

## Data and preprints

Journal requirements for data deposition and encouragement of deposition of preprints in a community preprint server are stated policy. Because many authors put data in a public repository only upon publication and many still ask about the status of preprints, here is a further statement of our position.

In justifying holding onto data, some authors have expressed fears of being scooped by users of their own data, even with attribution, or of not being able to submit a paper if a journal editor or referee considers the entire advance reported to lie in the resource value of the data set already released. We have policies for data release and preprint server use (<http://www.nature.com/authors/policies/availability.html> and <http://www.nature.com/authors/policies/duplicate.html>), and we stand behind these policies by accepting for review and eventual publication papers that contain data already released and by respecting preprint publications as desirable.

The genomics community has a history of releasing sequence data before publication. An impressive precedent was set in 2003 by the coordinated prepublication deposition of SARS coronavirus genomes from groups worldwide (for example, *Science* **301**, 309–310, 2003; *Science* **300**, 1394–1399, 2003; *Science* **300**, 1399–1404, 2003; and *N. Engl. J. Med.* **349**, 187–188, 2003; among other studies). Rapid responses to dangerous public health problems are promoted by immediate data release, resulting in an increase in the number of research groups able to work with the data and leading to tangible benefits in the characterization and treatment of disease.

A new benchmark was set for data release when whole-genome data from outbreaks of *Escherichia coli* (*N. Engl. J. Med.* **365**, 718–724, 2011) and Ebola virus (*Science* **345**, 1369–1372, 2014) stimulated the authors of these papers to release these sequence data as they were generated. These ‘real-time’ releases allow all researchers to keep up with the progression of the strains and not rely on outdated or incomplete data. Full data release will continue to be important for emerging infectious disease agents with little to no previously existing genomic information. Indisputably, the availability of these types of data is a public good, and we wish to design our policies to enable authors to release their data without jeopardizing subsequent publication.

Scientists who make data sets quickly available during times of health crisis are lauded for selfless good citizenship in putting public health considerations above their own personal ones. By extension, unwillingness to openly share important, time-sensitive data (as in the case of negative results from clinical trials; *Nat. Genet.*

**44**, 1171, 2012) may engender the opposite effect. For infectious disease genomics, the benefits of data deposition before publication are increasingly outweighing the detriments, and the collective community is driving the change. This shift may eventually apply to less urgent findings, and we want our policies to reflect that.

Currently, we require that data from all studies be made publicly available, with the mindset that the study’s value will be increased. In tune with maximizing the impact of the research for the community, we welcome and encourage uploading work to preprint servers. In our view, a willingness to share work and ideas in a public forum facilitates and often improves the progress of research. As such, publication on a preprint server will never preclude subsequent publication in a *Nature* journal. This policy, combined with our data deposition requirements, is intended to further a community-minded spirit without it being at the expense of authors.

Editors and referees may in exceptional cases recommend that authors make the data immediately publicly available. Factors such as public health interests will be carefully weighed, and the decision will be based on what is most mutually beneficial for authors, other researchers and the communal good. Journals can contribute to data access by posting hyperlinked accession codes to deposited data in a predictable place within the paper and making clear that, for genomics projects, assemblies, annotations and raw sequence deposited in a publicly funded and community-supported database accessible via a project accession code are required for publication. Standardizing these requirements will make it simpler for authors, referees and readers to easily describe, access or use underlying data sets to ensure maximum usefulness to the community.

Public health emergencies of international concern represent extreme situations that compel a timely response, and we have become comfortable with prepublication data releases under these circumstances. Having this type of data release become the norm for all types of data is a goal we think is worth working toward. We look forward to a genomics community that is confident to declare in every publication: “These data were deposited on <date> with <accession code> prior to submission of this article for publication.” ■