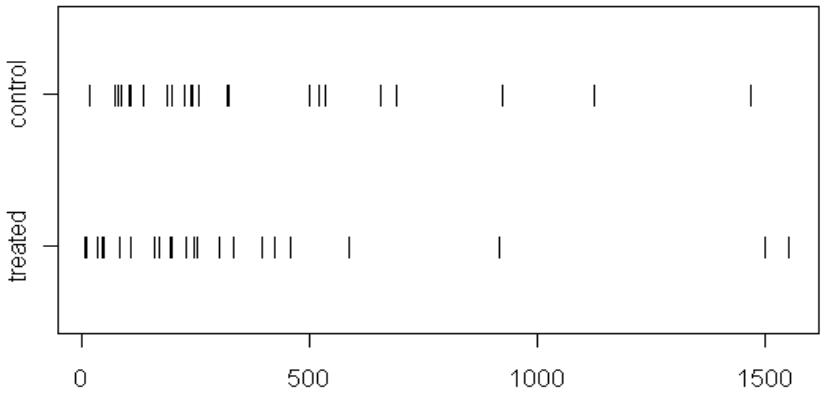
Name:

Question 1

Your lab mate has collected the data shown below in both tabular and graphical form for the number of granules observed in treated and untreated cells in his latest experiment. The data are not paired.

Treated	Control
234	928
427	92
173	539
591	696
175	326
462	1127
308	503
88	242
400	523
53	246
1556	229
918	76
14	138
1503	659
112	261
40	90
49	324
203	1472
11	113
198	85
461	192
164	203
336	90
250	20
256	108



Make a table giving the quartiles for each dataset.

Do these measurements look normally distributed to you? Justify your response.

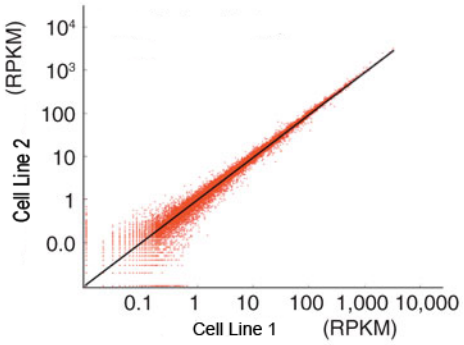
Discuss the appropriateness of using an unpaired t-test with the above data to compare the effect of treatment on the number of granules per cell in this experiment.

What advice would you give your lab mate on how to proceed?

Question 2

Your lab head has been asked to review a paper that claims to identify genes involved in a novel pathway. Since you've been telling her how much you've learned in your qBio class, she's asked you to comment on the statistical analysis in the paper.

To elucidate the pathway, the authors collected RNA-Seq data from ten different cell lines. For each cell line, two RNA-Seq experiments were performed, one in the presence of an siRNA known to lower production of the pathway's product, and another in the absence of the siRNA. For each RNA-Seq experiment, expression levels for each of the 25,354 genes detected were computed as RPKM values. The figure shows a comparison of the RPKM values from two of the cell lines in the absence of the siRNA.



The authors then identified a candidate set of 1,297 differentially expressed genes by selecting only those genes whose mean RPKM value in the absence of the siRNA differed by more than 100 from the mean RPKM value observed in the presence of the siRNA.

For each of the 1,297 candidate genes, a p-value for differential expression was computed using a paired t-test. The authors chose to be conservative in their analysis by applying the Bonferroni correction to the p-values, dividing each of them by 1,297. A total of 127 genes had corrected p-values below 0.05, and were reported as likely to be involved in the pathway under investigation.

What points would you suggest your lab head include in her review?

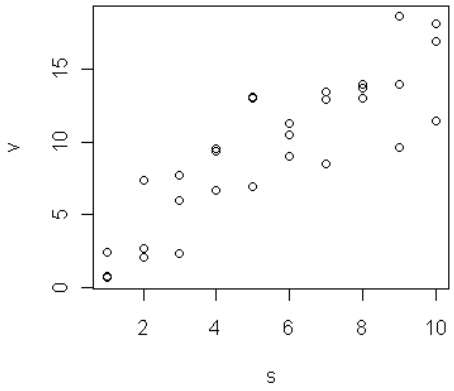
[more space to continue your response to Question 2]

Question 3

Your lab mate is fitting kinetic data to a Michaelis-Menten model.

$$v = \frac{v_{max} \cdot s}{K_m + s}$$

As shown in the figure below, he has measured the reaction rate for each concentration multiple times.



When fitting the Michaelis-Menten model to his data, your lab mate insists on averaging the measurements at each concentration before performing a non-linear fit in R. He claims that R "doesn't work" when the model is fit to the raw data with all of the points.

He shows you the fit using all the points (transcript 1, on page 8), and points out that R fails to estimate the upper bounds of the CIs for the kinetic parameters.

He then shows you that when he fits the model using averaged data, R can compute the CIs (transcript 2, on page 9).

Do you think it is important to fit the model to the raw data with all of the points in this case? Explain your reasoning?

Why do you think R failed to compute the CIs of the kinetic parameters in the first fit?

What advice would you give to your lab mate?

```
> q3
    s          v
1   1  2.4520454
2   2  2.1043420
3   3  2.3307356
4   4  9.3392197
5   5  6.9631963
6   6 10.4745642
7   7  8.4593253
8   8 13.6811145
9   9 18.5851897
10 10 18.0820365
11  1  0.6644538
12  2  7.3705949
13  3  7.7329379
14  4  6.6338494
15  5 12.9840912
16  6 11.2460105
17  7 13.4293666
18  8 13.9836722
19  9  9.5973956
20 10 11.4292015
21  1  0.7570696
22  2  2.6422438
23  3  5.9403375
24  4  9.5246435
25  5 13.0497241
26  6  8.9787538
27  7 12.9242244
28  8 12.9574417
29  9 13.9418355
30 10 16.8619078
> m1 <- nls(v ~ vmax * (s/(Km + s)), start=list(Km=25, vmax=40), data=q3)
> summary(m1)

Formula: v ~ vmax * (s/(Km + s))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
Km      18.16      11.78   1.541   0.1345
vmax    43.39      20.16   2.153   0.0401 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.419 on 28 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 7.131e-06

> confint(m1)
Waiting for profiling to be done...
          2.5% 97.5%
Km    7.301663    NA
vmax 24.343849    NA
```

Transcript 1

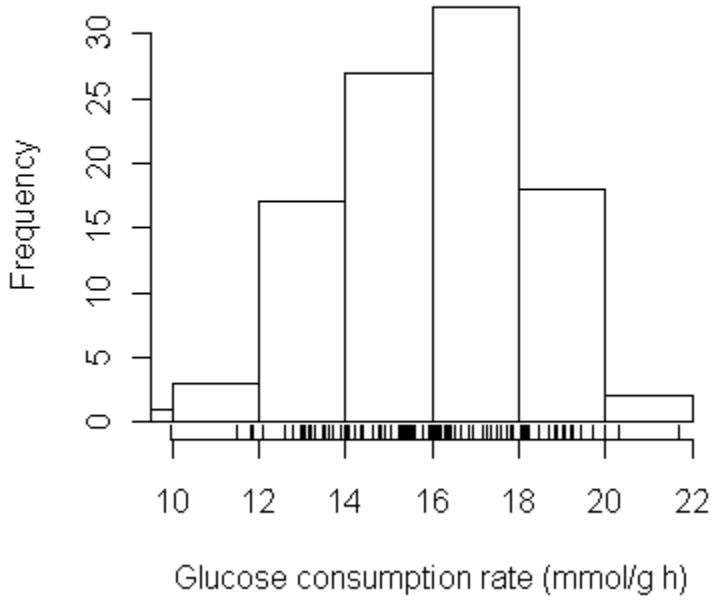
```
> q3b=aggregate(q3$v, by=list(s=q3$s), FUN="mean")
> q3b
    s         x
1   1  1.291190
2   2  4.039060
3   3  5.334670
4   4  8.499238
5   5 10.999004
6   6 10.233109
7   7 11.604305
8   8 13.540743
9   9 14.041474
10 10 15.457715
> m2 <- nls(x ~ vmax * (s/(Km + s)), start=list(Km=25, vmax=40), data=q3b)
> confint(m2)
Waiting for profiling to be done...
          2.5%     97.5%
Km    9.119785  65.86029
vmax 27.636742 123.97110
```

Transcript 2

Question 4

You are designing an experiment to test if a knockdown affects the rate of glucose consumption in yeast cells under anaerobic conditions. Your lab can make individual measurements of glucose metabolism for $25 each, and your lab head has given you a budget of $1,600 for this experiment.

To estimate the variance in glucose consumption under anaerobic conditions, you look at some old lab notebooks and gather 100 independent, prior measurements involving the same strain of yeast that you plan to use. A plot of these data is below.



Assuming you design an <u>unpaired</u> experiment, estimate the difference in glucose consumption rate that you have a reasonable chance of detecting (use power=50%, α=0.05)?

Is it possible to quantify the difference in glucose consumption rate that you have a reasonable chance of detecting (power = 50%, α=0.05) in a <u>paired</u> experiment from just this information? Explain your reasoning.

[extra space in case you need it]

[extra space in case you need it]

[extra space in case you need it]