Nature Editorial: To ensure their results are reproducible, analysts should show their workings.

Description

Tweet

See Devil in the Details, *Nature*, Volume:470, Pages: 305–306 , 17 February 2011.

How many aid agencies could do the same, when their projects manage to deliver good results? Are there lessons to learned here?

Article text:

As analysis of huge data sets with computers becomes an integral tool of research, how should researchers document and report their use of software? This question was brought to the fore when the release of e-mails stolen from climate scientists at the University of East Anglia in Norwich, UK, generated a media fuss in 2009, and has been widely discussed, including in this journal. The issue lies at the heart of scientific endeavour: how detailed an information trail should researchers leave so that others can reproduce their findings?

The question is perhaps most pressing in the field of genomics and sequence analysis. As biologists process larger and more complex data sets and publish only the results, some argue that the reporting of how those data were analysed is often insufficient.

Take a recent survey by comparative genomist Anton Nekrutenko at Pennsylvania State University in University Park and computer scientist James Taylor of Emory University in Atlanta, Georgia. The pair examined 14 sequencing papers published last year in Science, Nature and Nature Genetics, and found that the publications often lacked essential details needed to reproduce the analysis — the papers referenced merely bioinformatics software, for example, without noting the version used or the value of key parameters.

“Transparency is a laudable goal, but given the complexity of the analyses, is it realistic?”

The two researchers presented their findings at the Advances in Genome Biology and Technology meeting in Marco Island, Florida, on 2 February. Although their account has not been published, it does not seem to have surprised anyone in the field. Indeed, it builds on a 2009 paper in Nature Genetics that found similar omissions in published accounts of microarray experiments. (J. P A. Ioannidis et al. Nature Genet. 41, 149–155; 2009). In this case, findings from 10 of the 18 studies analysed could not be reproduced, probably because of missing information.

If genomics were as politicized as climate science, the authors of studies in which the information trail is missing would probably face catcalls, conspiracy charges and demands for greater transparency and openness. Instead, many in the field merely shrug their shoulders and insist that is how things are done. Bioinformatics is a fast-paced science in which software and standards for data analysis change rapidly and with them, the protocols and workflows of users.

Nature does not require authors to make code available, but we do expect a description detailed enough to allow others to write their own code to do a similar analysis.

Some in the field say that it should be enough to publish only the original data and final results, without providing detailed accounts of the steps in between. Others argue that it is pointless to document the version of the software used, as new incarnations of programs differ little. But that is not always the case. Edward McCabe, then at the California NanoSystems Institute at the University of California, Los Angeles, was so perturbed when different versions of the same bioinformatics software gave wildly different results that he published a paper on it (N. K. Henderson-Maclennan et al. Mol. Genet. Metab. 101, 134–140; 2010). Reviewers resisted its publication, asking what was new about the findings, as it was already common knowledge that different software versions could dramatically affect analyses. There is a troubling undercurrent here: that the problem lies not with the lack of information, but rather with those who find the incomplete information a problem, such as researchers who are new to the field.

Transparency is a laudable goal, but given the complexity of the analyses, is it realistic? There are certainly examples of stellar documentation. The 1000 Genomes Project, for example, a project to sequence and analyse more than a thousand genomes, has carefully detailed its workflows, and makes both its data and its procedures available for the world to see. It is perhaps easier for members of that project — which is essentially repeating the same procedure more than a thousand times — to practise good experimental hygiene than it is for individual scientists, who have more flexible and varied research goals. Nevertheless, tools are coming online to simplify documentation of the complex analyses required for genome analysis. These include freely available programs such as Taverna (http://www.taverna.org.uk) and Nekrutenko’s more user-friendly Galaxy (http://main.g2.bx.psu.edu). Neither of these is perfect, but they illustrate the level of detail that could enrich published reports.

As genome sequencing spreads from the large, centralized sequencing centres that largely pioneered the technique into smaller labs and clinics, it is important that the community consider such solutions.

Category

1. Uncategorized

Tags

1. evidence
2. replication
3. transparency

Date
25/05/2025
Date Created
17/02/2011
Author
admin