Bacterial Abundance qPCR

Nate Olson

2016-10-06

Summary

- Objective - quantify bacterial abundance in starting DNA

- 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
  - Clean no template controls

  - R^2 standard curve 0.99

  - Large fraction of samples (especially unmixed) with Ct values outside standard curve.
- Conclusions
  - R^2 for standard curve close to 1, but might want to consider only fitting to three highest concentration standard due to lower residual standard error.
  - High estimated DNA concentration for post treatment unmixed samples due to Ct values outside of the standard curve resulted in high post to pre treatment sample concentration ratios.
  - Should consider diluting and rerunning samples with multiple replicates outside of the standard curve to verify concentration estimates.

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 - for the assay
  - issue with fourth standard

Munging qPCR Data

```
bac_con_raw <- read_excel(path = "../data/MixStudy_Nate_20160919.xls",
                      sheet = "QDNA_20160919",
                      skip = 11, na = "Undetermined", col_names = FALSE)
colnames(bac_con_raw) <- c("well","sample_name",
                       "plate1_Ct","plate1_quant",
                       "plate2_Ct","plate2_quant",
                       "plate3_Ct","plate3_quant")

bac_con <- bac_con_raw %>% gather("id","value", -well, -sample_name) %>%
      separate(id, c("plate","var"), sep = "_") %>%
      spread(var,value) %>%
      mutate(sam_type = if_else(grepl('\\(', sample_name), "unmixed","titration"),
             sam_type = if_else(sample_name == "NTC", "NTC",sam_type)) %>%
      rename(stine_quant = quant)

bac_unmixed <- bac_con %>% filter(sam_type == "unmixed") %>%
            mutate(sample_name = str_replace(sample_name, ".*_", ""),
                   sample_name = str_replace(sample_name, '\\(',""),
                   sample_name = str_replace(sample_name, '\\)',""))

bac_con <- bac_con %>% filter(sam_type != "unmixed") %>% bind_rows(bac_unmixed)
```

```
bac_std <- read_excel(path = "../data/MixStudy_Nate_20160919.xls",
                      sheet = "QDNA_20160919",skip = 3, col_names = FALSE) %>%
      select(-X12, -X13, -X5,-X7,-X9) %>% filter(X2 %in% paste0("Std",1:7))
colnames(bac_std) <- c("well","sample_name","conc","plate1","plate2","plate3")
bac_std <- bac_std %>% gather("plate","Ct",-well, -sample_name, -conc) %>%
      mutate(conc = as.numeric(conc), Ct = as.numeric(Ct)) %>% filter(!is.na(Ct))
```

```
## Warning in eval(substitute(expr), envir, enclos): NAs introduced by
## coercion
```

```
## NAs introduced when converting samples with undetermined and omit Ct values.
bac_std <- mutate(bac_std, log_conc = log10(conc))
```

Results

NTC Check

For NTC (no template control) samples with assigned Ct values, the Ct values were significantly different from the sample Ct values. Excluding NTC from the rest of the analysis. Standad curve adjusted R^2 is 0.9947.

```
ntc_sam_ct <- bac_con %>%
      mutate(sample_type = if_else(sample_name == "NTC", "NTC","samples")) %>%
      ggplot() + geom_point(aes(x = sample_type, y = Ct, color = plate)) + theme_bw()
## Warning about missing value due to one NTC reaction with a Ct value of NA
```

```
## plot shows okay to exclude NTC
bac_con <- bac_con %>% filter(sample_name != "NTC")
```

```
ntc_sam_ct
```

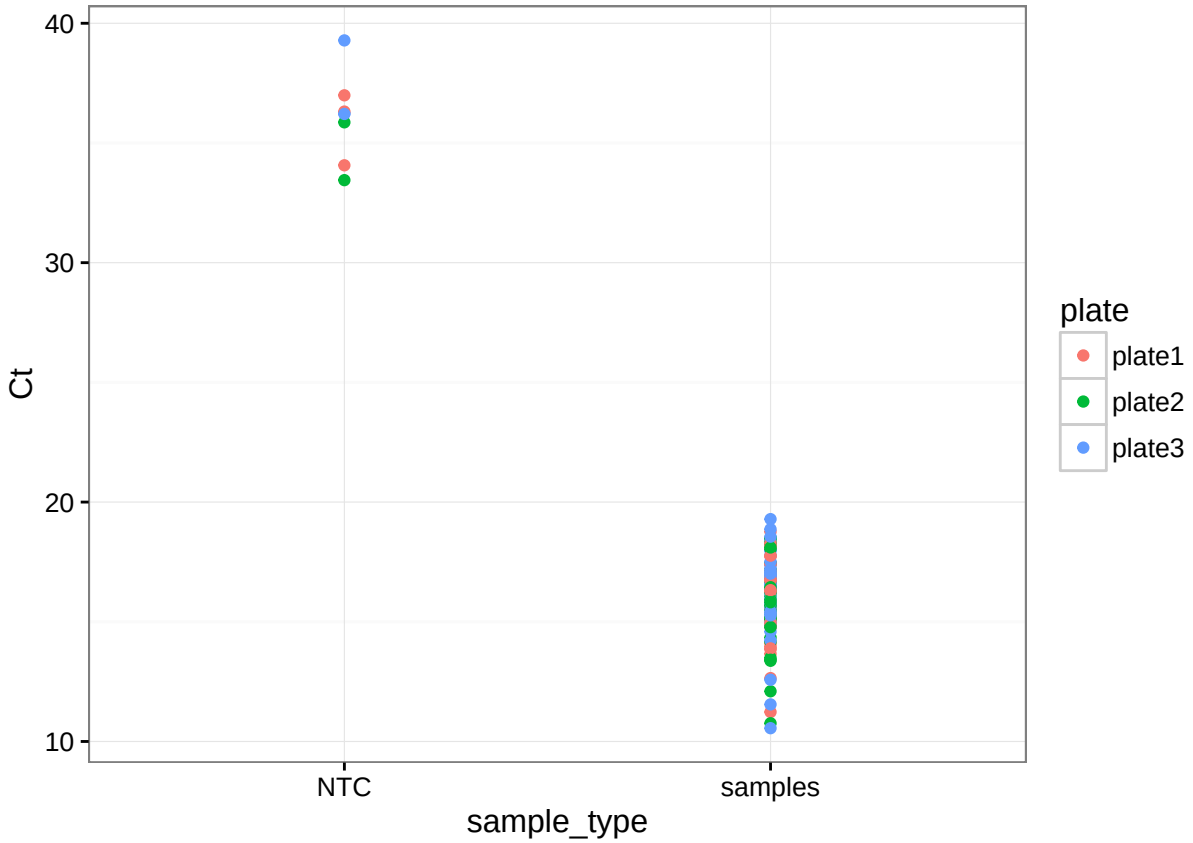


Figure 1: Comparison of sample and NTC Ct values.

Standard Curve Analysis

The standard curve was fit using a linear model with independent estimated slope and intercepts for the three replicate plates.

```
std_fit2 <- lm(log_conc~plate/Ct, data = bac_std)
summary(std_fit2)
```

```
##
## Call:
## lm(formula = log_conc ~ plate/Ct, data = bac_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19978 -0.07599  0.03538  0.06859  0.16901
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.417411   0.189273  23.339 1.21e-08 ***
```

```
## plateplate2     0.411203   0.313488   1.312    0.226
## plateplate3    -0.347403   0.260241  -1.335    0.219
## plateplate1:Ct -0.213979   0.006960 -30.743 1.36e-09 ***
## plateplate2:Ct -0.231415   0.010552 -21.930 1.97e-08 ***
## plateplate3:Ct -0.204254   0.006644 -30.741 1.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1347 on 8 degrees of freedom
## Multiple R-squared:  0.9967, Adjusted R-squared:  0.9947
## F-statistic: 487.5 on 5 and 8 DF,  p-value: 1.029e-09
```

Only fitting linear models to three highest concentration standard results in lower R^2 but also lower residual standard error.

```
bac_std_high <- bac_std %>% filter(conc > 0.1)
std_fit_high <- lm(log_conc~plate/Ct, data = bac_std_high)
summary(std_fit_high)
```

```
##
## Call:
## lm(formula = log_conc ~ plate/Ct, data = bac_std_high)
##
## Residuals:
##        1        2        3        4        5        6        7        8
## -0.03114  0.06956 -0.03842 -0.02511  0.05473 -0.02961  0.05037 -0.08883
##        9
##  0.03846
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.10966    0.30706  16.641 0.000472 ***
## plateplate2     0.18160    0.44213   0.411 0.708835
## plateplate3    -0.42726    0.41615  -1.027 0.380110
## plateplate1:Ct -0.25113    0.01581 -15.884 0.000543 ***
## plateplate2:Ct -0.25613    0.01611 -15.895 0.000541 ***
## plateplate3:Ct -0.23858    0.01504 -15.865 0.000544 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08887 on 3 degrees of freedom
## Multiple R-squared:  0.9961, Adjusted R-squared:  0.9895
## F-statistic: 151.3 on 5 and 3 DF,  p-value: 0.0008396
```

```
bac_fit <- bac_con %>% select(sample_name,plate, Ct, sam_type, stine_quant) %>% add_predictions(std_fit
std_coef <- std_fit2 %>% coefficients()
coef_df <- frame_data(
      ~plate, ~intercept, ~slope,
      "plate1",std_coef[1], std_coef[4],
      "plate2",std_coef[1] + std_coef[2], std_coef[5],
      "plate3",std_coef[1] + std_coef[3], std_coef[6]
)
knitr::kable(coef_df, caption = "Slope and intercepts for each plate infered using a linear model.")
```

Table 1: Slope and intercepts for each plate infered using a linear model.

plate	intercept	slope
plate1	4.417411	-0.2139791
plate2	4.828613	-0.2314146
plate3	4.070008	-0.2042542

```
fit2_df <- coef_df %>% mutate(mod = "log_conc~plate/Ct")
fit_coefs <- fit2_df
ggplot(fit_coefs) + geom_point(data = bac_std, aes(y = log_conc, x = Ct)) +
      geom_abline(aes(slope = slope, intercept = intercept), color = "grey60") +
      facet_grid(.~plate) + theme_bw() + labs(y = "DNA Concentration log10(ng/ul)")
```

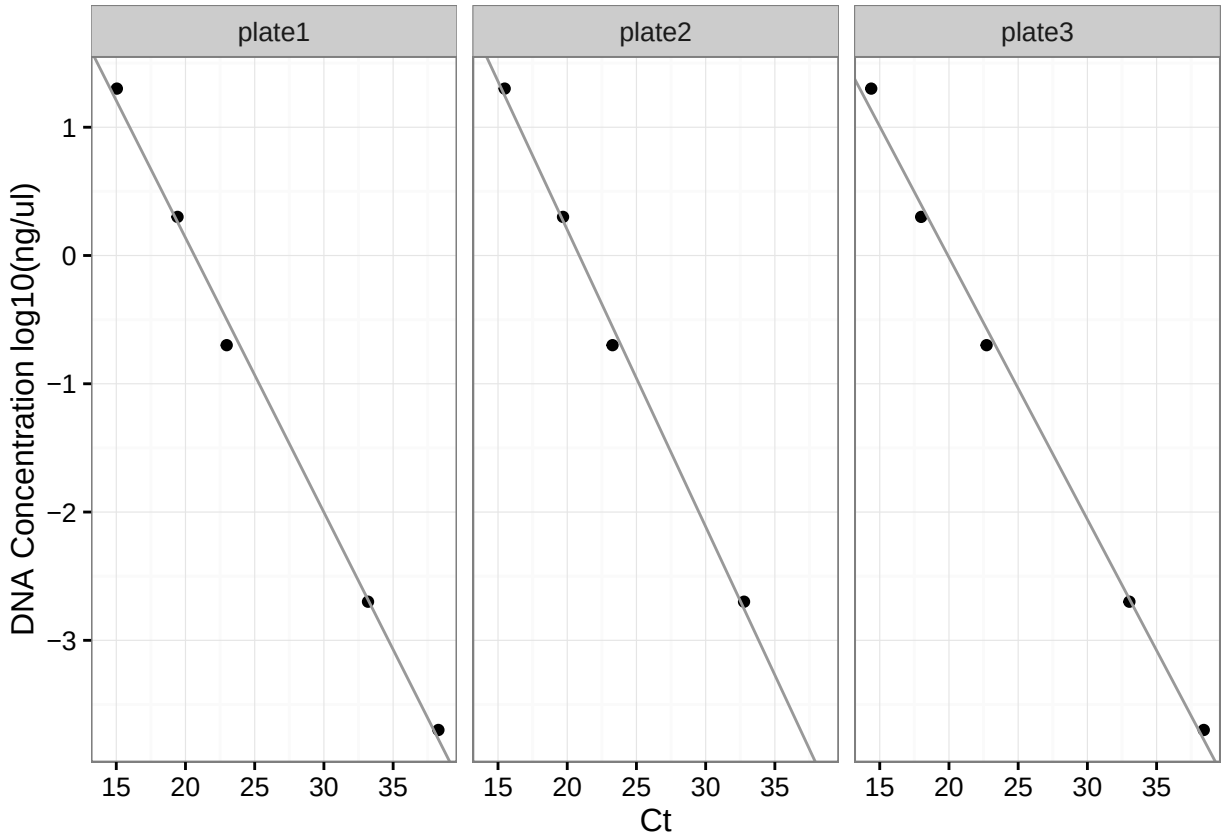


Figure 2: Standard curve fit using a linear model 'log(concentration)~Ct/plate'.

Predicted Sample Concentrations

The predicted sample concentration ranges from 1.35 ng/ul to over 218 ng/ul. Samples have Ct values lower then the standard curve samples and fall outside of the standard curve.

```
min_std <- bac_std %>% filter(conc == 20) %>% select(plate, Ct) %>% rename(min_ct = Ct)
bac_fit <- bac_fit %>% left_join(min_std) %>%
      mutate(relative_ct = if_else(Ct < min_ct, "outside","inside"))
```

```
## Joining, by = "plate"
```

```
ggplot(bac_std) + geom_point(aes(y = log_conc, x = Ct)) +
      geom_abline(data = coef_df, aes(slope = slope, intercept = intercept), color = "grey80") +
      geom_point(data = bac_fit, aes(y = pred, x = Ct, shape = relative_ct, color = relative_ct)) +
      facet_grid(plate~sam_type) +
      theme_bw() + labs(y = "DNA Concentration log10(ng/ul)",
                        color = "Ct relative to std. curve", shape = "Ct relative to std. curve") +
      theme(legend.position = "bottom")
```

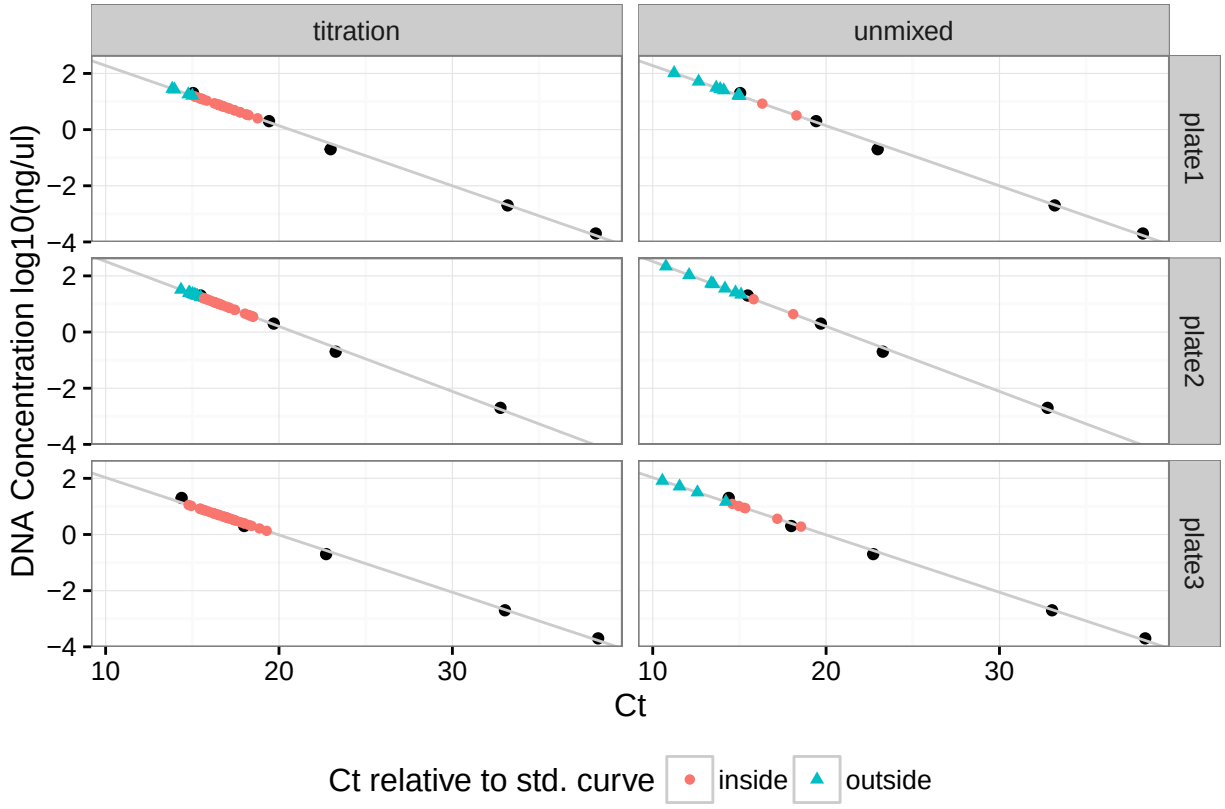


Figure 3: Estimated sample concentrations relative to standard curve and standards (black points).

Only 2 of the 10 unmixed pre and post treatment samples have Ct values within the standard curve for all three plates, and of the remaining 8 have Ct within the standard curve for 2 of the three plates. The linearity of the standard curve outside of the standard concentration range is unknown. Estimated concentrations less than 20 ng/ul (the highest concentration standard) but Ct values outside of the standard curve due to the deviation of the model from the 20 ng/ul standard.

```
bac_fit <- bac_fit %>%
      mutate(quant = 10^pred) %>%
      group_by(sample_name) %>%
      mutate(quant_min = min(quant), quant_max = max(quant)) #%>%
      #ungroup() %>%
      #mutate(sample_name = fct_reorder(sample_name, quant, fun = "median"))
bac_fit %>% filter(sam_type == "unmixed") %>%
      ggplot() +
            geom_hline(aes(yintercept = 20), linetype = 2) +
            geom_linerange(aes(x = sample_name,ymin = quant_min, ymax = quant_max), color = "grey60") +
            geom_point(aes(y = quant, x = sample_name, color = relative_ct, shape = plate)) +
            theme_bw() + labs(x = "Sample Names", y = "DNA concentration (ng/ul)") + coord_flip()
```

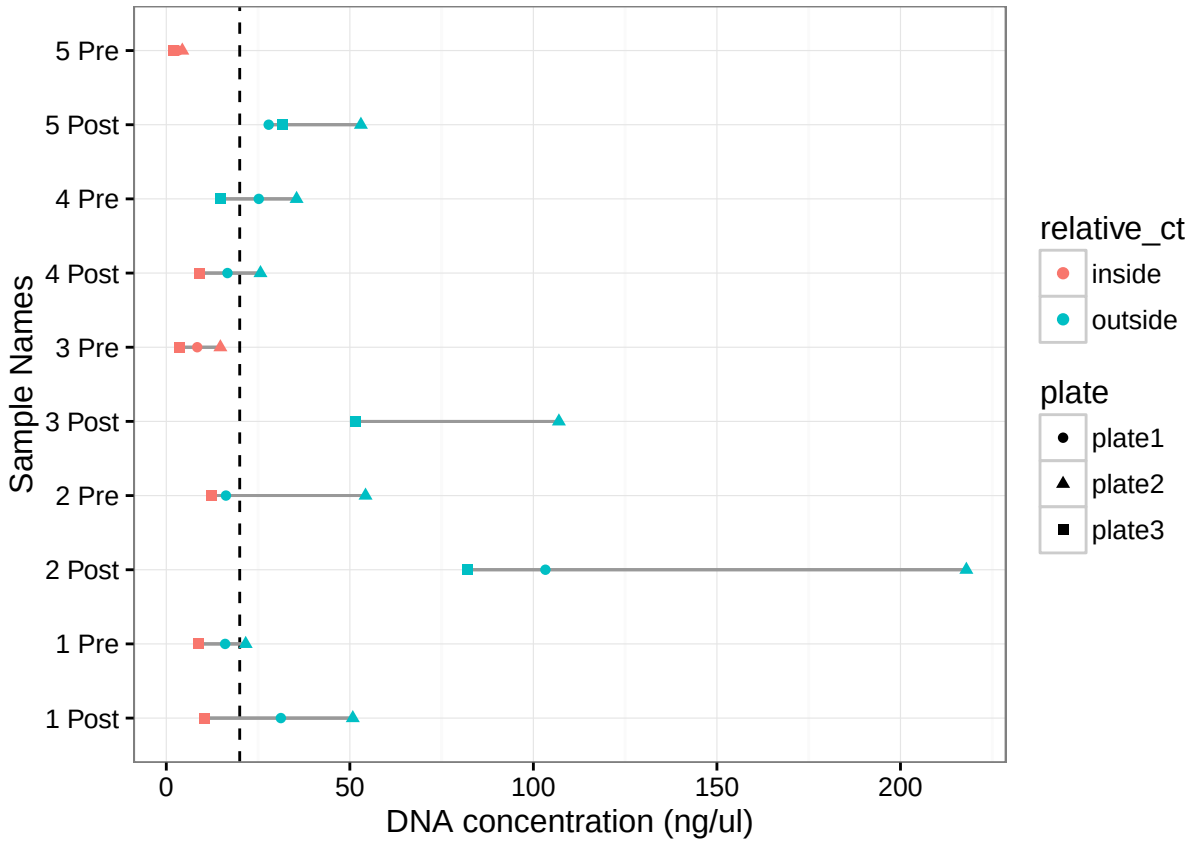
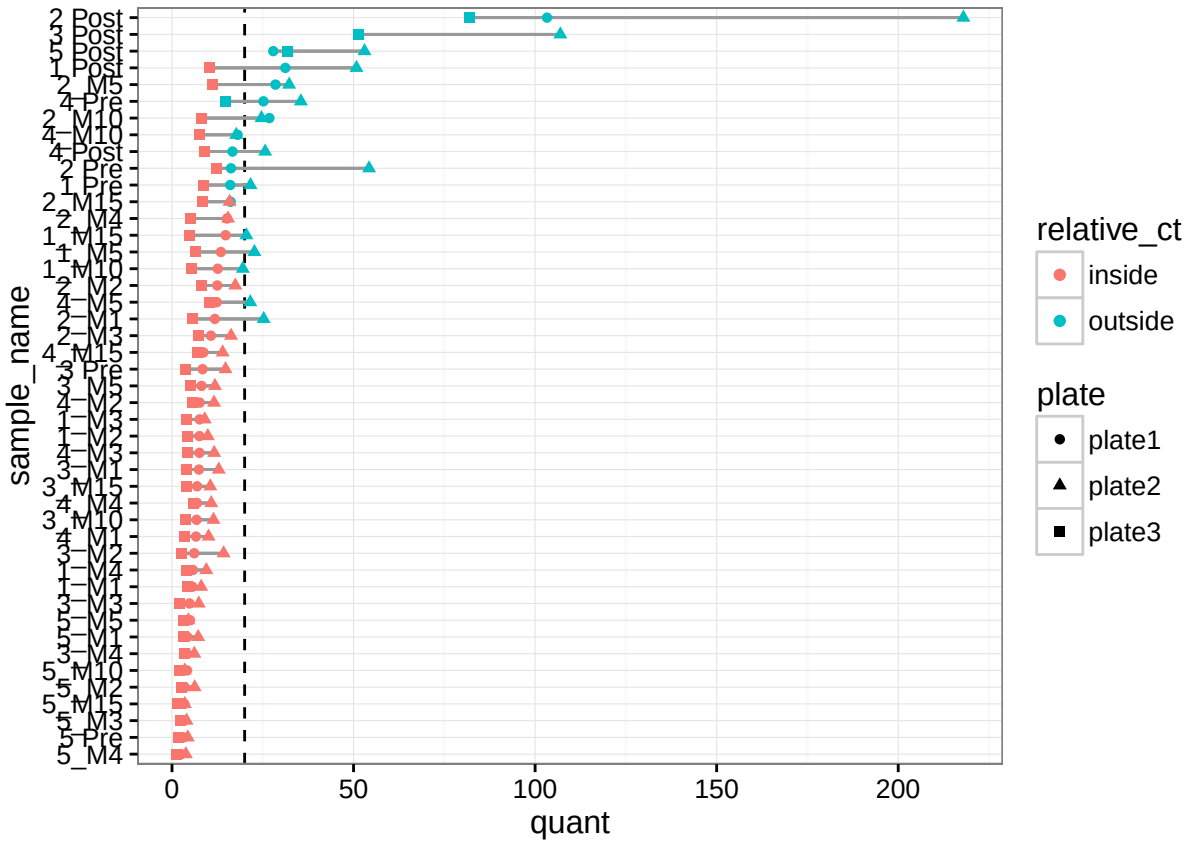


Figure 4: Variability of sample concentration relative to 20 ng/ul standard for unmixed pre and post samples.

The variability in concentration for the unmixed samples, especially post treatment biological replicates 2 ('2 Post') and 3 ('3 Post'), is likely due to the sample concentration falling outside the range of the standard curve. Dashed line is at 20ng/ul, to indicate the highest concentation standard.

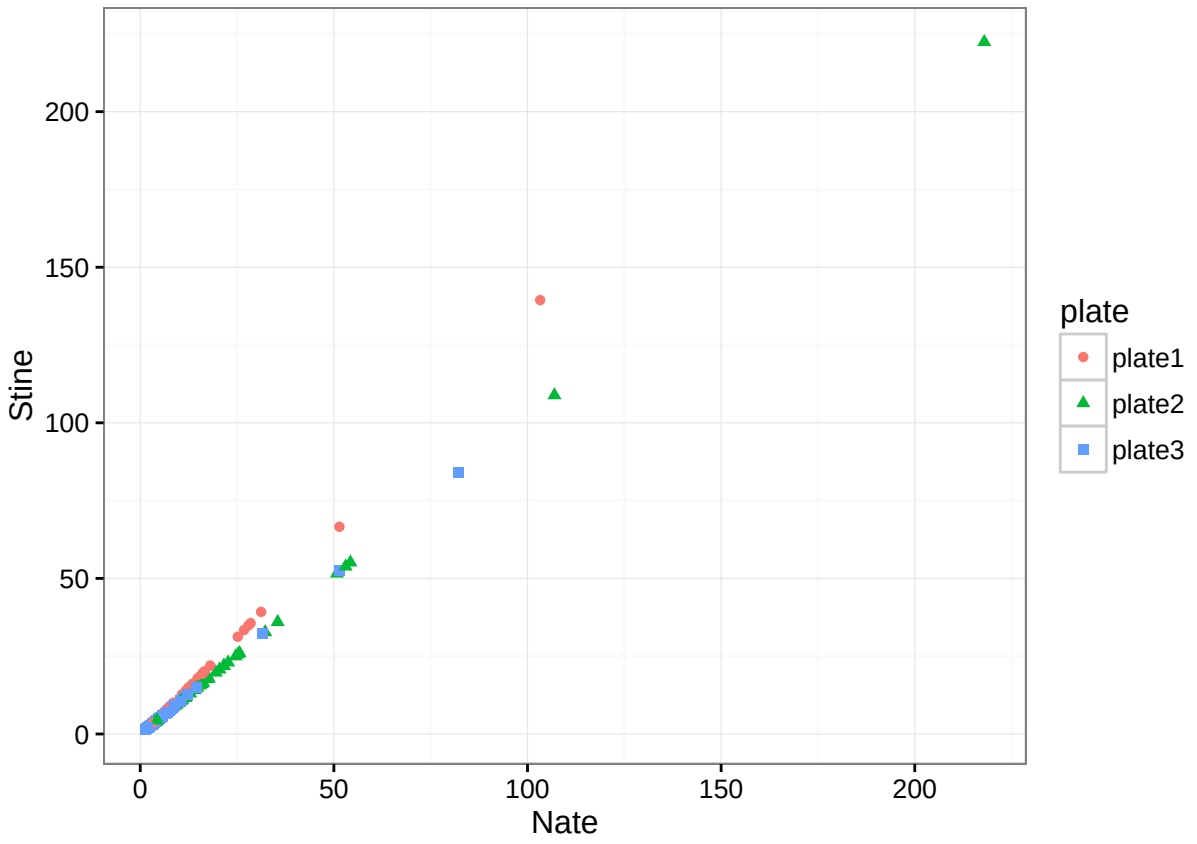
```
bac_fit %>% ungroup() %>%
      mutate(sample_name = fct_reorder(sample_name, quant, fun = "median")) %>%
      ggplot() +
            geom_hline(aes(yintercept = 20), linetype = 2) +
            geom_linerange(aes(x = sample_name,ymin = quant_min, ymax = quant_max),
                           color = "grey60") +
            geom_point(aes(y = quant, x = sample_name, color = relative_ct, shape = plate)) +
            theme_bw() + coord_flip()
```



Comparison of Concentration Estimates

Using methods shown above ("Nate") compared to values provided by Shan and Colin ("Stine").

```
bac_fit %>% ggplot() + geom_point(aes(x = quant, y = stine_quant, color = plate, shape = plate)) + theme
```



Unmixed Sample Concentration Ratios

```
bac_unmixed <- bac_fit %>% ungroup() %>% filter(sam_type == "unmixed") %>%
            separate(sample_name,into = c("bio_rep","titration"),remove = FALSE) %>%
            mutate(titration_factor = if_else(titration == "Pre", 20, 0)) %>%
            select(bio_rep, plate,pred, titration) %>%
            spread(titration,pred)
```

Estimated concentration ratios for pre and post samples (post/pre).

```
bac_unmixed %>% group_by(bio_rep) %>% summarise(con_ratio = 10^median(Post)/10^median(Pre)) %>% kable()
```

bio_rep	con_ratio
1	1.9467031
2	6.3573696
3	6.1095478
4	0.6614112
5	9.9251738

Tritration Concentration Deviation from Predicted Value

Calculating expected bacterial DNA concentration for titrations using estimated unmixed sample concentrations based on how the titrations were generated.

```
bac_mixed <- bac_fit %>% ungroup() %>% filter(sam_type == "titration") %>%
      separate(sample_name, into = c("bio_rep","titration"), remove = FALSE) %>%
      mutate(titration_factor = str_replace(titration, "M","") %>% as.numeric())
```

```
bac_fit2 <- left_join(bac_mixed, bac_unmixed) %>% select(-sample_name, -Ct, -min_ct, -sam_type) %>%
      mutate(exp_pred = Post*(2^-titration_factor) + Pre*(1-2^-titration_factor))
```

```
## Joining, by = c("bio_rep", "plate")
```

```
bac_fit2 %>% ggplot(aes(x = 10^exp_pred, y = 10^pred, color = plate, shape = plate)) +
      geom_point() +
      geom_abline(aes(intercept = 0, slope = 1), color = "grey60",linetype = 2) +
      facet_grid(~bio_rep, scales = "free") +
      labs(x = "Expected (ng/ul)", y = "Estimated (ng/ul)") + theme_bw()
```

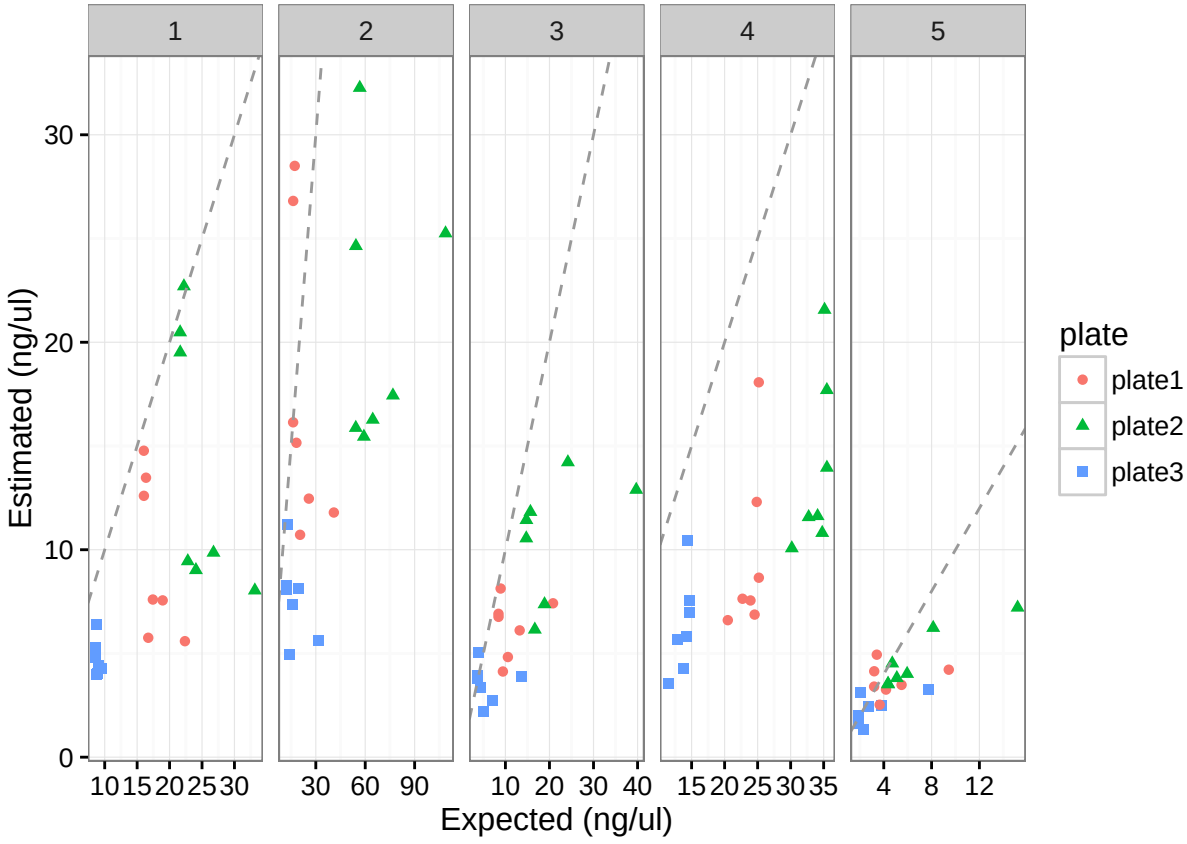


Figure 5: Correlation between predicted and estimated titration bacterial DNA concentrations. Predicted Concentration - calculation based on concentration measurements for pre and post unmixed samples. Estimated Concentration - calculated using standard curve. Dashed line indicates the expected 1:1 relationship.

Conclusions

- High R^2 for standard curve, might want to consider only fitting to three highest concentration standard due to lower residual standard error.
- High estimated DNA concentration for post treatment unmixed samples due to Ct values outside of the standard curve resulted in high post to pre treatment sample concentration ratios.

- Should consider diluting and rerunning samples with multiple replicates outside of the standard curve to verify concentration estimates.

Caveats

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