# Bacterial Abundance qPCR

*Nate Olson*

*2016-12-15*

## Summary

**Objective**: Determine the proportion of bacterial DNA in the mixture study samples for use in correcting for compositional biases.

**Method**: Bacterial abundance quantified using Zymo bacterial concentration assay. The assay targets the 16S rRNA gene and uses *E. coli* as a standard. *E. coli* genome is approximatly 5 Mb with 6 copies of the 16S rRNA gene.

**Results**

**Conclusions**

### Questions

- How much template was added to each reaction?
- Deviations in the average population genome size and 16S rRNA gene copy number per genome will bias the abundance estimate. May want to consider reporting abundance measurements as number of 16S rRNA gene copies rather than bacterial DNA concentration.

## Objective

The proportion of pre and post exposure samples in individual titrations is dependent on the ratios at which the two samples were mixed. This assumes that the pre and post samples have equivalent proportions of bacterial to non-bacterial DNA. To validate this assumption the concentration of bacterial DNA was assayed using qPCR. Additionally, the concentation of bacterial DNA in the titrations was assayed.

## Methods

- zymo qPCR assay - https://www.zymoresearch.com/dna/dna-analysis/femto-bacterial-dna-quantification-kit
- 45 Samples - all mixed and unmixed
- diluted samples - need to find out how they were diluted
- triplicates per sample - 135 reactions
  - three qPCR plates, one replicate of each sample ran on each plate
- 7 concentration standard curve
- Experimet was run twice (on seperate days) using the Zymo standard curve and once using an in-house standard curve. The in-house standard curve was 10-fold dilutions of genomic DNA from an *E. coli* strain.

## Results

### Standard Curve Analysis

Fitting the standard curve to a linear model, `Ct ~ log10(concentration)`.
The expected slope for the standard curve is -3.33 indicating a perfect doubling every PCR cycle, for a amplification factor (AF) of 2 and efficiency (E) of 1.

$$AF = 10^{-1/slope}$$

$$E = 10^{-1/slope} - 1$$

The model was fit using the full standard curve and only points in the standard curve with concentrations greater than 0.02 ng/ul. Fitting the regression to all concentrations in the standard curve resulted in a lower amplification efficiency and $R^2$ for all three standard curves.

```
fit_mod_full <- qpcrBacStd %>%
      filter(!is.na(Ct)) %>% mutate(log_conc = log10(conc)) %>%
      ## excluding standard curve outlier
      #filter(std != "zymo" | date != "2016-12-09" | conc != 0.00002 | plate != "plate3") %>%
      group_by(date, std) %>% nest() %>%
      mutate(fit = map(data, .f=~lm(Ct~log_conc ,data = .)), mod = "full")

fit_mod_sub <- qpcrBacStd %>%
      filter(!is.na(Ct)) %>% mutate(log_conc = log10(conc)) %>%
      ## excluding standard curve points outside of sample Ct value range
      filter(conc >= 0.2) %>%
      group_by(date, std) %>% nest() %>%
      mutate(fit = map(data, .f=~lm(Ct~log_conc ,data = .)), mod = "sub")

fit_mod <- bind_rows(fit_mod_full, fit_mod_sub)

fit_list <- fit_mod$fit %>% set_names(paste(fit_mod$date, fit_mod$std, fit_mod$mod))

fit_coefs <-fit_list %>% map_df(coefficients) %>%
      add_column(coefs = c("intercept","slope")) %>%
      gather("std","stat",-coefs) %>% spread(coefs, stat)

std_fit <- fit_list %>% map_df(broom::glance, .id = "std") %>%
      select(std, adj.r.squared) %>% left_join(fit_coefs) %>%
      separate(std, c("date","std","mod"), sep = " ") %>%
      mutate(amplification_factor = 10^(-1/slope),
             efficiency = (amplification_factor - 1) * 100)
```

```
## Joining, by = "std"
```

The efficiency and precision ($R^2$) were highest for the in-house standard curve. Fitting the regression model for the experiment run on 12/09 without the low concentration plate 3 outlier resulted in a lower efficiency (76% versus 80%) but higher $R^2$ (0.997 versus 0.986).
Fitting the regression to only the standard with concentrations within the observed range of the samples (20 ng/ul, 2 ng/ul, and 0.2 ng/ul) compared to all standard curve samples resulted in a higher amplification efficiency but slightly lower $R^2$ values for all three standard curves. The higher efficiency

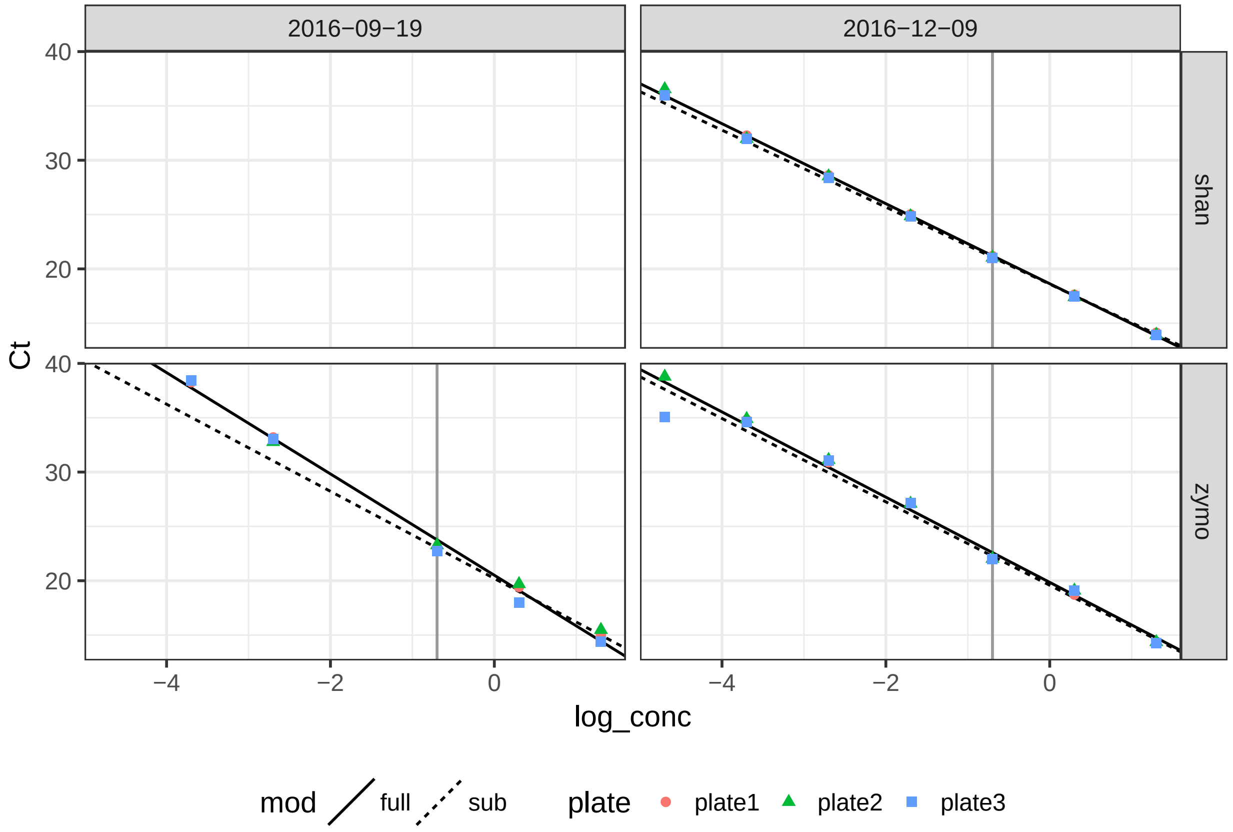


Figure 1: qPCR bacterial abundance standard curves. Using two different standards and performed on two different days. Two models were fit to the standard curves, one with all data point and a second with only the 20 ng/ul, 2 ng/ul, and 0.2 ng/ul standards. Grey vertical line indicates concentration cutoff for subset model.

```
std_fit %>% select(std, date, mod, efficiency, adj.r.squared) %>%
      arrange(date, std) %>% knitr::kable()
```

| std | date | mod | efficiency | adj.r.squared |
|---|---|---|---|---|
| zymo | 2016-09-19 | full | 63.86615 | 0.9933987 |
| zymo | 2016-09-19 | sub | 77.57709 | 0.9717054 |
| shan | 2016-12-09 | full | 86.86251 | 0.9992791 |
| shan | 2016-12-09 | sub | 91.49338 | 0.9990023 |
| zymo | 2016-12-09 | full | 79.89735 | 0.9859375 |
| zymo | 2016-12-09 | sub | 82.21117 | 0.9818836 |

```
qpcrBacStd %>% mutate(log_conc = log10(conc)) %>% ggplot(aes(y = Ct, x = log_conc)) +
      geom_vline(aes(xintercept = log10(0.2)), color = "grey60") +
      geom_abline(data = std_fit,
                  aes(intercept = intercept, slope = slope, linetype = mod)) +
      geom_point(aes(color = plate, shape = plate)) +
      facet_grid(std~date) + theme_bw() +
      theme(legend.position = "bottom")
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

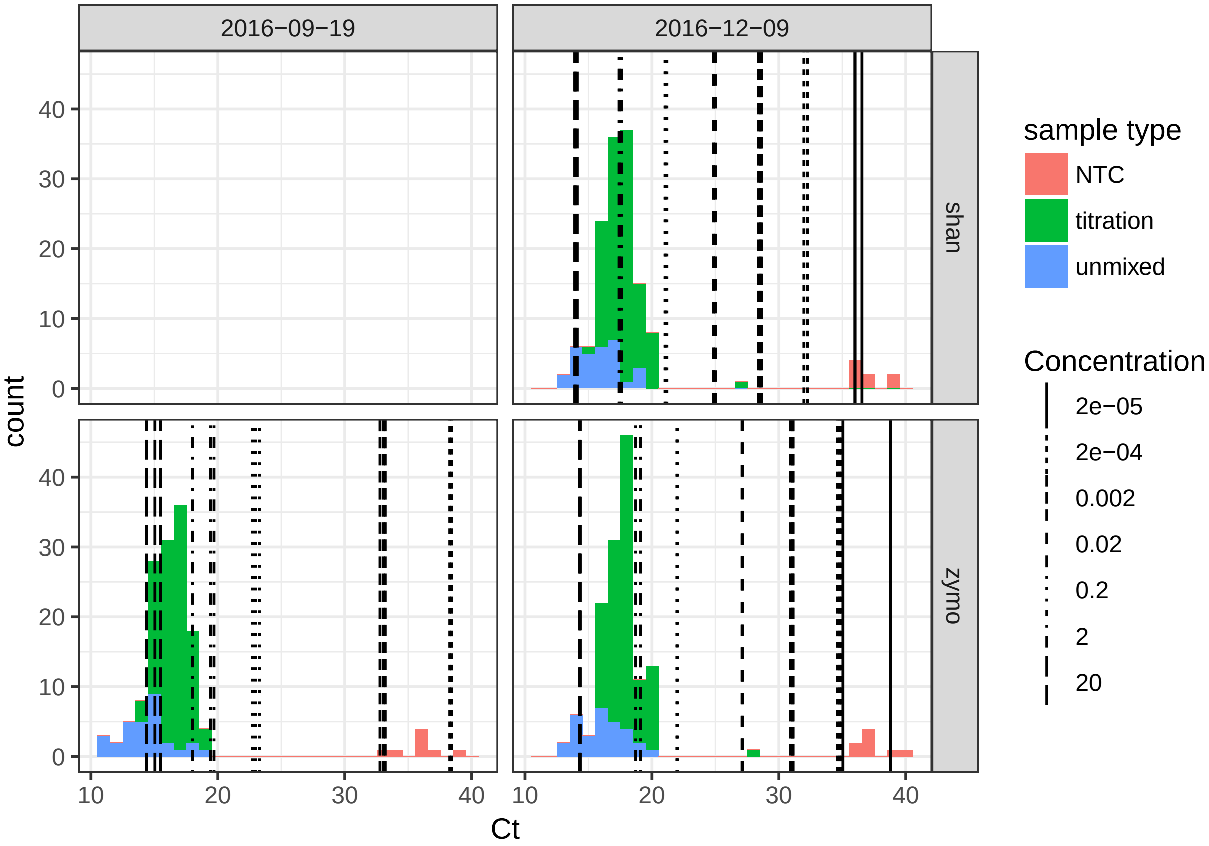


Figure 2: Distribution of unmixed and titration sample Ct values relative to Ct values for standard curve samples.

```
qpcrBacAbu %>% ggplot() +
      geom_histogram(aes(x = Ct, fill = sam_type)) +
      geom_vline(data = qpcrBacStd,aes(xintercept = Ct, linetype = factor(signif(conc,2)))) +
      facet_grid(std~date) + theme_bw() +
      labs(fill = "sample type", linetype = "Concentration")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 8 rows containing missing values (geom_vline).
```

**Conclusion:** The in-house standard curve with standard concentrations 20 ng/ul, 2ng/ul, and 0.2 ng/ul resulted in the best overall combination of precision ($R^2$) and amplification efficiency.

### Predicting Sample Concentration

Using in-house with standard concentrations 20 ng/ul, 2ng/ul, and 0.2 ng/ul to predict sample concentrations.

```
mod <- qpcrBacStd %>%
      filter(std == "shan", conc >= 0.2) %>% mutate(log_conc = log10(conc)) %>%
      {lm(log_conc~Ct, data = .)}

bac_abu <- qpcrBacAbu %>% filter(!is.na(Ct), std == "shan") %>% add_predictions(mod) %>%
```

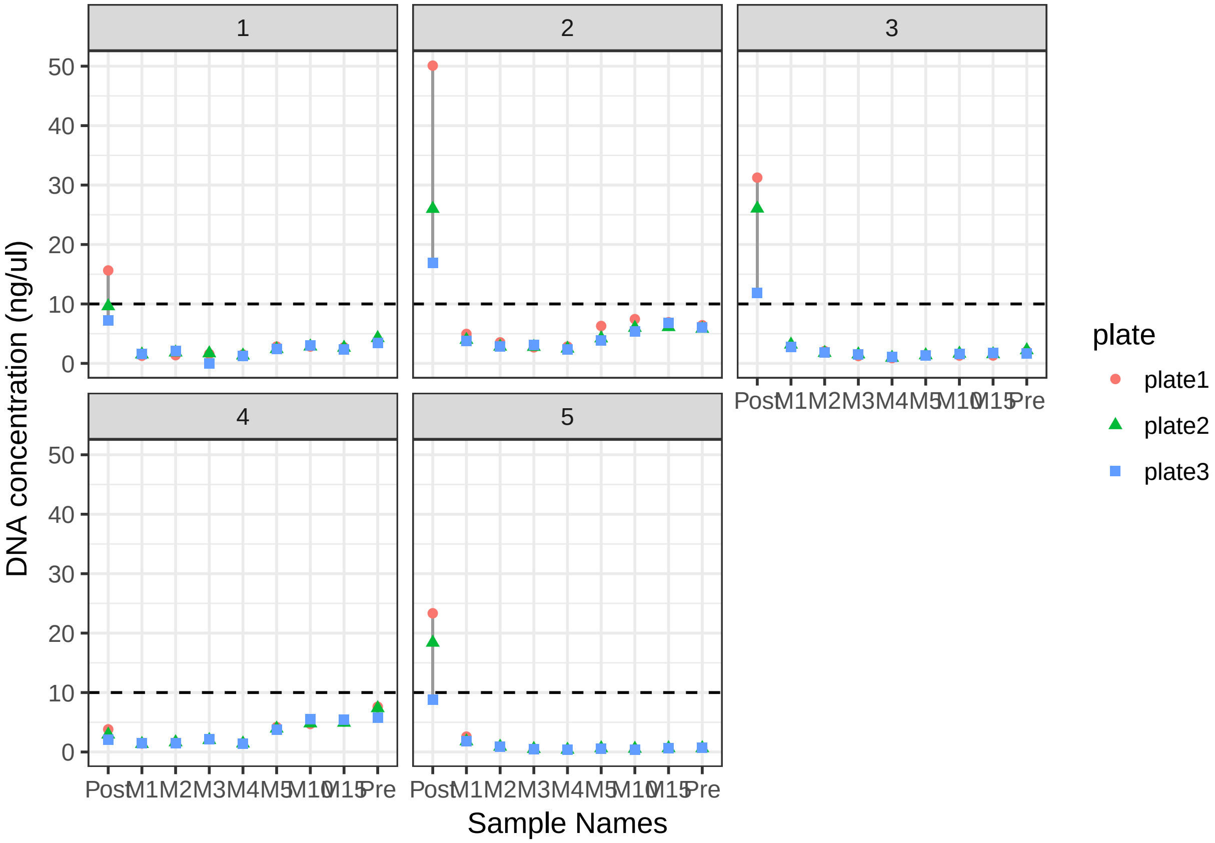


Figure 3: Predicted mixture study sample concentrations.

```
      mutate(quant = 10^pred) %>% group_by(sample_name) %>%
      mutate(quant_min = min(quant), quant_max = max(quant))
```

The unmixed sample concentrations were normalized to 10 ng/ul prior to making the titrations therefore all samples are expected to have concentrations less than 10 ng/ul.
**NEED TO FIGURE OUT WHY THE SAMPLE CONCENTRATIONS ARE GREATER THAN 10 NG/UL**

```
bac_abu %>% filter(sample_name != "NTC") %>% ungroup() %>%
      mutate(sample_name = gsub(" ","_", sample_name)) %>%
      separate(sample_name, c("bio_rep","titration"), sep = "_") %>%
      mutate(titration = fct_relevel(titration, c("Post",paste0("M",c(1,2,3,4,5,10,15)),"Pre"))) %>%
      ggplot() +
            geom_hline(aes(yintercept = 10), linetype = 2) +
            geom_linerange(aes(x = titration,ymin = quant_min, ymax = quant_max), color = "grey60") +
            geom_point(aes(y = quant, x = titration, color = plate, shape = plate)) +
            facet_wrap(~bio_rep) +
            theme_bw() + labs(x = "Sample Names", y = "DNA concentration (ng/ul)")
```

### Estimating Pre and Post Bacterial Proportions

Need to estimate the proportion of DNA in the unmixed sample that is bacterial to correct for differences in bacterial DNA proportions.

### Using qPCR bacterial quantification values

As the pre and post treatment samples were diluted to 10 ng/ul prior to generating the titration series the unmixed samples should have less than 10 ng/ul and the proportion of bacterial DNA in the samples is obtained by dividing the estimated concentration by 10. Due to the estimated post treatment sample concentrations greater than 10, we will set these samples at 10 ng/ul.

```
bac_abu %>% filter(sam_type == "unmixed") %>% ungroup() %>%
      mutate(sample_name = gsub(" ","_", sample_name), quant = if_else(quant > 10, 10, quant)) %>%
      separate(sample_name, c("bio_rep","titration"), sep = "_")  %>%
      group_by(bio_rep, titration) %>% summarise(bac_prop = median(quant)/10) %>%
      spread(titration, bac_prop)
```

```
## Source: local data frame [5 x 3]
## Groups: bio_rep [5]
##
##   bio_rep      Post        Pre
## *   <chr>     <dbl>      <dbl>
## 1       1 0.9619490 0.35332052
## 2       2 1.0000000 0.60633949
## 3       3 1.0000000 0.20336004
## 4       4 0.2944102 0.73765265
## 5       5 1.0000000 0.06824916
```

### TODO Estimated QUANT SANITY CHECK

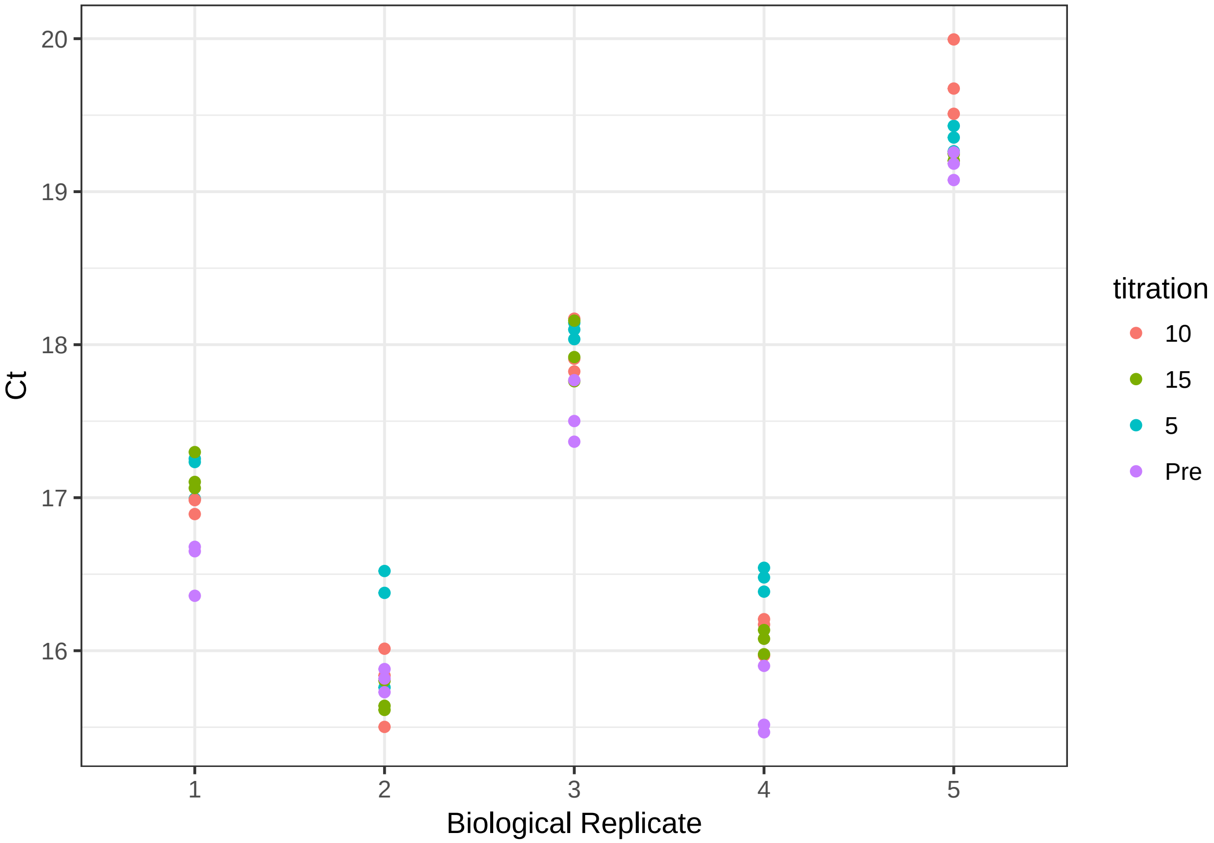
### Using titration samples predicted concentrations to infer unmixed sample concentrations.

For titrations $2^{-5}$ the maximum amount of bacterial DNA from the post treatment sample is 0.65 ng/ul. This is assuming an initial sample consentation of 10 ng/ul and all of the DNA is bacterial. Therefore, we can use the estimated concentrations for the $2^{-5}$, $2^{-10}$, and $2^{-15}$ two-sample titrations along with the unmixed pre-treatment sample to estimate the concentration of bacterial DNA in the pre-treatment samples.

```
pre_sams <- c(paste(1:5,"Pre"), paste0(rep(1:5, each =3),"_M",rep(c(5,10,15),5)))
bac_pre <- bac_abu %>% filter(sample_name %in% pre_sams, Ct < 25) %>% ungroup() %>%
      mutate(sample_name = gsub(" ","_", sample_name),
             sample_name = gsub("M","", sample_name)) %>%
      separate(sample_name, c("bio_rep","titration"), sep = "_")
bac_pre_mean <- bac_pre %>% group_by(bio_rep) %>%
      summarise(stine_quant = mean(stine_quant), quant = mean(quant))
```

The precision of qPCR is around 0.5 to 1 Ct, as the qPCRs from the titrations and mixtures are within 1 Ct we can use the estimated concentrations from all of the samples to estimate the concentration of the unmixed pre-treatment sample.

```
ggplot(bac_pre) + geom_point(aes(x = bio_rep,y = Ct, color = titration)) +
      theme_bw() + labs(x = "Biological Replicate")
```



```
ggplot(bac_pre) +
      geom_jitter(aes(x = bio_rep,y = quant, color = titration), width = 0.15) +
      geom_text(data = bac_pre_mean,
                aes(x = bio_rep, y = quant, label = signif(quant,3))) +
      theme_bw()+
      theme(legend.position = "bottom") +
      labs(x = "Biological Replicate", y = "Bacterial DNA (ng/ul)")
```

### Estimating Post Treatment Bacterial DNA Concentrations

Overall range of Ct values for different treatments. Two sets of distributions for titrations with titration factors 1-4, and titrations factors 5, 10, and 15. Due to the higher than expected bacterial DNA concentration estimate for post treatment sample will try to estimate the concentration value using the pre-treatment concentration estimated in the previous section and the estiamted concentrations for the titration factor 1-4 samples.

```
bac_abu %>% filter(sample_name != "NTC", Ct < 25) %>% ungroup() %>%
      mutate(sample_name = gsub(" ","_", sample_name)) %>%
      separate(sample_name, c("bio_rep","titration"), sep = "_") %>%
      mutate(titration = fct_relevel(titration, c("Post",paste0("M",c(1,2,3,4,5,10,15)),"Pre"))) %>%
      ggplot() +
            geom_point(aes(y = Ct, x = titration, color = plate, shape = plate)) +
            facet_wrap(~bio_rep) +
            theme_bw() + labs(x = "Sample Names")
```

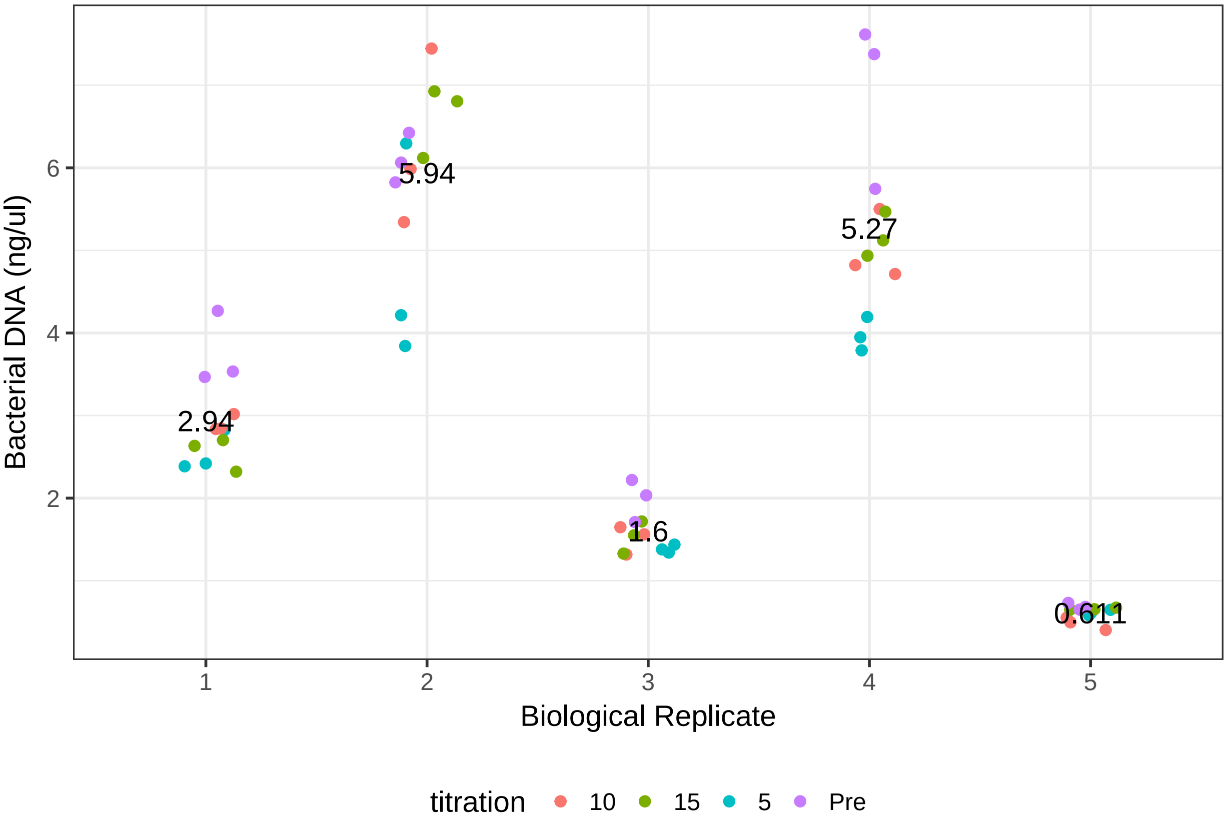
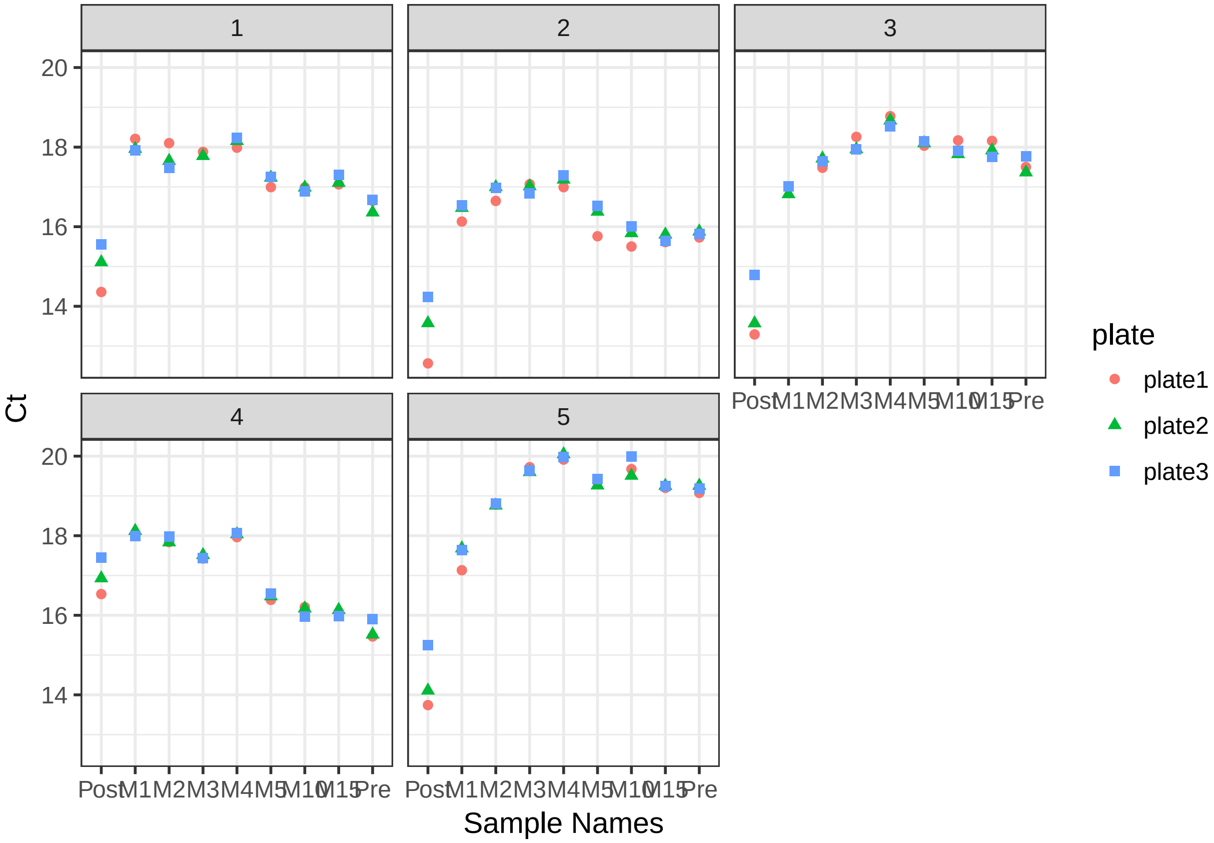


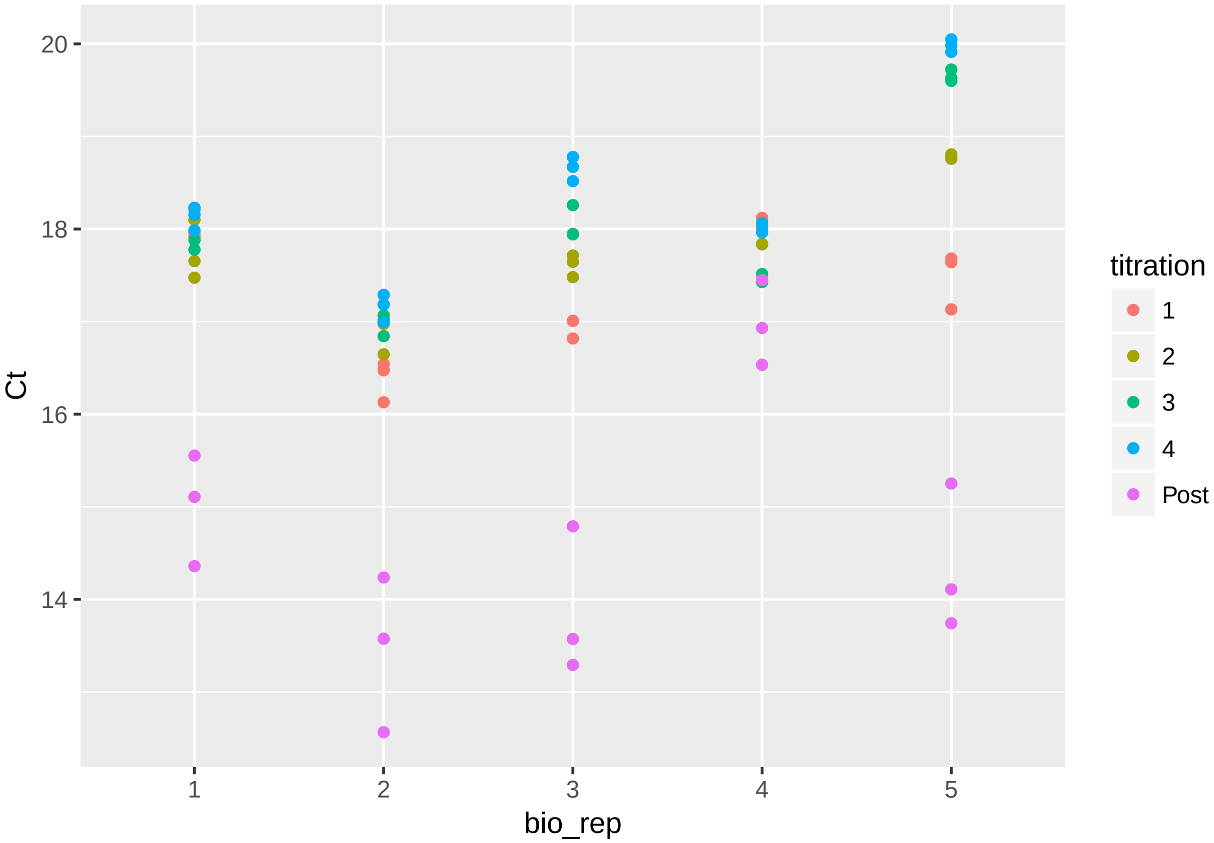
Figure 4: Bacterial DNA concentration for unmixed pre-treatment and titration samples used to estimate pre-treatment bacterial sample concentrations. Mean values indicated for each biological replicate.



Plot of titration concentrations for titrations $2^{-1}$, $2^{-2}$, $2^{-3}$, and $2^{-4}$

```
post_sams <- c(paste(1:5,"Post"), paste0(rep(1:5, each =4),"_M",rep(1:4,5)))
bac_post <- bac_abu %>% filter(sample_name %in% post_sams, Ct < 25) %>% ungroup() %>%
      mutate(sample_name = gsub(" ","_", sample_name),
             sample_name = gsub("M","", sample_name)) %>%
      separate(sample_name, c("bio_rep","titration"), sep = "_")
```

```
ggplot(bac_post) + geom_point(aes(x = bio_rep, y = Ct, color = titration))
```



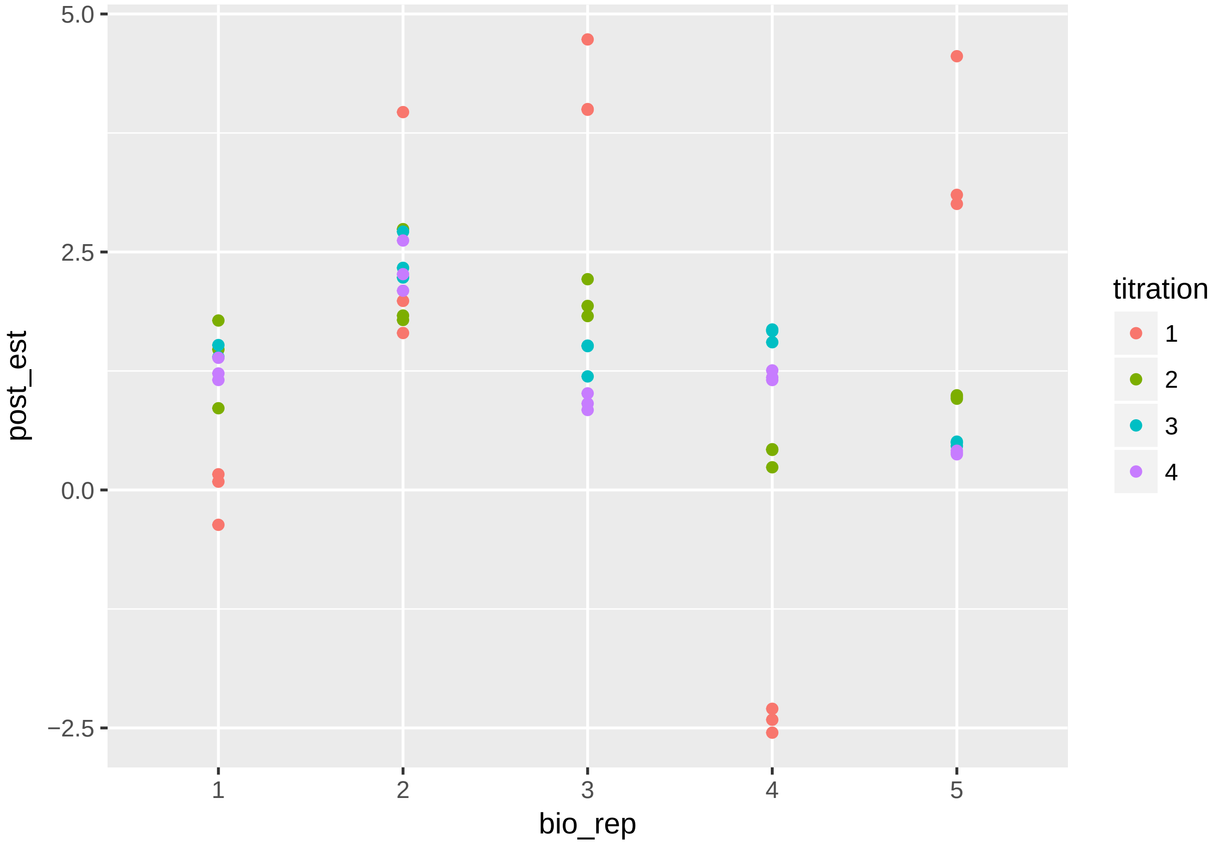
Calculating Post from pre estimate and titrations

```
bac_pre_mean <- bac_pre %>% group_by(bio_rep) %>% summarise(pre_quant = mean(quant))
bac_post_est <- bac_post %>% filter(titration != "Post") %>% select(bio_rep, titration, quant) %>%
      left_join(bac_pre_mean) %>%
      mutate(pre_tprop = 2^-as.numeric(titration), post_tprop = 1-pre_tprop,
             post_est = (quant - pre_quant*pre_tprop)/post_tprop)
```

```
## Joining, by = "bio_rep"
```

The estimated concentration values are low, less than 0 for a number of samples. If the concentations are so low, why are the measured concentations are so much higher for unmixed post treatment samples.

```
bac_post_est %>%
      ggplot() + geom_point(aes(x = bio_rep, y = post_est, color = titration))
```



## Conclusions

- qPCR is not precise enough to accurately measure the different in the proportion of bacterial between the pre and post mixture samples.

- The high estiamted concentration (low Ct) values for the post treatment samples in concering. Potential reasons are non-specific amplification from other DNA in the samples or due to edge effects (differences in cycling temperature at the edge of the plate).

### Caveats

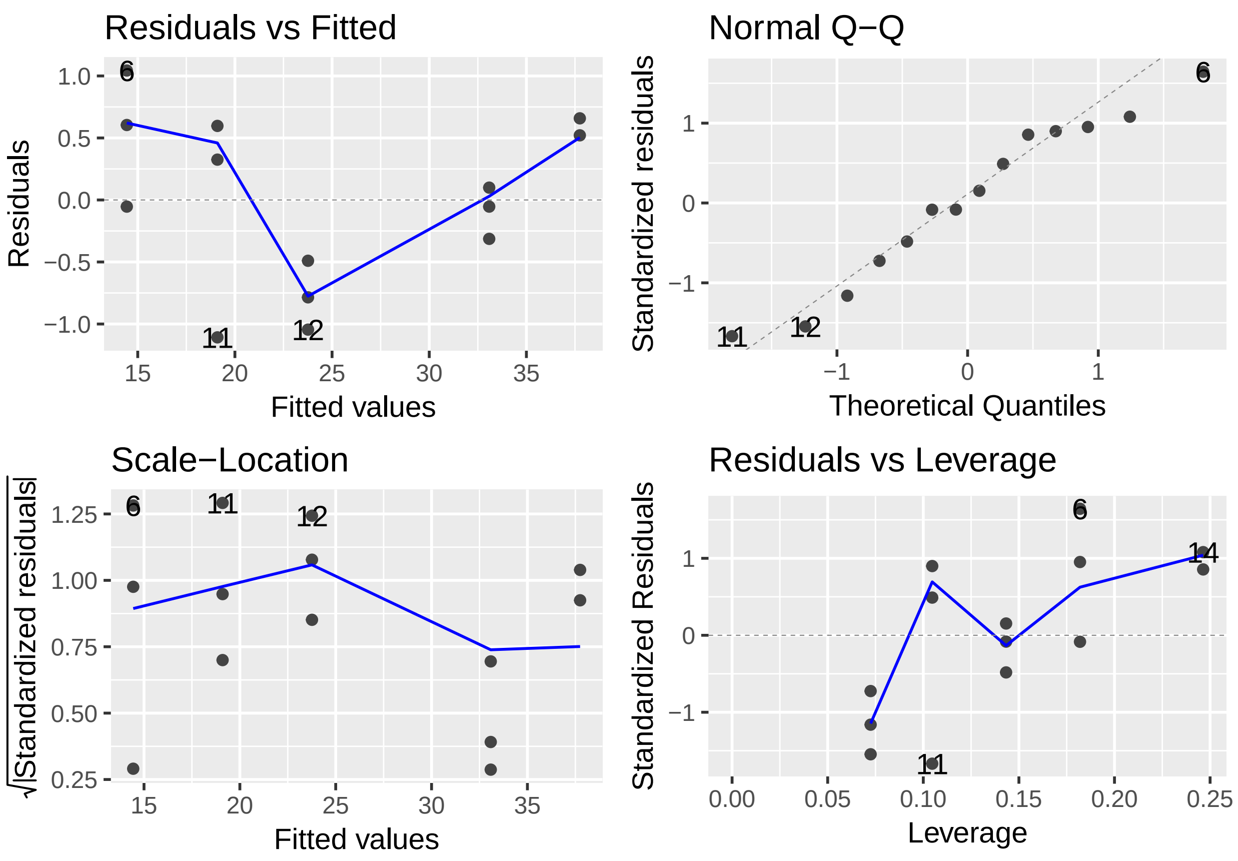
- To reduce experimental design complexity, sample name confounded with well except for negative controls (No Template Control, NTC).

## Appendix

### Linear regression fit diagnostic plots.

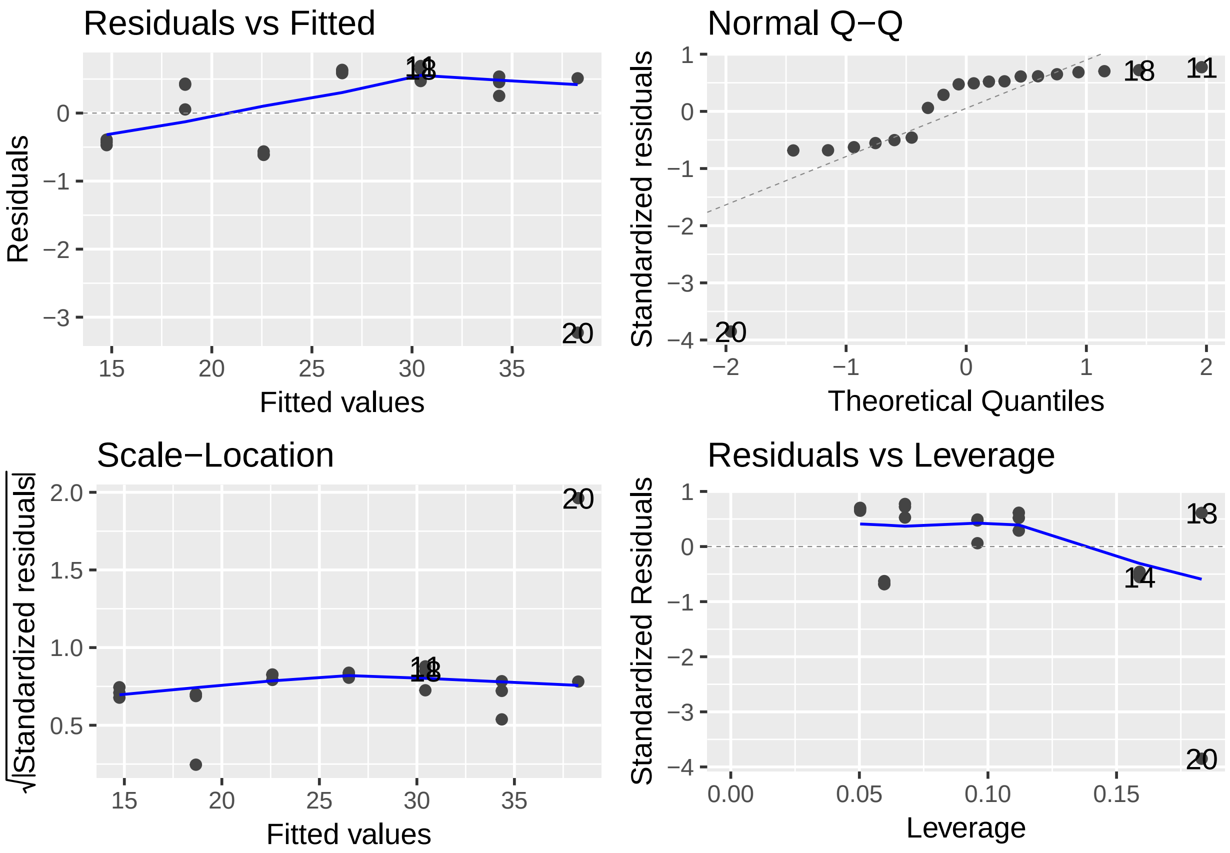
### 2016-09-19 zymo standard curve

```
fit_list$`2016-09-19 zymo full` %>% autoplot()
```



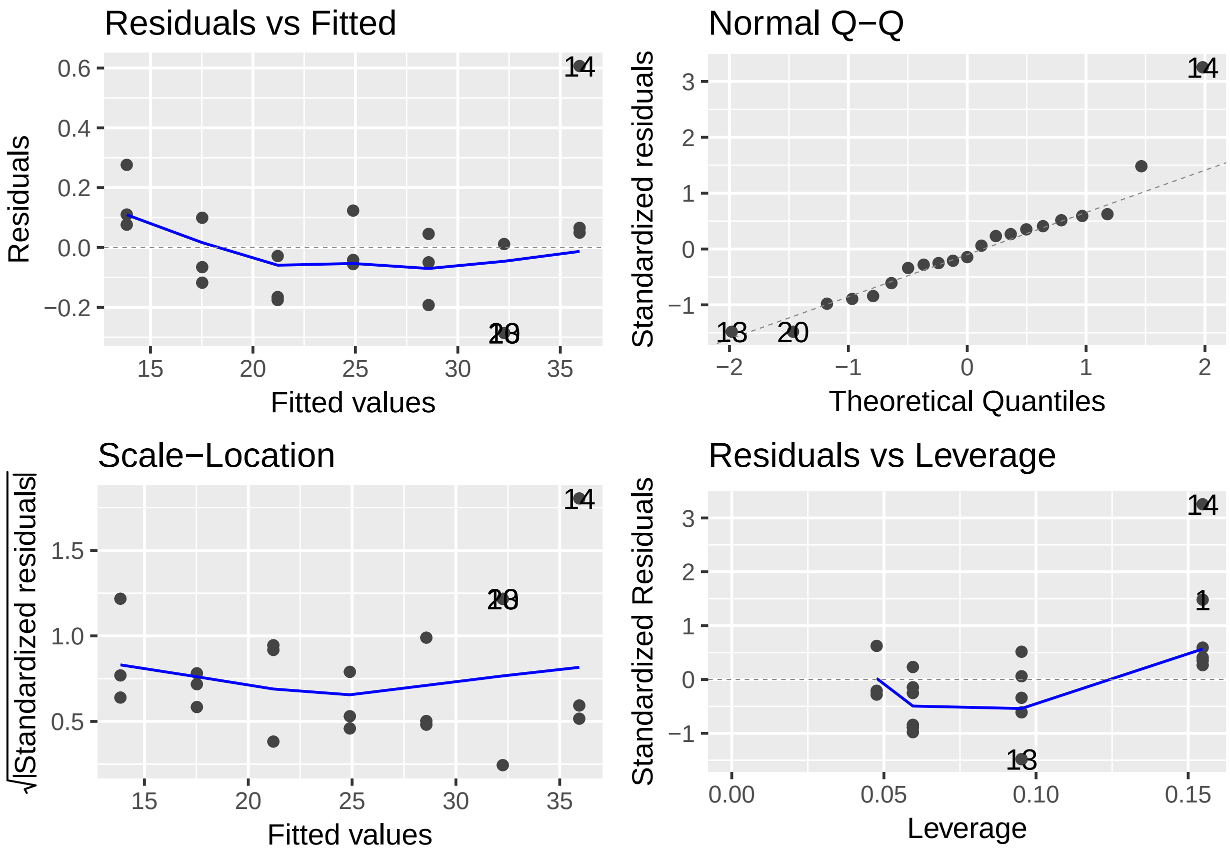
**2016-12-09 zymo standard curve**

```
fit_list$`2016-12-09 zymo full` %>% autoplot()
```



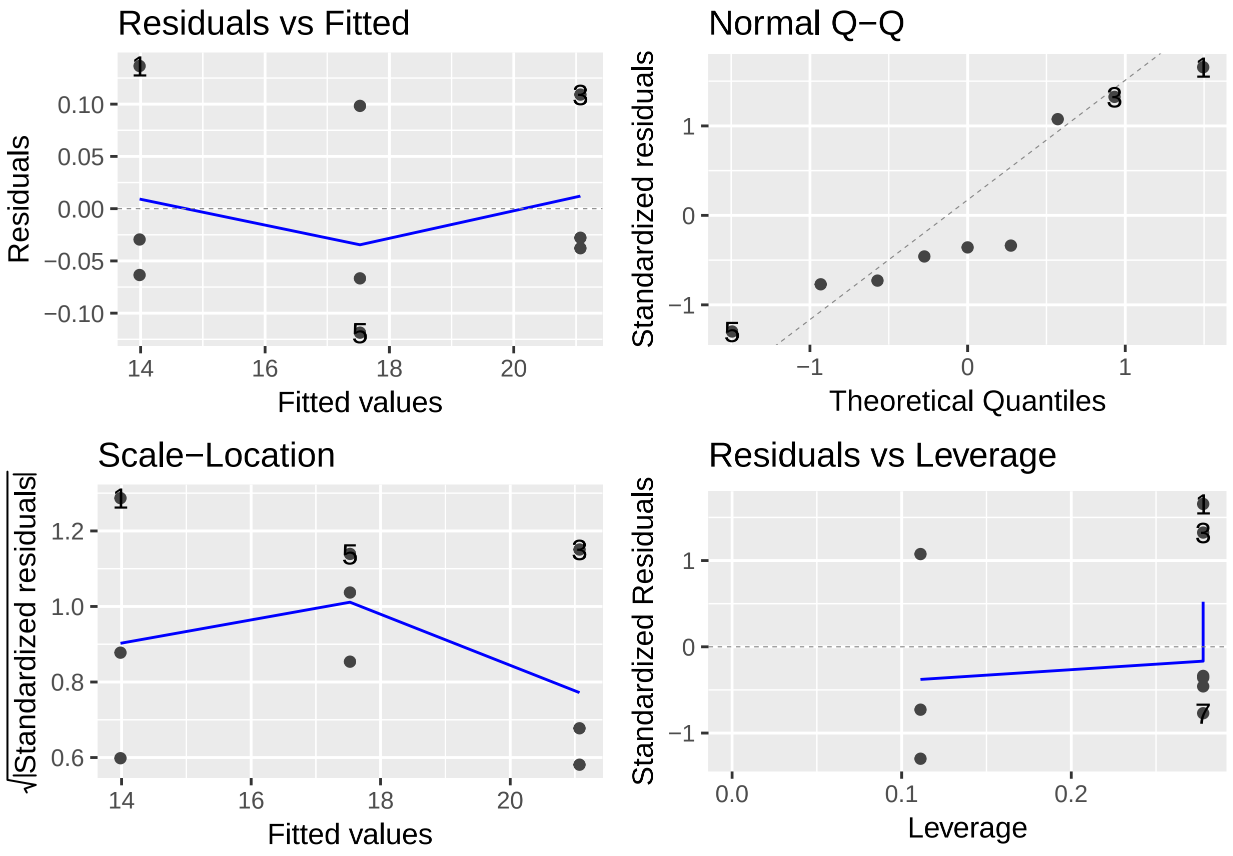
**2016-12-09 in-house standard curve**

```
fit_list$`2016-12-09 shan full` %>% autoplot()
```



**2016-12-09 in-house standard curve subset**

```
fit_list$`2016-12-09 shan sub` %>% autoplot()
```


Residuals vs Fitted
Residuals
0.10
0.05
0.00
−0.05
−0.10
14
16
18
20
Fitted values
1
3
5
Normal Q−Q
Standardized residuals
1
0
−1
−1
0
1
Theoretical Quantiles
5
3
1
Scale−Location
√|Standardized residuals|
1.2
1.0
0.8
0.6
14
16
18
20
Fitted values
1
5
3
Residuals vs Leverage
Standardized Residuals
1
0
−1
0.0
0.1
0.2
Leverage
1
3
7