nature genetics

Minimum compliance for a microarray experiment?

Results generated with the use of microarrays should be properly described and deposited in one of the public databases prior to review. We examine the last two years of the journal's publications in an attempt to ensure that community standards for data deposition are met.

This month our sister journal, *Nature Biotechnology*, publishes results of the Microarray Quality Control Consortium (http://www.nature.com/nbt/focus/maqc/index.html) comparing the performance of seven different array platforms in measuring RNA expression. Are the data generated from these microarrays being made available with the same attention to quality control that is lavished on the platforms themselves? Not quite. Although we have come a long way toward a community consensus in the five years since the launch of the Minimum Information About a Microarray Experiment concept (MIAME) (*Nat. Genet.* **29**, 365–371; 2001), there are a number of simple steps that can still be taken to improve the impact and accessibility of the published data sets.

Nature and the *Nature* Research Journals insist upon authors obtaining experiment accession codes to MIAME-compliant data. To ease this process, the public Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (AE) (http://www.ebi.ac.uk/aerep/?) databases that host the data can now be accessed via submission forms employing the mark-up language and ontology developed by the Microarray Gene Expression Data Society (MGED). The MIAME principles are simple: enough information should be made available and appropriately linked together to allow unambiguous interpretation of the data and potential verification of the conclusions.

Of 30 microarray papers published in *Nature Genetics* in 2005–2006, we judged all 11 with AE accessions to be MIAME compliant. The main criticisms of these entries were that they lacked bibliographic citation to and from the original publication, suggesting that communication between editors and curators is in need of improvement. We identified a number of commercial array designs bearing reporter entities that were not described on the array design by their complete oligonucleotide sequence or by a sequence accession number and sequence coordinates. MIAME originally provided a compromise on this issue. Unsurprisingly, those manufacturers who completely describe every feature of their product seem to be increasingly favored by researchers. Ten of nineteen papers with GEO accessions were fully MIAME compliant. Those that were not fully compliant variously lacked (i) 'raw' data output by the array scanner, (ii) oligonucleotide sequences or sequence coordinates on the array design, (iii) hybridization or extraction protocols or (iv) a complete array description. Overall, 15/19 GEO accessions were correctly linked in both directions to the original publication via PubMed. We expect to be contacted by alert readers should we fail to set right any remaining deficiencies.

GEO may seem a little quicker to use and more flexible at the data entry stage. Unfortunately, the data submission form carries optional protocol fields for treatment, growth, extraction, labeling, hybridization and scanning. This has led to varying interpretations as authors attempt to submit the minimal information: sometimes protocols are submitted; other times they are submitted via links to papers or websites or are even omitted. In contrast, the data entry procedure at ArrayExpress more than makes up for the time taken, since it ensures that protocol fields are either filled or linked to existing protocols in the database. Despite some large files that need to be negotiated, the resulting data entries are more uniform and more richly interconnected, allowing rapid and thorough checking of the basic data points. We endorse the suggestion of the original MIAME paper that the database entries should be a complete record of the experiment, even if this duplicates information available in the published paper. The strongest reason for this is that we have detected several cases where an accession cites a chain of published works, and the original methodology cannot be found. In the interests of verifying, using and integrating GEO and AE entries, the minimal record should be as complete as possible within the public database structure provided.

On balance, we continue to endorse both of the public databases wholeheartedly for their considerable strengths. Because accessible results mean more citations and greater impact for your papers, we emphasize that correct public submission of all microarray data is a good way to advance your scientific reputation (see the October 2004 Editorial (*Nat. Genet.* **36**, 1025; 2004)). To help, checklists covering array-based measurement of gene expression, chromatin immunoprecipitation (chIP-chip) and comparative genomic hybridization (aCGH) are provided by MGED (http://www.mged.org/Workgroups/MIAME/miame.html). We encourage authors to help with expeditious review and publication by making their MIAME-compliant GEO and AE accessions available under a private password to editors and referees at the time the paper is submitted for review. ∎