Common mistakes in data presentation and statistical analysis: how can the BioStat Decision Tool help?

Anne L. Bishop[1] and Siouxsie Wiles[2,3*]

[1]Independent scholar, Nottingham, United Kingdom; [2]Department of Molecular Medicine and Pathology, University of Auckland, Auckland, New Zealand; [3]Maurice Wilkins Centre for Molecular Biodiscovery, New Zealand.

* Corresponding author:

Dr. Siouxsie Wiles

Department of Molecular Medicine and Pathology, University of Auckland

85 Park Road, Grafton, Building 502, Level 3

Auckland 1023

New Zealand

Phone: +64 9 373 7599 ext 84284 (office)

E-mail: s.wiles@auckland.ac.nz

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

Abstract

As medical and molecular microbiologists who regularly read the scientific literature, it is our impression that many published papers contain data that is inappropriately presented and/or analysed. This is borne out by a number of studies which indicate that typically at least half of published scientific articles that use statistical methods contain statistical errors. While there are an abundance of resources dedicated to explaining statistics to biologists, the evidence would suggest that they are largely ineffective. These resources tend to focus on how particular statistical tests work, with reams of complicated-looking mathematical formulae. In addition, many statisticians are unfamiliar with the application of statistical techniques to molecular microbiology, instead telling us we need more samples, which can be difficult both ethically and practically in fields that include animal work and painstaking sample collection. In an age where performing a statistical test merely requires clicking a button in a computer programme, it could be argued that what the vast majority of biologists need is not mathematical formulae but simple guidance on which buttons to click. We have developed an easy to follow decision chart that guides biologists through the statistical maze. Our practical and user friendly chart should prove useful not only to active researchers, but also to journal editors and reviewers to rapidly determine if data presented in a submitted manuscript has been correctly analysed.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

It is estimated that around half of published papers in the biomedical literature contain mistakes in data presentation and analysis [1,2,3,4,5,6,7,8]. The most up-to-date review of such mistakes is for the journal "Infection and Immunity", in which Dr Cara Olsen looked at all 141 articles from two issues, January 2002 (volume 70, no. 1) and July 2002 (volume 70, no. 7) [8]; her conclusions are in line with those of other journals similarly reviewed since 1979 [1,2,3,4,5,6,7]. Our reading of current literature in many biomedical journals suggests that the situation remains largely the same and we are certainly not the only researchers to find this concerning [9,10]. In Box 1, we highlight some of the most common mistakes being made by biomedical researchers. Such mistakes appear to be particularly prevalent when it comes to analysis of small data sets, to which many commonly used statistical analysis tools, such as t-tests for statistical analysis and presentation of means and standard deviations, are not well suited.

Box 1. Common mistakes in data analysis and presentation in biomedical publications.

1. Failure to adjust or account for multiple comparisons, which could lead to the presentation of false positive results.
2. Reporting that a result is "significant" without conducting a statistical test.
3. Use of statistical tests that assume a normal distribution on data that is skewed.
4. Presenting data with unlabelled or inappropriate error bars/ measures of variability.
5. Failure to describe the tests performed.

To address this issue, we have developed a simple flow chart to help researchers avoid these common mistakes when handling their data. Called the BioStat Decision Tool (DT), the flow chart (summarised in Fig. 1) is freely available online

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

(http://flexiblelearning.auckland.ac.nz/biostat-tree/index.html), or for a small fee, as a smartphone application. The BioStat DT is a decision making tree, complete with handy tips and a glossary of terms to help scientists understand each step as they go along. The BioStat DT can be used to find out how best to analyse and present particular types of data, but could also be useful as a guide for journal reviewers and editors when assessing an author's data presentation and analysis choices. The tool is aimed at biologists with small data sets, which are often encountered in research involving human samples or animal models due to practical and/or ethical considerations. It is important to note that the BioStat DT is simply a decision making tool; it does not tell researchers how to carry out a particular test with their software package, or allow users to input their own data.

An example of using the BioStat DT to analyse and present a dataset

In this section, we will use a thought experiment and simulated data to explore how the BioStat DT could help researchers avoid making the mistakes outlined in Box 1. Imagine a group of microbiologists are interested in the effect of two different bacterial gene deletions (let's call them *Δmut1* and *Δmut2*) upon transcription of *geneX* in a mouse infection model. The design for the thought experiment to test the mutants is outlined in Fig. 2. The researchers want to know whether expression of *geneX* in a tissue of interest is significantly different between vehicle (saline)-inoculated mice (from here-on-in termed controls) and mice infected with a wild type (WT) bacterium. This may have been shown previously in the literature. Furthermore, the researchers also want to know whether expression of *geneX* differs between mice infected with the WT bacterium and *Δmut1* or *Δmut2*, and whether the expression of *geneX* differs between the two deletion strains.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

Experimental set up and data

The researchers use 4 mice per group in the first experiment, and repeat the entire experiment on a separate occasion, to give a total of 8 mice per test condition from two independent experiments and a total of 32 tissue samples to process, as outlined in Fig. 2. They prepare RNA from the tissue of interest and make cDNA with random primers. Although the PCR primers could be designed to cross introns in mammalian genes, so that they shouldn't give amplification products with genomic DNA as template, RNA mixes lacking reverse transcriptase (RT) enzyme are prepared and tested for background amplification. Quantitative (q) RT-PCR is used to determine the levels of *geneX* transcript in each cDNA sample, normalized to the levels of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) transcript, chosen as an example of a relatively stably expressed gene in the majority of mammalian tissues (see for instance [11] for a discussion of qRT-PCR and relative-expression analysis). The double-stranded DNA produced by PCR amplification is detected using SYBR-green. Standard curves of differing template concentration are used to determine the linear range and efficiency of the primers, but are not discussed further here, except to say that C_T (cycle threshold) values (where fluorescence comes above a background threshold) need to fall within the linear range to be reliable. GAPDH-normalized *geneX* transcription is expressed relative to one calibrator sample from the control group. Each qRT PCR plate contains samples lacking RT (just one reaction for each) and the cDNA qRT-PCR reactions run in triplicate. One WT-infected cDNA sample, which would be expected to give a positive signal for *geneX*, is also included on every plate to allow normalization for inter-plate variation.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

Having worked through the experiment, the researchers determine the C_T values that are below the background (the lower the cycle number at which product is detected, the more template was present in the sample) for both GAPDH and *geneX* for each sample. Now we will use the BioStat DT to avoid making the data presentation and analysis flaws that are rife in the literature.

Step 1. Identifying the type of data and replicates

The BioStat DT begins with a question about the type of data that the researcher is working with (*frequencies* or *measurements*) (Fig. 3A). In the case of our thought experiment it is *measurements*. Within the smartphone application, there is a Glossary in which terms like this are defined. Selecting the *measurements* option leads to a question about *ratios* (Fig. 3B). If the data is expressed as a ratio then the data is not continuous, a prerequisite for many statistical tests which assume a normal/Gaussian distribution (Box 2). As the researchers plan to analyze the normalized C_T values as relative amounts of gene expression compared to a calibrator sample, then the values will be ratios, so the answer to this question is YES. At this point, the BioStat DT suggests that the researcher should consider transforming the data to spread it into a more continuous distribution (Fig. 3C). We will come back to this later.

After selecting CONTINUE, the tool next asks about the experimental replicates (Fig. 3D), and whether these are *technical* as well as *biological* replicates. In our thought experiment, the technical replicates are the triplicates that were plated for bacterial quantification (given as colony forming units [CFU]) and the triplicates carried out for each qRT-PCR reaction. As these technical replicates are essentially a measure of pipetting accuracy, they should be averaged to get the most accurate value for each test sample within the experimental design. The replicates

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

can be plotted on a scatter plot or one could just eyeball the numbers to see if there were any outliers due to pipetting and/or homogenization errors. If there are the odd erroneous values, having 3 samples and taking a median of the technical replicates will effectively cancel out these outliers and give a middle-of-the-evidence data point to work with. This is what the BioStat DT suggests (Fig. 3D), but if the replicates are evenly distributed then the mean and the median will be very similar and could be used interchangeably. For an inter-plate correction, the researchers take the median of the triplicates from the same cDNA on each plate and use this ratio to correct C_Ts for all other samples on the plate. The biological replicates are the individual mice; in this case there are two different bacterial inputs and two different sets of mice, in order to check reproducibility across experiments. It is recommended to pool these biological replicates, to give *n*=8 for each test group. If a researcher was to have problems with inter-experimental variation then they could normalize to a control group or show all of the data, so two lots of n = 4 in this case, to show that the trends are consistent, even if the absolute values are not. Showing one "representative experiment" is not acceptable; all of the data will be needed in order to test for statistical significance and, if justified, to present something as a significant finding. This will avoid the common mistake of "Reporting that a result is "significant" without conducting a statistical test" (Box 1). Identification of biological and technical replicates and handling of these data appropriately can be a troublesome area and is covered specifically in reviews such as [12].

Step 2. Is the data normally distributed?

In the thought experiment, the researchers have taken the median value for each of the triplicates, checked the input CFUs are reasonably consistent between the different bacterial

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

strains being studied and within the two experimental repeats (for example, within 10%), normalized the data and calculated C_T values relative to both GAPDH levels and to a control calibrator sample, termed the $2^{-\Delta\Delta CT}$ method. Many excellent texts deal with the ins and outs of calculating $2^{-\Delta\Delta CT}$ and we won't go into that further [13]. The researchers now have 8 data points for each test condition. The next question asked by the BioStat DT regards the distribution of the data (Fig. 3E). Answering this question is important to avoid making a mistake which is rife in the biomedical literature: "Use of statistical tests that assume a normal distribution on data that is skewed" (Box 1). It is especially important for scientists working with small datasets, such as those generated by the experiment described here, not to assume normality and present the data as means and standard deviations/standard errors, as such datasets can be dramatically skewed by outliers. A quick reminder of how to test for normality is given in Box 2.

Box 2. Normal or not?

A normal/ Gaussian, distribution is perfectly symmetrical around the mean, with a bell shaped curve when you plot the frequency of each value, and stretches infinitely in each direction. Data with this distribution allows many powerful assumptions to be made for testing of differences between groups, for instance using a Student's t-test or One Way ANOVA.

There are a number of ways to test a dataset for normal distribution:

1. Using mathematical tests embedded within statistical software packages, in particular the D'Agostino-Pearson test is recommended, which looks at both how symmetrical (skewness) and how peaked or flat (kurtosis) the data distribution is compared to a perfect symmetrical Gaussian/normal distribution bell curve [14]. A large p value (close to 1) for these tests suggests that your sample is consistent with a Gaussian distribution,

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

i.e. it does not significantly deviate from normality. At least eight data points are required to carry out the D'Agostino-Pearson test. If you have less data you can't look at the distribution mathematically, so we suggest using non-parametric tests that do not assume a normal distribution.

2. Plotting the data as a scatter graph to see the shape of the distribution. Does it look like a bell-shaped normal curve?
3. Analysing the column statistics, for example, what are the means and medians of each group and are they almost identical, suggesting a normal distribution (for example see Table 2 and Fig. 4A where mean ≠ median for WT and Δ*mut2*)?

Analysing the simulated data (Table 1) from the thought experiment outlined in Fig. 2, which resembles the shape of many data sets that we have encountered for *in vivo* infection-induced host responses in our own experiments and in the literature, we can see that there are some individual mice with high responses that result in the WT and Δ*mut2* infected groups not being normally distributed (Table 1 and Fig. 4A). This variation results in quite different values for the means and medians and the data fails the D'Agostino-Pearson normality test (Fig. 4A and Table 2). For these two groups of data you can reject the null hypothesis that they conform to a normal distribution.

Answering NO to the question "Is your data normally distributed" the BioStat DT then asks if the number of samples in each group (*n*) is 8 or greater, the cut off for the D'Agostino-Pearson test for normality (Fig. 3F). The limit exists because it is difficult to predict mathematically what a theoretical continuous infinite distribution of the data would look like with such a small number of data points to work with. Some normality tests can work with less than *n*=8 group sizes, such

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

as Kolmogorov-Smirnov, but these are not as well respected as the D'Agostino-Pearson test [14]. In this case, using the simulated dataset, answering YES to this question leads the BioStat DT to ask what the non-normal data looks like, with several options to choose from (Fig 3G). From Fig. 4B the simulated data appears positively skewed in the groups that do not conform to a normal distribution. Selecting POSITIVELY SKEWED, the BioStat DT suggests performing a transformation (Fig. 3H). You may recall that a transformation of the data is also suggested because the data is in the form of ratios (Fig. 3C), so there are two reasons to transform the data prior to further analysis. In this case, carrying out a $\log_{10}$ transformation results in a tighter grouping of the data set on a scatter plot (Fig. 4B), the means and medians are closer and the transformed data now passes the D'Agostino-Pearson test for normality (Table 2). All subsequent statistical analysis should now be performed using $\log_{10}$-transformed data. Selecting CONTINUE leads the BioStat DT to ask if the transformed data is now normally distributed (Fig. 3I), to which the answer now is YES. This means that the transformed data can now be analysed using parametric tests and suggests presenting the data as means with either the 95% confidence interval (probably the most appropriate choice [9]) or standard deviation (Fig. 3J). With this advice, the BioStat DT attempts to address common mistake #4: "unlabelled or inappropriate error bars/measures of variability" (Box 1). For datasets such as the simulated one presented here, with a small number of samples, our preference is to present all of the data points individually, so that the reader can see the full spread of data for themselves.

If the transformed data had still not passed the D'Agostino-Pearson test for normality, selecting NO would result in the BioStat DT advising that the data be analysed using non-parametric tests, and presented as medians and inter-quartile ranges, rather than using the mean and standard deviation or 95% confidence intervals. With this advice, the BioStat DT attempts to address common mistake #3: "Use of statistical tests that assume a normal distribution on data that is

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

skewed" (Box 1). It also serves to remind the user of data presentation options, and helps them to avoid another common mistake of carrying out non-parametric tests, but presenting the mean rather than the median, so that the presentation does not reflect the analysis carried out.

Step 3. Selecting the appropriate statistical test to perform

Selecting CONTINUE on the BioStat DT leads to the question "Are you looking for differences or associations?" (Fig. 3K). For the simulated dataset we are looking for *differences*, so selecting this option takes us to a question about the number of groups there are in the dataset (Fig. 3L). For the simulated dataset there are four groups: controls, WT-infected, *Δmut1*-infected and *Δmut2*-infected. Selecting MORE THAN TWO, the next question the BioStat DT asks is "Are you examining the effect of one factor or two?" (Fig. 3M). For the thought experiment the answer is ONE, that being *geneX* transcript levels. Next comes the question: "Are your data from *independent* samples, or from *repeated measurements* on the same sample?" (Fig. 3N). In this case, the data is from independent samples, so selecting this option leads the BioStat DT to suggest the "One way Analysis of Variance (ANOVA)" is the appropriate test for analyzing the simulated dataset (Fig. 3O).

At this stage, the BioStat DT also explains why the selected test is appropriate, alongside a reminder of best presentation options and the need to state what test is carried out and what is presented (mean or median and the measure of variability) in figure legends and/or methods when publishing the data. In this particular example, the BioStat DT is addressing common mistake #1 "Failure to adjust or account for multiple comparisons…" (Box 1), by taking the user to the ANOVA and explaining why it is not appropriate to do multiple *t*-tests without correcting for false positives. The tool also helps the user to avoid mistakes #2 "Reporting that a result is

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

"significant" without conducting a statistical test", #4 "Presenting data with unlabelled or inappropriate error bars/ measures of variability" and #5 "Failure to describe the tests performed" (Box 1).

Step 4. Analysing and presenting the data

While the BioStat DT does not tell researchers how to carry out a particular test with their software package, or allow users to input their own data, we will finish by describing the analysis of the simulated dataset from the experiment outlined in Fig. 2. Performing an ANOVA on the dataset yields an over-all probability (p)-value of <0.0001, indicating there are significant differences between the groups within our thought experiment (Table 3). Teasing out the groups within the dataset that are different from each other requires post-hoc testing with corrections for multiple comparisons, such as Bonferroni's correction. These corrections are required to avoid false positives (type 1 errors) due to repeated testing of the same data [15]. Researchers can choose to make only the most biologically interesting comparisons, as every extra comparison results in an additional correction to the p-values. This increased stringency can therefore result in the researcher making what is known as a type 2 error, failing to detect a significant result where one exists.

From the post-hoc tests performed on the simulated dataset, we find that all of our infection groups have significantly higher expression of *geneX* when compared to the uninfected controls (Table 4). Furthermore, the tests indicate that the expression of *geneX* is significantly lower in the *Δmut1*-infected group when compared to the WT-infected group. However, *geneX* expression by the *Δmut2*-infected group does not differ significantly from either the WT or *Δmut1*-infected groups. Interestingly, if we compare the results in Table 4 with a similar analysis

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

carried out on the untransformed dataset (Table 5), we see that there is now no longer a significant difference between the uninfected controls and the *Δmut1*-infected group. In this case, using the wrong analysis, i.e. ignoring the lack of a normal distribution in all groups and leaving the data un-transformed, would have resulted in the researchers getting a false-negative, and accepting the null hypothesis that there was no significant difference between WT and Δ*mut1*, when in fact there was.

In the final presentation of our simulated dataset, we could plot the $2^{-\Delta\Delta CT}$ values on a $\log_{10}$ scale, so that readers can see the effect of the transformation that we carried out in order to make the data fit a more normal distribution, while retaining values that are easier to quickly understand (Fig. 5). We would state in the legend what statistical test was performed, what correction for multiple comparisons was chosen, and what the p values were (see Fig. 5 legend and Box 3). We would also state, in the legend or in the methods, the numbers of samples per group and more details of what we did to test reproducibility, such that two independent experiments were performed, with 4 mice in each to give a total of 8 samples per group, and that for data in the form of ratios statistical tests were carried out after $\log_{10}$-transformation.

Box 3: Degrees of significance?

More often than not, when scientists present a statistical analysis of their results they do so using adjectives such as “very significant” and “extremely significant”, or different numbers of asterisks. In contrast, many statisticians feel strongly that once a threshold significance level has been set (usually 0.05), a result can only be "statistically significant" or not "statistically significant", and so oppose the use of adjectives or asterisks to describe levels of statistical

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

significance. In reality, it can be useful to see whether the data would have passed a more stringent threshold for significance, which is perhaps why this practice persists.

Conclusion

In summary, the BioStat DT leads users step-by-step through data analysis and presentation decisions, describing in simple terms what the next step should be and why, as well as giving tips on how best to present the data. We hope that by providing user-friendly, maths-free statistical support to researchers, the BioStat DT will help raise the standards in data presentation and analysis and improve adherence to the guidelines provided for authors by many journals.

It should be noted that the fact that many published papers fall short of the standards described in journal guidelines, suggests that many reviewers and editors are failing to identify errors during the peer review process. The BioStat DT could also provide a means for reviewers and editors to assess whether the guidelines for authors have been followed, and indeed, whether the appropriate statistical test has been performed. The tool also provides a means by which researchers can justify the analysis they performed by allowing them to generate a summary of the decision tree choices that were taken.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

References

1. White SJ (1979) Statistical errors in papers in the British Journal of Psychiatry. Br J Psychiatry 135: 336-342.
2. Emerson JD, Colditz GA (1983) Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 309: 709-713.
3. Felson DT, Cupples LA, Meenan RF (1984) Misuse of statistical methods in Arthritis and Rheumatism. 1982 versus 1967-68. Arthritis Rheum 27: 1018-1022.
4. MacArthur RD, Jackson GG (1984) An evaluation of the use of statistical methodology in the Journal of Infectious Diseases. J Infect Dis 149: 349-354.
5. Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM (1985) Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg 64: 607-611.
6. Cruess DF (1989) Review of use of statistics in The American Journal of Tropical Medicine and Hygiene for January-December 1988. Am J Trop Med Hyg 41: 619-626.
7. Welch GE, 2nd, Gabbe SG (1996) Review of statistics usage in the American Journal of Obstetrics and Gynecology. Am J Obstet Gynecol 175: 1138-1141.
8. Olsen CH (2003) Review of the use of statistics in infection and immunity. Infect Immun 71: 6689-6692.
9. Cumming G, Fidler F, Vaux DL (2007) Error bars in experimental biology. The Journal of Cell Biology. pp. 7-11.
10. Vaux DL (2012) Research methods: Know when your numbers are significant. Nature 492: 180-181.
11. VanGuilder HD, Vrana KE, Freeman WM (2008) Twenty-five years of quantitative PCR for gene expression analysis. Biotechniques 44: 619-626.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

12. Vaux DL, Fidler F, Cumming G (2012) Replicates and repeats--what is the difference and is it significant? A brief discussion of statistics and experimental design. EMBO Rep 13: 291-296.

13. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402-408.

14. D'Agostino RB, Belanger A, D'Agostino Jr RB (1990) A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician 44: 316-321.

15. Streiner DL, Norman GR (2011) Correction for multiple testing: is there a resolution? Chest 140: 16-18.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

Acknowledgments/ Financial Disclosure/ Competing Interests

S.W. is supported by a Sir Charles Hercus Fellowship (09/099) from the Health Research Council of New Zealand (www.hrc.govt.nz). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. S.W. declares that her company Lucy Ferrin Ltd supports development of smartphone application versions of the BioStat Decision Tool that are available for purchase. A.B. and S.W. will therefore benefit from any profits generated from the sale of the smartphone applications.

Tables

Table 1. Simulated data for relative expression of *geneX* given as $2^{-\Delta\Delta CT}$ and $\log_{10} 2^{-\Delta\Delta CT}$.

Table 2. Column statistics for the untransformed and transformed simulated data of the relative expression of *geneX*. CI – confidence interval.

Table 3. One-way analysis of variance (ANOVA) of the transformed and untransformed simulated data. F – ratio of between group variability to within group variability; R square – proportion of the variation in the dependent variable accounted for by the independent variable.

Table 4. Post-hoc testing of the transformed simulated data using Bonferroni's correction for multiple comparisons. t – ratio of the departure of an estimated parameter from its notional value; CI – confidence interval.

Table 5. Post-hoc testing of the untransformed simulated data using Bonferroni's correction for multiple comparisons. t – ratio of the departure of an estimated parameter from its notional value; CI – confidence interval.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

Figure legends

Figure 1. Summary of the Biostat DT tree.

Figure 2. Experimental design to determine the effect of two different bacterial gene deletions (Δ*mut1* and Δ*mut2*) upon transcription of *geneX* in a mouse infection model. Key: GADPH, Glyceraldehyde 3-phosphate dehydrogenase; qRT PCR, quantitative reverse transcriptase PCR; C_T, cycle threshold; $2^{-\Delta\Delta CT}$, a method of relative gene expression where, for each sample, efficiency of amplification (for perfect amplification this is 2) is raised to the negative power of $\Delta\Delta C_T$. Where, $\Delta\Delta C_T$ is the test gene C_T expressed relative to the control gene (in this case GAPDH) C_T and relative to a calibrator sample arbitrarily chosen from a control condition (in this case saline).

Figure 3. Summary of the BioStat DT questions and choices made to analyse the simulated dataset.

Figure 4. Expression of *geneX* as determined by qRT-PCR normalized to GAPDH expression and relative to an arbitrarily chosen calibrator sample within the saline-control group.

Data is shown before (A) and after (B) $\log_{10}$ transformation. Saline-inoculated negative controls (controls) are compared with mice X hours after infection wild type (WT), Δ*mut1* or Δ*mut2* bacterial strains. Mean values per group are denoted by solid lines, while median values are

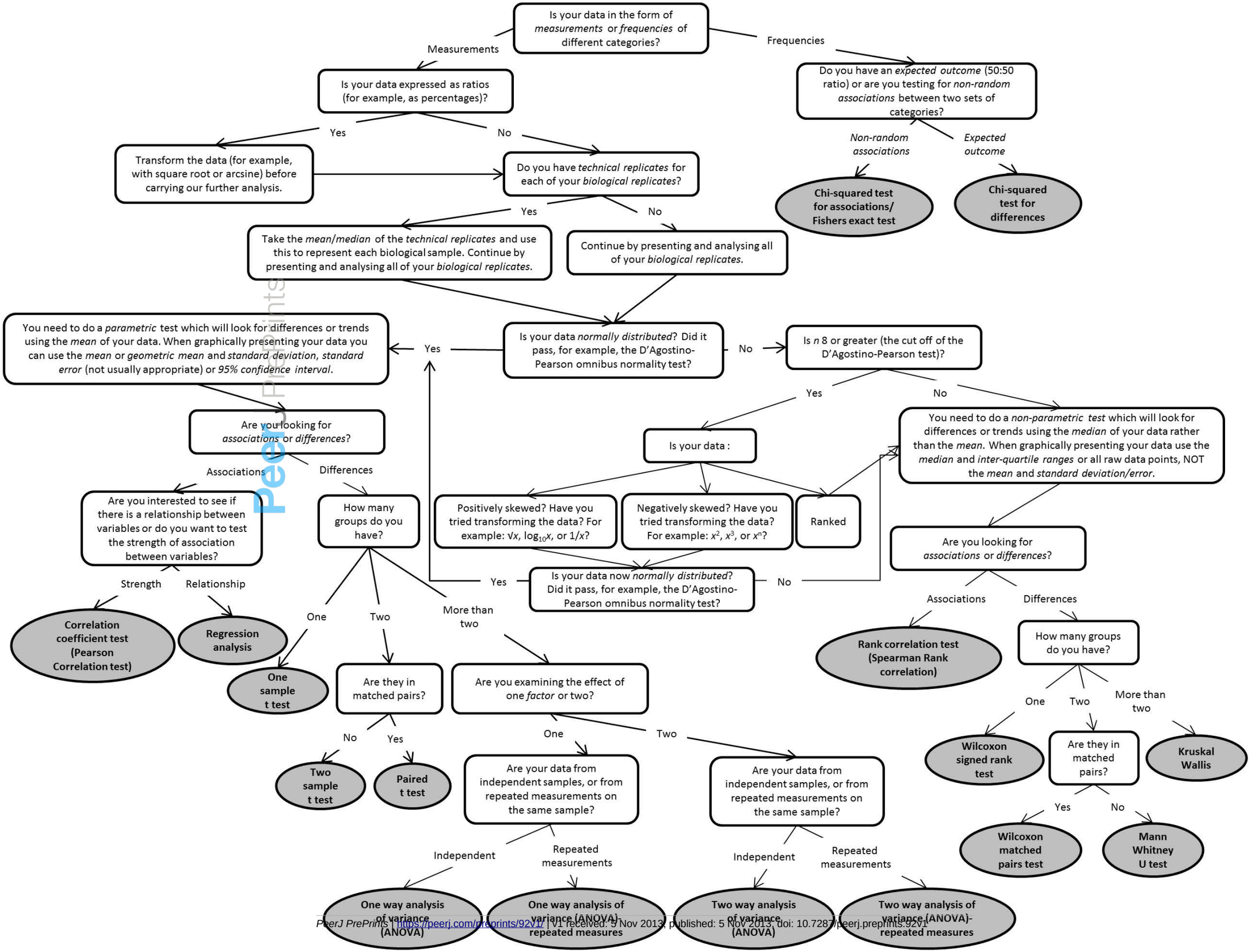
PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

denoted by dashed lines. Each symbol represents a sample taken from an individual animal. Data is pooled from two independent experiments with four animals per group per experiment.

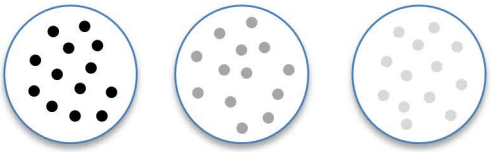
Figure 5. Expression of *geneX* as determined by qRT-PCR normalized to GAPDH expression and relative to an arbitrarily chosen calibrator sample within the saline-control group.

Saline-inoculated negative controls (controls) are compared with mice X hours after infection wild type (WT), Δ*mut1* or Δ*mut2* bacterial strains. Mean values per group are denoted by solid lines. Each symbol represents sample taken from an individual animal. Data is pooled from two independent experiments with four animals per group per experiment. Differences between groups were tested on $\log_{10}$-transformed data with One Way Analysis of Variance (ANOVA) and post-hoc t-tests using Bonferroni's correction for multiple comparisons. Probability values are shown for significant ($p<0.05$) differences.

PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1

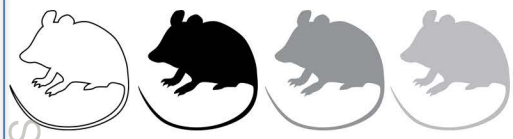


1. Preparation of bacterial inoculum



i. Bacterial cultures grown overnight on solid media.
ii. Numerous bacterial colonies scraped from into saline and diluted to give equivalent optical densities at 600nm.
iii. Each diluted preparation plated in triplicate for viable counts (as colony forming units [CFU]).

2. Infection of mice

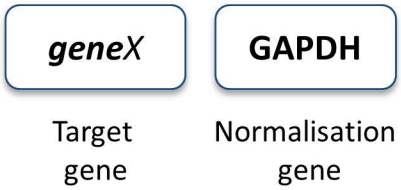


i. Each group of 4 mice is infected with one of 4 inocula – saline, WT bacteria, Δ*mut1* bacteria or Δ*mut2* bacteria.
ii. At a given time point, animals are euthanised and the appropriate organs harvested.
iii. Each organ is divided – one half for bacterial viable counts, the other for extraction of RNA.

3. Repeat steps 1 and 2

Infections repeated on two independent occasions to give n=8 animals per experimental condition

4. Quantify expression of *geneX*



i. Prepare cDNA (random primed) and control samples without RT enzyme.
ii. Run qRT PCR for *geneX* and normalisation gene
iii. From C_T data calculate relative expression values ($2^{-\Delta\Delta CT}$) for each experimental group of animals

A. Is your data in the form of *measurements* or *frequencies* of different categories?
= MEASUREMENTS

B. Is your data expressed as ratios (for example, as percentages)?
= YES

C. Data expressed as ratios or percentages are inherently non-continuous in distribution. Transform the data before carrying our further analysis.
= CONTINUE

D. Do you have *technical replicates* for each of your *biological replicates*?
= YES Take the *mean*/*median* of the *technical replicates* and use this to represent each biological sample. Continue by presenting and analysing all of your *biological replicates*.

E. Is your data *normally distributed*?
= NO

F. Is *n* equal to or greater than 8 (the cut off of the D'Agostino-Pearson test)?
= YES

G. Is your data positively skewed, negatively skewed or ranked?
= POSITIVELY SKEWED

H. Have you tried transforming the data? For example: $\sqrt{x}$, $\log_{10}x$, or $1/x$?
= CONTINUE

I. Is your data now *normally distributed*?
= YES

J. You need to do a *parametric* test which will look for differences or trends using the *mean* of your data. When graphically presenting your data you can use the *mean* or *geometric mean* and *standard deviation*, *standard error* (not usually appropriate) or *95% confidence interval*.
= CONTINUE

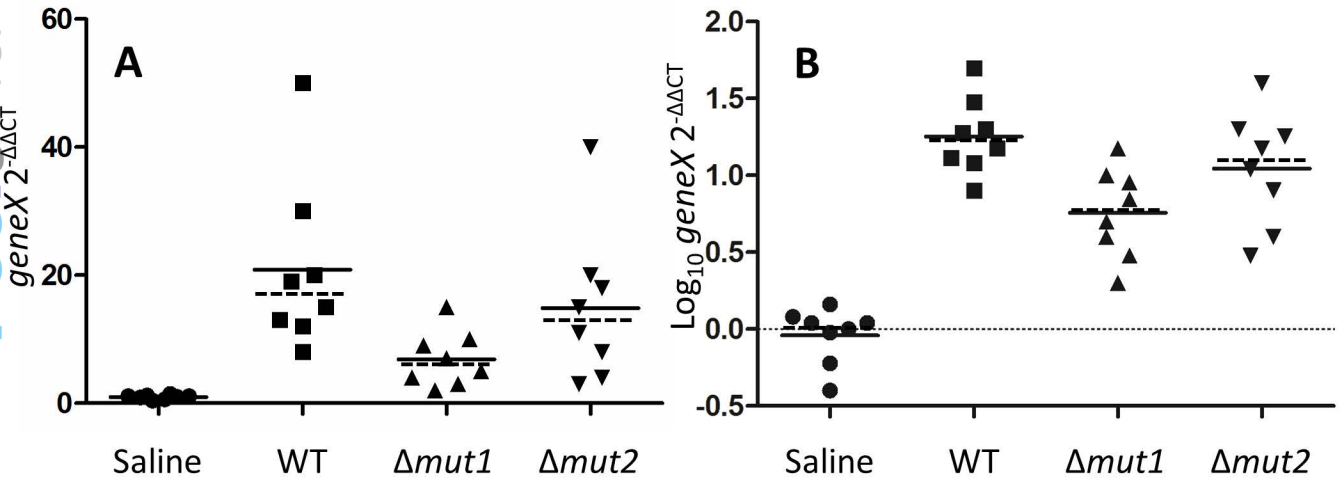
K. Are you looking for *differences* or *associations*?
= DIFFERENCES

L. How many groups do you have?
= MORE THAN TWO

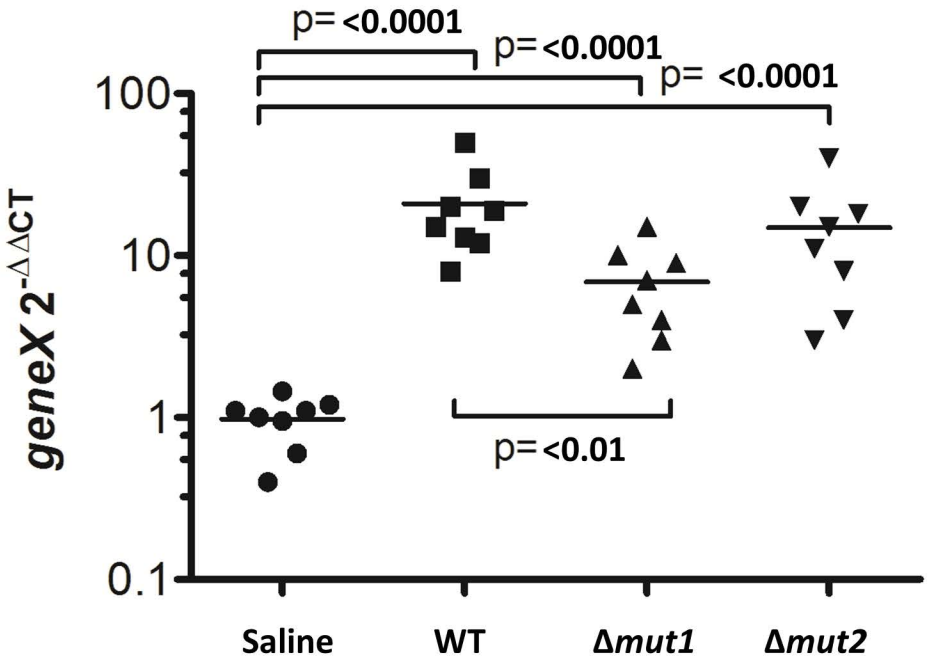
M. Are you examining the effect of one *factor* or two?
= ONE

N. Are your data from *independent* samples, or from *repeated measurements* on the same sample?
= INDEPENDENT SAMPLES

O. Test = One way Analysis of Variance (ANOVA)
This is a *parametric* test to determine if you have significant differences between the *means* of more than two un-matched independent groups.



PeerJ PrePrints | https://peerj.com/preprints/92v1/ | v1 received: 5 Nov 2013, published: 5 Nov 2013, doi: 10.7287/peerj.preprints.92v1



Mouse	Relative expression of *geneX* as $2^{-\Delta\Delta CT}$				Relative expression of *geneX* as $\log_{10} 2^{-\Delta\Delta CT}$			
	Saline-inoculated	WT-inoculated	Δ*mut1*-inoculated	Δ*mut2*-inoculated	Saline-inoculated	WT-inoculated	Δ*mut1*-inoculated	Δ*mut2*-inoculated
1	1.00	30.0	5.0	20.0	0.00	1.48	0.70	1.30
2	1.10	19.0	2.0	4.0	0.04	1.28	0.30	0.60
3	0.95	8.0	7.0	15.0	-0.02	0.90	0.85	1.18
4	0.40	20.0	4.0	3.0	-0.40	1.30	0.60	0.48
5	0.60	12.0	15.0	11.0	-0.22	1.08	1.18	1.04
6	1.20	15.0	3.0	40.0	0.08	1.18	0.48	1.60
7	1.10	50.0	9.0	8.0	0.04	1.70	0.95	0.90
8	1.45	13.0	10.0	18.0	0.16	1.11	1.00	1.26

	Untransformed data				Log_{10} transformed data			
	Saline-inoculated	WT-inoculated	Δ*mut1*-inoculated	Δ*mut2*-inoculated	Saline-inoculated	WT-inoculated	Δ*mut1*-inoculated	Δ*mut2*-inoculated
25% Percentile	0.6875	12.25	3.250	5.000	-0.1720	1.088	0.5084	0.6773
Median	**1.050**	**17.00**	**6.000**	**13.00**	**0.02070**	**1.227**	**0.7720**	**1.109**
75% Percentile	1.175	27.50	9.750	19.50	0.06973	1.433	0.9886	1.290
Mean	**0.9750**	**20.88**	**6.875**	**14.88**	**-0.03984**	**1.254**	**0.7568**	**1.045**
Standard Deviation	0.3338	13.51	4.324	11.89	0.1817	0.2478	0.2915	0.3734
Lower 95% CI of mean	0.6959	9.584	3.260	4.938	-0.1917	1.046	0.5131	0.7326
Upper 95% CI of mean	1.254	32.17	10.49	24.81	0.1121	1.461	1.001	1.357
P value for D'Agostino & Pearson omnibus normality test	0.6814	0.0182	0.4614	0.0467	0.1454	0.6528	0.8651	0.9094
Passed normality test (alpha=0.05)?	**Yes**	**No**	**Yes**	**No**	**Yes**	**Yes**	**Yes**	**Yes**

One-way analysis of variance	Log_{10} transformed data	Untransformed data
P value	< 0.0001	0.0010
P value summary	****	**
Are means significantly different? (P < 0.05)	Yes	Yes
Number of groups	4	4
F	32.25	7.164
R square	0.7756	0.4342

Groups	Mean Difference	t	Significant? P < 0.05?	**Summary**	95% CI of diff
Saline vs WT	-1.293	9.162	**Yes**	<0.0001	-1.694 to -0.8926
Saline vs Δ*mut1*	-0.7967	5.643	**Yes**	<0.0001	-1.197 to -0.3959
Saline vs Δ*mut2*	-1.085	7.683	**Yes**	<0.0001	-1.485 to -0.6838
WT vs Δ*mut1*	0.4967	3.518	**Yes**	<0.01	0.09591 to 0.8975
WT vs Δ*mut2*	0.2088	1.479	No	Not significant	-0.1920 to 0.6095
Δ*mut1* vs Δ*mut2*	-0.2879	2.040	No	Not significant	-0.6887 to 0.1128

Groups	Mean Difference	t	Significant? P < 0.05?	**Summary**	95% CI of diff
Saline vs WT	-19.90	4.301	**Yes**	<0.01	-33.03 to -6.765
Saline vs Δ*mut1*	-5.900	1.275	No	Not significant	-19.03 to 7.235
Saline vs Δ*mut2*	-13.90	3.004	**Yes**	<0.05	-27.03 to -0.7653
WT vs Δ*mut1*	14.00	3.026	**Yes**	<0.05	0.8653 to 27.13
WT vs Δ*mut2*	6.000	1.297	No	Not significant	-7.135 to 19.13
Δ*mut1* vs Δ*mut2*	-8.000	1.729	No	Not significant	-21.13 to 5.135