EDITORIAL

nature cell biology

Reproducing data

Traditionally, the scientific process has been based on prepublication screening through peer review and postpublication validation by independent reproduction. Of course all data published in reputable journals must be reproducible within the original lab as a condition of publication. However, before a finding can attain 'dogma' status, it must pass the independent reproducibility test.

When publication pressures were more manageable, and before the old adage 'publish or perish' emerged as a primary driving force in the molecular biosciences, reproducibility was still a key step in the scientific process. Scientists at the beginning of their academic career often learned their craft by reproducing published data — a bit like teaching art by copying the great masters.

These days, scientists request reagents from each other more than ever; however, the primary aim is not to reproduce, but to move to the next step. It is exceedingly difficult to convince a postdoc to spend months reproducing a complex set of experiments when the outcome is either an unpublishable confirmation, or a lack of confirmation, which would require much more work to ensure that the case made is watertight and often result in the publication of an abbreviated refutation (see editorials, May and November 2005). The PI will worry about the significant drain on resources that a rock solid refutation requires, and the drain on morale that may result from a protracted fight for acceptance of negative data by the original author and the broader community.

Consequently, competitive labs are not often motivated to reproduce data; more importantly, it is not something they are encouraged to do. One way to address this would be allocate a percentage of the time of each lab and researcher solely for independent data confirmation. Granting agencies should take these endeavours seriously and give credit for documented evidence of data reproduction. Initiating an online repository for this data would also be worthwhile and the confirmatory nature of the data may allow for curation without full-blown peer review.

Nowadays, somewhat ironically, most data reproduction occurs through related studies published in a similar timeframe. Increased competitiveness has meant that many researchers prefer the relatively safe option of working on predictable projects that are likely to result in publications, in favour of more esoteric research. The result is increased redundancy as everyone jumps on the next obvious question (occasionally primed by exposure to unpublished data). Copublication can seem frustratingly redundant and is certainly not the most efficient way to spend limited resources. However, in the current system, parallel research is increasingly the favoured means of data validation and in the absence of grass roots policy changes, we may have to live with it.

Nothing to hide

Many readers will recall frustration on reaching the last paragraphs of an exciting paper only to stumble across an anticlimactic referral to enticing 'unpublished data' or 'data not shown' in the discussion section. Individual papers can only achieve conceptual advances of finite size and a good study will raise as many questions as it answers. We like when the discussion of a paper clearly outlines some of the next key questions to be addressed, and realize that these questions are often already being addressed. However, adding 'data not shown' to key implications can come across as somewhat disingenuous: if the data adds to the conceptual advance presented in the published paper, it should be included in the paper and formally peer reviewed. Conversely, if it is not essential, why mention it at all? The reader is left feeling that the authors are trying to stake a claim on a topic without having data that would hold up to the scrutiny of their peers. It is much better to just flavour the discussion with the next big questions — even if the work is already in progress.

An altogether different category of 'data not shown' relates to the core data of the paper; for example, data for one cell line is presented and the supporting data from four others is 'not shown'. Why, especially when many journals host online supplementary information? If the data warrants being mentioned to support the conclusions made, then it should usually be presented to back up the claim. How else can the reader judge if the generality of the conclusions hold, or if the authors have viewed their supporting data through rose tinted glasses. Even worse is when controls (such as for antibody specificity) are 'not shown'. This is as unacceptable as the excuse 'your format restrictions are too tight'.

This leaves data that is submitted for the eyes of peer reviewers and editors only. Again, we would discourage this approach. If a referee requests unnecessary or irrelevant data, surely the author should argue their case, rather than trying to appease an off-base referee (unless the data happens to be available). If the data is important, but part of another planned study, the author should publish the other study first or 'bite the bullet' and transfer the data. Our readers have a right to judge the same data set that was evaluated by our referees.

Finally, a note on 'personal communications'; assuming these are attributed and permission has been obtained, these have to be admissible, but only where they underscore further reaching statements in the discussion, or as an entry point into a study — not to bolster core data in the paper.

Online publishing should have made 'data not shown' largely a thing of the past and we encourage our authors to show what is necessary to present a definitive and well rounded piece of work of the sort we hope you expect to find in this journal.