EMBO reports

science & society

Replicates and repeats—what is the difference and is it significant?

A brief discussion of statistics and experimental design

David L. Vaux, Fiona Fidler & Geoff Cumming

Science is knowledge gained through repeated experiment or observation. To be convincing, a scientific paper needs to provide evidence that the results are reproducible. This evidence might come from repeating the whole experiment independently several times, or from performing the experiment in such a way that independent data are obtained and a formal procedure of statistical inference can be applied—usually confidence intervals (CIs) or statistical significance testing. Over the past few years, many journals have strengthened their guidelines to authors and their editorial practices to ensure that error bars are described in figure legends—if error bars appear in the figures—and to set standards for the use of image-processing software. This has helped to improve the quality of images and reduce the number of papers with figures that show error bars but do not describe them. However, problems remain with how replicate and independently repeated data are described and interpreted. As biological experiments can be complicated, replicate measurements are often taken to monitor the performance of the experiment, but such replicates are not independent tests of the hypothesis, and so they cannot provide evidence of the reproducibility of the main results. In this article, we put forward our view to explain why data from replicates cannot be used to draw inferences about the validity of a hypothesis, and therefore should not be used to calculate CIs or *P* values, and should not be shown in figures.

…replicates are not independent tests of the hypothesis, and so they cannot provide evidence of the reproducibility of the main results

Let us suppose we are testing the hypothesis that the protein Biddelonin (BDL), encoded by the *Bdl* gene, is required for bone marrow colonies to grow in response to the cytokine HH-CSF. Luckily, we have wild-type (WT) and homozygous *Bdl* gene-deleted mice at our disposal, and a vial of recombinant HH-CSF. We prepare suspensions of bone marrow cells from a single WT and a single $Bdl^{-/-}$ mouse (same sex littermates from a $Bdl^{+/-}$ heterozygous cross) and count the cell suspensions by using a haemocytometer, adjusting them so that there are 1×10^5 cells per millilitre in the final solution of soft agar growth medium. We add 1 ml aliquots of the suspension to sets of ten 35 × 10 mm Petri dishes that each contain 10 µl of either saline or purified recombinant mouse HH-CSF.

We therefore put in the incubator four sets of ten soft agar cultures: one set of ten plates has WT bone marrow cells with saline; the second has $Bdl^{-/-}$ cells with saline; the third has WT cells with HH-CSF, and the fourth has $Bdl^{-/-}$ cells with HH-CSF. After a week, we remove the plates from the incubator and count the number of colonies (groups of >50 cells) in each plate by using a dissecting microscope. The number of colonies counted is shown in Table 1.

We could plot the counts of the plates on a graph. If we plotted just the colony counts of only one plate of each type (Fig 1A shows the data for plate 1), it seems clear that HH-CSF is necessary for many colonies to form, but it is not immediately apparent whether the response of the $Bdl^{-/-}$ cells is significantly different to that of the WT cells. Furthermore, the graph does not look 'sciency' enough; there are no error bars or *P*-values. Besides, by showing the data for only one plate we are breaking the fundamental rule of science that all relevant data should be reported and subjected to analysis, unless good reasons can be given why some data should be omitted.

To make it look better, we could add the mean numbers of colonies in the first

Table 1 | Bone marrow colonies per plate

	Plate number									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
WT + saline	0	0	0	1	1	0	0	0	0	0
$Bdl^{-/-}$ + saline	0	0	0	0	0	1	0	0	0	2
WT + HH-CSF	61	59	55	64	57	69	63	51	61	61
$Bdl^{-/-}$ + HH-CSF	48	34	50	59	37	46	44	39	51	47

1×10^5 WT or $Bdl^{-/-}$ bone marrow cells were plated in 1 ml soft agar cultures in the presence or absence of 1 µM HH-CSF. Colonies per plate were counted after 1 week. WT, wild type.

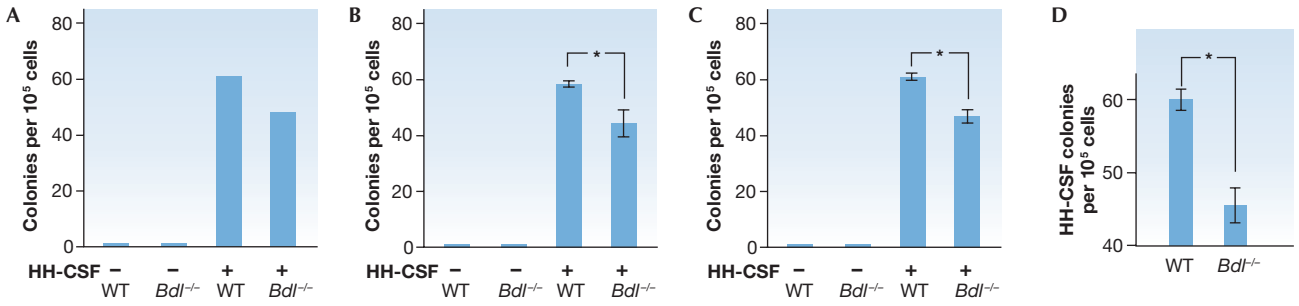

Fig 1 | Displaying data from replicates—what not to do. (**A**) Data for plate 1 only (shown in Table 1). (**B**) Means ± SE for replicate plates 1–3 (in Table 1), $*P > 0.05$. (**C**) Means ± SE for replicate plates 1–10 (in Table 1), $*P < 0.0001$. (**D**) Means ± SE for HH-CSF-treated replicate plates 1–10 (in Table 1). Statistics should not be shown for replicates because they merely indicate the fidelity with which the replicates were made, and have no bearing on the hypothesis being tested. In each of these figures, $n = 1$ and the size of the error bars in (**B**), (**C**) and (**D**) reflect sampling variation of the replicates. The SDs of the replicates would be expected to be roughly the square root of the mean number of colonies. Also, axes should commence at 0, other than in exceptional circumstances, such as for log scales. SD, standard deviation; SE, standard error.

three plates of each type to the graph (Fig 1B), with error bars that report the standard error (SE) of the three values of each type. Now it is looking more like a figure in a high-profile journal, but when we use the data from the three replicate plates of each type to assess the statistical significance of the difference in the responses of the WT and *Bdl*$^{-/-}$ cells to HH-CSF, we find $P > 0.05$, indicating they are not significantly different.

As we have another seven plates from each group, we can plot the means and SEs of all ten plates and re-calculate *P* (Fig 1C). Now we are delighted to find that there is a highly significant difference between the *Bdl*$^{-/-}$ and WT cells, with $P < 0.0001$.

However, although the differences are highly statistically significant, the heights of the columns are not dramatically different, and it is hard to see the error bars. To remedy this, we could simply start the *y*-axis at 40 rather than zero (Fig 1D), to emphasize the differences in the response to HH-CSF. Although this necessitates removing the saline controls, these are not as important as visual impact for high-profile journals.

With a small amount of effort, and no additional experiments, we have transformed an unimpressive result (Fig 1A,B) into one that gives strong support to our hypothesis that BDL is required for a response to HH-CSF, with a highly significant *P*-value, and a figure (Fig 1D) that looks like it could belong in one of the top journals.

Sidebar A | Fundamental principles of statistical design

Fundamental principle 1
Science is knowledge obtained by repeated experiment or observation: if $n = 1$, it is not science, as it has not been shown to be reproducible. You need a random sample of independent measurements.

Fundamental principle 2
Experimental design, at its simplest, is the art of varying one factor at a time while controlling others: an observed difference between two conditions can only be attributed to Factor A if that is the only factor differing between the two conditions. We always need to consider plausible alternative interpretations of an observed result. The differences observed in Fig 1 might only reflect differences between the two suspensions, or be due to some other (of the many) differences between the two individual mice, besides the particular genotypes of interest.

Fundamental principle 3
A conclusion can only apply to the population from which you took the random sample of independent measurements: so if we have multiple measures on a single suspension from one individual mouse, we can only draw a conclusion about that particular suspension from that particular mouse. If we have multiple measures of the activity of a single vial of cytokine, then we can only generalize our conclusion to that vial.

Fundamental principle 4
Although replicates cannot support inference on the main experimental questions, they do provide important quality controls of the conduct of experiments. Values from an outlying replicate can be omitted if a convincing explanation is found, although repeating part or all of the experiment is a safer strategy. Results from an independent sample, however, can only be left out in exceptional circumstances, and only if there are especially compelling reasons to justify doing so.

So, what is wrong? The first problem is that our data do not confirm the hypothesis that BDL is required for bone marrow colonies to grow in response to HH-CSF, they actually refute it. Clearly, bone marrow colonies are growing in the absence of BDL, even if the number is not as great as when the *Bdl* genes are intact. Terms such as 'required', 'essential' and 'obligatory' are not relative, yet are still often incorrectly used when partial effects are seen. At the very least, we should reformulate our hypothesis, perhaps to "BDL is needed for a full response of bone marrow colony-forming cells to the cytokine HH-CSF".

The second major problem is that the calculations of *P* and statistical significance are based on the SE of replicates, but the ten replicates in any of the four conditions were each made from a single suspension of bone marrow cells from just one mouse. As such, we can at best infer a statistically significant difference between the concentration of colony-forming cells in the bone marrow cell suspension from that particular WT mouse and the bone marrow suspension from that particular gene-deleted mouse. We have made just one comparison, so $n = 1$, no matter how many replicate plates we count. To make an inference that can be generalized to all WT mice and *Bdl*$^{-/-}$ mice, we need to repeat our experiments a number of times, making several independent comparisons using several mice of each type.

…by showing the data for only one plate we are breaking the fundamental rule of science that all relevant data should be reported and subjected to analysis…

Rather than providing independent data, the results from the replicate plates are linked because they all came from the same suspension of bone marrow cells. For example, if we made any error in determining the concentration of bone marrow cells, this error would be systematically applied to all of the plates. In this case, we determined the initial number of bone marrow cells by performing a cell count using a haemocytometer, a method that typically only gives an accuracy of ±10%. Therefore, no matter how many plates are counted, or how small the error bars are in Fig 1, it is not valid to conclude that there is a difference between the WT and *Bdl*$^{-/-}$ cells. Moreover, even if we had used a flow cytometer to sort exactly the same number of bone marrow cells into each of the plates, we would still have only tested cells from a single *Bdl*$^{-/-}$ mouse, so *n* would still equal 1 (see Fundamental principle 1 in Sidebar A).

To be convincing, a scientific paper describing a new finding needs to provide evidence that the results are reproducible. While it might be argued that a hypothetical talking dog would represent an important scientific discovery even if $n = 1$, few people would be convinced if someone claimed to have a talking dog that had been observed on one occasion to speak a single word. Most people would require several words to be spoken, with a number of independent observers, on several occasions. The cloning of Dolly the sheep represented a scientific breakthrough, but she was one of five cloned sheep described by Campbell *et al* [1]. Eight fetuses and sheep were typed by microsatellite analysis and shown to be identical to the cell line used to provide the donor nuclei.

Inferences can only be made about the population from which the independent samples were drawn. In our original experiment, we took individual replicate aliquots from the suspensions of bone marrow cells (Fig 2A). We can therefore only generalize our conclusions to the 'population' from which our sample aliquots came: in this case the population is that particular suspension of bone marrow cells. To test our hypothesis, it is necessary to carry out an experiment similar to that shown in Fig 2B. Here, bone marrow has been independently isolated from a random sample of WT mice and another random sample of *Bdl*$^{-/-}$ mice. In this case, we can draw conclusions about *Bdl*$^{-/-}$ mice in general, and compare them withWT mice (in general). In Fig 2A, the number of *Bdl*$^{-/-}$ mice that have been compared with WT mice (which is the comparison relevant to our hypothesis) is one, so $n = 1$, regardless of how many replicate plates are counted. Conversely, in Fig 2B we are comparing three *Bdl*$^{-/-}$ mice with WT controls, so $n = 3$, whether we plate three replicate plates of each type or 30. Note, however, that it is highly desirable for statistical reasons to have samples larger than $n = 3$, and/or to test the hypothesis by some other approach, for example, by using antibodies that block HH-CSF or BDL, or by

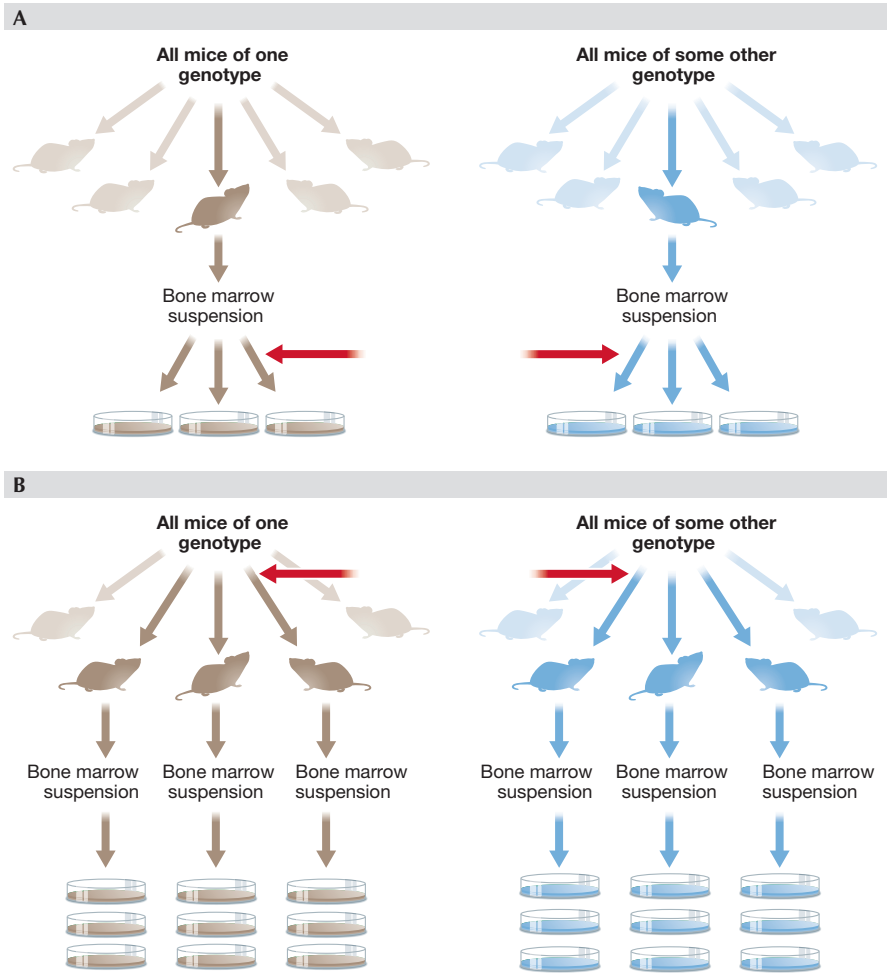

Fig 2 | Sample variation. Variation between samples can be used to make inferences about the population from which the independent samples were drawn (red arrows). For replicates, as in (**A**), inferences can only be made about the bone marrow suspensions from which the aliquots were taken. In (**A**), we might be able to infer that the plates on the left and the right contained cells from different suspensions, and possibly that the bone marrow cells came from two different mice, but we cannot make any conclusions about the effects of the different genotypes of the mice. In (**B**), three independent mice were chosen from each genotype, so we can make inferences about all mice of that genotype. Note that in the experiments in (**B**), $n = 3$, no matter how many replicate plates are created.

To be convincing, a scientific paper needs to provide evidence that the results are reproducible

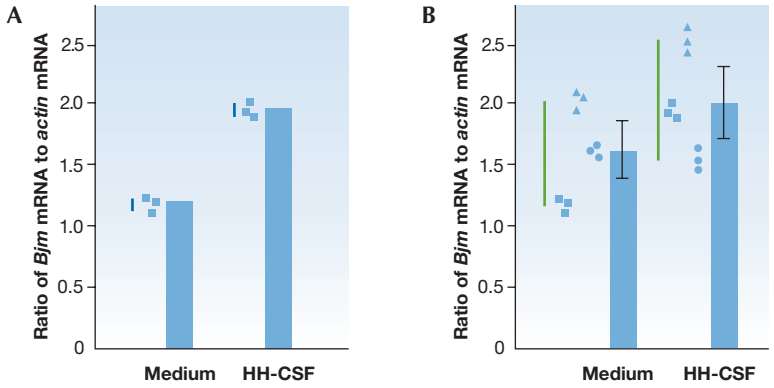

Fig 3 | Means of replicates compared with means of independent samples. (**A**) The ratios of the three-replicate *Bjm* PCR reactions to the three-replicate *Actin* PCR reactions from the six aliquots of RNA from one culture of HH-CSF-stimulated cells and one culture of unstimulated cells are shown (filled squares). The means of the ratios are shown as columns. The close correlation of the three replicate values (blue lines) indicates that the replicates were created with high fidelity and the pipetting was consistent, but is not relevant to the hypothesis being tested. It is not appropriate to show *P*-values here, because $n = 1$. (**B**) The ratios of the replicate PCR reactions using mRNA from the other cultures (two unstimulated, and two treated with HH-CSF) are shown as triangles and circles. Note how the correlation between the replicates (that is, the groups of three shapes) is much greater than the correlation between the mean values for the three independent untreated cultures and the three independent HH-CSF-treated cultures (green lines). Error bars indicate SE of the ratios from the three independent cultures, not the replicates for any single culture. $P > 0.05$. SE, standard error.

re-expressing a *Bdl* cDNA in the *Bdl*$^{-/-}$ cells (see Fundamental principle 2 in Sidebar A).

One of the most commonly used methods to determine the abundance of mRNA is real-time quantitative reverse transcription PCR (qRT-PCR; although the following example applies equally well to an ELISA or similar). Typically, multi-well plates are used so that many samples can be simultaneously read in a PCR machine. Let us suppose we are going to use qRT-PCR to compare levels of *Boojum* mRNA (*Bjm*) in control bone marrow cells (treated with medium alone) with *Bjm* levels in bone marrow cells treated with HH-CSF, in order to test the hypothesis that HH-CSF induces expression of the *Bjm* gene.

We isolate bone marrow cells from a normal mouse, and dispense equal aliquots containing a million cells into each of two wells of a six-well plate. For the moment we use only two of the six wells. We then add 4 ml of plain medium to one of the wells (the control), and 4 ml of a mixture of medium supplemented with HH-CSF to the other well (the experimental well). We incubate the plate for 24 h and then transfer the cells into two tubes, in which we extract the RNA using TRizol. We then suspend the RNA in 50 µl TRIS-buffered RNAse-free water.

We put 10 µl from each tube into each of two fresh tubes, so that both *Actin* (as a control) and *Bjm* message can be determined in each sample. We now have four tubes, each with 10 µl of mRNA solution. We make two sets of 'reaction mix' with the only difference being that one contains *Actin* PCR primers and the other *Bjm* primers. We add 40 µl of one or the other 'reaction mix' to each of the four tubes, so we now have 50 µl in each tube. After mixing, we take three aliquots of 10 µl from each of the four tubes and put them into three wells of a 384-well plate, so that 12 wells in total contain the RT-PCR mix. We then put the plate into the thermocycler. After an hour, we get an Excel spreadsheet of results.

…should we dispense with replicates altogether? The answer, of course, is 'no'. Replicates serve as internal quality checks on how the experiment was performed

We then calculate the ratio of the *Bjm* signal to the *Actin* signal for each of the three pairs of reactions that contained RNA from the HH-CSF-treated cells, and for each of the three pairs of control reactions.

Replicates […] cannot be used to infer conclusions

In this case, the variation among the three replicates will not be affected by sampling error (which was what caused most of the variation in colony number in the earlier bone marrow colony-forming assay), but will only reflect the fidelity with which the replicates were made, and perhaps some variation in the heating of the separate wells in the PCR machine. The three 10 µl aliquots each came from the same, single, mRNA preparation, so we can only make inferences about the contents of that particular tube. As in the previous example, in this case n still equals 1, and no inferences about the main experimental hypothesis can be made. The same would be true if each RNA sample were analysed in 10 or 100 wells; we are only comparing one control sample to one experimental sample, so $n = 1$ (Fig 3A). To draw a general inference about the effect of HH-CSF on *Bjm* expression, we would have to perform the experiment on several independent samples derived from independent cultures of HH-CSF-stimulated bone marrow cells (Fig 3B).

For example, we could have put the bone marrow cells in all six wells of the tissue culture plate, and performed three independent cultures with HH-CSF, and three independent control cultures in medium without HH-CSF. mRNA could then have been extracted from the six cultures, and each split into six wells to measure *Actin* and *Bjm* mRNA levels by using qRT-PCR. In this case, 36 wells would have been read by the machine. If the experiment were performed this way, then $n = 3$, as there were three independent control cultures, and three independent HH-CSF-dependent cultures, that were testing our hypothesis that HH-CSF induces *Bjm* expression. We then might be able to generalize our conclusions about the effect of that vial of recombinant HH-CSF on expression of *Bjm* mRNA. However, in this case (Fig 3B) $P > 0.05$, so we cannot exclude the possibility that the differences observed were just due to chance, and that HH-CSF has no effect on *Bjm* mRNA expression. Note that we also cannot conclude that it has no effect; if $P > 0.05$, the only conclusion we can make is that we cannot make any conclusions. Had we calculated and shown errors and *P*-values for

replicates in Fig 3A, we might have incorrectly concluded, and perhaps misled the readers to conclude that there was a statistically significant effect of HH-CSF in stimulating *Bjm* transcription (see Fundamental principle 3 in Sidebar A).

Why bother with replicates at all? In the previous sections we have seen that replicates do not allow inferences to be made, or allow us to draw conclusions relevant to the hypothesis we are testing. So should we dispense with replicates altogether? The answer, of course, is 'no'. Replicates serve as internal quality checks on how the experiment was performed. If, for example, in the experiment described in Table 1 and Fig 1, one of the replicate plates with saline-treated WT bone marrow contained 100 colonies, you would immediately suspect that something was wrong. You could check the plate to see if it had been mislabelled. You might look at the colonies using a microscope and discover that they are actually contaminating colonies of yeast. Had you not made any replicates, it is possible you would not have realized that a mistake had occurred.

Fig 4 shows the results of the same qRT-PCR experiment as in Fig 3, but in this case, for one of the sets of triplicate PCR ratios there is much more variation than in the others. Furthermore, this large variation can be accounted for by just one value of the three replicates—that is, the uppermost circle in the graph. If you had results such as those in Fig 4A, you would look at the individual values for the *Actin* PCR and *Bjm* PCR for the replicate that had the strange result. If the *Bjm* PCR sample was unusually high, you could check the corresponding well in the PCR plate to see if it had the same volume as the other wells. Conversely, if the *Actin* PCR value was much lower than those for the other two replicates, on checking the well in the plate you might find that the volume was too low. Alternatively, the unusual results might have been due to accidentally adding two aliquots of RNA, or two of PCR primer-reaction mix. Or perhaps the pipette tip came loose, or there were crystals obscuring the optics, or the pipette had been blocked by some debris, etc., etc., etc. Replicates can thus alert you to aberrant results, so that you know when to look further and when to repeat the experiment. Replicates can act as an internal check of the fidelity with which the experiment was performed. They can alert you to problems with plumbing, leaks, optics, contamination, suspensions, mixing or mix-ups. But they cannot be used to infer conclusions.

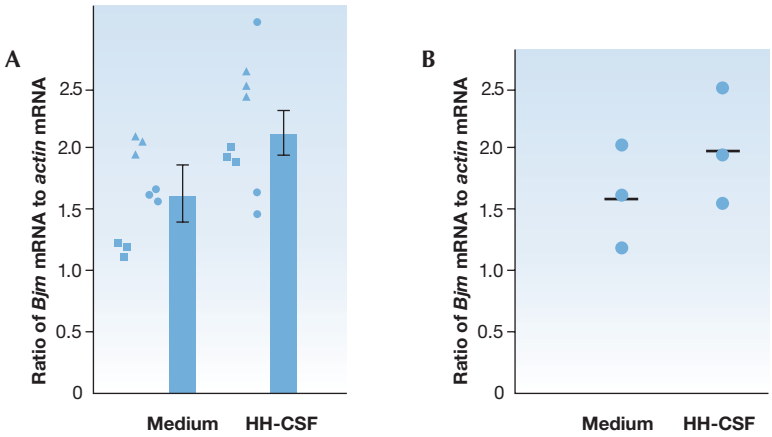

Fig 4 | Interpreting data from replicates. (**A**) Mean ± SE of three independent cultures each with ratios from triplicate PCR measurements. $P > 0.05$. This experiment is much like the one in Fig 3B. However, notice in this case, for one of the sets of replicates (the circles from one of the HH-CSF-treated replicate values), there is a much greater range than for the other five sets of triplicate values. Because replicates are carefully designed to be as similar to each other as possible, finding unexpected variation should prompt an investigation into what went wrong during the conduct of the experiment. Note how in this case, an increase in variation among one set of replicates causes a decrease in the SEs for the values for the independent HH-CSF results: the SE bars for the HH-CSF condition are shorter in Fig 4A than in Fig 3B. Failure to take note of abnormal variation in replicates can lead to incorrect statistical inferences. (**B**) *Bjm* mRNA levels (relative to *Actin*) for three independent cultures each with ratios from triplicate PCR measurements. Means are shown by a horizontal line. The data here are the same as those for Fig 3B or Fig 4A with the aberrant value deleted. When *n* is as small as 3, it is better to just plot the data points, rather than showing statistics. SE, standard error.

Because replicate values are not relevant to the hypothesis being tested, they—and statistics derived from them—should not be shown in figures. In Fig 4B, the large dots show the means of the replicate values in Fig 4A, after the aberrant replicate value has been excluded. While in this figure you could plot the means and SEs of the mRNA results from the three independent medium- and HH-CSF-treated cultures, in this case, the independent values are plotted and no error bars are shown. When the number of independent data points is low, and they can easily be seen when plotted on the graph, we recommend simply doing this, rather than showing means and error bars.

What should we look for when reading papers? Although replicates can be a valuable internal control to monitor the performance of your experiments, there is no point in showing them in the figures in publications because the statistics from replicates

Sidebar B | Error checklist when reading papers

(i) If error bars are shown, are they described in the legend?
(ii) If statistics or error bars are shown, is *n* stated?
(iii) If the standard deviations (SDs) are less than 10%, do the results come from replicates?
(iv) If the SDs of a binomial distribution are consistently less than $\sqrt{(np(1-p))}$—where *n* is sample size and *P* is the probability—are the data too good to be true?
(v) If the SDs of a Poisson distribution are consistently less than $\sqrt{(\text{mean})}$, are the data too good to be true?
(vi) If the statistics come from replicates, or from a single 'representative' experiment, consider whether the experiments offer strong support for the conclusions.
(vii) If *P*-values are shown for replicates or a single 'representative' experiment, consider whether the experiments offer strong support for the conclusions.

…if statistics, error bars and *P*-values for replicates are shown, they can mislead the readers of a paper who assume that they are relevant to the paper's conclusions

are not relevant to the hypothesis being tested. Indeed, if statistics, error bars and *P*-values for replicates are shown, they can mislead the readers of a paper who assume that they are relevant to the paper's conclusions. The corollary of this is that if you are reading a paper and see a figure in which the error bars—whether standard deviation, SE or CI—are unusually small, it might alert you that they come from replicates rather than independent samples. You should carefully scrutinize the figure legend to determine whether the statistics come from replicates or independent experiments. If the legend does not state what the error bars are, what *n* is, or whether the results come from replicates or independent samples, ask yourself whether these omissions undermine the paper, or whether some knowledge can still be gained from reading it.

You should also be sceptical if the figure contains data from only a single experiment with statistics for replicates, because in this case, $n = 1$, and no valid conclusions can be made, even if the authors state that the results were 'representative'—if the authors had more data, they should have included them in the published results (see Sidebar B for a checklist of what to look for). If you wish to see more examples of what not to do, search the Internet for the phrases 'SD of one representative', 'SE of one representative', 'SEM of one representative', 'SD of replicates' or 'SEM of replicates'.

ACKNOWLEDGEMENTS

This work was made possible through Victorian State Government Operational Infrastructure Support, and Australian Government NHMRC IRIISS and NHMRC grants 461221and 433063.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCE

1. Campbell KH, McWhir J, Ritchie WA, Wilmut I (1996) Sheep cloned by nuclear transfer from a cultured cell line. *Nature* **380:** 64–66

EMBO *reports* (2012) **13,** 291–296; published online 16 March 2012; doi:10.1038/embor.2012.36

David L. Vaux [top left] is at The Walter and Eliza Hall Institute and the Department of Experimental Biology, University of Melbourne, Melbourne, Australia.
E-mail: vaux@wehi.edu.au
Fiona Fidler and Geoff Cumming are at La Trobe University School of Psychological Science, Melbourne, Australia.
E-mails: f.fidler@latrobe.edu.au; g.cumming@latrobe.edu.au