Statistics for data from Fig 2h Caillaud *et al* PLoS Biology

Dan MacLean

13 August 2015

Pre-processing

Marie-Cecille Caillaud (MCC) sent me an Excel file of all the haemocytometry measurements she made - file `raw/MCC-Dan corrected.xslx`. The file annotates figures from the same biological replicates as colours which I can't parse programmatically. I therefore added columns to the sheet stating the replicate number. I also removed spaces in column headers and saved the file as `raw/MCC-Dan corrected Reps added.xlsx` and exported the sheet with the data to a csv file `fig_2h_data_manual.csv` which I can operate on programmatically and will use as my input.

Use some Python to get the data file into better shape

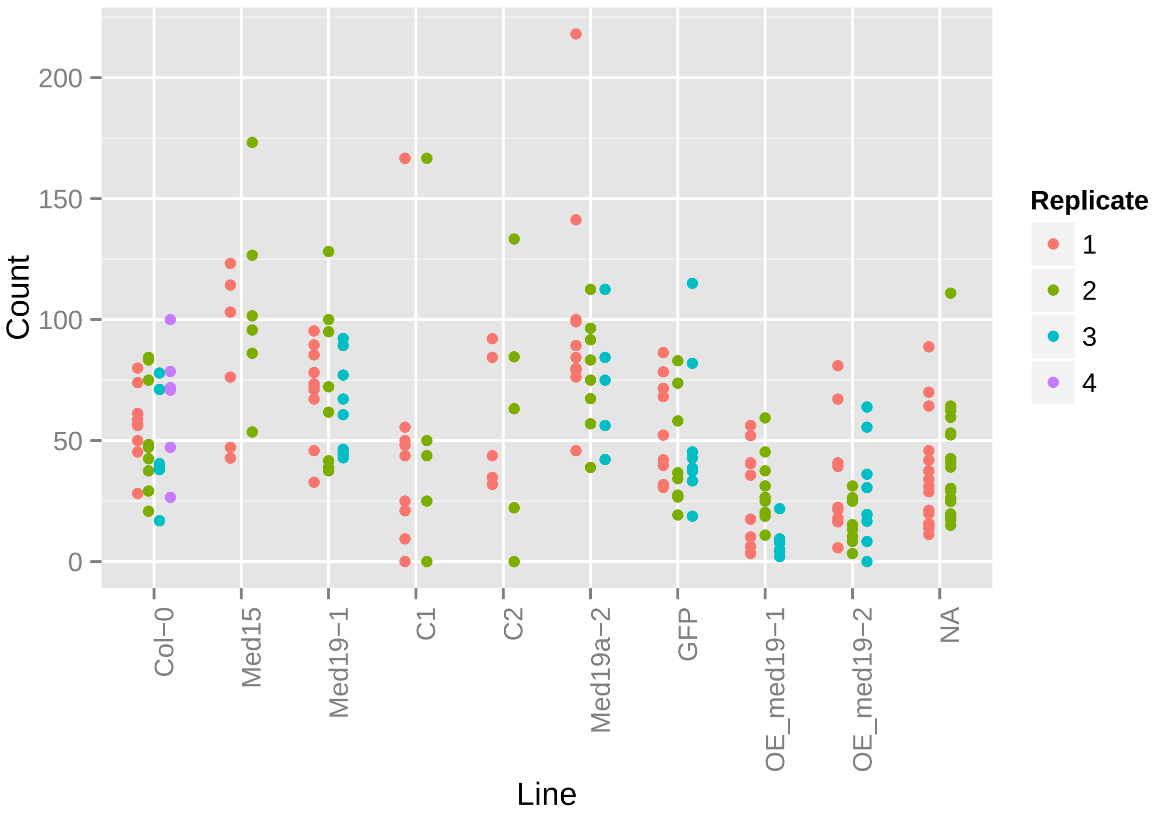
```
header = []
results = []
with open('raw/fig_2h_data_manual.csv', 'r') as file:
  for l in file:
    l = l.rstrip('\r\n')
    a = l.split(',')
    if l.startswith("Rep"):
      header = a
    else:
      for i in range(0,len(header),2):
        rep,line,count = a[i],header[i+1],a[i+1]
        if rep and line and count: ## if we have no empty values
          results.append([rep,line,count])

with open('data/reshaped_data.csv','w') as outfile:
  outfile.write("Replicate,Line,Count\n")
  for r in results:
    outfile.write(",".join(r) + "\n")
```

Load data, reorder as per Figure 2H and do a straightforward plot

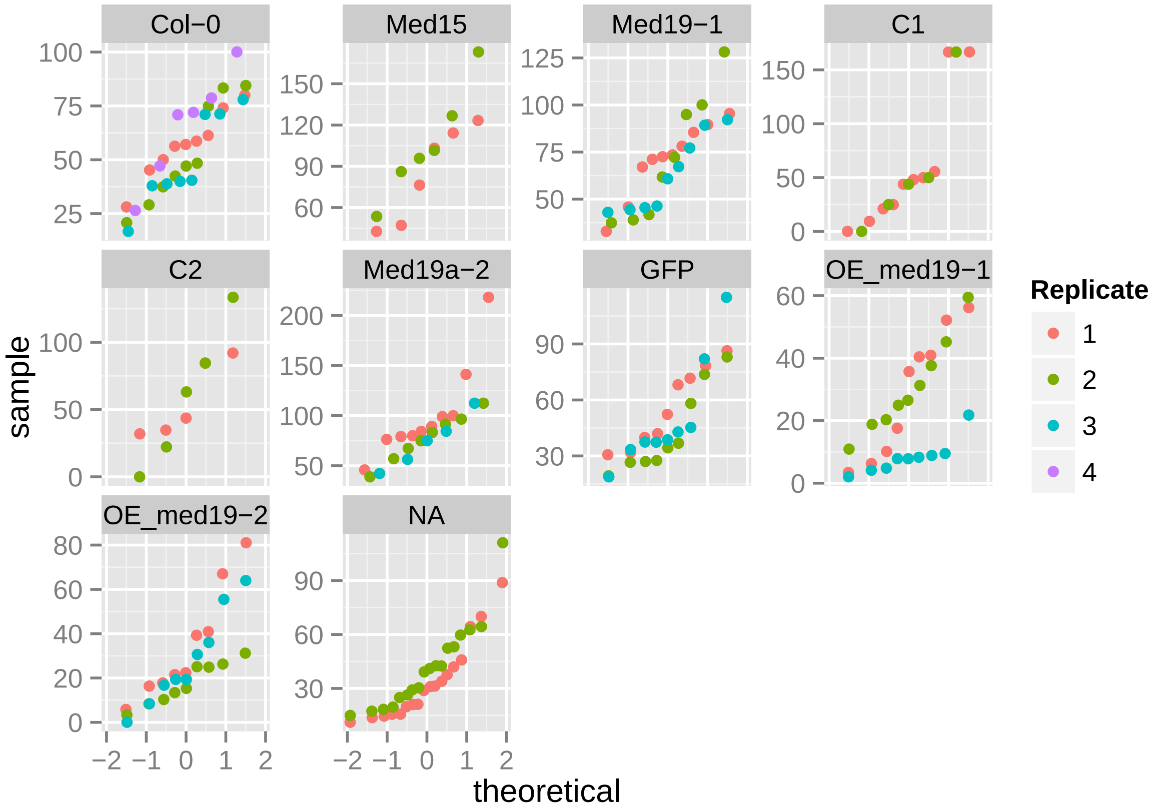
```
library(ggplot2)
data <- read.csv('data/reshaped_data.csv', header=TRUE)
data$Replicate <- as.factor(data$Replicate)
data$Line <- factor(data$Line, c("Col-0","Med15", "Med19-1","C1","C2","Med19a-2","GFP", "OE_med19-1","O
basic <- ggplot(data, aes(Line,Count))
scatter <- basic + geom_jitter(aes(colour=Replicate),position = position_dodge(width=0.5)) + theme(axis
scatter
```

```
## ymax not defined: adjusting position using y instead
```



The data look ok, a few outliers in `Med19a-2` and `C1` that could affect summary statistics. Let's do some `qqplots` and see how they lie.

```
#qnorm is default distribution - we are testing for a normal distribution
ggplot(data, aes(sample=Count)) + geom_jitter(stat="qq", aes(colour=Replicate) ) + facet_wrap( ~ Line, 
```



Those outliers could mess up summary statistics, they're off the curve, we have no good reason to ditch them though. I suppose they mean that occasionally the method used (spore counting) throws up some very extreme numbers. Overall these plots are ok, the variation seems normally distributed on the whole.

Let's have a look at summary statistics:

```
library(plyr)
summary <- ddply(data,"Line",summarise, mean=mean(Count),median=median(Count),diff=abs(mean(Count) - me
summary
```

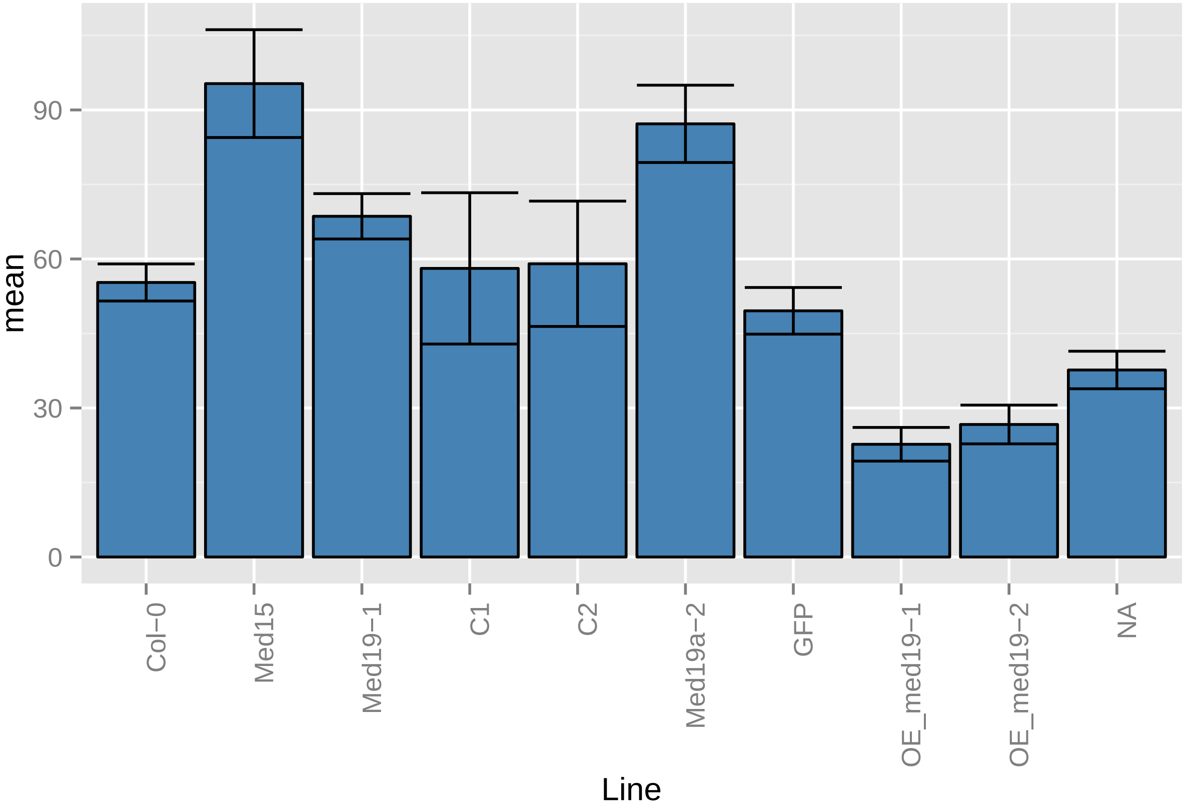
```
##           Line     mean median      diff  std_dev   std_err
## 1        Col-0 55.26875 53.125  2.143750 21.11359  3.732392
## 2        Med15 95.30083 98.630  3.329167 37.59931 10.853987
## 3      Med19-1 68.58889 71.140  2.551111 23.73107  4.567046
## 4           C1 58.10267 43.750 14.352667 58.98184 15.229045
## 5           C2 59.03000 53.455  5.575000 39.89419 12.615651
## 6     Med19a-2 87.20130 83.330  3.871304 37.32340  7.782467
## 7          GFP 49.56074 39.770  9.790741 24.40441  4.696630
## 8   OE_med19-1 22.69333 18.750  3.943333 17.61271  3.389568
## 9   OE_med19-2 26.68185 21.430  5.251852 20.25470  3.898019
## 10        <NA> 37.64722 31.115  6.532222 22.67504  3.779174
```

The summary stats seem fine overall, similar SD and SE and not much drift of the median from the mean, the concern again is Med19a-2 and C1 with the high standard deviation and mean dragged up by that couple of points.

Does a bar chart imply a higher effect than we see generally?

Let's make a bar graph with error bars on that first scatter to see how using a standard bar chart might be misleading our thinking.

```
ggplot(summary, aes(x=Line, y=mean)) + geom_bar(position=position_dodge(), stat="identity", fill="steel
```



The barchart is definitely suggesting a higher overall effect than we see from the individual replicates in the scatter plot for `Med19a-2` and `C1` My conclusion here is that although the mean is calculated correctly, it's just that the mean is a slightly misleading number to boil our data down to in this case. Also that very slight increase in standard error isn't giving us a clue as to that messy single outlier. Taken together the mean and SE plotted like this convince of us a bigger effect in general so the plot style isn't helpful.

Significance Tests

I need to boil down the data to the biological replicates.

```
library(reshape)
```

```
##
## Attaching package: 'reshape'
##
## The following objects are masked from 'package:plyr':
##
##     rename, round_any
```

```
bioreps <- cast(data, Line~Replicate, mean)
```

```
## Using Count as value column.  Use the value argument to cast to override this choice
```

```
bioreps <- melt(bioreps)
bioreps
```

```
##             Line      value Replicate
## X1        Col-0  56.742222         1
## X1.1      Med15  84.476667         1
## X1.2    Med19-1  71.135000         1
## X1.3         C1  58.612000         1
## X1.4         C2  57.394000         1
## X1.5   Med19a-2 101.320000         1
## X1.6        GFP  55.681111         1
## X1.7 OE_med19-1  29.183333         1
## X1.8 OE_med19-2  34.675556         1
## X2        Col-0  52.041111         2
## X2.1      Med15 106.125000         2
## X2.2    Med19-1  71.896250         2
## X2.3         C1  57.084000         2
## X2.4         C2  60.666000         2
## X2.5   Med19a-2  77.763750         2
## X2.6        GFP  42.908889         2
## X2.7 OE_med19-1  30.555556         2
## X2.8 OE_med19-2  17.592222         2
## X3        Col-0  49.311250         3
## X3.1      Med15        NaN         3
## X3.2    Med19-1  62.820000         3
## X3.3         C1        NaN         3
## X3.4         C2        NaN         3
## X3.5   Med19a-2  74.064000         3
## X3.6        GFP  50.092222         3
## X3.7 OE_med19-1   8.341111         3
## X3.8 OE_med19-2  27.777778         3
## X4        Col-0  65.843333         4
## X4.1      Med15        NaN         4
## X4.2    Med19-1        NaN         4
## X4.3         C1        NaN         4
## X4.4         C2        NaN         4
## X4.5   Med19a-2        NaN         4
## X4.6        GFP        NaN         4
## X4.7 OE_med19-1        NaN         4
## X4.8 OE_med19-2        NaN         4
```

I'll do an ANOVA and Tukey's HSD for multiple comparisons.

```
### ANOVA and Tukey's HSD on all pairwise - though really only interested in VS Col-0 the control
fit <- aov(lm(value ~ Line,data=bioreps))
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(value ~ Line, data = bioreps))
##
## $Line
##                             diff         lwr         upr     p adj
## Med15-Col-0            39.316354   10.229357  68.4033509 0.0045353
## Med19-1-Col-0          12.632604  -13.019716  38.2849241 0.7096161
## C1-Col-0                1.863521  -27.223476  30.9505175 0.9999996
## C2-Col-0                3.045521  -26.041476  32.1325175 0.9999813
## Med19a-2-Col-0         28.398104    2.745784  54.0504241 0.0243390
## GFP-Col-0              -6.423738  -32.076058  19.2285815 0.9903778
## OE_med19-1-Col-0      -33.291146  -58.943466  -7.6388259 0.0065687
## OE_med19-2-Col-0      -29.302627  -54.954947  -3.6503074 0.0191326
## Med19-1-Med15         -26.683750  -57.344137   3.9766366 0.1150995
## C1-Med15              -37.452833  -71.039604  -3.8660626 0.0230374
## C2-Med15              -36.270833  -69.857604  -2.6840626 0.0292616
## Med19a-2-Med15        -10.918250  -41.578637  19.7421366 0.9273692
## GFP-Med15             -45.740093  -76.400479 -15.0797060 0.0017453
## OE_med19-1-Med15      -72.607500 -103.267887 -41.9471134 0.0000079
## OE_med19-2-Med15      -68.618981  -99.279368 -37.9585948 0.0000164
## C1-Med19-1            -10.769083  -41.429470  19.8913033 0.9322777
## C2-Med19-1             -9.587083  -40.247470  21.0733033 0.9636822
## Med19a-2-Med19-1       15.765500  -11.657983  43.1889835 0.5371185
## GFP-Med19-1           -19.056343  -46.479826   8.3671409 0.3125190
## OE_med19-1-Med19-1    -45.923750  -73.347233 -18.5002665 0.0005174
## OE_med19-2-Med19-1    -41.935231  -69.358715 -14.5117480 0.0013572
## C2-C1                   1.182000  -32.404771  34.7687708 1.0000000
## Med19a-2-C1            26.534583   -4.125803  57.1949700 0.1186121
## GFP-C1                 -8.287259  -38.947646  22.3731274 0.9845255
## OE_med19-1-C1         -35.154667  -65.815053  -4.4942800 0.0185798
## OE_med19-2-C1         -31.166148  -61.826535  -0.5057615 0.0448115
## Med19a-2-C2            25.352583   -5.307803  56.0129700 0.1499194
## GFP-C2                 -9.469259  -40.129646  21.1911274 0.9661187
## OE_med19-1-C2         -36.336667  -66.997053  -5.6762800 0.0142627
## OE_med19-2-C2         -32.348148  -63.008535  -1.6877615 0.0346130
## GFP-Med19a-2          -34.821843  -62.245326  -7.3983591 0.0079653
## OE_med19-1-Med19a-2   -61.689250  -89.112733 -34.2657665 0.0000154
## OE_med19-2-Med19a-2   -57.700731  -85.124215 -30.2772480 0.0000357
## OE_med19-1-GFP        -26.867407  -54.290891   0.5560761 0.0571574
## OE_med19-2-GFP        -22.878889  -50.302372   4.5445946 0.1434128
## OE_med19-2-OE_med19-1   3.988519  -23.434965  31.4120020 0.9997776
```

A long table, but it's showing the overexpressers `OE_med19-1` and `OE_med19-2` are different from the `Col-0` control, as is the one with the noted high outliers `Med19a-2` and also `Med15`.

P-Hacking

Let's see how removing those high (`>=150`) outliers affects the *p*-values, see if any signficance we have is coming from one or two atypical data.

```r
under_150 <- data[data$value < 150, ]
bioreps_under150 <- cast(data, Line~Replicate, mean)
```

```
## Using Count as value column.  Use the value argument to cast to override this choice
```

```r
bioreps_under150 <- melt(bioreps_under150)
fit <- aov(lm(value ~ Line,data=bioreps_under150))
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(value ~ Line, data = bioreps_under150))
##
## $Line
##                             diff         lwr         upr     p adj
## Med15-Col-0            39.316354   10.229357  68.4033509 0.0045353
## Med19-1-Col-0          12.632604  -13.019716  38.2849241 0.7096161
## C1-Col-0                1.863521  -27.223476  30.9505175 0.9999996
## C2-Col-0                3.045521  -26.041476  32.1325175 0.9999813
## Med19a-2-Col-0         28.398104    2.745784  54.0504241 0.0243390
## GFP-Col-0              -6.423738  -32.076058  19.2285815 0.9903778
## OE_med19-1-Col-0      -33.291146  -58.943466  -7.6388259 0.0065687
## OE_med19-2-Col-0      -29.302627  -54.954947  -3.6503074 0.0191326
## Med19-1-Med15         -26.683750  -57.344137   3.9766366 0.1150995
## C1-Med15              -37.452833  -71.039604  -3.8660626 0.0230374
## C2-Med15              -36.270833  -69.857604  -2.6840626 0.0292616
## Med19a-2-Med15        -10.918250  -41.578637  19.7421366 0.9273692
## GFP-Med15             -45.740093  -76.400479 -15.0797060 0.0017453
## OE_med19-1-Med15      -72.607500 -103.267887 -41.9471134 0.0000079
## OE_med19-2-Med15      -68.618981  -99.279368 -37.9585948 0.0000164
## C1-Med19-1            -10.769083  -41.429470  19.8913033 0.9322777
## C2-Med19-1             -9.587083  -40.247470  21.0733033 0.9636822
## Med19a-2-Med19-1       15.765500  -11.657983  43.1889835 0.5371185
## GFP-Med19-1           -19.056343  -46.479826   8.3671409 0.3125190
## OE_med19-1-Med19-1    -45.923750  -73.347233 -18.5002665 0.0005174
## OE_med19-2-Med19-1    -41.935231  -69.358715 -14.5117480 0.0013572
## C2-C1                   1.182000  -32.404771  34.7687708 1.0000000
## Med19a-2-C1            26.534583   -4.125803  57.1949700 0.1186121
## GFP-C1                 -8.287259  -38.947646  22.3731274 0.9845255
## OE_med19-1-C1         -35.154667  -65.815053  -4.4942800 0.0185798
## OE_med19-2-C1         -31.166148  -61.826535  -0.5057615 0.0448115
## Med19a-2-C2            25.352583   -5.307803  56.0129700 0.1499194
## GFP-C2                 -9.469259  -40.129646  21.1911274 0.9661187
## OE_med19-1-C2         -36.336667  -66.997053  -5.6762800 0.0142627
## OE_med19-2-C2         -32.348148  -63.008535  -1.6877615 0.0346130
## GFP-Med19a-2          -34.821843  -62.245326  -7.3983591 0.0079653
## OE_med19-1-Med19a-2   -61.689250  -89.112733 -34.2657665 0.0000154
## OE_med19-2-Med19a-2   -57.700731  -85.124215 -30.2772480 0.0000357
## OE_med19-1-GFP        -26.867407  -54.290891   0.5560761 0.0571574
## OE_med19-2-GFP        -22.878889  -50.302372   4.5445946 0.1434128
## OE_med19-2-OE_med19-1   3.988519  -23.434965  31.4120020 0.9997776
```

Looks good! The same Lines come up as significant - the outliers aren't messing with the overall significance result.

More P-Hacking - ditching data originally in Figure 2H!

According to MCC and JJ then the lines of interest are really the `med19-1`, `Med19a-2`, `OE_med19-1` and `OE_med19-2`. Let's do the same tests for the restricted set and see if it substantially affects the result.

```
of_interest <- bioreps_under150[bioreps_under150$Line %in% c("Col-0", "Med19-1", "Med19a-2", "OE_med19-
fit <- aov(lm(value ~ Line,data=of_interest))
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(value ~ Line, data = of_interest))
##
## $Line
##                             diff        lwr        upr     p adj
## Med19-1-Col-0          12.632604 -12.105257  37.370466 0.4979780
## Med19a-2-Col-0         28.398104   3.660243  53.135966 0.0228532
## OE_med19-1-Col-0      -33.291146 -58.029007  -8.553284 0.0081105
## OE_med19-2-Col-0      -29.302627 -54.040489  -4.564766 0.0188384
## Med19a-2-Med19-1       15.765500 -10.680386  42.211386 0.3584632
## OE_med19-1-Med19-1    -45.923750 -72.369636 -19.477864 0.0011733
## OE_med19-2-Med19-1    -41.935231 -68.381118 -15.489345 0.0024215
## OE_med19-1-Med19a-2   -61.689250 -88.135136 -35.243364 0.0000889
## OE_med19-2-Med19a-2   -57.700731 -84.146618 -31.254845 0.0001638
## OE_med19-2-OE_med19-1   3.988519 -22.457368  30.434405 0.9869299
```

The result is not substantially different from before, the same lines show up as significantly different, that is `Med19a-2`, `OE_med19-1`, `OE_med19-2` and `Med15` are signifcantly different from the `Col-0` control. `Med19-1` is not.

Conclusion

The Med19-2 and Med15 lines get significantly more spores than the Col-0 wild-type and the two over-expressors of Med19 show significantly fewer spores than Col-0. There is no evidence for difference from the wild-type and other lines.

```
## Saving 6.5 x 4.5 in image
## ymax not defined: adjusting position using y instead
```