BaitCap Analysis

Jessica Rowell

9/8/2021

Background on the problem

Objective: Evaluate how well a laboratory method captures Shiga toxin-producing *E. coli* (STEC) from complex samples that also contain commensal (non-pathogenic) *E. coli* strains. Pathogenic and commensal *E. coli* are very similar and detecting STEC when it exists in along with commensal *E. coli* is an important problem for outbreak detection and surveillance. Bait capture method: My team members would like to evaluate a laboratory method called "bait capture" in comparison to a gold standard (shotgun sequencing). Briefly, bait capture involves first designing "baits" that bind to specific pieces of DNA. The desired DNA targets are captured, amplified, and then sequenced. The result is a set of sequences that are enriched for the DNA targets of interest.

Shotgun method: Shotgun sequencing involves chopping up long strands of DNA into small fragments and then sequencing all these fragments. In a sample with many different species, the small fragment sequences (called "short reads") get all mixed up and we have to use computational techniques (like binning) to assemble the genomes and determine what species were in the original sample. Theoretically, this is an unbiased technique because everything gets sequenced equally (no enrichment of certain targets).

We are comparing the bait capture method against this method to determine how good bait capture is at capturing STEC in samples containing STEC and commensal *E. coli* strains.

Replicates: The baitcap vs. shotgun comparison was made in triplicate. The need for this analysis (i.e. this entire problem statement) stems from how the replicates were made. Each replicate was sourced from a separate solution containing STEC and commensal *E. coli* in a specified ratio. For example, if the specified ratio is 1:1, then a lab technician made 3 separate solutions of 1:1. Because the quantities are very small, there is substantial room for measurement error. We no longer have a known starting ratio for comparison. That's why we need a method that compares baitcap to shotgun per replicate.

This contrasts with the other way this experiment has been done. That time, 1 main solution was made containing STEC and commensal in the prespecified ratio. 3 separate samples were taken from that solution, and then baitcap and shotgun were run on each of those 3 samples. In this scenario, the ratio in the samples taken from the main solution should follow the normal distribution (I wouldn't expect it to be exactly the prespecified ratio every time), with a small variance. In that case, there isn't much need for a special approach. Since it was done differently this time, we must do the analysis differently to take into account that we don't have a reliable, known ratio for each of our 3 replicates.

Exploratory analysis

Taking a look at the data

Data tables

Input data for graphs and models

type = C2C1, comparison = C2 means this row compares C2 to C1
rep = replicate number
z = logY(baitcap) - logY(shotgun), where Y = reads baitcap / reads shotgun

Table 1: Table 1. Input data for graphs and models

type	comparison	rep	z
C2C1	C2	1	-0.6899549
C2C1	C2	2	-0.3926774
C2C1	C2	3	-0.1783091
FC1	F	1	0.2034167
FC1	F	2	0.2547066
FC1	F	3	0.5885684
KC1	K	1	0.1115875
KC1	K	2	0.1851309
KC1	K	3	-0.0719776
NC1	N	1	-0.8062599
NC1	N	2	-0.4437322
NC1	N	3	-0.1707488

Raw input data that can be used to check the numbers for Y, logY, and z above.

type = group indicator (C2C1 indicates that’s a C1 or C2 comparison row)
comparison = this couples with “type”; type = C2C1 means it’s a C2 vs. C1, and comparison indicates which row it is
bc = indicates baitcap (1) vs. shotgun (0)
rep = replicate number
outcome = number of reads

Table 2: Table 2. Raw input data

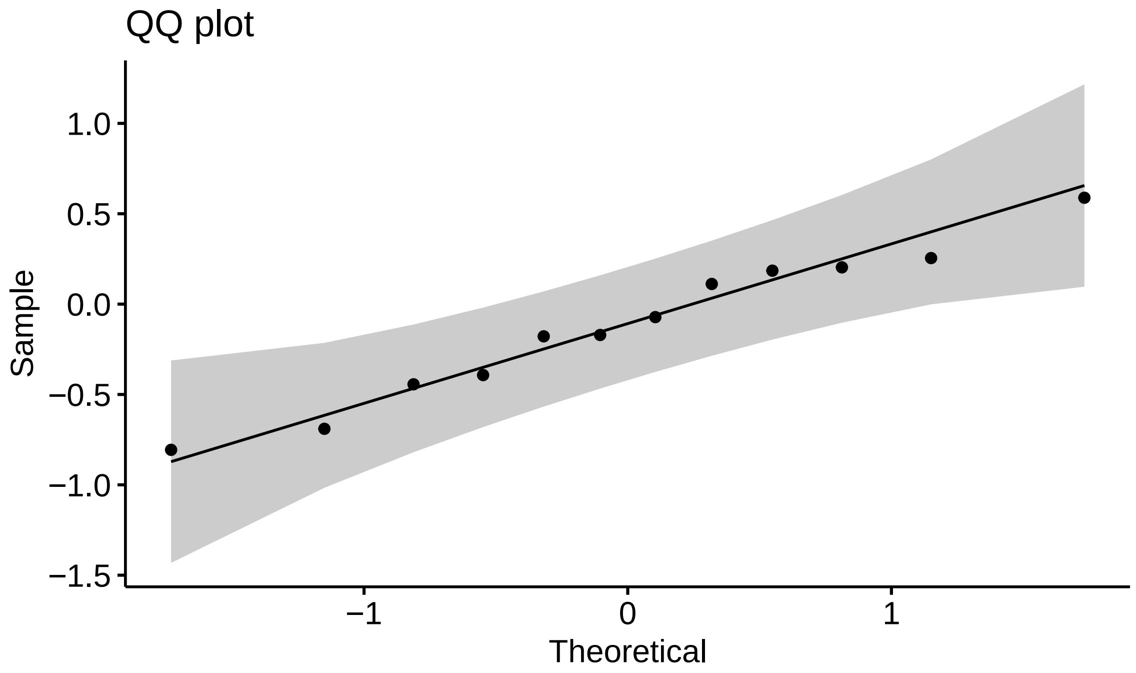
type	comparison	bc	rep	outcome	y	logY
C2C1	C1	0	1	436478	NA	NA
C2C1	C2	0	1	1234094	2.8273911	1.0393544
C2C1	C1	1	1	749924	NA	NA
C2C1	C2	1	1	1063554	1.4182157	0.3493995
C2C1	C1	0	2	726636	NA	NA
C2C1	C2	0	2	1716316	2.3620024	0.8595097
C2C1	C1	1	2	1040912	NA	NA
C2C1	C2	1	2	1660186	1.5949341	0.4668324
C2C1	C1	0	3	472832	NA	NA
C2C1	C2	0	3	1048416	2.2173119	0.7962956
C2C1	C1	1	3	1270758	NA	NA
C2C1	C2	1	3	2357496	1.8551888	0.6179865
FC1	C1	0	1	498234	NA	NA
FC1	F	0	1	774642	1.5547755	0.4413311
FC1	C1	1	1	814320	NA	NA
FC1	F	1	1	1551692	1.9055064	0.6447478
FC1	C1	0	2	609786	NA	NA
FC1	F	0	2	664352	1.0894839	0.0857041
FC1	C1	1	2	1557814	NA	NA
FC1	F	1	2	2189546	1.4055247	0.3404107
FC1	C1	0	3	647186	NA	NA

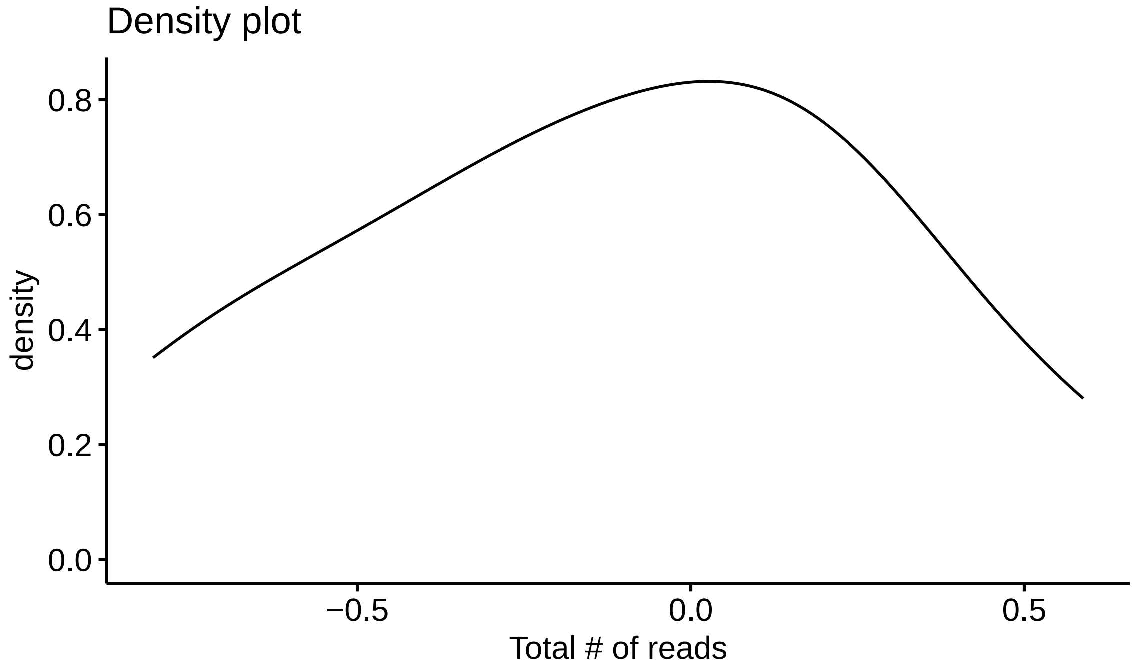
type	comparison	bc	rep	outcome	y	logY
FC1	F	0	3	395658	0.6113513	-0.4920835
FC1	C1	1	3	1401322	NA	NA
FC1	F	1	3	1543266	1.1012929	0.0964849
KC1	C1	0	1	542488	NA	NA
KC1	K	0	1	363722	0.6704701	-0.3997761
KC1	C1	1	1	947848	NA	NA
KC1	K	1	1	710526	0.7496202	-0.2881886
KC1	C1	0	2	513918	NA	NA
KC1	K	0	2	799854	1.5563845	0.4423655
KC1	C1	1	2	1198652	NA	NA
KC1	K	1	2	2244974	1.8729156	0.6274963
KC1	C1	0	3	584424	NA	NA
KC1	K	0	3	883666	1.5120289	0.4134524
KC1	C1	1	3	1460214	NA	NA
KC1	K	1	3	2054552	1.4070212	0.3414748
NC1	C1	0	1	514756	NA	NA
NC1	N	0	1	870196	1.6905019	0.5250255
NC1	C1	1	1	1174822	NA	NA
NC1	N	1	1	886816	0.7548514	-0.2812344
NC1	C1	0	2	631414	NA	NA
NC1	N	0	2	1330742	2.1075586	0.7455302
NC1	C1	1	2	871974	NA	NA
NC1	N	1	2	1179160	1.3522880	0.3017980
NC1	C1	0	3	422726	NA	NA
NC1	N	0	3	1294646	3.0626127	1.1192684
NC1	C1	1	3	546734	NA	NA
NC1	N	1	3	1411604	2.5818844	0.9485195

Our outcome variable of interest is the ratio of STEC reads to commensal reads. We want to compare this value for baitcap vs. shotgun sequencing. We want to account for the effect, if any, of replicate number, given the wet lab methodology.

First we visually check whether our outcome variable is normally distributed in our dataset. These look reasonable, especially considering the small number of observations (n = 12). For the QQ plot, the points should fall within the gray area, preferably as close to the line as possible. For the density plot, we're looking for a nice "normal looking" curve, without much skew or bumps.

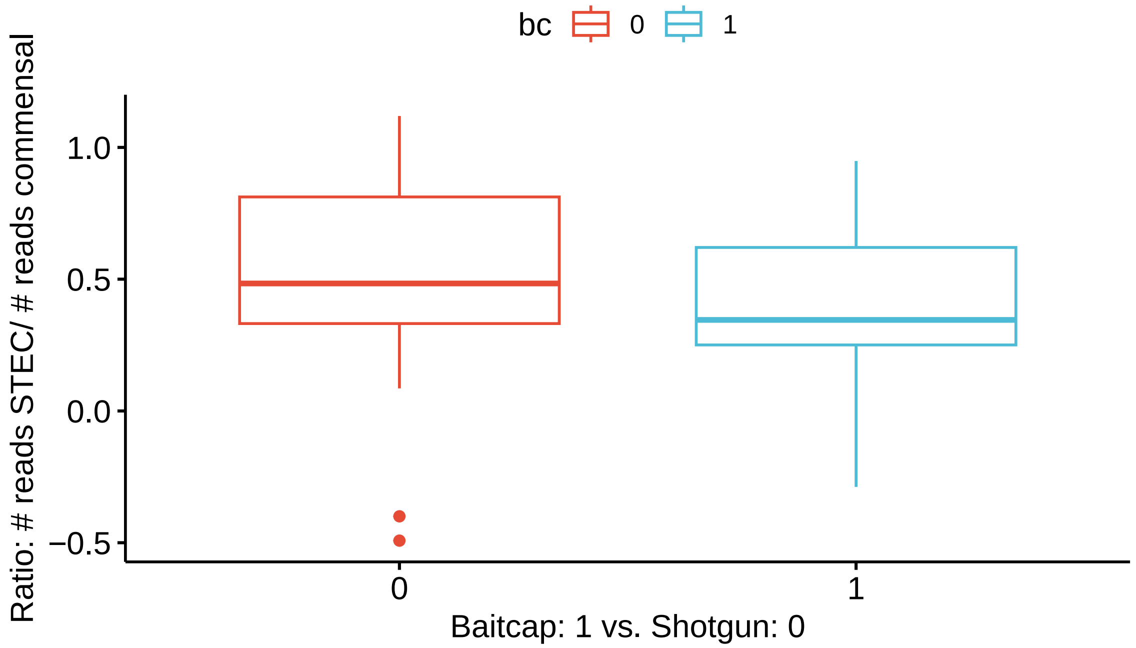
```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



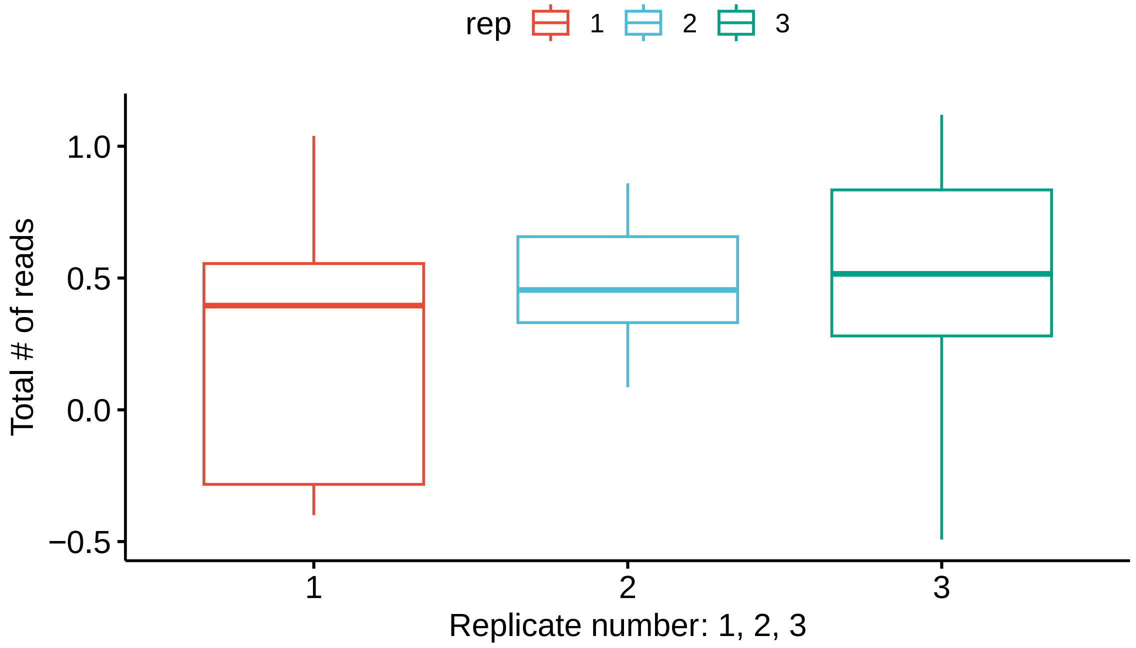


We may want to satisfy our curiosity by taking a quick look at how our outcome of interest compares for baitcap vs. shotgun. Here our outcome of interest is the ratio of total number of reads for STEC vs. commensal. In the plots below, the bold line in the box is the median and the box represents the interquartile range. So the middle half of the data resides within the box. The lines extend to cover 99.3% of the data. Any dots outside the lines are outliers. Very long boxes indicate a large variance, and low precision. For our small number of observations, a larger variance is not unexpected.

Here we observe that baitcap seems to capture slightly more STEC than commensal, compared to shotgun, but there is a lot of overlap in the two distributions.



Now we visually check the distribution of that same ratio (# reads STEC vs. # reads commensal) across replicate number. While the medians are consistent across replicates, there is more variability in replicates 1 and 3 and the distributions look different.



First linear regression analysis: assume 12 independent samples

In this first analysis, we will treat the replicates as biological samples. So, we have 4 different biological sample comparisons: C2 vs. C1, F vs. C1, K vs. C1, N vs. C1. For each of those, there are 3 replicates.

Let's call our outcome "z". Z is the **difference** of the ratio of STEC-to-commensal for baitcap vs. shotgun. For analysis the ratios are log-transformed (as you see below).

$$y = \frac{reads_{STEC}}{reads_{commensal}}$$

$$z = \log(y_{baitcap}) - \log(y_{shotgun})$$

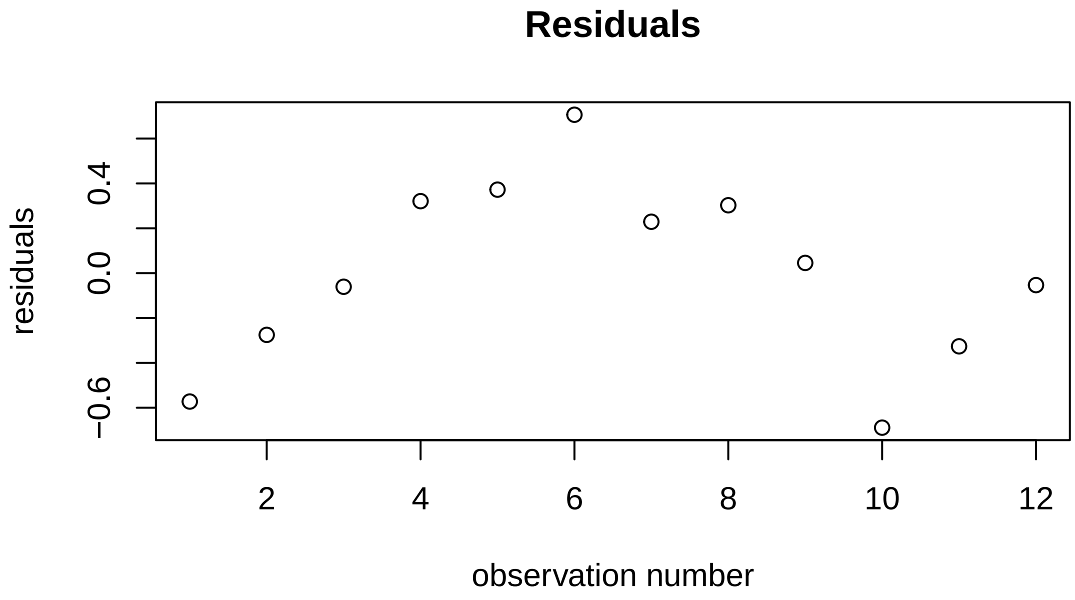
If we first consider each replicate as a separate, independent sample, we can just observe the average of z (μ_z) in our whole dataset. This is equivalent to saying we did 12 separate tests of baitcap vs. shotgun, and we want to see what's our best estimate of "z".

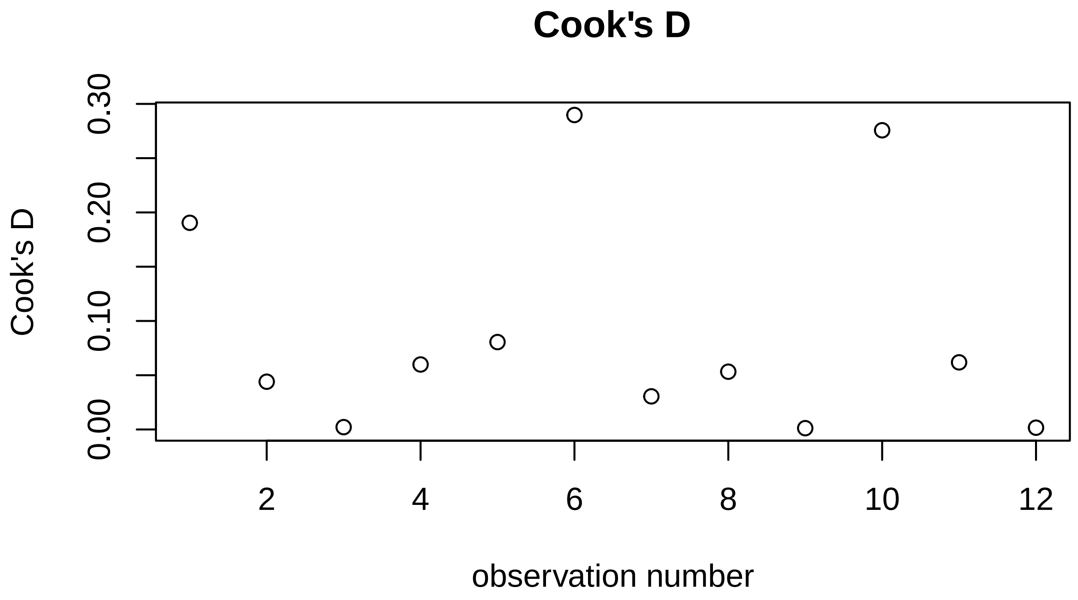
The line we are fitting looks like this:

$$Z_i = \beta_0 + \epsilon, i = 1, 2, ..., 12$$

Some details to note: the residuals are not randomly distributed. Observations 6 and 10 could be influential. These two observations correspond to F vs. C1 (replicate 3) and N vs. C1 (replicate 1).

```
##
## Call:
## lm(formula = z ~ 1, data = d4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68874 -0.28792 -0.00384  0.30722  0.70609
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1175     0.1192  -0.986    0.346
##
## Residual standard error: 0.4131 on 11 degrees of freedom
```





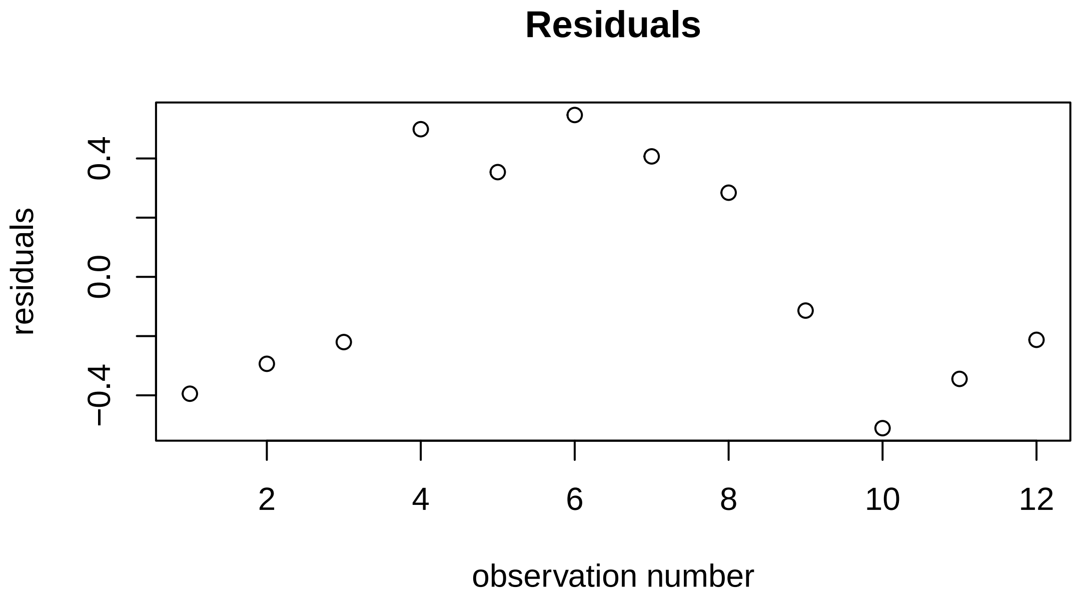
```
## [1] "Inverse ln(intercept): 0.889122002617685"
```

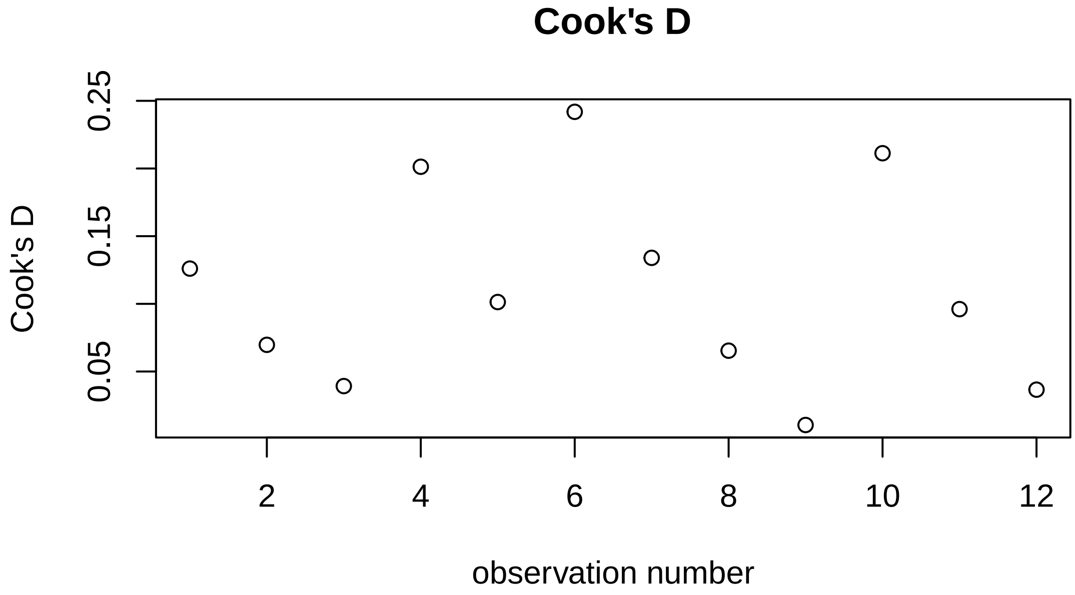
Our estimate of μ_z is 0.89. If we treat all replicates as independent biological samples, then our estimate is that the ratio of STEC to commensal *E. coli* is 0.88 higher for baitcap than for shotgun. This is not statistically significant. It also considers all 12 biological samples the same (and I am questioning whether that is correct, especially because I think C2 is also commensal).

Second linear regression analysis: assess the importance of replicate

Now we will introduce a variable for the effect of replicate on Z. The results of the linear regression suggest that replicate number does not significantly predict whether baitcap performs better than shotgun for detecting STEC vs. commensal *E. coli*.

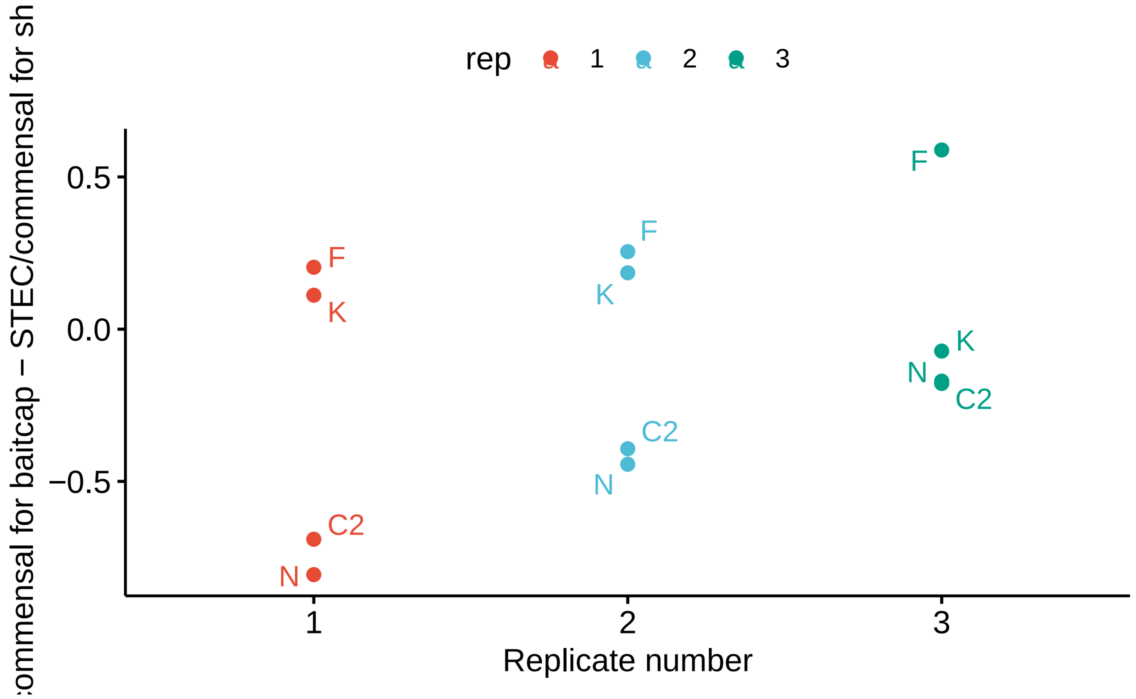
```
##
## Call:
## lm(formula = z ~ as.factor(rep), data = d4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5110 -0.3063 -0.1633  0.3671  0.5467
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.2953     0.2139  -1.380    0.201
## as.factor(rep)2   0.1962     0.3025   0.648    0.533
## as.factor(rep)3   0.3372     0.3025   1.115    0.294
##
## Residual standard error: 0.4279 on 9 degrees of freedom
## Multiple R-squared:  0.1222, Adjusted R-squared:  -0.07283
## F-statistic: 0.6266 on 2 and 9 DF,  p-value: 0.5562
```





Summary of linear regression analyses

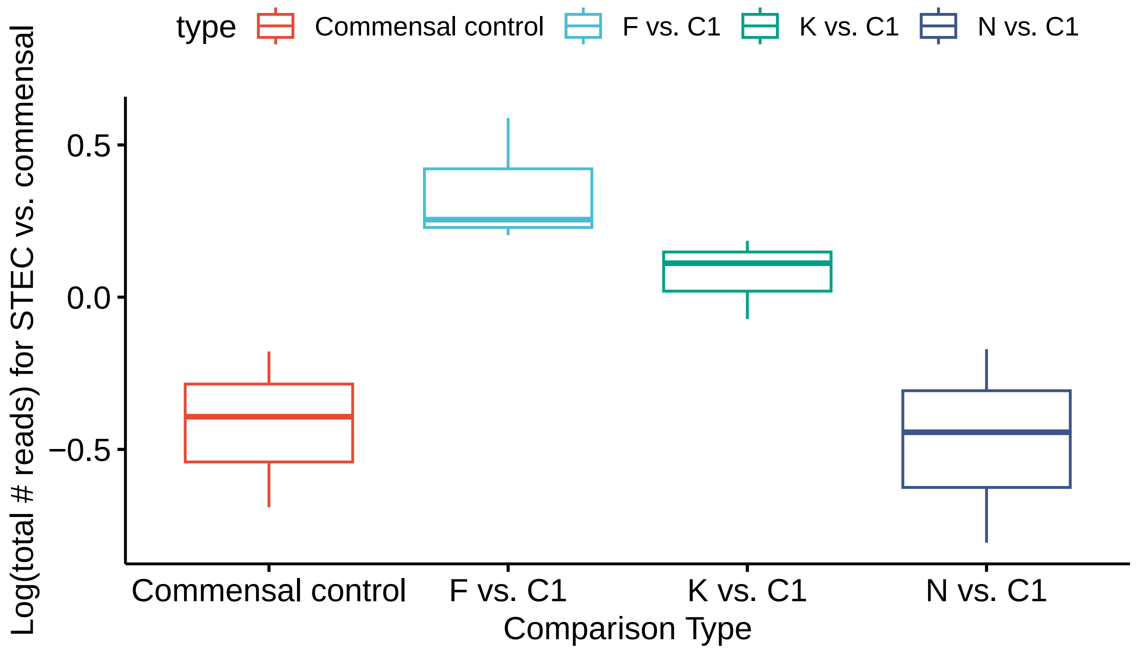
In the figure below, we can see the following minor trend: baitcap performs a little better than shotgun for K vs. C1 and F vs. C1, but a little worse for N vs. C1 and C2 vs. C1. I'd say there is also a trend towards an effect of replicate number, and I would probably change the wet lab methodology if it's not costly in time or other resources. With larger samples, independent replicates might become a problem. It would be interesting to repeat this analysis with replicates drawn from a single mixture, and compare the results with what we observe here.



Analysis of pilot baitcap study data

The figure below shows a comparison among different types: C2 vs. C1, F vs. C1, K vs. C1, and N vs. C1. Here our outcome of interest is the ratio of total number of reads for one type (C2, F, K, or N) vs. commensal (C1). In the plots below, the bold line in the box is the median and the box represents the interquartile range. So the middle half of the data resides within the box. The lines extend to cover 99.3% of the data. Any dots outside the lines are outliers. Very long boxes indicate a large variance, and low precision. For our small number of observations, a larger variance is not unexpected.

Here we observe that baitcap seems to capture more of its target STEC (F and K types) than commensal, compared to shotgun, but only when used on STEC for which the baits have been designed. The nonspecific target STEC, N, performed the same as the control group (C2 vs. C1).



log(STEC/commensal) - BC log(STEC/commensal) - SS

outcome: log[(STEC/commensal)-BC / (STEC/commensal)-SS] **same as: log(STEC/commensal-BC) - log(STEC/commensal-SS)**

Comparing target STEC vs. controls

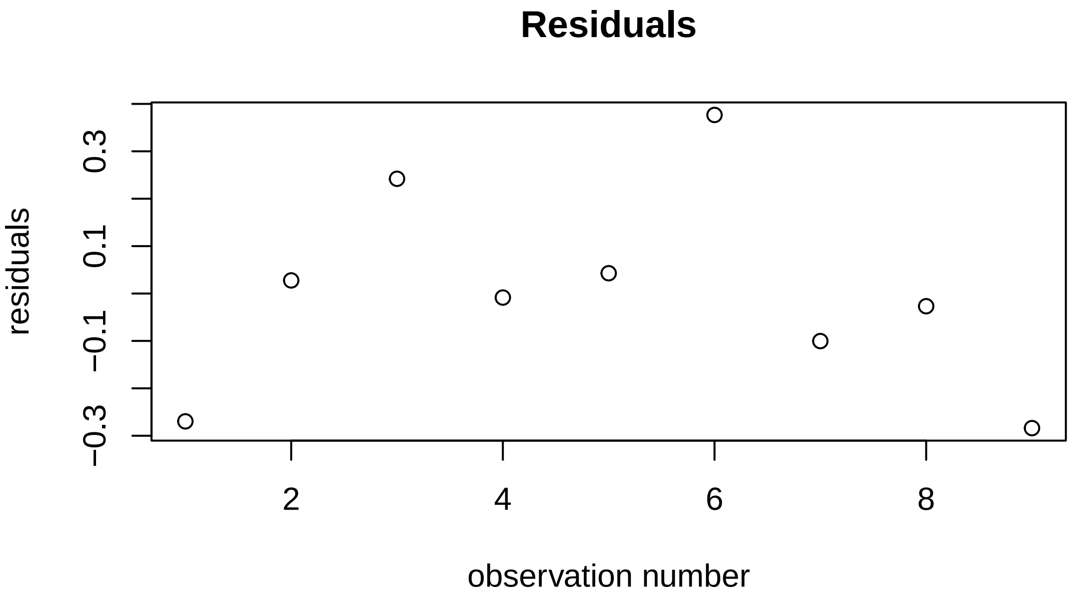
Now I have collapsed the target STEC types (F and K) together and removed the non-target STEC group (N) to compare baitcap vs. shotgun performance in target STEC vs control.
So now we are comparing two groups: one control group of 3 replicates (C2 vs. C1) and one experimental group comprising two sets of 3 replicates.

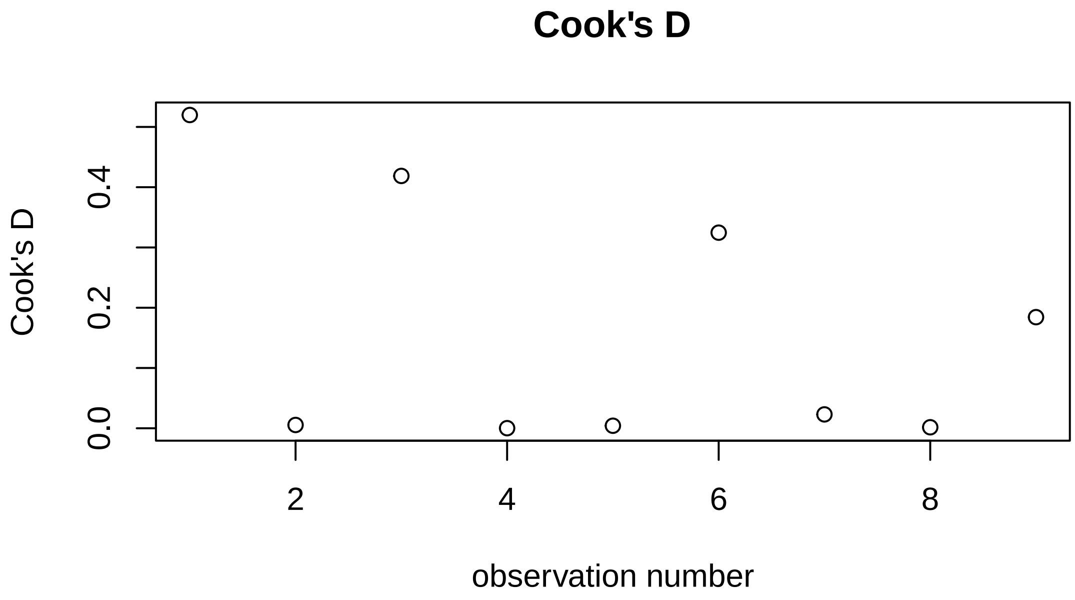
The model suggests that the ratio of reads captured by baitcap vs. shotgun method is 88% higher for target STEC vs. the commensal control. It also suggests that this value is statistically significant (p = 0.006). However, a quick evaluation of the diagnostic plots for the model show that our normality assumptions are not valid. There are some outliers in our residuals (top left plot) and the QQ plot deviates dramatically from a straight line (top right plot). We have so little data (n = 9) that it's hard to say what we see is largely due to sample size. That said, I would **not** use the results of this linear model in a paper or presentation. I would just show the graphs and say we can see "trends" of an effect but cite the limited sample size in this small pilot study.

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

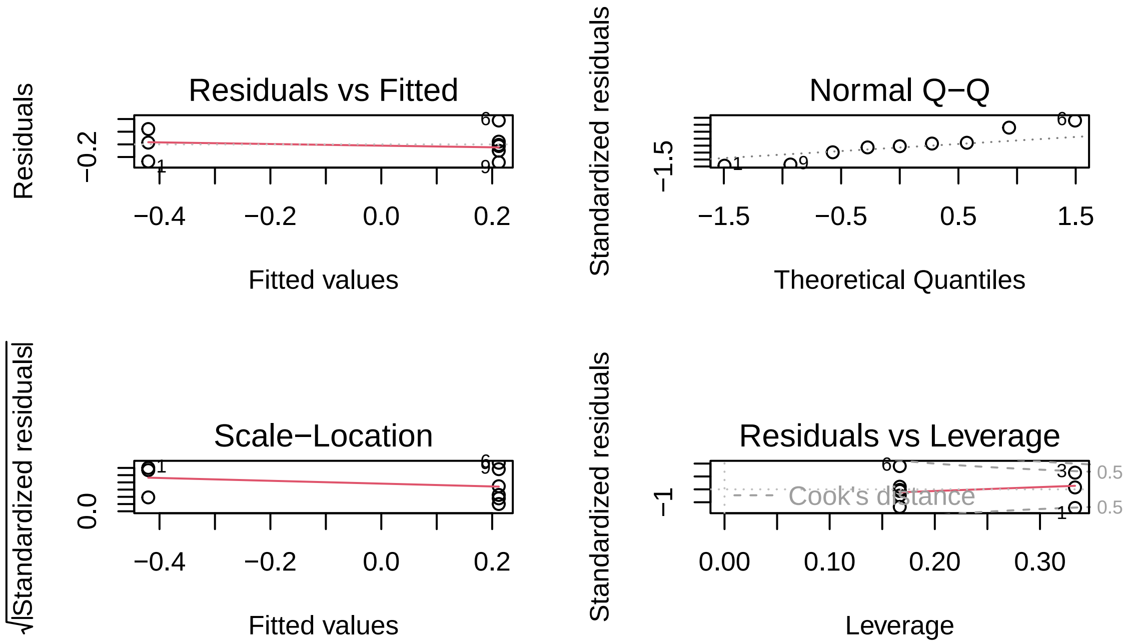
```
##
## Call:
## lm(formula = z ~ as.factor(type), data = d_sub)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28388 -0.10032 -0.00849  0.04280  0.37666
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.4203     0.1322  -3.179  0.01552 *
## as.factor(type)FKC1    0.6322     0.1619   3.904  0.00587 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.229 on 7 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6403
## F-statistic: 15.24 on 1 and 7 DF,  p-value: 0.00587
```

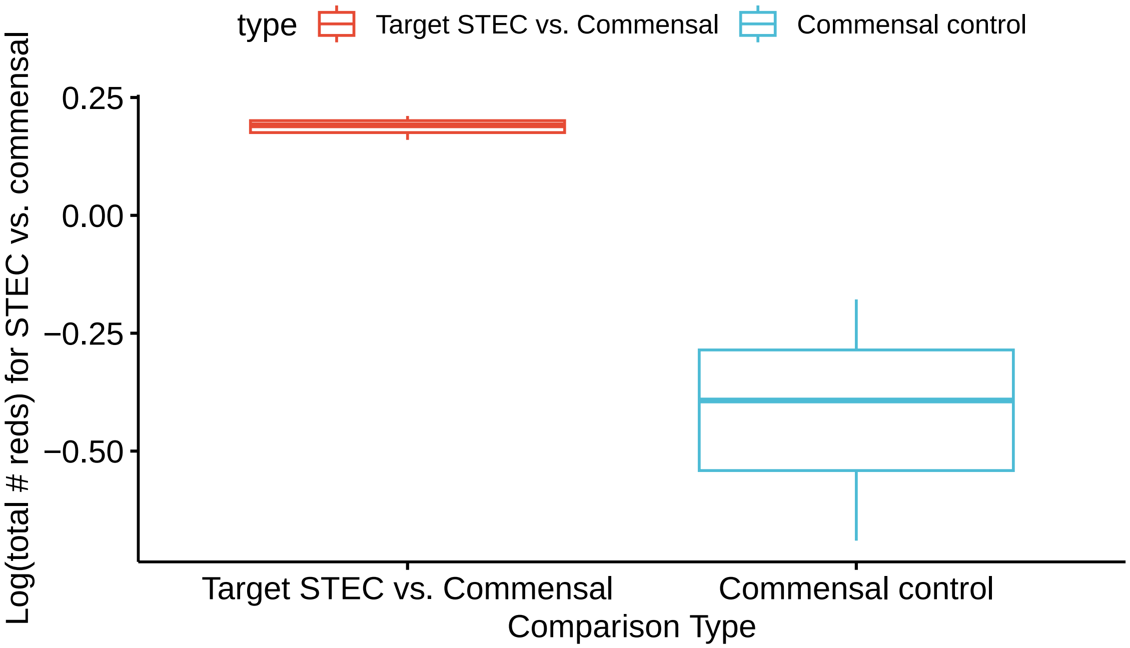




```
##         (Intercept) as.factor(type)FKC1
##           0.6568407           1.8817820
```



The boxplot figure below


type
Target STEC vs. Commensal
Commensal control
Log(total # reds) for STEC vs. commensal
0.25
0.00
−0.25
−0.50
Target STEC vs. Commensal
Commensal control
Comparison Type