Mandatory submission of microarray data to public repositories: how is it working?

Beverly Ventura
Managing Editor, Physiological Genomics

ONE OF THE UNDERLYING PRINCIPLES of scientific research and publication in peer-reviewed journals is the concept of reproducibility, within studies and between laboratories. Since January 1, 2003, Physiological Genomics and the American Physiological Society have adopted the microarray data standard developed by the Microarray Gene Expression Data Society (MGED), requiring that all authors using microarray data analysis in their research submit a complete data set to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus database (GEO; http://www.ncbi.nlm.nih.gov/geo/) prior to manuscript submission. Any paper submitted to an APS publication that uses microarray data analysis must comply with the “minimum information about microarray experiments” (MIAME) standard. Access to the GEO database is free and open for all, and the database is maintained by NCBI (3).

The American Physiological Society believes that microarrays have become an important tool in molecular genetics and physiology research within a short time period. “For microarray analysis of gene expression to have any long-term impact, it is crucial that the issue of reproducibility be adequately addressed. In addition, since microarray analytic standards are certain to change, it is crucial that authors identify the nature of the experimental conditions prevalent at the time of their research. If today’s research is to be relevant tomorrow, the core elements that are immune to obsolescence must be made clear” (http://www.the-aps.org/publications/i4a/prep_manuscript.htm).

GEO became operational in July of 2000 as a way of managing and sharing the huge amounts of scientific data generated in microarray experiments (7). The difficulty with microarray experiments lies not only in management and interpretation of the large amount of data involved but also with including the other information needed to make sense of that data in a way that can be used by others (1). A microarray study that looks at 40,000 genes from 10 different samples under 20 different conditions produces at least 8 million data points (2). Such a huge volume of data would make no sense without information about the tissue type and collection conditions. (See sidebar, below.)

Almost two years after instituting the requirement for submitting microarray studies to GEO, Physiological Genomics decided to take a look at the process. In September 2004, we surveyed 76 scientists who have had microarray studies published in this journal and 237 more who reviewed microarray papers for the journal. Overall, 29% of them responded to our survey (Fig. 1).

Do you consider the depositing of all microarray data published in this journal to be of significant value to the scientific community? According to Ralph A. Meyer, Jr., Director of the Biology Division, Orthopaedic Research Laboratory, Carolinas Medical Center, Charlotte, NC, the potential is there for a valuable contribution.

“We study bone fracture healing in which almost all genes change in expression at some time point after fracture,” Meyer said. “The process of healing of this tissue is quite complex. We use microarrays with 8,700 genes. There must be threads of data pointing to new metabolic pathways involved in this process that we do not appreciate now. A more adventurous soul might use these deposited data to explore the roles of presently unknown factors. A public repository would be a valuable way for others to examine possible novel factors.

“The real question is whether the repositories are used in this fashion. Are there folks reviewing other people’s arrays and writing papers on the further analysis of deposited data? I have not seen this presented at meetings, nor have I seen manuscripts like this. It may be happening, but I am unaware of it.

“The mining of existing data would need to happen quickly after the deposit,” Meyer added. “Technology continues to improve. Microarrays from five years ago had very few genes and were so expensive that few arrays could be used in a project. I suspect such data would be of limited value today, since repeating the experiment with modern arrays would yield so much more information. Similarly, most of us use microarray measurement of RNA levels as a surrogate for protein synthesis. In ten years, proteomics will advance to the point of being able to measure most proteins in a tissue in one assay, so deposited data about the measurement of RNA levels will be of less interest.”

The value of GEO submission is undeniable to many researchers. Adrian Recinos III of the Department of Internal Medicine, Division of Endocrinology at the University of Texas Medical Branch explained, “Having access to the raw data affords the reader an opportunity to judge through his/her own statistical methods the significance of what the author is purporting. Also, these data are expensive and painstaking to obtain, and having access to such experiments already performed is useful to readers conducting similar types of studies.”

While most (92% of the authors) said they believed the depositing of microarray data to be of significant value to the scientific community, some questioned whether the process is really allowing for its intended purpose: to facilitate the meaningful sharing of data between researchers.

“Because of the lack of the standardization and the existence of multiple microarray platforms,” said Yixin Wang, Director of Discovery Research and Molecular Diagnostics at Veridex, LLC in San Diego, “direct comparisons of data sets from different sources are difficult at this stage. Unless the unprocessed raw data are deposited into the database with sufficiently detailed sample descriptions, it is hard to reproduce the analysis in the published paper. If the journal decides to ask for
deposit of the data, it is imperative for the editors and reviewers to ensure that the manuscripts describe the details of data analysis so that the readers can take full advantage of the data in the database.”

For some, the system works just the way it was intended. “We have had inquiries about our own paper that led us to refer the inquirer to the public database,” said Peter F. Davies, Professor of Pathology and Laboratory Medicine and Director of the Institute for Medicine and Engineering, University of Pennsylvania. “In our case, investigators had candidate gene findings using a similar experimental design as ours and they wished to obtain a fuller picture of related gene expression changes.”

A full 55% of the reviewers said they were aware of researchers making use of the data that has been deposited in GEO or other general data depositories (see Refs. 3, 6, 7, 9, for example). The uses vary, with some mining the microarray data and others using the site for statistical validation or testing of software tools in development.

According to Myriam Gorospe of the RNA Regulation Section, Laboratory of Cellular and Molecular Biology of the National Institute on Aging at NIH, “My laboratory has recently published several papers with array data that was deposited in GEO. Despite the fact that these papers were only published several months ago, we have already received inquiries on these data. My sense is that researchers do see and make use of such data, probably to see if/how the data agree with their own findings. However, I should add that we have received comparable number and type of inquiries from other scientists regarding other array data that was not submitted to GEO, but was instead available through our own institutional databases.”

Grier Page, Assistant Professor of Statistical Genetics at the University of Alabama-Birmingham, has used the data in GEO and the Nottingham Arabidopsis Stock Centre for several projects. One of these, Page said, used the Power Atlas (http://www.poweratlas.org), “a web-based resource that is still in development that allows investigators to choose a public data set that is similar (in organism, tissues, and/chip type), to use as pilot data to estimate the sample size required for a future experiment. For this experiment we downloaded all the data in GEO.” Page’s group also uses GEO for quality control, mining it for normative data for developing standards and for network estimation and reconstruction. “We have taken the public data and used it for a variety of purposes: looking for orthologous genes that are similar in expression across species and tissues; paralogous genes that are similar and different in expression across tissues and experiments; and seeded cluster analysis, taking genes with known function and using the large public data resources to identify new genes of potential functional relationship to these genes.

“The incredible power of public data makes many of these projects possible,” Page said. “We are trying to solve many highly complex problems for which it is critical to have very large data sets, such as GEO, to assist in the interpretation of studies.”

John D. Porter, Professor of Neurology at Case Western Reserve University in Cleveland agreed. “My lab has been using data deposited in GEO,” he said. “The newer feature of being able to search for expression patterns of individual genes across many data sets or for multiple genes within restricted sets of data has made GEO increasingly useful. For some time, you needed to have access to algorithms within your own group to download and evaluate GEO data, but the new tools
available on the web site have made it much more useful to a broader audience. The alternatives for entire data sets are lab web sites, which often disappear from the public view, and supplemental material on journal web sites, which are currently and should remain restricted to processed data rather than the original raw data sets that would take up a lot of space.

“Depositing data to GEO should not be an obstacle, as the procedure is rather painless once one becomes acquainted with it,” Porter added. “Overall, it would be hard to argue against a central repository that is trying to improve access and analytic tools.”

In contrast, another researcher said he is opposed to mandatory deposit of microarray data into public repositories, because a group may be continuing their analysis of the data, with other publications expected from that analysis. He likened it to putting a lab notebook on the internet and allowing colleagues access to it before the project is complete.

Those who work in private industry, where proprietary information is their bread and butter, are often reluctant to submit data to public repositories. These investigators may be prohibited from submitting manuscripts to journals that require data deposit by their pending patents and commercial agreements.

According to a researcher who asked not to be identified, “In my own case, a potential manuscript describing microarray data was not submitted, as our collaborator was an industry partner, and publishing primary data, owned by the industry partner, might have presented intellectual property concerns.

“We ended up passing on the opportunity to publish our observations (and were scooped by others), but instead took the approach of following up on the microarray data using real-time PCR and publishing those data,” she said. “So the requirement for depositing primary data probably does inhibit some authors. On the other hand, the mindless publication of papers with data limited to those derived from microarray analysis has very limited value anyway. I think overall I would favor maintaining the strict requirement to deposit data.”

Do authors for Physiological Genomics feel that the current databases [GEO at NCBI, ArrayExpress at the European Bioinformatics Institute (EBI), and the Center for Information Biology Gene Expression database (CIBEX) in Japan] provide a satisfactory platform for this data collection, or should efforts be made to set up one universal database for all microarray information? Many in our survey favored the idea of a universal database, if a less cumbersome system could be developed for depositing and searching the data. Others felt that the databases in existence (5, 8) are sufficient.

“It [setting up a universal database] would be a worthy goal,” said Steven P. Suchyta, Assistant Professor of Animal Sciences at Michigan State University. “As it stands now, GEO is fine for storing the raw data, but due to the complexity of microarray experiments and the data associated, entering data and retrieval of data from GEO is not as straightforward as the sequence databases for example.”

“Having a combined database also greatly expands the possibilities for those doing metadata analyses,” added Joshua Spin of Stanford University. “I think that GEO has access limitations, in terms of searching and downloading data en bloc. The more comprehensive the database, and the more searchable and categorizable, the better.”

“Submission of microarray data [in GEO], at least for Affymetrix users, is easy and straightforward,” noted Miroslav Blumenberg, Associate Professor of Dermatology and Biochemistry at New York University Medical Center. “Home-made cDNA arrays are more problematic.”

Jerry Wright of the Department of Physiology at Johns Hopkins Medical School pointed out some of the difficulties with GEO use. “Getting raw data out of GEO is currently difficult because of the obscure syntax and confusing interface; not that other sites are much better. I’ve found it easier to send people CDs with my data files than to explain to them how to locate and download from NCBI,” he said.

Eric P. Hoffman, Director of the Research Center for Genetic Medicine and James Clark Professor of Pediatrics at Children’s National Medical Center in Washington, DC, feels some of the difficulty people have experienced with GEO submission is because the organization is understaffed in comparison to its European counterpart at ArrayExpress, i.e., 6 or 7 full-time people as opposed to about 25. He feels it is important to increase support for GEO. (See sidebar, below.)

Martin Flueck of Universit ä t Bern in Switzerland, defended the GEO submission process, saying, “the procedure for deposition of the files is straightforward. Beginners may accidentally submit wrong data as they do not yet know the correct syntax. However, the curators do their job very well and immediately respond with e-mail in case a set of data is odd and has to be removed.”

Researchers do face challenges in the day-to-day use of public repositories. Both entering one’s own data and making use of others’ data sets can require a researcher to acquire and learn to use new and sometimes complex software. Without the proper program, the process of submitting data to GEO or ArrayExpress can be quite time-consuming and laborious (10, 11). Although there is some grumbling, 67% of those who responded said they did not find the deposit of microarray data into GEO to be an obstacle to submission or review of articles in Physiological Genomics.

Should data deposits in GEO continue to be required? Some feel that requiring the deposit of data prior to publication is the only way it will become a standardized process. Others think its time has either not yet arrived, or should never arrive.

“I would strongly favor discontinuing the requirement that data be deposited in GEO,” said one researcher at a government institution. “The one good paper I know of that actually used the repository as its primary source of data was published about a year or two ago in Physiological Genomics itself and concluded that cross-platform comparisons using databases such as GEO are hazardous at best, invalid at worst.

“Additionally, both as a reviewer and an author, I concur with the sentiment that uploading data into GEO is an obstacle to publication in PG. Unlike sequence data, microarray data are highly sensitive to experimental context, which makes cross-experiment comparisons based on unprocessed data a very risky business indeed. Authors are often reluctant to give unrestricted access to primary data without a compelling reason, and given the very limited utility of public databases in their current and foreseeable state of development, it is hard for some to see why microarray data should be treated any differently than, say, the raw output of ELISA plate readers.”

On the other hand, Brian Black, Assistant Professor at the Cardiovascular Research Institute, University of California,
San Francisco, said, “I feel strongly that microarray and other large-volume supplemental data should be deposited in a public database as a prerequisite to publication. In my view, this is no different from the requirement to make large annotated sequence files available through GenBank. I do not see this as an obstacle to publication or review, and I believe these databases are being used to see lists of targets, etc. I also believe that as these data are more consistently deposited, the information will begin to be more widely used than it is now.”

“My lab has generated a large chunk of the Affymetrix data out there (6,000 internal profiles, about 1,500 public at this point),” added Eric Hoffman. “We’ve also been working with Oracle databases quite a bit to improve public access. We also do pipelines to NCBI GEO, and work closely with them. GEO can indeed be labor intensive; however, without your (nice) requirement, the raw data would likely never see the light of day. So I strongly urge you to continue this requirement, with the note that submission processes will become considerably easier (e.g., our GEO submission process internally is a single keystroke process, but we have put four years of database programming behind it).”

“The time and effort required to submit the data to GEO will easily pay off with standardized results useful for the public,” agreed Thomas Langmann of the Institute of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Germany. “Also, I think that GEO submission does not keep away researchers submitting to PG. In my case, GEO helped to prevent duplicate publication of similar data sets without additional data and thereby helped the review process.”

The Editorial Board of Physiological Genomics thanks all the scientists who responded to this survey. Based on these results, the mandatory submission policy of Physiological Genomics will not only be sustained, but the Editorial Board has requested that the APS Publications Committee extend this policy to allow submission of the data to ArrayExpress, or CIBEX, as well as to GEO, in order to make submission easier for all authors.

SIDEBAR
Upon completion of this survey, the results were sent to the staff at the Gene Expression Omnibus. Following is the response from Ron Edgar, Ph.D., Gene Expression Team Leader at NCBI.

“I agree with most of what your reviewers and authors had to say, especially the observation that journals play a critical role in making such data . . . ever see the light of day. I think we (the community) are sadly missing out on proteomic and other technologies which are being generated in large quantities but not submitted anywhere because it is not enforced by the journals. GEO is fully capable of accepting virtually any type of high-throughput data.”

“Microarray data is inherently more complex than sequence or publication records, not merely because of the sheer volume, but because the true meaning of the data is only apparent in conjunction with the biological, experimental and statistical contexts it was captured, whereas a sequence is a sequence. It is therefore inevitable that submission of such will be a more complex task.”

“We put a lot of emphasis on improving the services we provide, from submission tools to search and data mining. Everyone wants effective data mining but there can be no mining without proper annotations. Proper annotation takes effort and time from submitters but the scientific community will benefit in the long run.”

Edgar also provided the following information about GEO: GEO holds more than 35,000 submitted accessions, which account for about half a billion individual measurements. They receive 2,000–3,000 new array measurements every month.

Typically, GEO records are accessed more than 15,000 times each weekday, by more than 1,000 unique users.

Bulk FTP downloads average 30,000 per month (http://www.ncbi.nlm.nih.gov/geo/).

REFERENCES