Heredity (2011) 106, 709
© 2011 Macmillan Publishers Limited All rights reserved 0018-067X/11
www.nature.com/hdy

EDITORIAL

Data archiving

Heredity (2011) **106,** 709; doi:10.1038/hdy.2010.43

Data are the building blocks of science, the basic observations around which we construct our theories. When a paper is published, we may doubt its interpretations, but a fundamental principle of the scientific method is that we should be able to re-examine the data and form our own opinion of their meaning. We should be able to extend existing data sets in the hope of making more powerful tests of our ideas. We might like to gather the data from many studies in order to test hypotheses or search for patterns. All of these rely on the assumption that the data underlying scientific publications are freely available, and remain available for a long time after the original project is completed.

In practice, data are not always easily accessible. A striking exception is DNA sequence data. Since the early days of sequencing, the whole international community has recognized the value of depositing sequence data in publicly available databases. The result is a fantastic resource, which is growing exponentially both in size and in value. One of the reasons for the success of EMBL/GenBank/DDBJ has been the near-universal insistence by journals that deposition of sequence data is a condition of publication. This principle has been extended to some other data types in genetics, especially microarray data, but until now has not been applied to a wide range of data types.

During 2009, *Heredity* introduced a new data policy, which extended the principle of archiving data in publicly accessible repositories to all data types. Our current policy statement is as follows:

> Authors are strongly encouraged to follow established minimum guidelines for the reporting of biological data, wherever appropriate. Guidelines for many relevant data types are available from MIBBI: Minimum Information for Biological and Biomedical Investigations (www.mibbi.org).
> DNA sequences published in *Heredity* must be deposited in a publicly available database, usually EMBL/GenBank/DDBJ, and accession numbers must be included in the final version of the paper. Where public databases exist for other data types, such as microarray data (see http://www.ebi.ac.uk/Databases/microarray.html, for example), they must be used and the relevant reference should be included in the paper. Where no public database exists, authors are strongly encouraged to provide the data on which their analyses are based as Electronic Supplementary Information. The data should be formatted for use in a relevant, readily available software package, ideally one which allows data export in a variety of formats (such as CREATE for population genetic data: https://bcrc.bio.umass.edu/pedigreesoftware/node/2). Sufficient metadata (such as sample locations, individual identities, etc.) should be provided to allow easy repetition of analyses presented in the paper.
> *Heredity* proposes to make public archiving of data a requirement for publication in the near future and welcomes feedback from authors on this proposal (please address comments to heredity@shef.ac.uk).

Three things prevented us from making data archiving a requirement for publication. The first was a concern that repositories for data storage, suitable for the wide range of data types that appear in *Heredity*, were not yet available. Second, some authors reasonably feel that they need to retain control of hard-won data for some time after publication of the first set of analyses, while they prepare further output. This requires a system that allows time-limited embargoes on access to archived data. Finally, we were concerned about being too far out of step with other major journals that publish in similar areas.

Two initiatives have now substantially changed the situation. *Heredity* has joined a group of journals dedicated to promoting the preservation and full use of data. These journals, including *The American Naturalist*, *Molecular Ecology, The Journal of Evolutionary Biology* and *Evolution*, will introduce parallel data archiving policies based on the same principles that are set out in Whitlock *et al*. (2010). Second, the National Science Foundation has sponsored the development of a flexible data repository called Dryad (see http://datadryad.org). This will act as a data library, capable of storing any data type, accepting data submissions through links to journal paper submission systems, assigning unique identifiers to data sets and making those data readily available. It will also allow embargoes. These will probably be available routinely for up to 1 year after publication and possibly for longer at the Editor's discretion. In the long term, Dryad is likely to be supported by a consortium of journals, but it has the support of the National Science Foundation until 2012.

As the major obstacles have now been removed, *Heredity*, along with other journals participating in the joint data-archiving initiative, will move to making data archiving a requirement for publication during 2010. Data types for which there are established archives, such as sequences, microarray data or phylogenetic trees, will still be submitted to those databases. Dryad is likely to be a preferred, but not mandatory, repository for other data types. These data will be required at a level suitable for re-analysis: at the individual level rather than as summary statistics, but not in raw form (for example, microsatellite genotypes for all individuals, with the metadata required to identify populations and treatments, but not chromatograms).

The journal, and the Genetics Society, firmly believes that promoting a culture of sharing data is in the long-term interests of our scientific endeavour. Now that the technology is available to make routine data archiving, searching and retrieval straightforward, we should embrace the opportunity and make our data available to all, now and for the future.

Conflict of interest

The author declares no conflict of interest.

R Butlin
Managing Editor, Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield, South Yorkshire, UK
E-mail: r.k.butlin@sheffield.ac.uk

References

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010). Data archiving. *Am Nat* **175**: 145–146.