Editorial

THE EMBO JOURNAL

Reproducibility blues

Bernd Pulverer

Research findings advance science only if they are significant, reliable and reproducible. Scientists and journals must publish robust data in a way that renders it optimally reproducible. Reproducibility has to be incentivized and supported by the research infrastructure but without dampening innovation.

Why all the fuss?

All research builds on the preceding literature—knowledge advances by sharing ideas, findings and tools. The purpose of scientific communication and publishing is to share and archive information that accurately describes nature—ideally generalizable paradigms. The data has to be robust and significant, and the methods by which it was obtained have to be described in sufficient detail to allow others to build on the work. The scientific literature has grown enormously, fuelled by investment but also research assessment. The information torrent carries with it too much unreliable research that is only partially cleared over time. Prominent research papers that are unreliable can be toxic—they waste resources and can mislead scientists and the public.

More reliable quality assurance mechanisms must be implemented to filter the unprecedented volumes of research output. Papers must include rich data, less ambiguous descriptions of materials and methods, and improved linking to related content. Reproducibility is a central tenet of the scientific process. We must incentivize the sharing of reproducible research, not restrict reward criteria to citation rates. At the same time, we must be careful to define what level of reproducibility we actually strive for. The ongoing discussion about reproducibility is important, but crying wolf for the wrong reasons risks undermining a research environment that is yielding an unprecedented rate of discovery.

Reproducible papers, reproducible data, reproducible conclusions

Reproducibility is the topic du jour—but commentators often fail to define precisely what they mean. At the basic level, we need to ensure that the way we report science is *in principle* reproducible. At a higher level, we expect the *conclusions* of a body of work to be reproducible—that is, the biological insight to be factually correct and generalizable. At a more specific level, we expect to be able to *replicate* the specific experiments reported to yield consistent data.

> *"If a methods section offers little more detail than the average recipe book, is it surprising that the reproducibility of this paper is not readily demonstrated?"*

The research paper in many journals still largely reflects the one-dimensional world of print—editorial guidelines and author habits conspire to yield minimal information to support elaborate conclusions. Data, its description and methods are readily sacrificed to meet editorial space and word limits. As a result, vestigial materials and methods sections abound that miss important information or that cite previous methods with equally imprecise details. Compact writing is important to navigate the information wall facing us, but we should not save space in the wrong places: at least for online publication, journals ought to exclude methods sections from format restrictions (and, for that matter, arbitrary limits for references). If a methods section offers little more detail than the average recipe book, is it surprising that the reproducibility of this paper is not readily demonstrated? After all, two cooks will invariably create widely different meals from the same recipe depending on ambiguities in the description, different sources of ingredients, their experience and predilections.

Shaken, not stirred

However, extensive methods alone will not guarantee reproducibility. Experiments are finicky—we all have a favourite story of how long it took to nail the tiny changed variable that rendered an experiment temporarily unreproducible. Sometimes the variable turns out to be tap water, humidity or the lunar cycle, sometimes it is never identified. For example, an experimental discrepancy between the Bissell and Polyak labs turned out to be down to the agitation of cell cultures, uncovered after one year of trials and tribulations (Hines *et al*, 2015). In this respect, it can be useful to enhance methods with video based protocols which can capture variables that are hard to describe in words, as pioneered by the journal *JoVE*. Much more effort has to be invested in using validated reagents and in identifying them unambiguously. The same goes for hardware and data acquisition. It has been known for years that many antibodies and small molecule reagents do not—or do not only—detect what it says on the box, yet antibody validation efforts such as antibodypedia are still not routine (Björling & Uhlén, 2008; Baker, 2015). We have known since the 1960s that many cell lines in common use are contaminated or mislabeled (Yu *et al*, 2015), yet authentication is not a standard publishing requirement. A precondition for reproducibility is to capture as many experimental variables as precisely as possible.

Statistics

The confidence in the reproducibility and generalizability of research findings ought to

EMBO, Heidelberg, Germany. E-mail: bernd.pulverer@embojournal.org
DOI 10.15252/embj.201570090

be assessed by statistical tests. Biological data reflects complex systems that are inherently noisy, yet the cost and time investment often limits how often an experiment is reproduced before publication of generalized conclusions. Statistical rigour and formally unbiased data acquisition and reporting are not always a strength in molecular cell biology—for example blinded experimentation to limit cognitive bias remains rare. It is all but trivial to evaluate objectively when one can be sure an experimental finding has a sufficiently high probability of being reproducible to warrant publication and researchers tend to rely on instinct, well honed as that may be.

At this journal, we have the policy to present only appropriate statistics and only when this is warranted—it can be completely justified to present data without statistics, but not to present data with inappropriate statistics. We encourage the presentation in figures of data points alongside any statistics where this helps interpret the data; more importantly, we encourage the posting of source data alongside figures for replicates (Pulverer, 2014).

In this issue, we launch a new series on statistics—in a series of articles, biostatistician Bernd Klaus provides hands on advice tailored to the molecular cell biologist (Klaus, 2015).

Reproducibility by orthogonal experimentation

A paper may well produce reliable and generalizable conclusions, while replicating the exact experiments reported may be hard, even if the experiments are described in some detail. One crucial test is whether orthogonal experimentation comes to similar conclusions. In fact, ensuring that different experimental approaches and experiments in unrelated systems yield the same conclusions remains a gold standard for reporting reliable, reproducible biological insight. We need to avoid overreaching in our goals for a reliable literature and very clearly differentiate between robust conclusions and robust experimental data.

Correction, Retraction, Withdrawal

Branding papers publically as unreproducible is only warranted when all means of sharing information and reagents have been exhausted in the attempts to replicate experiments—sometimes this has to involve researchers travelling to the laboratory first reporting a result. If an experiment cannot be reproduced in the source lab, with the same infrastructure and expertise, it certainly has to be classed as unreproducible and the authors of the study have an obligation to correct the literature without delay. Depending on the severity of the problem, a correction or a retraction of the paper—or minimally the affected experimental data—is appropriate (Pulverer, 2015b). A retraction is a daunting prospect and it would, in our view, be advisable to distinguish between author-initiated removal of a paper due to a lack of reproducibility or experimental flaws that were unpredictable, and cases of removal due infringement of research ethics or scientific integrity. We suggest to apply the terms *withdrawal* and *retraction*, respectively.

Often, however, while the reproducibility of findings or the generality of claims are in question and the evidence is sufficiently definitive to warrant informing the community of the issue, the criteria for revoking publication are not met. In such cases, peer reviewed comments on papers by named individuals presenting concrete evidence are recommended; this should naturally include a response by the original authors where appropriate.

As I have argued previously, there needs to be a cultural shift to embrace correction of the literature as a core part of the scientific process (Pulverer, 2015b). Scientific knowledge evolves and we need to view the literature in a more fluid manner that can be corrected and reversed, otherwise we leave trails of misinformation that at minimum reduce research efficiency. Academic assessment structures must reward investment to reproduce important research findings and such data should be published in a visible forum.

Industrializing reproducibility

The issue of reproducibility in the biosciences came into the limelight in the wake of recent commentaries from reputable researchers in the biotech and pharmaceutical industries that claimed remarkably low rates of reproducibility for notable research papers (Prinz *et al*, 2011; Begley & Ellis 2012). While these commentaries were not supported by actual data, the conclusions are certainly echoed by others in industry and beyond. For example, a recent survey by ASCB reported 72% of respondents "had trouble replicating another lab's published results" and 40% of issues "were not resolved". In half the cases, the issue was deemed "not important enough to pursue". This has fueled a debate that has escalated to somewhat of a watershed moment, where the focus of attention for some has shifted from discussing research findings to chronicling the apparently sizeable undercurrent of unreliable papers that potentially mislead scores of researchers and waste research funds.

> *"…ensuring that different experimental approaches and experiments in unrelated systems yield the same conclusions remains a gold standard…"*

The *Open Science Collaboration* is one initiative that aims to quantify how reliable the literature really is. It recently reported that 39 of 100 papers assessed in the psychological sciences could be "replicated unambiguously" (Open Science Collaboration, 2015). The not-for-profit *Reproducibility Project* has turned its attention to cancer, with an ongoing effort to replicate 50 prominent cancer papers in contract laboratories (Morrison, 2014). The project is investing $1.3 million USD of philanthropic support into the independent validation of key experiments. Each validation attempt is carefully outlined in a peer reviewed, published 'registered report' that identifies specific experiments that are to be reproduced, defines protocols and discusses the related literature. This is certainly a formidable and well planned undertaking, but the proof will be in the pudding: will a negative result mean a paper is not reproducible? As Sean Morrison noted in an editorial announcing the registered reports: "It's a credible effort to address an important question… In principle, the findings of the Reproducibility Project could be undermined by the same sources of error it is attempting to address… The findings….will often defy binary categorization into right and wrong" (Morrison, 2014).

The initiative will certainly be careful to report how many papers were reproduced in

this project, but the media may be tempted to turn the argument around to declare cancer research unreliable. Not all of those whose papers are being reproduced have embraced the project. Richard Young noted: "If the project does match the results, it will be unsurprising—the paper's findings have already been reproduced. If it doesn't, a lack of expertise in the replicating lab may be responsible." (Kaiser, 2015).

Beyond this specific stress test, the pursuit of systematic reproducibility through outsourcing to contract laboratories would be an expensive and likely non-scalable solution. Moreover, it is unclear if it would lead to formal improvements to the reliability of the literature: papers that are reproduced would benefit from a stamp of approval, but papers that are not reproduced experimentally by another laboratory must not conversely fall victim to being marked by default as unreliable or suspect. Nonetheless, the reproducibility of findings certainly has to be guaranteed before entering clinical trials. There are initiatives in the research community to address this. For example, Precison Pancreas, a multicenter pancreatic cancer network in the UK, has established a cross-validation approach. Owen Sansom, a member of the project, notes: "We placed cross-institutional validation at the centre of the preclincal work. If an exciting result is found in one centre, before it can be progressed onto a human trial we have committed to cross-validation within the network using the same protocols. Our rationale is that human cancer is much more complex than any of our model systems and if a result is not reproducible across the network, then it is highly unlikely to work in patients. This is analogous to late stage clinical trials, which will never recruit in a single centre." Confirmatory or divergent data certainly deserves to be published in the quality literature and ought to be bidirectionally linked to the original paper.

Solutions

Milestone findings attract attention and attempts to build on such findings will usually uncover reproducibility problems. This kind of scrutiny-though-use will certainly not apply to the majority of the over 1 million papers published annually in the biosciences; admittedly many are essentially archival material and receive limited if any attention anyway. The problem is rather the flawed papers that receive attention, but are not corrected formally—such papers can derail research progress. Reproducibility issues may spread by word of mouth, but not everyone in the community will know and this is restricted to prominent cases; also, anecdotal evidence can be very damaging: the strong opinion of one leader in a field may suffice to undermine the work of a newcomer.

"…research assessment policies will continue to undermine the very thing that they aim to promote: efficient research progress."

What can be done realistically by the journals? An initial goal is to ensure that the existing literature is better interlinked. This can be achieved in two ways: by adding systematic forward links to correcting or corroborating literature, and by the curated versioning of papers. The former would alert or reassure readers by systematically linking them to publications that have built on or have contradicted the paper they are reading. Versioning would allow authors to update their papers in a tractable manner to reflect state of the art knowledge (see also Pulverer, 2015b). Journals also need to apply better standards that are set out in detailed guidelines and author checklists; the publication of source data and enhanced methods sections needs to become standard. Editors must actively engage in policing reagent sharing and correction of the literature.

Its not just down to the journals: we need better mechanisms to encourage and support reproducing work. The exclusive pursuit of 'breakthrough science' by funders, institutions and journals alike leave researchers little choice but to focus on the next steps ahead and hope for the best. Experienced staff scientists are the constituency best placed with the task of reproducing published findings—alas, such positions are continuously eroded.

Risk

The collateral damage in overextending reproducibility requirements may be to lower risk taking in reporting research. The more innovative an approach and the more novel a finding, the higher the risk that the data or the conclusions may not be easily reproducible. This is not in any way intended to imply that the most novel findings are based on poor experimentation, but that such work is harder to reproduce because reagents are not available or methods are new, and because such findings cannot rely on a body of corroborating evidence. If such research was state of the art, the data apparently compelling and the interpretations carefully framed, but ultimately the work does not hold up, that is just fine—it is part of the scientific process and it must not transpire as a failure. We must not throw out the baby with the bathwater in the pursuit of a definitive literature, but rather ensure the self-correcting mechanisms of the literature work (Pulverer, 2015b).

Nevertheless, it is right to expect that particular papers that promulgate extraordinary claims must also be based on the highest levels of evidence and must be subjected to and rise to an extraordinary level of validation. In particular, claims that affect public health or policy but repeatedly fail to be validated must be classed as not reproducible and removed from the literature without delay.

De-pressurized publishing

The aim has to be to formally publish only those scientific findings for which we have compelling experimental support. It is right to do all we can to underpin the quality, reliability and reproducibility of published research findings without sacrificing risk taking and the sharing of provocative findings and thought. The reproducibility debate is everywhere now and that is good. If it spills over to undermine trust in basic research, it would be devastating. The current uneasy juxtaposition of reproducibility concerns and the barrage of reports on breaches of scientific integrity could snowball into a general erosion of trust in and support for the scientific enterprise. Funders may surmise that the apparent return on their investment is diminishing. Governments may feel that tax-payer funds are better invested elsewhere. However, policy makers and funders also need to reflect on the fact that research assessment in a hypercompetitive research environment is fuelling the escalating publication rates and the incentives to publish in high Impact Factor

journals at all costs. As articulated by the San Francisco Declaration on Research Assessment (DORA), this cycle can only be broken through the action of all the stakeholders: funders, institutions, journals and above all the researchers themselves, who are broadly in charge of self-governing research assessment (Pulverer, 2015a). Senior researchers at the top of their game understandably argue that their reputation is their most valuable asset and that they would not endanger it with unreliable, unreproducible research publications. Alas, funding is scarce and the large number of PhD students and postdocs passing through the system in pursuit of an academic position based on publication in only a handful prestigious journals means that unless we ensure that research assessment looks at performance in a more differentiated manner, and unless non-academic research based careers are made more palatable, we will not be able to de-pressurize the system anytime soon. The tail will keep wagging the dog in that narrow-minded research assessment policies will continue to undermine the very thing that they aim to promote: efficient research progress.

We would all do well to support the ethos of only publishing research that to the best of our knowledge will be reproducible—and to support journals that are committed to such rigour and transparency. We also need to publish, collaborate and share reagents openly so that our science can be reproduced. At the same time, we need to both encourage correction and protect risk taking and innovation.

References

Baker M (2015) Reproducibility crisis: Blame it on the antibodies. *Nature* 521: 274 – 276

Björling E, Uhlén M (2008) Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol Cell Proteomics* 7: 2028 – 2037

Begley CG, Ellis LM (2012) Raise standards for preclinical cancer research. *Nature* 483: 531 – 533

Hines CH, Su Y, Kuhn I, Polyak K, Bissell MJ (2015) Sorting out the FACS: a devil in the details. *Cell Rep* 6: 779 – 781

Kaiser J (2015) The cancer test. *Science* 348: 1411 – 1413

Klaus B (2015) Statistical relevance—relevant statistics, part I. *EMBO J* 34: 2727 – 2730

Morrison S (2014) Time to do something about reproducibility. *Elife* 3: e03981

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349: aac4716

Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 10: 712

Pulverer B (2014) Transparent, reproducible data. *EMBO J* 33: 2597

Pulverer B (2015a) Dora the Brave. *EMBO J* 34: 1601 – 1602

Pulverer B (2015b) When things go wrong: correcting the scientific record. *EMBO J* 34: 2483 – 2485

Yu M, Selvaraj SK, Liang-Chu MM, Aghajani S, Busse M, Yuan J, Lee G, Peale F, Klijn C, Bourgon R, Kaminker JS, Neve RM (2015) A resource for cell line authentication, annotation and quality control. *Nature* 520: 307 – 311