EDITORIAL

nature cell biology

# On data availability, reproducibility and reuse

**Data sharing is an inherent principle of research publication, and information on how data may be accessed is key for the replication and furthering of scientific findings. Here we revisit the policies of *Nature Cell Biology* on data availability.**

A fundamental principle of science is that others should be able to reproduce and build on the findings reported in research publications. Sharing the materials, protocols and data that underlie these findings is a requirement of many funding agencies, including the National Science Foundation and National Institutes of Health in the United States, and the Wellcome Trust and Research Councils in the United Kingdom. It is also a long-standing policy of the *Nature* journals, including *Nature Cell Biology*. At the time of submission authors are required to provide information on reagents, methodology and experimental reporting, and to make the data that support the paper's conclusions available to editors and referees for the peer-review process. Moreover, publication in the *Nature* journals is conditional on authors making reagents, protocols, computational code and the minimal dataset underlying the findings available to readers without undue qualifications.

In principle, a paper's data could be provided in full in the main display items and supplementary information. However, given the complexity and multidisciplinary nature of the majority of *Nature Cell Biology* publications, this is very rare. For the typical *Nature Cell Biology* paper, the minimal dataset would include not only the data presented in the paper itself, but also the raw numerical data underlying graphical representations and quantitative assessments, the independent repeats of representative experiments that provide support for the reproducibility of the results, large-scale datasets generated for that particular study, and previously published datasets that have been reanalysed. In practice, some of these data would be present in the paper (in many cases including source data for graphs and independent repeats), however, large-scale datasets would be deposited in public databases with accession number details given in the Methods for both previously unpublished and publicly available datasets, and other pieces of data would be available directly from the authors. For data types such as microarrays, protein and nucleic acid sequences, DNA/RNA sequencing data, and macromolecular structures, deposition in discipline-specific, community-endorsed public repositories is mandatory for peer review and publication in *Nature Cell Biology*. We also strongly encourage deposition of other data types for which established public repositories exist, including proteomics and metabolomics. In the absence of data-type-specific repositories, data may be uploaded in generalist, unstructured repositories such as figshare and Dryad. Apart from ensuring that the policy for mandatory data deposition is followed, *Nature Cell Biology* editors reserve the right to request the public provision (for example, in the paper or through a repository) of any type of data deemed to be central to the paper's conclusions, or particularly important for the reproducibility of the findings and for transparency of experimental reporting.

Although many authors may be aware of institutional and funder mandates regarding data sharing, navigating the public repository landscape can often be complex, especially in cases where the most appropriate repository is not immediately obvious, or where multiple different choices exist. To help authors comply with the requirements of institutes, funding bodies, and with journal policy, we provide detailed information in the 'Availability of data' section of our editorial policies website (http://go.nature.com/2eG4GcL). Further information on accepted and recommended repositories for different types of datasets can be found through our sister publication, *Scientific Data* (http://go.nature.com/2eLHBFP), and additional advice can be obtained by contacting our Research Data Support Helpdesk (http://go.nature.com/2lufNqY).

Typically a research paper would include information on how the data supporting the conclusions may be accessed, but such details may be scattered in different parts of the manuscript, complicating data discovery for readers. To address this, starting in March 2016 *Nature Cell Biology*, *Nature Communications*, *Nature Geoscience*, *Nature Neuroscience*, and *Nature Physics* piloted a standardized approach to providing this information by including data availability statements at the end of the Methods sections of all published papers. The success of this pilot policy resulted in its rollout for all *Nature* journals as of September 2016. The data availability section consolidates all details on whether and how the minimal dataset underlying the findings is available to readers, and also states any restrictions to availability (for example, due to third-party rights, or privacy limitations for human data). Depending on the paper this section can be a minimal statement declaring that the data are available from the authors on reasonable request, or a lengthier paragraph detailing the location of source data, accession details for primary large-scale data, and how to access public datasets reanalysed in the paper, typically by noting accession numbers, links, digital object identifiers, or other unique identifiers with references cited if available. It should be noted that this policy does not increase the requirements of data availability for *Nature Cell Biology* papers, but rather enhances the transparency of how this information is conveyed in the paper by consolidating it in a single, clearly worded section. To aid authors, *Nature Cell Biology* editors give specific guidance on what information to include in the data availability statement by providing an example statement tailored to each particular paper in the decision letter offering publication. Further details and examples can be found in our guidance for authors (http://go.nature.com/2bF4vQn) and Frequently Asked Questions (http://go.nature.com/2nEvJHq) documents.

Data sharing and information on data accessibility is instrumental for transparent scientific reporting, and the discoverability, validation and reuse of datasets with correct attribution and citation. Ensuring that this information is available in the structured, consistent manner of data availability statements is an important step towards increased reproducibility and transparency in research.

As always, we welcome comments on this and our other policies from authors, referees and readers at ncb@nature.com.



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.