# Statisitics 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 stacked the data and 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

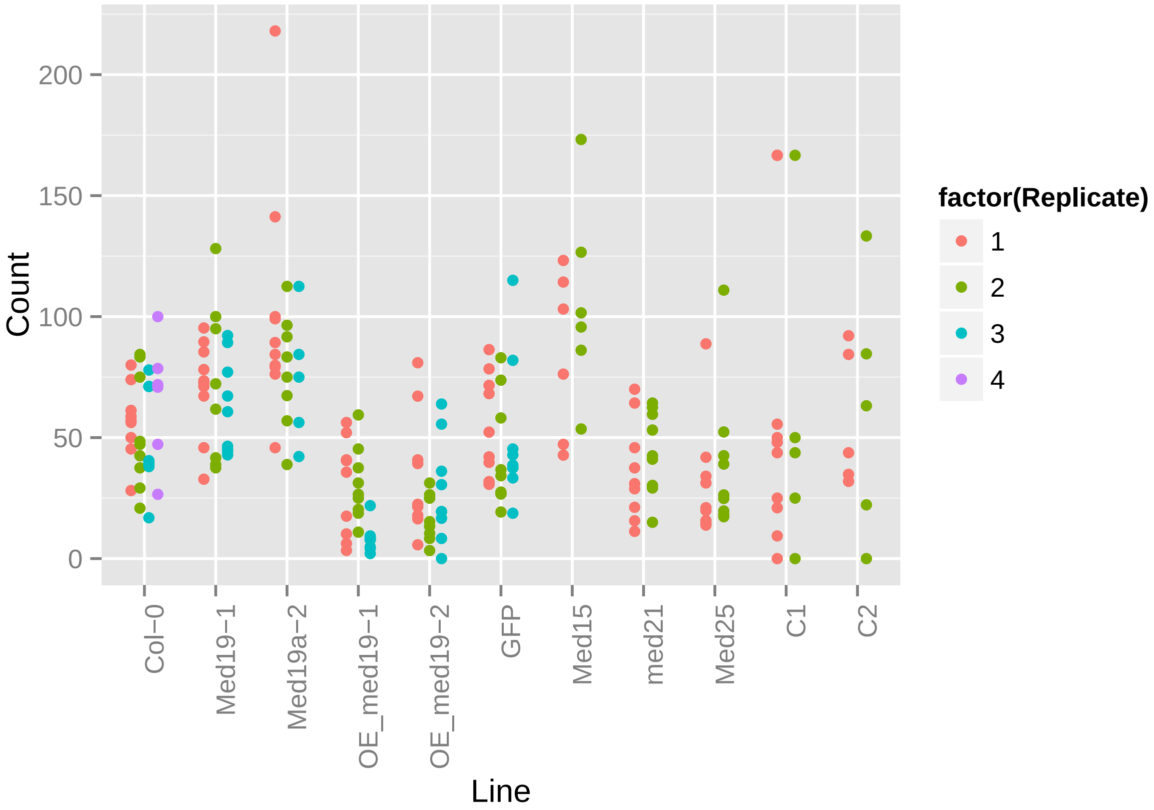
```
header = []
results = []
with open('raw/fig_2h_data_manual.csv', 'r') as file:
  for l in file:
    l = l.rstrip('\r\n')
    a = l.split(',')
    if l.startswith("Rep"):
      header = a
    else:
      for i in range(0,len(header),2):
        rep,line,count = a[i],header[i+1],a[i+1]
        if rep 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 for our preferred order and do a straightforward plot

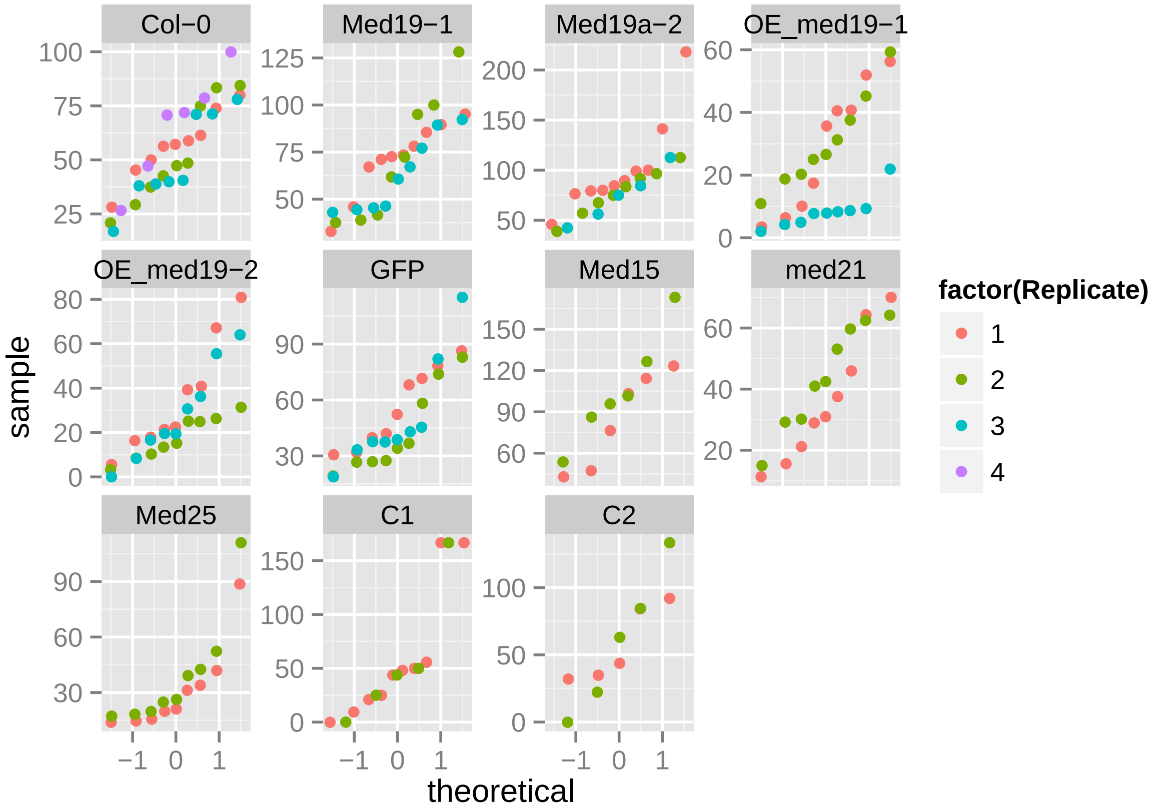
```
library(ggplot2)
data <- read.csv('data/reshaped_data.csv', header=TRUE)
data$Line <- factor(data$Line, c("Col-0","Med19-1","Med19a-2","OE_med19-1","OE_med19-2","GFP","Med15","
basic <- ggplot(data, aes(Line,Count))
scatter <- basic + geom_jitter(aes(colour=factor(Replicate)),position = position_dodge(width=0.5))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



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
```

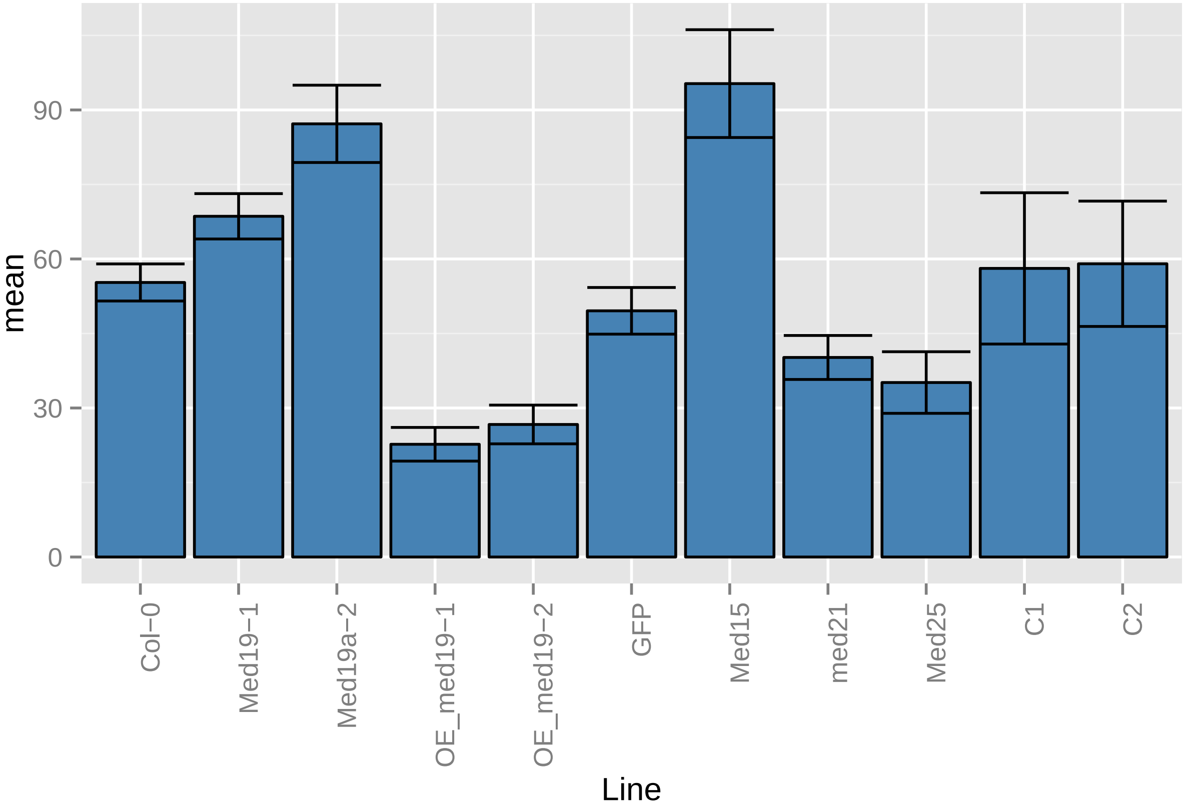
```
##          Line     mean median       diff  std_dev   std_err
## 1       Col-0 55.26875 53.125  2.1437500 21.11359  3.732392
## 2     Med19-1 68.58889 71.140  2.5511111 23.73107  4.567046
## 3    Med19a-2 87.20130 83.330  3.8713043 37.32340  7.782467
## 4  OE_med19-1 22.69333 18.750  3.9433333 17.61271  3.389568
## 5  OE_med19-2 26.68185 21.430  5.2518519 20.25470  3.898019
## 6         GFP 49.56074 39.770  9.7907407 24.40441  4.696630
## 7       Med15 95.30083 98.630  3.3291667 37.59931 10.853987
## 8       med21 40.17056 39.285  0.8855556 18.79839  4.430822
## 9       Med25 35.12389 25.580  9.5438889 26.30013  6.199001
## 10         C1 58.10267 43.750 14.3526667 58.98184 15.229045
## 11         C2 59.03000 53.455  5.5750000 39.89419 12.615651
```

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'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(Count ~ Line,data=data))
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(Count ~ Line, data = data))
##
```

```
## $Line
##                              diff        lwr         upr     p adj
## Med19-1-Col-0          13.3201389 -11.354590  37.9948680 0.8045170
## Med19a-2-Col-0         31.9325543   6.120314  57.7447945 0.0037161
## OE_med19-1-Col-0      -32.5754167 -57.250146  -7.9006875 0.0012945
## OE_med19-2-Col-0      -28.5868981 -53.261627  -3.9121690 0.0094382
## GFP-Col-0              -5.7080093 -30.382738  18.9667199 0.9996113
## Med15-Col-0            40.0320833   8.069345  71.9948221 0.0030740
## med21-Col-0           -15.0981944 -42.918188  12.7217988 0.7992672
## Med25-Col-0           -20.1448611 -47.964854   7.6751321 0.4006612
## C1-Col-0                2.8339167 -26.712959  32.3807927 0.9999999
## C2-Col-0                3.7612500 -30.447162  37.9696621 0.9999996
## Med19a-2-Med19-1       18.6124155  -8.180653  45.4054844 0.4656981
## OE_med19-1-Med19-1    -45.8955556 -71.594564 -20.1965466 0.0000012
## OE_med19-2-Med19-1    -41.9070370 -67.606046 -16.2080281 0.0000145
## GFP-Med19-1           -19.0281481 -44.727157   6.6708608 0.3662911
## Med15-Med19-1          26.7119444  -6.047993  59.4718815 0.2291335
## med21-Med19-1         -28.4183333 -57.150699   0.3140321 0.0554495
## Med25-Med19-1         -33.4650000 -62.197365  -4.7326345 0.0087936
## C1-Med19-1            -10.4862222 -40.893700  19.9212552 0.9891500
## C2-Med19-1             -9.5588889 -44.513320  25.3955423 0.9983367
## OE_med19-1-Med19a-2   -64.5079710 -91.301040 -37.7149021 0.0000000
## OE_med19-2-Med19a-2   -60.5194525 -87.312521 -33.7263835 0.0000000
## GFP-Med19a-2          -37.6405636 -64.433633 -10.8474947 0.0004130
## Med15-Med19a-2          8.0995290 -25.525506  41.7245637 0.9994442
## med21-Med19a-2        -47.0307488 -76.745700 -17.3157979 0.0000306
## Med25-Med19a-2        -52.0774155 -81.792366 -22.3624645 0.0000020
## C1-Med19a-2           -29.0986377 -60.436222   2.2389463 0.0955031
## C2-Med19a-2           -28.1713043 -63.937793   7.5951848 0.2755097
## OE_med19-2-OE_med19-1   3.9885185 -21.710490  29.6875275 0.9999902
## GFP-OE_med19-1         26.8674074   1.168398  52.5664163 0.0319816
## Med15-OE_med19-1       72.6075000  39.847563 105.3674370 0.0000000
## med21-OE_med19-1       17.4772222 -11.255143  46.2095877 0.6642257
## Med25-OE_med19-1       12.4305556 -16.301810  41.1629210 0.9456321
## C1-OE_med19-1          35.4093333   5.001856  65.8168108 0.0088159
## C2-OE_med19-1          36.3366667   1.382235  71.2910979 0.0339482
## GFP-OE_med19-2         22.8788889  -2.820120  48.5778978 0.1311902
## Med15-OE_med19-2       68.6189815  35.859044 101.3789185 0.0000000
## med21-OE_med19-2       13.4887037 -15.243662  42.2210692 0.9094081
## Med25-OE_med19-2        8.4420370 -20.290328  37.1744025 0.9969692
## C1-OE_med19-2          31.4208148   1.013337  61.8282923 0.0361271
## C2-OE_med19-2          32.3481481  -2.606283  67.3025794 0.0980800
## Med15-GFP              45.7400926  12.980156  78.5000296 0.0004655
## med21-GFP              -9.3901852 -38.122551  19.3421803 0.9928496
## Med25-GFP             -14.4368519 -43.169217  14.2955136 0.8660148
## C1-GFP                  8.5419259 -21.865552  38.9494034 0.9979139
## C2-GFP                  9.4692593 -25.485172  44.4236905 0.9984645
## med21-Med15           -55.1302778 -90.320095 -19.9404605 0.0000391
## Med25-Med15           -60.1769444 -95.366762 -24.9871272 0.0000041
## C1-Med15              -37.1981667 -73.768498  -0.6278358 0.0423676
## C2-Med15              -36.2708333 -76.700855   4.1591886 0.1241273
## Med25-med21            -5.0466667 -36.521396  26.4280627 0.9999867
## C1-med21               17.9321111 -15.078864  50.9430858 0.7983374
## C2-med21               18.8594444 -18.381958  56.1008465 0.8601184
## C1-Med25                22.9787778 -10.032197  55.9897525 0.4624844
## C2-Med25                23.9061111 -13.335291  61.1475132 0.5873699
## C2-C1                    0.9273333 -37.621180  39.4758467 1.0000000
```

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$Count < 150, ]
fit <- aov(lm(Count ~ Line,data=under_150))
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(Count ~ Line, data = under_150))
##
## $Line
##                              diff        lwr        upr     p adj
## Med19-1-Col-0           13.320139  -6.610891  33.251169 0.5258190
## Med19a-2-Col-0          25.984432   4.860675  47.108189 0.0040583
## OE_med19-1-Col-0       -32.575417 -52.506447 -12.644386 0.0000139
## OE_med19-2-Col-0       -28.586898 -48.517928  -8.655868 0.0002752
## GFP-Col-0               -5.708009 -25.639039  14.223021 0.9975410
## Med15-Col-0             32.949432   6.291681  59.607183 0.0037754
## med21-Col-0            -15.098194 -37.569814   7.373425 0.5175703
## Med25-Col-0            -20.144861 -42.616481   2.326758 0.1247418
## C1-Col-0               -24.307917 -50.125842   1.510009 0.0853414
## C2-Col-0                 3.761250 -23.870619  31.393119 0.9999972
## Med19a-2-Med19-1        12.664293  -9.241825  34.570411 0.7293053
## OE_med19-1-Med19-1     -45.895556 -66.653948 -25.137163 0.0000000
## OE_med19-2-Med19-1     -41.907037 -62.665430 -21.148644 0.0000000
## GFP-Med19-1            -19.028148 -39.786541   1.730245 0.1056068
## Med15-Med19-1           19.629293  -7.652580  46.911166 0.4103521
## med21-Med19-1          -28.418333 -51.626922  -5.209744 0.0043459
## Med25-Med19-1          -33.465000 -56.673589 -10.256411 0.0002471
## C1-Med19-1             -37.628056 -64.089918 -11.166193 0.0003273
## C2-Med19-1              -9.558889 -37.793356  18.675578 0.9905581
## OE_med19-1-Med19a-2    -58.559848 -80.465967 -36.653730 0.0000000
## OE_med19-2-Med19a-2    -54.571330 -76.477448 -32.665212 0.0000000
## GFP-Med19a-2           -31.692441 -53.598559  -9.786323 0.0002308
## Med15-Med19a-2           6.965000 -21.200009  35.130009 0.9992978
## med21-Med19a-2         -41.082626 -65.323207 -16.842046 0.0000053
## Med25-Med19a-2         -46.129293 -70.369873 -21.888712 0.0000002
## C1-Med19a-2            -50.292348 -77.663818 -22.920879 0.0000005
## C2-Med19a-2            -22.223182 -51.311878   6.865515 0.3190343
## OE_med19-2-OE_med19-1    3.988519 -16.769874  24.746911 0.9999274
## GFP-OE_med19-1          26.867407   6.109014  47.625800 0.0018176
## Med15-OE_med19-1       65.524848  38.242975  92.806722 0.0000000
## med21-OE_med19-1       17.477222  -5.731367  40.685811 0.3404375
## Med25-OE_med19-1       12.430556 -10.778033  35.639144 0.8118336
## C1-OE_med19-1           8.267500 -18.194363  34.729363 0.9949865
## C2-OE_med19-1          36.336667   8.102200  64.571134 0.0019963
## GFP-OE_med19-2         22.878889   2.120496  43.637282 0.0176890
## Med15-OE_med19-2       61.536330  34.254457  88.818203 0.0000000
## med21-OE_med19-2       13.488704  -9.719885  36.697293 0.7228484
## Med25-OE_med19-2        8.442037 -14.766552  31.650626 0.9837431
## C1-OE_med19-2           4.278981 -22.182881  30.740844 0.9999855
## C2-OE_med19-2          32.348148   4.113681  60.582615 0.0109372
## Med15-GFP              38.657441  11.375568  65.939314 0.0003510
## med21-GFP              -9.390185 -32.598774  13.818404 0.9651976
## Med25-GFP             -14.436852 -37.645441   8.771737 0.6327817
## C1-GFP                -18.599907 -45.061770   7.861955 0.4469212
## C2-GFP                  9.469259 -18.765208  37.703726 0.9912224
## med21-Med15           -48.047626 -77.237150 -18.858102 0.0000116
## Med25-Med15           -53.094293 -82.283817 -23.904769 0.0000007
## C1-Med15              -57.257348 -89.094746 -25.419951 0.0000010
## C2-Med15              -29.188182 -62.513470   4.137107 0.1469392
## Med25-med21            -5.046667 -30.470402  20.377069 0.9999021
## C1-med21               -9.209722 -37.634322  19.214878 0.9933008
## C2-med21               18.859444 -11.222325  48.941214 0.6215967
## C1-Med25               -4.163056 -32.587656  24.261545 0.9999943
## C2-Med25               23.906111  -6.175658  53.987880 0.2630186
## C2-C1                  28.069167  -4.588213  60.726546 0.1665859
```

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.

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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(Count ~ Line, data = of_interest))
##
## $Line
##                             diff         lwr       upr     p adj
## Med19-1-Col-0          13.320139  -4.3000316  30.94031 0.2301396
## Med19a-2-Col-0         31.932554  13.5000897  50.36502 0.0000429
## OE_med19-1-Col-0      -32.575417 -50.1955872 -14.95525 0.0000108
## OE_med19-2-Col-0      -28.586898 -46.2070686 -10.96673 0.0001502
## Med19a-2-Med19-1       18.612415  -0.5204569  37.74529 0.0607313
## OE_med19-1-Med19-1    -45.895556 -64.2471620 -27.54395 0.0000000
## OE_med19-2-Med19-1    -41.907037 -60.2586435 -23.55543 0.0000000
## OE_med19-1-Med19a-2   -64.507971 -83.6408434 -45.37510 0.0000000
## OE_med19-2-Med19a-2   -60.519452 -79.6523248 -41.38658 0.0000000
## OE_med19-2-OE_med19-1   3.988519 -14.3630879  22.34012 0.9746800
```

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.