**Community Page**

# Submission of Microarray Data to Public Repositories

**Catherine A. Ball, Alvis Brazma\*, Helen Causton, Steve Chervitz, Ron Edgar, Pascal Hingamp, John C. Matese, Helen Parkinson, John Quackenbush, Martin Ringwald, Susanna-Assunta Sansone, Gavin Sherlock, Paul Spellman, Chris Stoeckert, Yoshio Tateno, Ronald Taylor, Joseph White, Neil Winegarden**

A fundamental principle guiding the publication of scientific results is that the data supporting any scholarly work must be made fully available to the research community, in a form that allows the basic conclusions to be evaluated independently. In the context of molecular biology, this has typically meant that authors of a paper describing a newly sequenced genome, gene, or protein must deposit the primary data in a permanent, public data repository, such as the sequence databases maintained by the DNA Data Bank of Japan (DDBJ), European Bioinformatics Institute (EBI), and National Center for Biotechnology Information (NCBI). Similarly, we, members of the Microarray Gene Expression Data Society (MGED; http://www.mged.org), believe that all scholarly scientific journals should now require the submission of microarray data to public repositories as part of the process of publication. While some journals have already made this a condition of acceptance, we feel that submission requirements should be applied consistently and that journals should recognize ArrayExpress (Brazma et. al. 2003), Gene Expression Omnibus (GEO) (Edgar et. al 2002), and the Center for Information Biology Gene Expression Database (CIBEX) (Ikeo et. al. 2003) as acceptable public repositories.

To this end, the members of MGED propose the following as a new paradigm for the publication of microarray-based studies. (1) Authors should continue to take primary responsibility for ensuring that all data collected and analyzed in their experiments adhere to the "Minimum Information about a Microarray Experiment" (MIAME) guidelines and should continue to use the MIAME checklist (http://www.mged.org/Workgroups/MIAME/miame_checklist.html) as a means of achieving this goal. (2) Scientific journals should require that all primary microarray data are submitted to one of the public repositories—ArrayExpress, GEO, or CIBEX—in a format that complies with the MIAME guidelines. (3) Public databases should work with authors and scientific journals to establish data submission and release protocols to assure compliance with MIAME guidelines. (4) To assist with the review process, the databases should continue to work in collaboration with publishers to provide qualified referees with secure means of accessing prepublication data. Authors should be strongly encouraged to submit data to the databases during review.

Naturally, data should be protected from general release prior to either publication or authorization from the data submitters, whichever comes first. At a minimum, journals should require valid accession numbers for microarray data as a requirement for publication, and these accession numbers should be included in the text of the manuscript to allow members of the community to find and access the underlying data.

Since its inception in 1999, MGED has been working with the broader scientific community to establish standards for the exchange and annotation of microarray data. In December 2001, we proposed the MIAME guidelines (Brazma et al. 2001) and requested that interested parties provide feedback on its relevance and utility. The feedback from both researchers and scientific journals was overwhelmingly positive, yet almost everyone who responded also asked for help in implementing these guidelines.

Catherine A. Ball is in the Department of Biochemistry, Stanford University School of Medicine, Stanford, California, United States of America; Alvis Brazma is at the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom; Helen Causton is at the Clinical Sciences Centre/Imperial College Microarray Centre, Imperial College, London, United Kingdom; Steve Chervitz is Bioinformatics Engineer at CIS Enterprise Data Group, Affymetrix, Emeryville, California, United States of America; Ron Edgar is at the National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, United States of America; Pascal Hingamp is at the Laboratoire Technologies Avancées pour le Génome et la Clinique, Faculté des Sciences de Luminy, Centre d'Immunologie de Marseille Luminy, Universite Aix-Marseille-II, Marseille Cedex, France; John C. Matese is at the Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, New Jersey, United States of America; Helen Parkinson is at the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom; John Quackenbush is at The Institute for Genomic Research, Rockville, Maryland, United States of America; Martin Ringwald is at the Jackson Laboratory, Bar Harbor, Maine, United States of America; Susanna-Assunta Sansone is at the European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, United Kingdom; Gavin Sherlock is in the Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America; Paul Spellman is Computational Scientist at Lawrence Berkeley National Laboratory, Berkeley, California, United States of America; Chris Stoeckert is Research Associate Professor in the Department of Genetics, Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; Yoshio Tateno is Professor at the Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima, Japan; Ronald Taylor is Senior Research Scientist II at Computational BioSciences Group, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, United States of America; Joseph White is at The Institute for Genomic Research, Rockville, Maryland, United States of America; and Neil Winegarden is at the University Health Network, Microarray Centre, Toronto, Ontario, Canada.

\*To whom correspondence should be addressed. E-mail: brazma@ebi.ac.uk

Subsequently, in the summer of 2002, we submitted an open letter to various journals (e.g., Ball et al. 2002a, 2002b) urging the community to adopt the MIAME requirements for microarray data publication. We provided a checklist so that authors could ensure that sufficient information to allow their data to be re-analyzed by others would be available. Again, the response from the community was extremely positive, and most of the major scientific journals now require publications describing microarray experiments to comply with the MIAME standards. While the adoption of these standards has greatly improved the accessibility of microarray data, much of it remains on individual authors' websites in a variety of formats; consequently, obtaining and comparing datasets remains a significant challenge. Clearly we need additional requirements for publication that include submission of expression data to public data repositories.

Though one might ask why this requirement was not part of the original MIAME recommendation, the answer is quite simple—MIAME was ahead of its time. While NCBI and the EBI had developed nascent microarray data repositories, and work was underway to create a similar database at the DDBJ, submitting data to these databases was a considerable burden for authors. However, since that time, improvements in the data-entry utilities available for GEO (http://www.ncbi.nlm.nih.gov/geo), ArrayExpress (http://www.ebi.ac.uk/arrayexpress), and CIBEX (http://cibex.nig.ac.jp), as well as a growing number of commercial and academic software packages capable of writing MAGE-ML documents (Spellman et al. 2002) that can be directly submitted to these public databases, have lowered the barriers for data submission to the point

where we as a community **must** now reconsider that submission to one of these databases be a requirement.

Requiring authors to submit microarray data to a public database will provide a number of distinct advantages to the entire research community. (1) These established repositories have a commitment to continued community service and to providing some level of assurance that published gene expression datasets will continue to be available into the future. (2) Having the data available in these public repositories in a standardized format will not only make them more accessible, but it will allow expression data to be integrated with other relevant data, including the available genome sequences, single nucleotide polymorphism and haplotype mapping information, the literature, and other resources that can aid in further interpretation of expression patterns. Although many authors now provide some or all of this information, the established databases are much more likely to assure that these links are maintained and current. (3) Curation of data submitted to public data repositories will assist authors, reviewers, and publishers in assuring that the data comply with the MIAME requirements, further enhancing their utility. (4) The standardization of microarray data formats will enable the development of additional data analysis and integration tools and makes it easier for scientists to access, query, and share data. (5) Finally, submission prior to publication will make it easier for referees to access the data confidentially, facilitating the review and publication process.

In the same way that availability of sequence data had a profound impact on a wide range of disciplines, we believe that requiring that microarray data be deposited in public repositories as a necessity for publication will

accelerate the rate of scientific discovery.

What this proposal requires is a change in the way in which we approach the publication of microarray-based studies. Both authors and journals have a responsibility to assure that the requisite data are available, and because submitting MIAME-compliant data can take considerable time and effort, this process should be factored into review and publication timelines. However, while this process may be time consuming and painful at first, we believe that the benefits of building an open repository of microarray data will far outweigh any initial disadvantages. As always, it is our sincere hope that these suggestions stimulate discussion within the community and that together we can arrive at a consensus that ensures that microarray data are widely and easily accessible. Finally we would like to urge the DDBJ, EBI, and NCBI to work together towards exchanging all MIAME-compliant microarray data. ∎

**References:**
Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, et al. (2002a) Standards for microarray data. Science 298: 539.
Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, et al. (2002b) The underlying principles of scientific publication. Bioinformatics 18: 1409.
Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365–371.
Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. (2003) ArrayExpress—A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31: 68–71.
Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y (2003) CIBEX: Center for information biology gene expression database. C R Biol 326: 1079–1082.
Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3: RESEARCH0046.