Show me the data!

The potential and power of gene expression analysis using DNA microarrays has lead to the widespread use of this technology. These expression or 'profiling' studies (as they are commonly known) are providing a new and unprecedented view of complex biological systems[1–9]. The common practice of most scientific endeavors is that upon publication, the data that are the foundation of the paper are made available to the public. For example, when reporting newly discovered genes, scientists are required to submit the gene/protein sequence to GenBank before submitting the paper that describes the cloning and characterization of that gene. Why, then, is the standard different for the gene expression 'profiling' community?

In the past six months, I have reviewed five different microarray papers in a row (for five different journals), where in all cases, the primary microarray data tables were not provided anywhere. How can one judge the scientific integrity of the data or the conclusions drawn if one can not see the data? The question of what microarray data are to be made available and with what annotations is the subject of an article on minimal information about a microarray experiment (MIAME; see page 365; ref. 10); however, what is not being discussed is the need for data on all genes on the relevant microarrays to be made available to reviewers at the time of submission and to the community upon publication.

Many microarray experiments generate large data sets that can contain tens to hundreds of samples, each of which might contain thousands of individual data points[5,7]. Data obtained on interrogating thousands of genes in hundreds of samples can be extremely complex, with patterns of co-expressed genes contained within other patterns of genes[11]. Therefore, different biological insights may be uncovered using different analyses of the same data. To ensure that the most biological insights can be extracted from a complex data set, the data generated in one laboratory must be made available to others, as each researcher has a unique perspective and brings different analytical methods to the table that will help to extract insights beyond those identified by the original set of authors. For example, the data sets described by Alizadeh *et al.*[1] and Golub *et al.*[6] represent rich resources and have fuelled publications of other groups.

In most cases, in place of the primary and/or processed microarray data tables, authors provide subjective interpretations of the data and limited sets of genes that they believe encompass the important aspects of their work. At the very least, some data value for every gene on the microarray must be made available, as the identity of genes whose expression levels do not change is often as important as those that do; in addition, many researchers are only interested in one or a few genes. Their interests are ill served if these genes do not make it into these 'selected' gene lists.

It is clear that new policies need to be instituted to ensure that microarray data are made publicly available. Currently, this is voluntary and many choose not to release the primary microarray data. My suggestion is to place much of the enforcement on the journals who publish the studies and on the granting agencies who fund the studies. In this scenario, the policy could be that once a paper has been accepted for peer review, the primary microarray data tables would be included with the text and figures of the manuscript, or that when a Primary Investigator accepts public funds to perform microarray experiments, she or he would agree to release some form of the primary data upon publication.

The question of where to house the data, and in what format, requires and is receiving serious attention[10]; however, there are a number of short-term considerations. First, an obvious location to place the data is in the NCBI Gene Expression Omnibus, the DNA Database of Japan or the European Bioinformatics Institute, where data can be housed and given an Accession Number, and where upon publication or some predetermined date, the data 'attached' to that Accession Number are released into the public domain.

A second measure, which should at least be adopted during the peer review process, is that each journal provides a protected site on its server where data can be housed and made available to the reviewers. In today's modern computer-based age, journals must adapt: they must be able to handle and distribute large electronic data sets to reviewers. One method of data distribution is to place primary microarray data on the private web sites/servers of individual researchers. This method can, and does compromise, reviewer anonymity. As many know, when logging into a website to retrieve data, the identity of the location from which one is logging in is recorded by the server that houses the data, and the location of the reviewer is thus divulged. The current standard is that the reviewer remains anonymous unless he or she requests otherwise; therefore, the exclusive housing of large data sets on private web sites during the review process must be changed to protect those reviewers who choose not to make their identities known. Regardless of where the data are initially housed, it should be understood that upon publication, the data will be made available as Supplementary Information either at the journal's web site, the author's web site, the NCBI Gene Expression Omnibus, some other public repository or a combination of any of these.

The development of standards for data release of gene expression studies should help other large-format biological studies, but there are additional challenges specific to some of these other technologies. For example, what are we to do with tissue microarray data where the primary data may consist of hundreds to thousands of individual tumor images, each of which might be several megabytes in size? In this case, the storage of primary data may require the ability to host a Gigabyte of data. Some in the microarray field, including most involved with the MIAME effort, would argue that microarray image files should be provided along with the tables of numeric data. What are we to do with the upcoming flow of proteomics data? Are we to ask that the separation methodology data (that is, 2-D and/or chromatography data) and mass spectroscopy data be made available? Are authors' interpretations of these data sufficient? These are some immediate challenges that must be dealt with soon. As for gene expression studies, we can and should adopt the standards set forth in the MIAME report[10]. In some ways, though, this is putting the cart before the horse, because public release of such data is not yet required.

Charles M. Perou

Lineberger Comprehensive Cancer Center and Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. Correspondence should be addressed to C.M.P. (e-mail: cperou@med.unc.edu).

1. Alizadeh, A.A. *et al. Nature* **403**, 503–511 (2000).
2. Alon, U. *et al. Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
3. DeRisi, J.L., Iyer, V.R. & Brown, P.O. *Science* **278**, 680–686 (1997).
4. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
5. Gasch, A.P. *et al. Mol. Biol. Cell* **11**, 4241–4257 (2000).
6. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
7. Hughes, T.R. *et al. Cell* **102**, 109–126 (2000).
8. Perou, C.M. *et al. Nature* **406**, 747–752 (2000).
9. Winzeler, E.A., Lee, B., McCusker, J.H. & Davis, R.W. *Parasitol.* **118**, S73–S80 (1999).
10. Brazma, A. *et al. Nature Genet.* **29**, 365–371 (2001).
11. Brown, P.O. & Botstein, D. *Nature Genet.* **21**, 33–37 (1999).