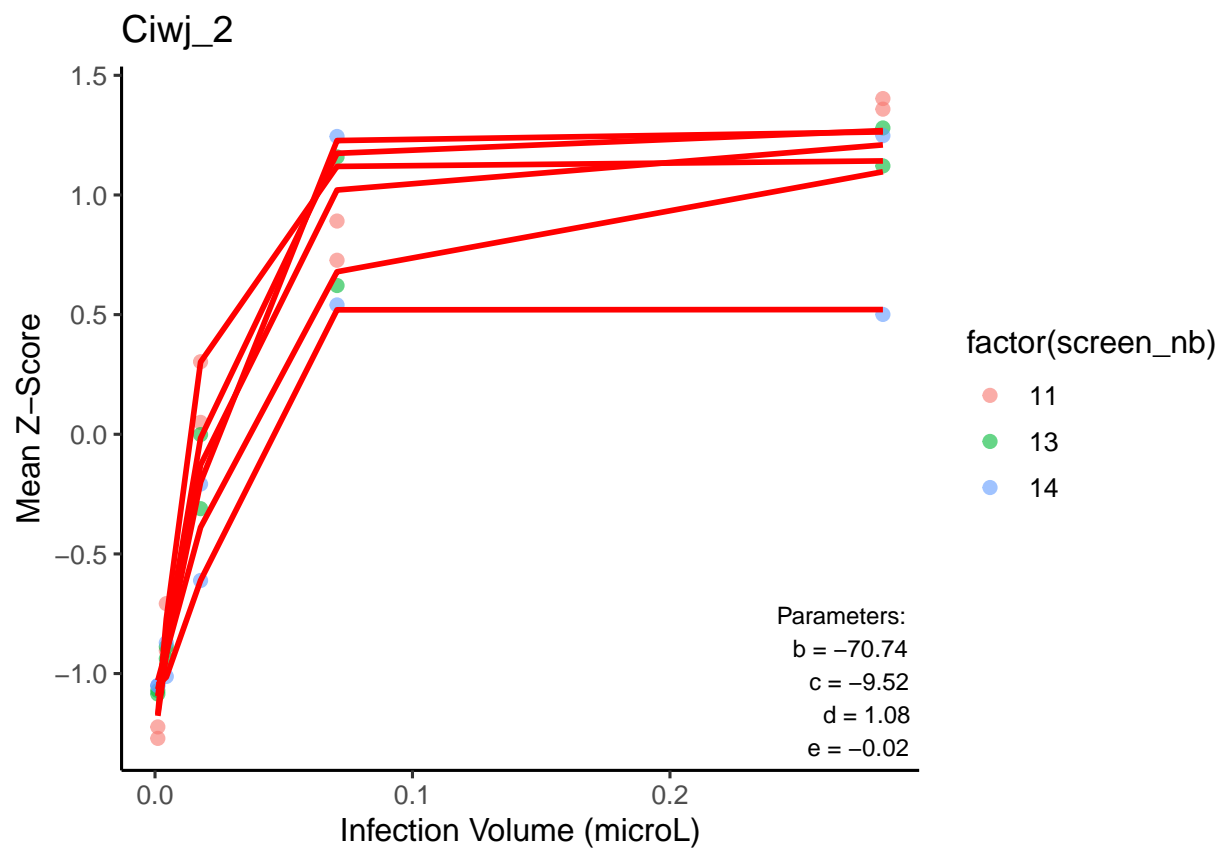
```
##  
## $Cicb_2
```



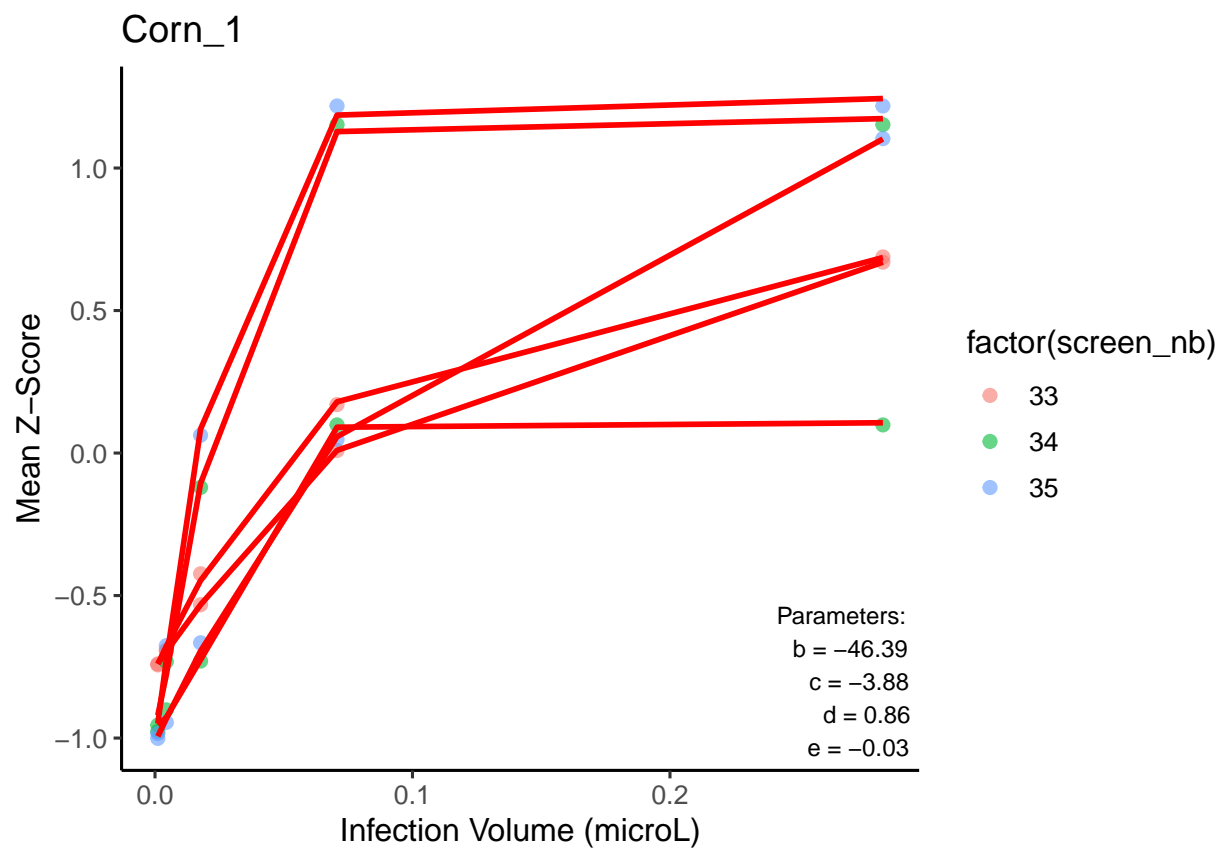
 ## \$Civh_3



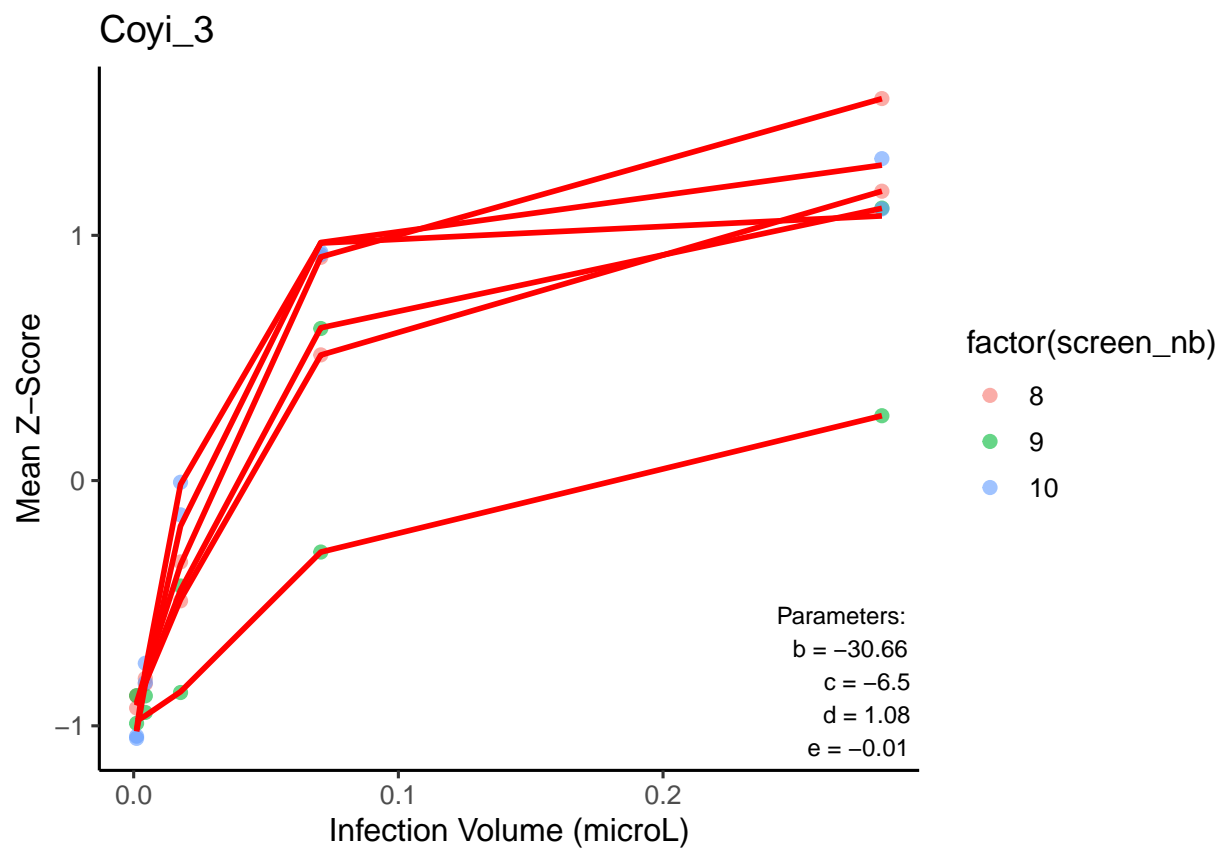
\$Ciwj_2



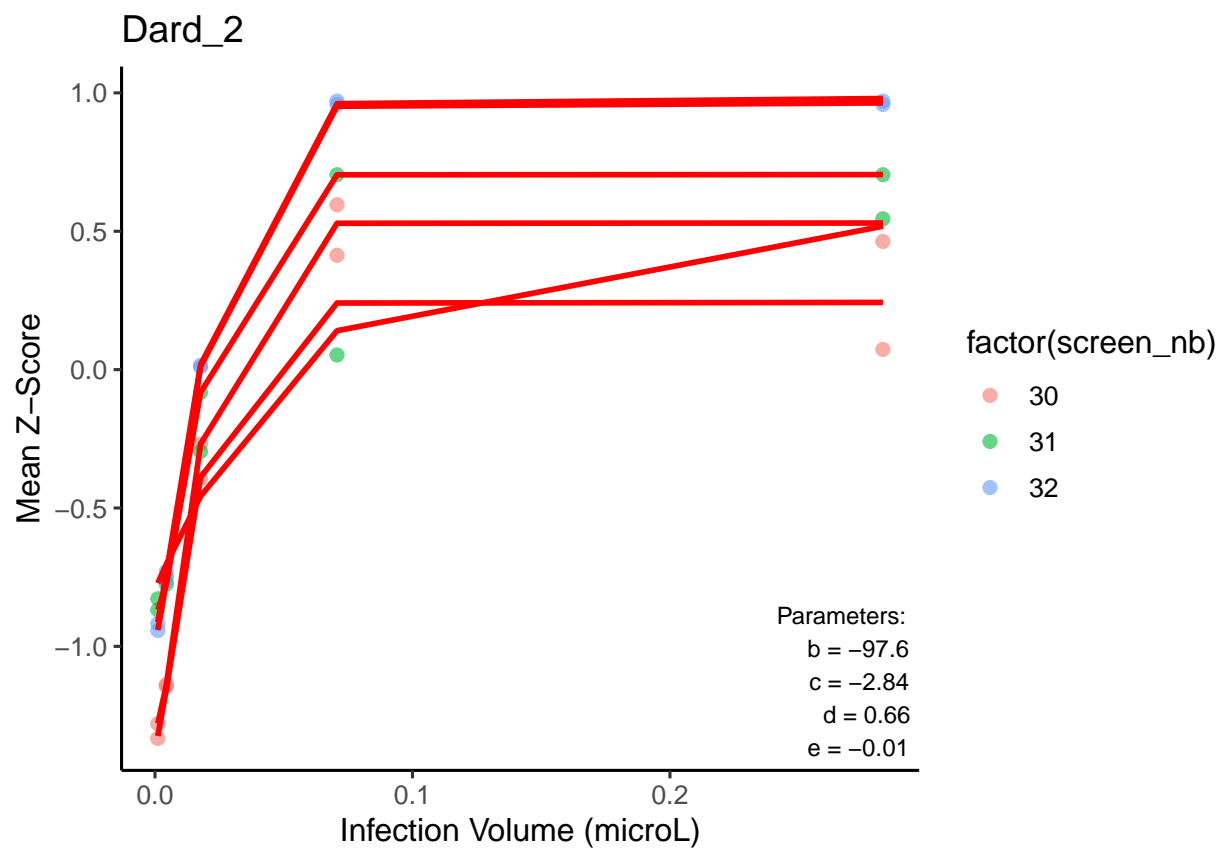
\$Corn_1



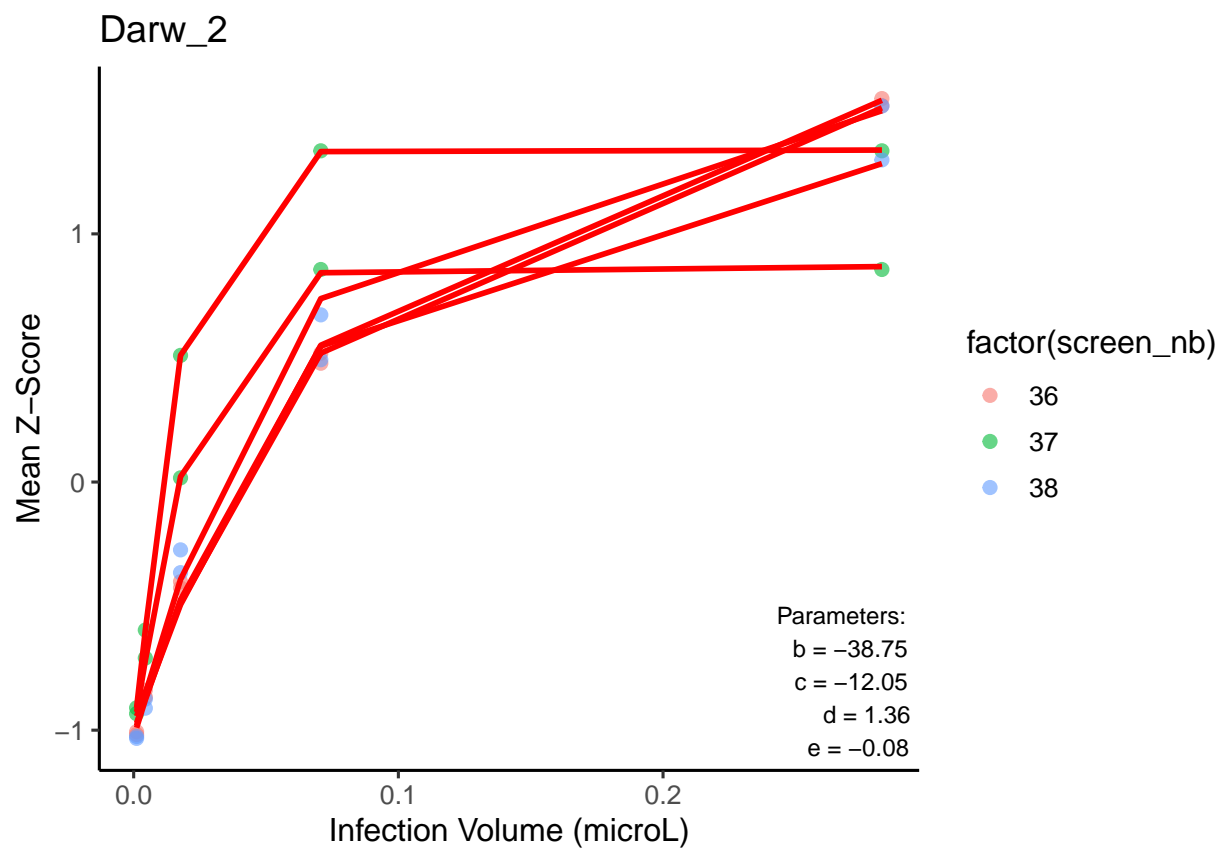
 ## \$Coyi_3



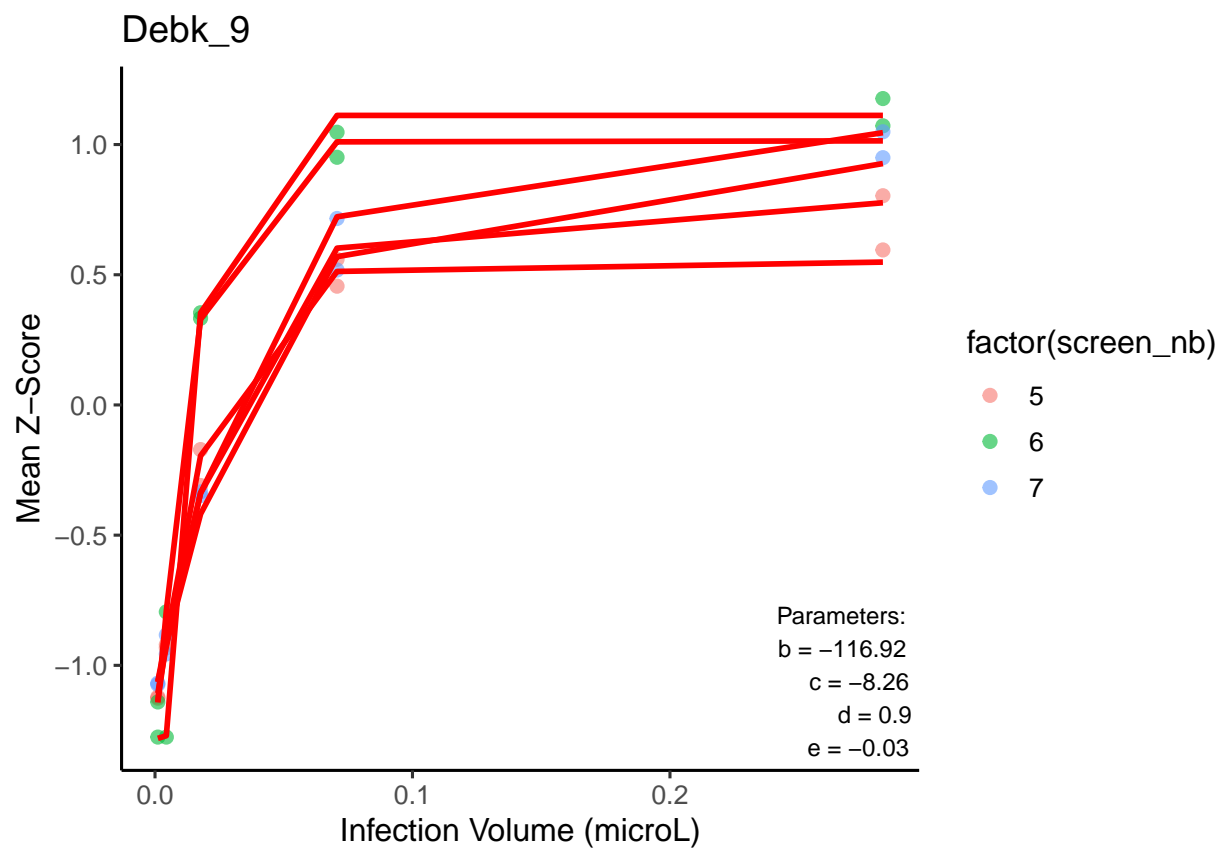
 ## \$Dard_2



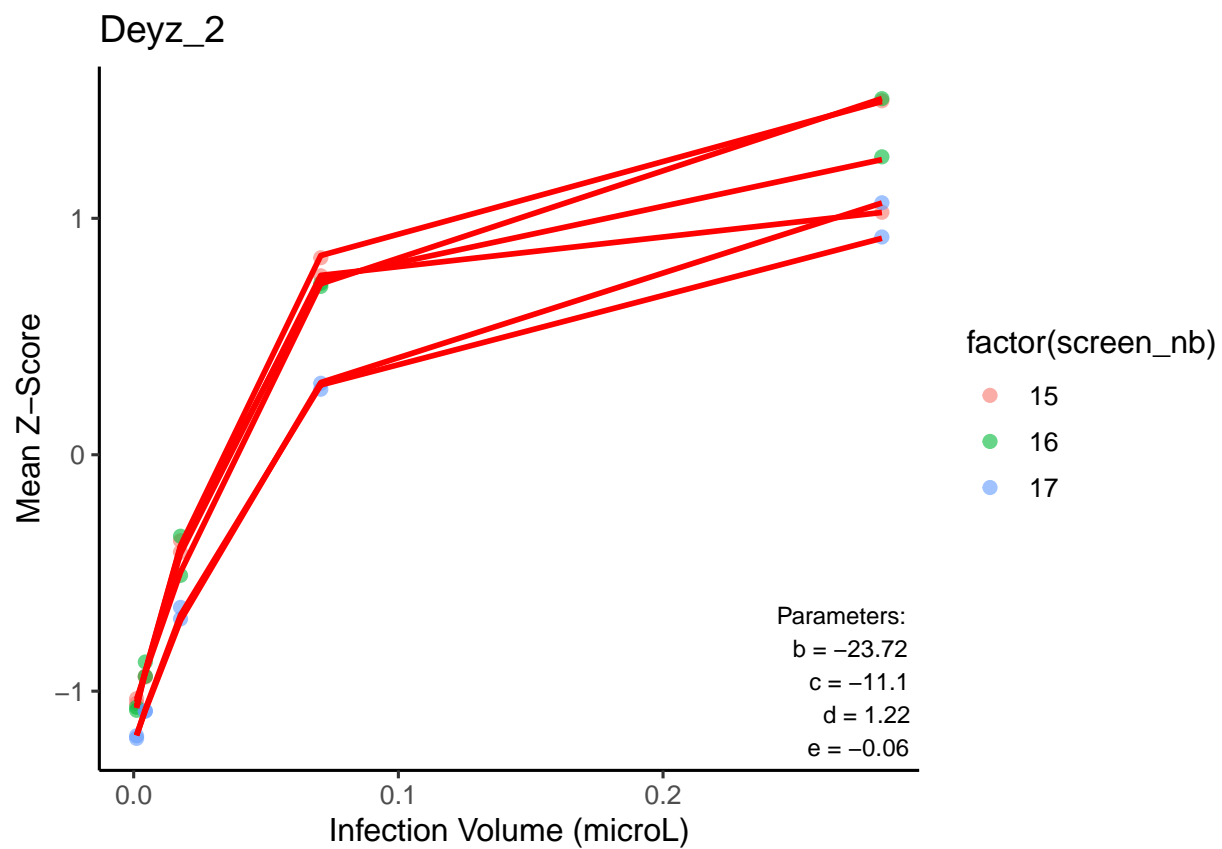
```
##
## $Darw_2
```



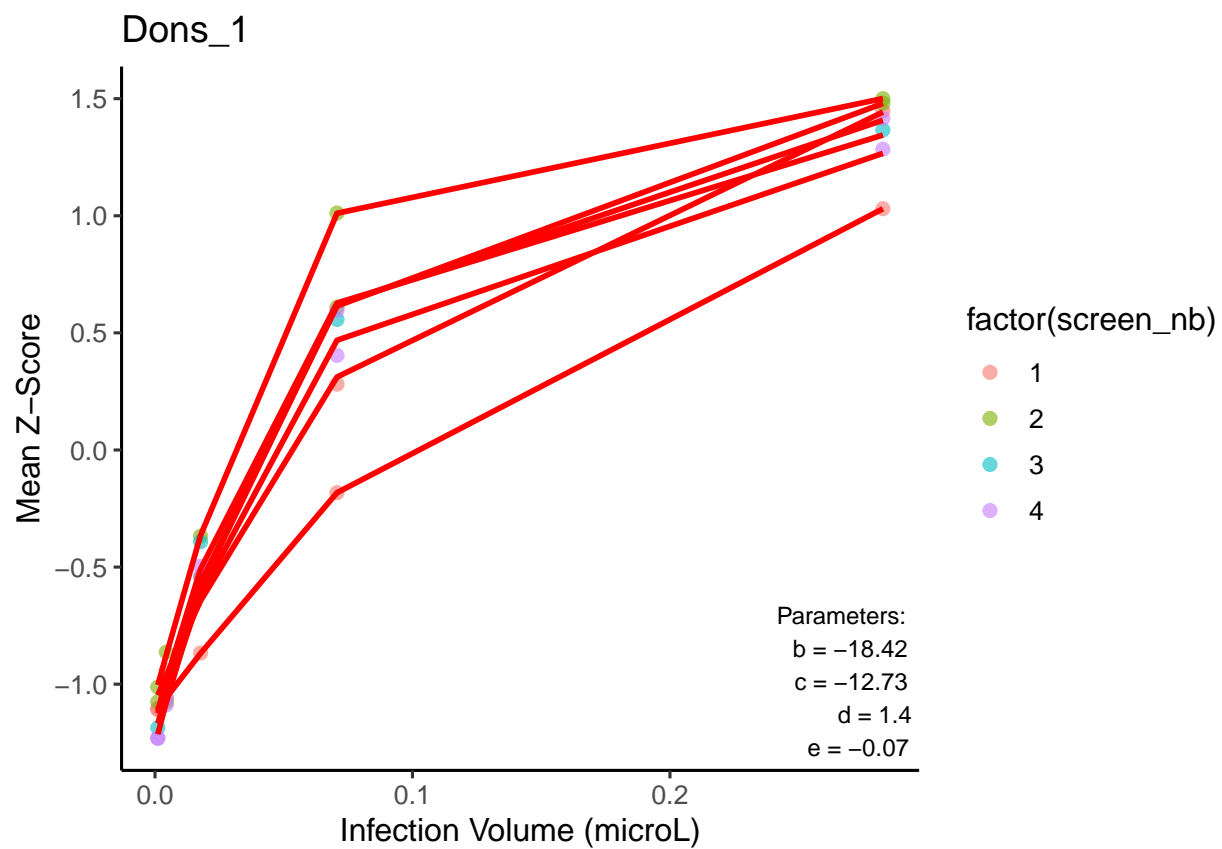
 ## \$DebK_9



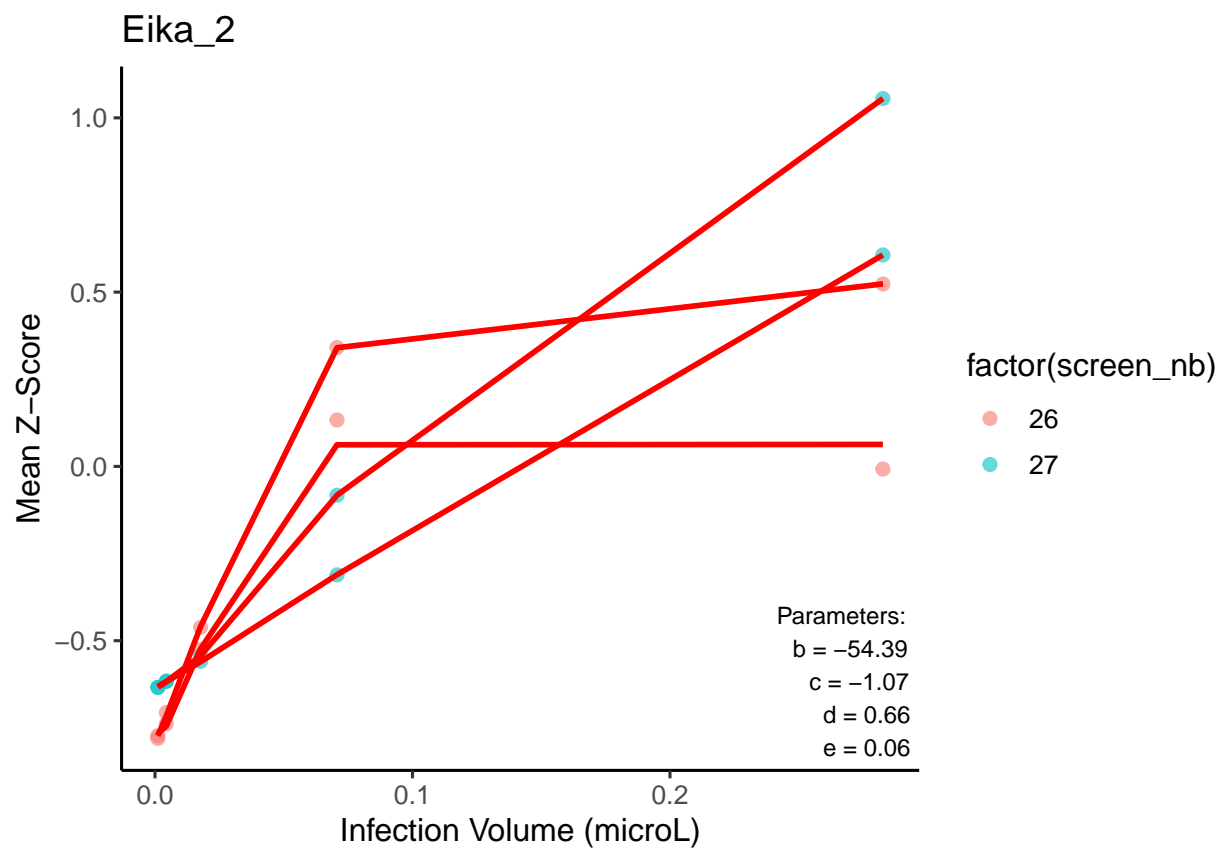
\$Deyz_2



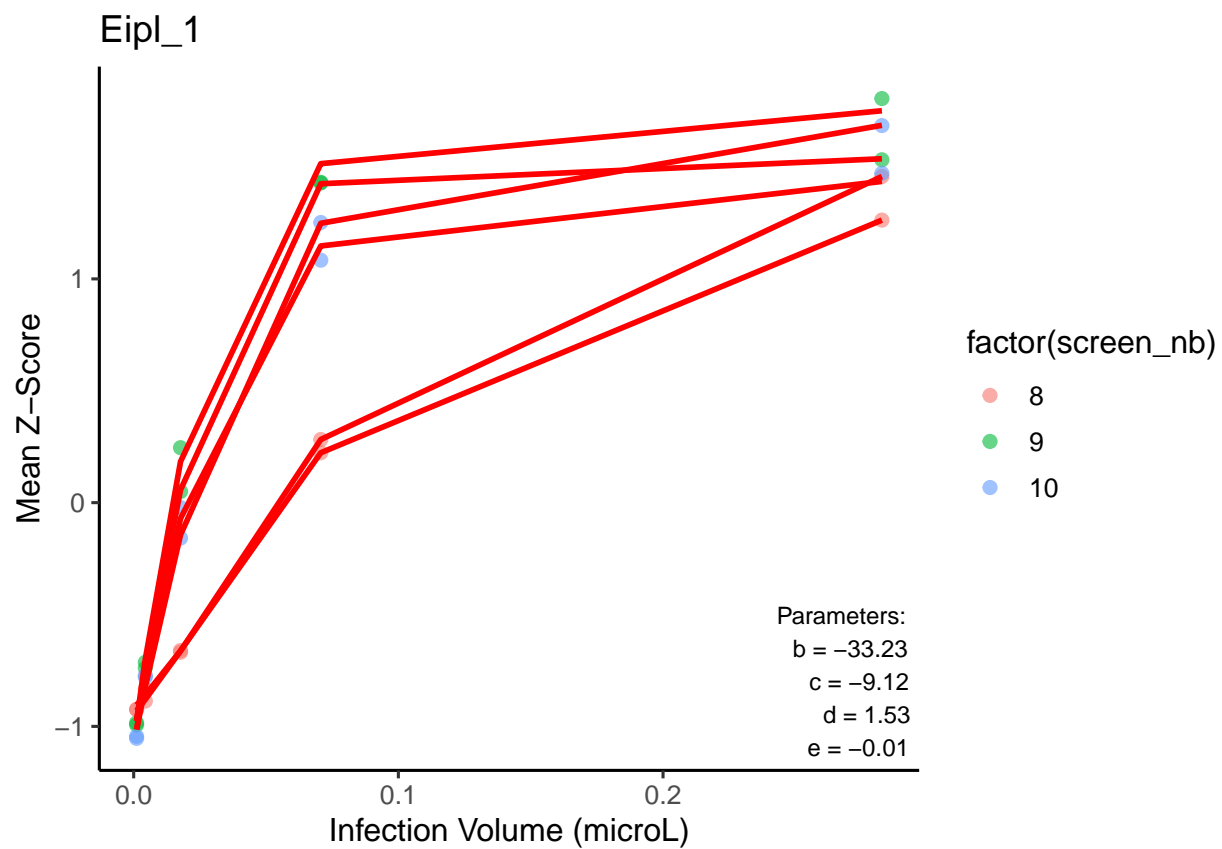
\$Dons_1



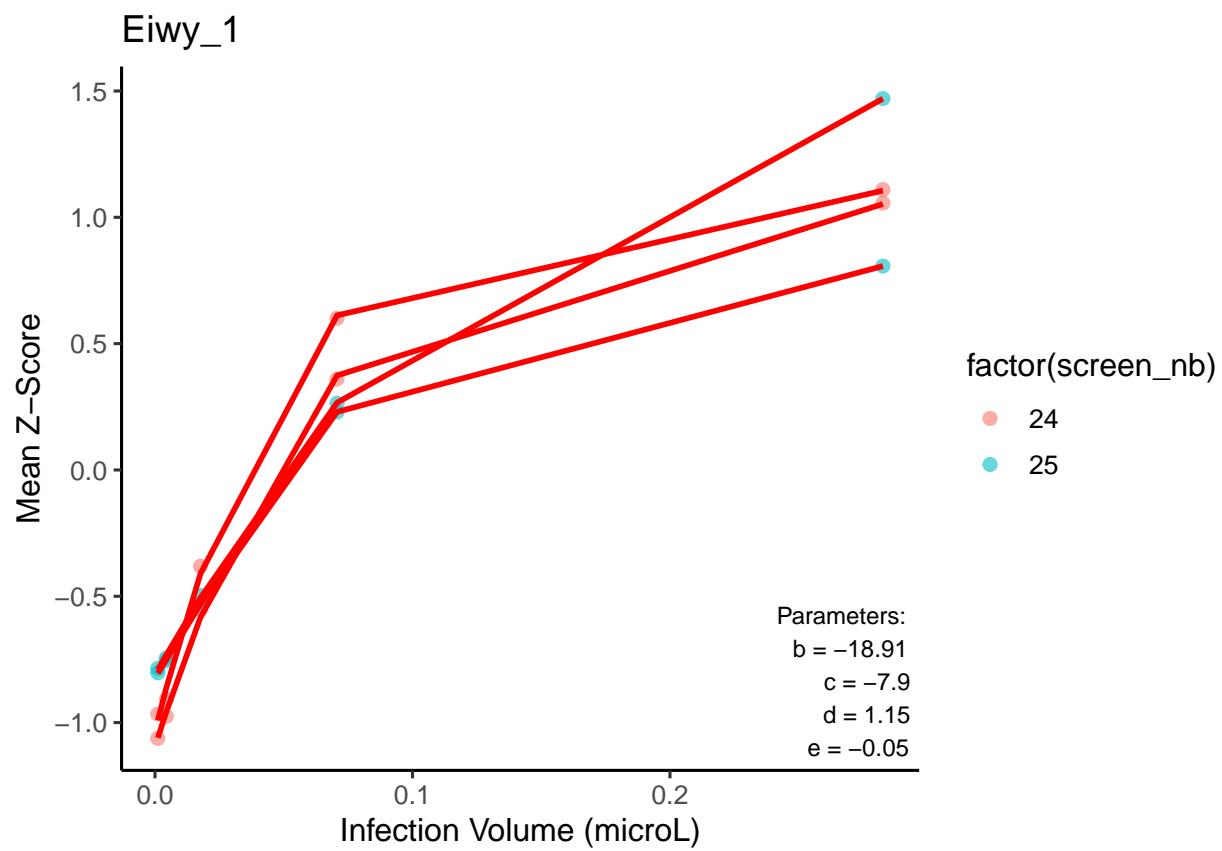
```
##  
## $Eika_2
```



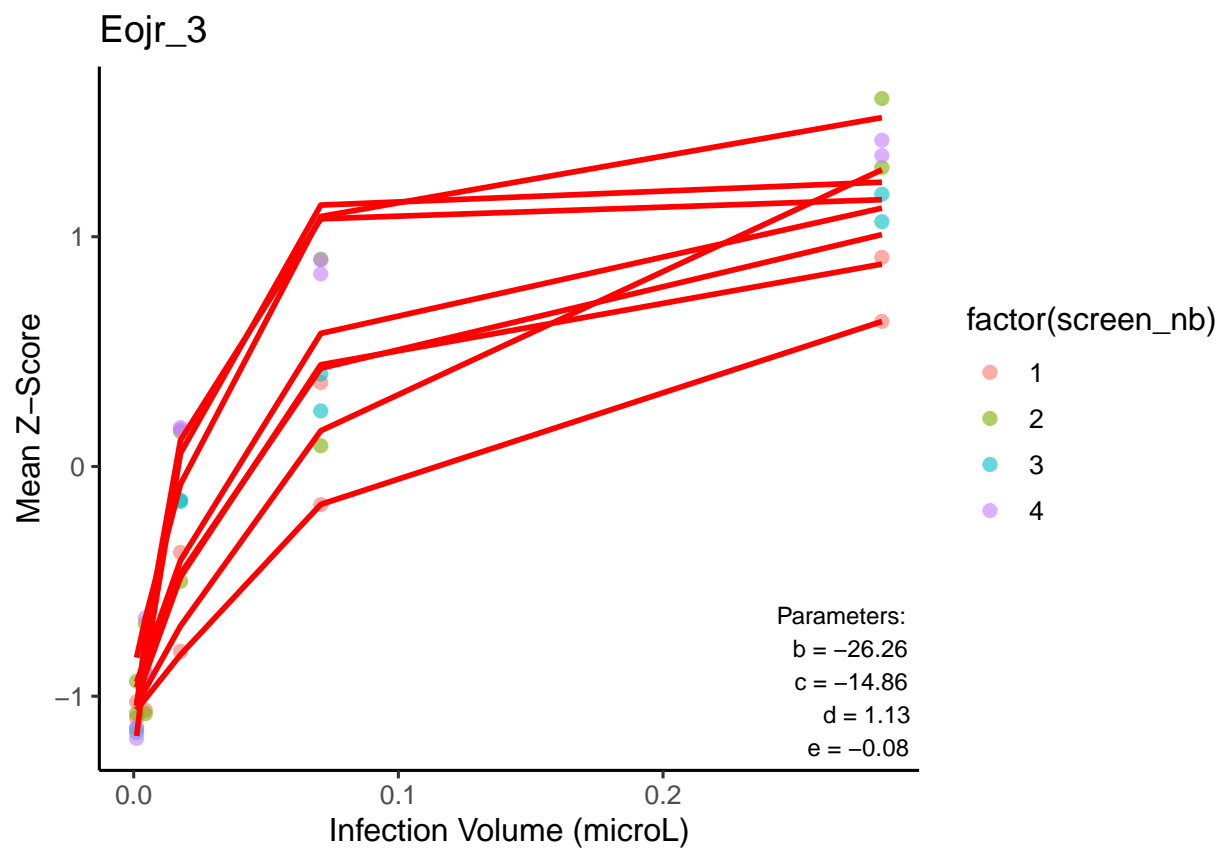
\$Eipl_1



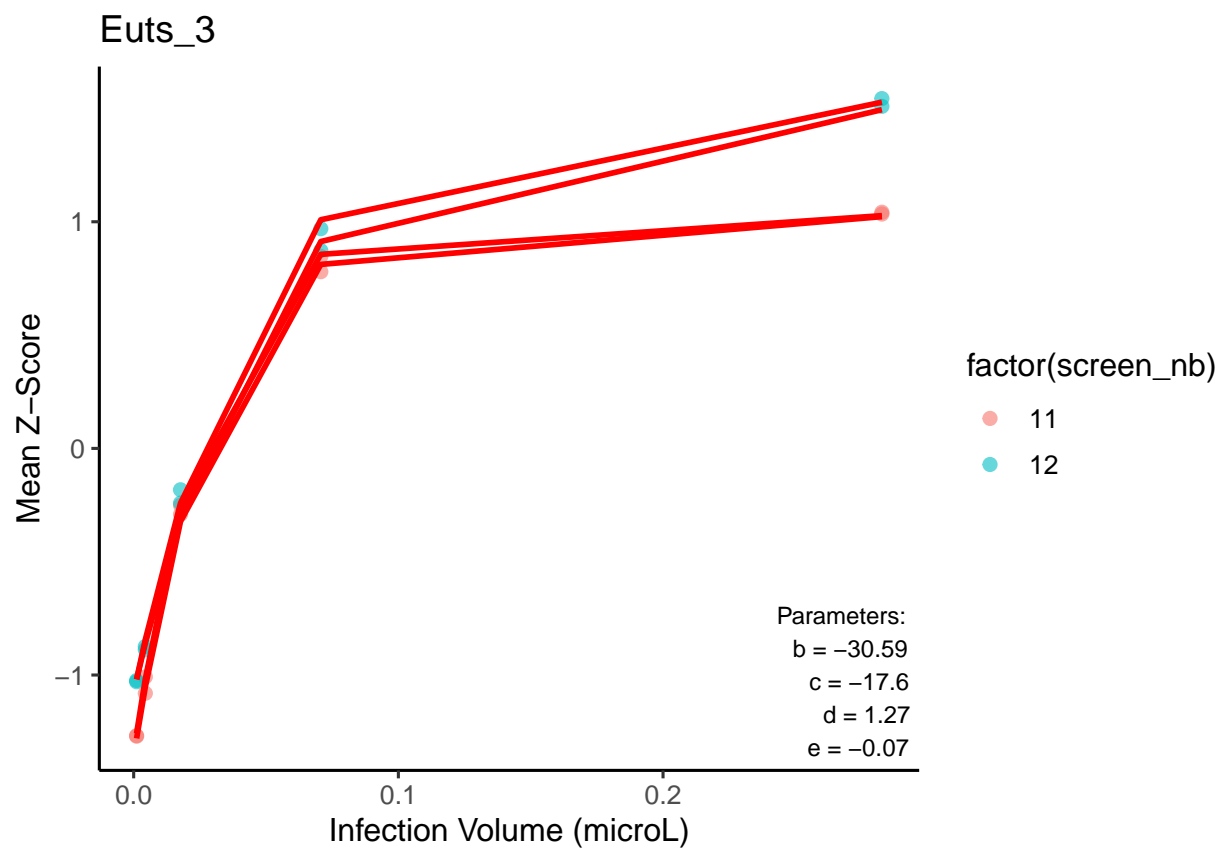
\$Eiwy_1



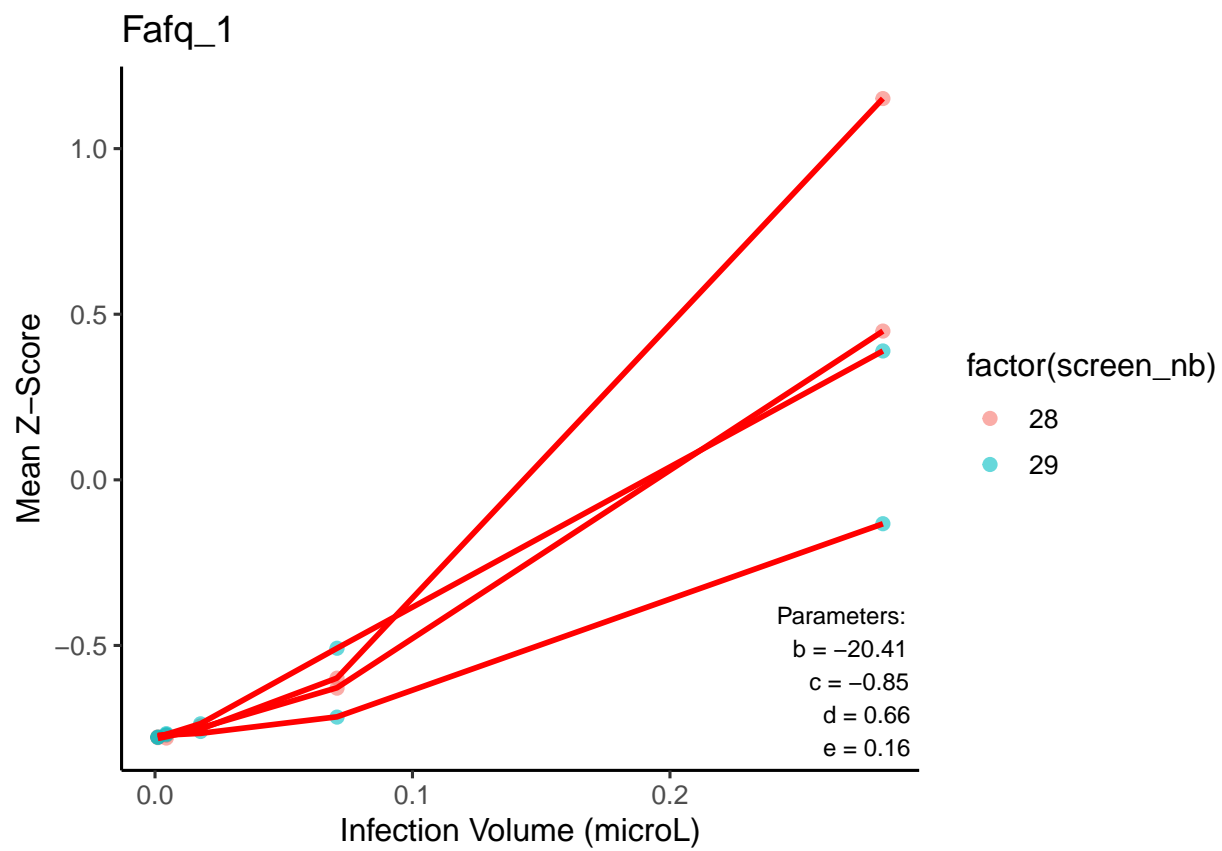
\$Eojr_3



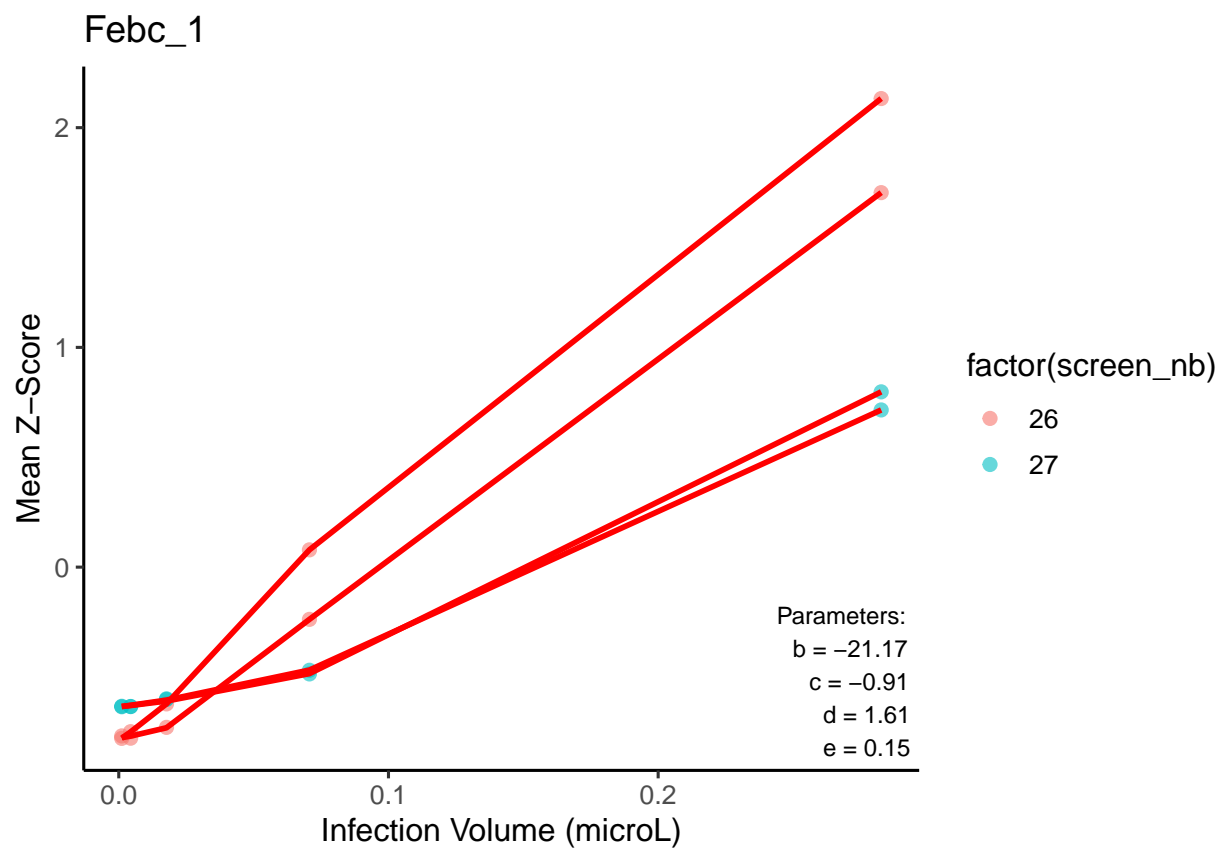
\$Euts_3



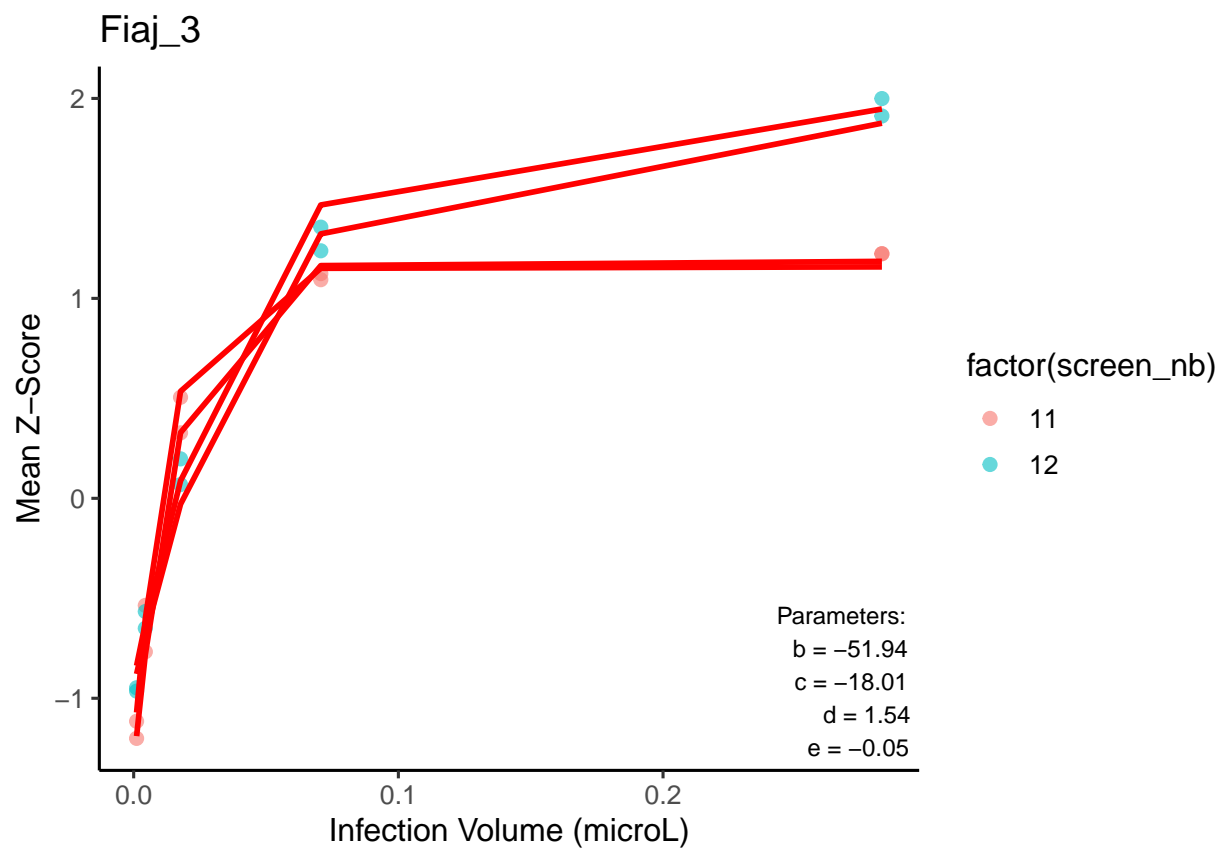
```
##  
## $Fafq_1
```



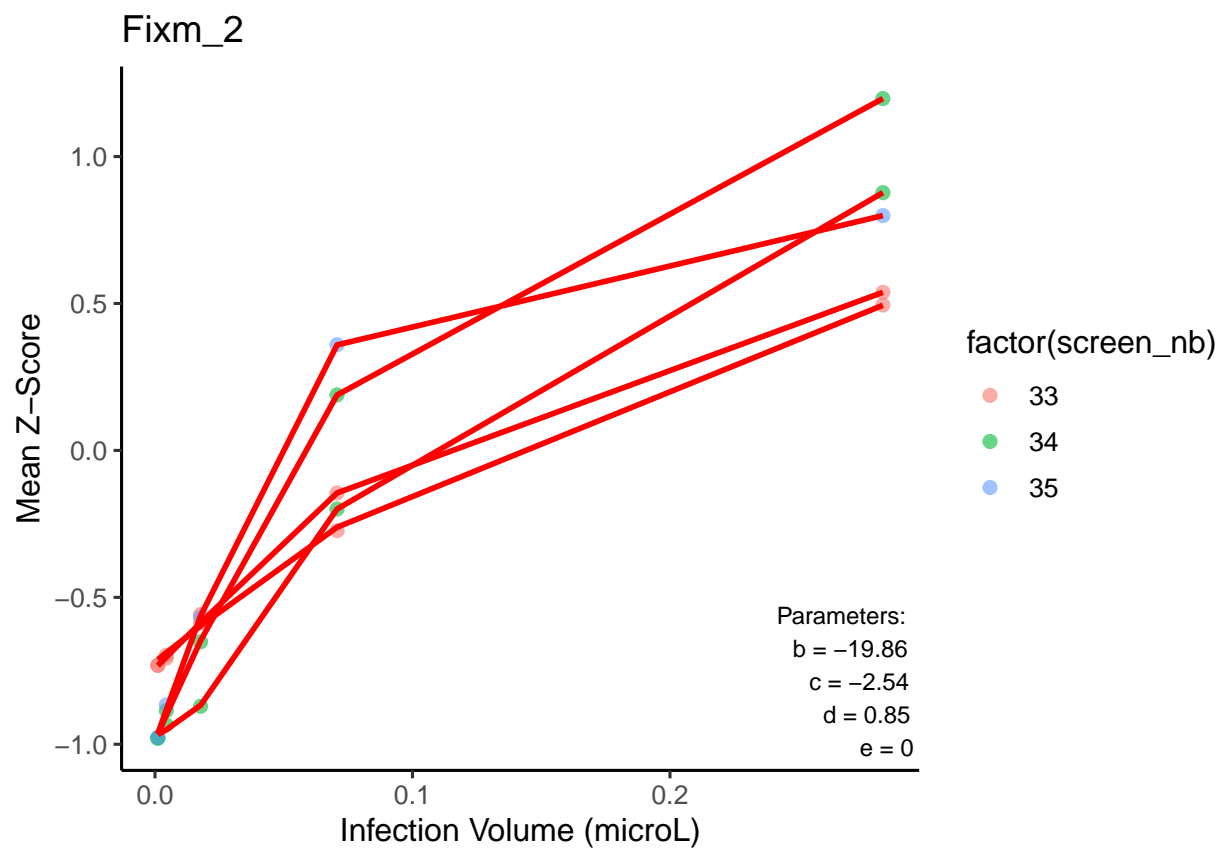
 ## \$Febc_1



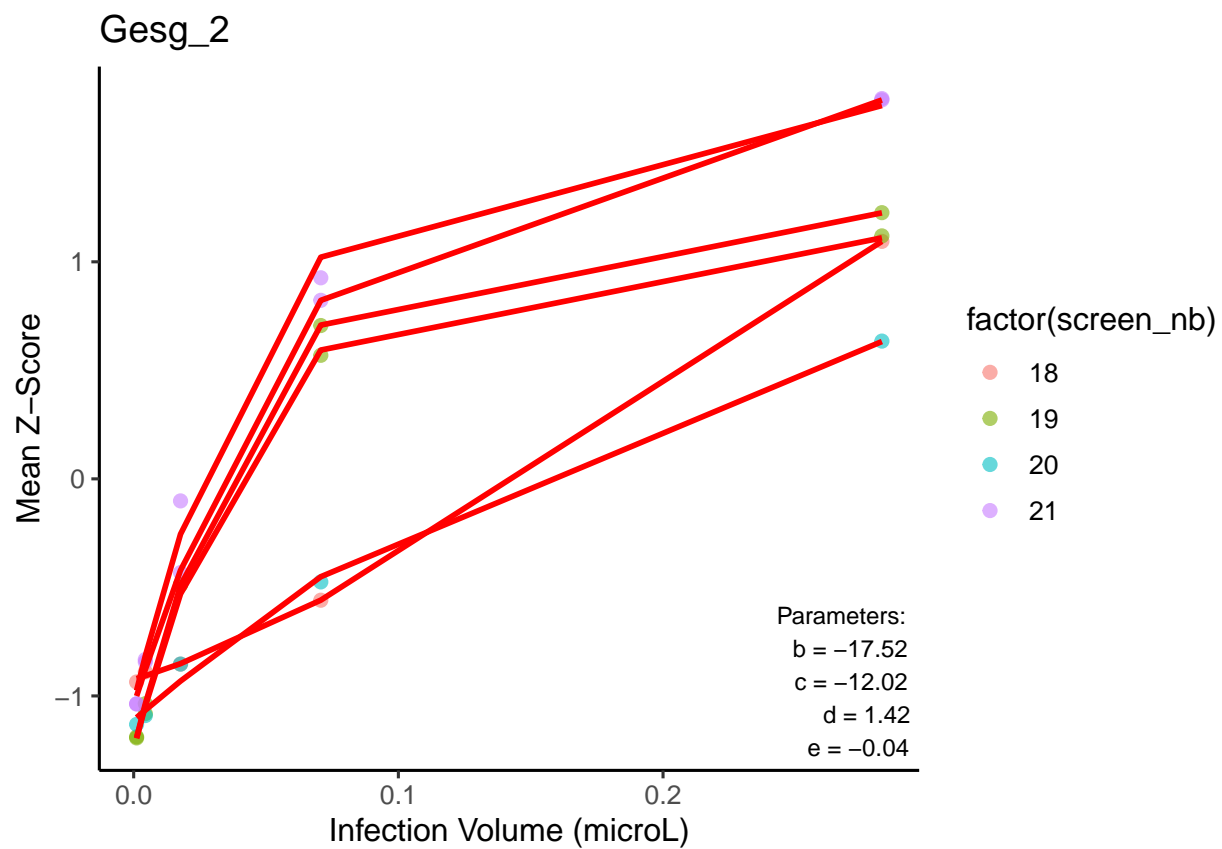
```
##  
## $Fiaj_3
```

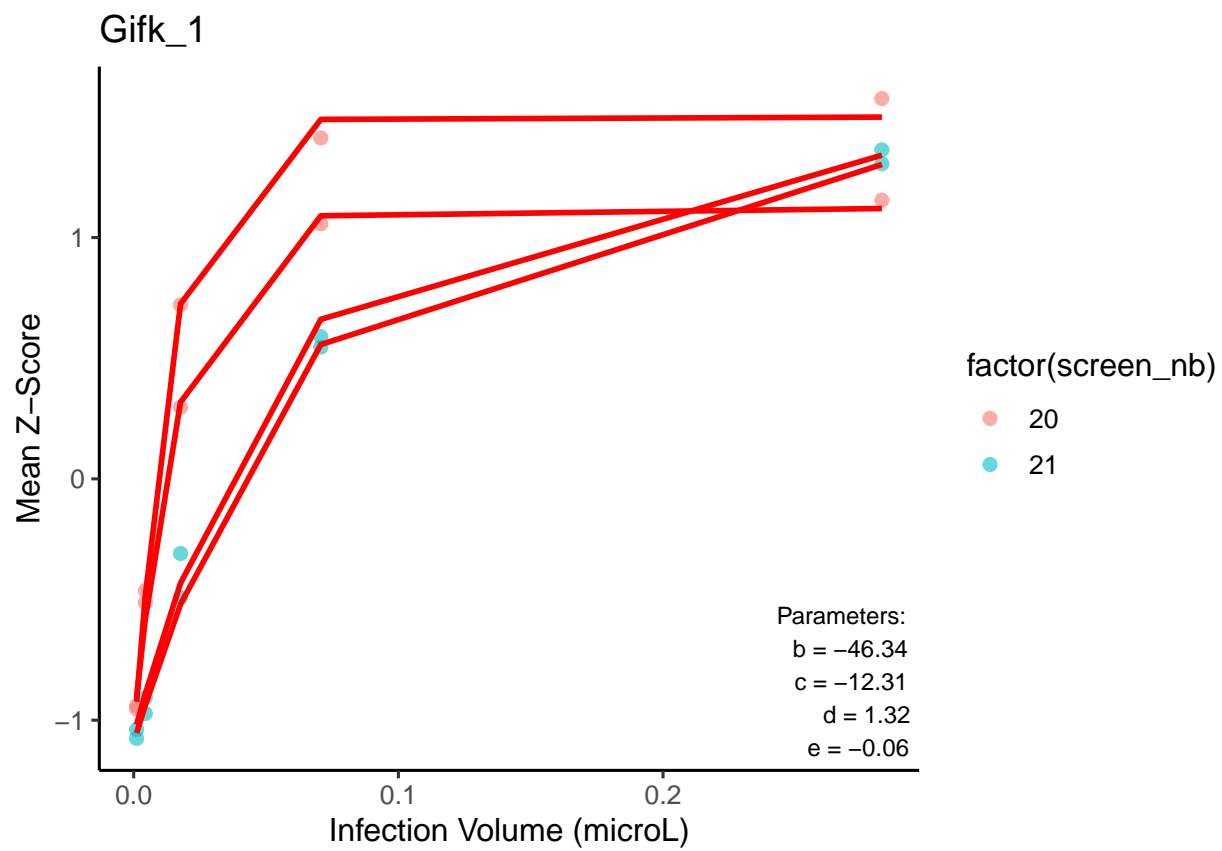
\$Fixm_2



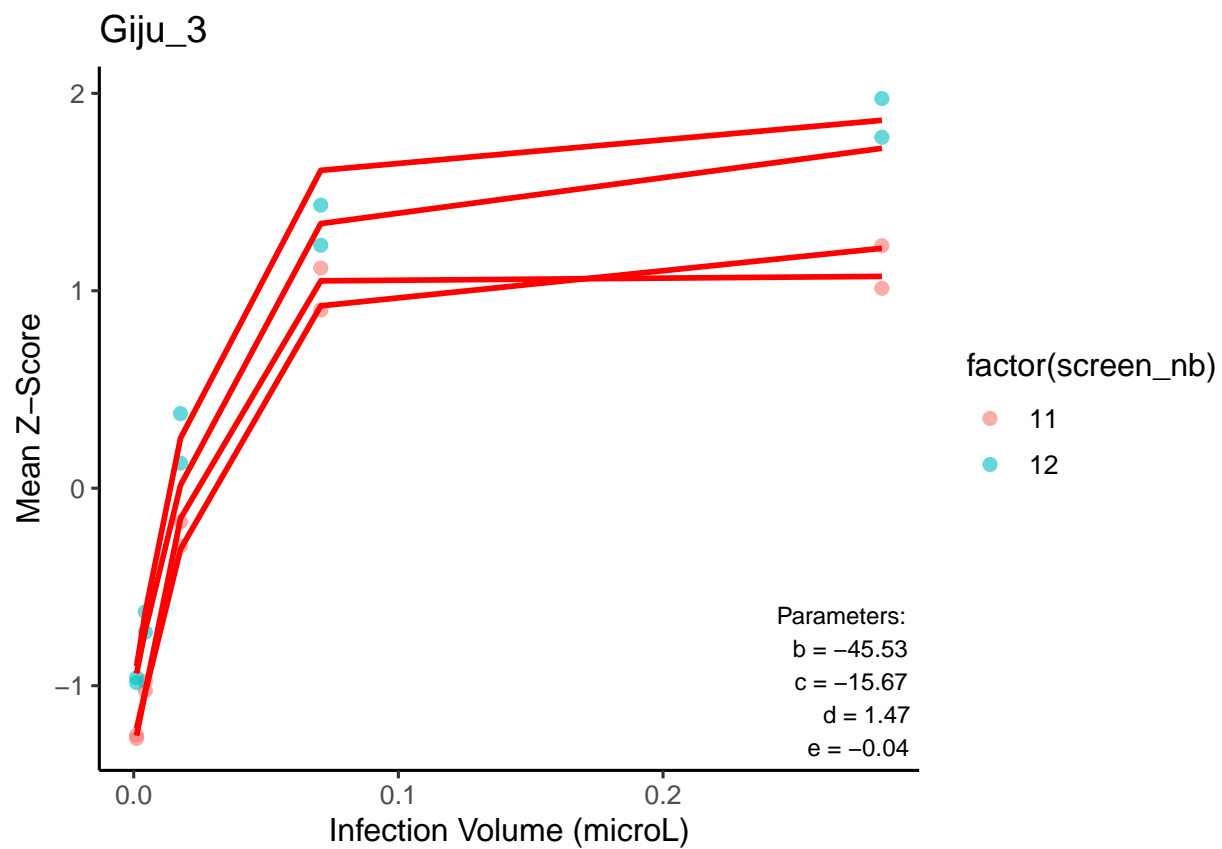
 ## \$Gesg_2



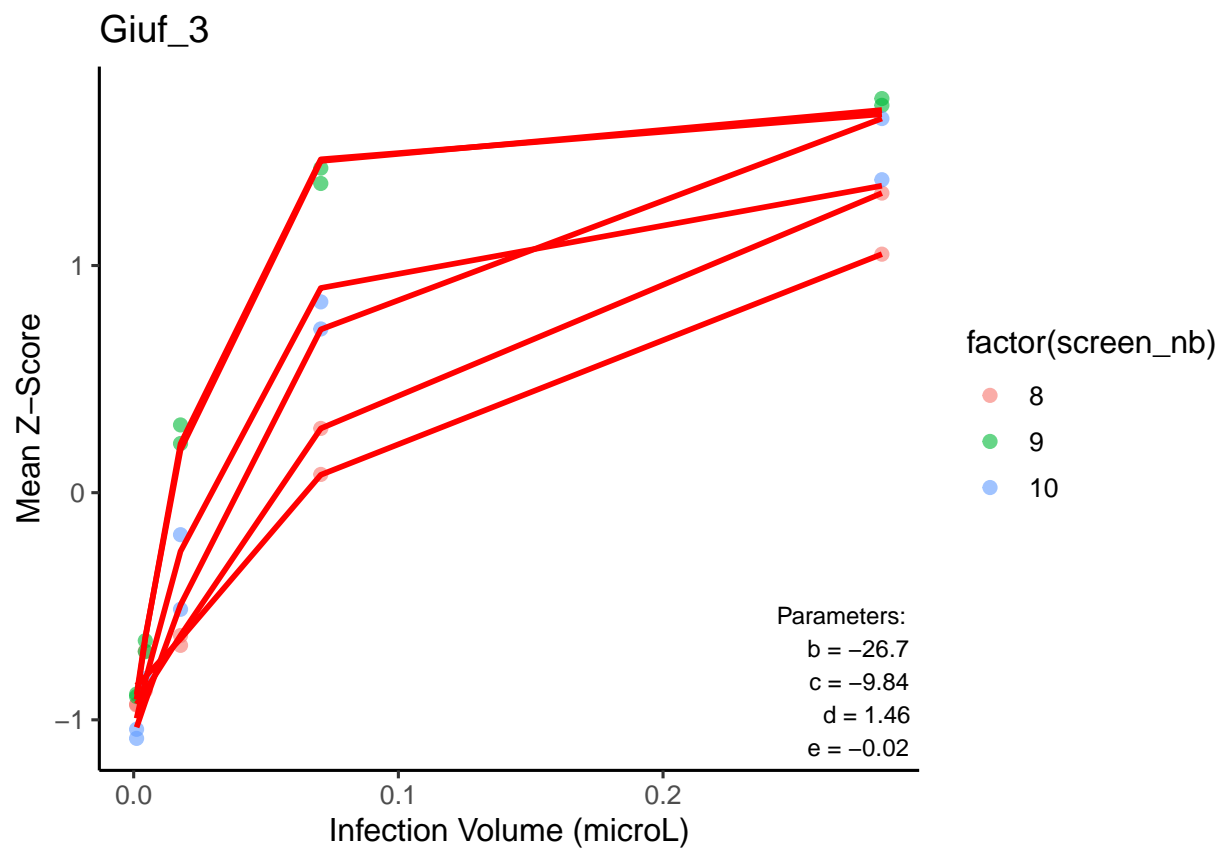
\$Gifk_1



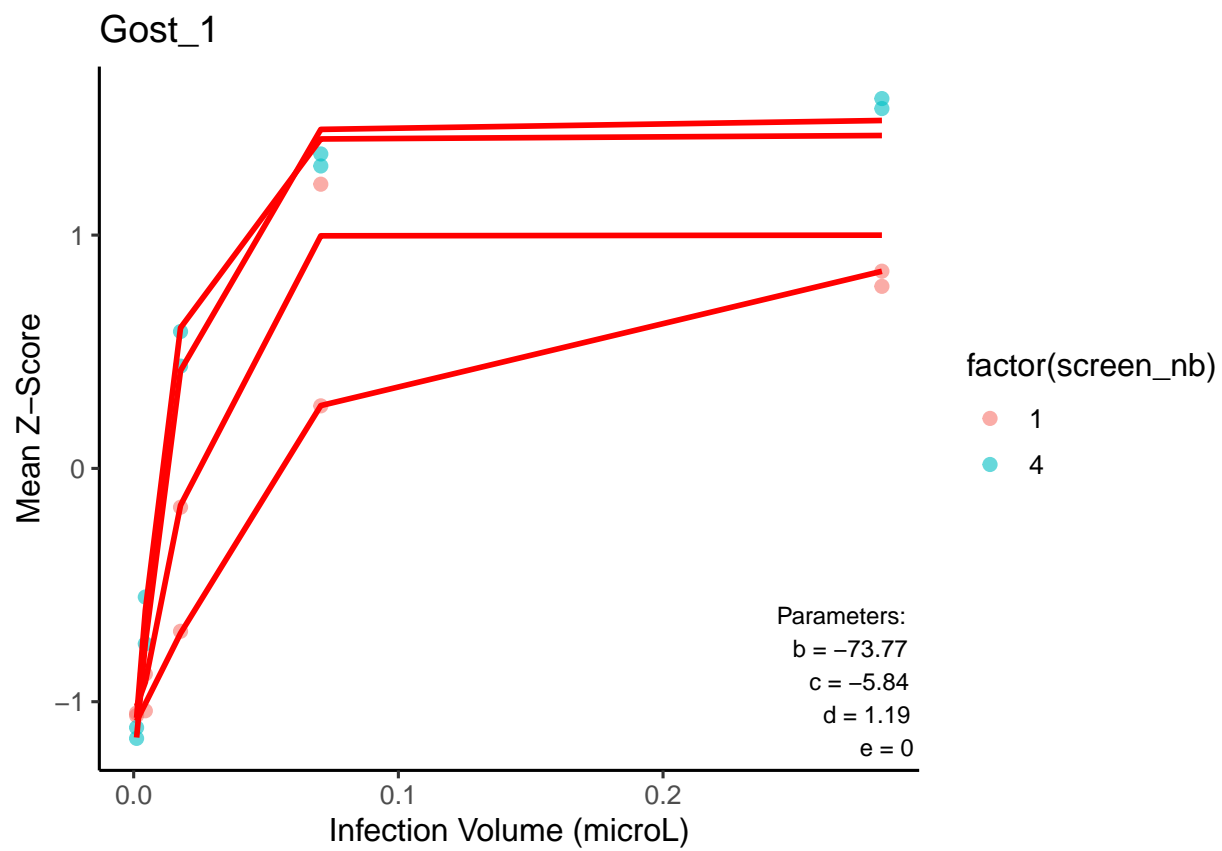
\$Giju_3



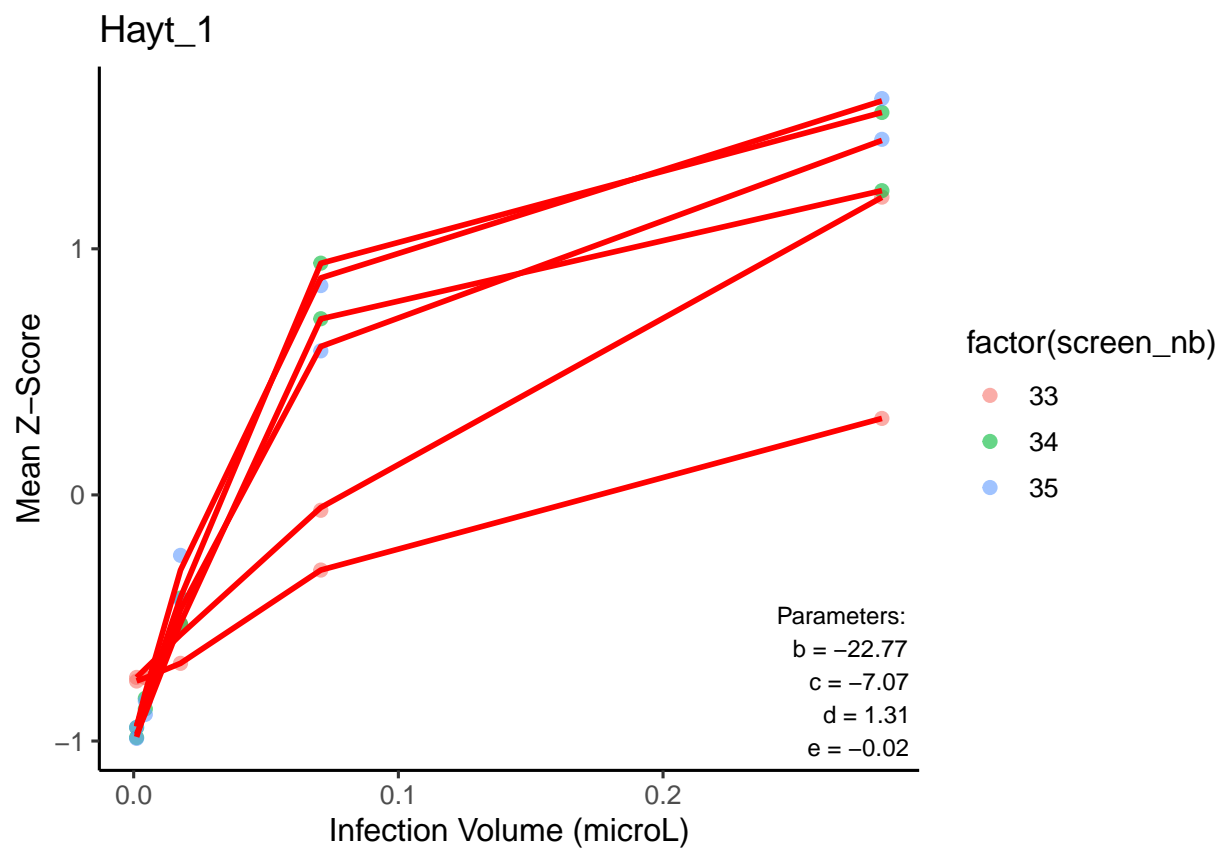
 ## \$Giuf_3



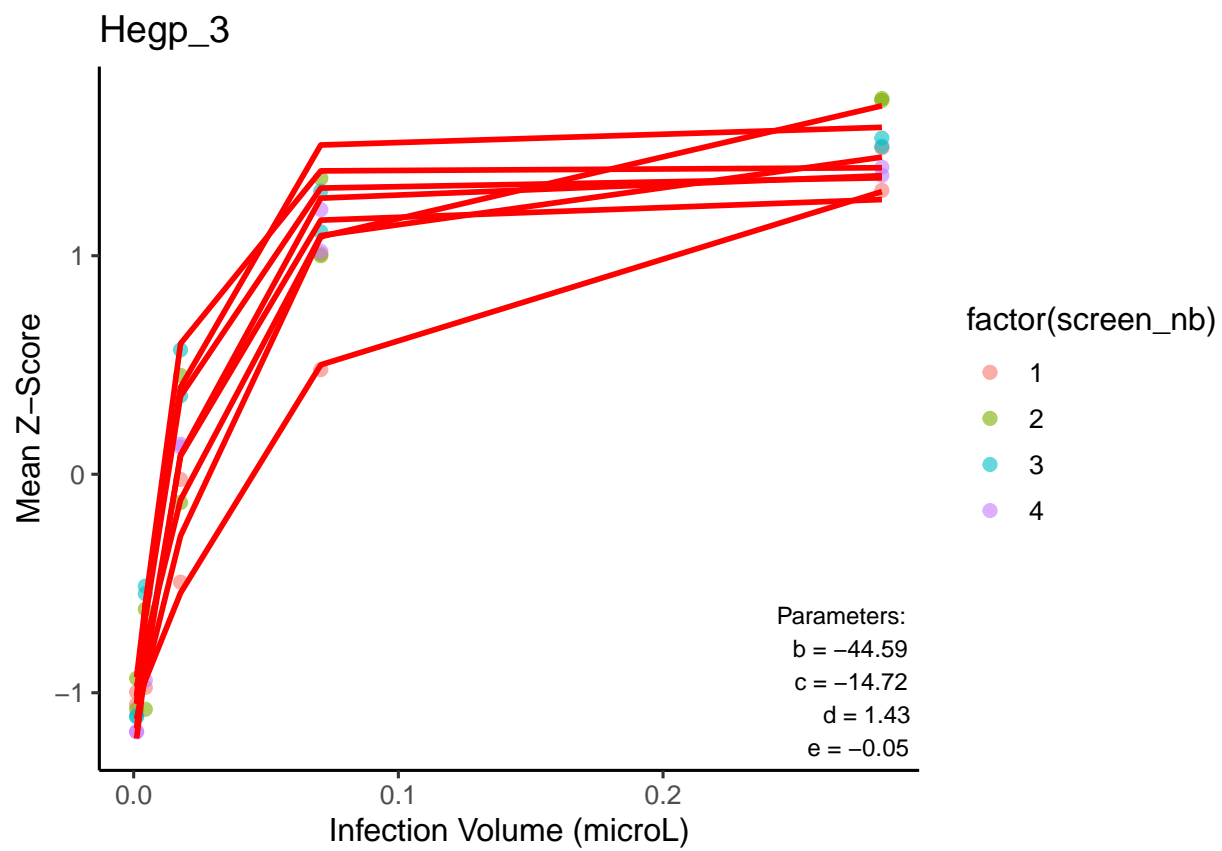
\$Gost_1



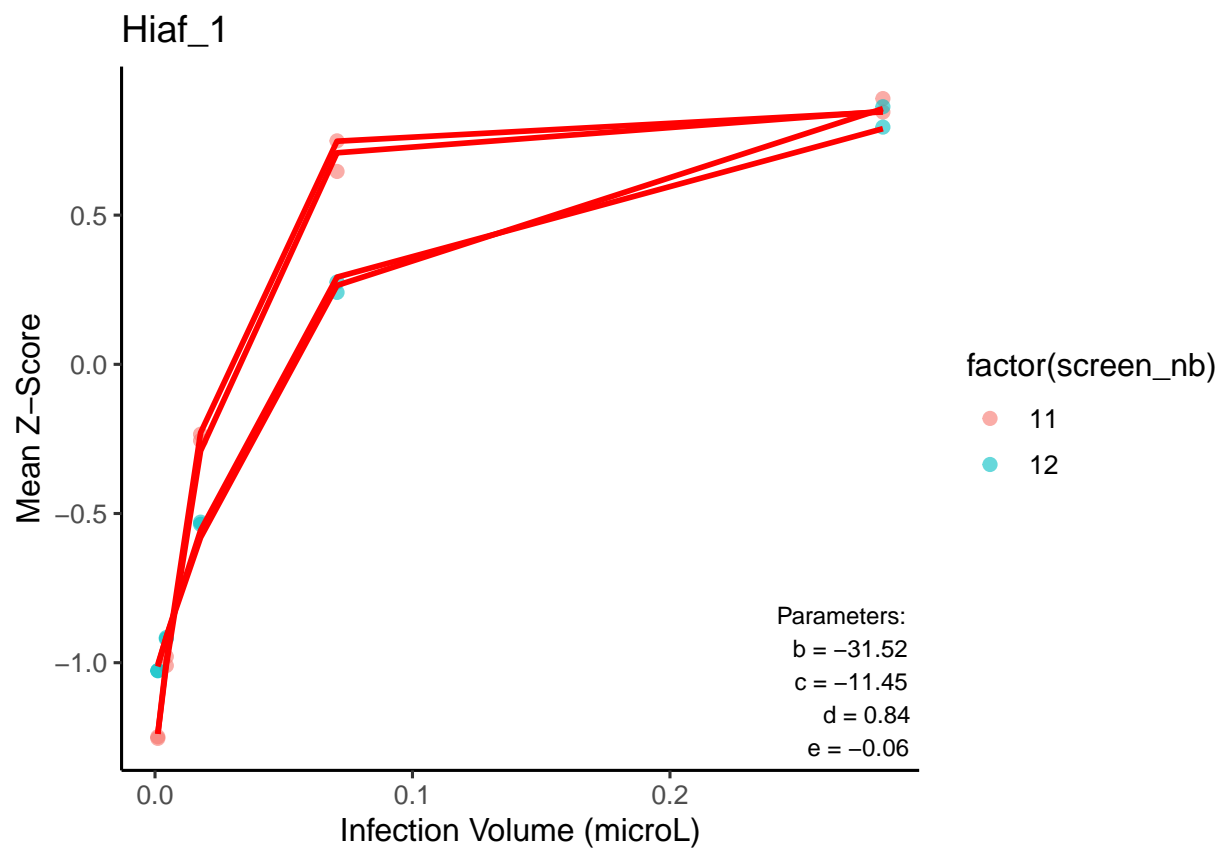
 ## \$Hayt_1



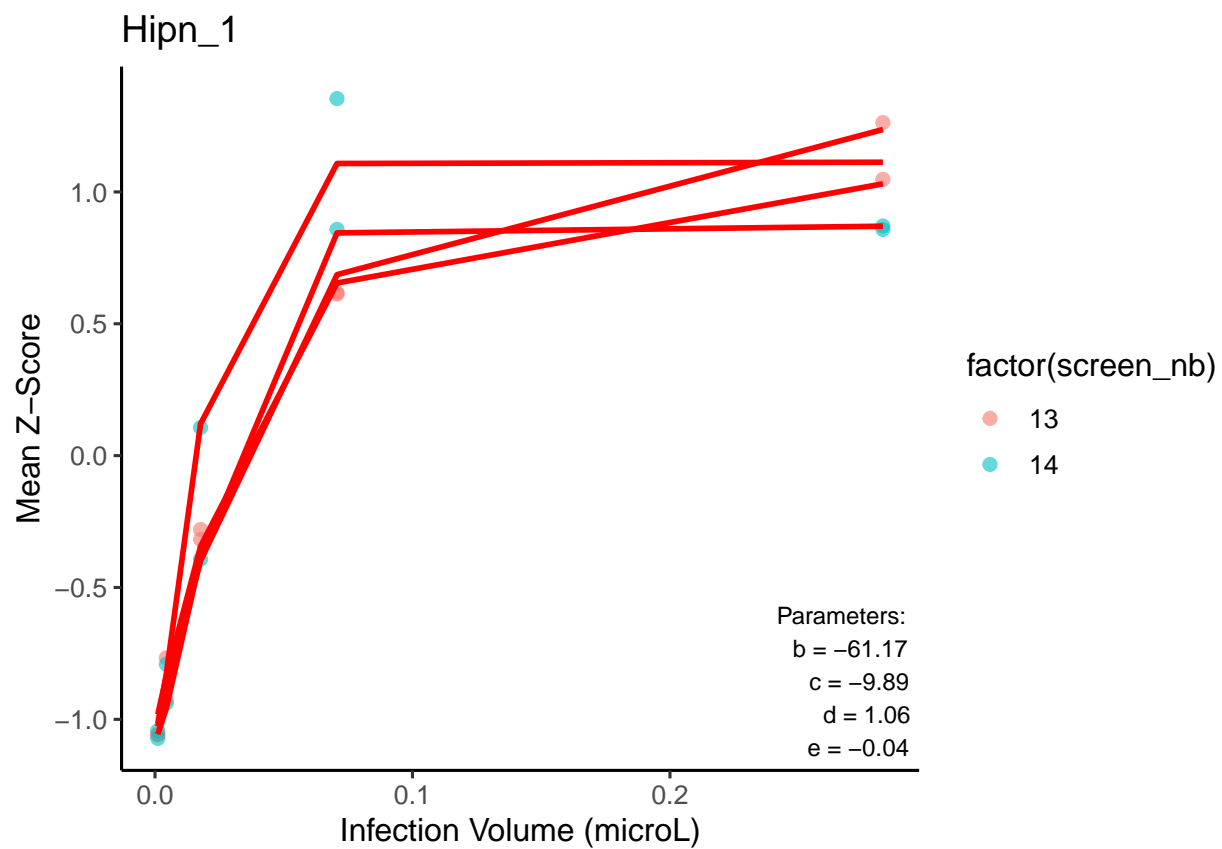
\$Hegp_3



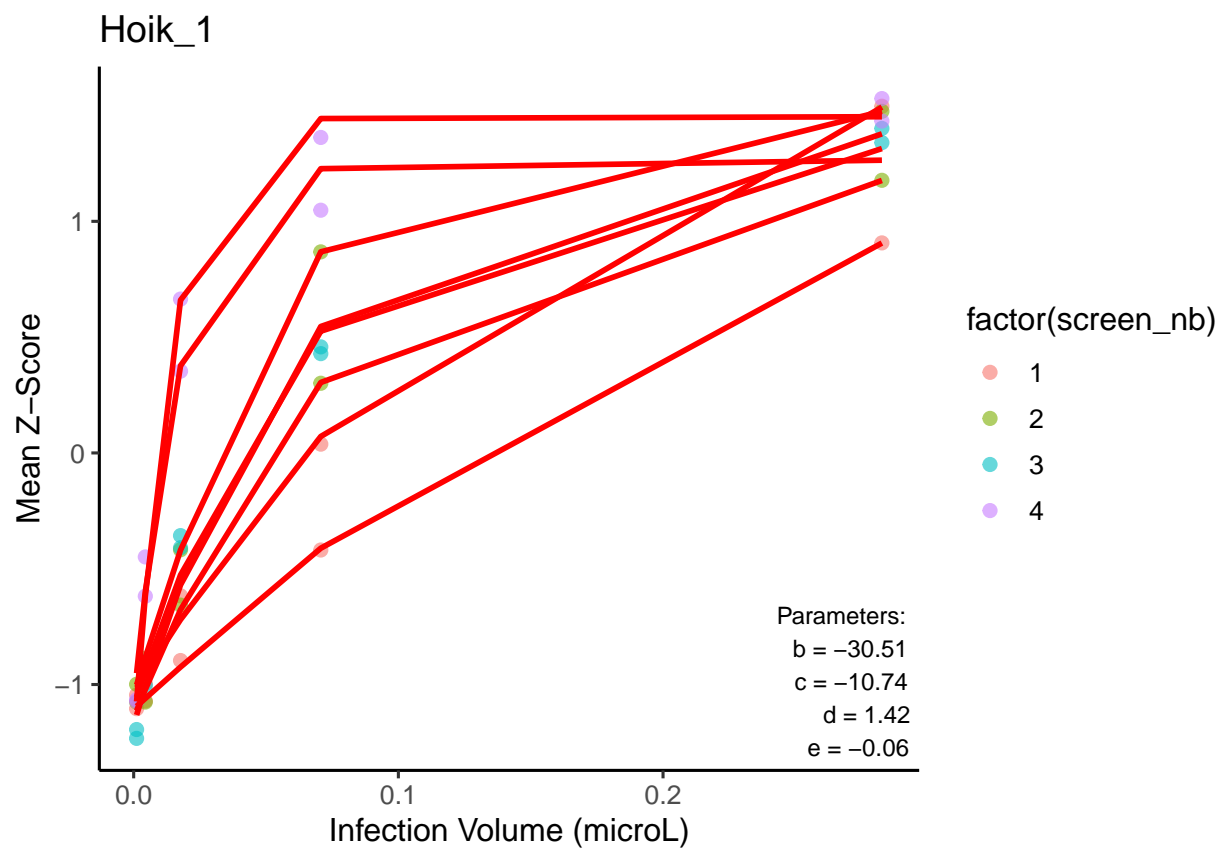
\$Hiaf_1



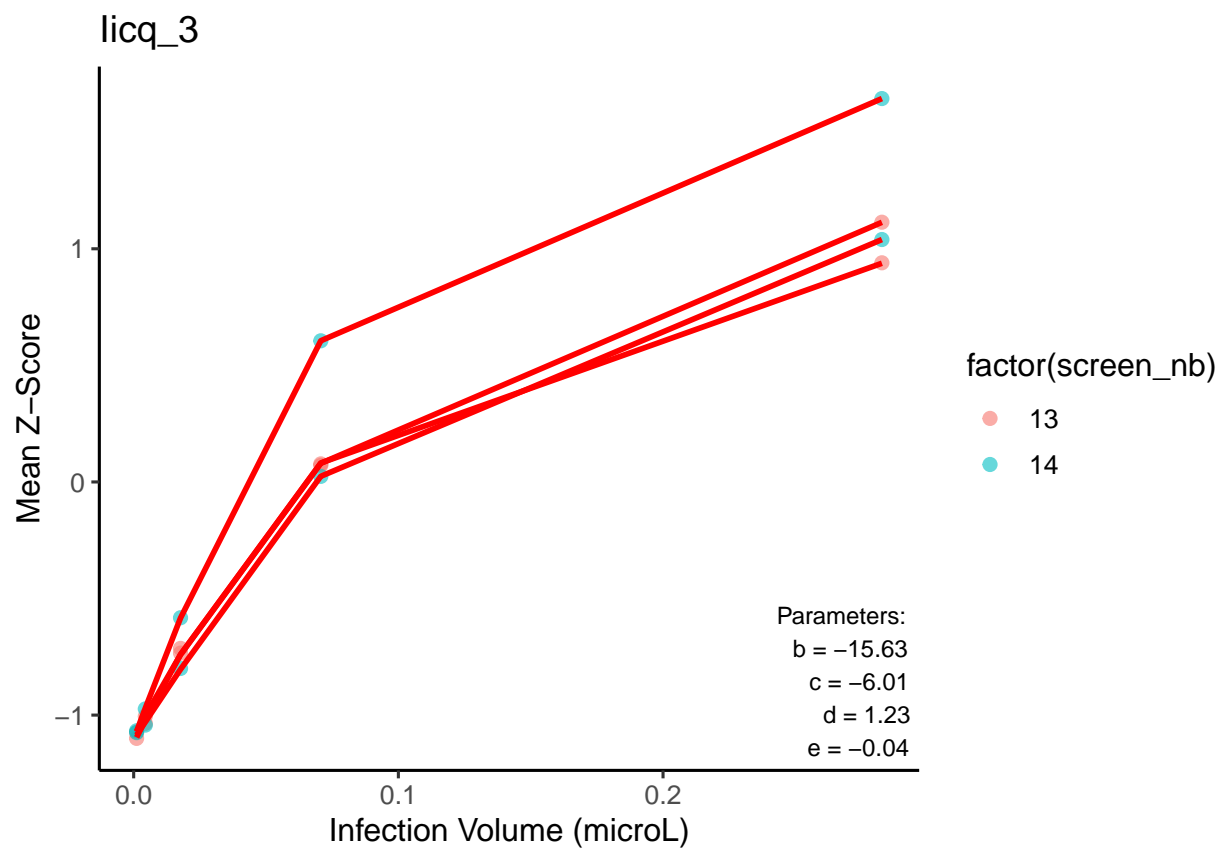
```
##  
## $Hipn_1
```



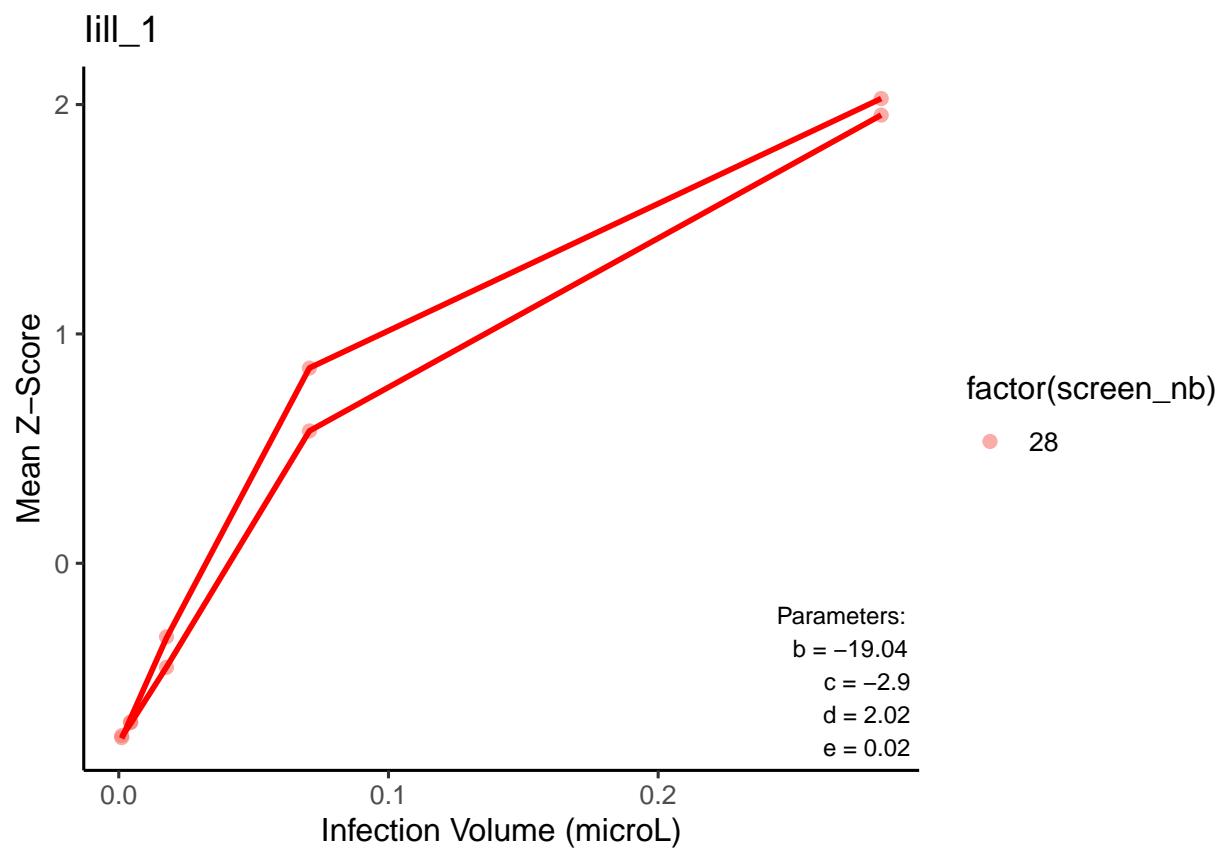
\$Hoik_1



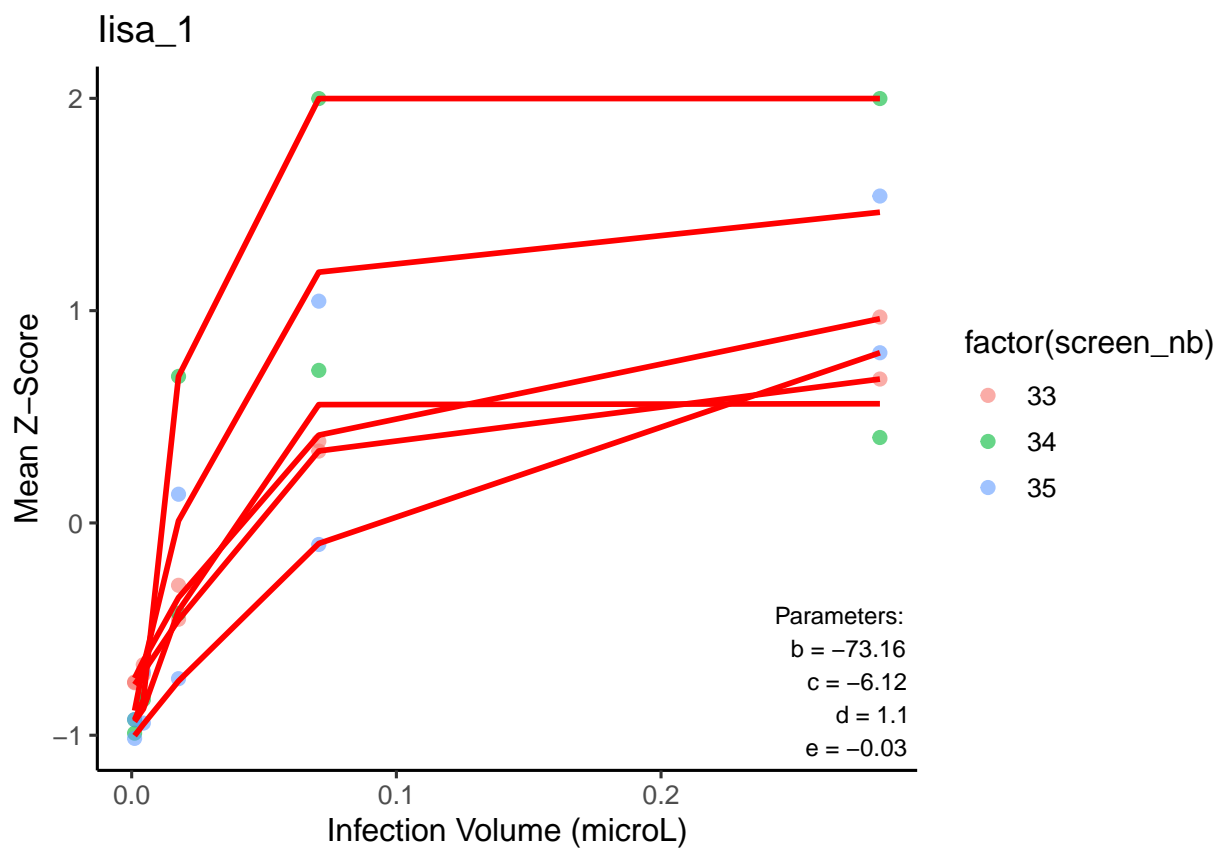
```
##
## $Iicq_3
```



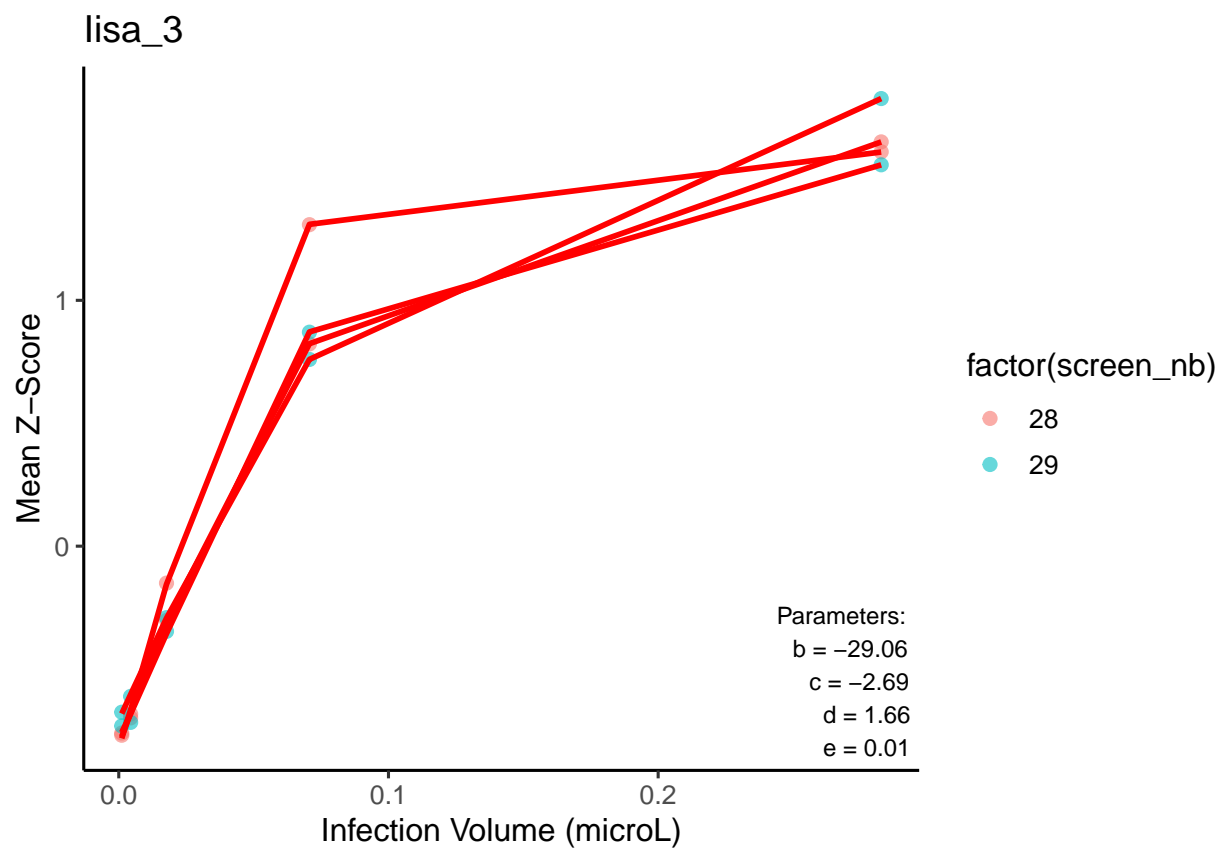
\$Iill_1



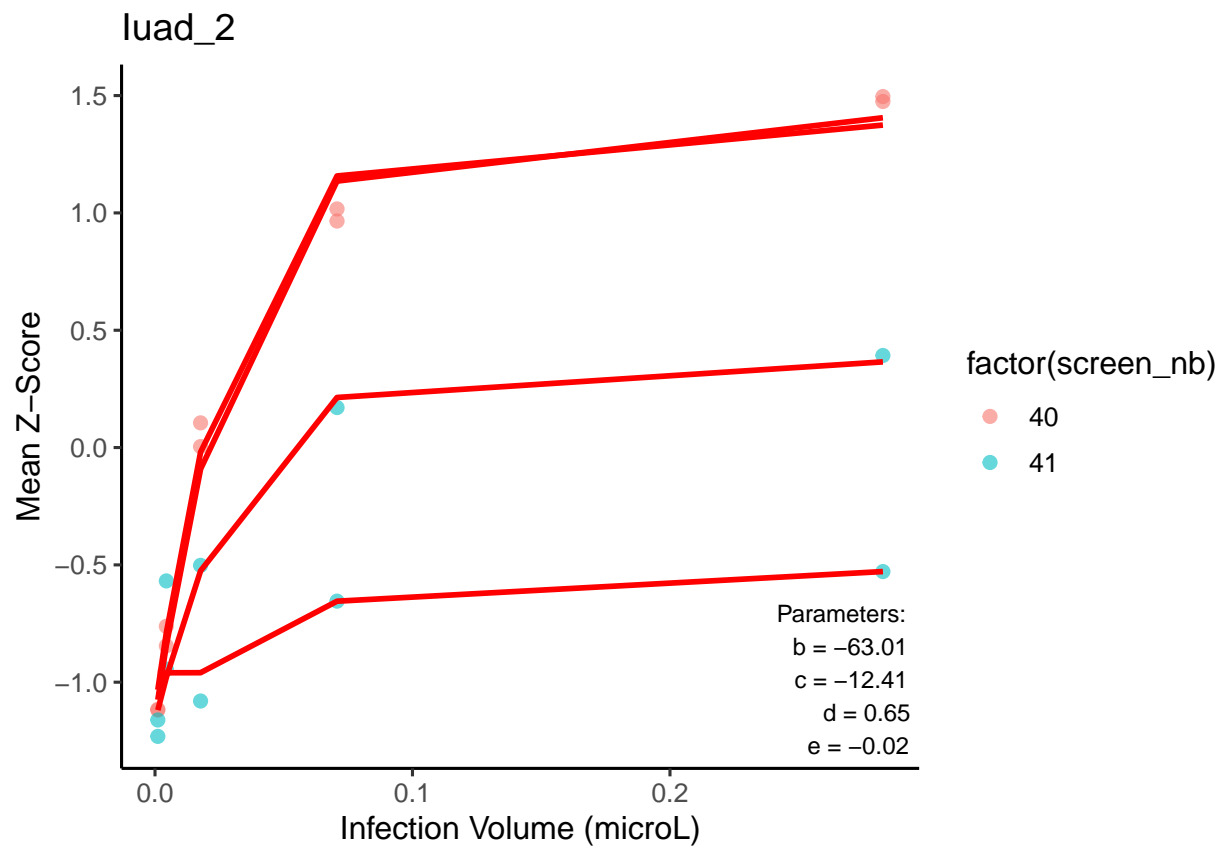
 ## \$Iisa_1



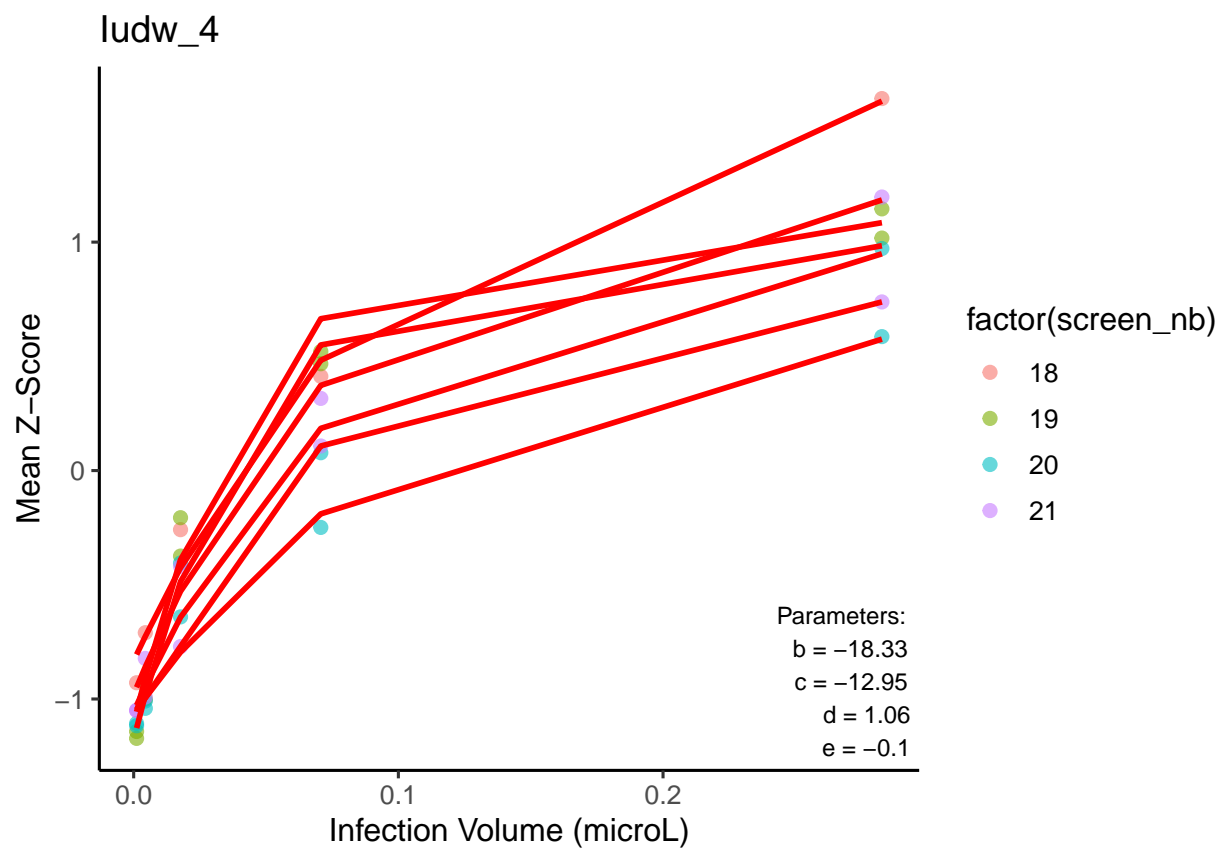
```
##
## $Iisa_3
```



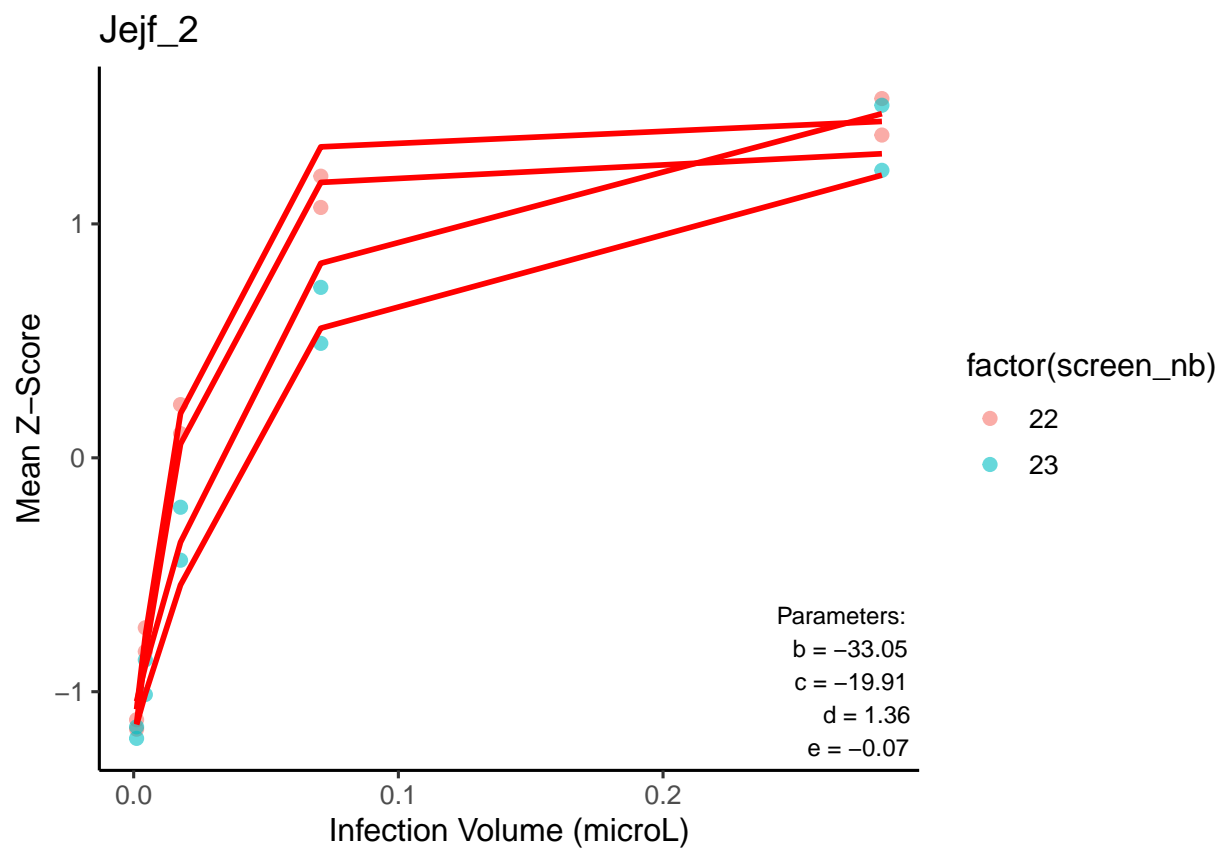
\$Iuad_2



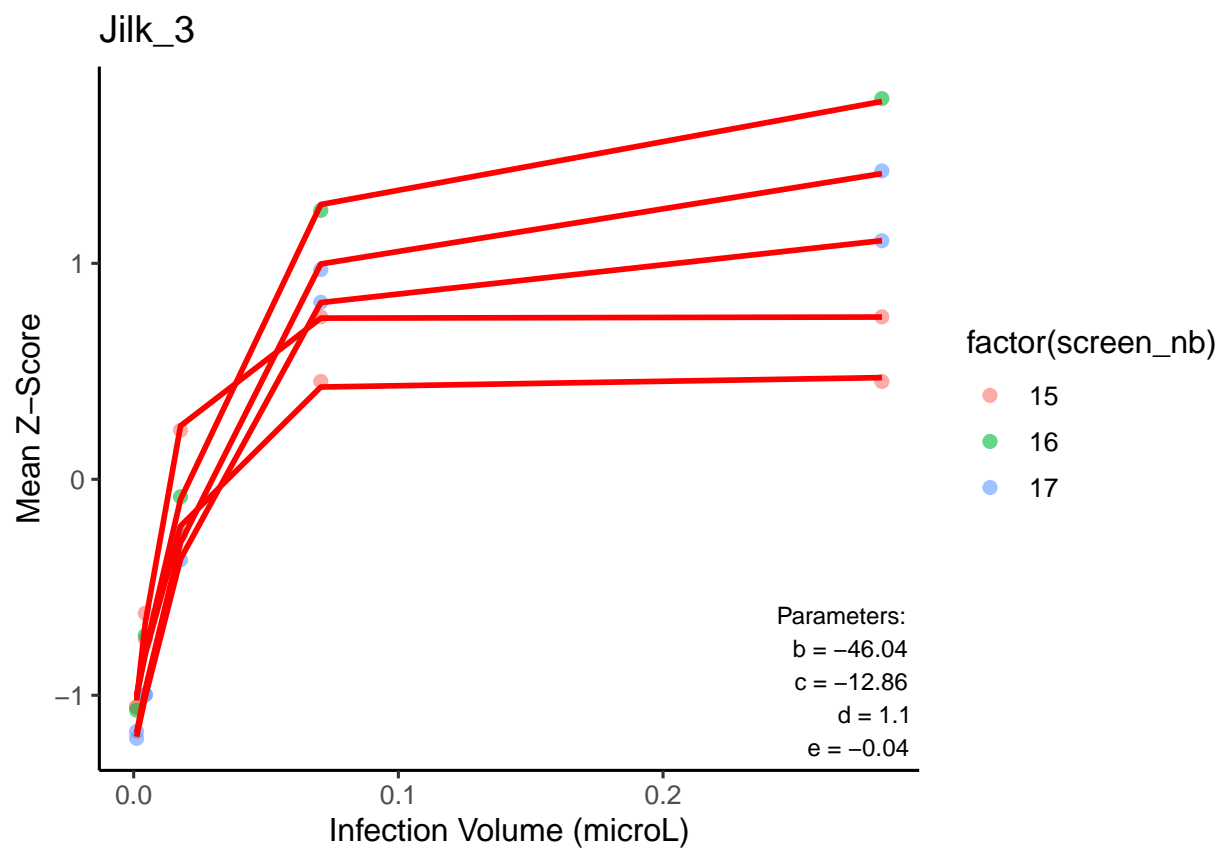
```
##
## $Iudw_4
```



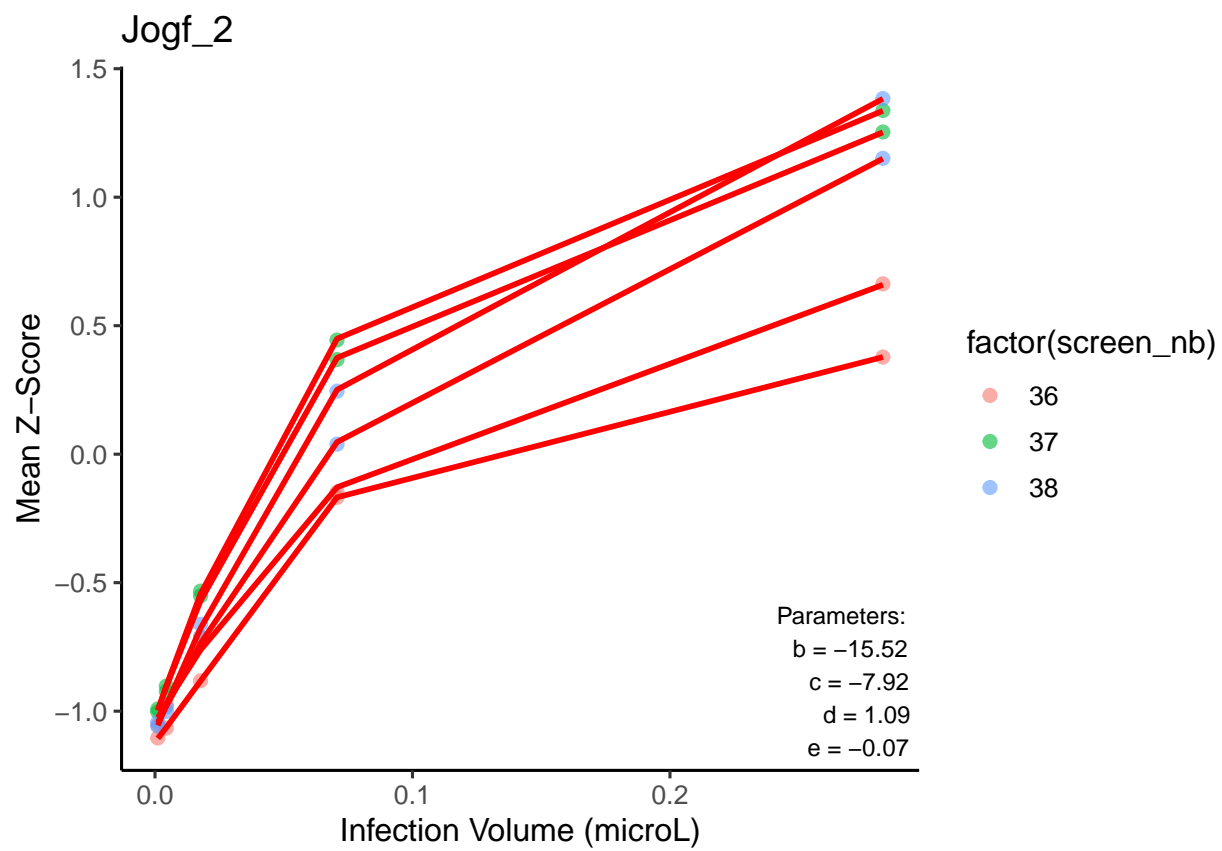
```
##  
## $Jejf_2
```



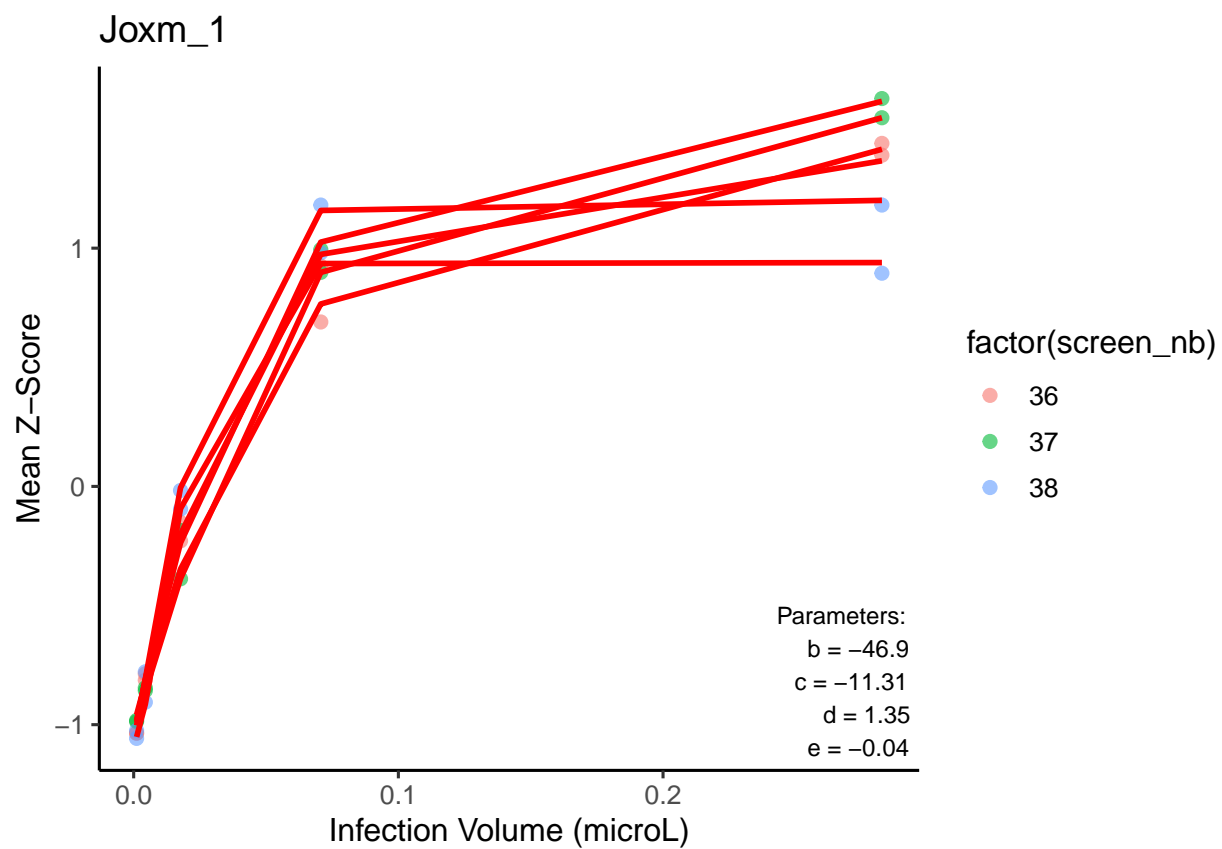
\$Jilk_3



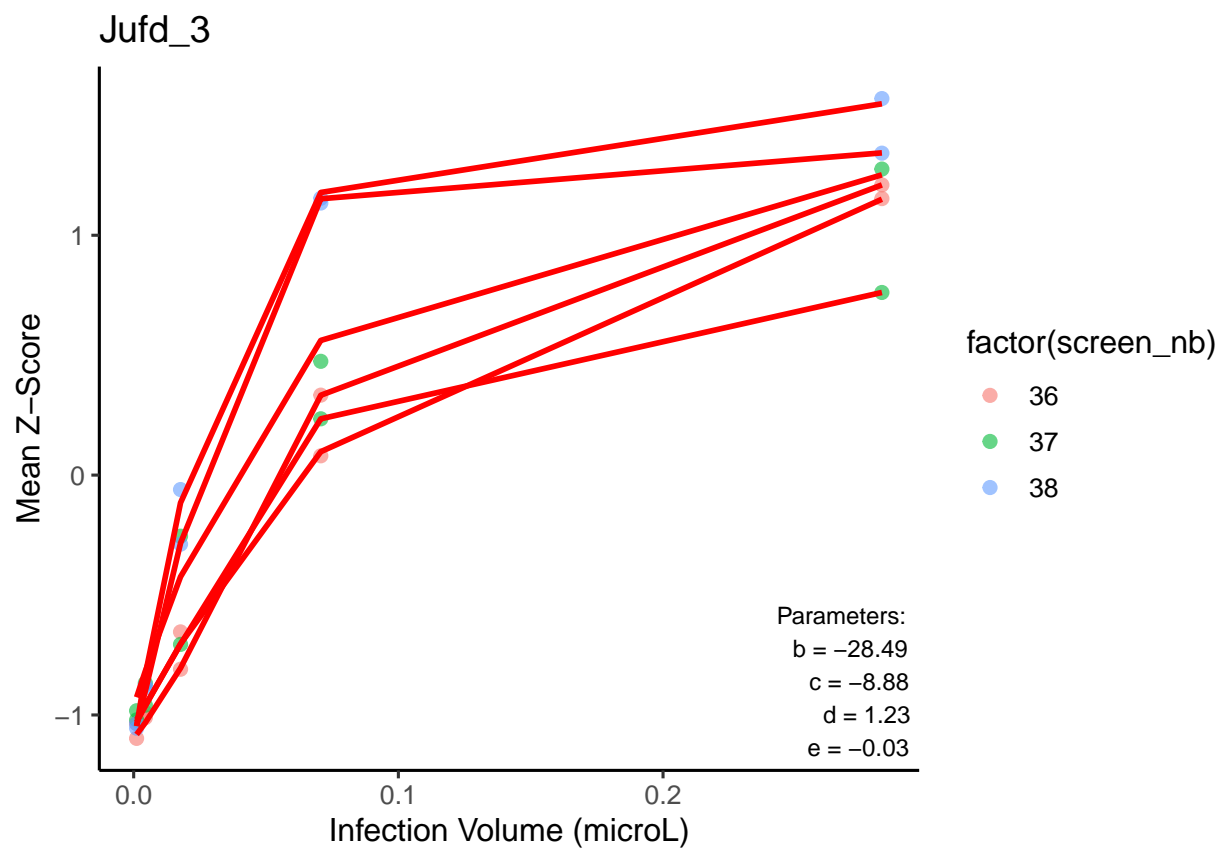
 ## \$Jogf_2



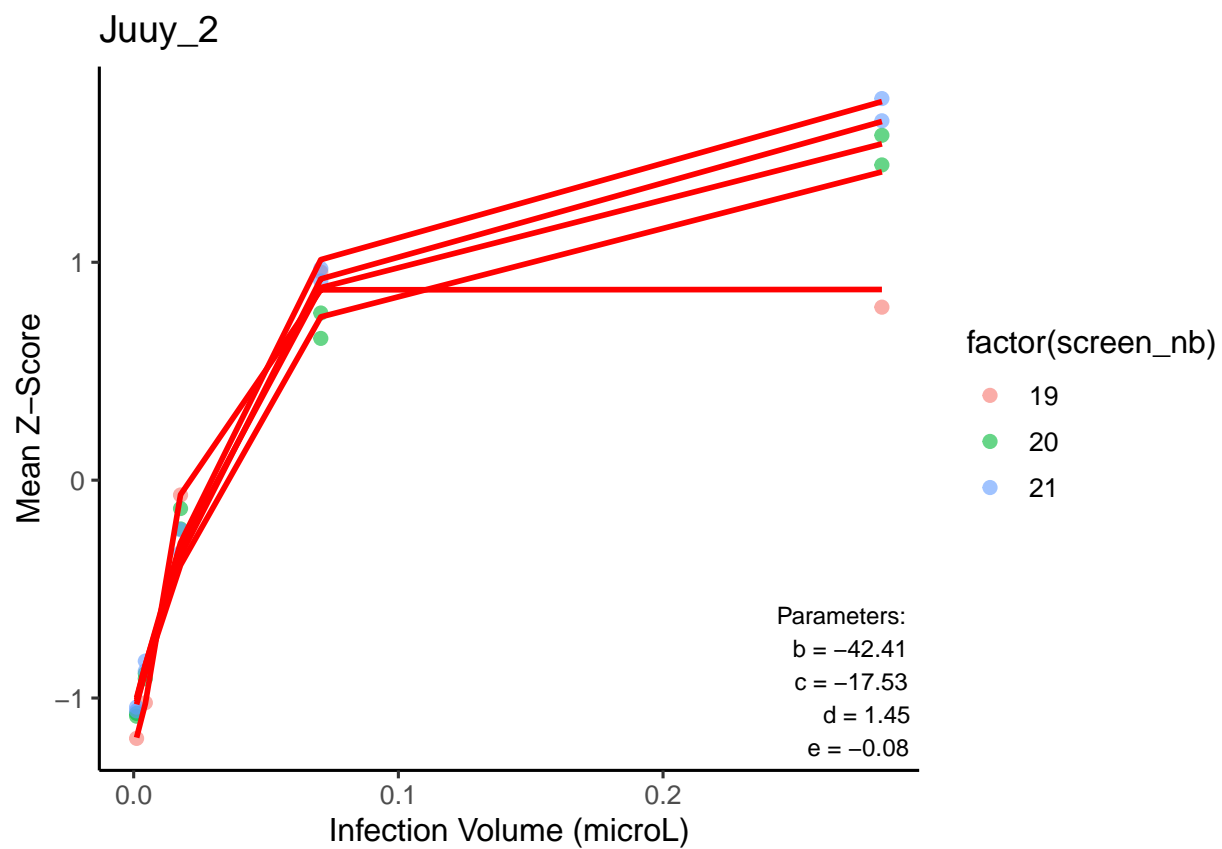
 ## \$Joxm_1



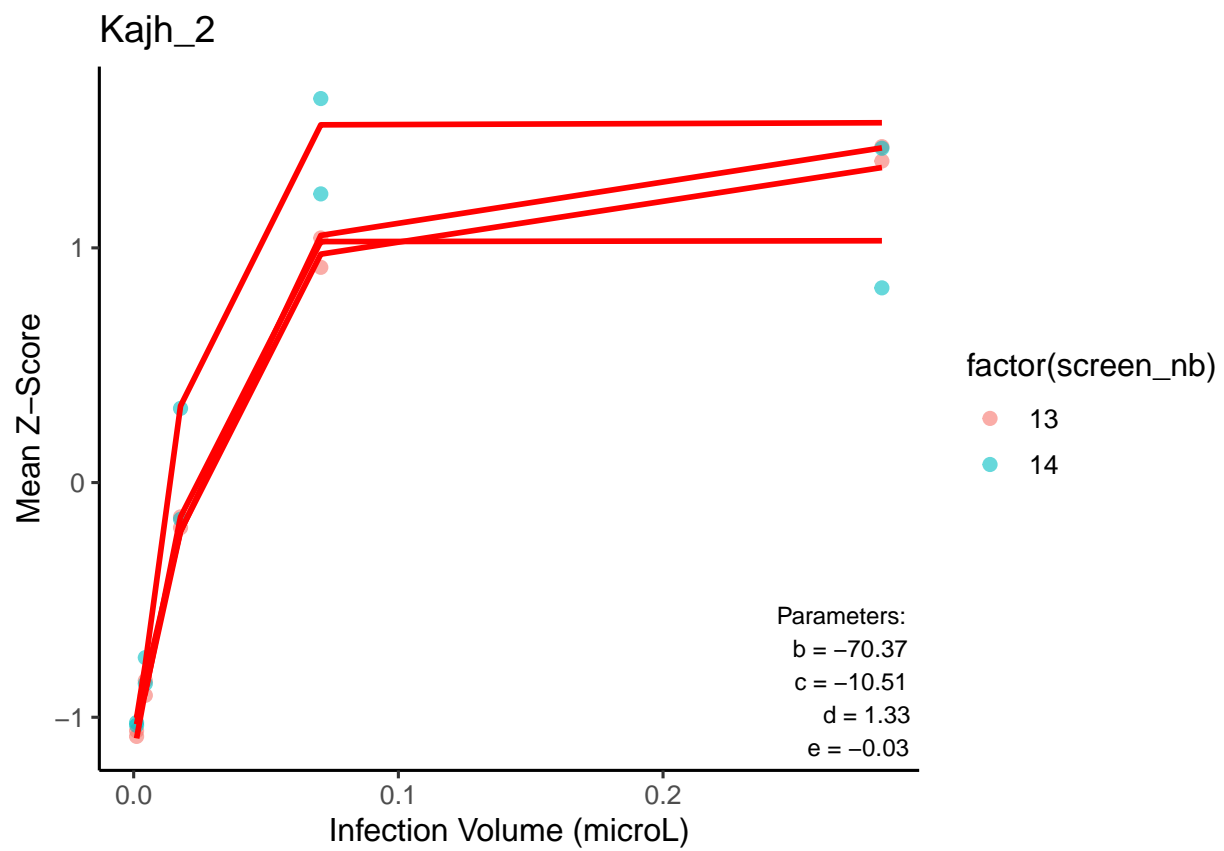
 ## \$Jufd_3



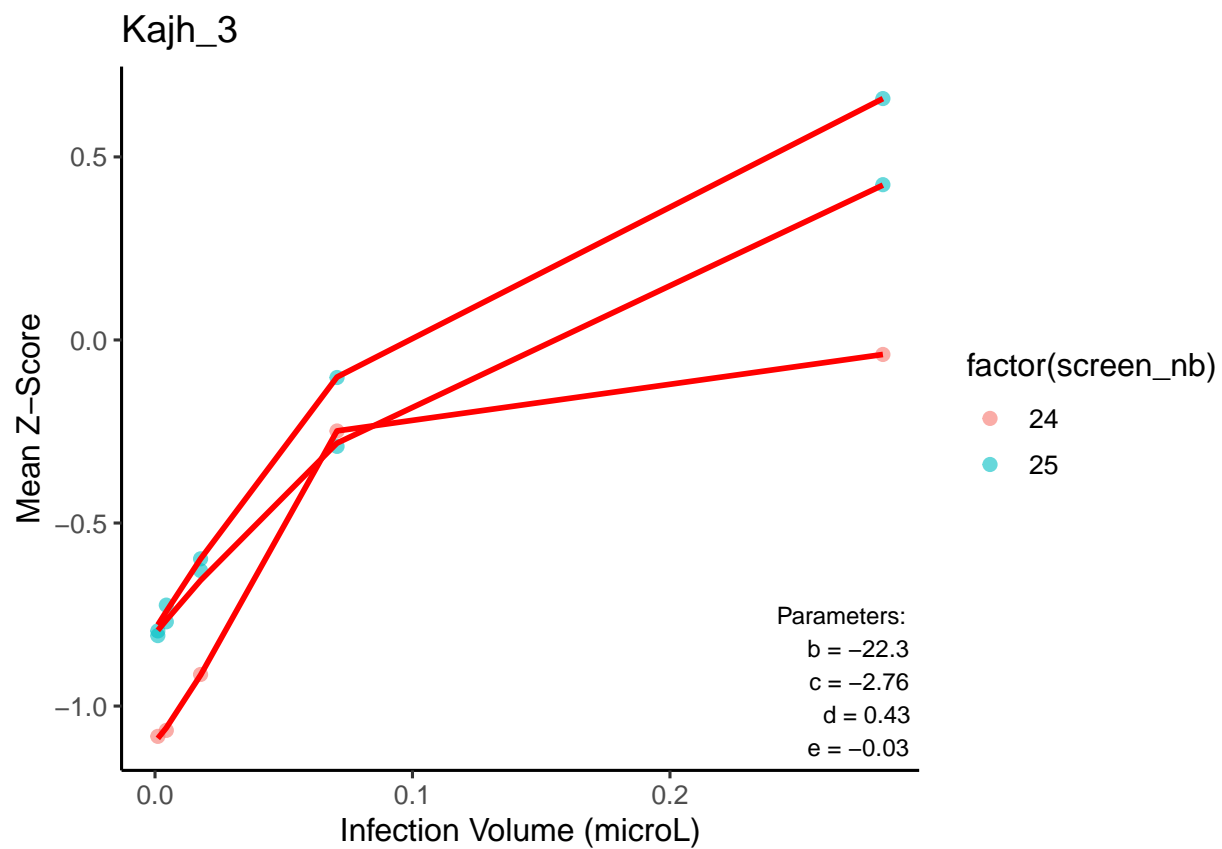
 ## \$Juuy_2



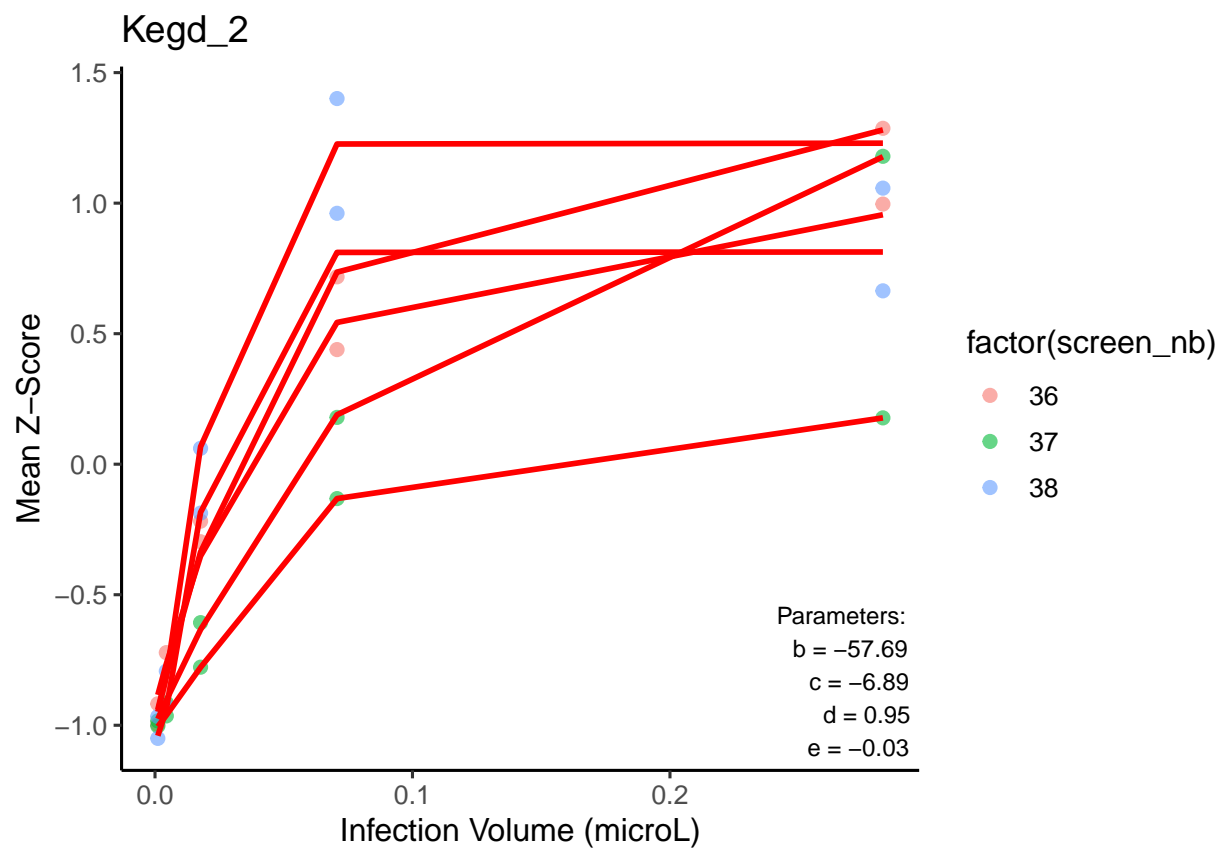
\$Kajh_2



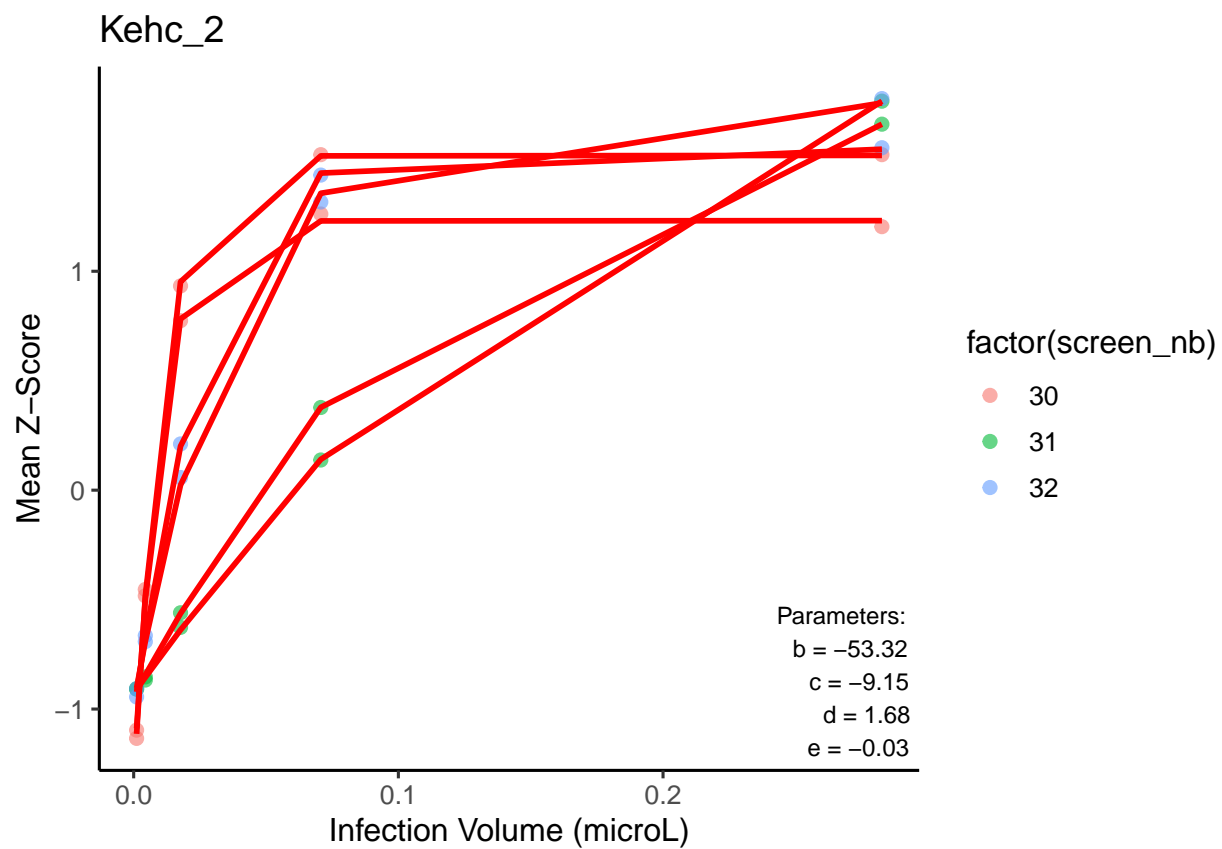
 ## \$Kajh_3



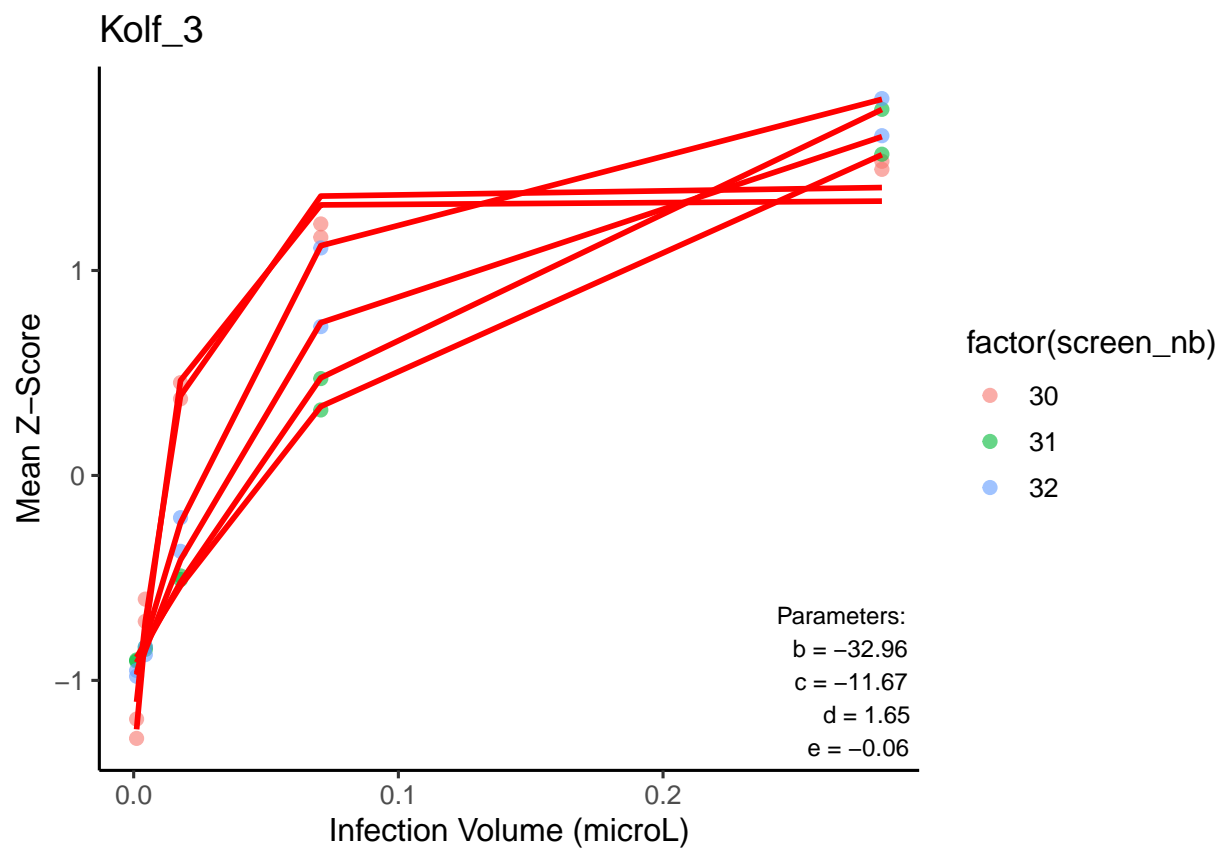
\$Kegd_2



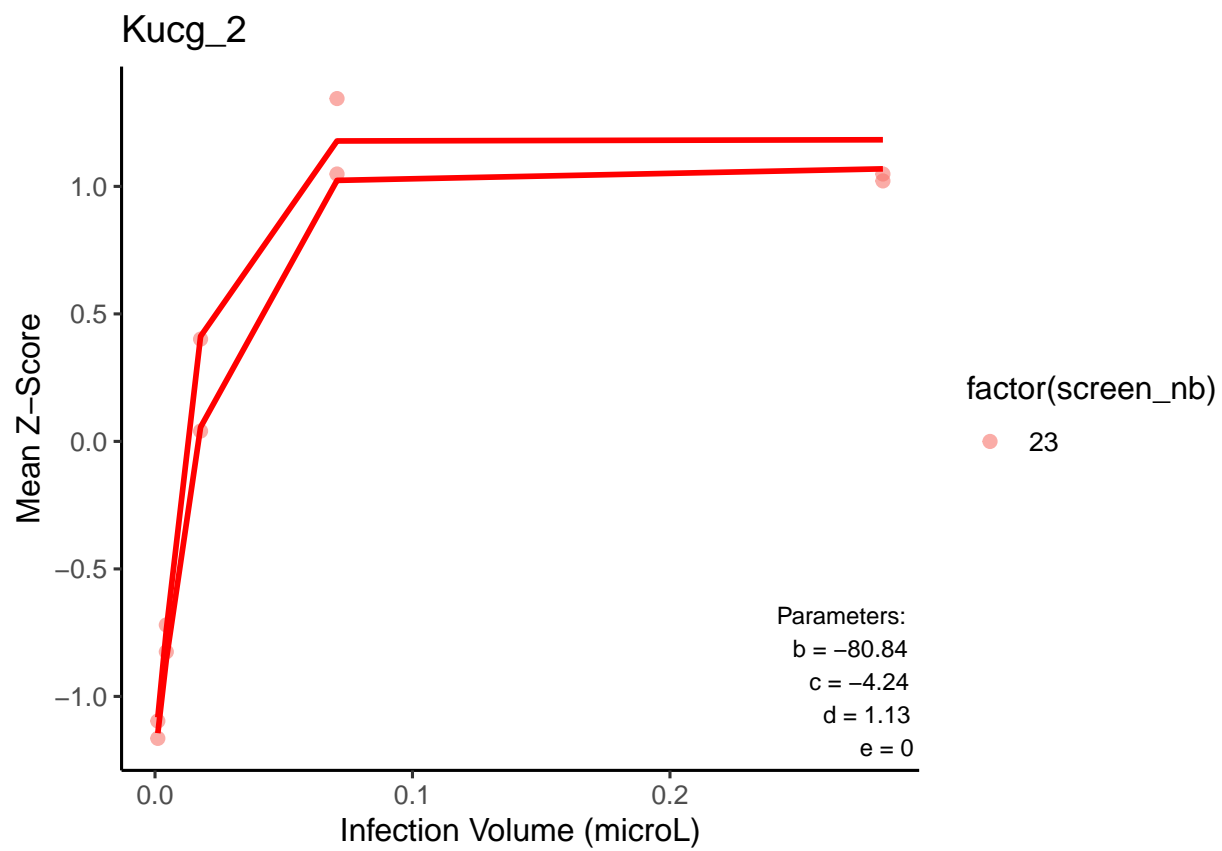
 ## \$Kehc_2



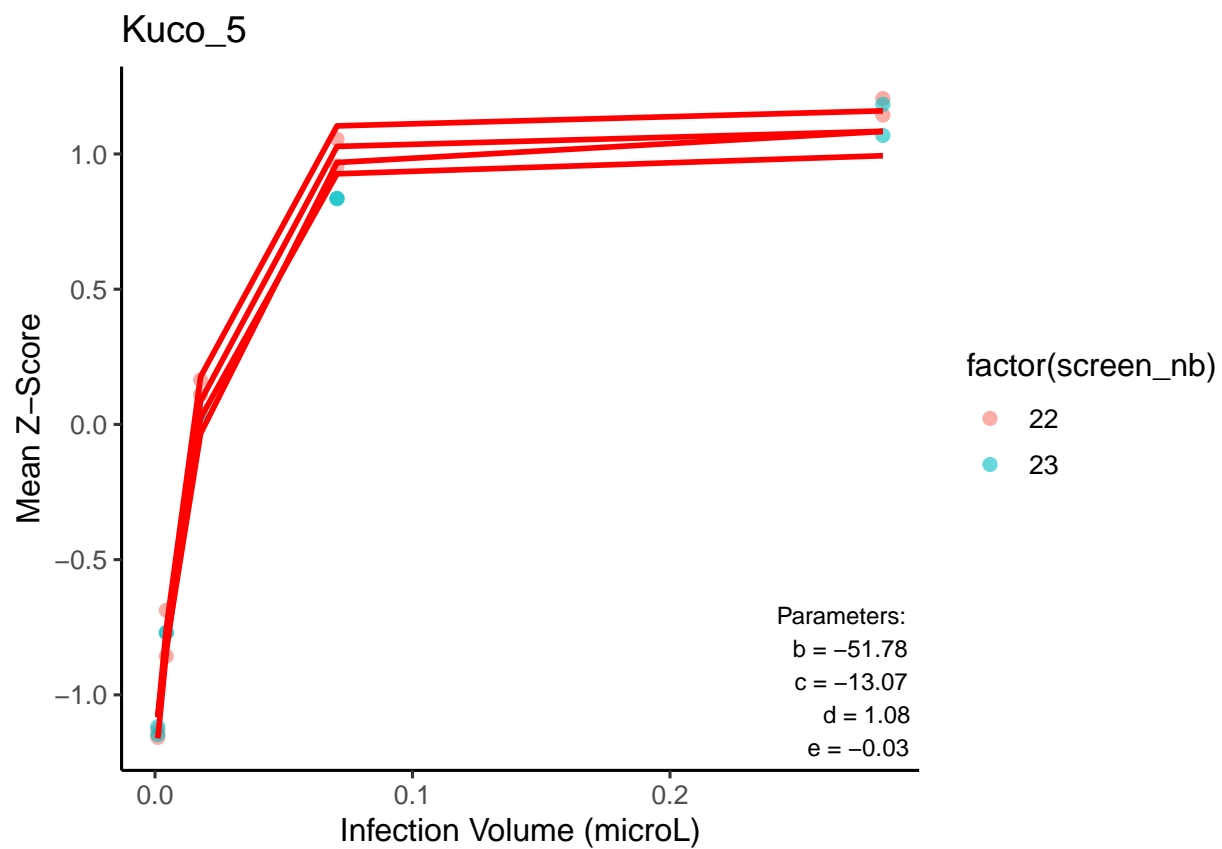
```
##  
## $Kolf_3
```



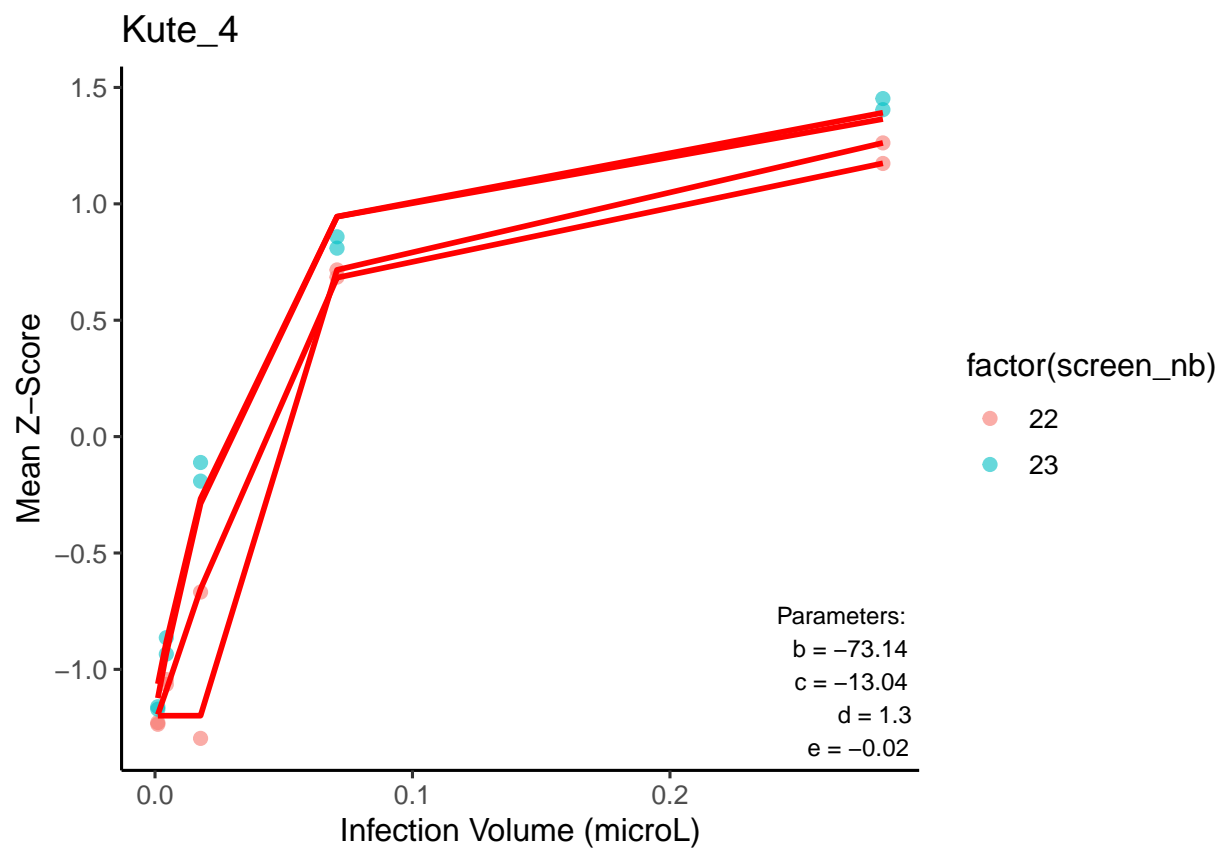
 ## \$Kucg_2



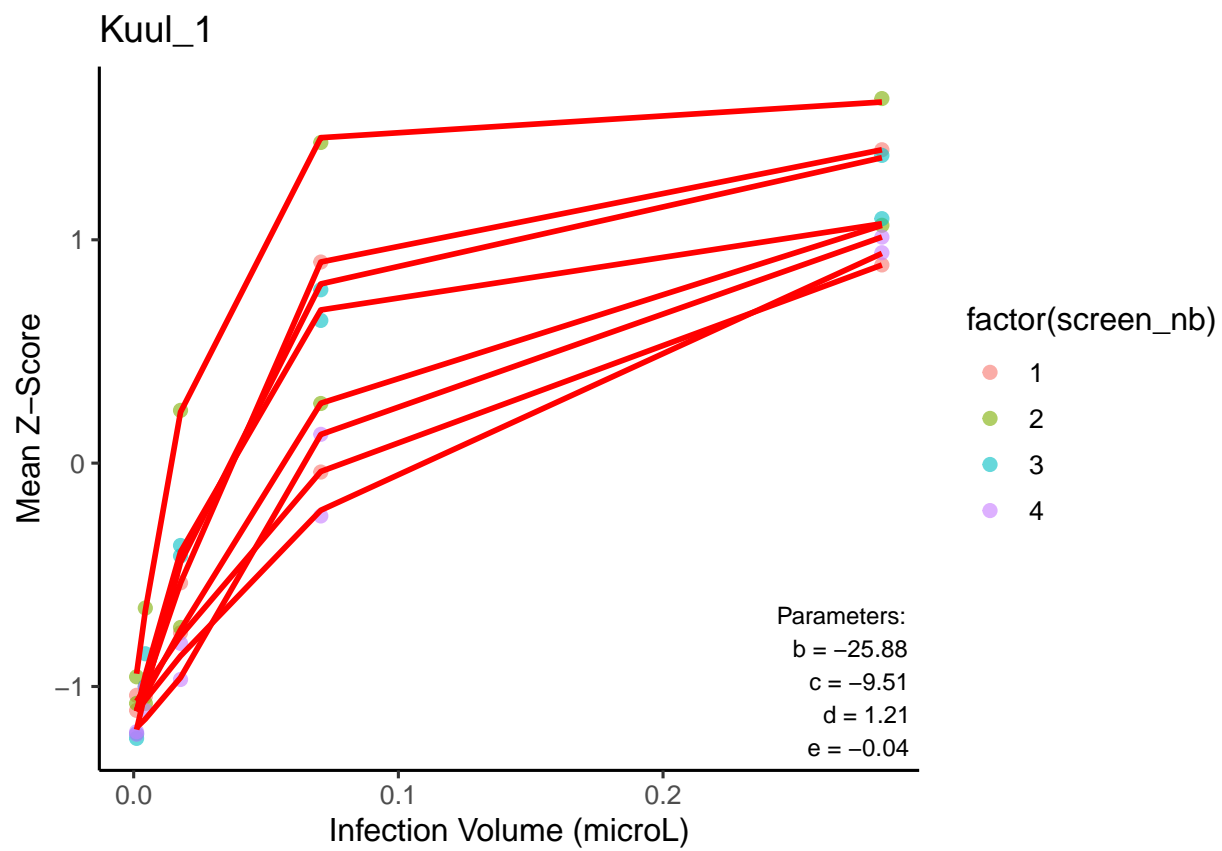
\$Kuco_5



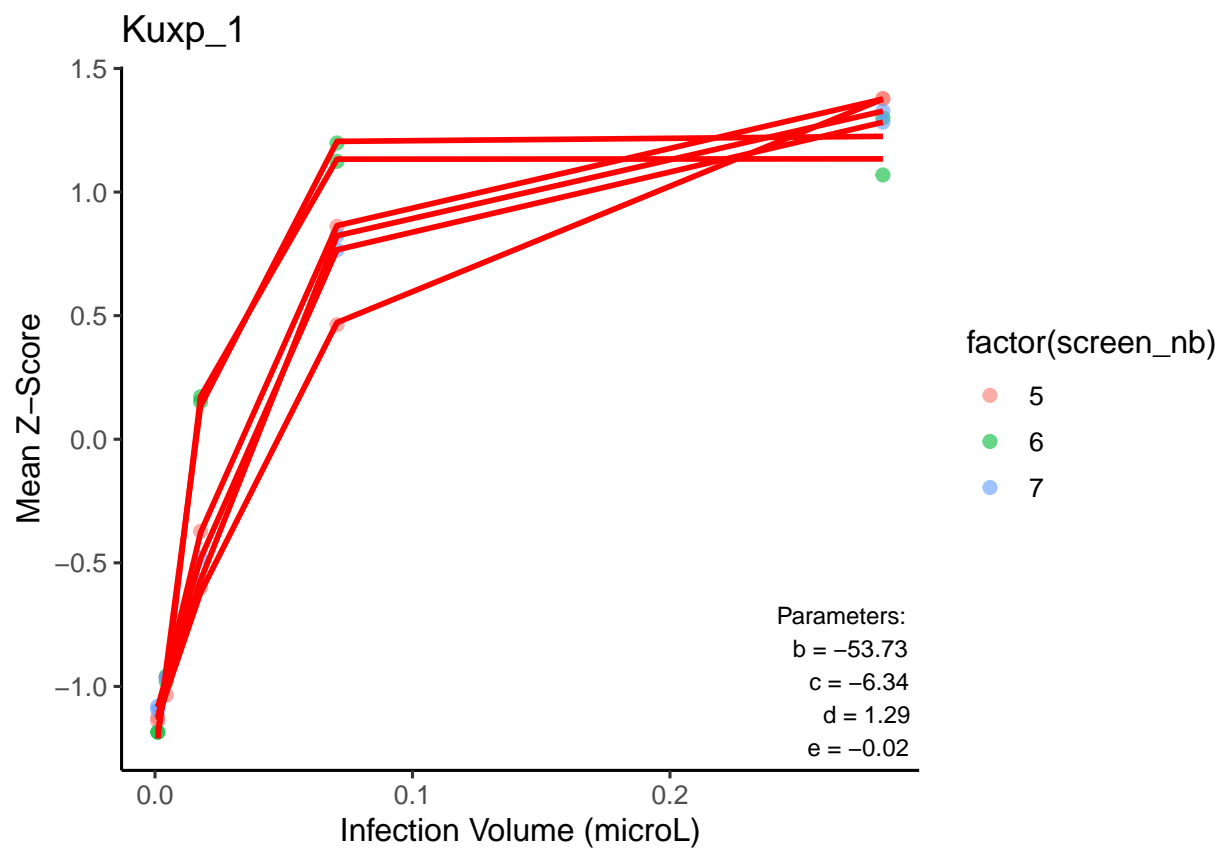
\$Kute_4



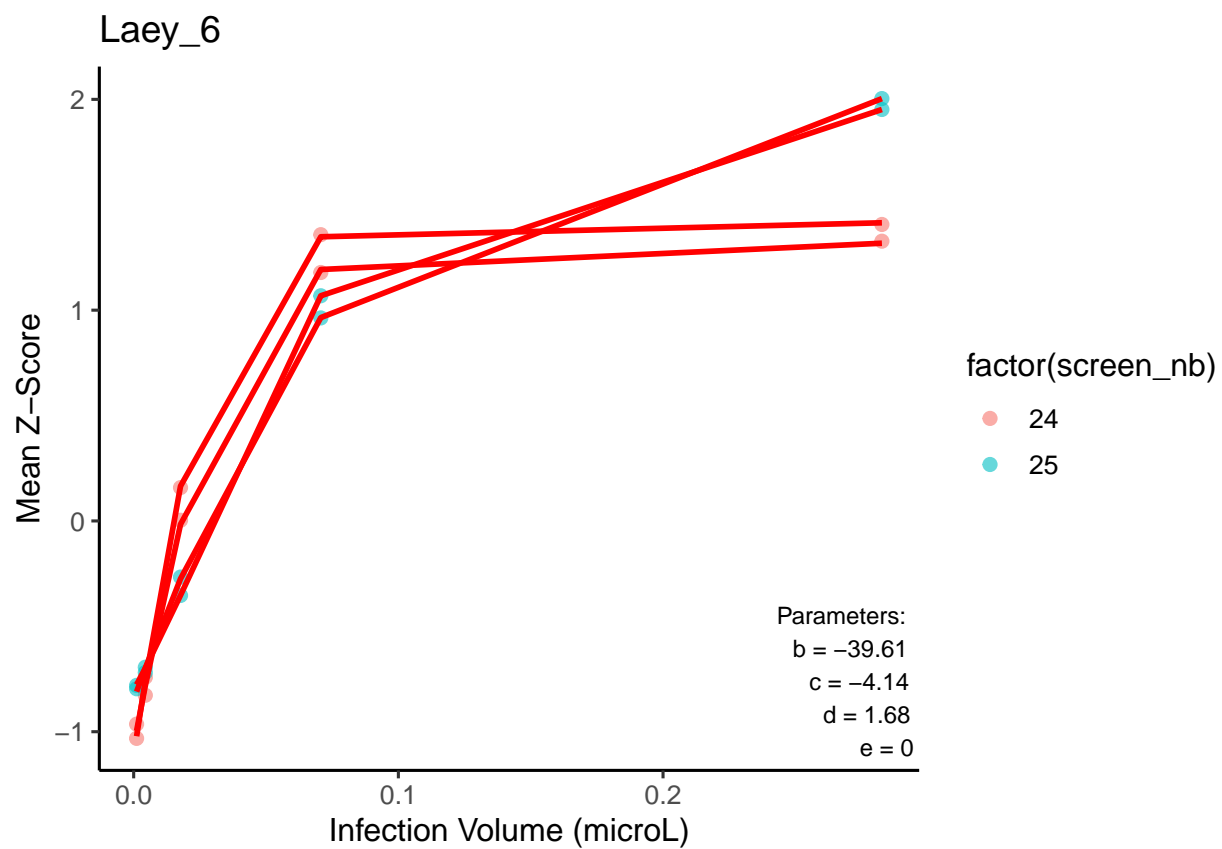
\$Kuul_1



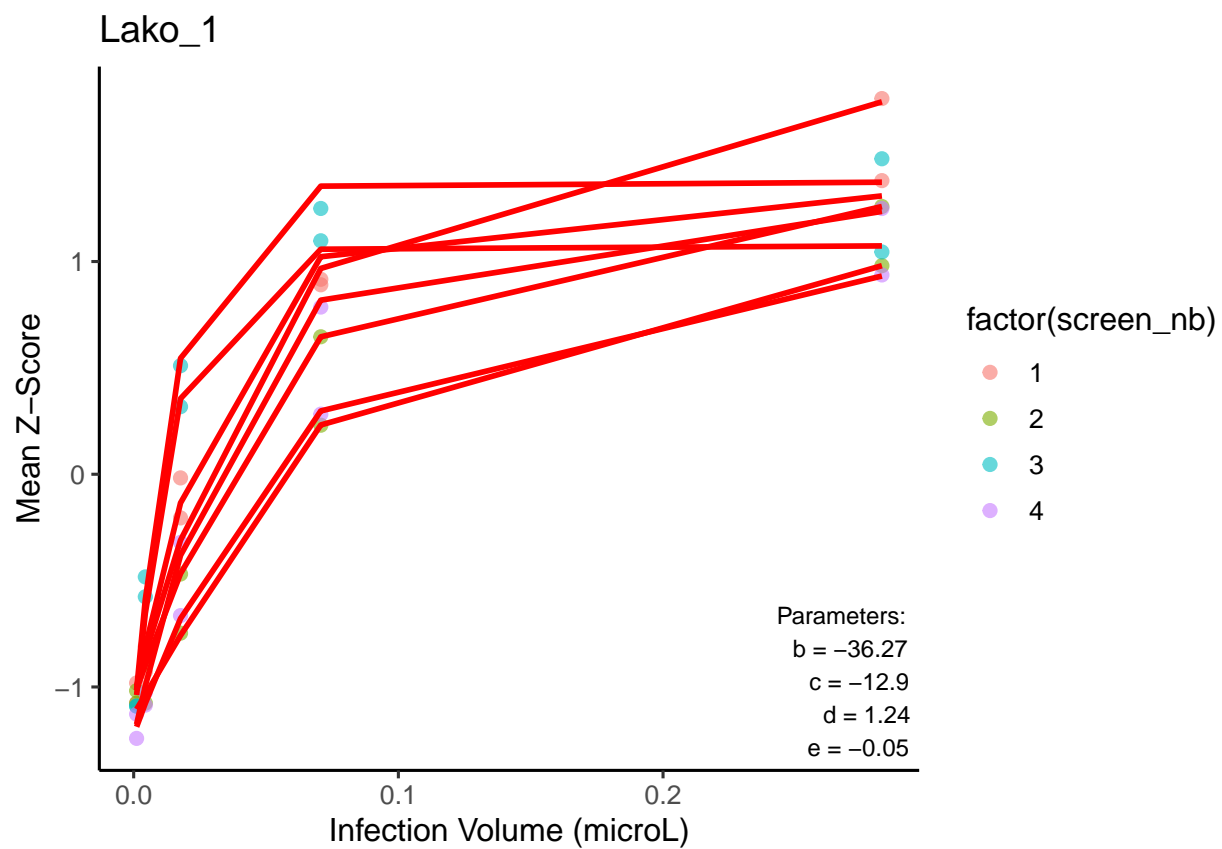
 ## \$Kuxp_1



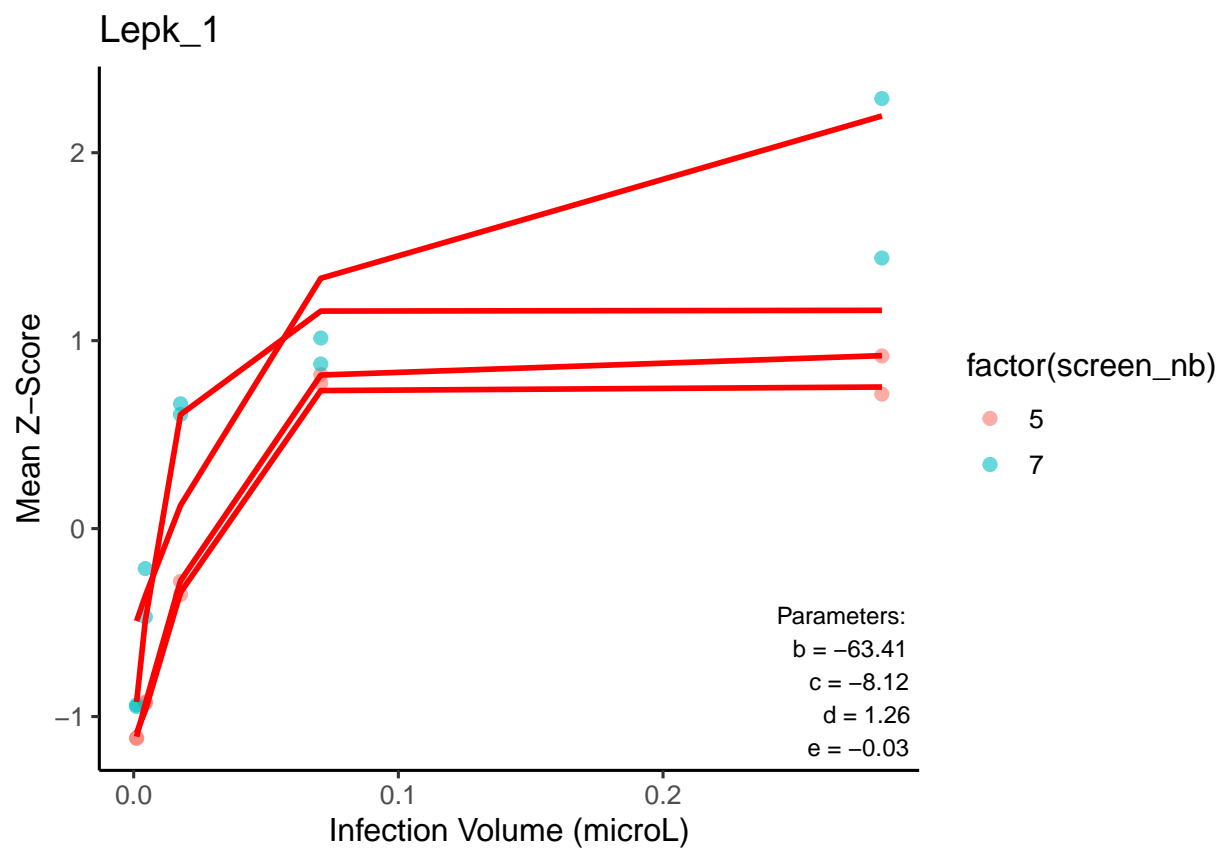
\$Laey_6



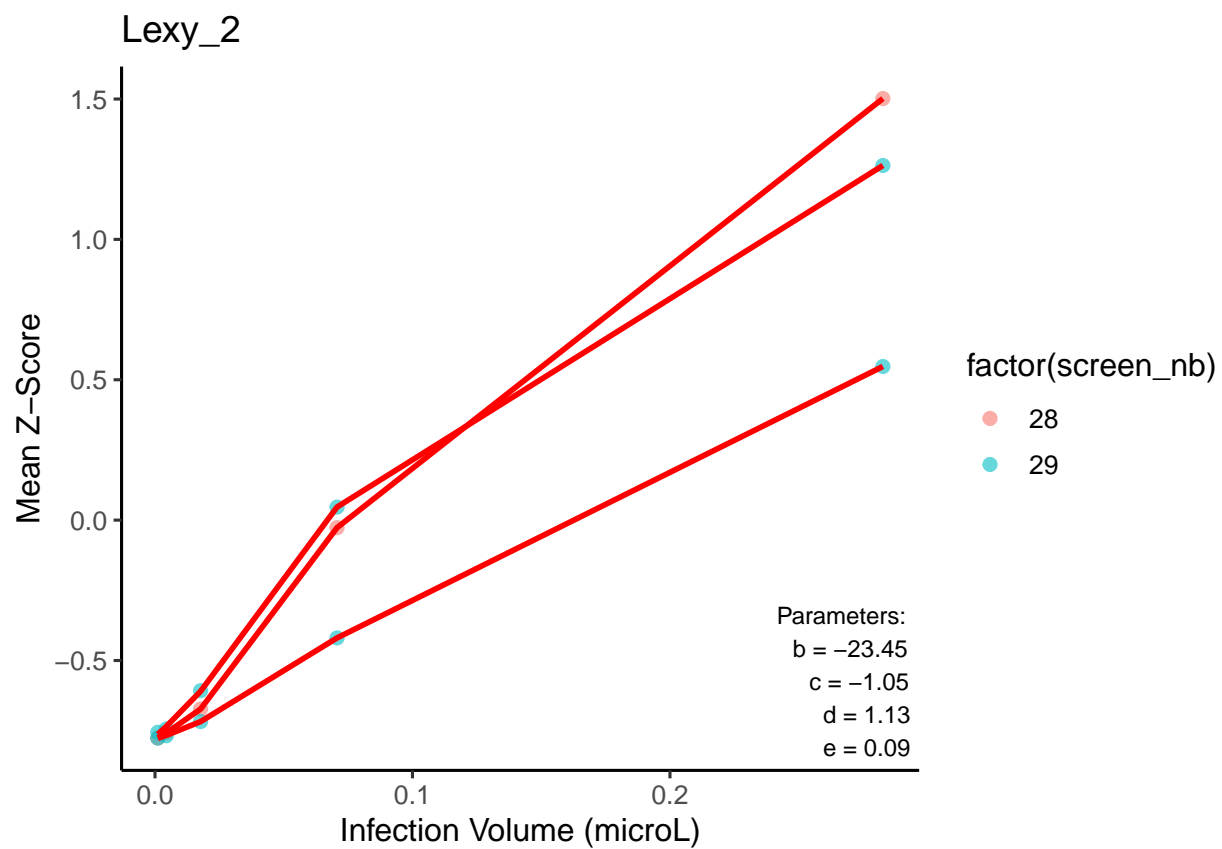
 ## \$Lako_1



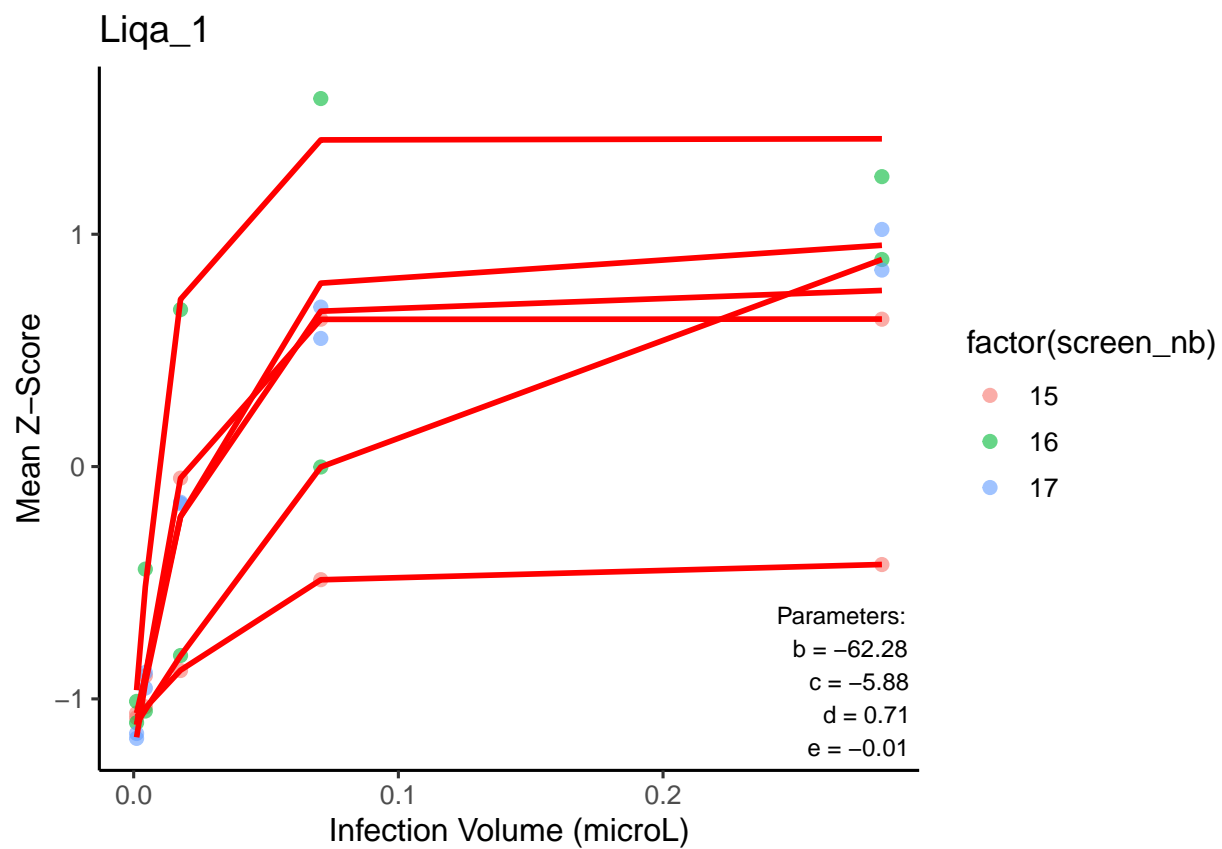
 ## \$Lepk_1



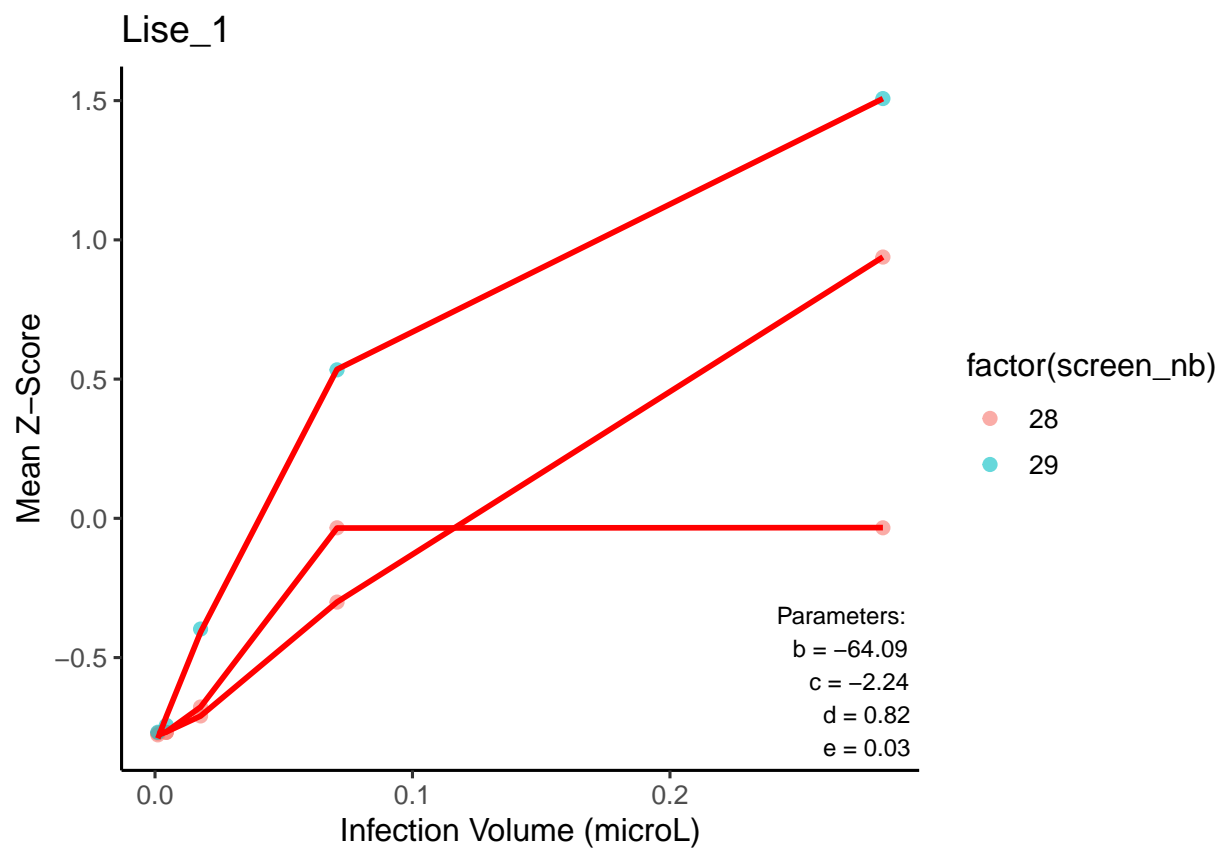
 ## \$Lexy_2



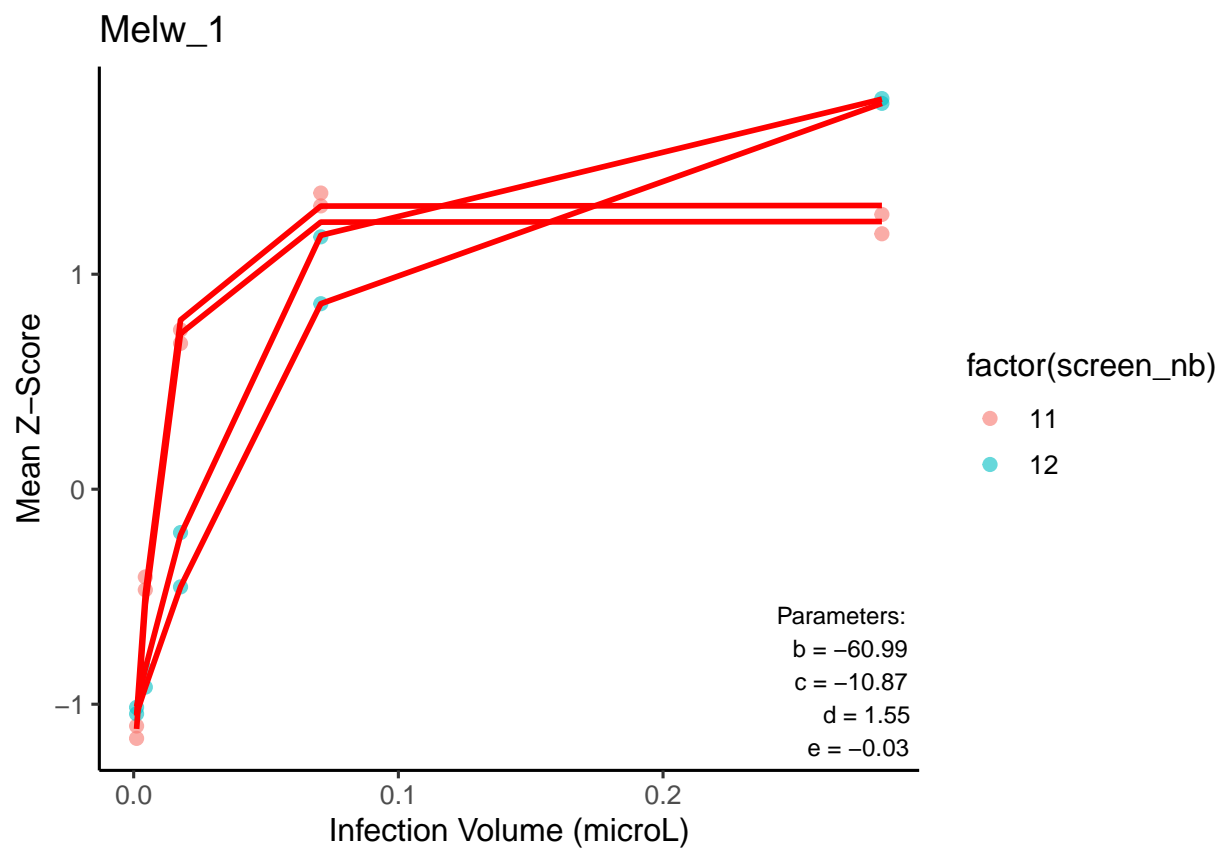
\$Liqa_1



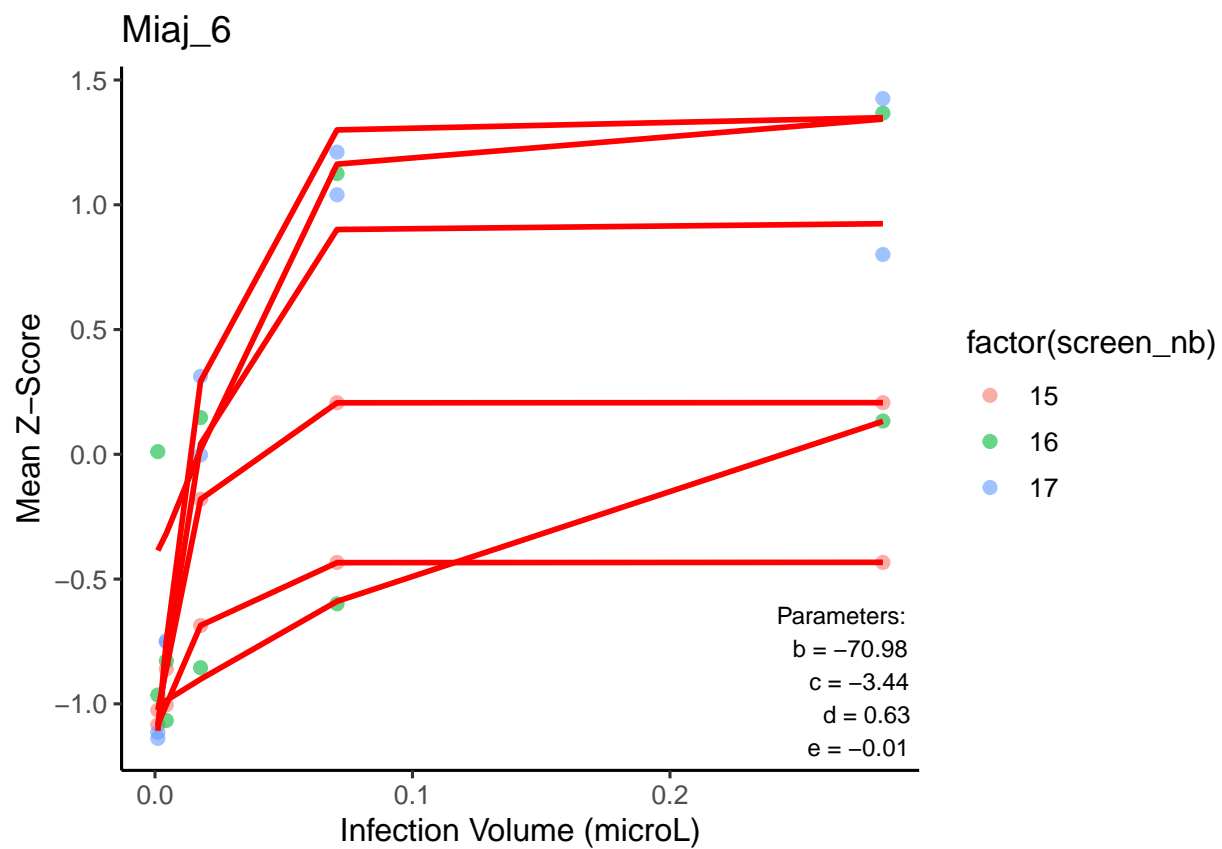
 ## \$Lise_1



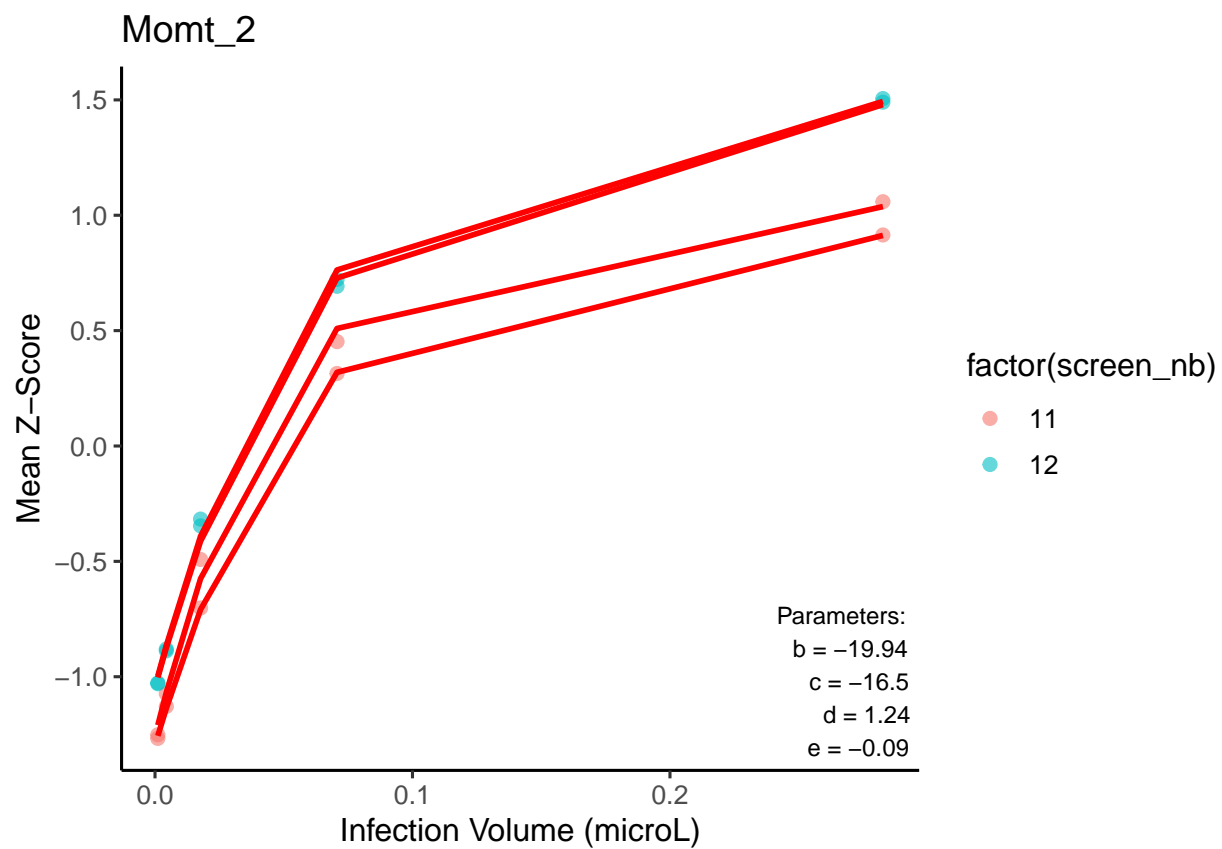
\$Melw_1



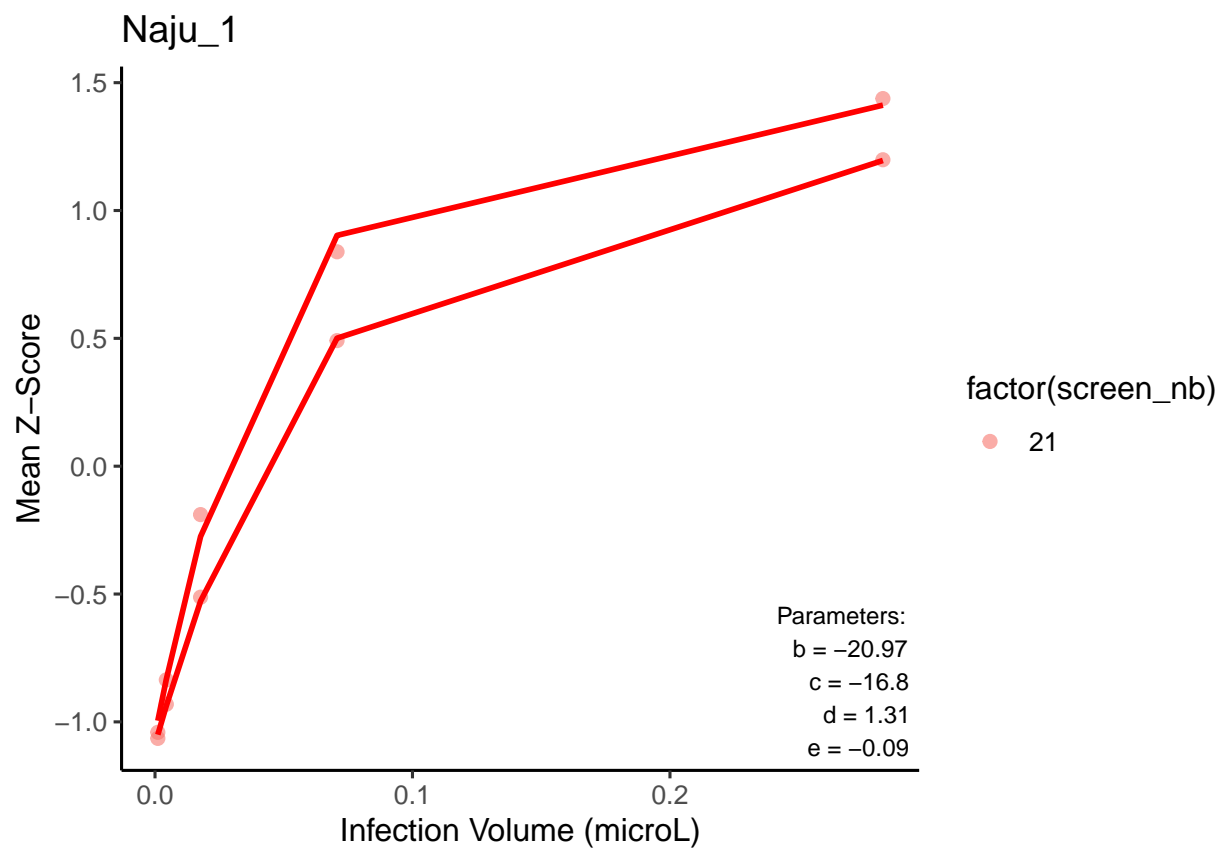
\$Miaj_6



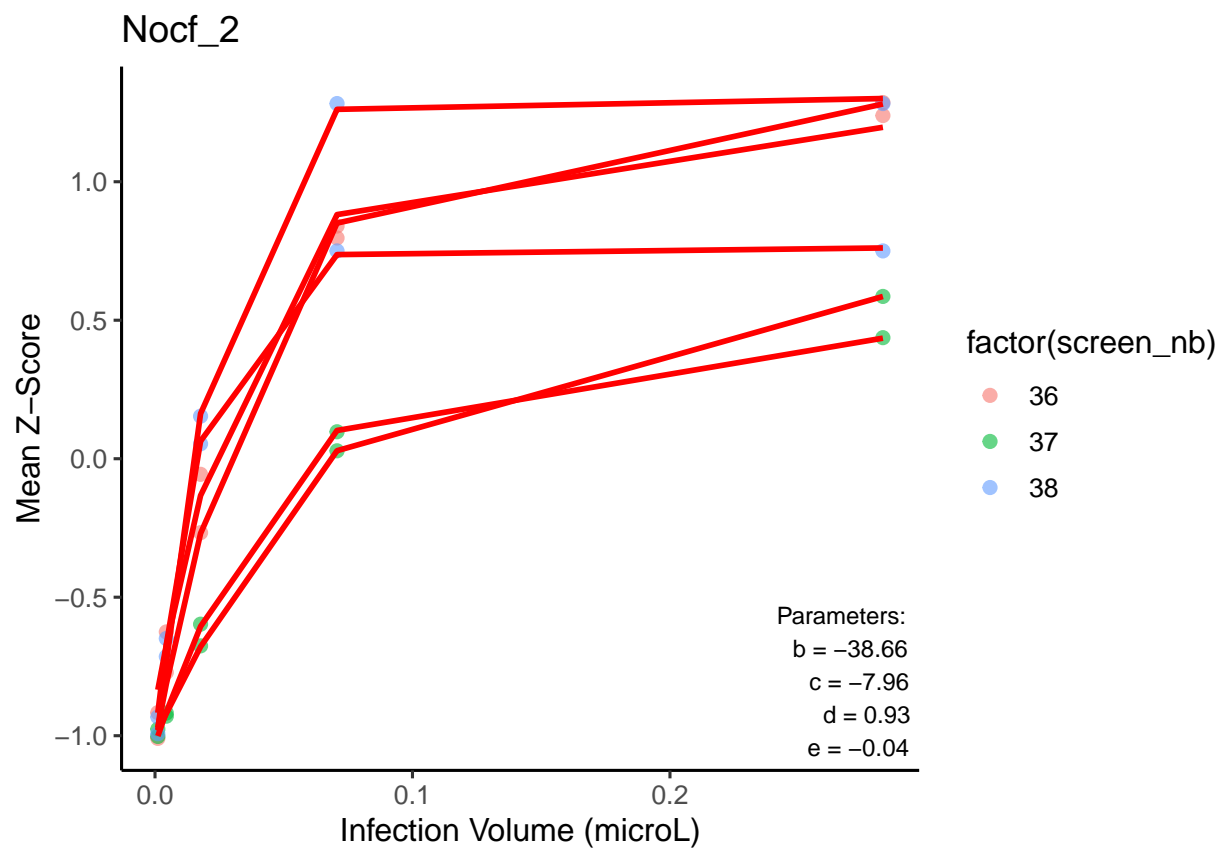
 ## \$Momt_2



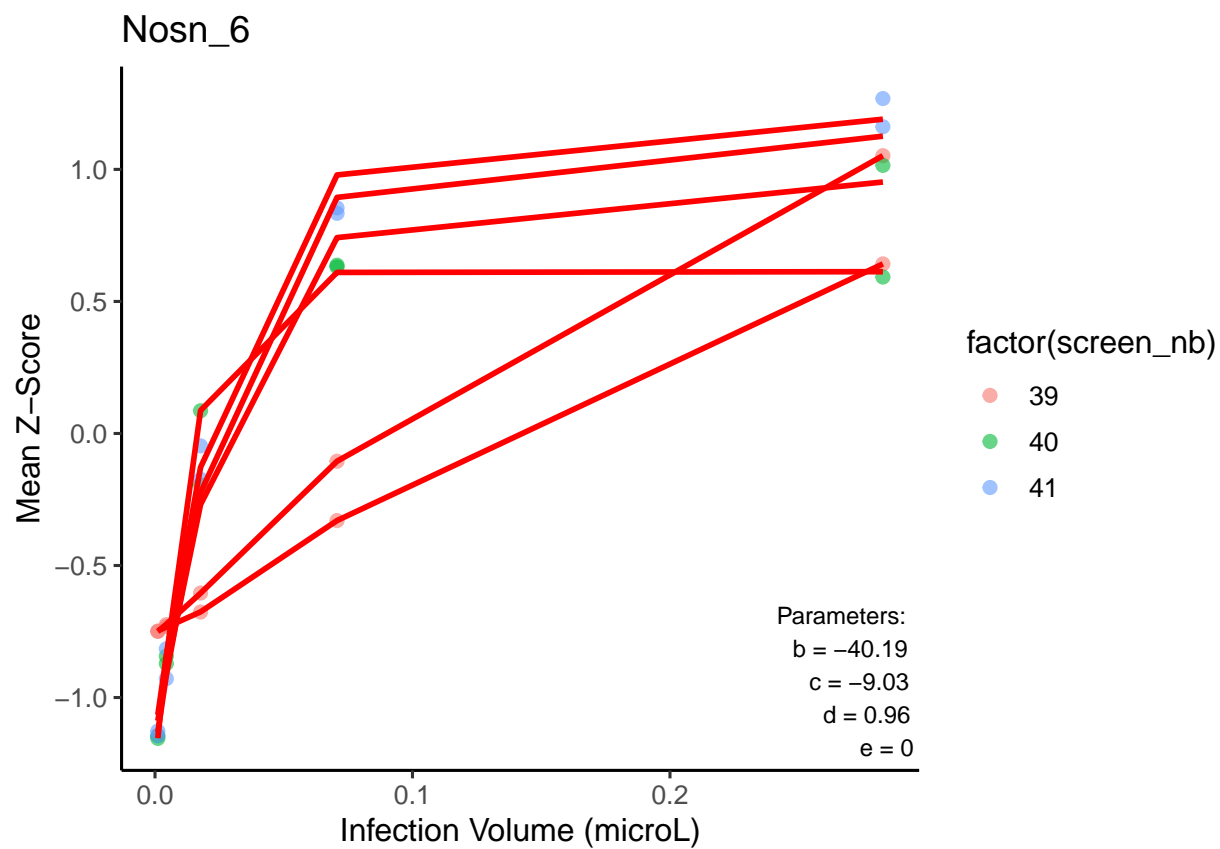
 ## \$Naju_1



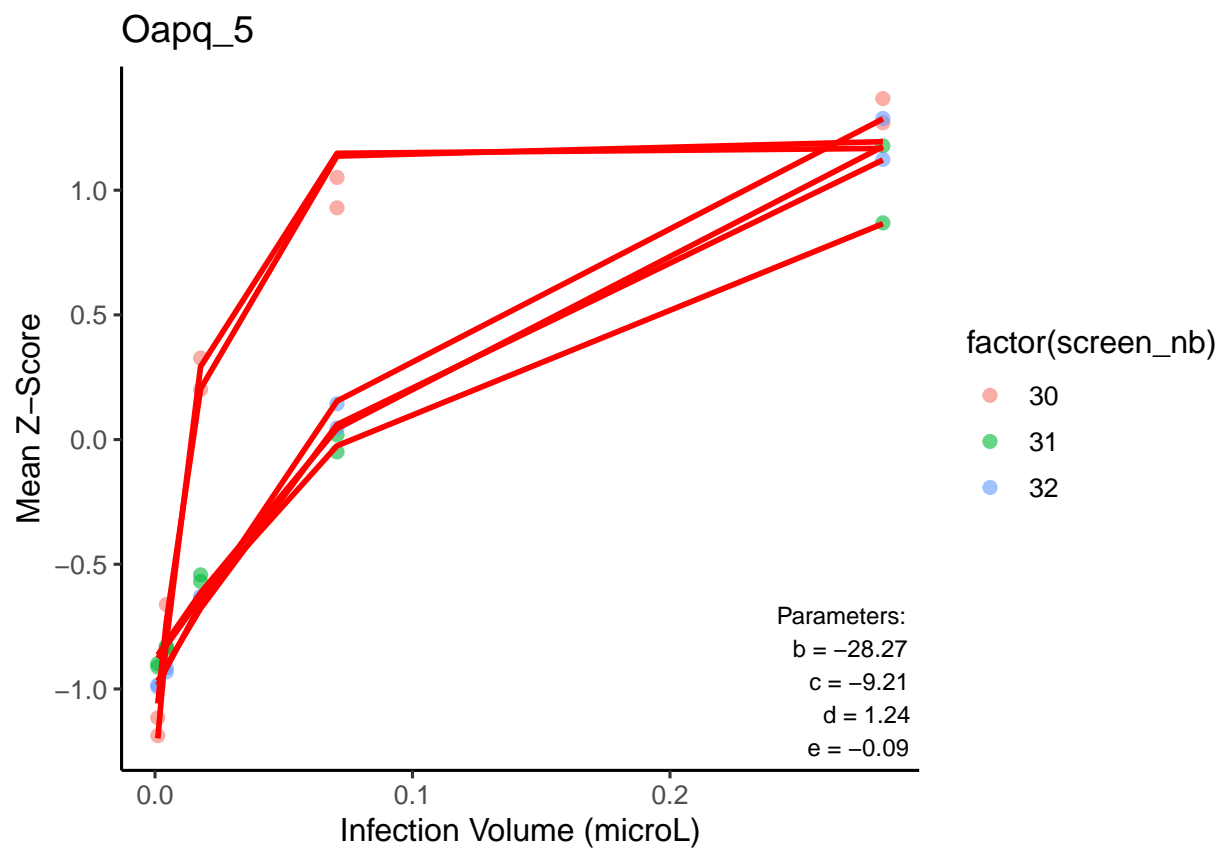
\$Nocf_2



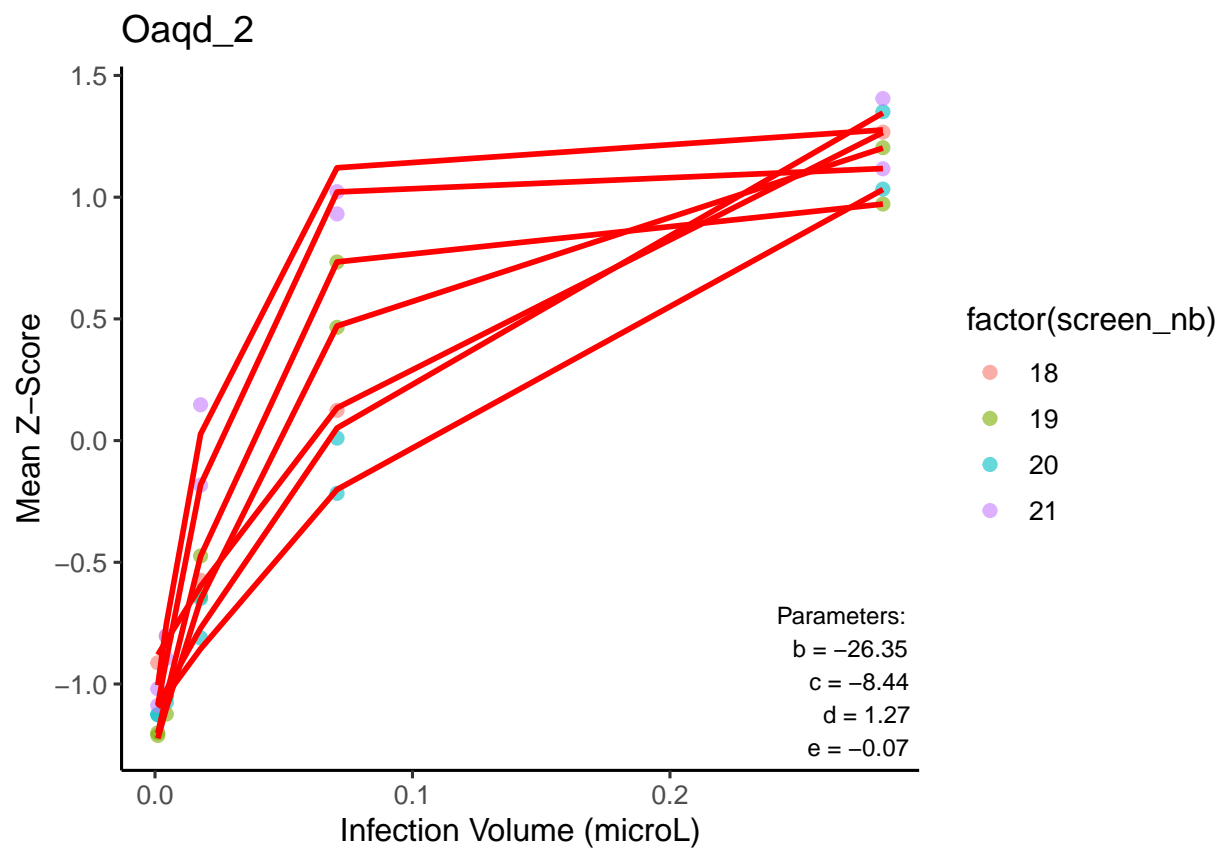
 ## \$Nosn_6



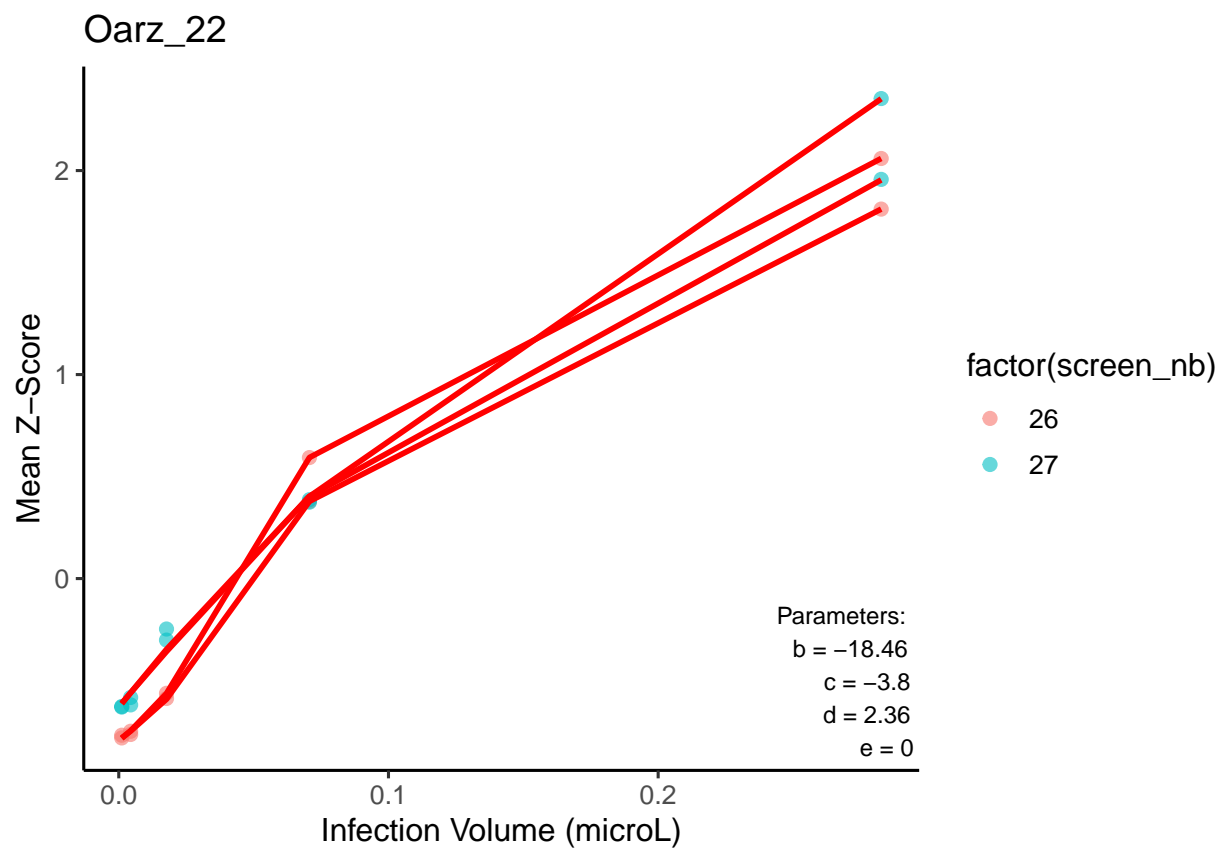
 ## \$0apq_5



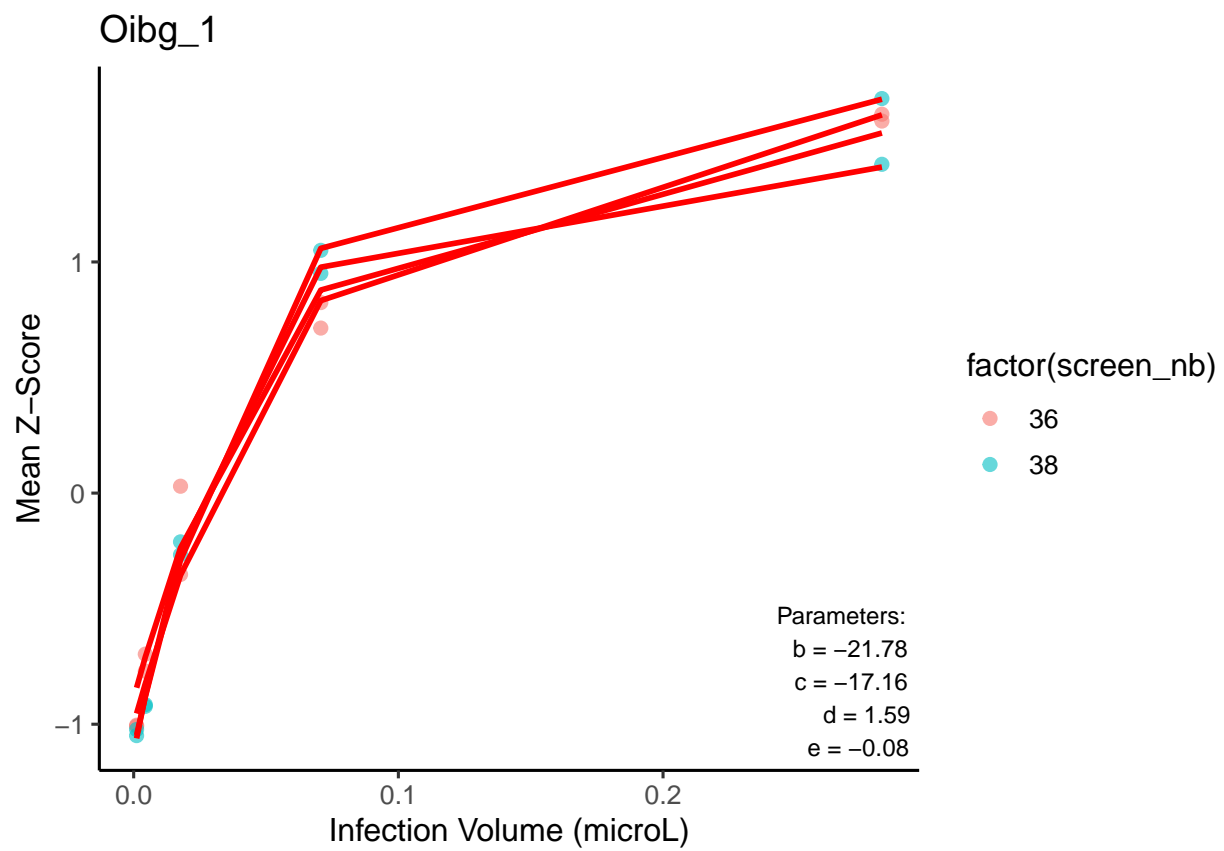
\$0aqd_2



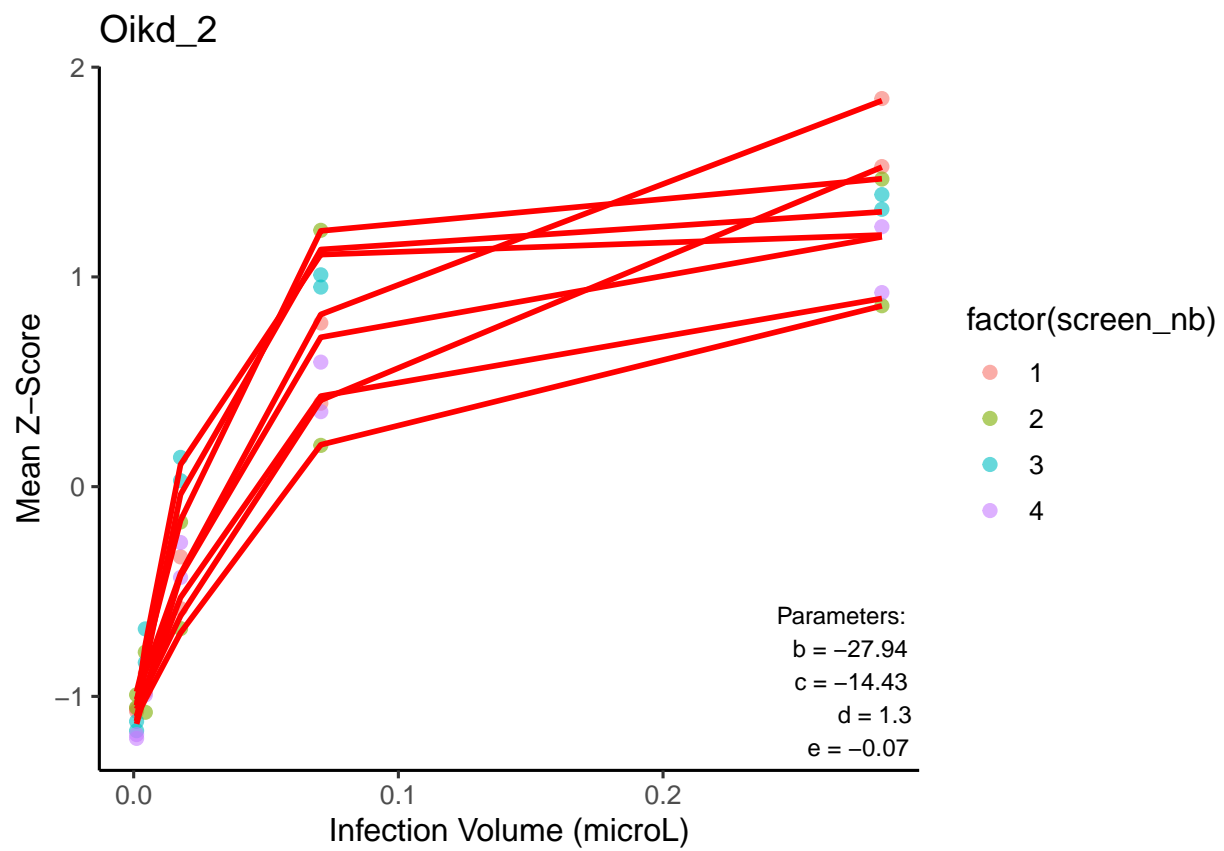
 ## \$0arz_22



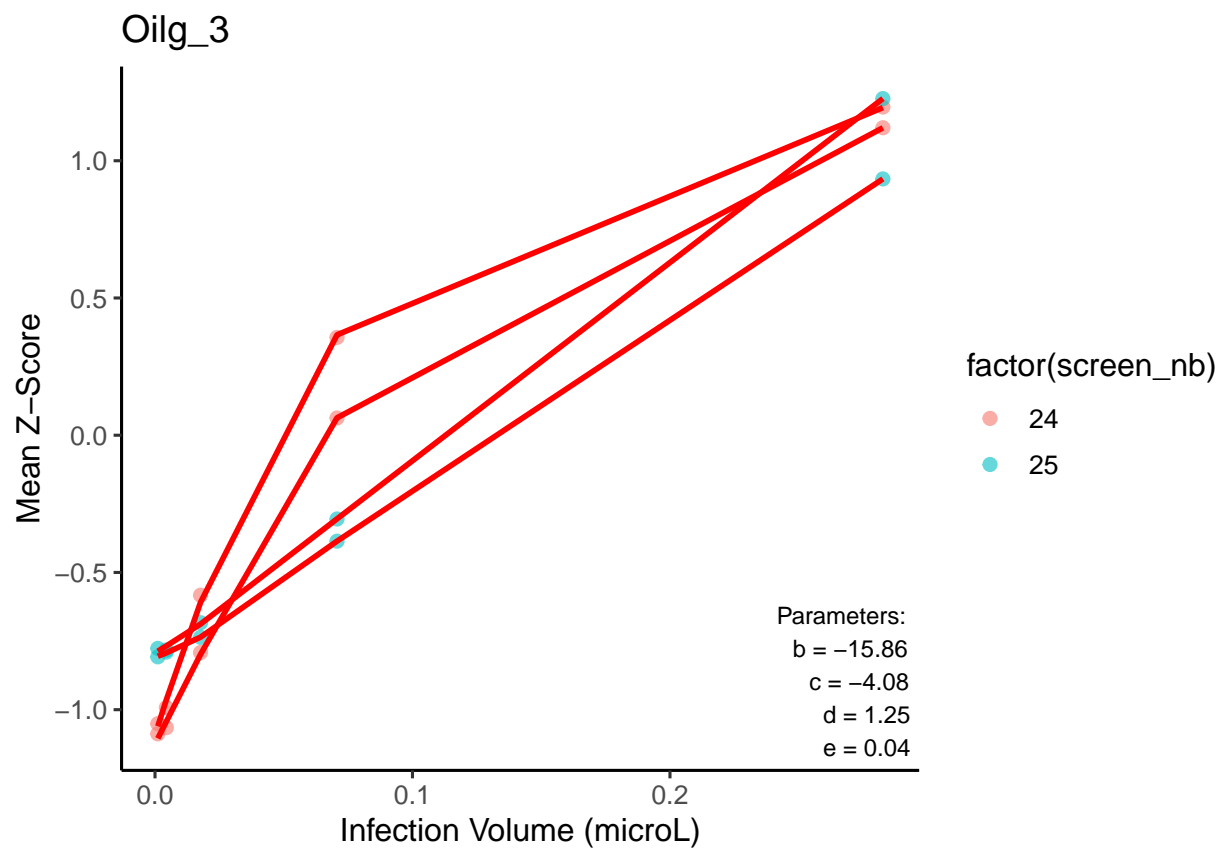
\$0ibg_1



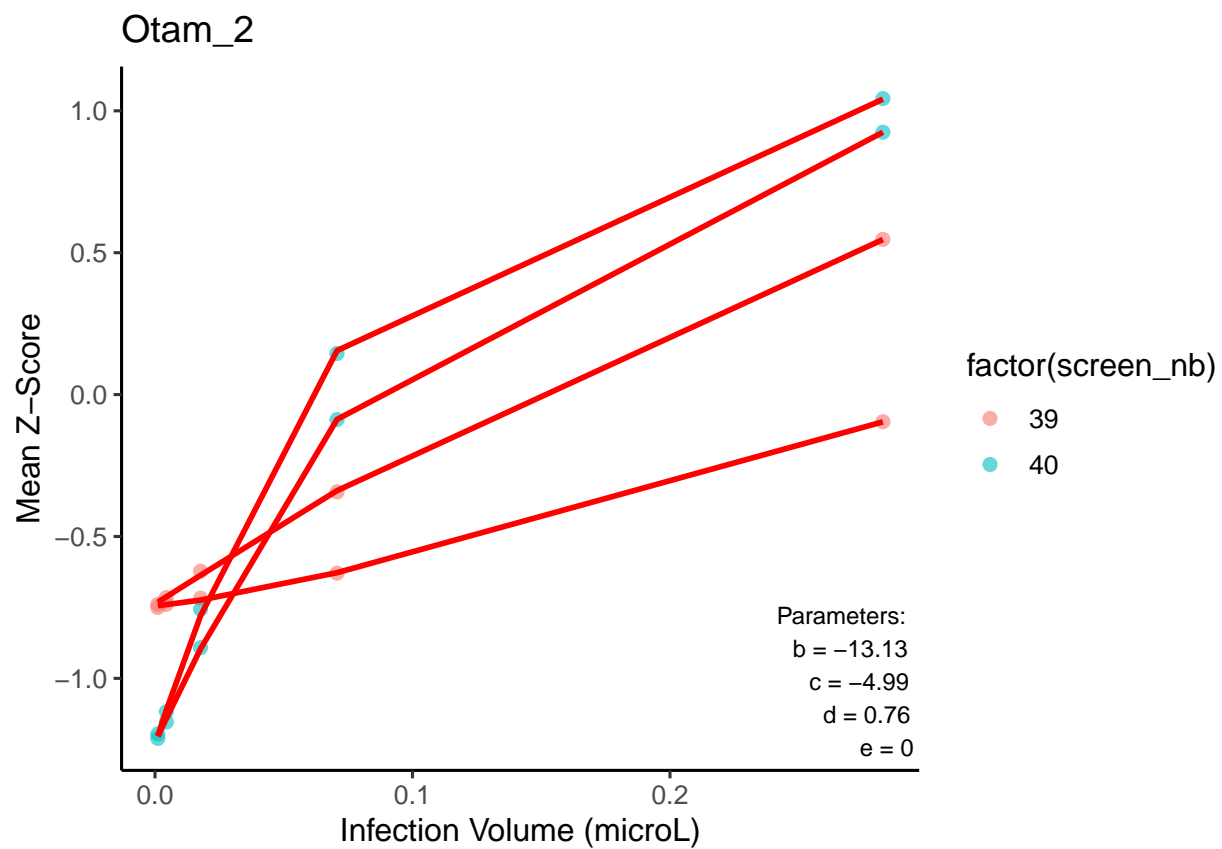
\$Oikd_2



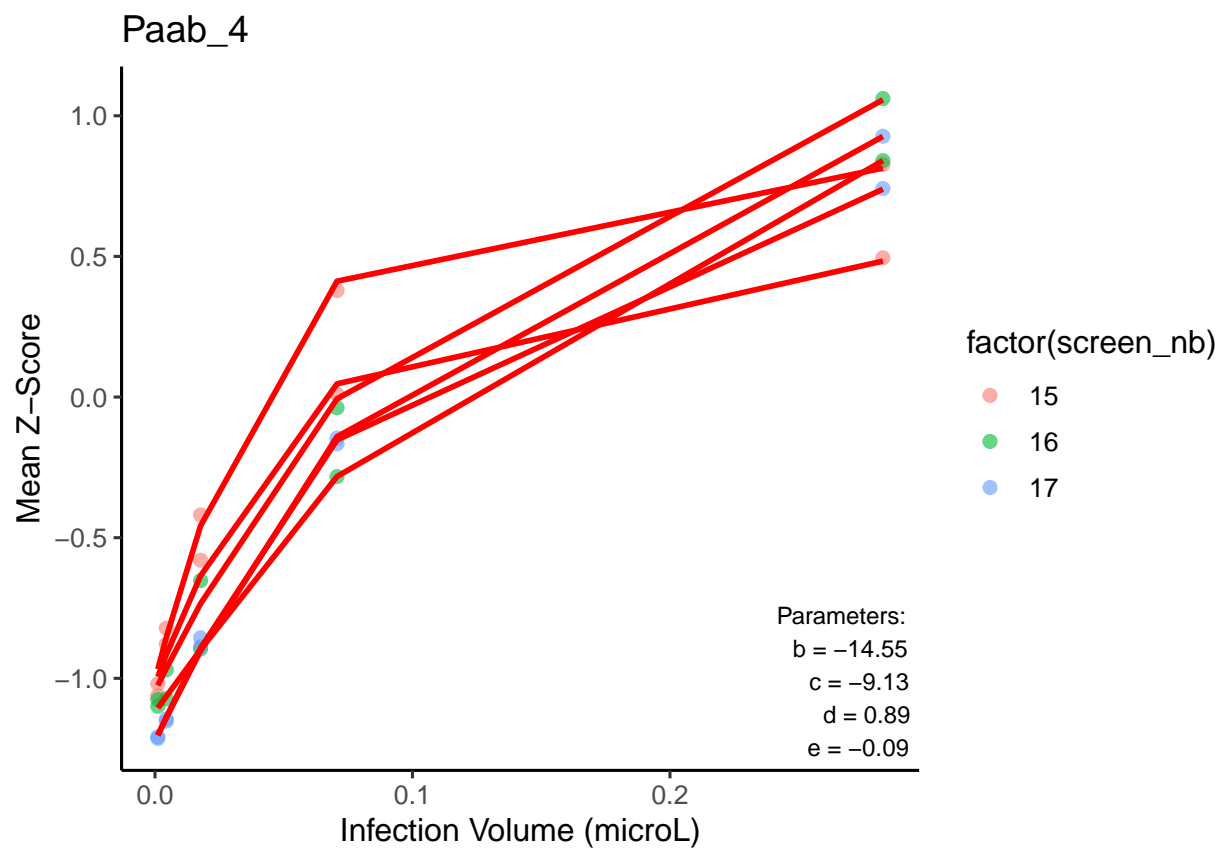
 ## \$Oilg_3



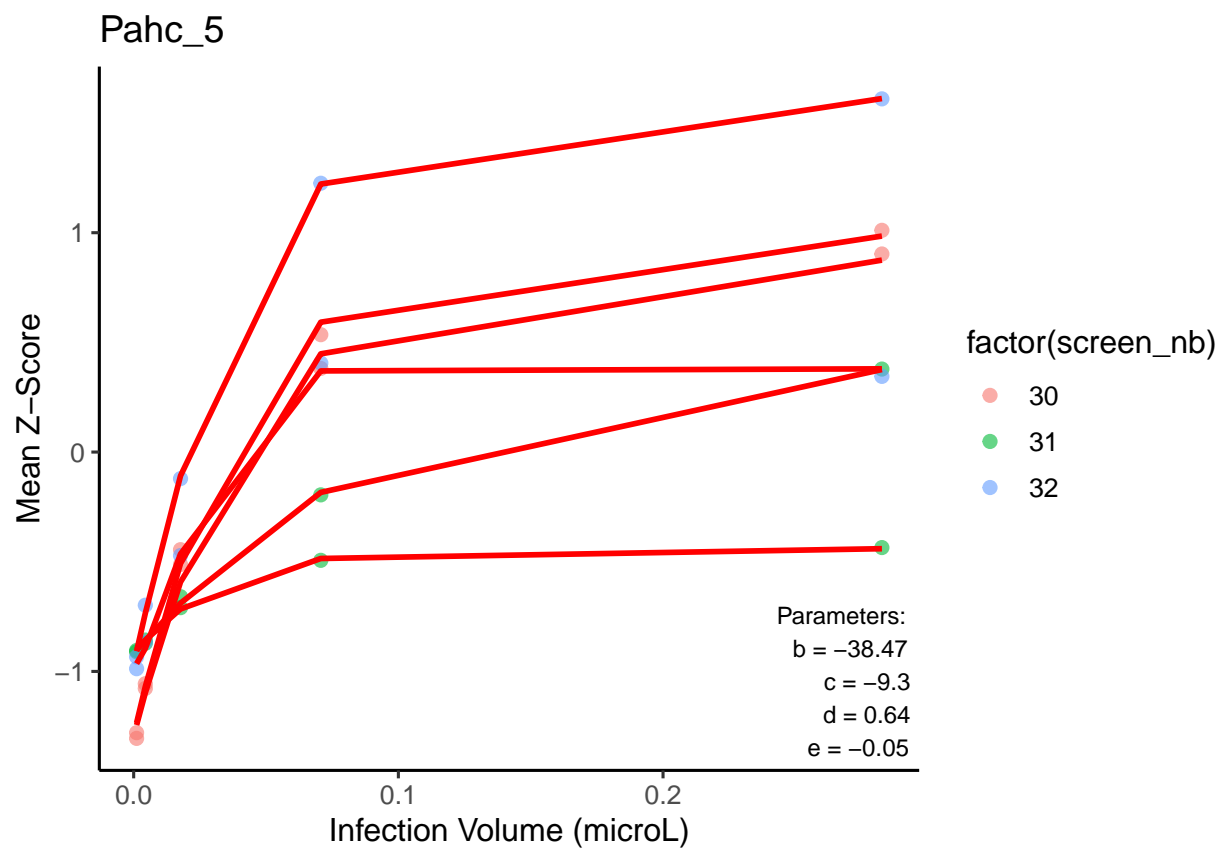
\$0tam_2



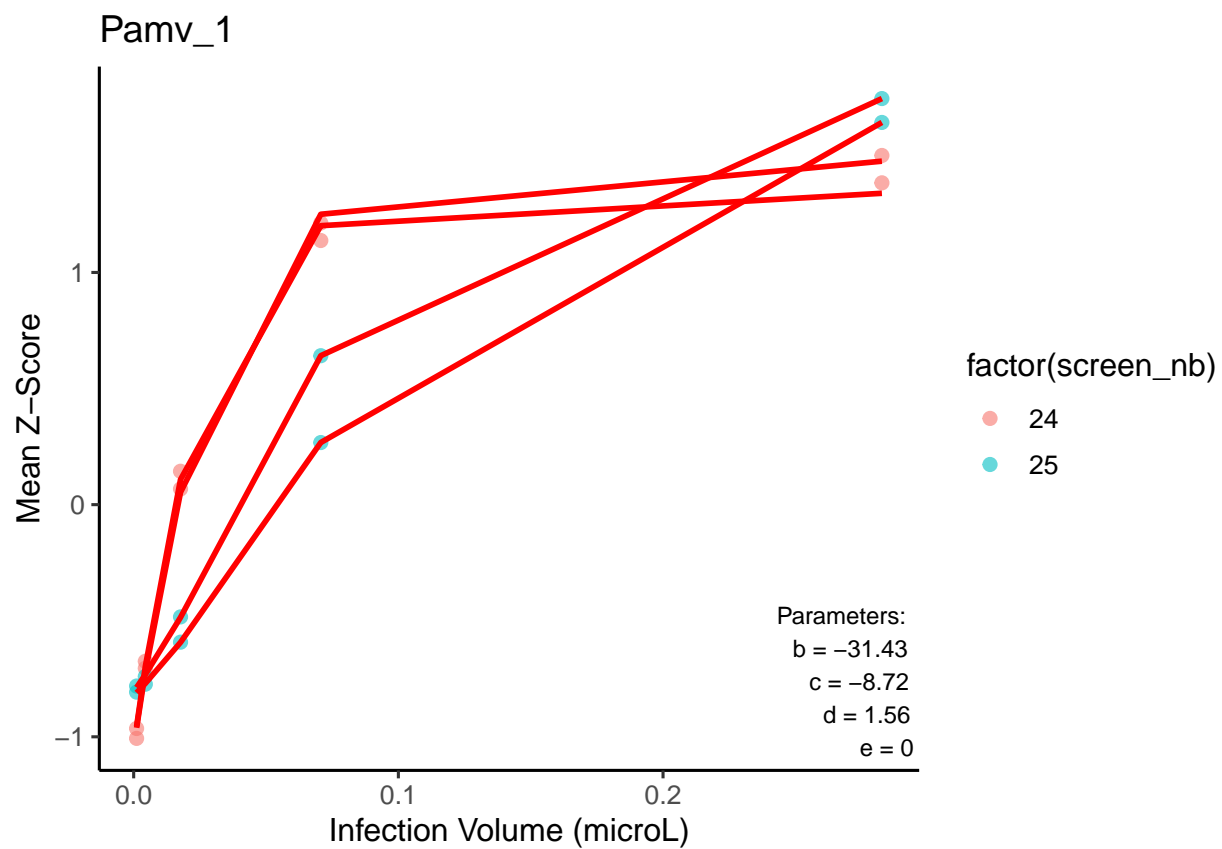
\$Paab_4



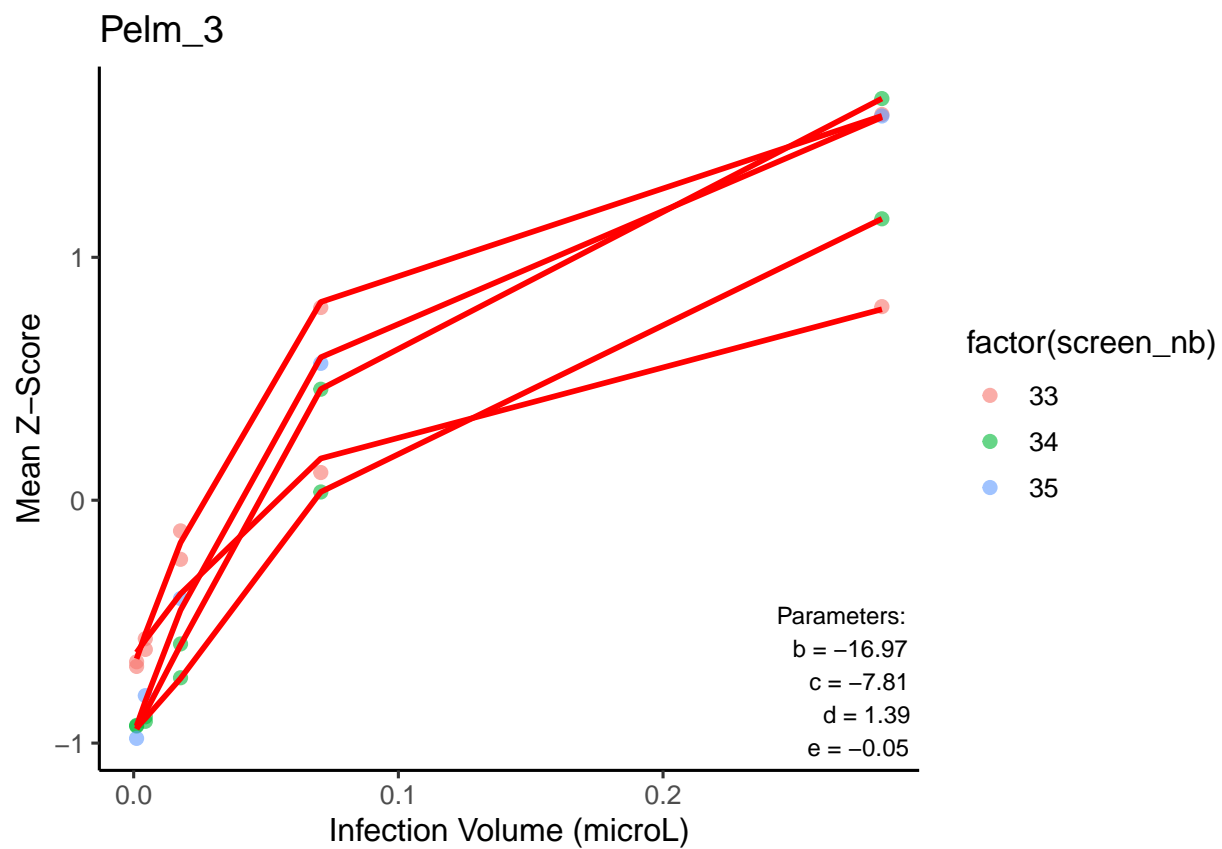
 ## \$Pahc_5



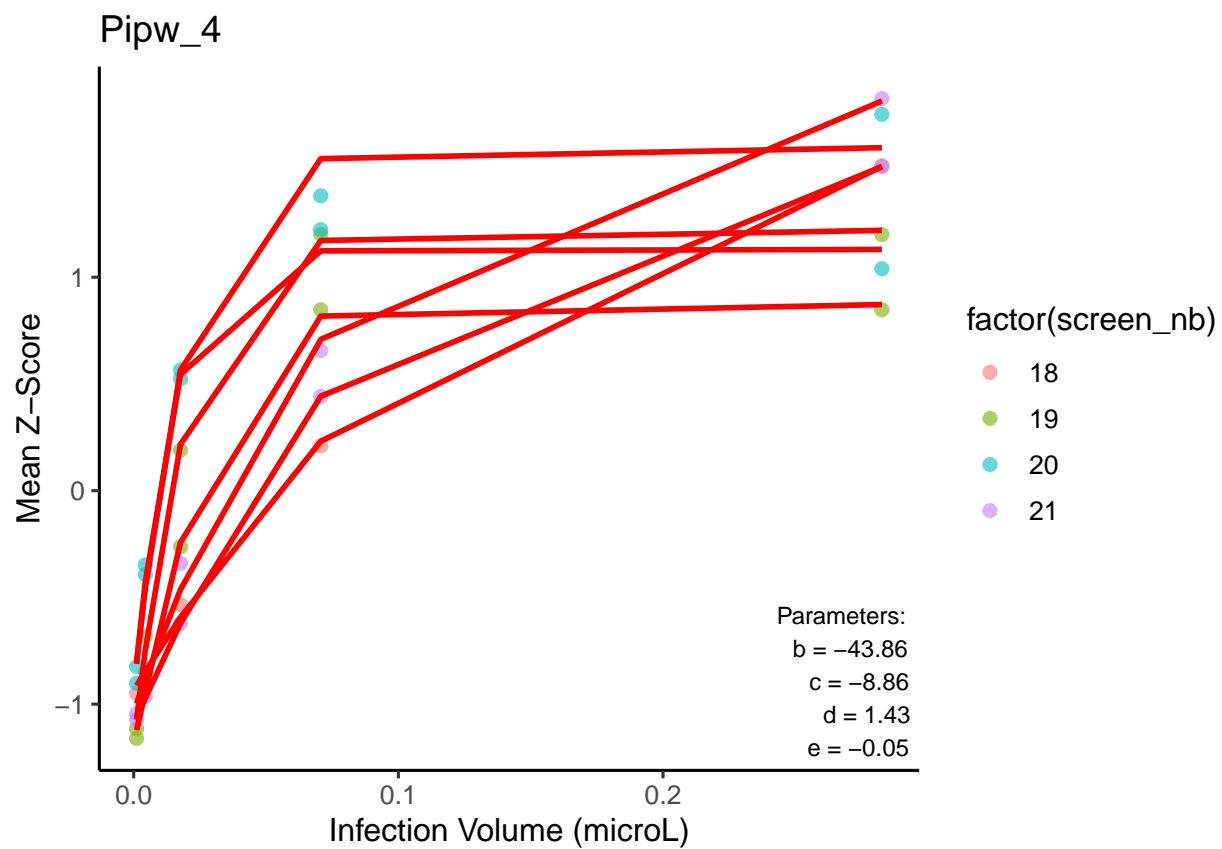
 ## \$Pamv_1



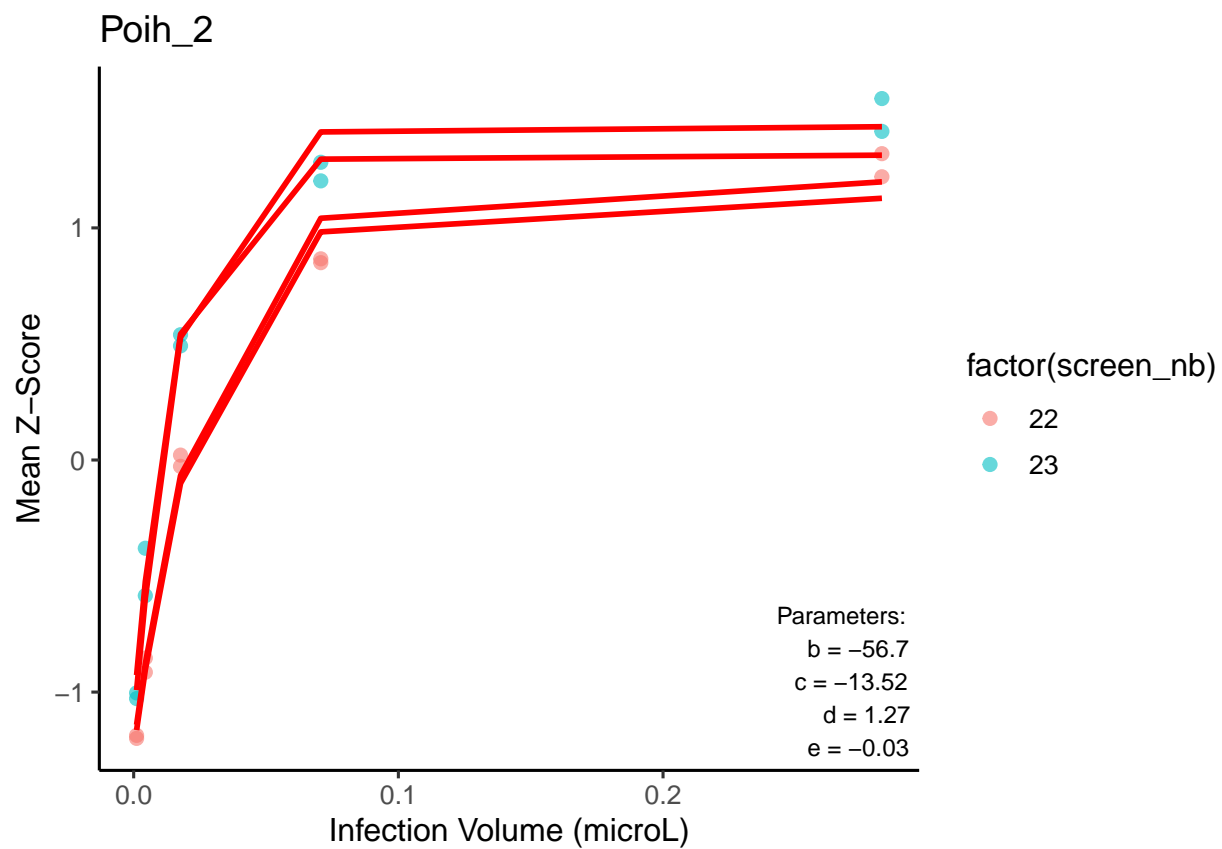
 ## \$Pelm_3



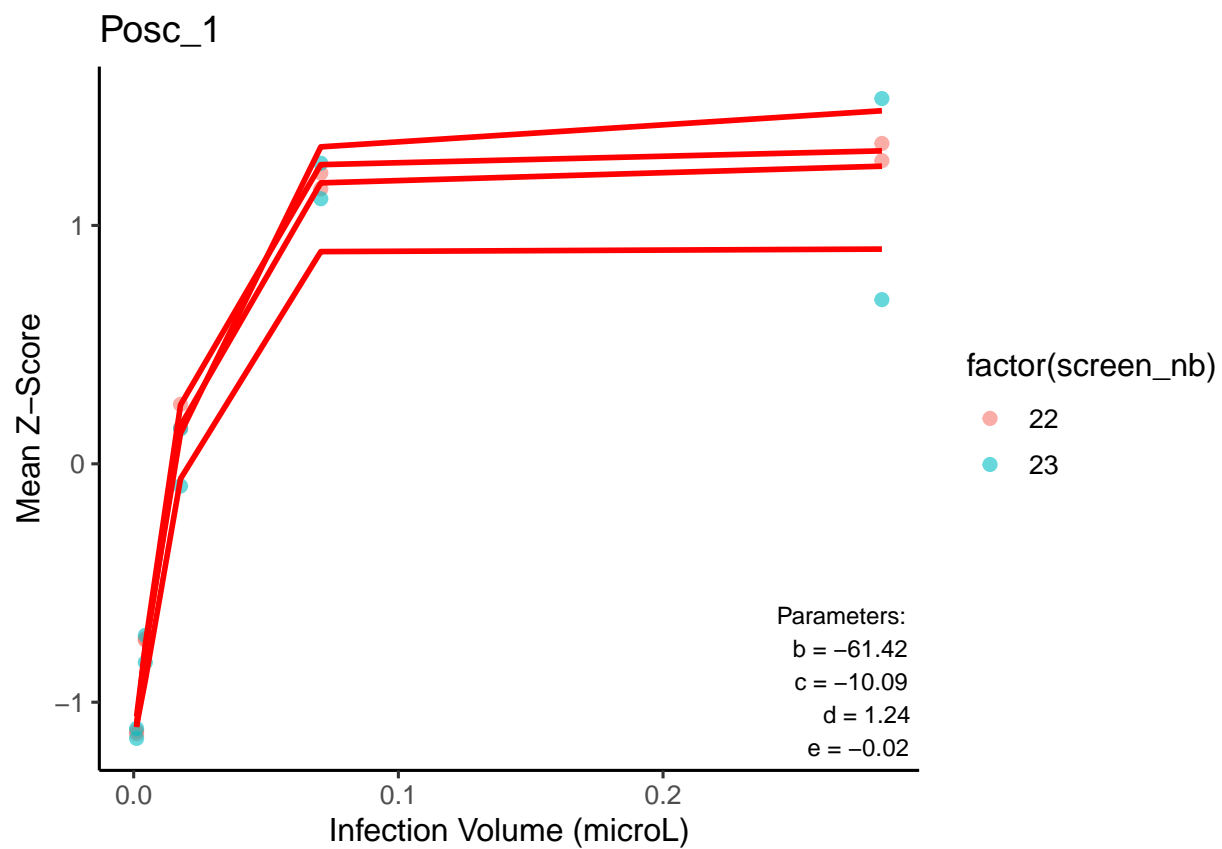
\$Pipw_4



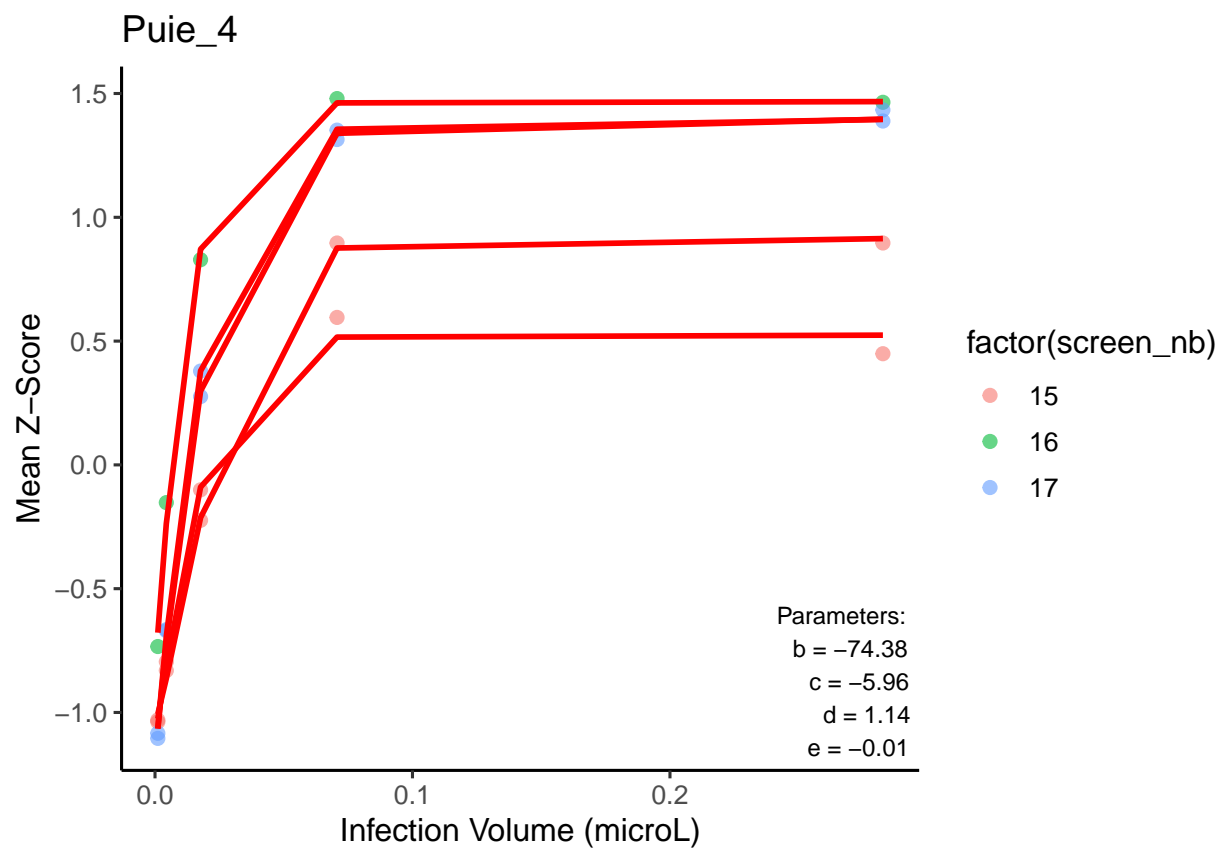
```
##  
## $Poih_2
```



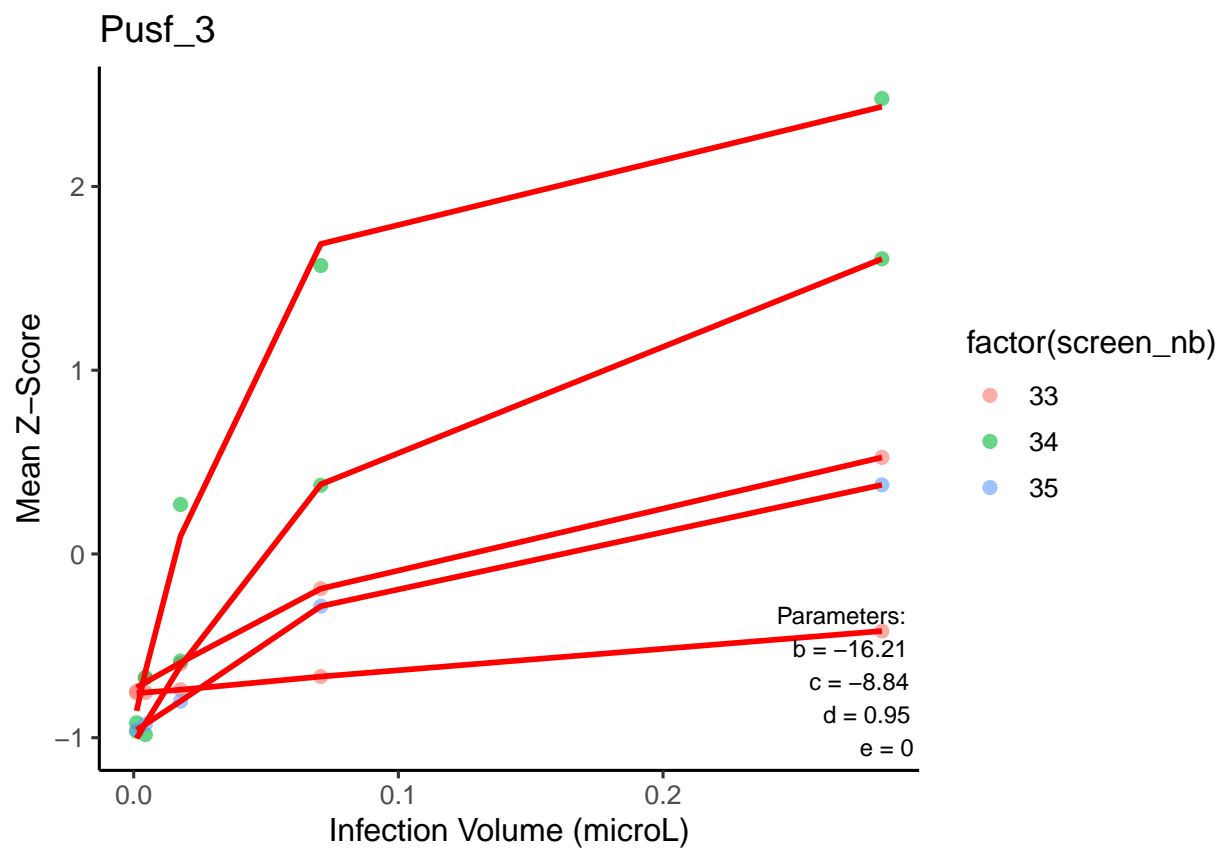
\$Posc_1



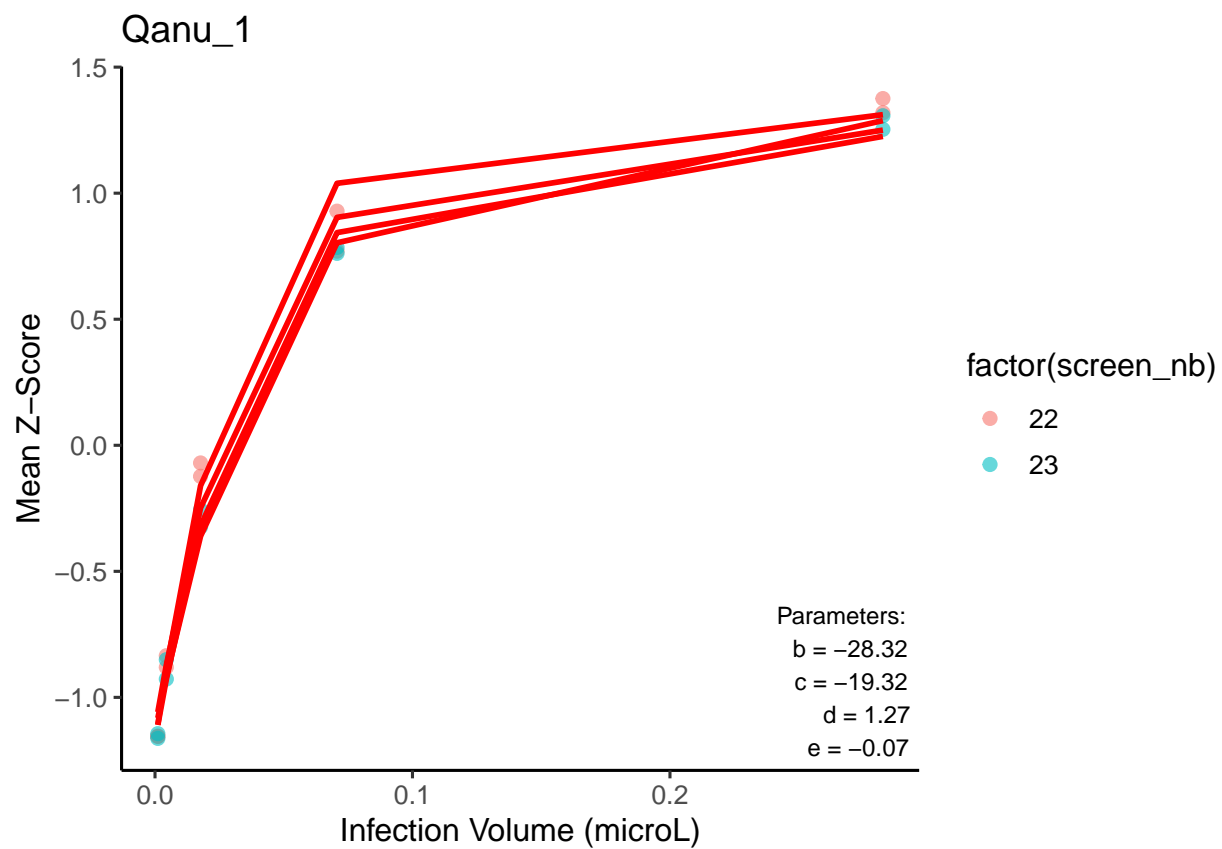
 ## \$Puie_4



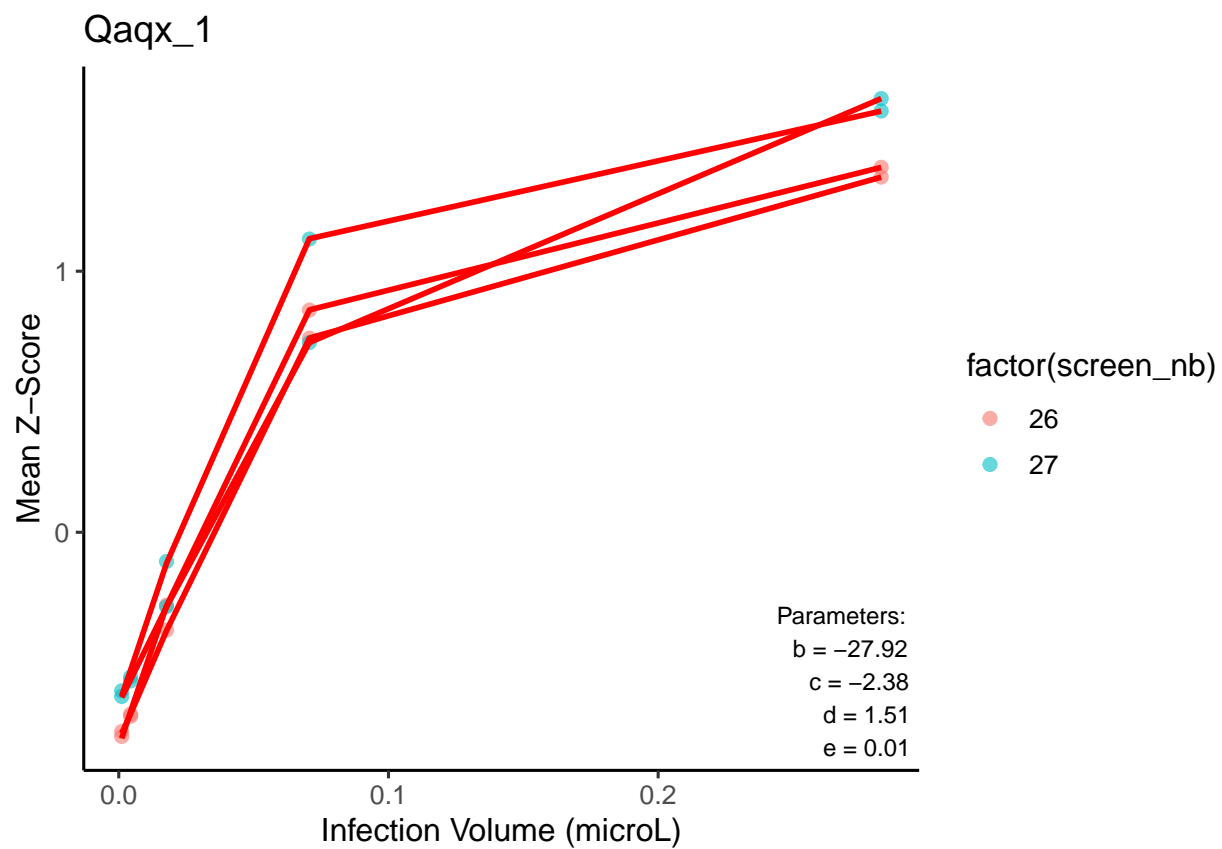
\$Pusf_3



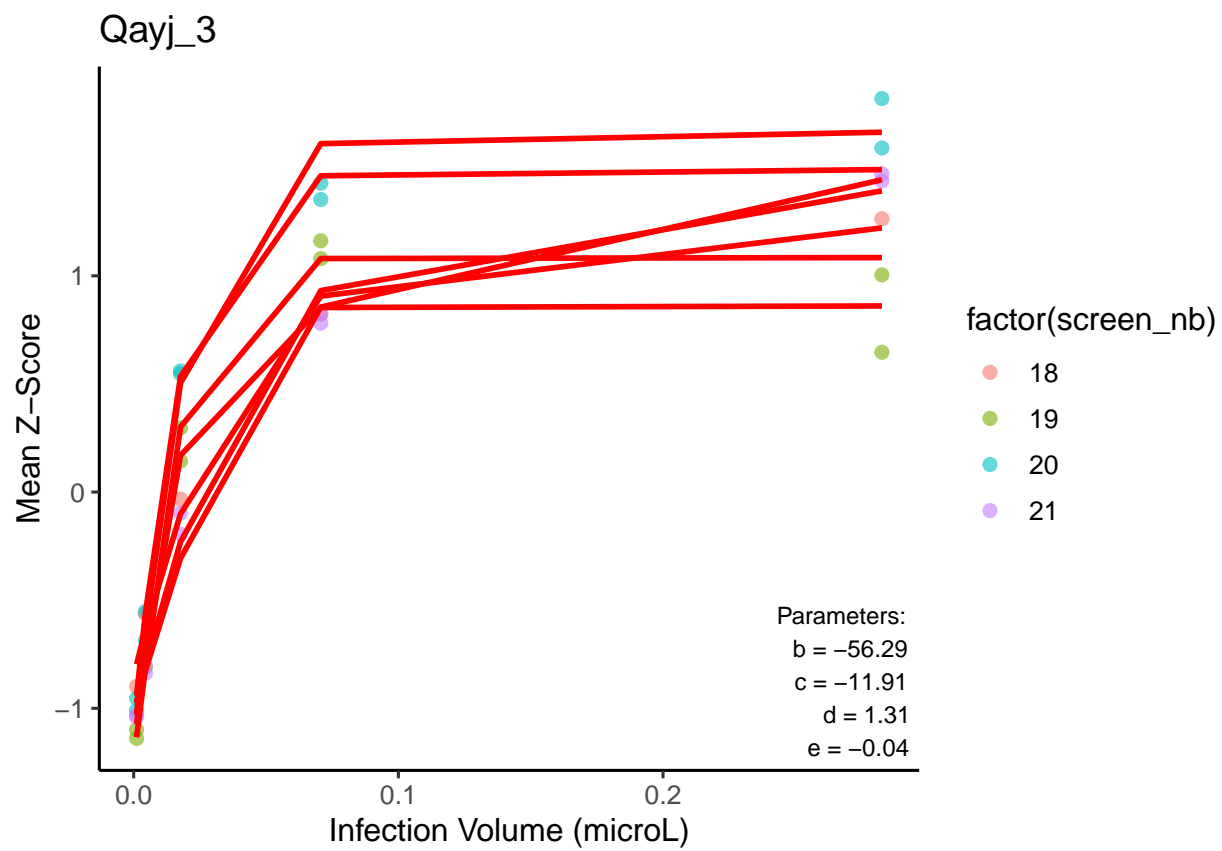
\$Qanu_1



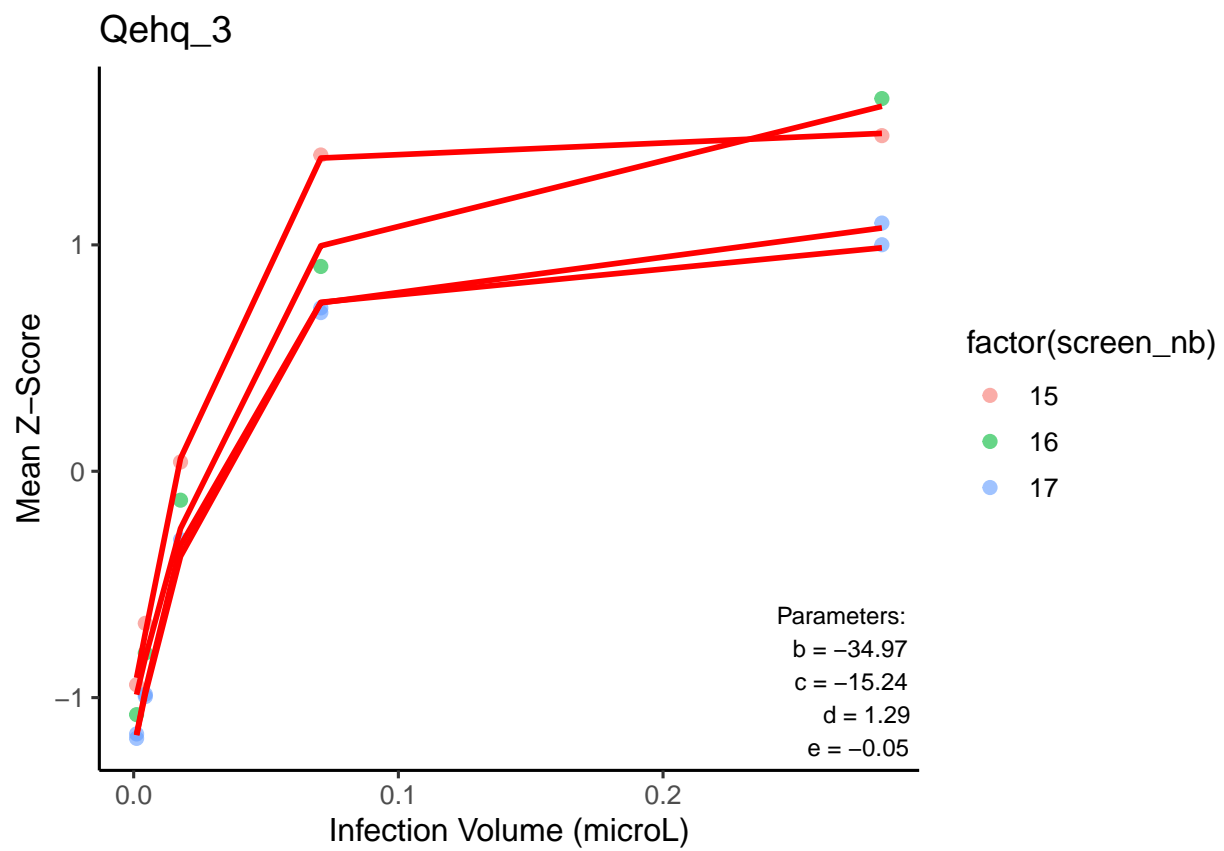
\$Qaax_1



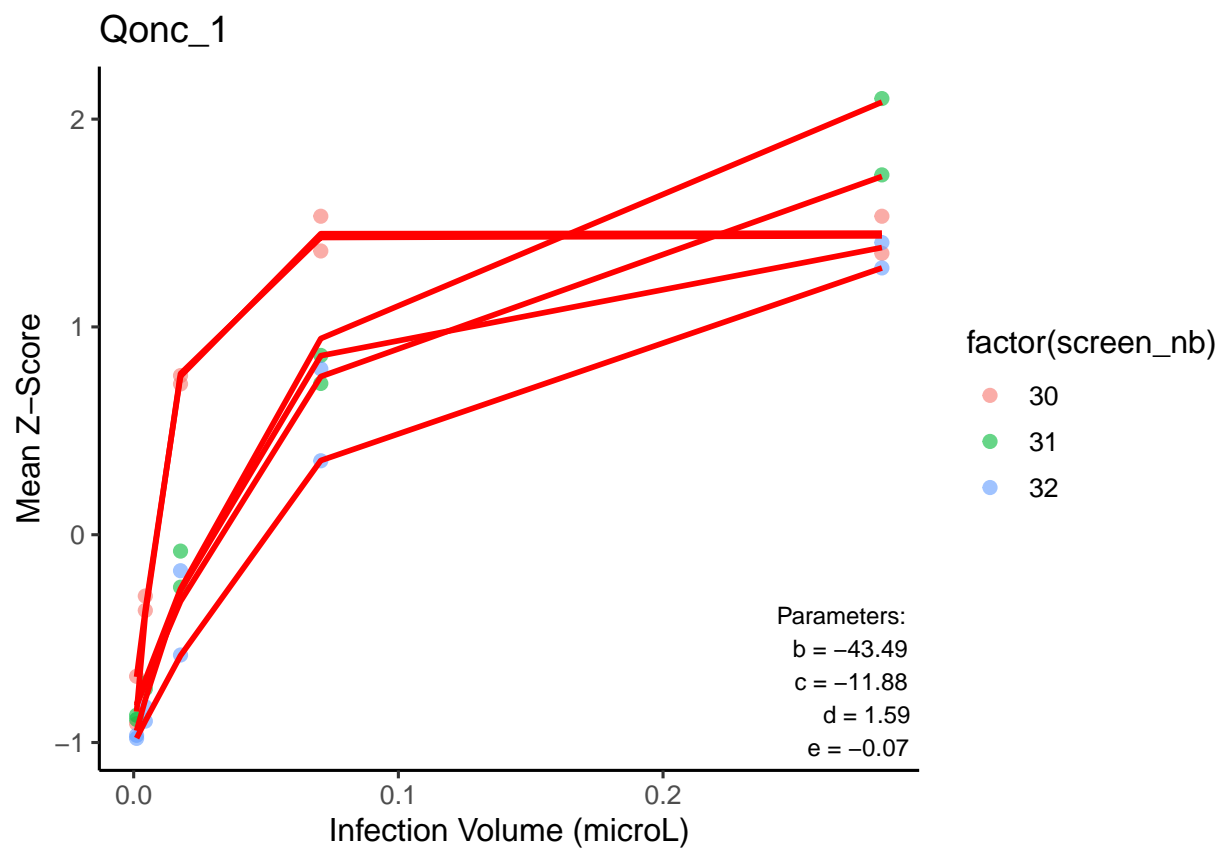
\$Qayj_3



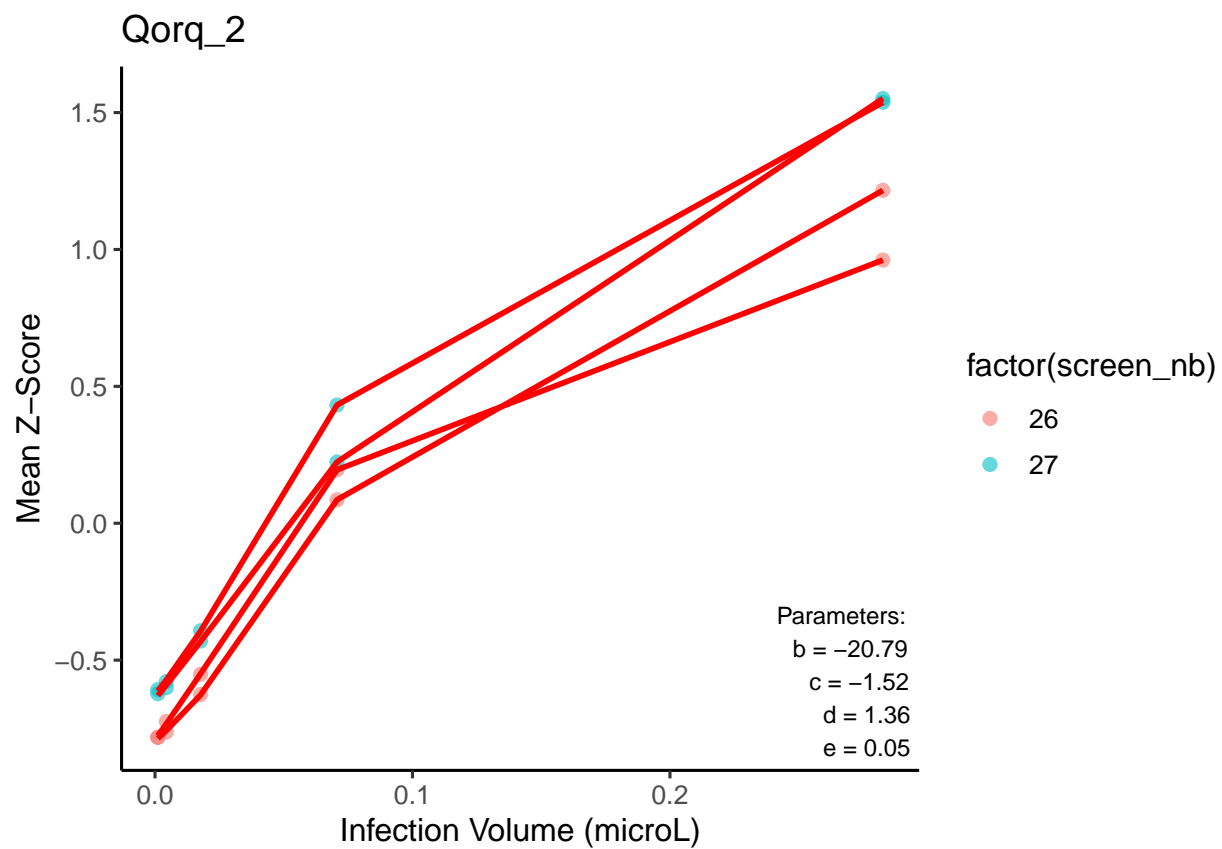
 ## \$Qehq_3



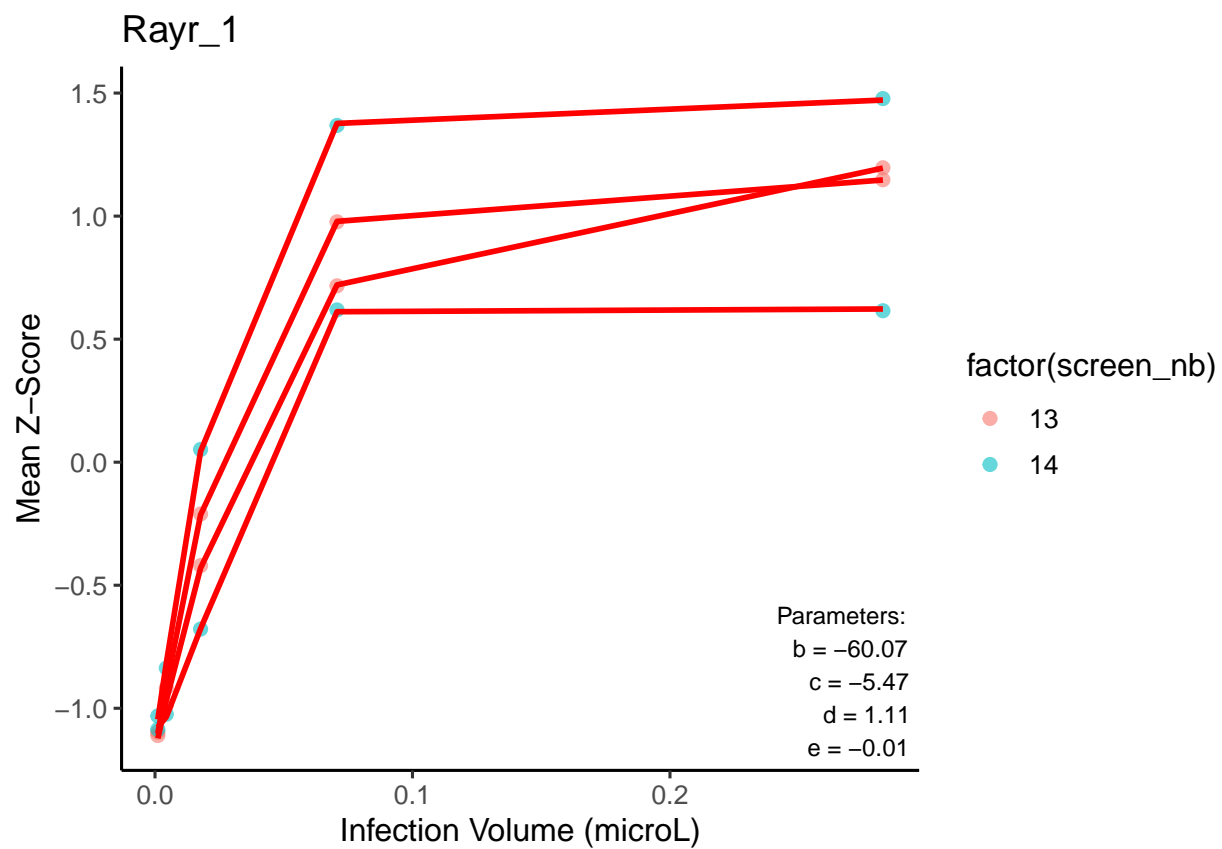
\$Qonc_1



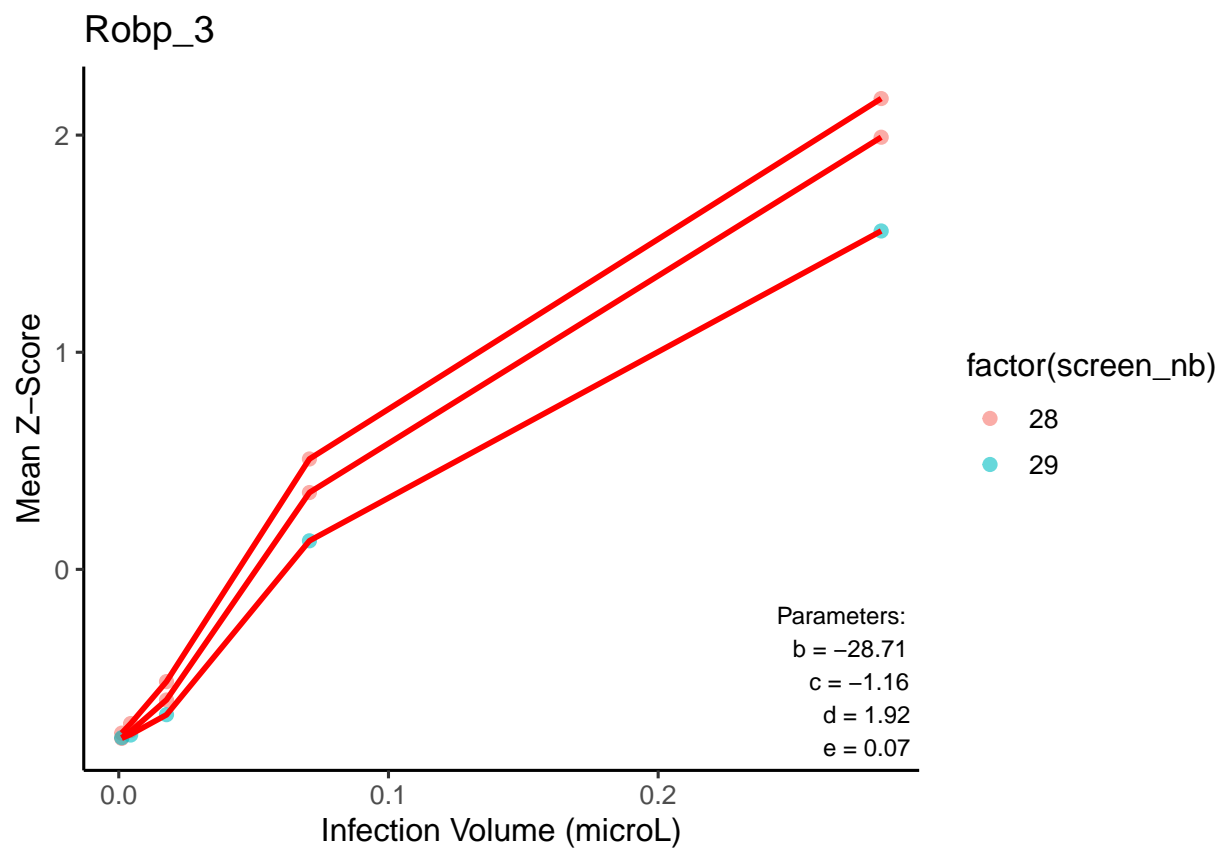
 ## \$Qorq_2



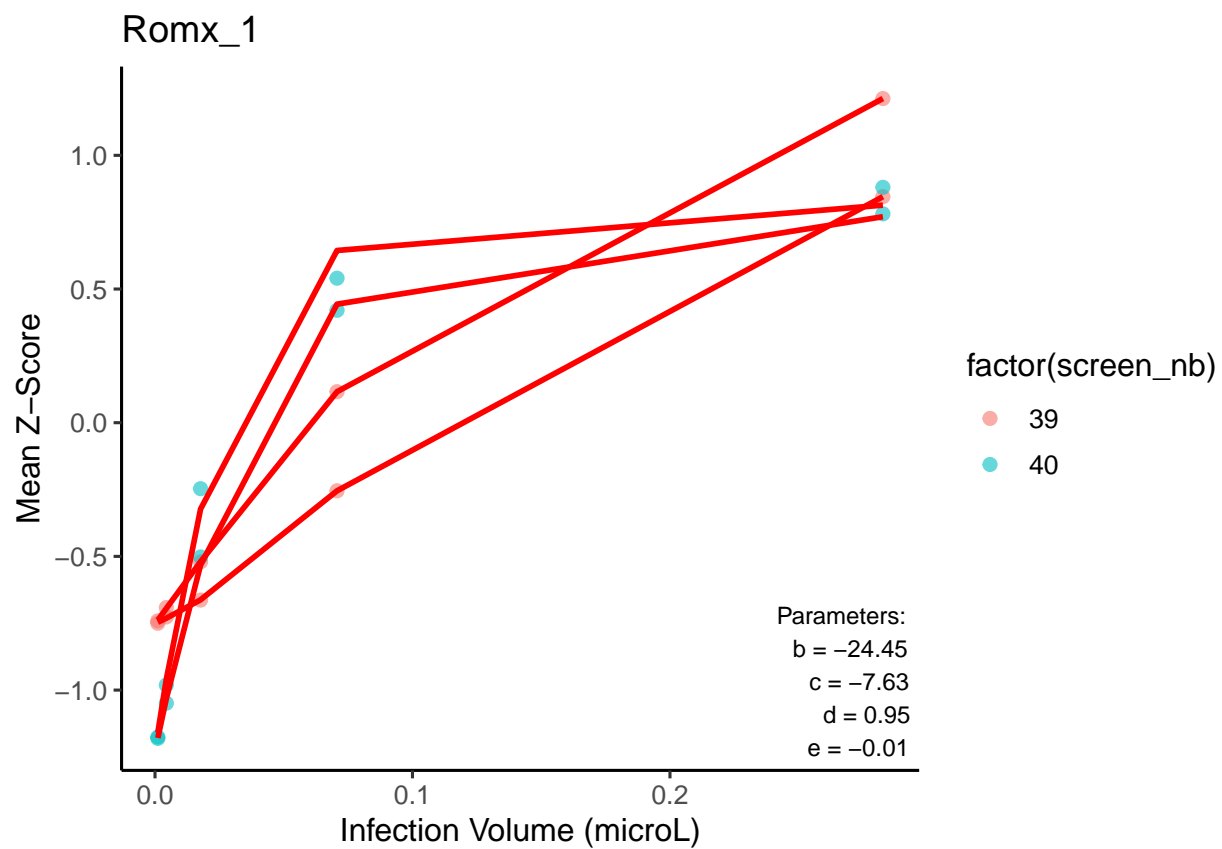
\$Rayr_1



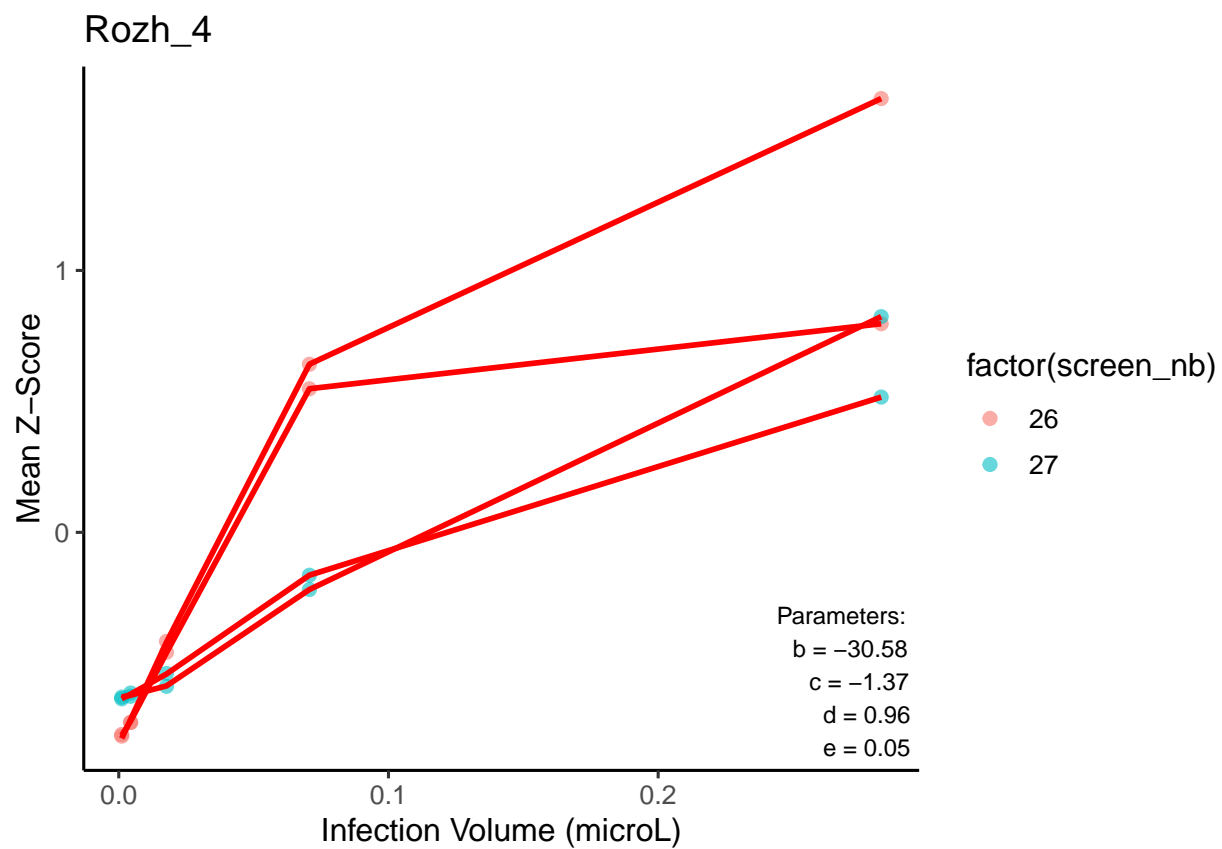
 ## \$Robp_3



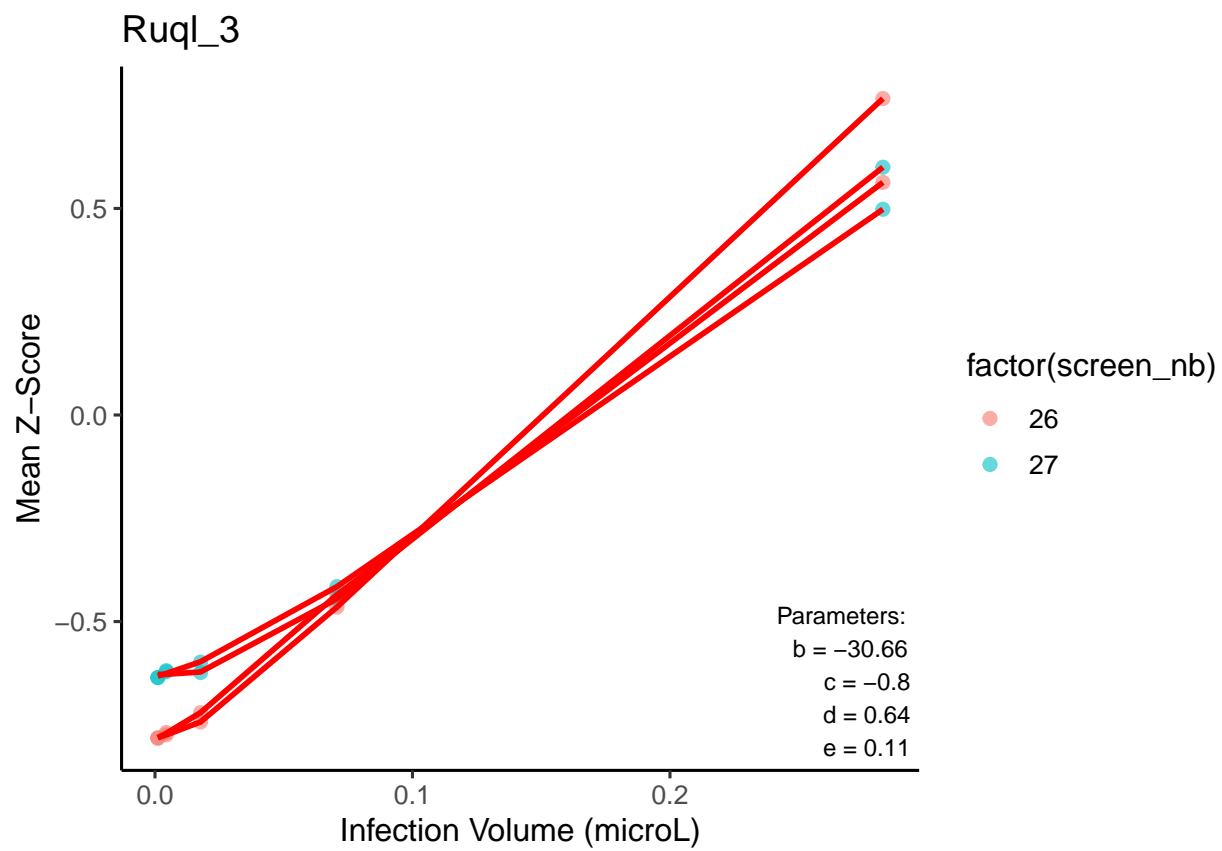
\$Romx_1



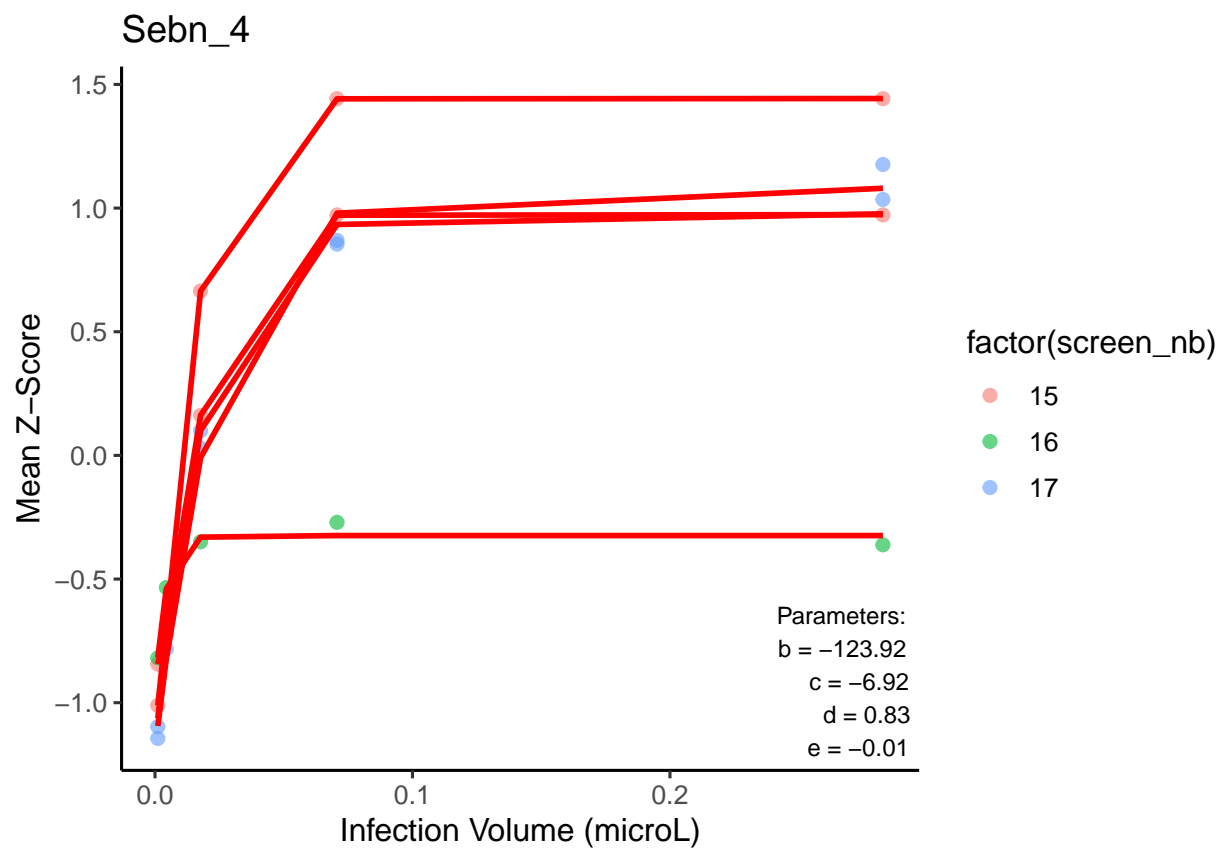
 ## \$Rozh_4



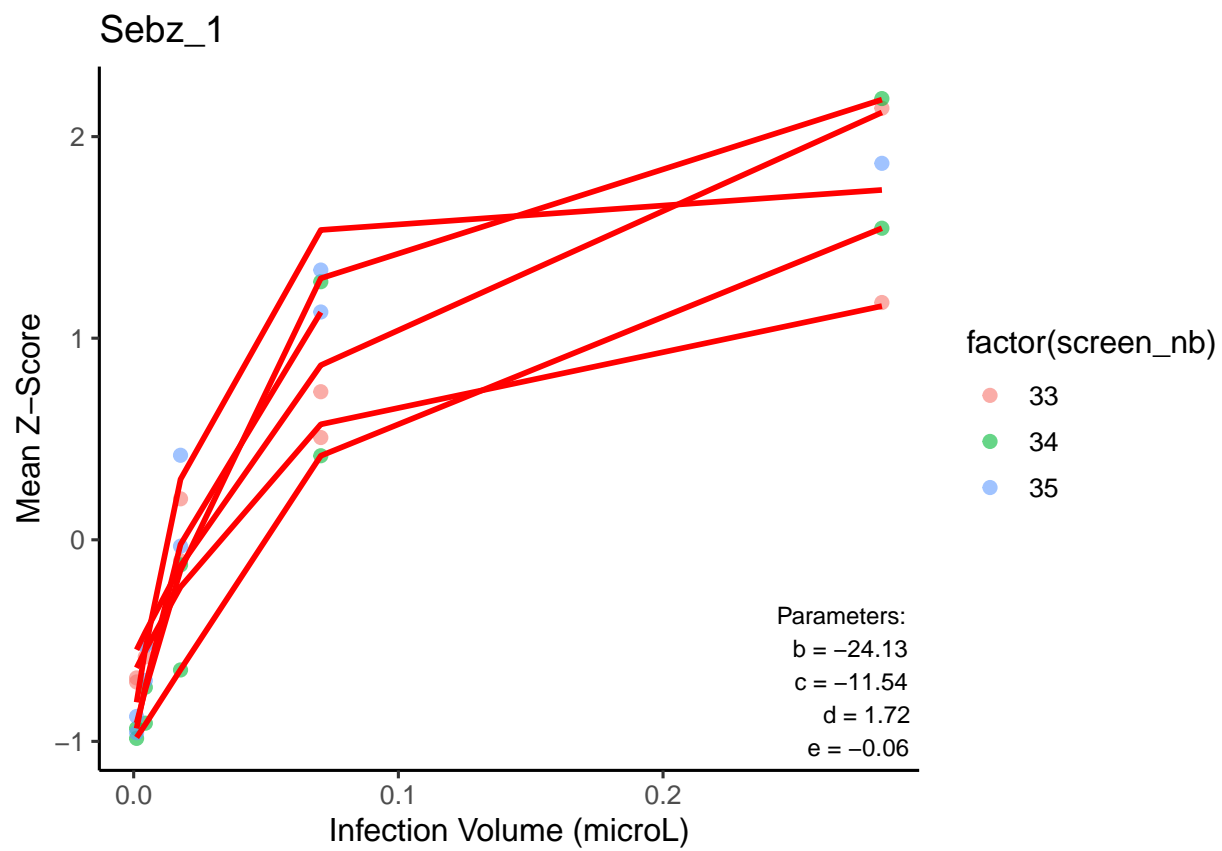
\$Ruq1_3



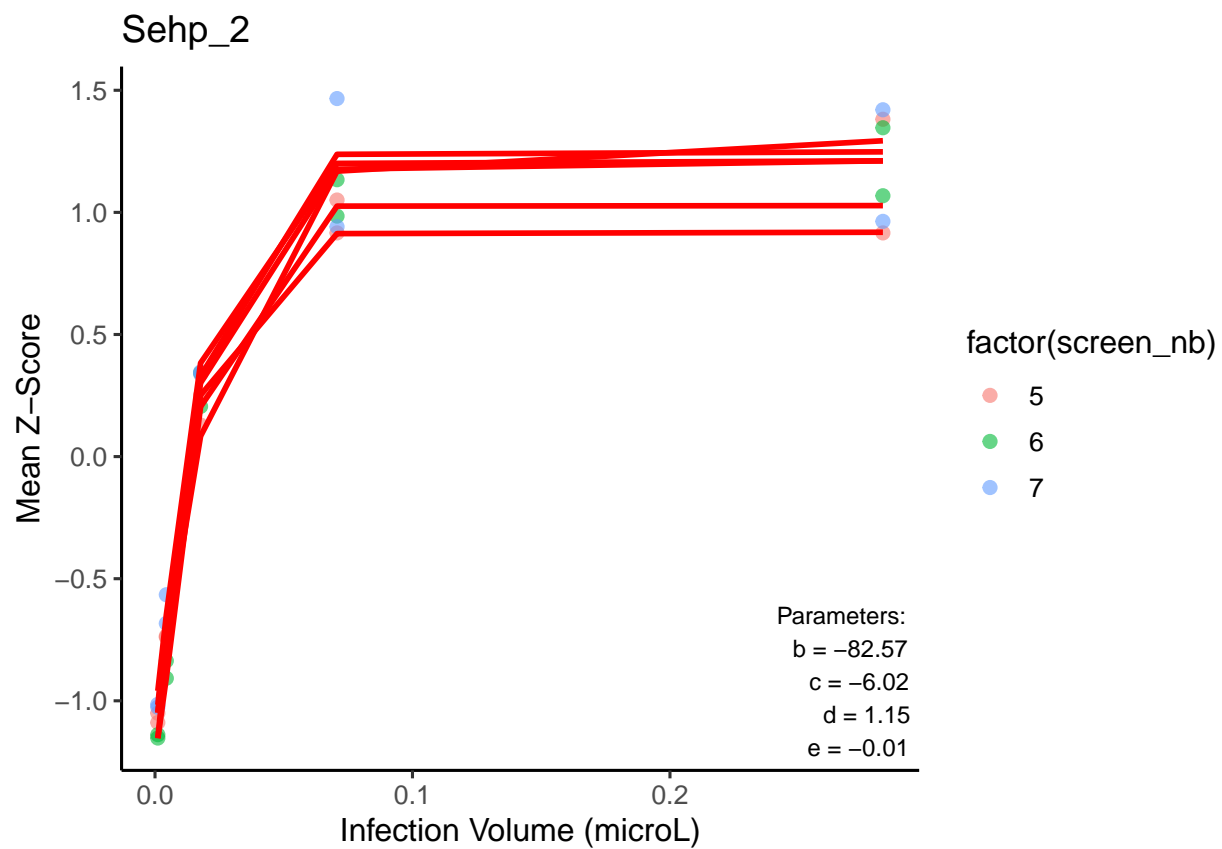
\$Sebn_4



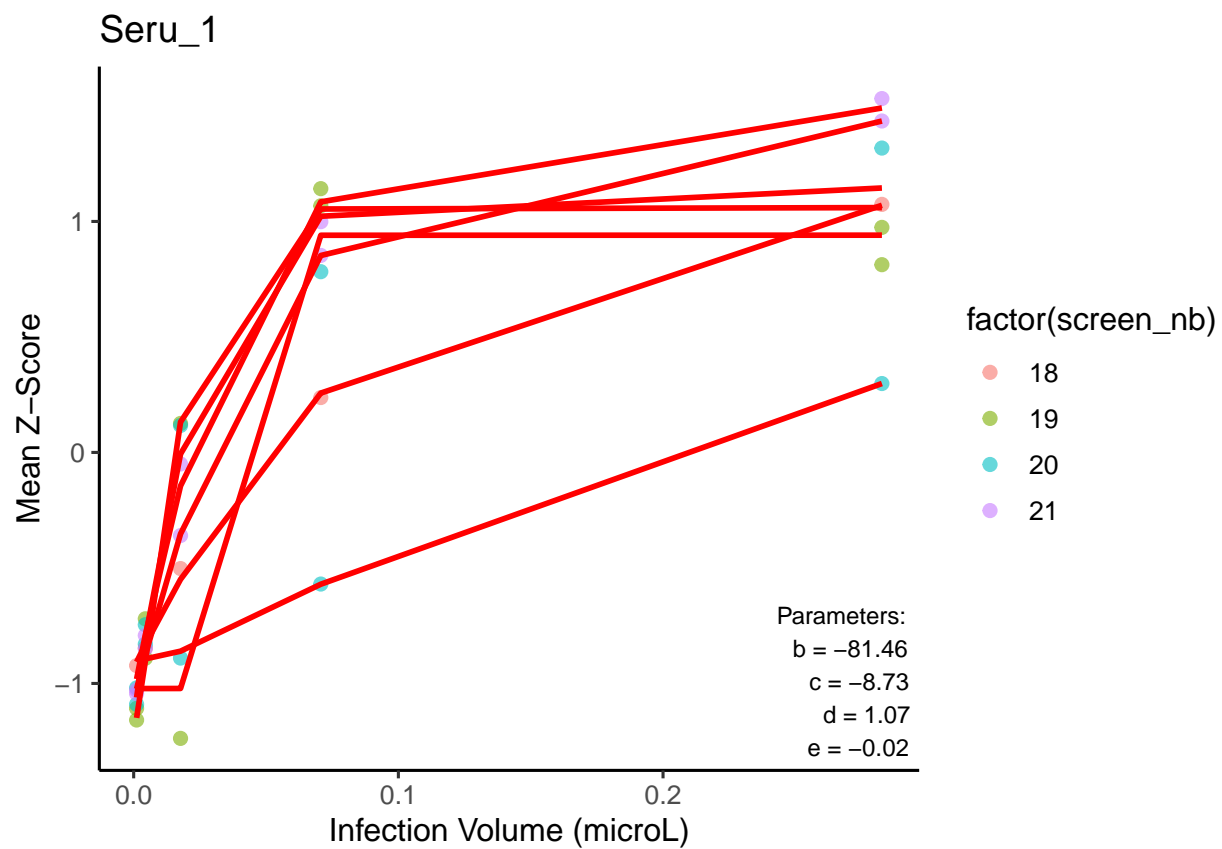
 ## \$Sebz_1



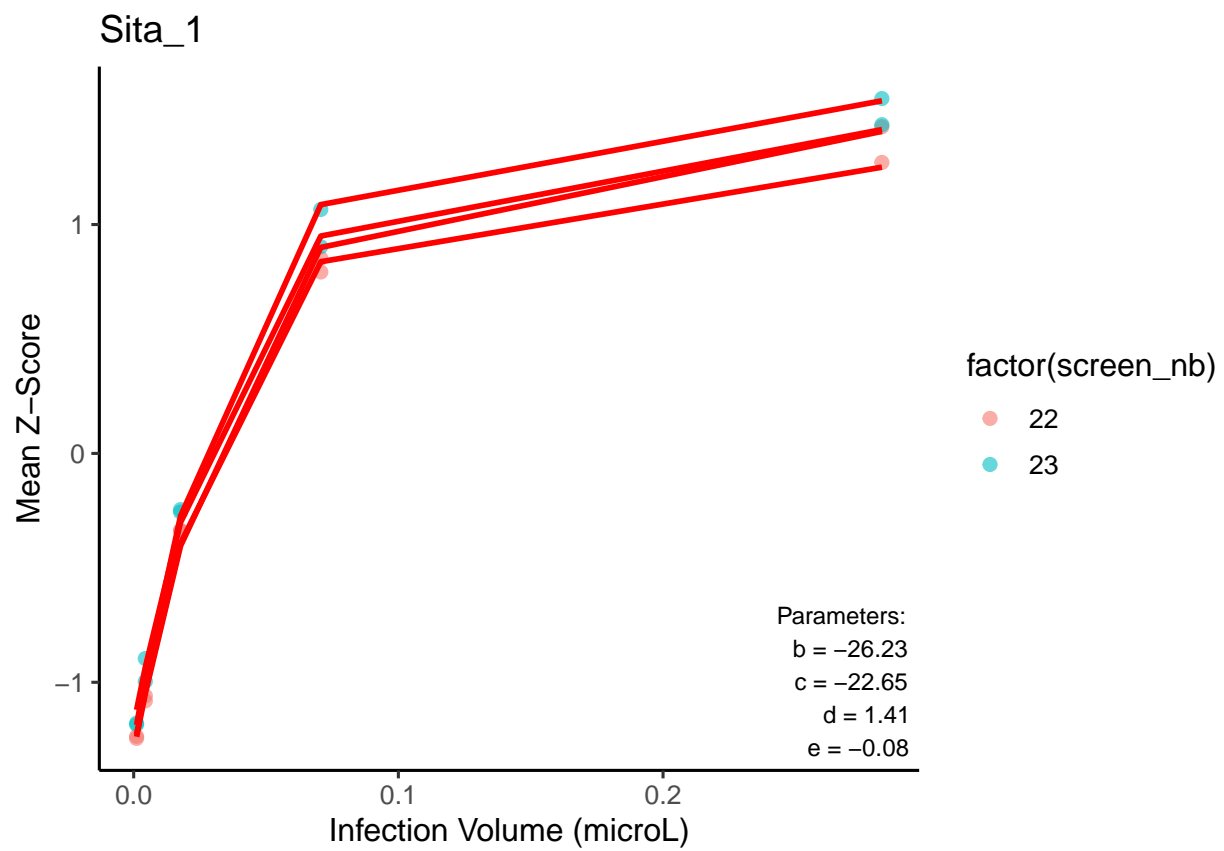
 ## \$Sehp_2



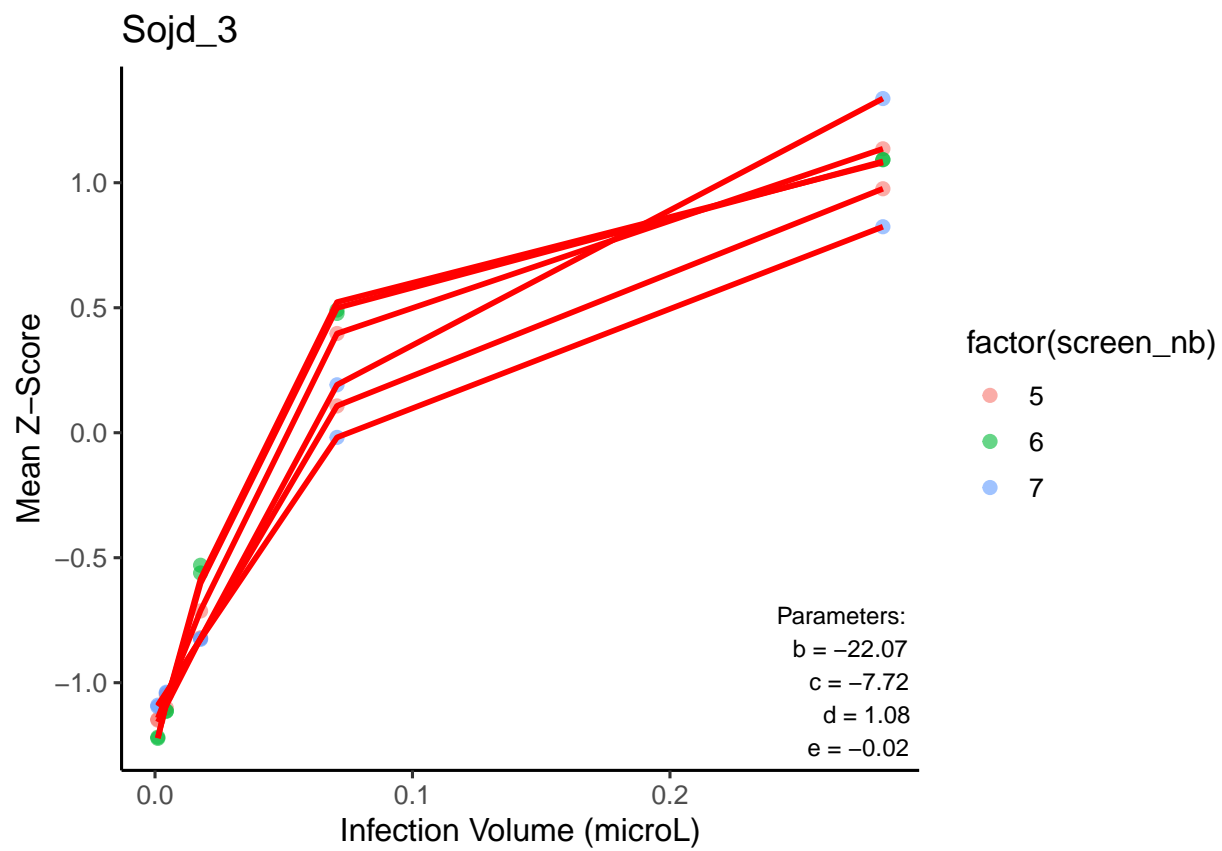
\$Seru_1



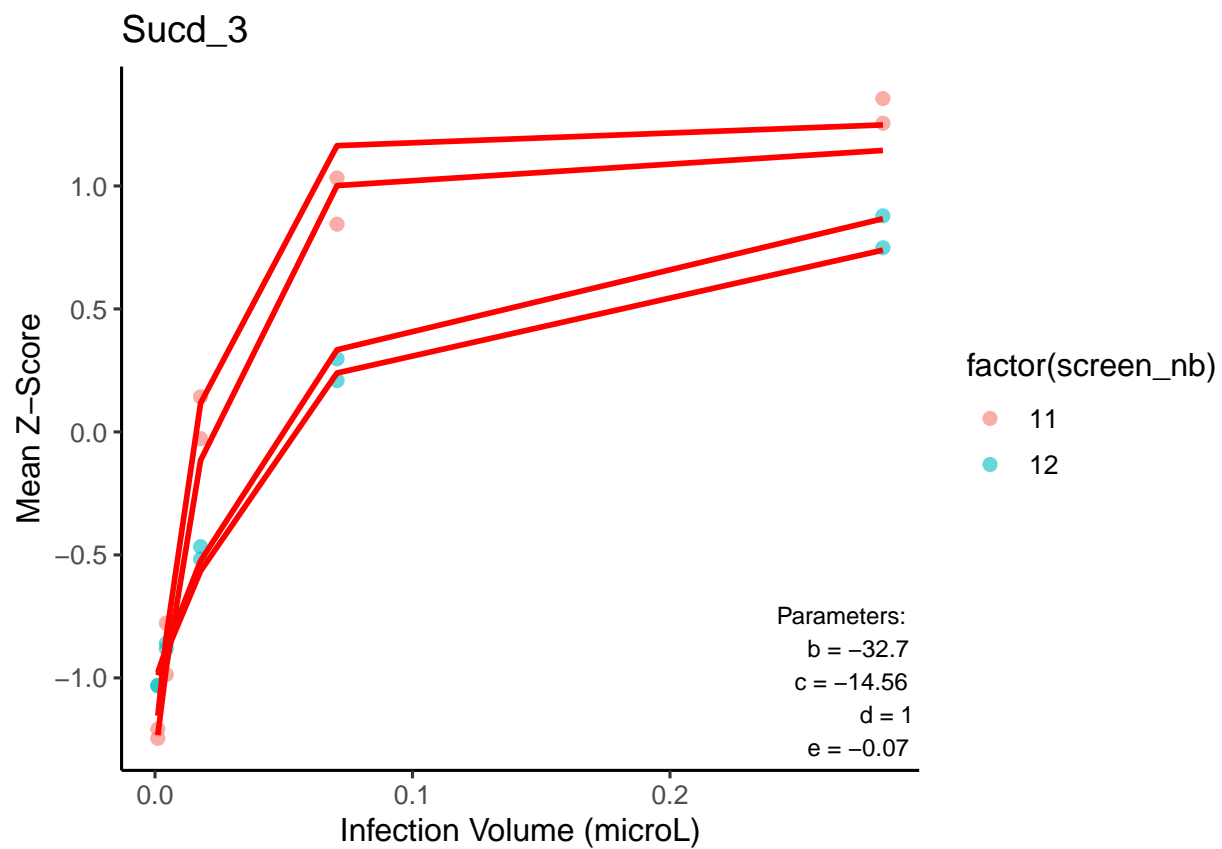
 ## \$Sita_1



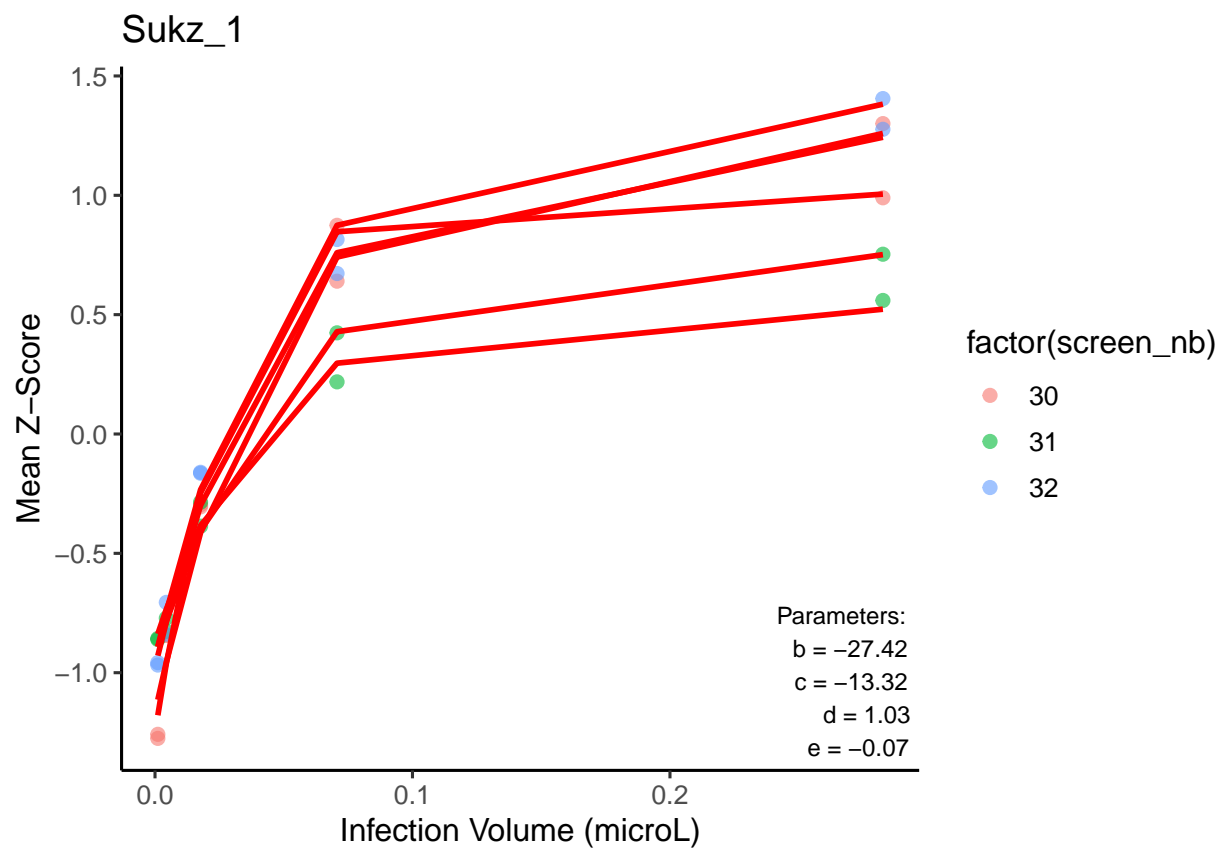
\$Sojd_3



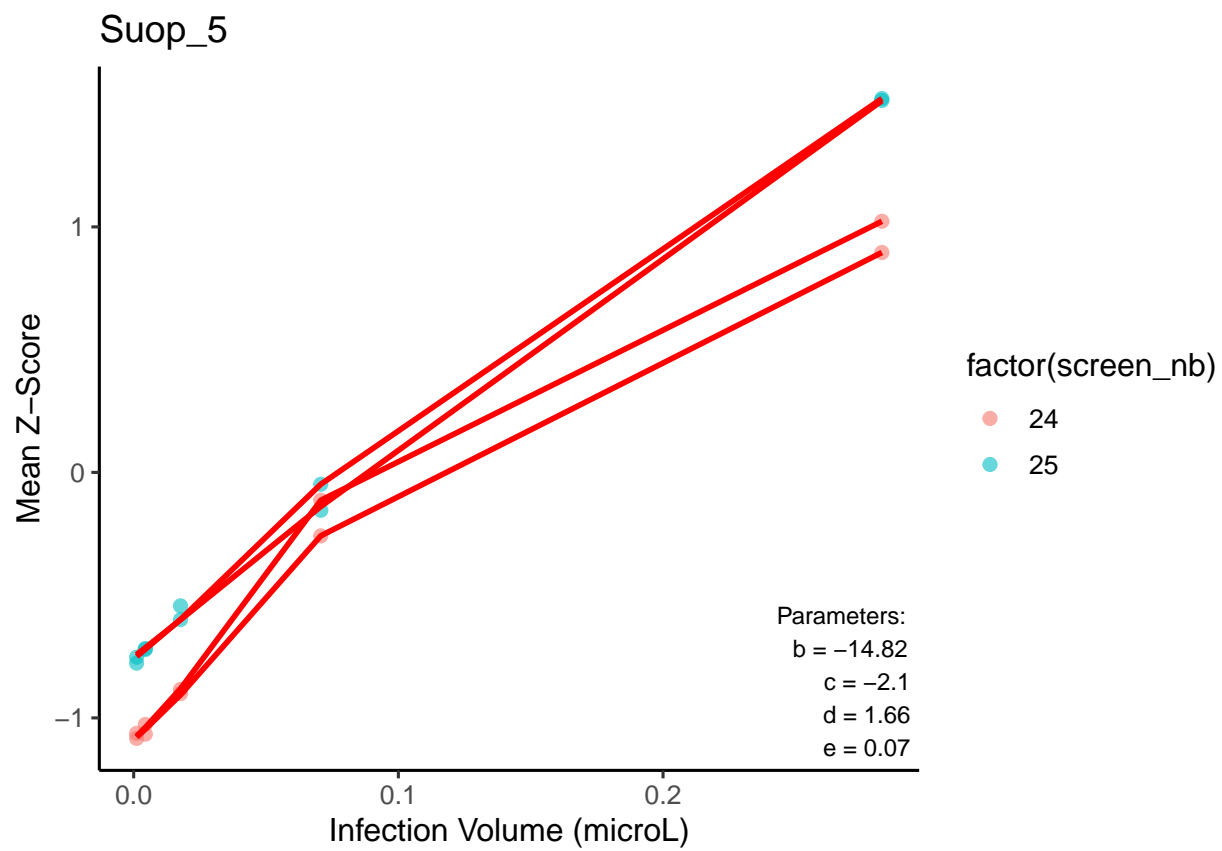
 ## \$Succd_3



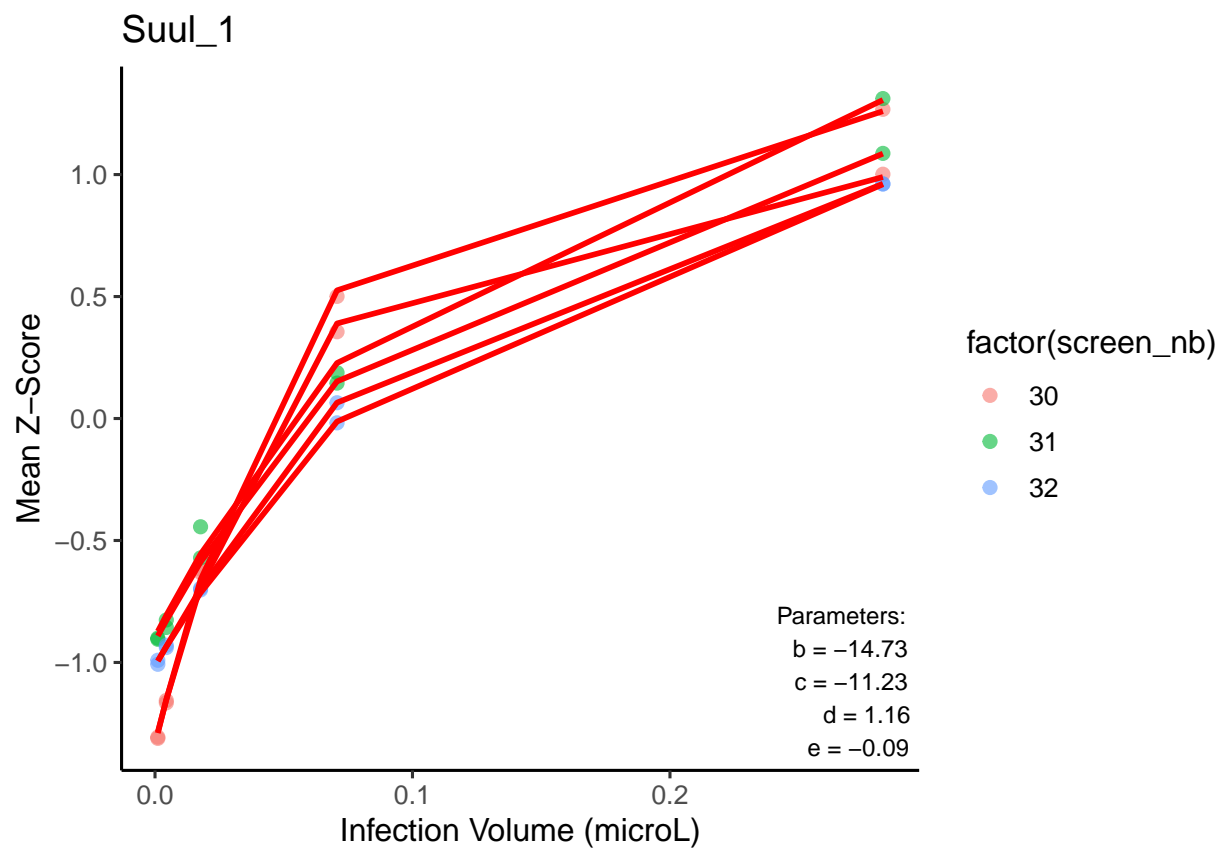
 ## \$Sukz_1



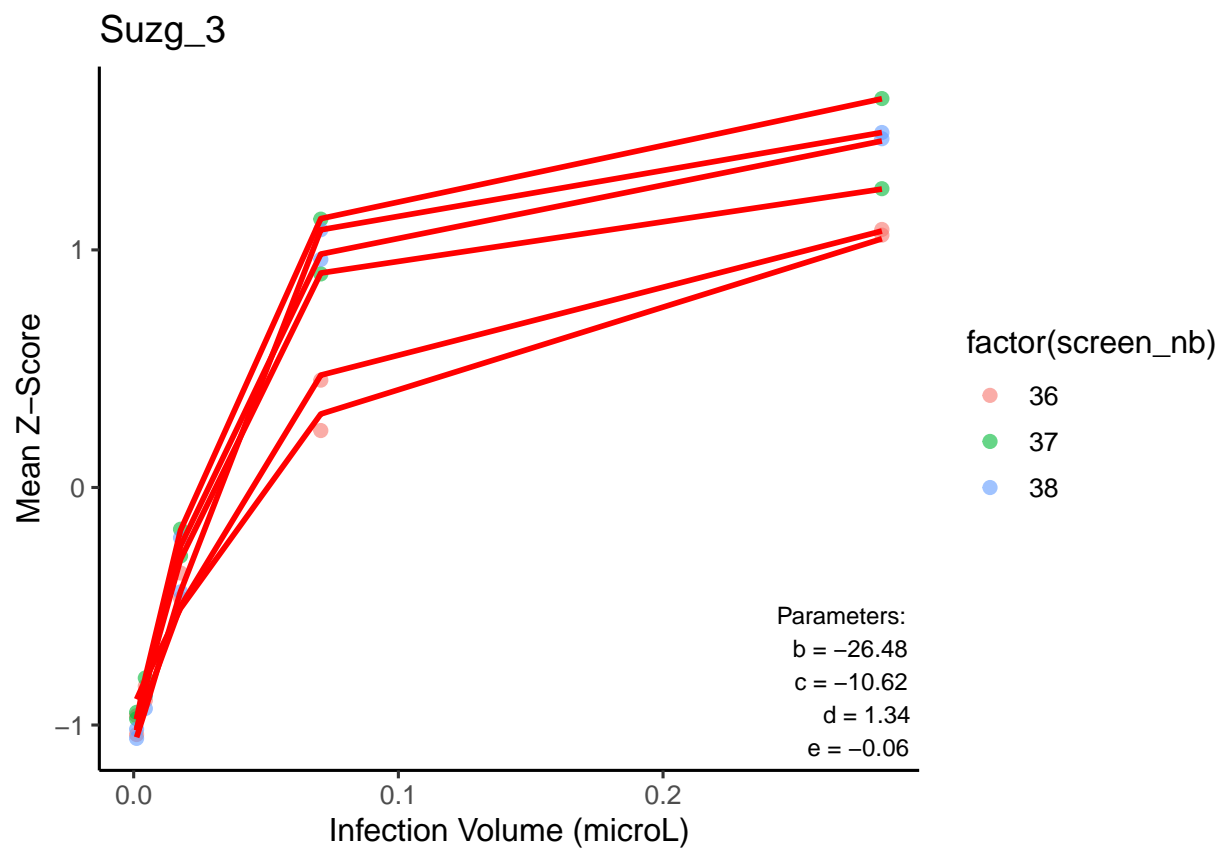
 ## \$Suop_5



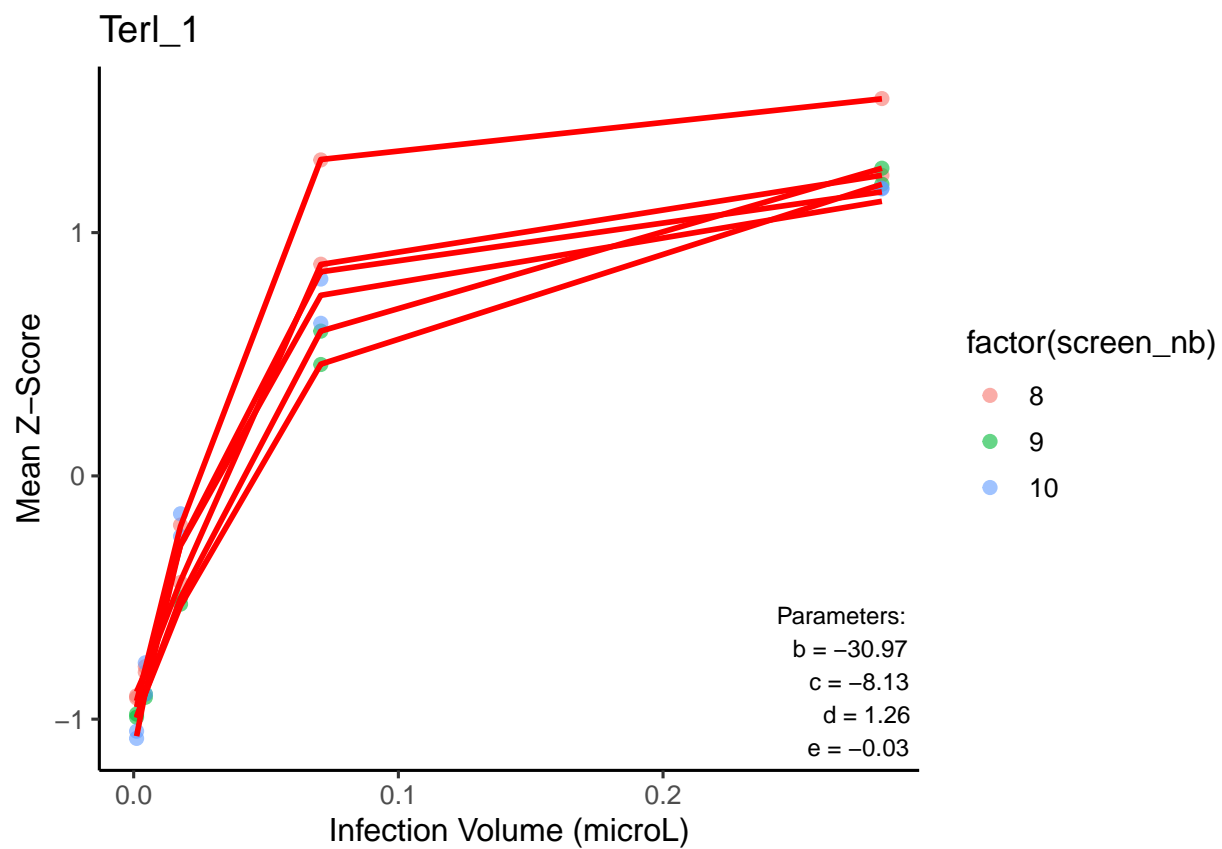
\$Suul_1



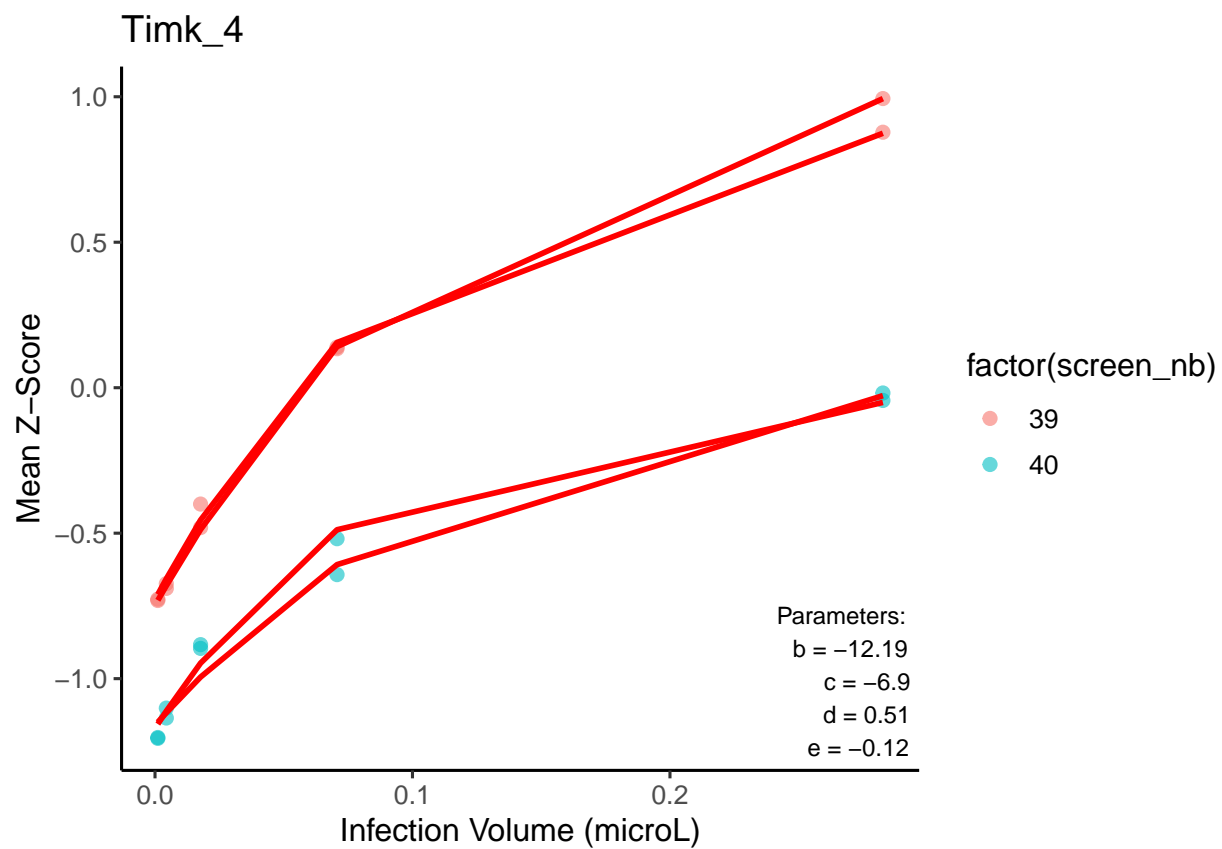
\$Suzg_3



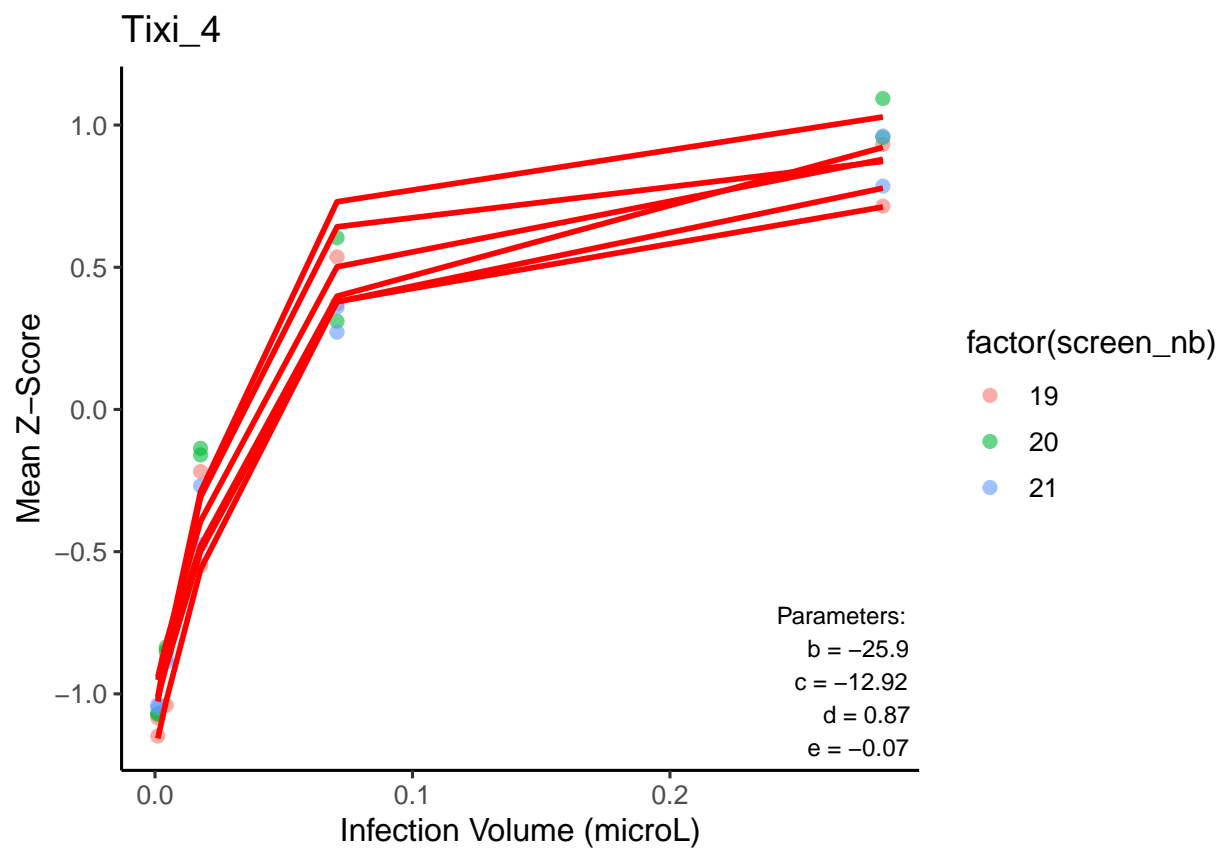
\$Terl_1



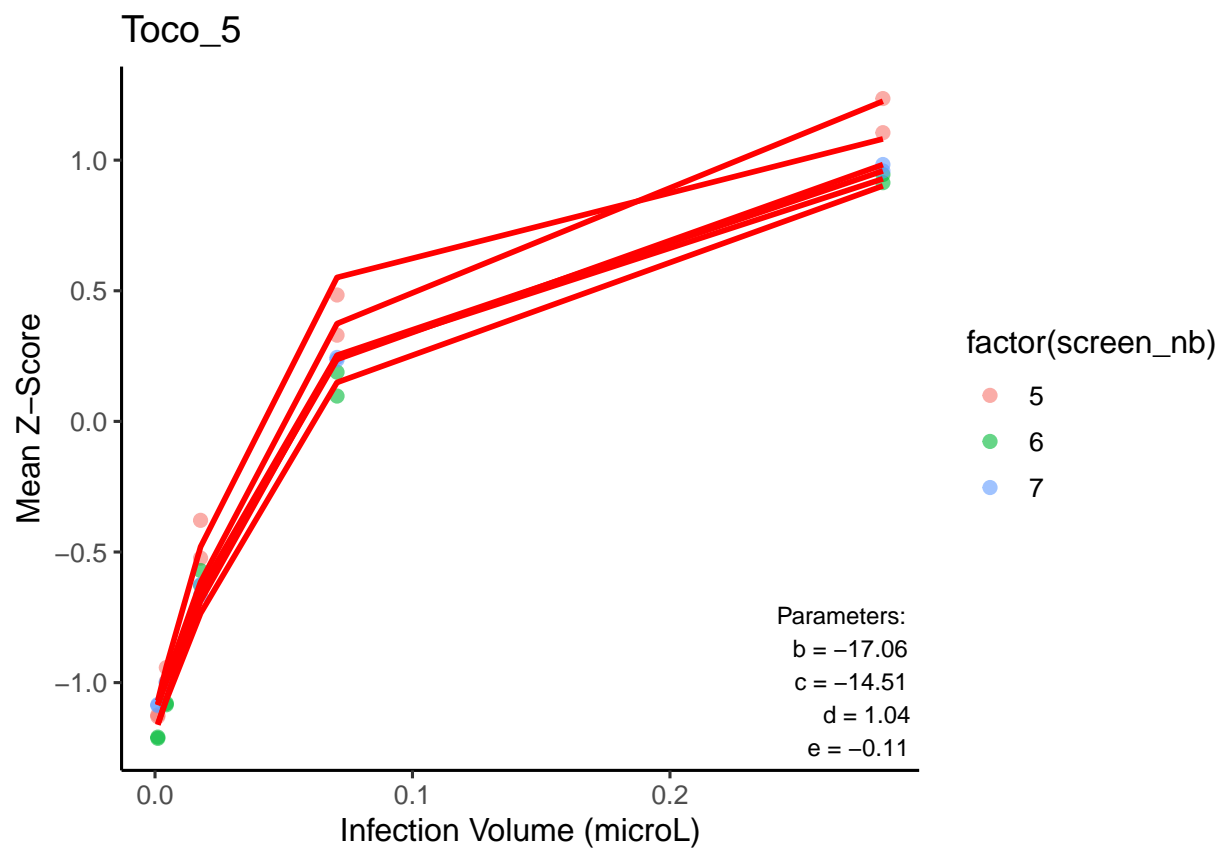
 ## \$Timk_4



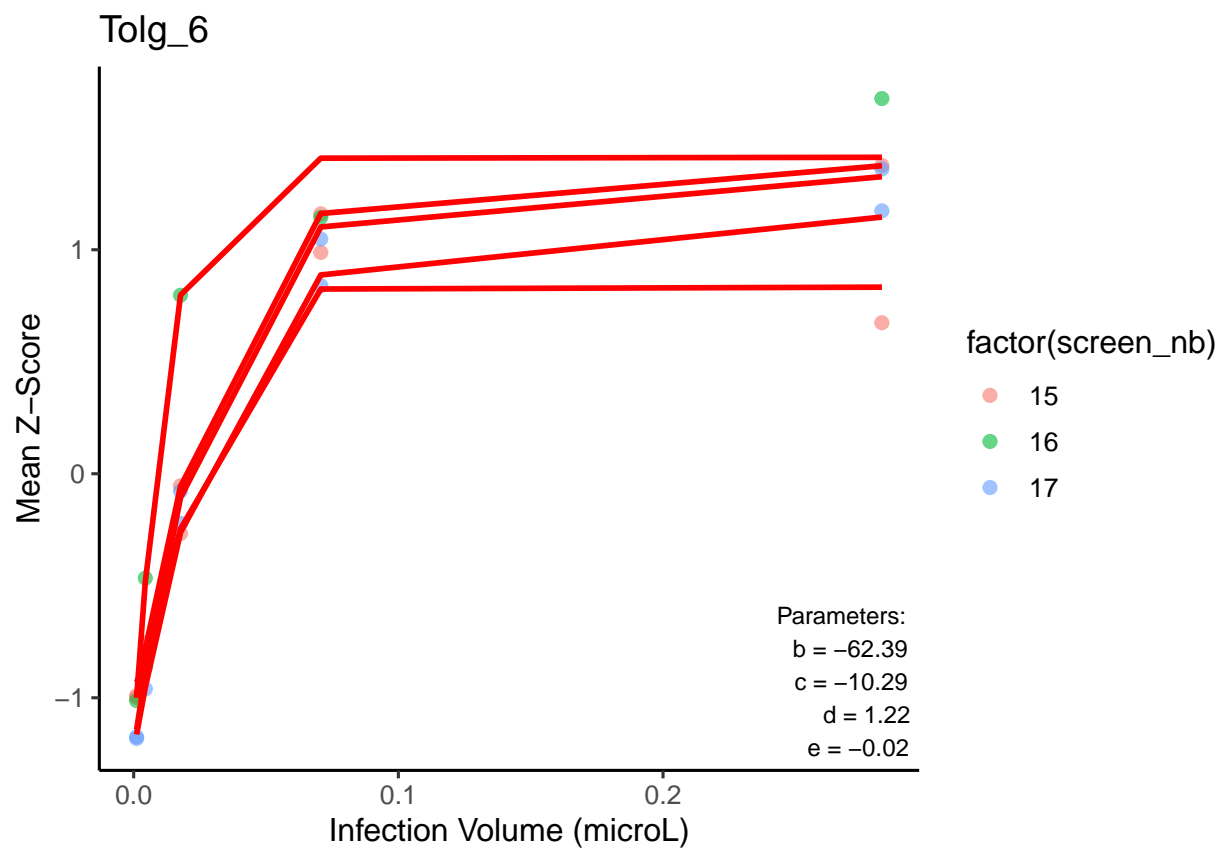
```
##  
## $Tixi_4
```



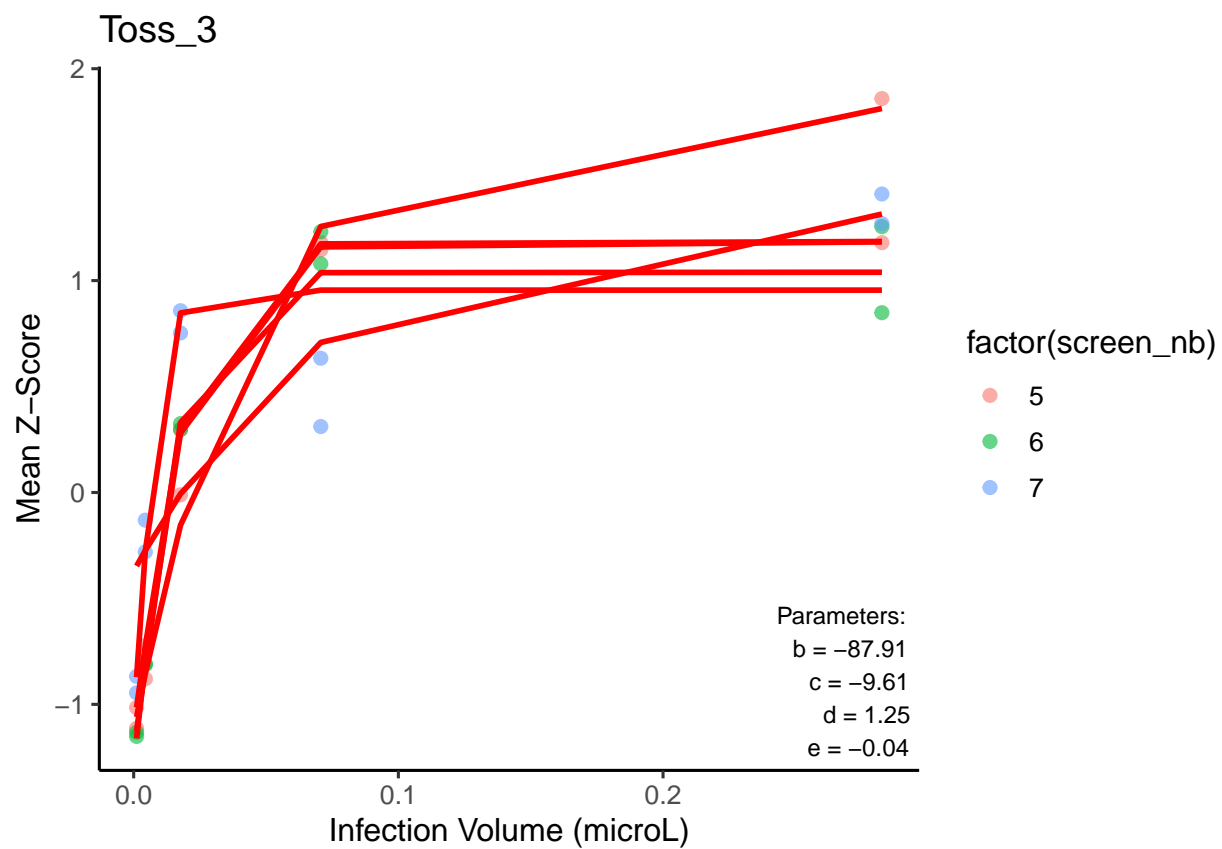
 ## \$Toco_5



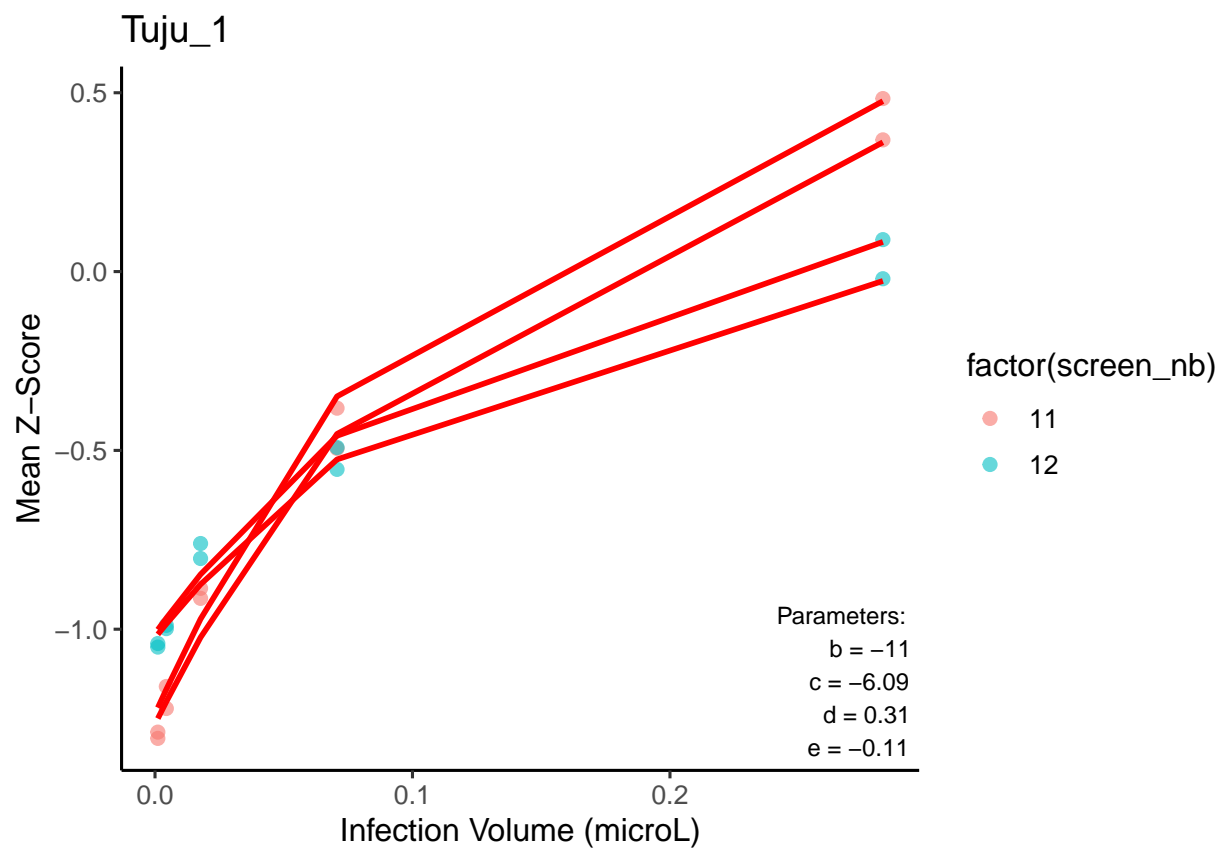
 ## \$Tolg_6



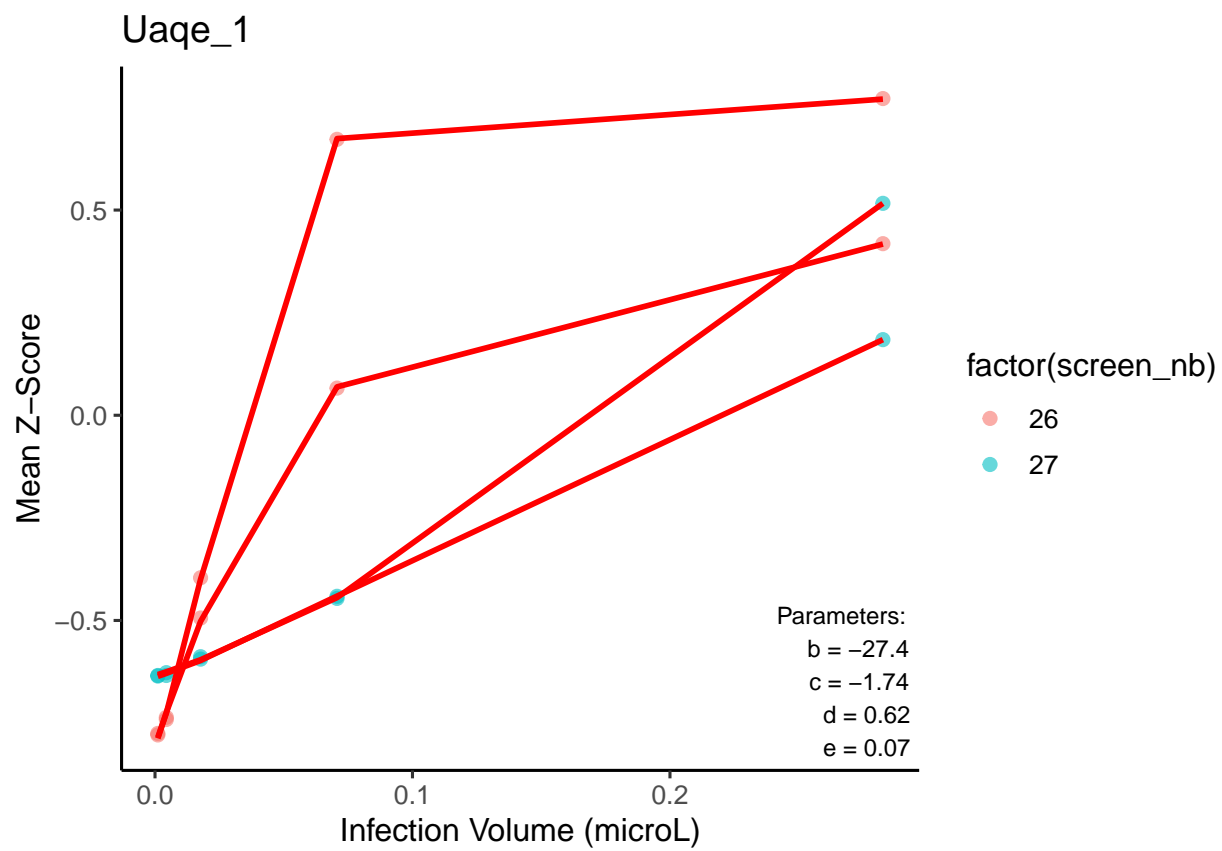
 ## \$Toss_3



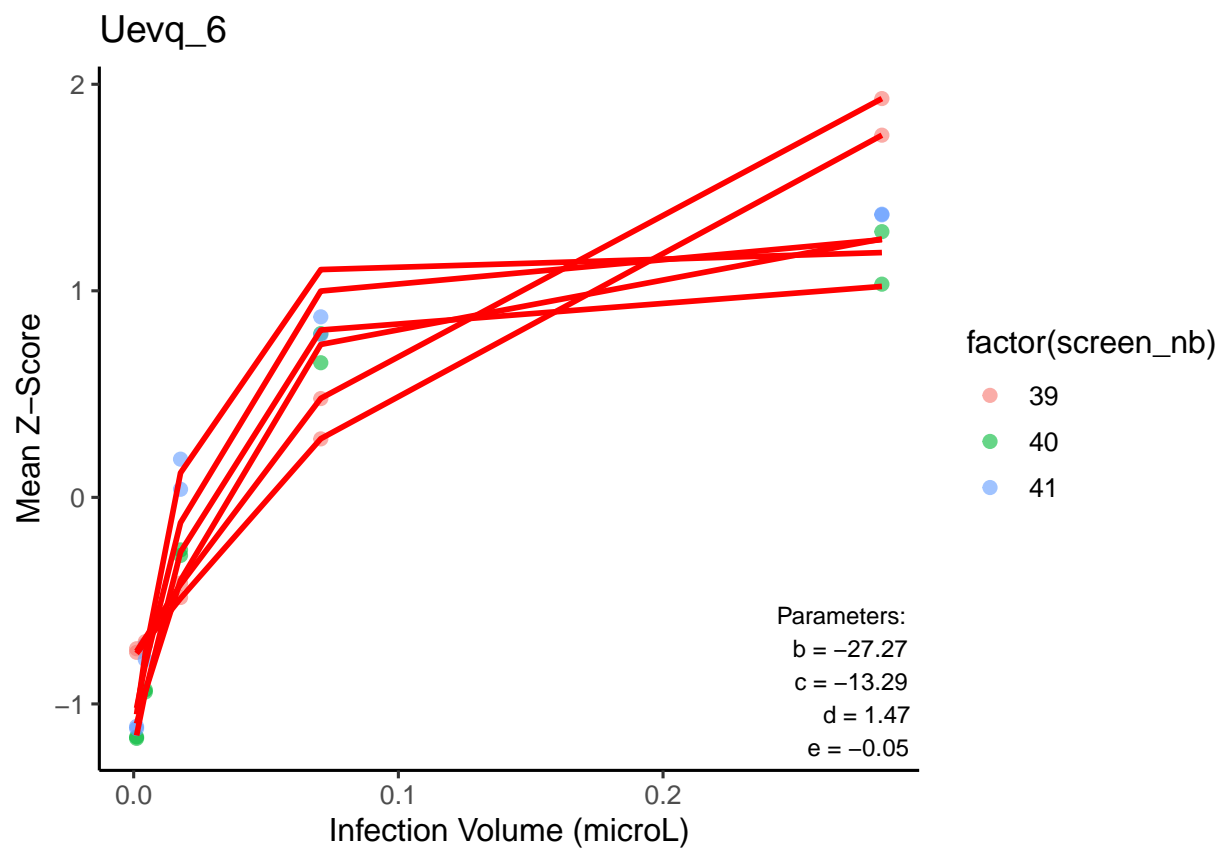
 ## \$Tuju_1



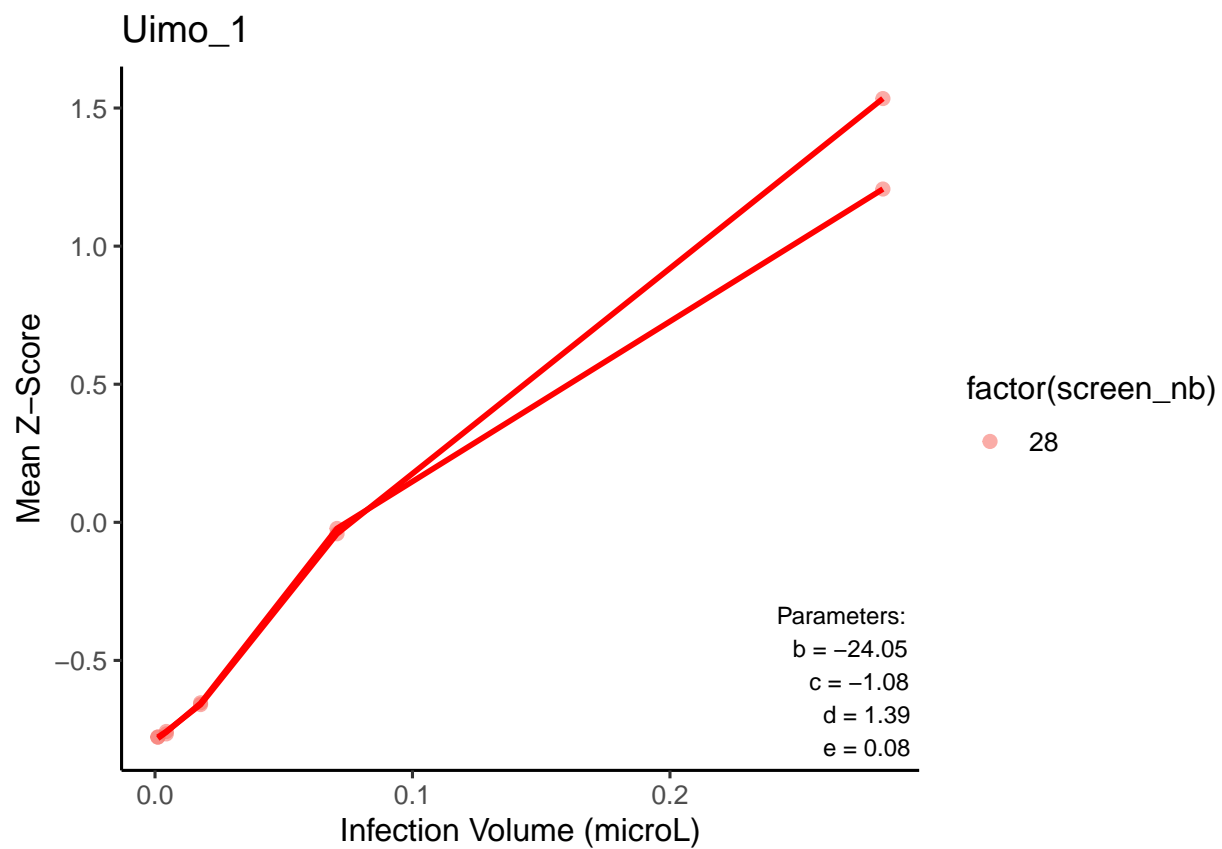
\$Uaqe_1



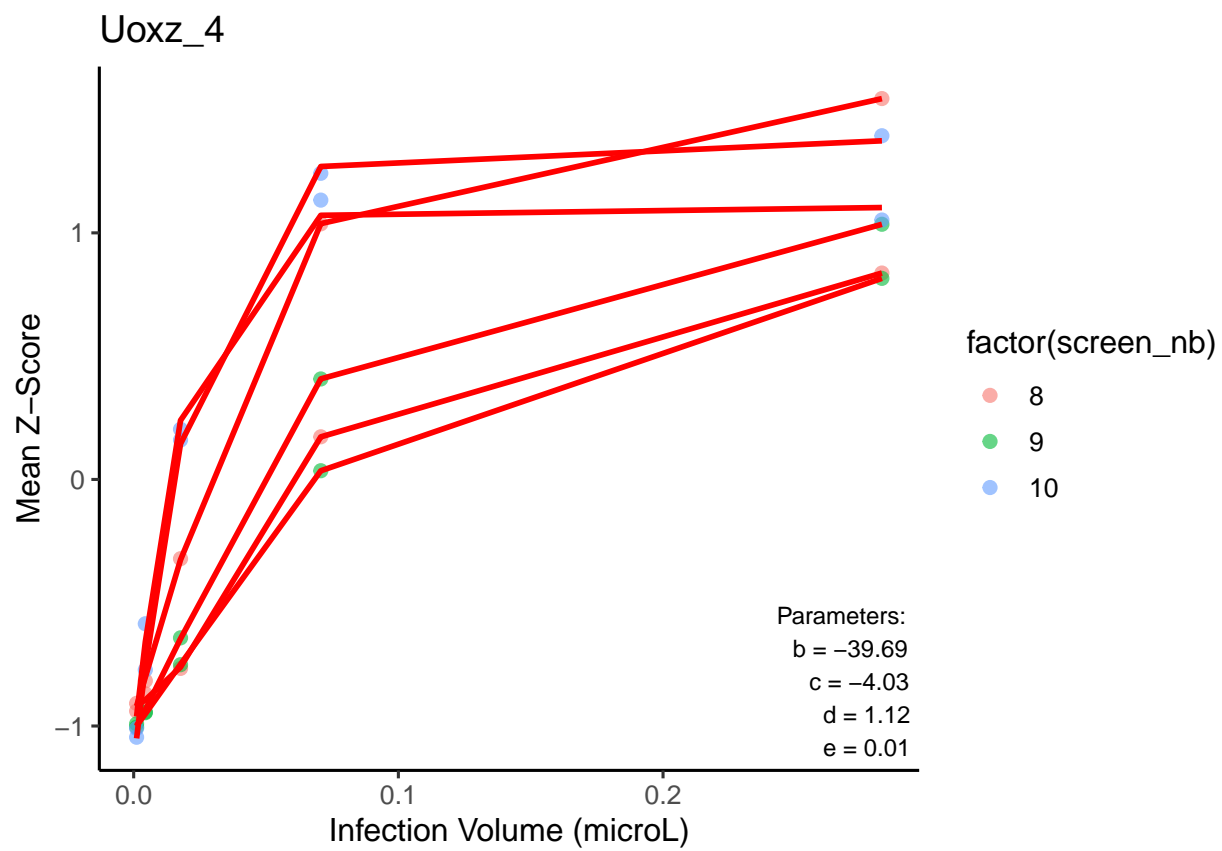
 ## \$Uevq_6



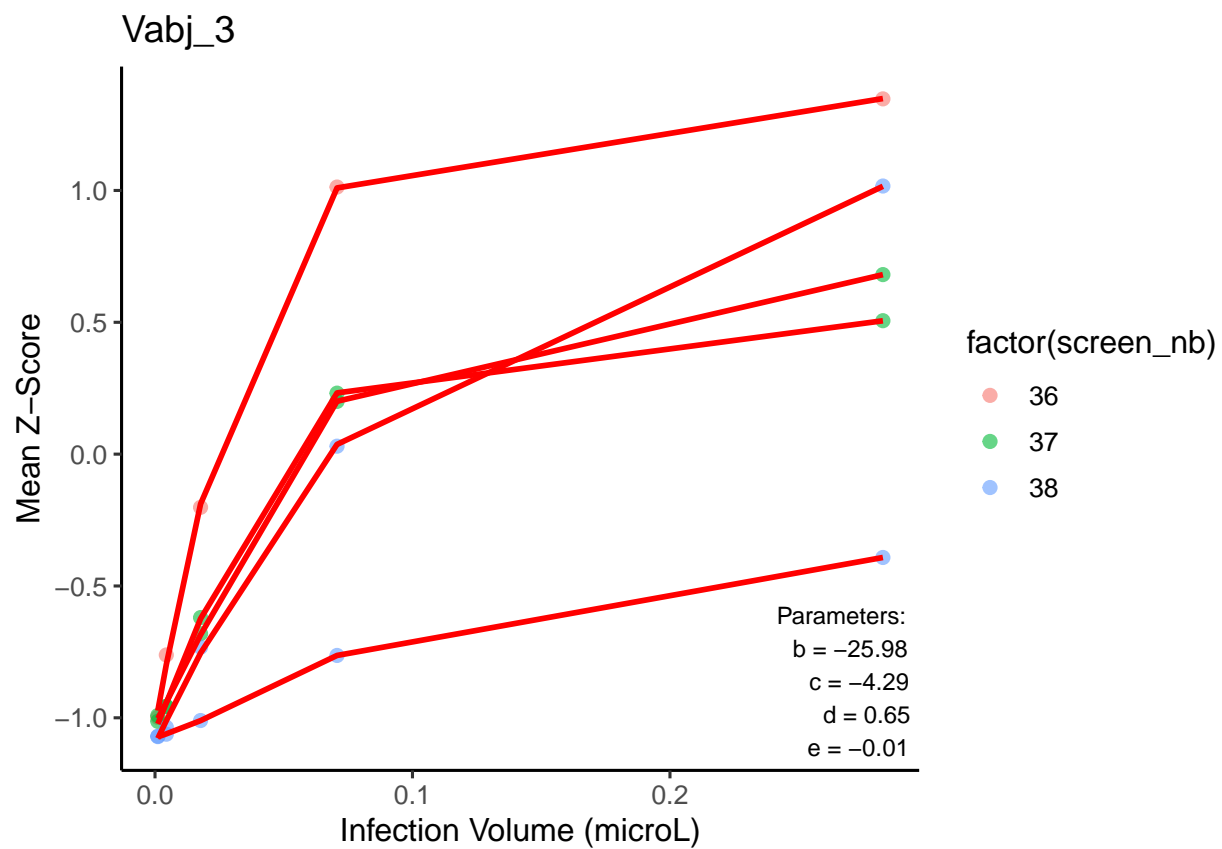
 ## \$Uimo_1



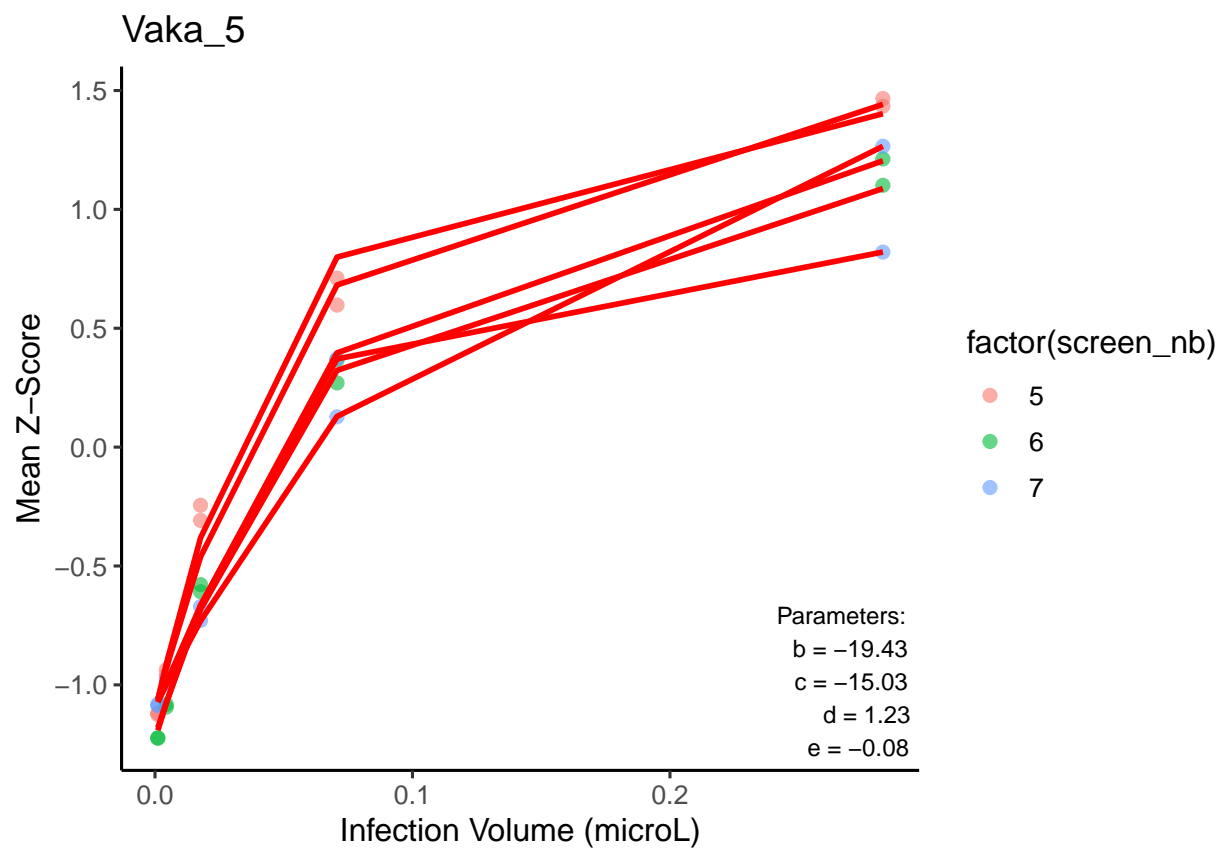
\$Uoxz_4



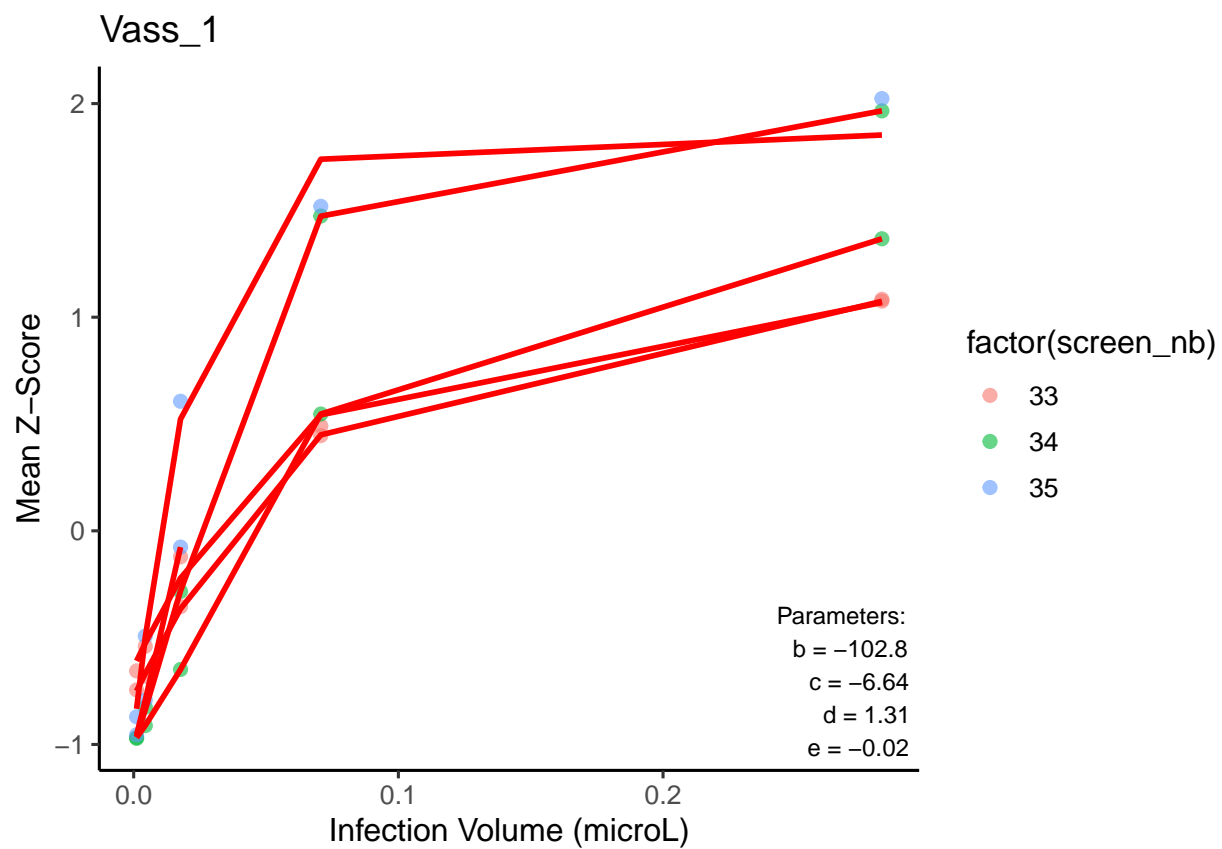
 ## \$Vabj_3



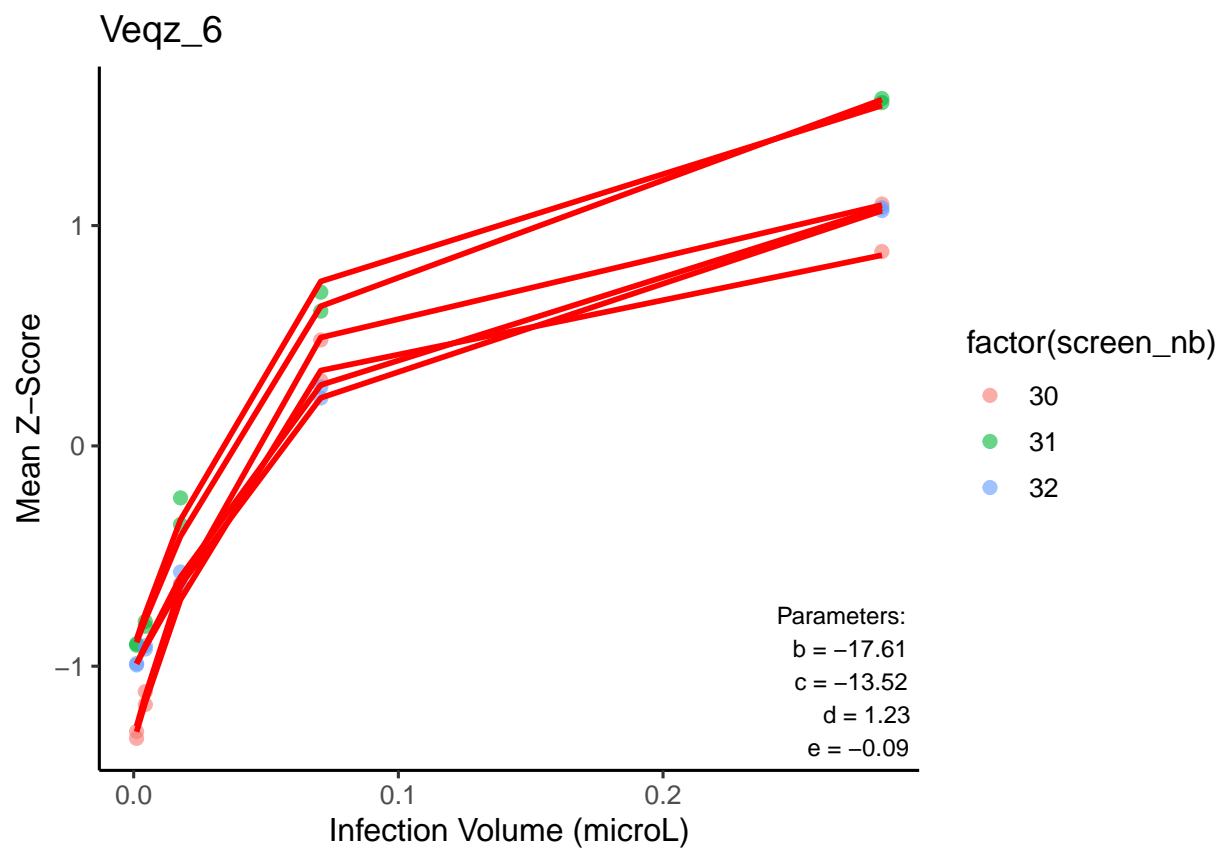
 ## \$Vaka_5



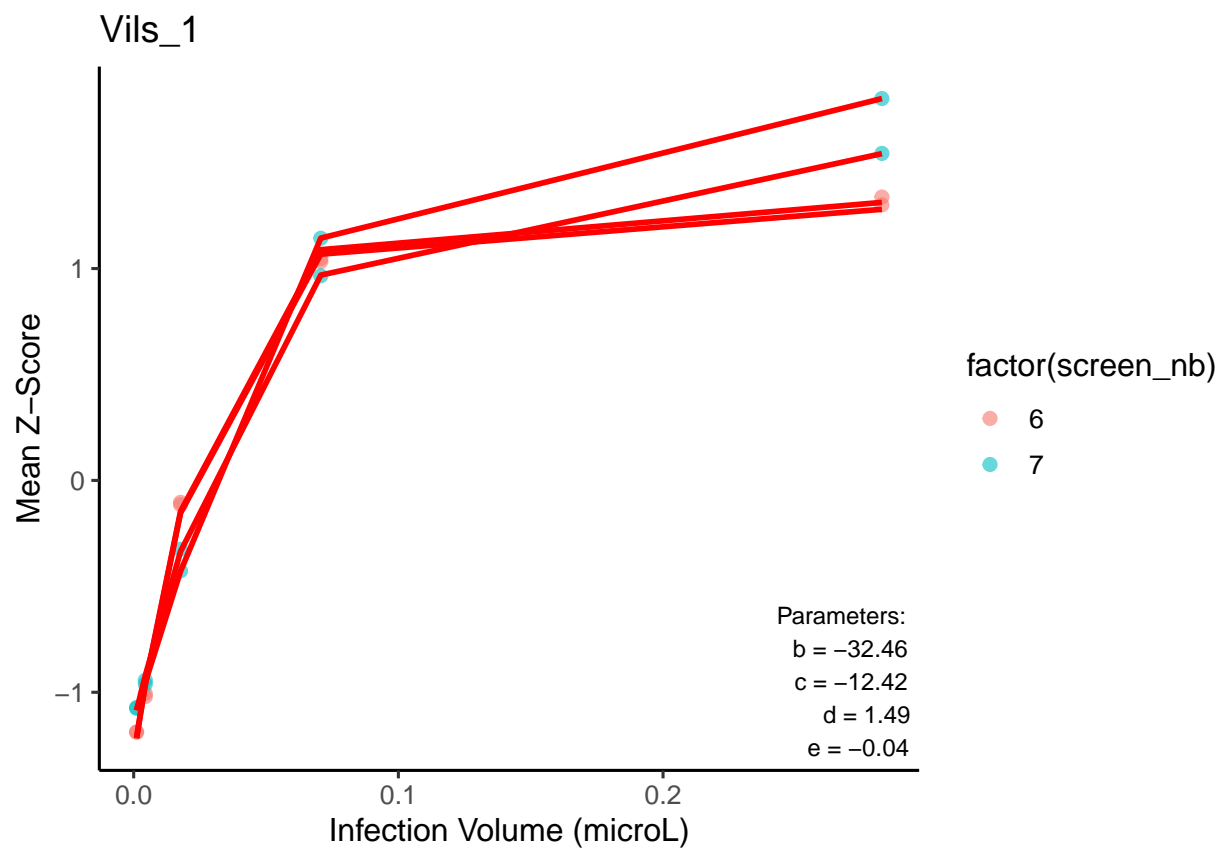
\$Vass_1



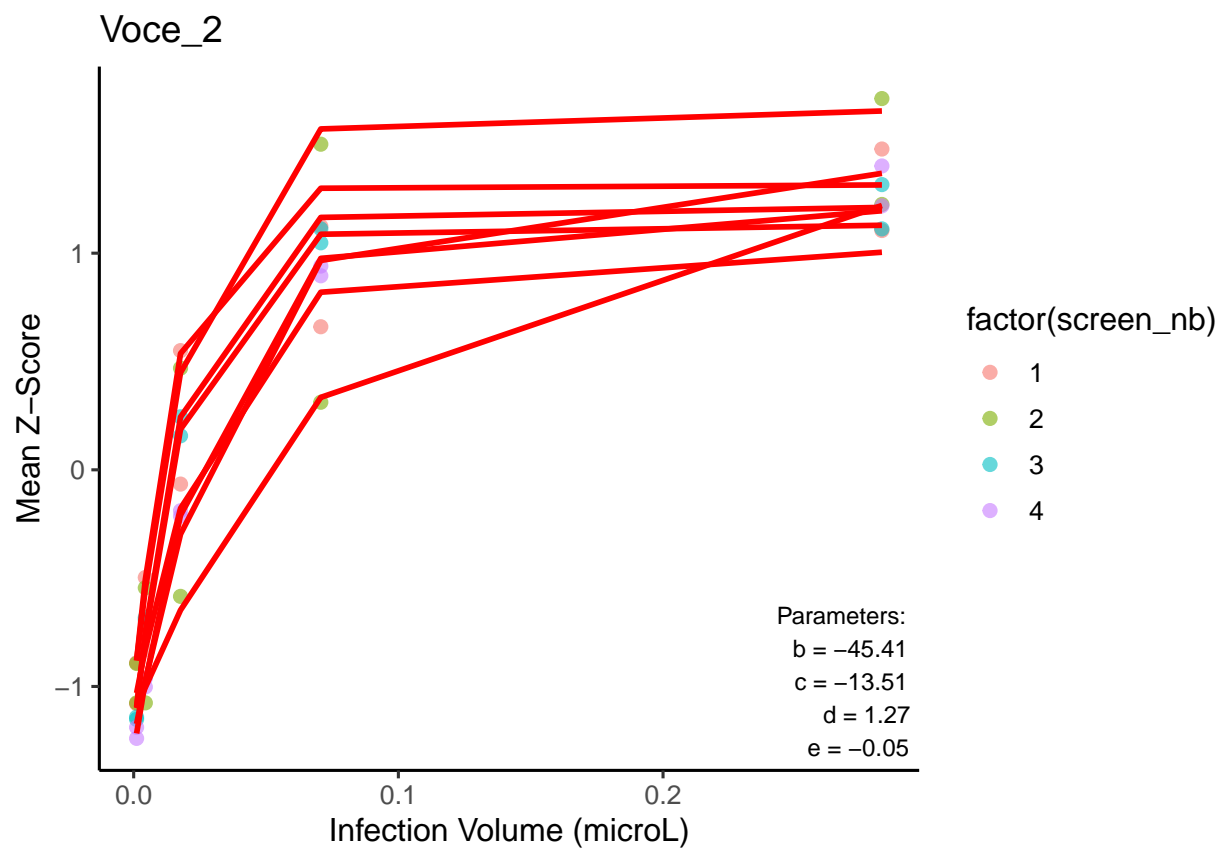
 ## \$Veqz_6



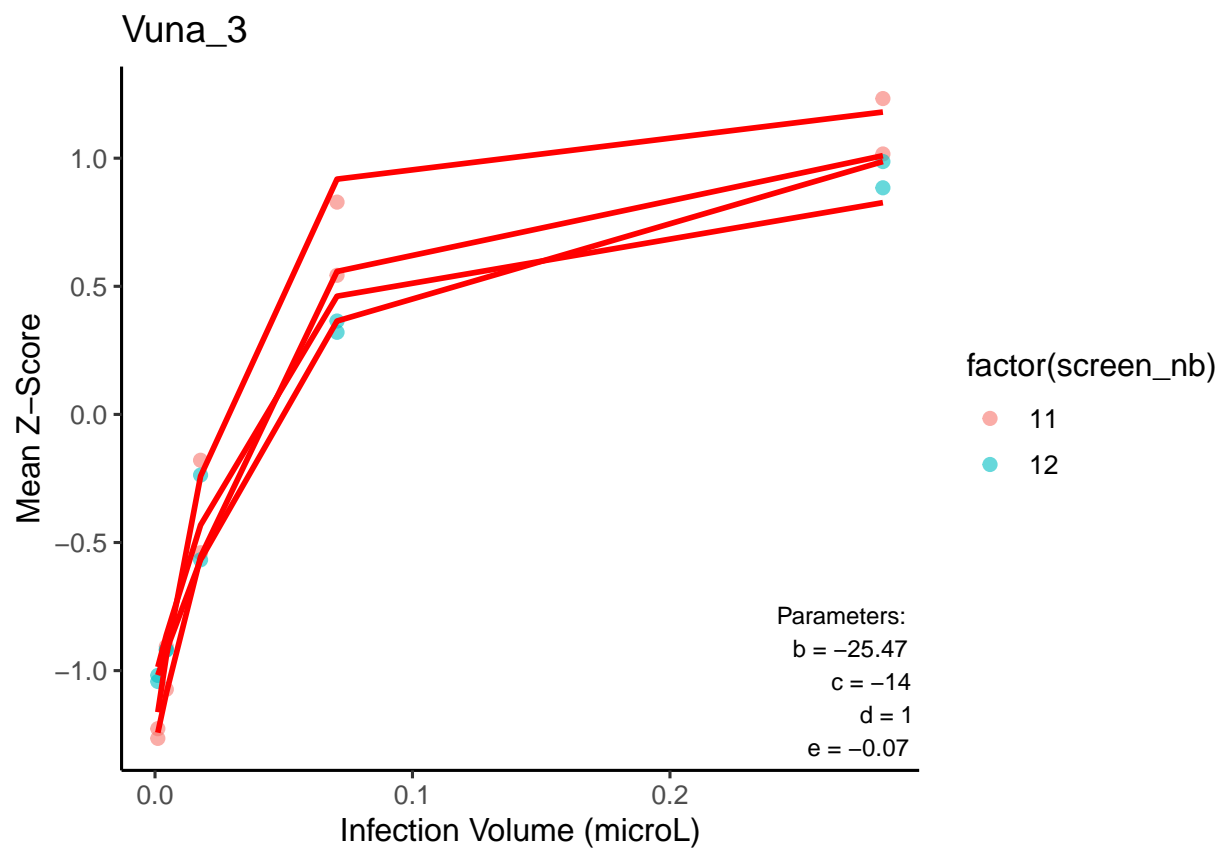
 ## \$Vils_1



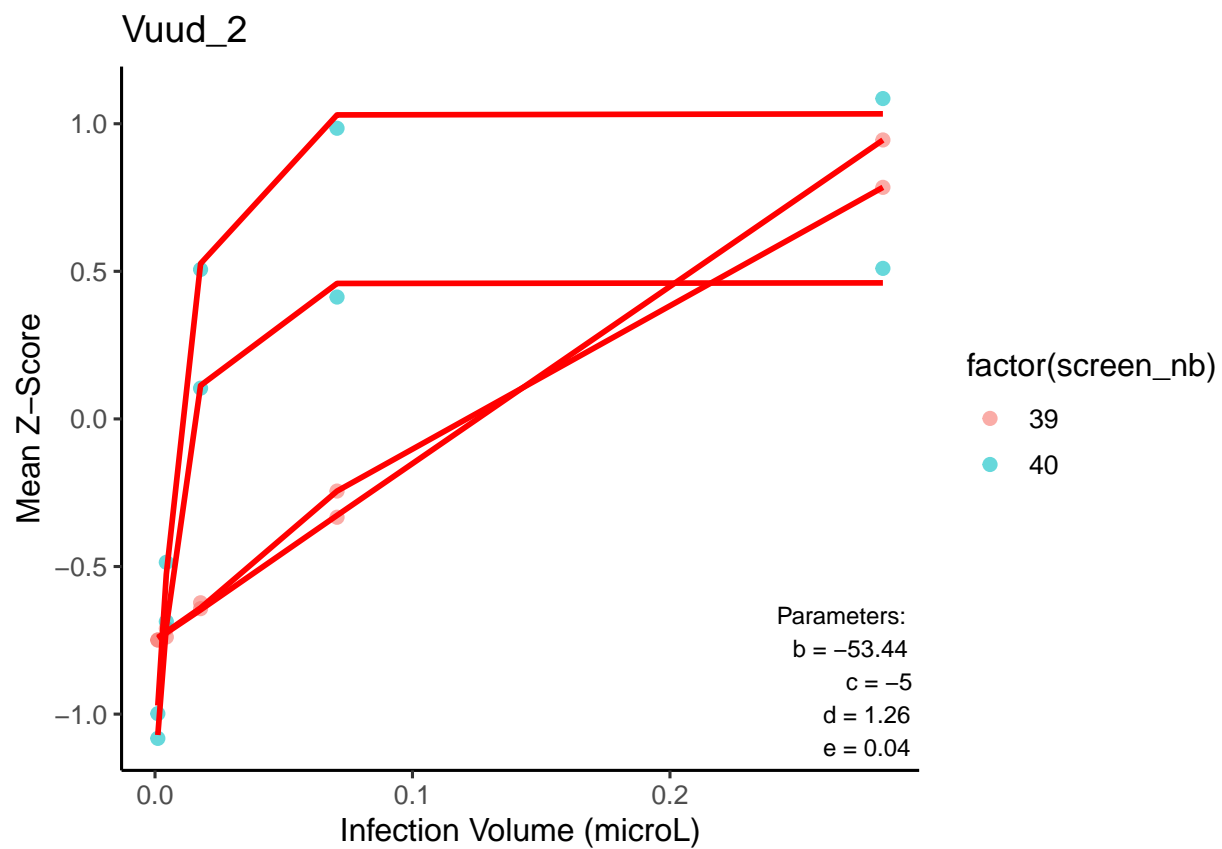
 ## \$Voce_2



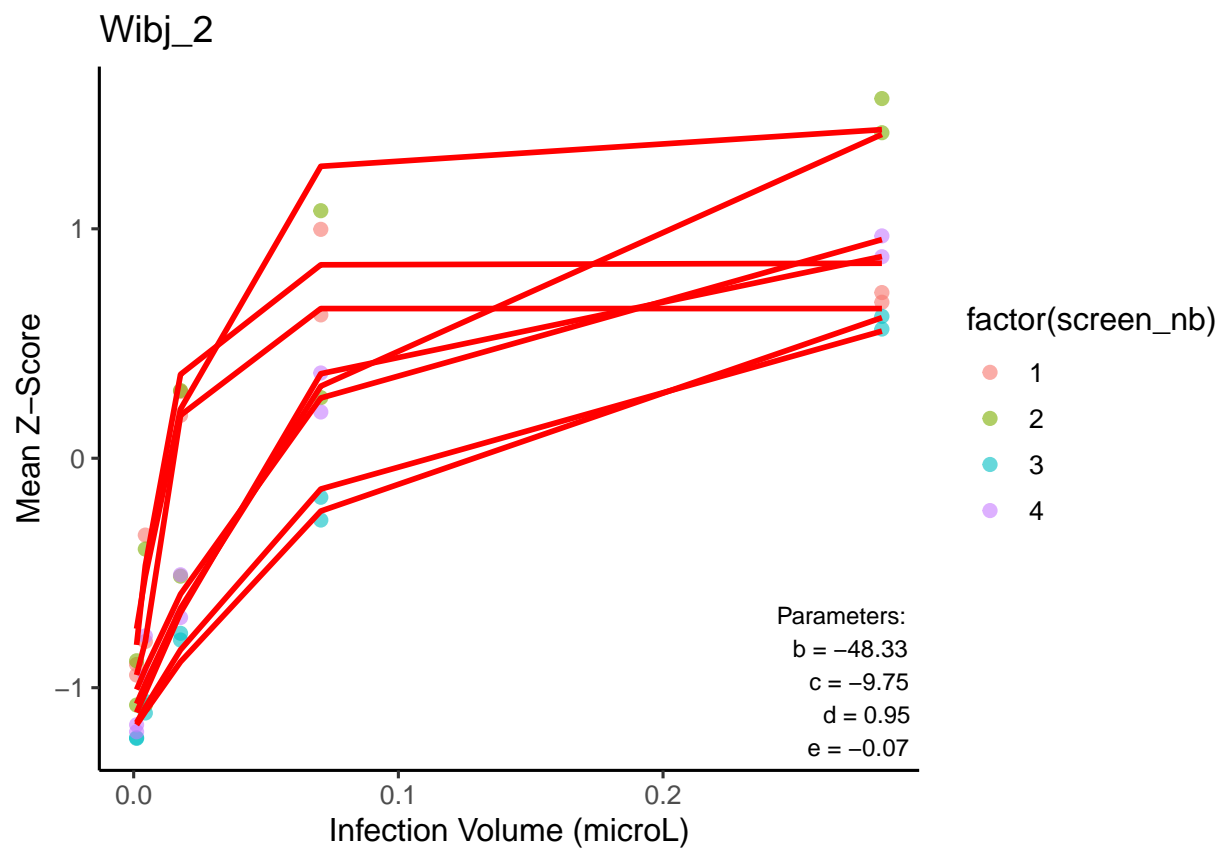
 ## \$Vuna_3



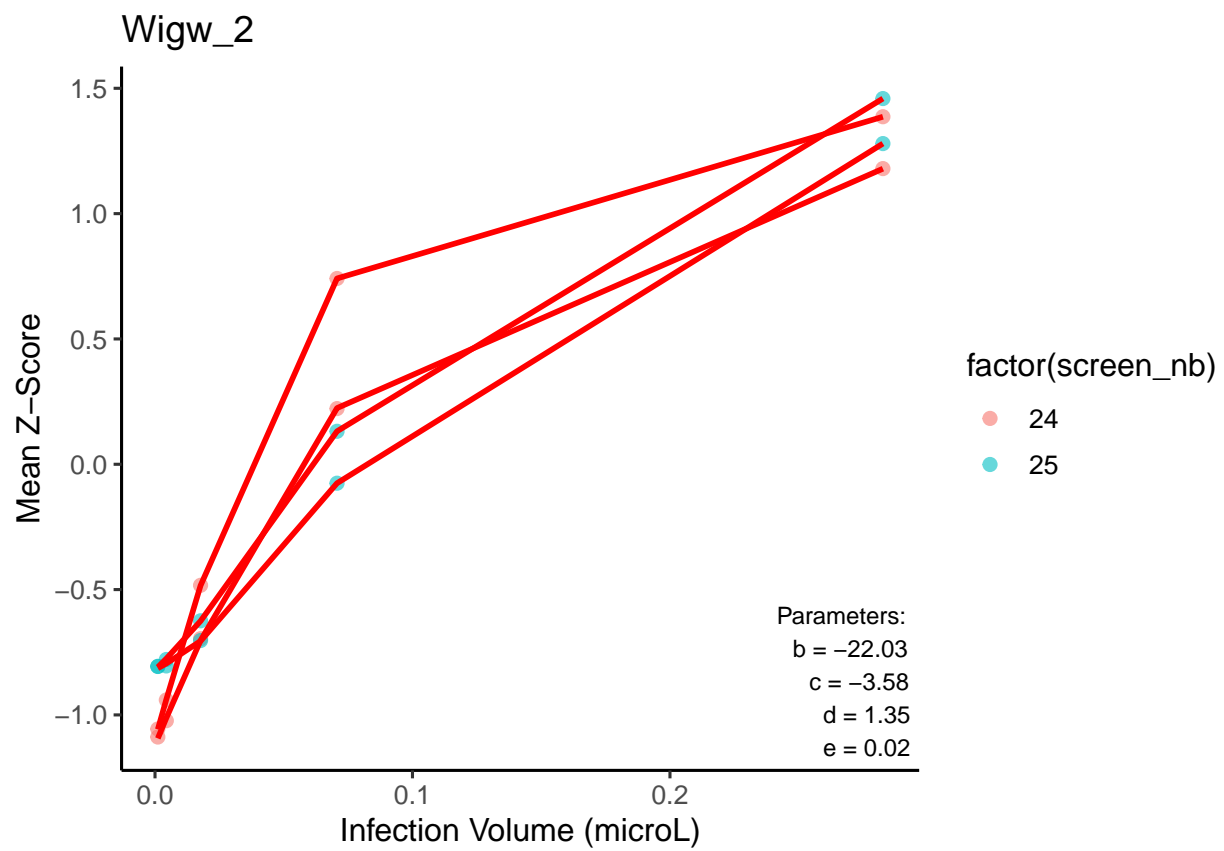
 ## \$Vuud_2



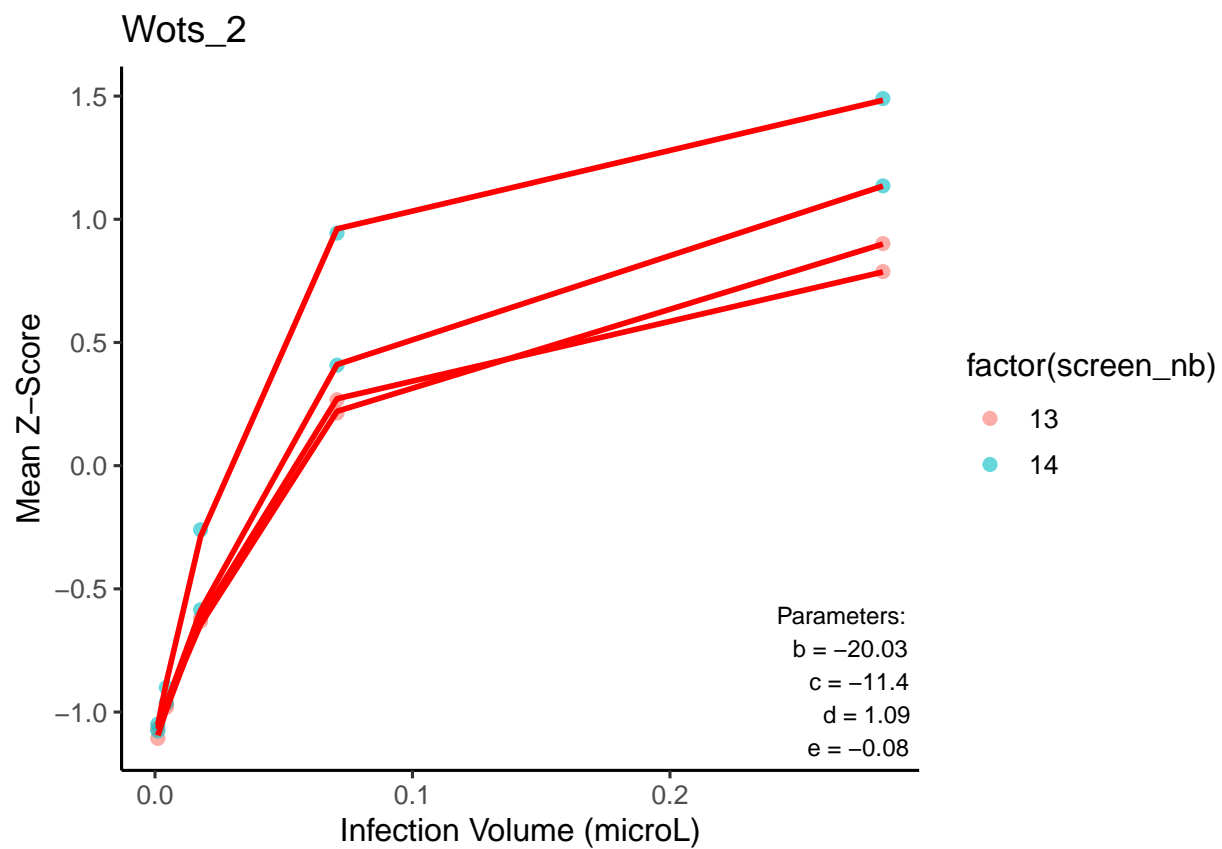
```
##
## $Wibj_2
```



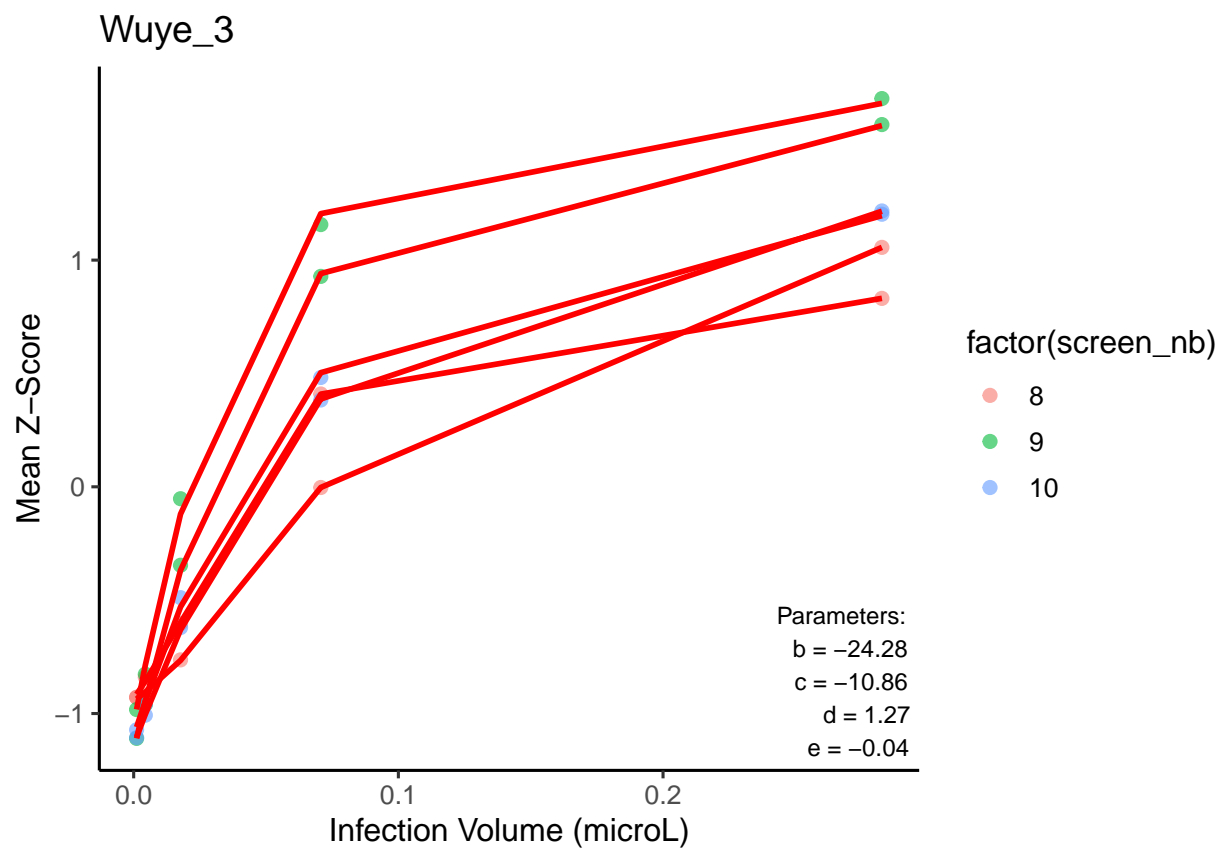
```
##
## $Wigw_2
```

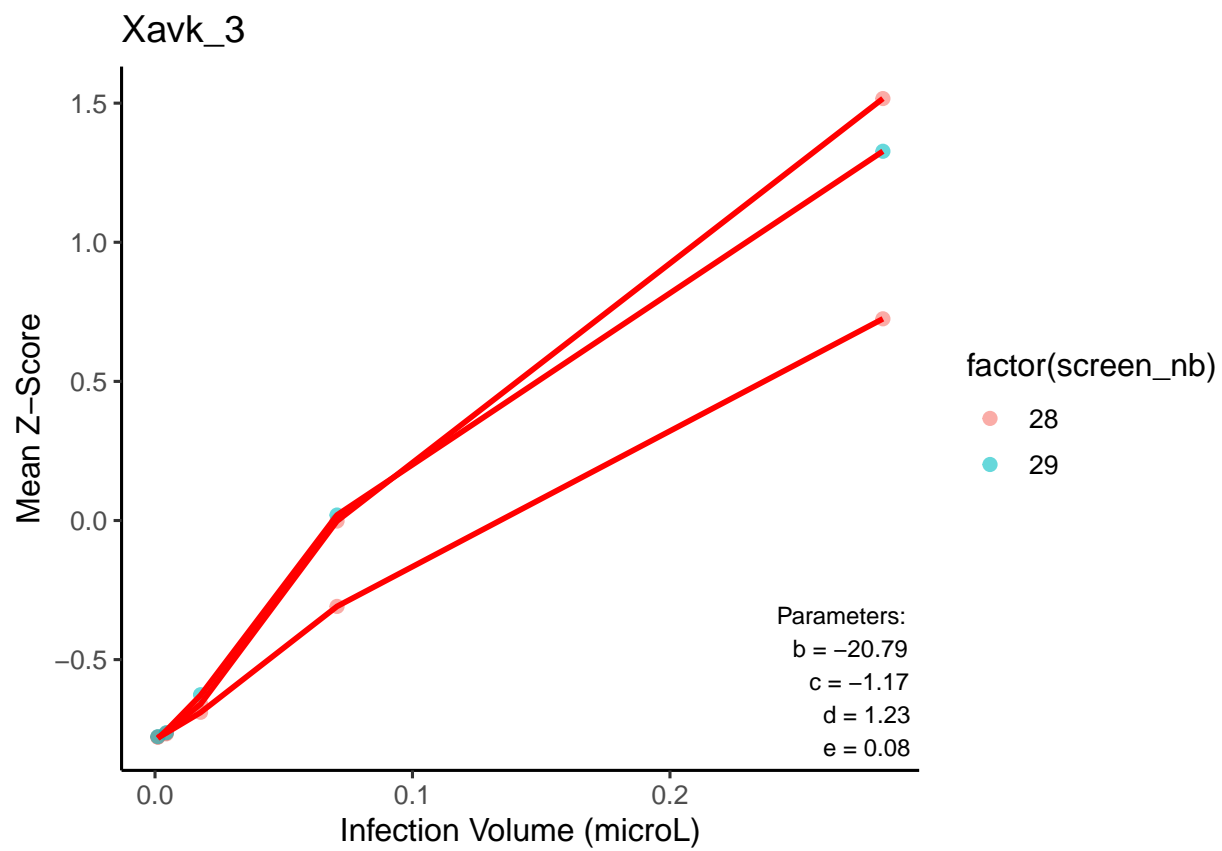
```
##  
## $Wots_2
```



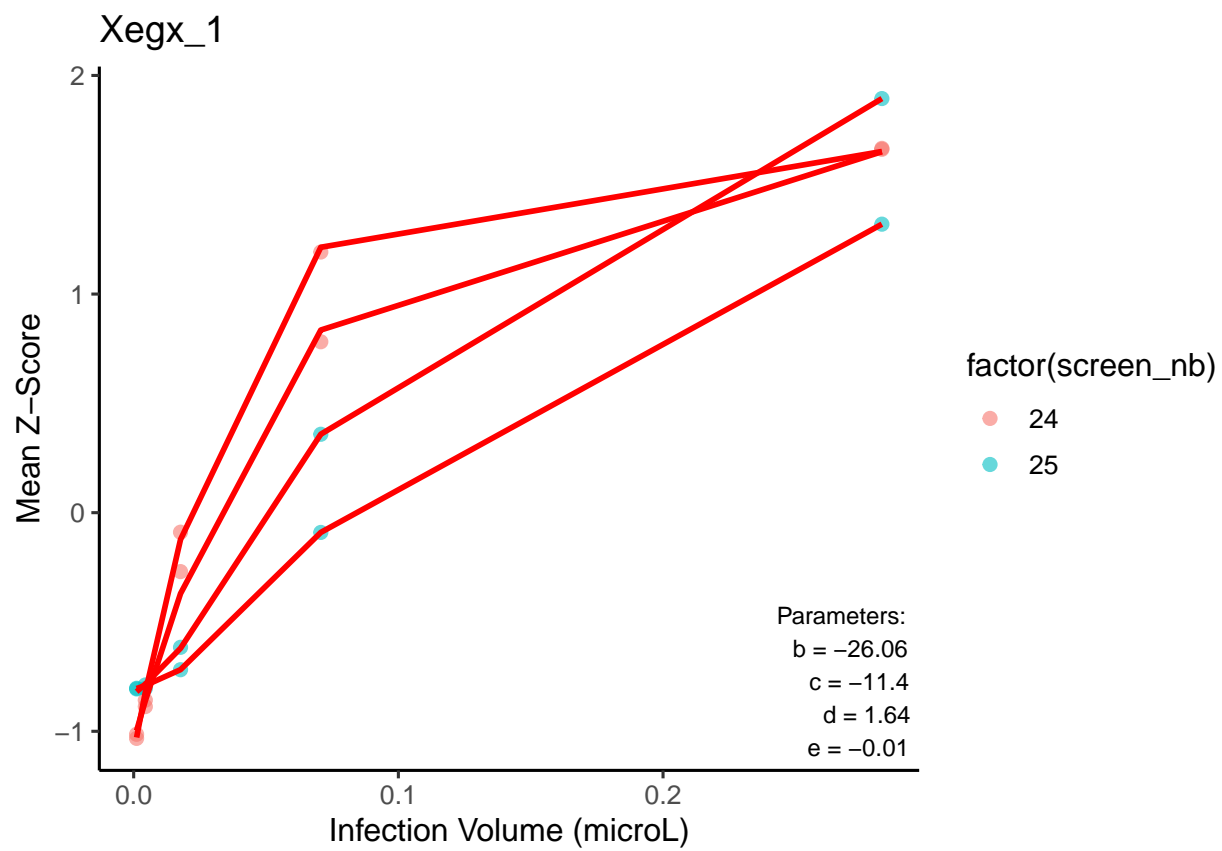
\$Wuye_3



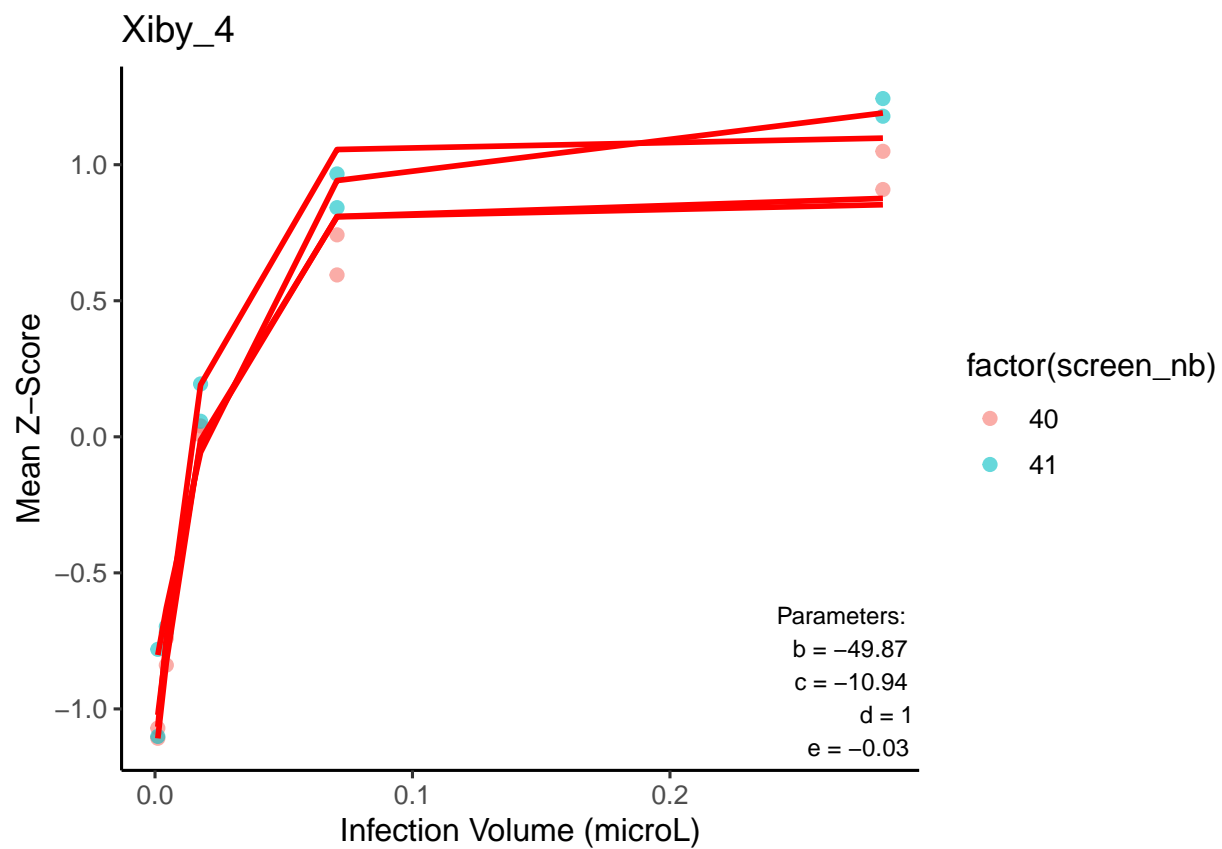
\$Xavk_3



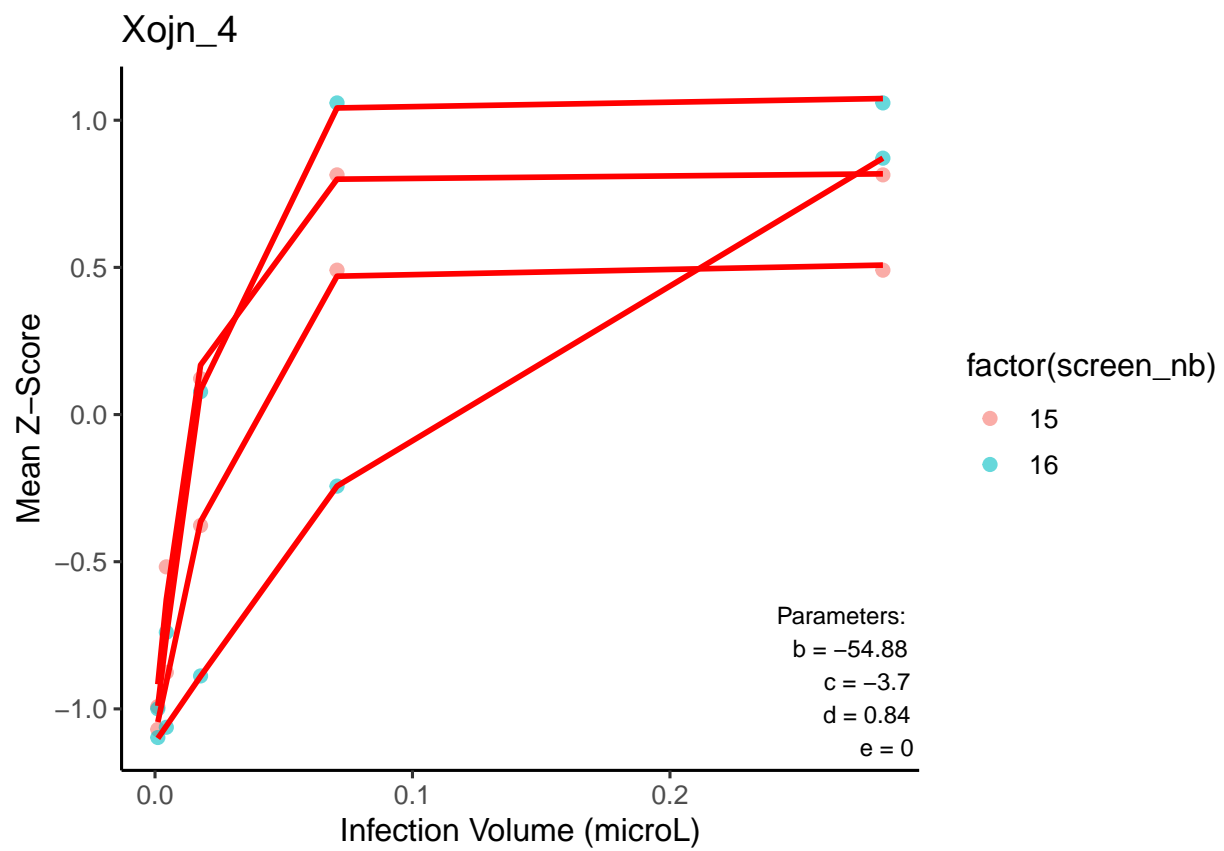
\$Xegx_1



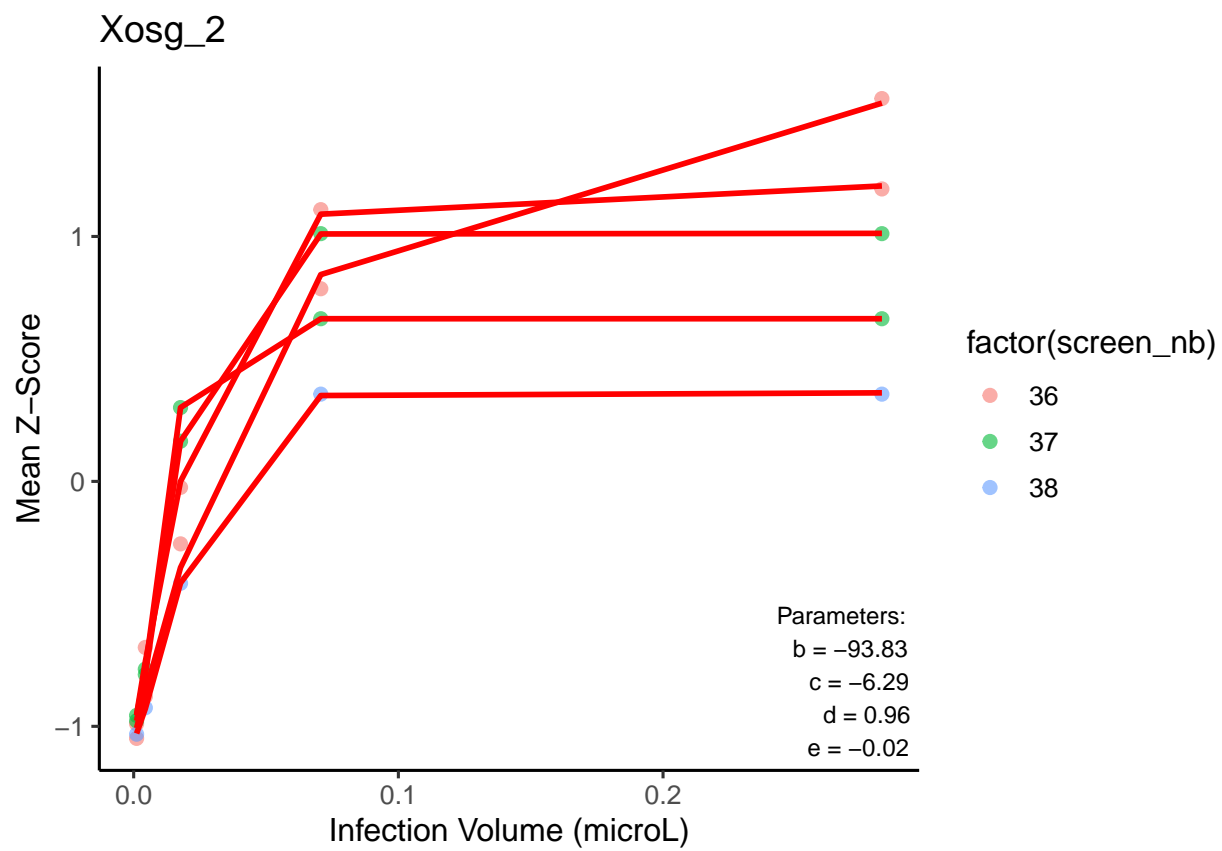
\$Xiby_4



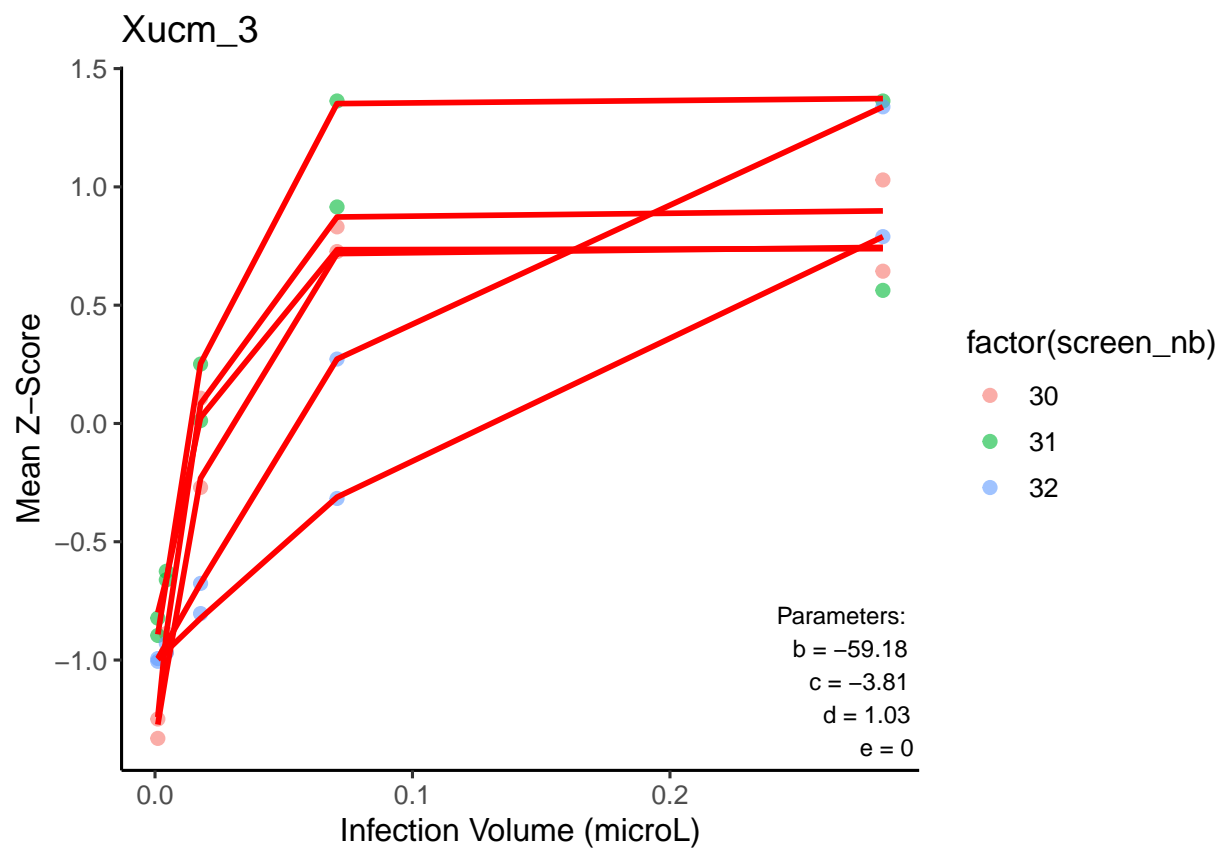
\$Xojn_4



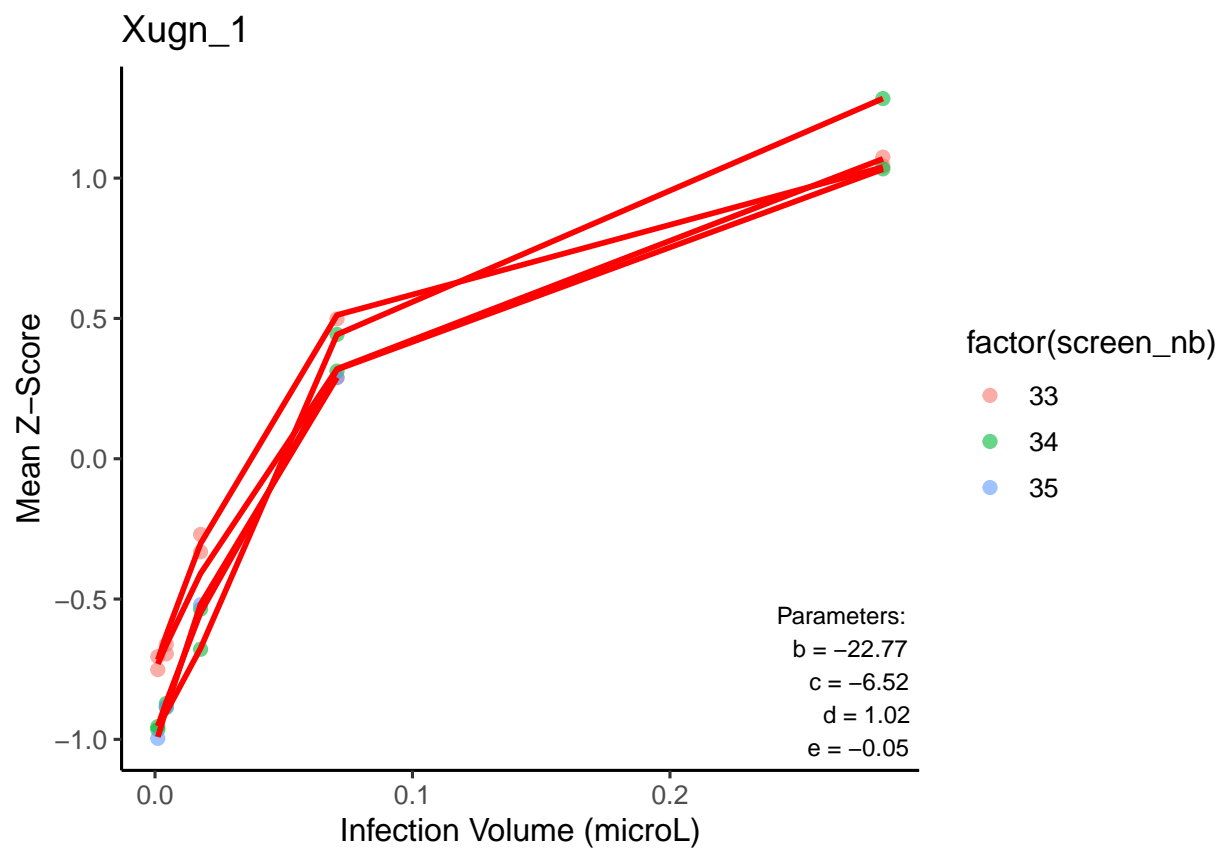
\$Xosg_2



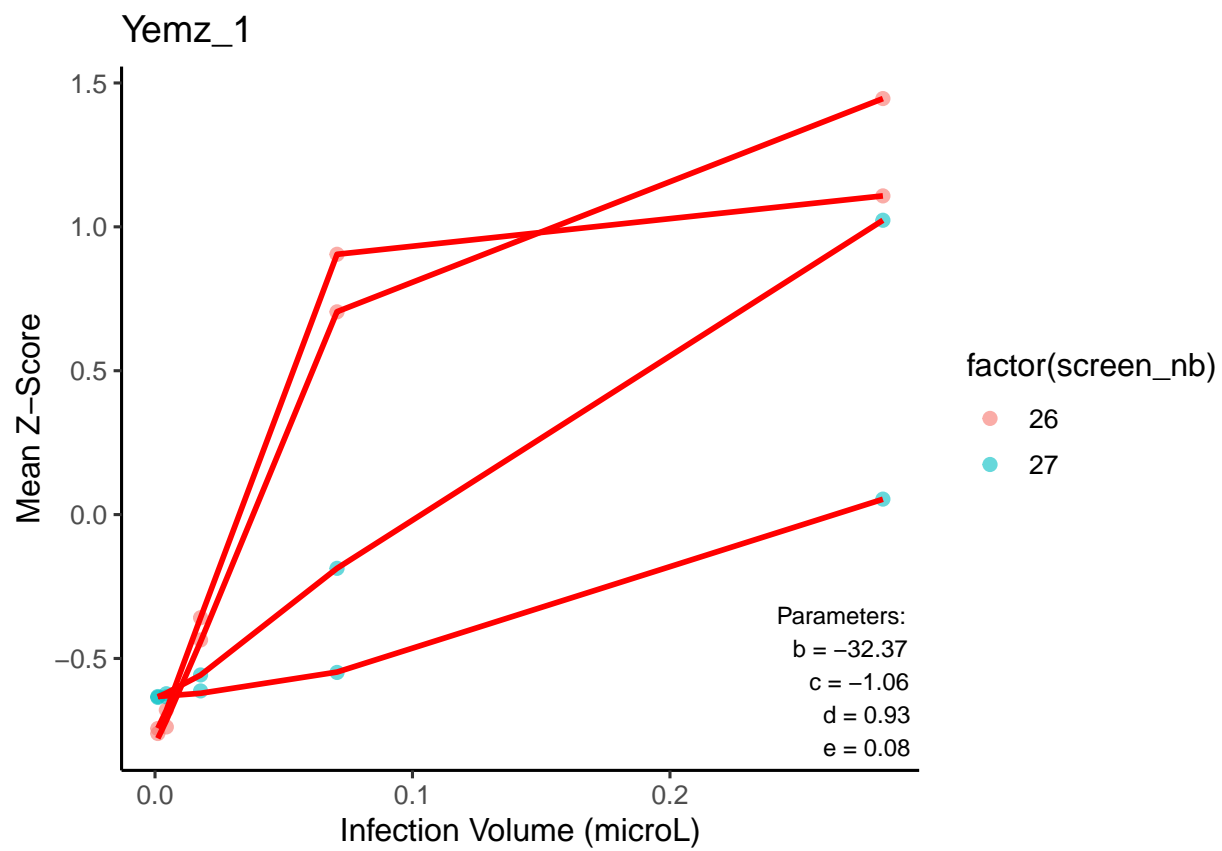
 ## \$Xucm_3



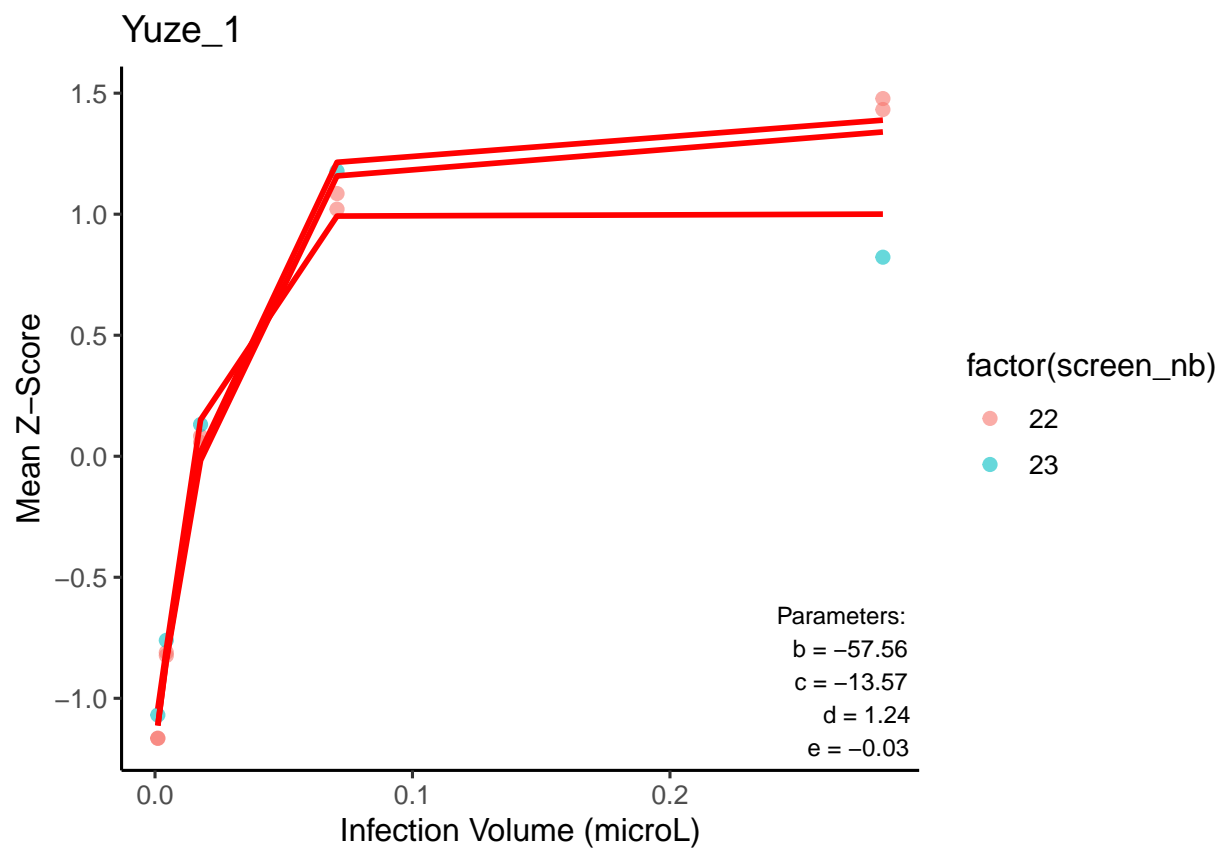
 ## \$Xugn_1



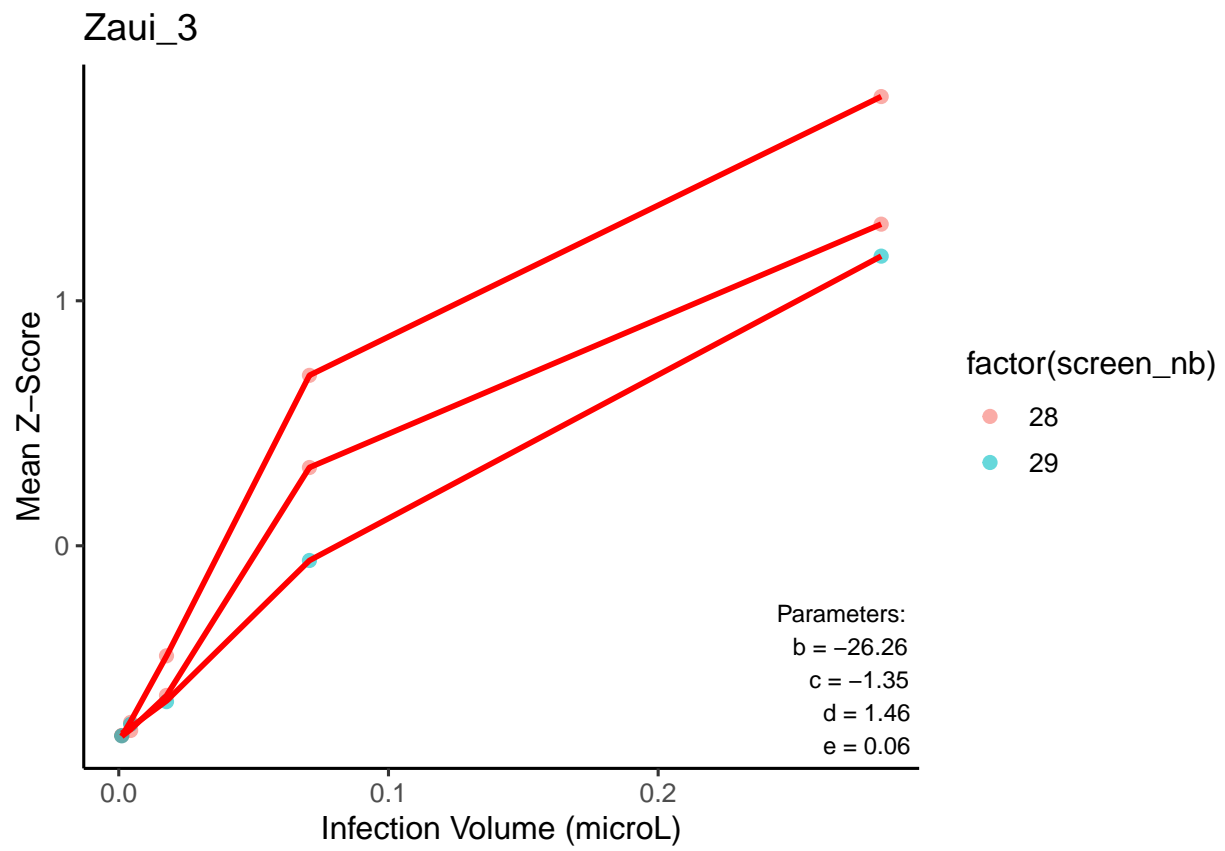
\$Yemz_1



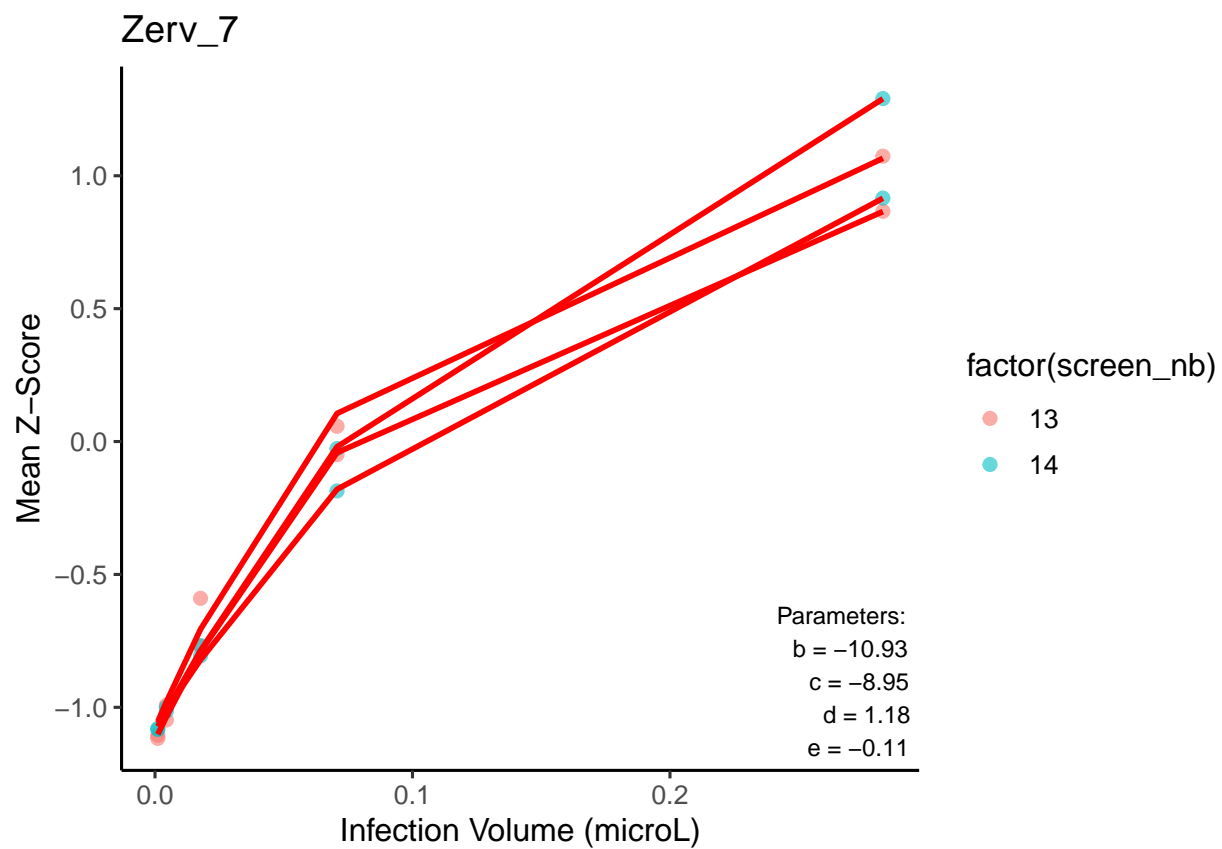
\$Yuze_1



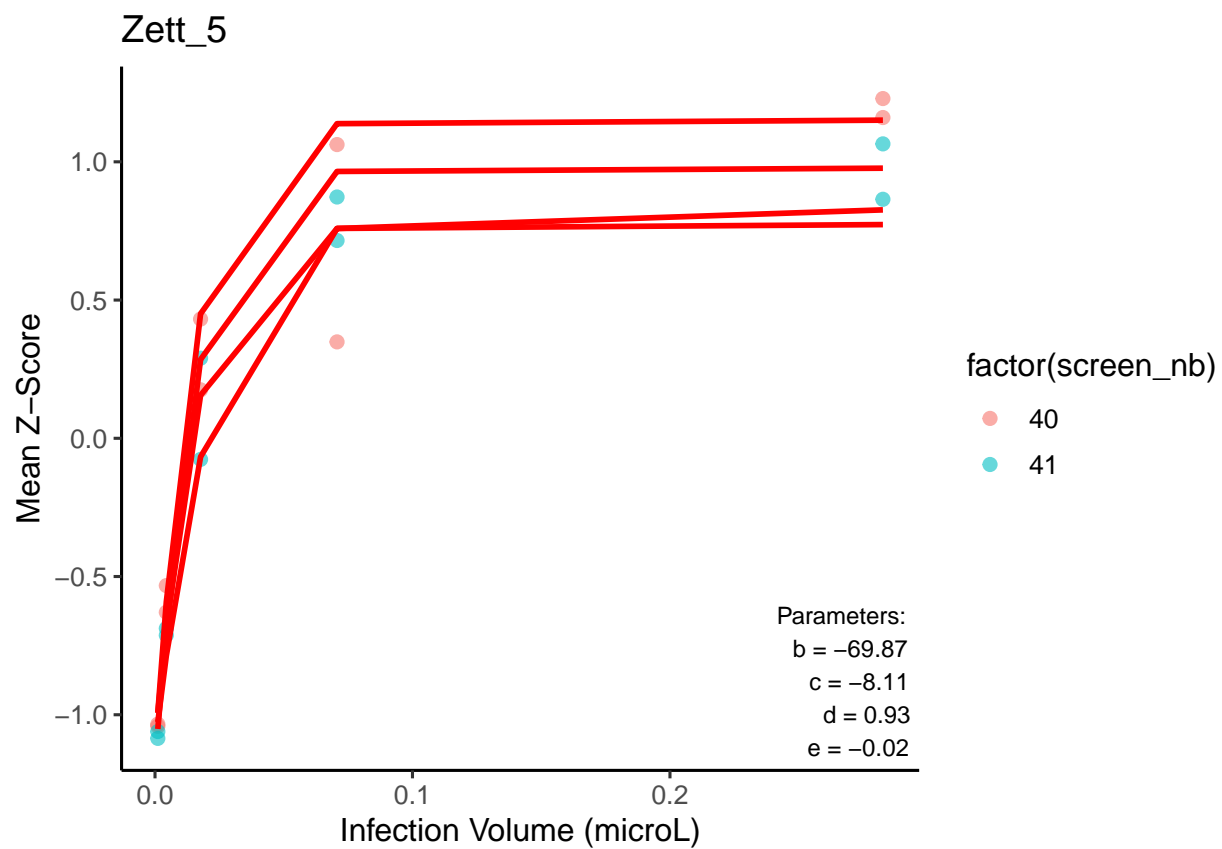
\$Zaui_3



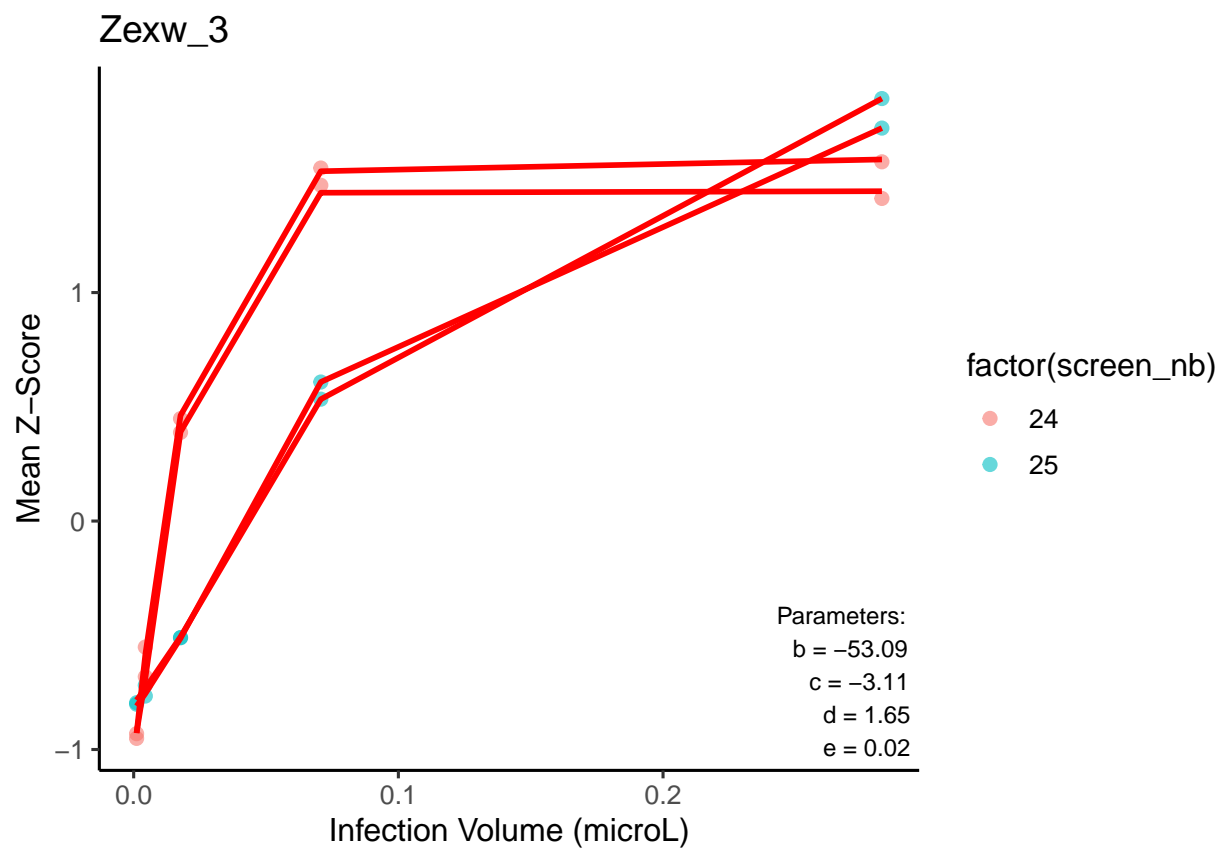
\$Zerv_7



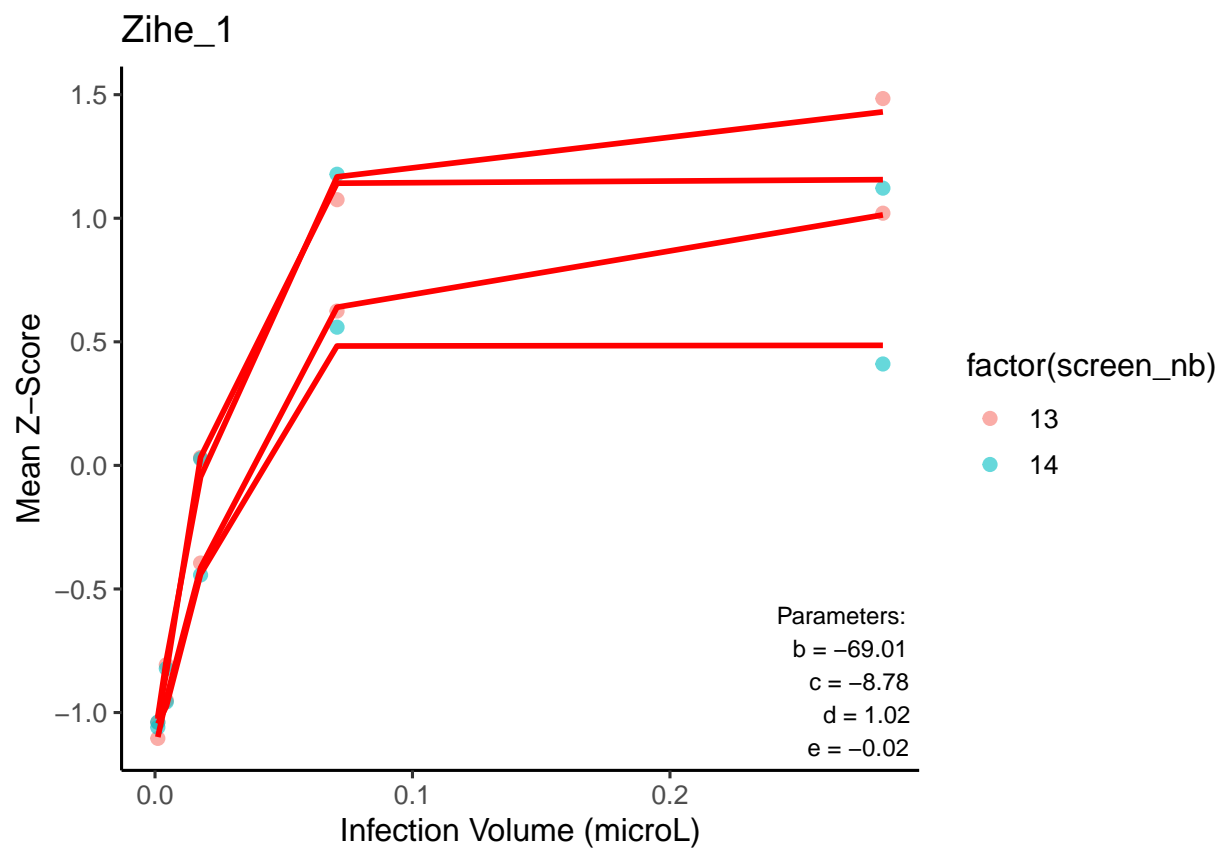
\$Zett_5



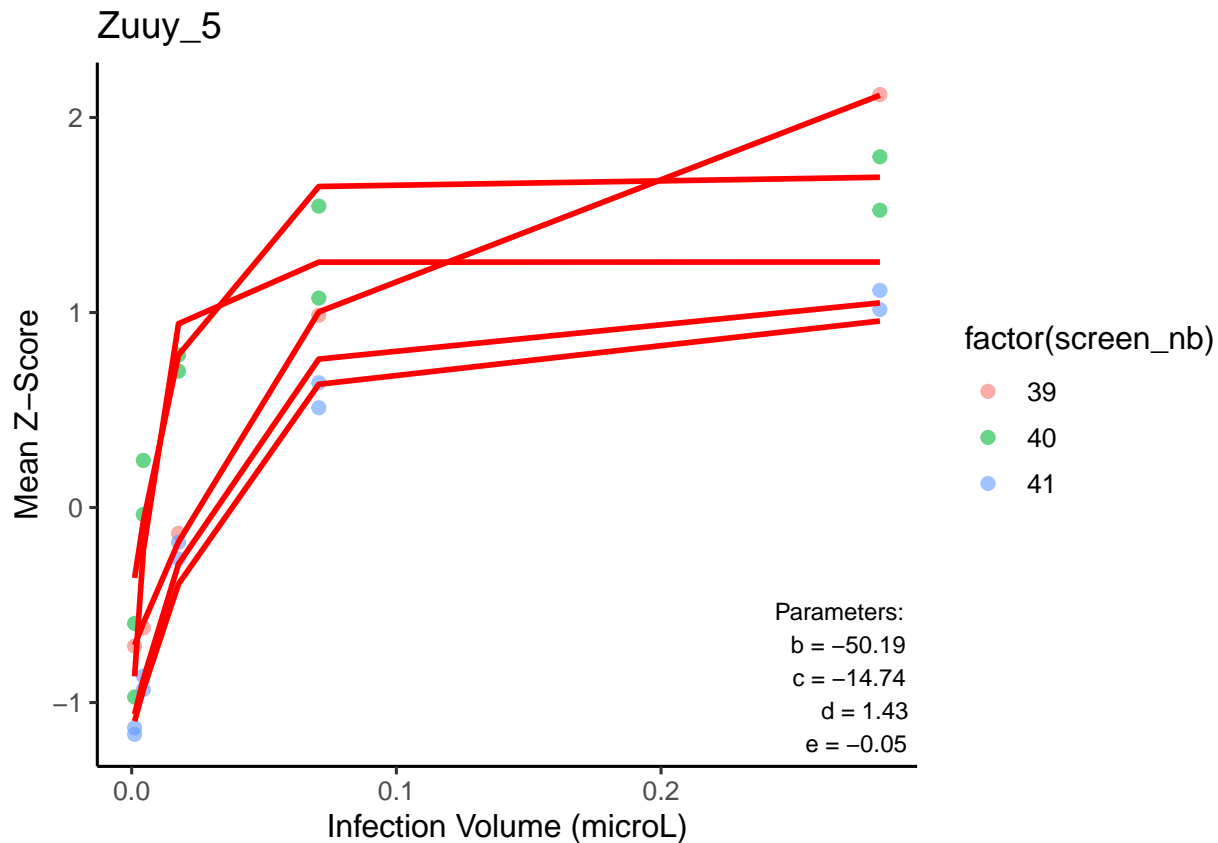
\$Zexw_3



\$Zihe_1



\$Zuuy_5



averaging and cleaning data

I now bind rows from the cell line dataframe list to get a dataframe containing all my parameters, i then group it and summarise it to get the mean for all the parameters across replicates

```
HIV1_vector_data=vector_data_per_cell_line%>%unnest()
```

```
## Warning: `cols` is now required when using `unnest()`.
## i Please use `cols = c(data)`.
```

```
HIV1_vector_data_mean=HIV1_vector_data%>%
  group_by(batch,cell_line,screen,screen_nb,titre,infection_volume_ul)%>%
  summarise(assay_output=mean(assay_output),
            zscore=mean(zscore),a_log=mean(a),
            b_log=mean(b),c_log=mean(c),
            logis_b=mean(logis_b),
            logis_d=mean(logis_d),
            logis_c=mean(logis_c),
            logis_e=mean(logis_e),
            area_under_curve=mean(area_under_curve))%>%
  ungroup()
```

```
## `summarise()` has grouped output by 'batch', 'cell_line', 'screen',
## 'screen_nb', 'titre'. You can override using the `.groups` argument.
```

```
HIV1_vector_data_mean <- HIV1_vector_data_mean %>%
  mutate(across(c(zscore, a_log, b_log,c_log, logis_b, logis_d,logis_c,logis_e, area_under_curve), as.numeric))
```

i then create a dataframe containing the parameter information for each cell screen, so one row per screen and cell line

```
HIV1_vector_data_PCA=unique(HIV1_vector_data_mean%>%
                             group_by(cell_line,screen_nb)%>%
                             summarise(assay_output=max(zscore),
                                       a_log=a_log,
                                       b_log=(b_log),
                                       c_log=c_log,
                                       logis_b=logis_b,
                                       logis_d=logis_d,
                                       logis_c=logis_c,
                                       logis_e=logis_e,
                                       area_under_curve=area_under_curve))
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
## always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `summarise()` has grouped output by 'cell_line', 'screen_nb'. You can override
## using the `.groups` argument.
```

Then averaging between screens,so i get a dataframe with each cell line by itself with its parameters

```
HIV1_vector_data_PCA_sum <- HIV1_vector_data_PCA %>%
  group_by(cell_line) %>%
  summarise(
    assay_output      = mean(assay_output, na.rm = TRUE),
    a_log             = mean(a_log, na.rm = TRUE),
    b_log             = mean(b_log, na.rm = TRUE),
    c_log             = mean(c_log, na.rm = TRUE),
    logis_b           = -mean(logis_b, na.rm = TRUE),
    logis_d           = mean(logis_d, na.rm = TRUE),
    logis_c           = mean(logis_c, na.rm = TRUE),
    logis_e           = mean(logis_e, na.rm = TRUE),
    area_under_curve  = abs(mean(area_under_curve, na.rm = TRUE))
  )
```

i Then rescale it for the PCA, allowing the different parameters to be included(by scaling it one parameter with a different range doesn't have a skewing effect on the data)

i then turn it into a matrix to set the row names as the cell names,

```
HIV1_vector_data_PCA_1=HIV1_vector_data_PCA_sum%>%mutate_if(is.numeric,scale)

HIV1_vector_data_PCA_1=subset(HIV1_vector_data_PCA_1, select=-cell_line)

HIV1_vector_data_PCA_1=as.matrix(HIV1_vector_data_PCA_1)

row.names(HIV1_vector_data_PCA_1)=HIV1_vector_data_PCA_sum$cell_line
```

here i am checking the correlation but there is a visualization for this later

```
HIV1_vector_data_PCA_1=as.data.frame(HIV1_vector_data_PCA_1)
cor(x=HIV1_vector_data_PCA_1$c_log,y=HIV1_vector_data_PCA_1$area_under_curve)
```

```
## [1] 0.3553378
```

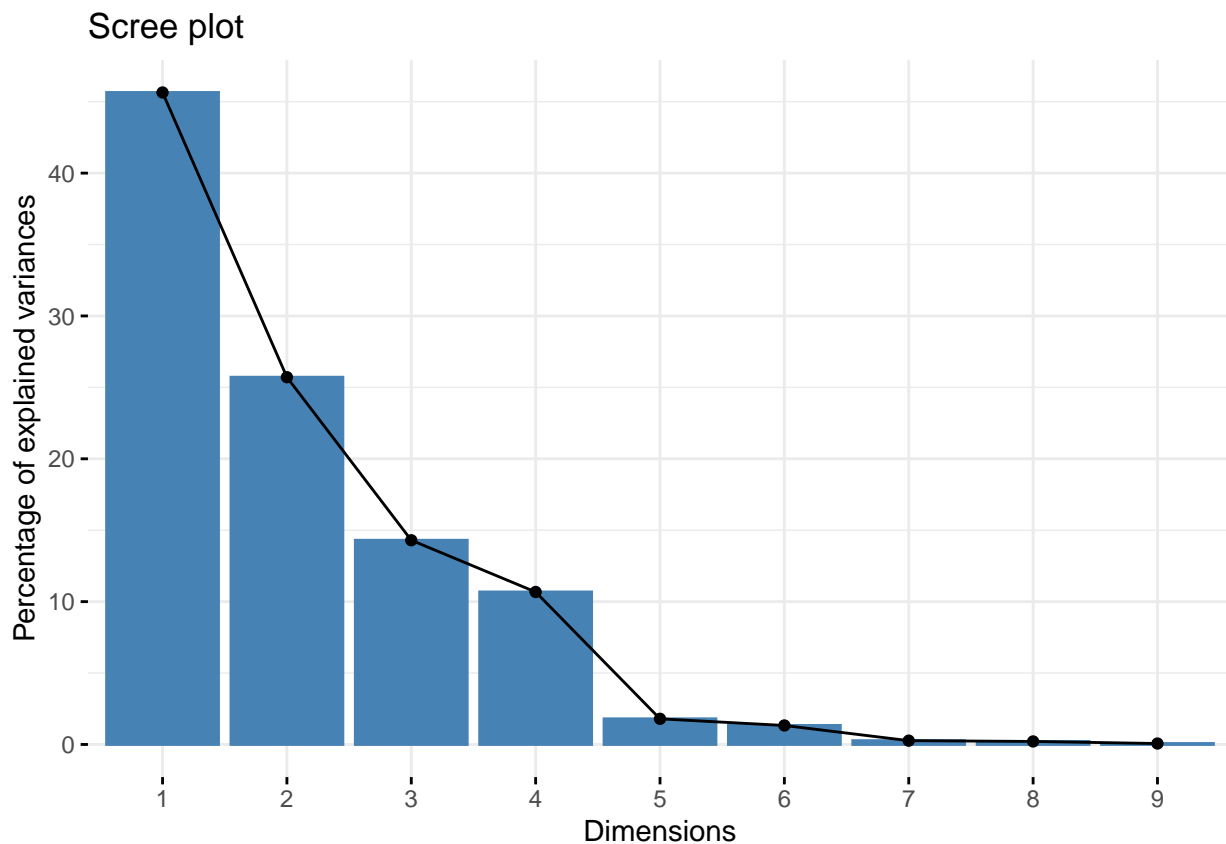
```
#PCA
```

i used prcomp for my pca analysis, also used factoextra which gives me some more tools to help visualise the data, it generates an object that allows me to visualise data

```
PCA=prcomp(x = HIV1_vector_data_PCA_1)
```

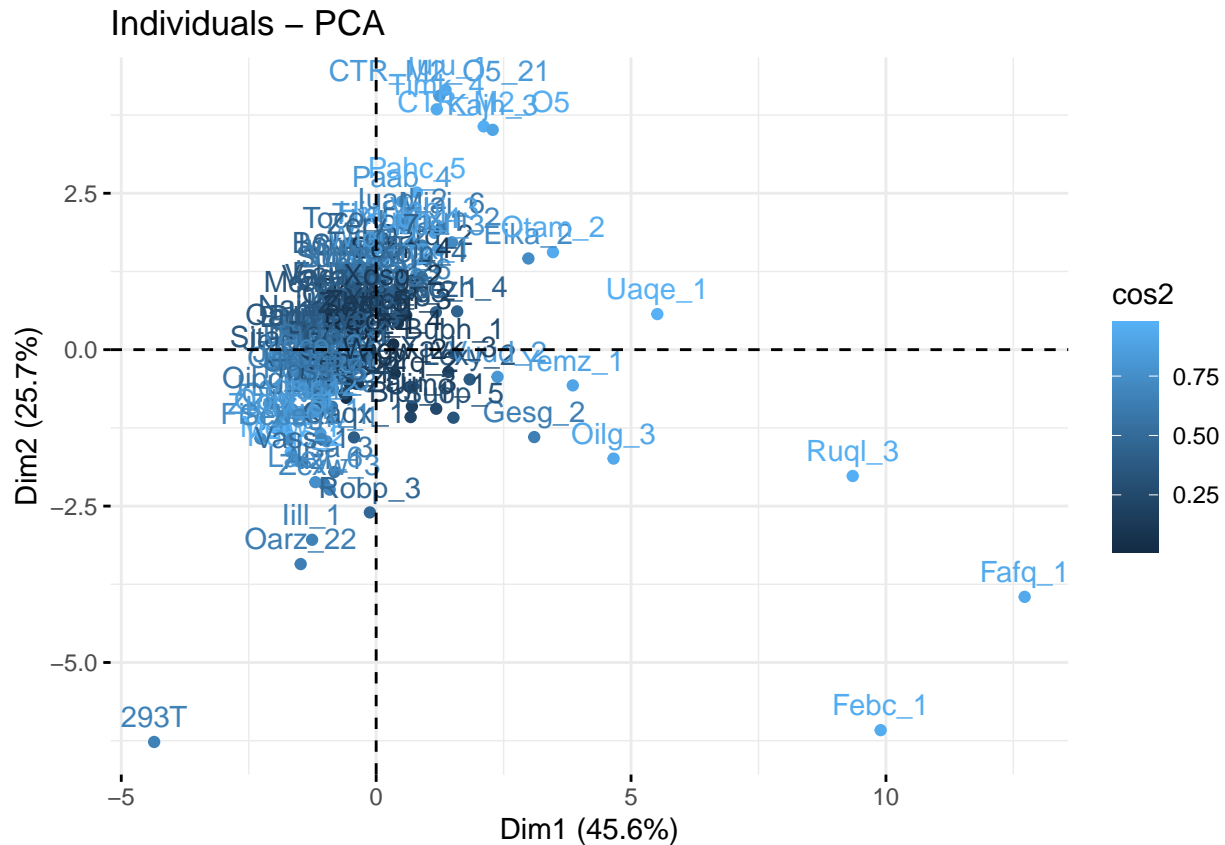
I generate a graph to show me my contributions for each principle component, normally majority of difference in variance is explained by PC1 and PC2

```
library(factoextra)
fviz_eig(PCA)
```



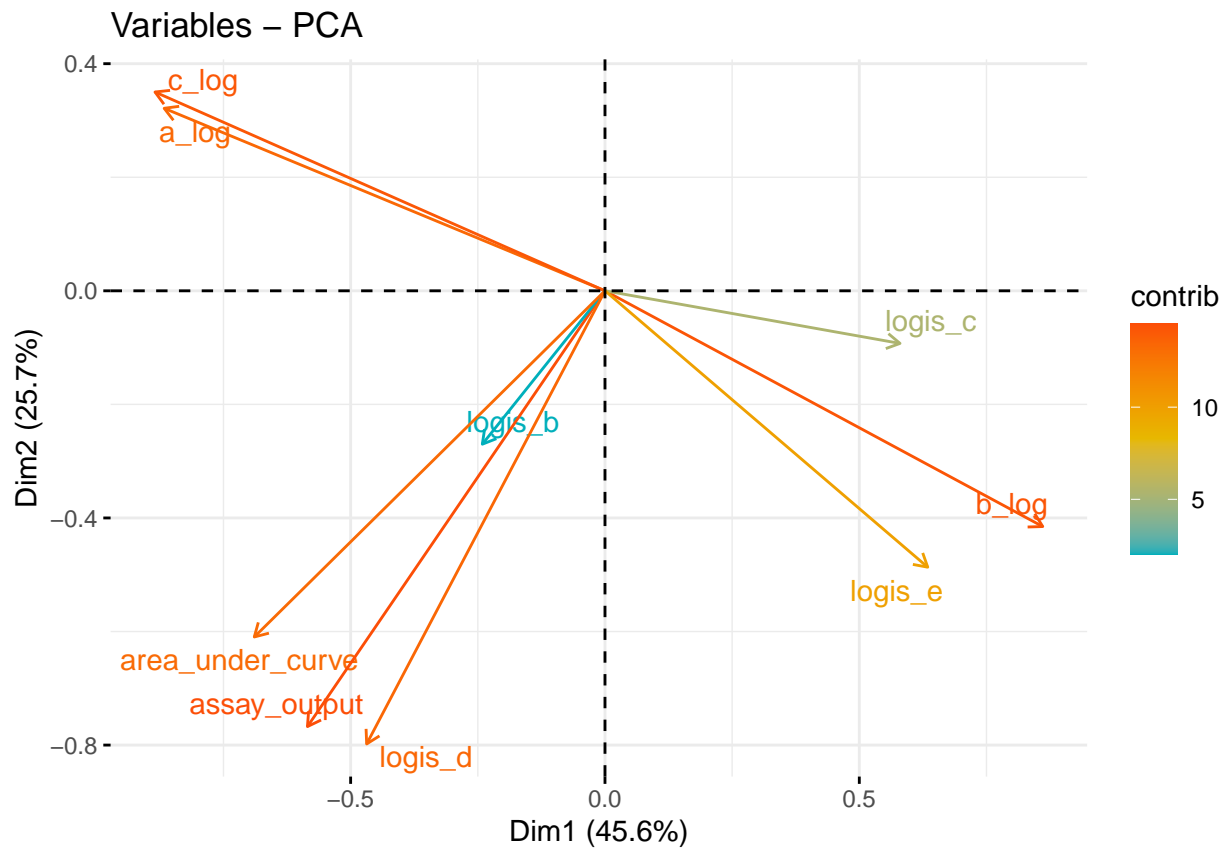
This is a plot of the points on my graph by PC1 and PC2 , can visualise the spread

```
fviz_pca_ind(PCA,col.ind = "cos2",repel = F)
```



This shows how each of my parameters contribute to the variance explained in PC1 and PC2, the more direction a parameter points to the more that parameter contributes to the variance explained by that Principle component. Contributions are shown by colour, with blue meaning it has little contribution to the variance in the PCs

```
fviz_pca_var(PCA,
  col.var = "contrib", # Color by contributions to the PC
  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
  repel = TRUE )
```

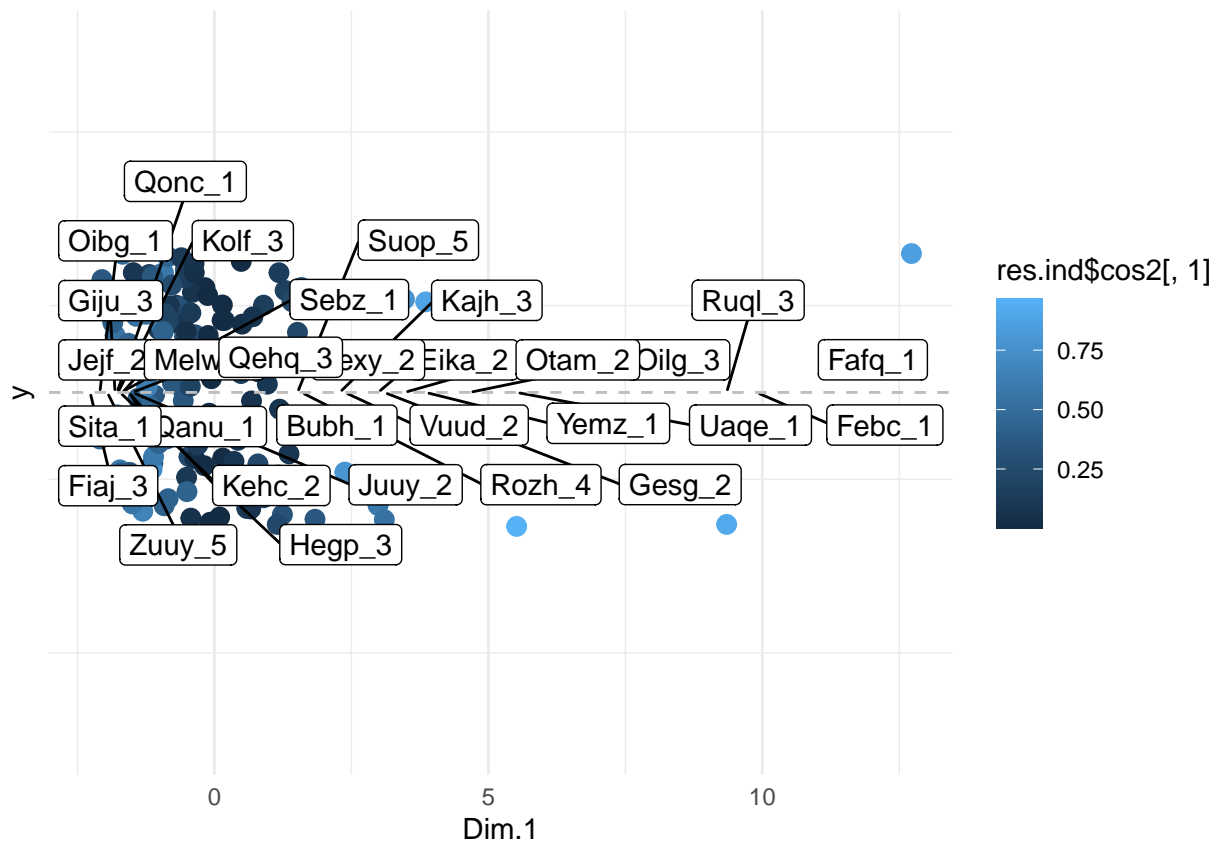


Overlays data points with the parameters to see how each datapoint is in parameter and can see which parameters are causing the split

```
fviz_pca_biplot(PCA, repel=F,
  col.var = "#2E9FDF", # Variables color
  col.ind = "#696969"  # Individuals color
)
```

I then plot a graph showing how data is organised only on the PC with the highest percentage of variance (PC1), this gives me a linear graph of points. I added jitter to make the points more readable but the y axis in the graph doesn't represent anything at all

```
ggplot(data = ind_coord, aes(x=Dim.1, y = 0, label=row.names(ind_coord)))+
  geom_point(size = 3, aes(colour = res.ind$cos2[,1]), position = position_jitter(height = 0.02, width = 0.02))+
  geom_label_repel(aes(label=ifelse(Dim.1 < quantile(Dim.1, 0.1) | Dim.1 > quantile(Dim.1, 0.9), row.names(ind_coord)[1:nrow(ind_coord)]),
    size = 10, color = "red", fontface = "bold", angle = 45))+
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray")+
  theme_minimal() +
  coord_cartesian(ylim = c(-0.05, 0.05))+
  theme(axis.ticks.y = element_blank(),
    axis.text.y = element_blank(),
    panel.grid.major.y = element_blank())
```



#classification

Here i work on my classification, because the PC1 is only about 50% means that i can't trust it properly classify my data, but just in case i keep the upper and lower 10% of the points for dimension 1

but i also create a weighted score, combining both PC1 and PC2 to properly split up my data. I try to use it by taking the percentages of PC1 and PC2 and rescaling it to 100% and then multiplying the PC by its scaled percentage. When you look at the directions of parameters and the spread of the cell lines, assay output and area under curve seem to account for most of the variation, so by weighting the PC, you can classify more accurately the top and bottom quantiles.

I also plot the points with k means clustering algorithm, and i label based on the top quantiles for my combined score, i save the top and bottom 10% to their own dataframes for storage

```
susceptible_vector_cell=ind_coord[(ind_coord$Dim.1 < quantile(ind_coord$Dim.1, 0.1)),]
Resistant_vector_cell=ind_coord[(ind_coord$Dim.1 > quantile(ind_coord$Dim.1, 0.9)),]
```

```
HIV_vector_data_susceptible=HIV1_vector_data_PCA_1[match(row.names(susceptible_vector_cell), (HIV1_vector_data_PCA_1$cell_line)),]
HIV_vector_data_resistant=HIV1_vector_data_PCA_1[match(row.names(Resistant_vector_cell), (HIV1_vector_data_PCA_1$cell_line)),]
```

```
set.seed(45)
```

```
clusters <- kmeans(ind_coord[, c("Dim.1", "Dim.2")], centers = 8)
```

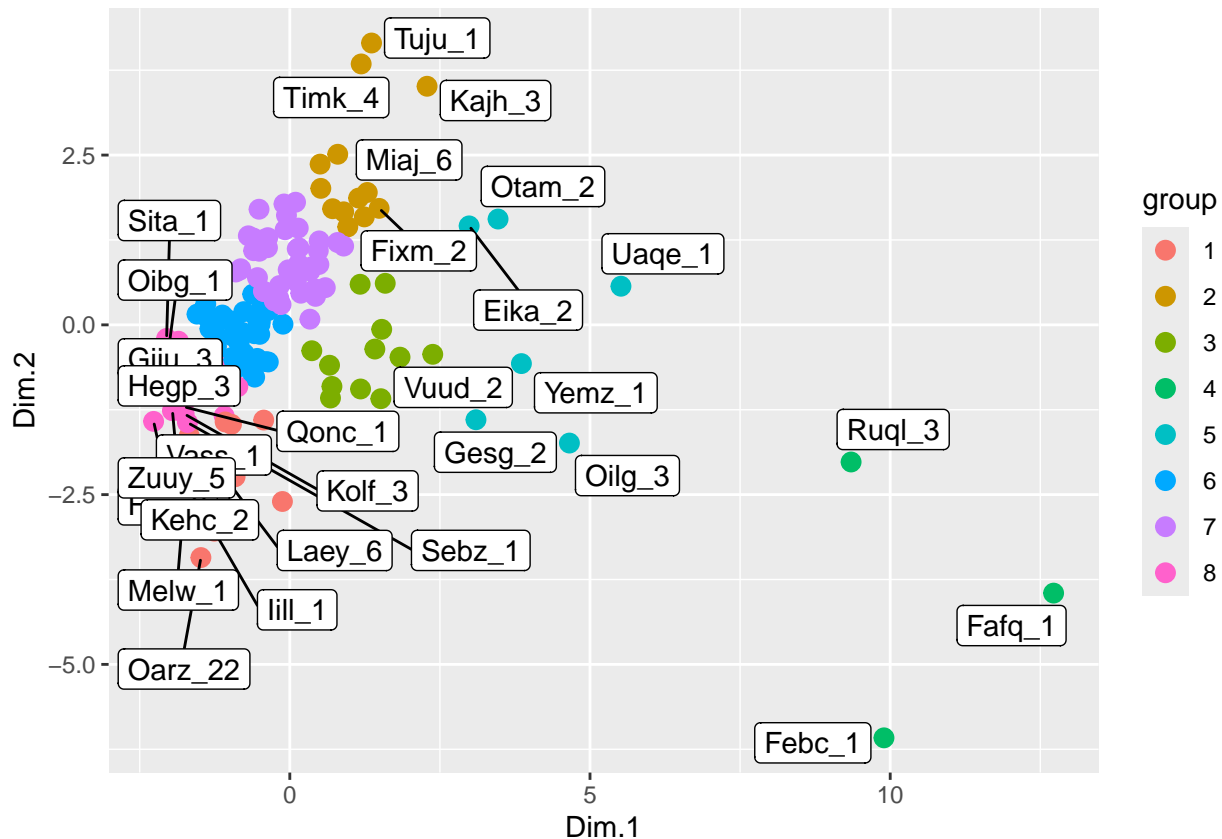
```
ind_coord$group <- as.factor(clusters$cluster)
```

```
ind_coord$pc_combined=0.70*ind_coord$Dim.1+ 0.30*ind_coord$Dim.2
```

```
ggplot(data = ind_coord,aes(x=Dim.1,y = Dim.2,label=row.names(ind_coord)))+
```

```
  geom_point(size = 3,aes(colour = group))+
```

```
  geom_label_repel(aes(label=ifelse(pc_combined < quantile(pc_combined, 0.1) | pc_combined > quantile(pc_combined, 0.9), label, "")))
```

```
susceptible_vector_cell=ind_coord[(ind_coord$pc_combined < quantile(ind_coord$pc_combined, 0.1)),]
Resistant_vector_cell=ind_coord[(ind_coord$pc_combined > quantile(ind_coord$pc_combined, 0.9)),]

HIV1_vector_data_notable <- as.data.frame(rbind(HIV_vector_data_resistant, HIV_vector_data_susceptible))
mutate(phenotype = ifelse(row.names(.) %in% row.names(HIV_vector_data_resistant),
                          "resistant",
                          "susceptible"))
```

here i create the classification based on the clusters, i set the clusters and choose the most extreme groups, NOTE- k means code changes every time so i set the seed. if the seed is changed this code would have to be changed also

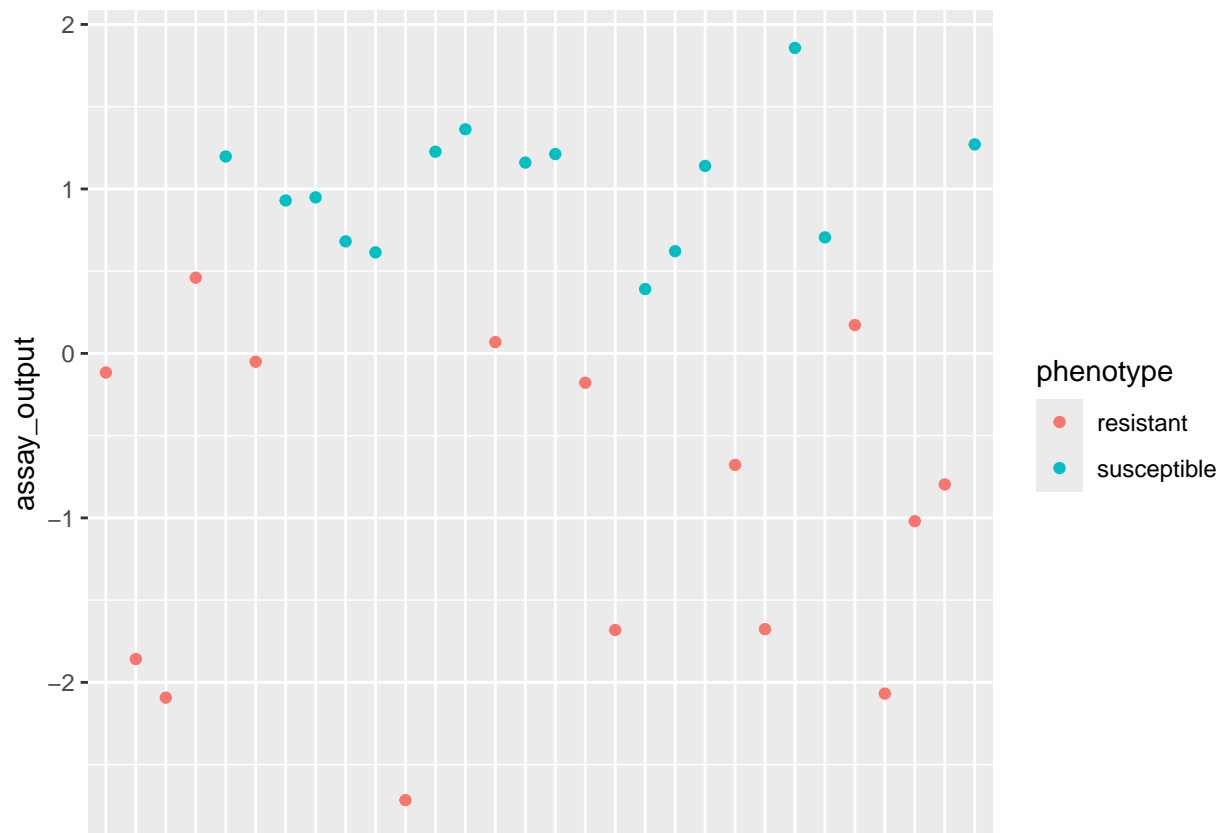
```
susceptible_vector_clusters=ind_coord[(ind_coord$group==1),]
Resistant_vector_clusters=ind_coord[(ind_coord$group==5|ind_coord$group==6|ind_coord$group==3),]
```

visualising data split

here i visualise the extremes i found, and see the split between them in a graph, can see how well the separation is by splitting between the parameters, can also see how well parameters align to biological expectations

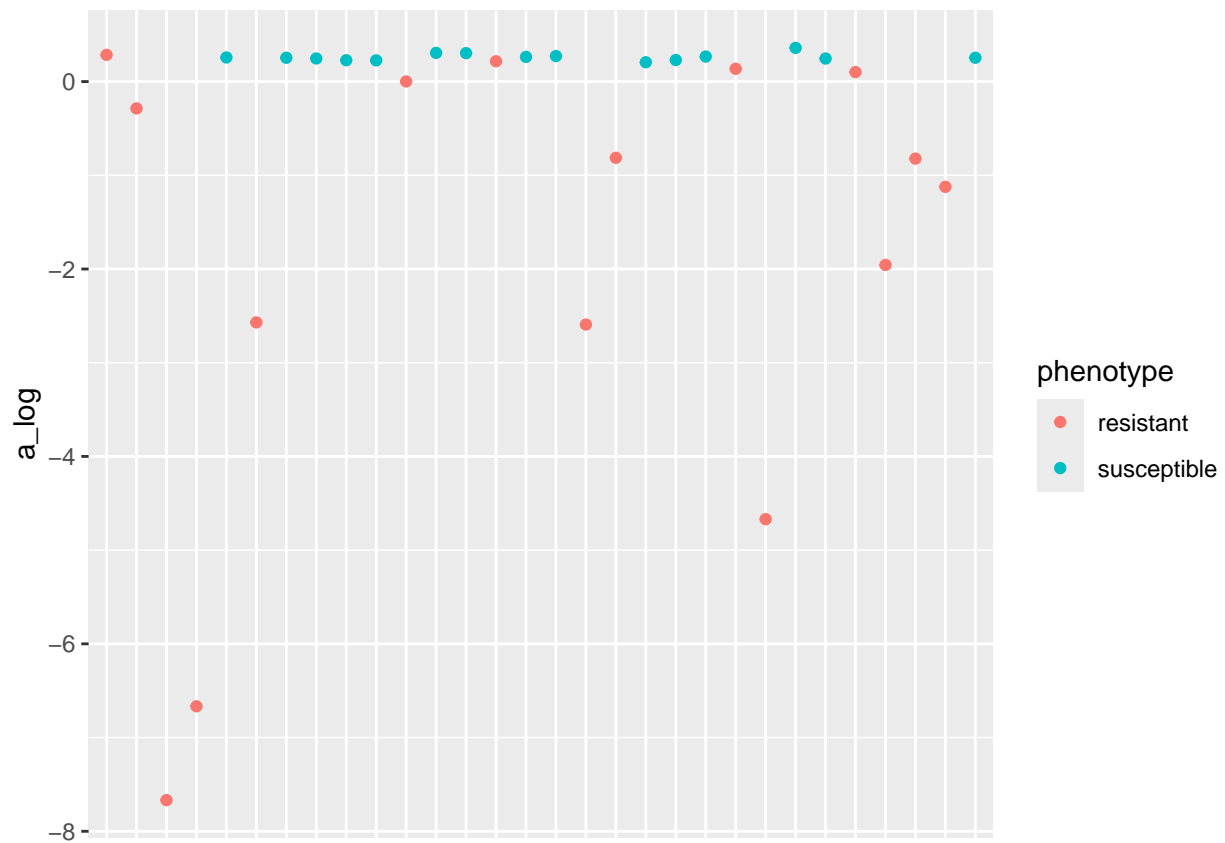
here i look at assay output(zscore) there is a solid split between the susceptible and resistant with susceptibles consistently having higher output (which is the max output seen)

```
ggplot(HIV1_vector_data_notable,aes(x=row.names(HIV1_vector_data_notable),y=assay_output,colour = phenotype))
  geom_point()+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

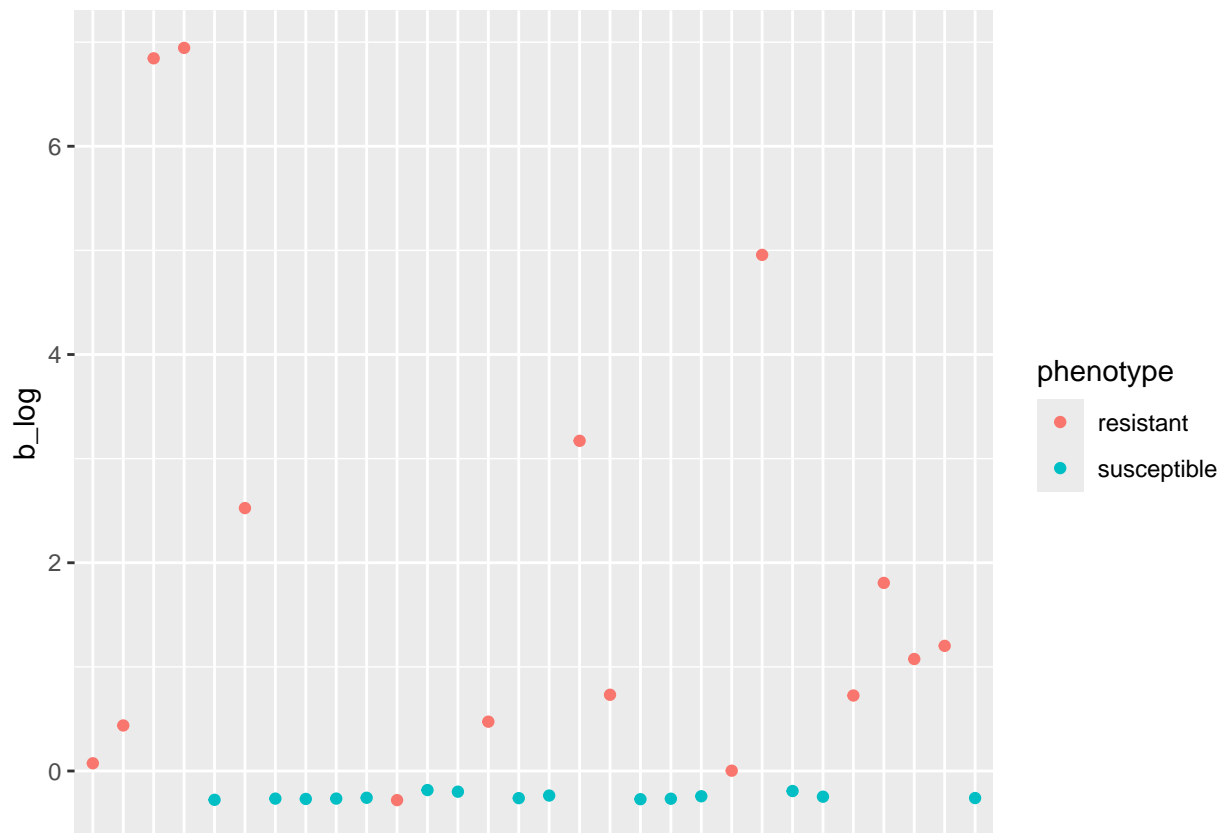


this looks at the a log (the y offset in our logarithmic function), there is a clear difference between the values with the resistant having lower y offsets which is what you would expect though some are near 0 but all the susceptible are above zero meaning showing they would be more permissible

```
ggplot(HIV1_vector_data_notable,aes(x=row.names(HIV1_vector_data_notable),y=a_log,colour = phenotype))+
  geom_point()+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

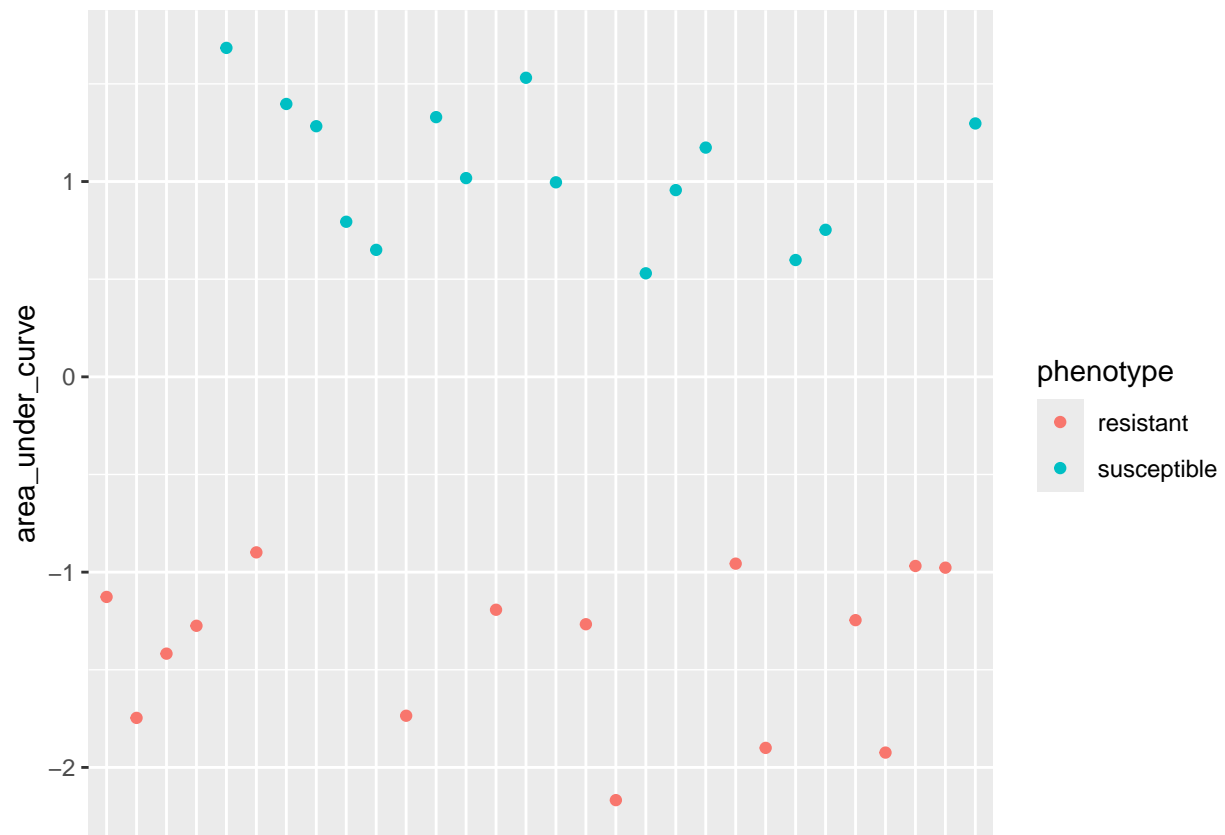


```
ggplot(HIV1_vector_data_notable,aes(x=row.names(HIV1_vector_data_notable),y=b_log,colour = phenotype))+
  geom_point()+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

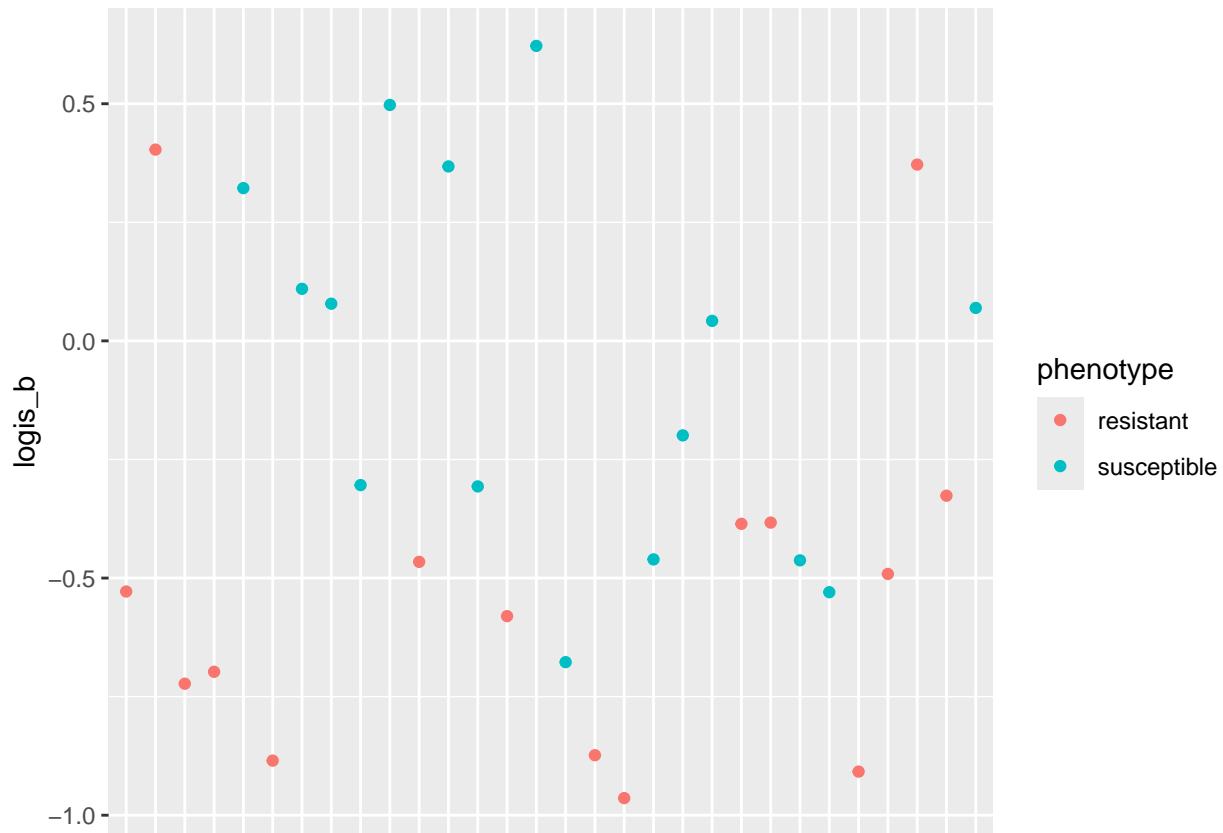


This also has a clear split which biologically makes sense, though biologically having a high absolute AUC doesn't necessarily mean that it is an extremely susceptible, but it does imply something about the dynamics of the and how it responds vector

```
ggplot(HIV1_vector_data_notable, aes(x=row.names(HIV1_vector_data_notable), y=area_under_curve, colour = phenotype)) +
  geom_point() +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



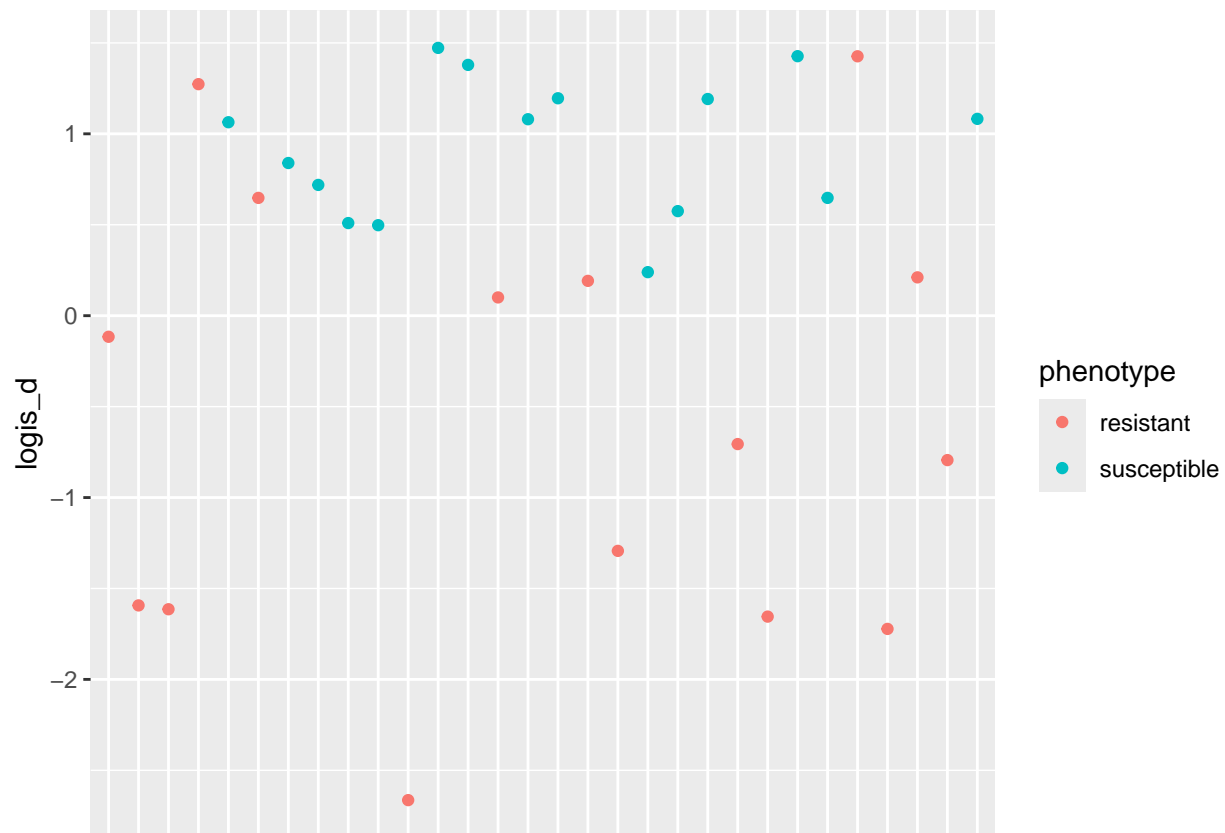
```
ggplot(HIV1_vector_data_notable,aes(x=row.names(HIV1_vector_data_notable),y=logis_b,colour = phenotype))
  geom_point()+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



logis_d is the max from the 4 parameter logistic, it mostly would align with the assay output, but the logistic may find a higher max depending on the shape of the fitted curve, e.g if the curve was increasing but hadn't reached a plateau the model would model a theorised plateau that may be higher than assay output, so in this way it measures both the max output and the dynamics of the gfp production

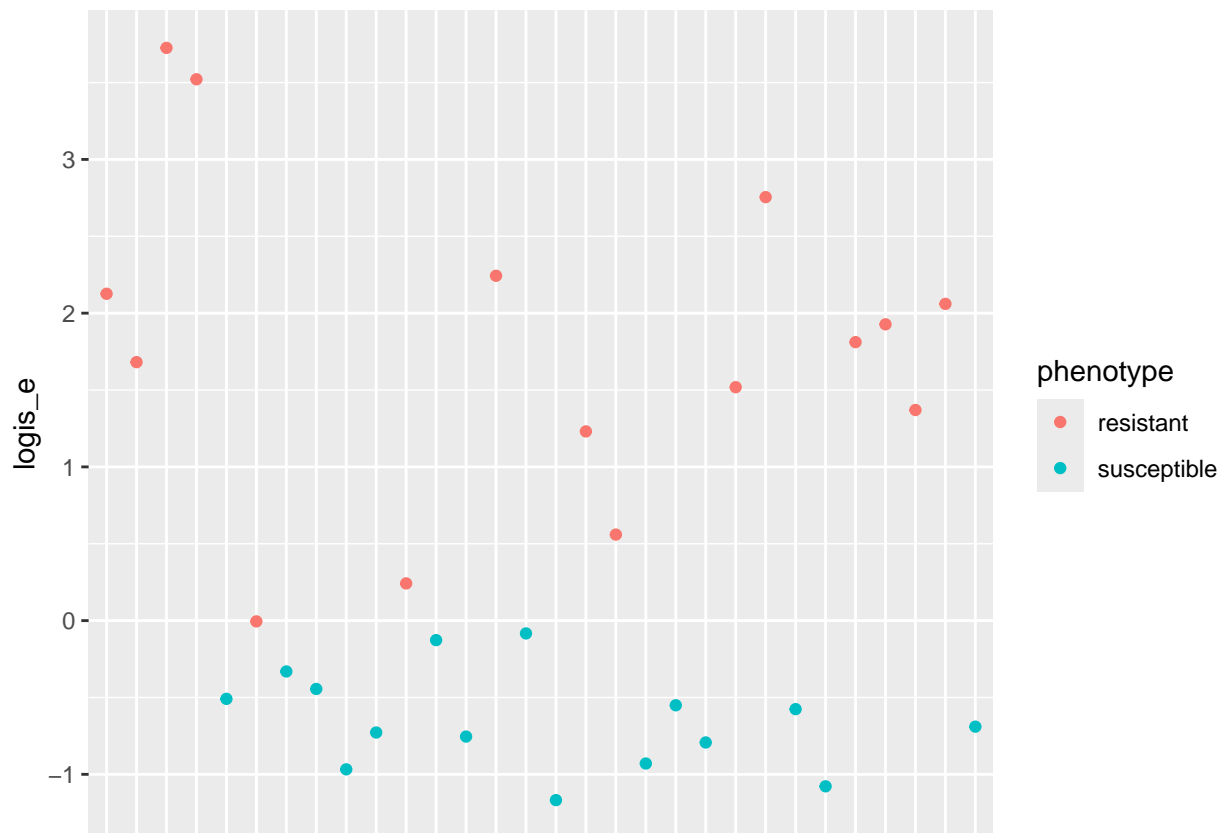
we see a good split except for one cell line Febc which is resistant but has a high value. This could mean the curve of that cell followed a more gradual increase but still was modelled a high theorised plateau.

```
ggplot(HIV1_vector_data_notable, aes(x=row.names(HIV1_vector_data_notable), y=logis_d, colour = phenotype))
  geom_point() +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



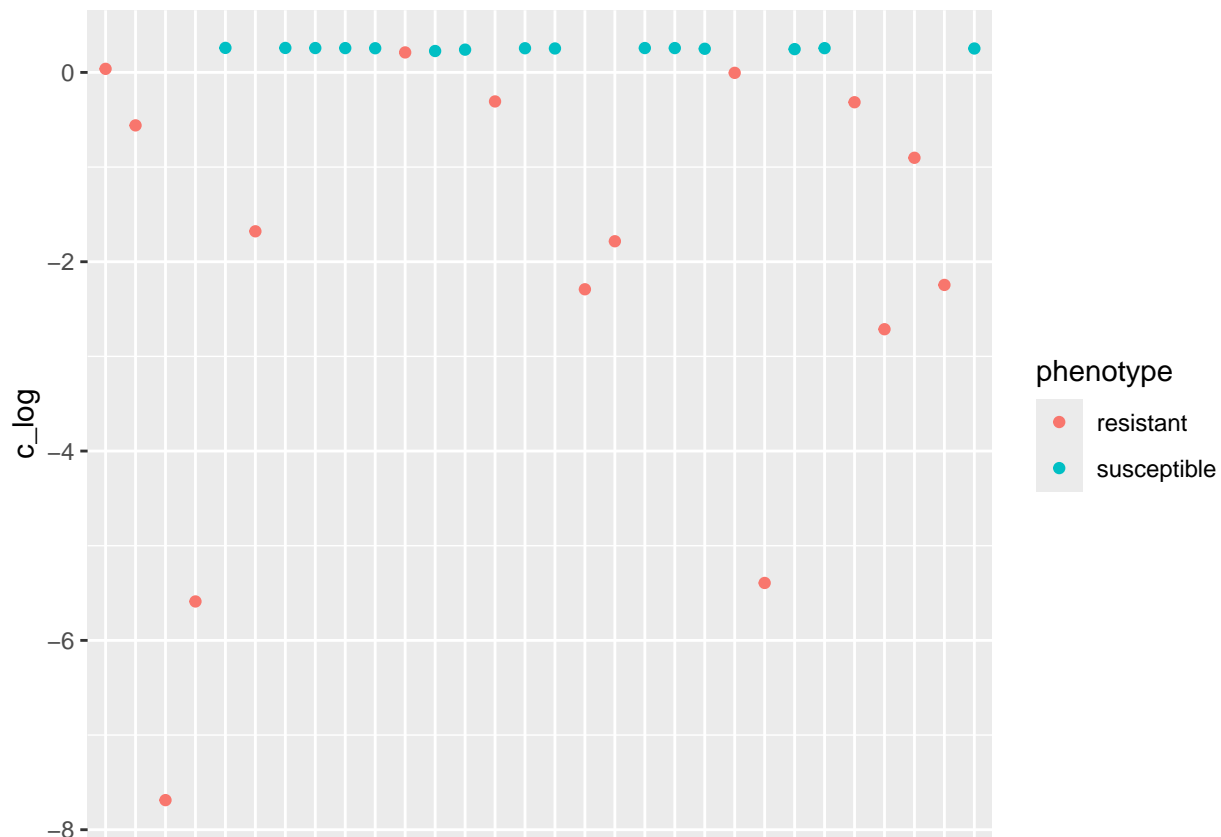
Logis e works as the x offset, so having a high logis e shifts the graph to the right which would theoretically make it more resistant . the graph is not as split, with a lot of overlap between susceptible and resistant but many of the resistant have higher logis e than susceptible which would be expected.

```
ggplot(HIV1_vector_data_notable,aes(x=row.names(HIV1_vector_data_notable),y=logis_e,colour = phenotype))
  geom_point()+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



c log is the x offset in the logarithmic model, this means that theoretically the susceptible would have lower c log but this is not what is seen, instead we see the opposite that only the resistant have low c log, However c_log is highly correlated with a_log which does fit what would be expected. It could be that due to a quirk in the modelling, the shape of the highly susceptible affected the way c log was modeled to have a correct fit, compared to the more flat growth of the resistant curves which may have been modeled to have much lower c_log.

```
ggplot(HIV1_vector_data_notable,aes(x=row.names(HIV1_vector_data_notable),y=c_log,colour = phenotype))+
  geom_point()+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

ranking list

Here i rank and gain the top cell lines in each parameter to see which cell lines appear the most,this is another way of separating the data but i dont prefer it

```
vector_ranking=list()
for(col in names(HIV1_vector_data_PCA_1[,c(1:3,6,7,9)])) {

  # Extract the column values
  vec <- HIV1_vector_data_PCA_1[[col]]

  # Calculate the 10th and 90th quantiles (ignoring NAs)
  low_quant <- quantile(vec, 0.1, na.rm = TRUE)
  high_quant <- quantile(vec, 0.9, na.rm = TRUE)

  # Identify rows where the value is either below the 10th percentile or above the 90th percentile
  extreme_idx <- which(vec < low_quant | vec > high_quant)

  # Subset the original data frame for these extreme values
  subset_data <- HIV1_vector_data_PCA_1[extreme_idx, ]

  # Add a phenotype column: if the value is below the 10th percentile, call it "resistant",
  # if above the 90th percentile, label it "susceptible"
  # (This uses the current column's values from the subset.)
  subset_data$phenotype <- ifelse(vec[extreme_idx] < low_quant, "resistant", "susceptible")
  subset_data$cell_line=rownames(subset_data)
  rownames(subset_data) <- NULL
}
```

```

# Store the subset in the vector_ranking list using the column name as the key
vector_ranking[[col]] <- subset_data
}

for(col in names(HIV1_vector_data_PCA_1[,c(4,5,8)])) {

  # Extract the column values
  vec <- HIV1_vector_data_PCA_1[[col]]

  # Calculate the 10th and 90th quantiles (ignoring NAs)
  low_quant <- quantile(vec, 0.1, na.rm = TRUE)
  high_quant <- quantile(vec, 0.9, na.rm = TRUE)

  # Identify rows where the value is either below the 10th percentile or above the 90th percentile
  extreme_idx <- which(vec < low_quant | vec > high_quant)

  # Subset the original data frame for these extreme values
  subset_data <- HIV1_vector_data_PCA_1[extreme_idx, ]

  # Add a phenotype column: if the value is below the 10th percentile, call it "resistant",
  # if above the 90th percentile, label it "susceptible"
  # (This uses the current column's values from the subset.)
  subset_data$phenotype <- ifelse(vec[extreme_idx] < low_quant, "susceptible", "resistant")
  subset_data$cell_line=rownames(subset_data)
  rownames(subset_data) <- NULL

  # Store the subset in the vector_ranking list using the column name as the key
  vector_ranking[[col]] <- subset_data
}

vector_ranking_df=bind_rows(vector_ranking)
cell_line_counts <- vector_ranking_df %>%
  count(cell_line,phenotype)

# 3. Get the cell lines that appear more than 9 times.
common_vector_cell_lines <- cell_line_counts[cell_line_counts$n>2,]

# 4. Filter the original data to include only those common_vector cell lines.
most_common_vector <- inner_join(vector_ranking_df,common_vector_cell_lines,join_by(x$cell_line==y$cell_line))

# Optionally, view the result:
unique(most_common_vector)

```

```

##      assay_output      a_log      b_log      c_log      logis_b      logis_d
## 1      3.24097893  0.3748034774 -0.30781438  0.26138590  6.67223990  3.17627500
## 2     -2.57523359 -0.0879936874 -0.11119842  0.07599256 -0.29951576 -2.63097122
## 3     -2.85411206  0.0171429525 -0.20613934  0.14320224 -1.15564974 -2.26783460
## 4     -1.40992169  0.0658324205 -0.32986337  0.25897950  1.83518570 -1.61186269
## 5     -1.85840178 -0.2863665039  0.43734457 -0.55976652  0.40332070 -1.59302648
## 6     -2.09287464 -7.6676476946  6.84457353 -7.68744361 -0.72271654 -1.61416545
## 7      2.36916357  0.4193862857 -0.08536256  0.21435862 -0.76819494  2.51267355
## 8      1.38929740  0.3037583952 -0.20473071  0.24231684 -0.43622101  1.43180882
## 9     -1.37827486  0.0570889699 -0.32799989  0.25949518  0.68907729 -1.62212331
## 10    -2.71602072  0.0013993340 -0.27986231  0.21113800 -0.46586836 -2.66384223

```

## 11	1.22633192	0.3060168663	-0.18304142	0.22607237	0.36780635	1.47261332
## 12	1.36286095	0.3035023592	-0.19860596	0.24027786	-0.30682589	1.37907413
## 13	1.43988187	0.3075732782	-0.22072534	0.24961443	-0.08653376	1.47590502
## 14	-1.39445273	0.0850566514	-0.30911863	0.25317759	0.66486663	-1.45685588
## 15	-1.73242591	0.0619139941	-0.30066114	0.22611018	0.95296422	-1.70405913
## 16	2.52913979	0.5054921755	0.14637025	0.11908930	-0.78743269	3.52827716
## 17	1.21211365	0.2723062888	-0.23580085	0.25313008	-0.67732479	1.19557544
## 18	-1.68158912	-0.8132760111	0.73215632	-1.78352495	-0.96399161	-1.29327586
## 19	-1.59239517	0.0639166064	-0.30910632	0.25265297	-0.12438428	-1.67003051
## 20	1.86819223	0.4628868862	0.11585090	0.11820841	-0.39688066	1.92577528
## 21	-1.67647306	-4.6698456853	4.95673736	-5.39322724	-0.38305668	-1.65468700
## 22	-1.45993359	0.0514286500	-0.34824396	0.26083884	3.49534120	-1.67281892
## 23	1.85760367	0.3595578882	-0.19246268	0.24587119	-0.46265959	1.42704415
## 24	-2.12726256	0.0367872400	-0.27495988	0.22544646	-0.99518913	-2.05062439
## 25	-2.77665484	-0.0007974928	-0.27401590	0.22080205	-1.03461904	-2.67770617
## 26	-2.06828665	-1.9568383087	1.80640660	-2.71304016	-0.49132411	-1.72225360
## 27	1.51534046	0.5971670379	0.21784745	0.20771823	2.54338422	0.55271857
## 28	1.33229906	0.3556537449	-0.02365003	0.15300191	-0.53547278	1.35242026
## 29	1.36239158	0.3156877808	-0.19259045	0.23579691	0.36020012	1.40192866
## 31	0.65215884	0.3577270850	0.20193453	0.01691024	-0.83989282	1.17188901
## 36	0.46029706	-6.6663853870	6.94478242	-5.58865183	-0.69762344	1.27314742
## 52	0.55718702	0.3634516355	0.28083267	-0.02947392	-0.60231356	0.60727133
## 54	0.12840515	0.3179874209	0.17324594	-0.01181657	-0.71696573	0.21919760
## 57	3.24097893	0.3748034774	-0.30781438	0.26138590	6.67223990	3.17627500
## 59	-0.28206288	0.1474555485	-0.33066477	0.26119292	4.00029372	-0.33059901
## 61	-1.85840178	-0.2863665039	0.43734457	-0.55976652	0.40332070	-1.59302648
## 62	-2.09287464	-7.6676476946	6.84457353	-7.68744361	-0.72271654	-1.61416545
## 63	0.46029706	-6.6663853870	6.94478242	-5.58865183	-0.69762344	1.27314742
## 64	-0.05052663	-2.5698552174	2.52514515	-1.67860286	-0.88485676	0.64777719
## 66	0.04819839	0.1665838352	-0.31204175	0.26069700	1.27986811	-0.19676331
## 67	0.06878914	0.2167553886	0.47359174	-0.30664353	-0.58032373	0.10046606
## 69	-0.17807664	-2.5928096321	3.17100719	-2.29107048	-0.87353800	0.19152622
## 70	-1.68158912	-0.8132760111	0.73215632	-1.78352495	-0.96399161	-1.29327586
## 72	0.09438870	0.1739567374	-0.31785954	0.26058283	1.15602044	0.00965968
## 73	-1.67647306	-4.6698456853	4.95673736	-5.39322724	-0.38305668	-1.65468700
## 75	0.01926630	0.1801361881	-0.30676506	0.26039653	1.33727150	-0.11852813
## 76	0.17260198	0.1013427982	0.72491903	-0.31511514	-0.90806671	1.42640360
## 77	0.46108104	0.1802200625	-0.31109608	0.26028477	1.51428686	0.18707478
## 78	-2.06828665	-1.9568383087	1.80640660	-2.71304016	-0.49132411	-1.72225360
## 82	-0.79663115	-1.1231189555	1.20168263	-2.24497106	-0.32634908	-0.79409691
##	logis_c	logis_e	area_under_curve	phenotype	cell_line	n
## 1	0.04505523	0.345586931	4.7274676	susceptible	293T	4
## 2	0.78717776	-0.315687810	-1.8792011	resistant	CTR_M2_05	4
## 3	0.78266138	-2.236166817	-1.1502481	resistant	CTR_M2_05_21	3
## 4	1.22438116	0.395132973	-0.6222876	resistant	Dard_2	4
## 5	1.59782705	1.681207451	-1.7463770	resistant	Eika_2	5
## 6	1.64470105	3.725626638	-1.4174523	resistant	Fafq_1	5
## 7	1.21075476	0.879023564	1.1674719	susceptible	Iill_1	4
## 8	1.25511634	0.686906345	1.1487306	susceptible	Iisa_3	4
## 9	-0.79636031	0.019445941	-0.9541170	resistant	Iuad_2	4
## 10	1.31682276	0.241950524	-1.7353425	resistant	Kajh_3	4
## 11	-0.10937860	-0.126953081	1.3297774	susceptible	Kehc_2	4
## 12	-0.64112753	-0.754235686	1.0176264	susceptible	Kolf_3	3
## 13	0.94886744	0.444008505	1.5340430	susceptible	Laey_6	4

## 14	0.58125381	0.321340421	-0.7298805	resistant	Liq_1 3
## 15	1.09769605	0.374412810	-0.9742222	resistant	Miaj_6 3
## 16	1.02116571	0.488407311	0.7882949	susceptible	Oarz_22 3
## 17	-1.79889938	-1.167779192	0.9959481	susceptible	Oibg_1 3
## 18	0.76919060	0.559721593	-2.1674507	resistant	Otam_2 4
## 19	-0.14114231	-0.515159392	-1.2007986	resistant	Pahc_5 3
## 20	1.59002928	1.995830815	0.1920732	susceptible	Robp_3 4
## 21	1.65398666	2.754689436	-1.9000064	resistant	Ruql_3 5
## 22	0.38515513	0.316469877	-0.5032298	resistant	Sebn_4 6
## 23	-0.58471376	-0.575847350	0.5981263	susceptible	Sebz_1 3
## 24	0.36700953	-1.853139886	-2.0241708	resistant	Timk_4 4
## 25	0.53858240	-1.763225675	-1.2587220	resistant	Tuju_1 4
## 26	1.45710477	1.927259027	-1.9239447	resistant	Uaq_1 5
## 27	0.04826202	0.004393533	1.0115718	susceptible	Vass_1 3
## 28	-0.58428260	0.348187762	0.4598859	susceptible	Xegx_1 3
## 29	1.16730007	0.955117249	1.3658417	susceptible	Zexw_3 4
## 31	1.41337320	1.529270897	-0.6694225	susceptible	Bipt_1 5
## 36	1.63109434	3.521387835	-1.2747026	resistant	Febc_1 3
## 52	1.59490788	2.171862647	-0.8661276	susceptible	Uimo_1 4
## 54	1.56910714	2.084708589	-1.1012058	susceptible	Xavk_3 4
## 57	0.04505523	0.345586931	4.7274676	resistant	293T 3
## 59	0.81885650	0.420738436	0.9293527	resistant	Ceik_1 3
## 61	1.59782705	1.681207451	-1.7463770	susceptible	Eika_2 3
## 62	1.64470105	3.725626638	-1.4174523	susceptible	Fafq_1 3
## 63	1.63109434	3.521387835	-1.2747026	susceptible	Febc_1 4
## 64	-0.26892680	-0.005039933	-0.8988614	susceptible	Gesg_2 3
## 66	0.92778720	0.440295599	0.9516241	resistant	Kucg_2 3
## 67	1.60563476	2.243183738	-1.1927011	susceptible	Lexy_2 3
## 69	0.96112383	1.230846284	-1.2667539	susceptible	Oilg_3 3
## 70	0.76919060	0.559721593	-2.1674507	susceptible	Otam_2 3
## 72	0.53414350	0.364156552	1.1870694	resistant	Puie_4 3
## 73	1.65398666	2.754689436	-1.9000064	susceptible	Ruql_3 3
## 75	0.55267657	0.356658125	0.9796917	resistant	Sehp_2 3
## 76	1.37997093	1.811536597	-1.2458728	susceptible	Suop_5 5
## 77	-0.20648674	-0.204807949	0.9660652	resistant	Toss_3 3
## 78	1.45710477	1.927259027	-1.9239447	susceptible	Uaq_1 3
## 82	1.59916664	2.060554373	-0.9771668	susceptible	Yemz_1 3

```
common_vector_list=unique(most_common_vector)
```