Genome Medicine

Commentary

Data reporting standards: making the things we use better

John Quackenbush

Address: Department of Biostatisics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Biostatisics 44 Binney Street, Sm822, Boston, MA 02115, USA. Email: johnq@jimmy.harvard.edu

Abstract
Genomic data often persist far beyond the initial study in which they were generated. But the true value of the data is tied to their being both used and useful, and the usefulness of the data relies intimately on how well annotated they are. While standards such as MIAME have been in existence for nearly a decade, we cannot think that the problem is solved or that we can ignore the need to develop better, more effective methods for capturing the essence of the meta-data that is ultimately required to guarantee utility of the data.

There was a time when making one's data publicly available meant publishing the image of a gel as part of a manuscript. After all, anyone else could look at the evidence in the picture, judge the quality of the data, and draw a conclusion about whether the data supported the conclusions presented in the manuscript. As DNA sequence data began to become more common in published research articles, authors regularly included figures or tables that presented the base sequence they had determined, and other scientists could use those data by manually transcribing the sequence and performing their own analysis. But as the complexity of sequence data grew and the scale of sequencing expanded with improvements in technology, it quickly became obvious that other, more systematic solutions were necessary.

And hence was born GenBank and the other international sequence repositories. GenBank started at Los Alamos National Laboratory as little more than a research project on how to index and archive DNA sequence, and quickly became an international resource and one of the major products of the National Library of Medicine and its National Center for Biotechnology Information (which was largely created to deal with just this type of data). In time, authors were no longer required to 'publish' their sequence data in research articles, but instead simply had to provide an 'accession number' that essentially guaranteed that other scientists could download the data and re-analyze them to verify the published results. As sequencing technologies evolved, the genome sequence databases adapted to provide sequence-quality data and other information that was essential to understand and interpret the data. And in most instances, all one needed to analyze the data was a minimal amount of information about the source. Even submitting data to these public repositories was relatively easy - the instruments that generated the data generally reported them in a standard format that could easily be uploaded with the appropriate annotation.

However, the development of DNA microarrays presented new challenges in defining what one meant by data availability, one minor and one major. The minor problem, from the outside, seems like the most significant, and that is the sheer quantity of data that microarray assays produce. Assays look at expression of more than 24,000 'genes' or 50,000 'probes (or probesets)' or 1,000,000 variant positions in the genome. And when collected across hundreds or thousands of samples, the absolute data volume can be staggering. But instruments and software produce data in tabular format that most public databases, such as GEO [1] or ArrayExpress [2] or CIBEX (the major efforts at NCBI, EBI and DDBJ, respectively, to capture functional genomics data), accept and so while the size of the data is big, the problem of dealing with it is not. Instead, the greatest problem is in simply describing what the data represent: what experiment was done, what hypotheses were tested, and what ancillary sample parameters describe the data.

In 2001, a number of colleagues and I published a description of a Minimum Information About a Microarray Experiment (MIAME) standard [3] that attempted to address the issue of data reporting in these types of experiments, and most of the scientific journals jumped, requiring MIAME-compliant submissions and accession numbers from the major databases (which themselves were involved in developing the standards). MIAME was supported by a data-reporting format, MAGE-ML [4], and later the more human-friendly and readable MAGE-TAB format [5]. And the efforts of the Microarray Gene Expression Data (MGED) society in developing MIAME spurred other communities to develop their own reporting standards and formats; many of these public repositories now exist, making data from proteomics and metabolomics and other high-throughput studies

MIBBI, Minimum Information for Biological and Biomedical Investigations; MGED, Microarray Gene Expression Database; MIAME, Minimum Information About a Microarray Experiment; PCR, polymerase chain reaction.

available. In fact, there are now more than 30 different reporting standards, and even an effort, the Minimum Information for Biological and Biomedical Investigations (MIBBI), that aims to collect standards as a first step to rationalizing and standardizing them.

Despite a growing realization that standards are necessary to enable re-use of data, the problem with all of these standards efforts is that they have failed to fully solve the problem of experimental annotation - describing the how and what behind the work. And this failure is reflected in the somewhat sorry state of the data in public databases. While anyone can now download a dataset from GEO, to analyze it one generally has to go back to the article and manually assign phenotypes to samples represented in the dataset. What is more frustrating is that the annotation for samples provided in an article is often incomplete, making the search for potential confounding effects a challenge, and causing one to wonder about the validity of the analysis that is presented.

A colleague and I recently published a re-analysis of a public gene-expression study in breast cancer that had searched for a predictive signature for metastasis to the lung. We found that the reported signature [6] was a much better predictor of molecular subtype, something that is known to be extremely important in breast cancer outcome, than it was of lung metastasis [7]. While this was certainly a simple oversight on the part of the original authors, there was no information on molecular subtype for the samples and we only discovered it by doing a 'forensic' analysis using other published signatures to infer subtype. And this raises another issue beyond the scope of this commentary - the lack of reporting standards for signatures inferred from the analysis, making the results from any analysis something that is often lost in supplemental tables or figures rather than something that is easily accessible and standardized.

So there are a few simple questions that we, as a community, must address. First, do we need data-reporting standards? I think the answer is yes. Two recent commentaries from international workshops concluded not only that standards are needed, but that they should expand to be more comprehensive [8,9]. Their view is that the value of the data is not in the individual studies, but in their use and re-use beyond the initial publication. And this is something that has clearly proven itself to be true for many, many well-annotated genomic datasets.

Second, have existing standards failed us? Here, I think, the answer is no. They represent important first steps of capturing data and information whose complexity is far beyond anything we have had to wrestle with before. And their shortcomings have set the stage for moving forward, provided we do this in an intelligent fashion.

Then third, how do we make standards work better? And the answer here is that we have to, as a community, recognize their value. And I think this is of fundamental importance. While I might argue that this means more funding for standards development, that is not going to be realistic without other changes in the way we see standards development efforts. Standards are not going to cure cancer or show us the evolutionary history of life on earth or give us drought-resistant plants. But standards are going to get us to those endpoints faster. There used to be an advertising campaign for BASF with the tagline, 'We don't make a lot of the products you buy; we make a lot of the products you buy better'. Well, that is what effective standards do: they make the data we have available better, making them something we can use. And that is something we cannot forget. It means that we need to make a conscious decision to invest in standards and to make an effort to reward those who choose to contribute through their creation, implementation and development. Standards development is not going to end up on the front page of *Le Monde* or the *New York Times*. But it is an academic endeavor requiring intellectual investment in the same way that molecular biology requires thought and careful planning.

Standards are built on the idea that the data we capture will be stored in a database. And many in the community do not fully understand what a database is. Put simply, a database is a model. And it is a model of two things - the relationships between various elements in the data we want to capture and the ways in which we want to use those data. The instantiation of the data model is, in part, the standards for communicating the data, because those standards capture the essential elements within the data to make them useful.

Kary Mullis won the Nobel Prize for his 1983 invention of polymerase chain reaction (PCR). Developing the technique required bringing together all of the pieces - all of which were known - and making an intellectual leap in understanding what they could do. Today, you can buy a kit from numerous vendors and teach a secondary school student to use PCR to analyze DNA. But molecular biology was, and remains, an accepted academic discipline. In genomic data meta-analysis, we are still in the early days; we have all the pieces and we are looking for the right ways to put them together to drive science forward faster than is now possible, and in ways that we are only starting to imagine. But without standards and the well-annotated data that they would provide, we're left without the basic tools we need to make progress. The entirety of the data is much more valuable than individual studies, but only if we know what the data represent. To paraphrase Shakespeare, a rose by any other name is still a rose; you just can't find it in your database.

Fourth, and finally, how do we enforce standards? This is in many ways the most complex question to answer, and one that will involve a community approach. The culture of biological research has been based on protecting data, releasing them only at the time of publication, and often releasing only the minimum mandated as necessary for acceptance by a particular journal. The human genome project used a different model, with rapid data release, but that model has neither been widely adopted beyond the genomics community nor survived the advent of broad genomic studies. The journals that do have data-reporting standards, and the referees who serve those journals, often do not have the time to fully explore the extent of the annotation provided by authors. Funding agencies mandate data release, but often fall to a minimum standard and fail to actively enforce this requirement. The only way to fully address this question is to forge a partnership in which all interested parties - authors, funders, and publishers - commit to making fully annotated data available. But that, again, will require the commitment of resources. Short of that, we need to reward those who make the effort to make data available, but that too will require that we develop new measures of value in science that go beyond counting impact factor or citations and consider things like downloads of datasets and web hits as measures of the importance of a particular dataset.

Science has come a long way from the time when cutting-edge molecular biology was running gels. With the rapid advances in technology that we have seen, we need to assure that the investment we make in generating data is not wasted. That means we need to make sure that our data-reporting standards keep up with our ability to generate data. The age of the $1,000 genome is likely not more than five years off, and we need to guarantee that the data tsunami that is about to wash over us is both useful and used. We need to make effective data-reporting standards as well as the software tools to implement them and facilitate their use. And we need to recognize the value of data and those who produce it by enforcing data-reporting standards. Because in the end we will all benefit. Standards truly do make the data we generate better by making sure they have a life beyond their initial publication.

Competing interests

John Quackenbush is a member of the board of directors of the MGED society, a not-for-profit organization dedicated to the development of genomic data standards.

References

1. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.
2. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA: **ArrayExpress - a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31:**68-71.
3. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29:** 365-371.
4. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A: **Design and implementation of microarray gene expression markup language (MAGE-ML).** *Genome Biol* 2002, **3:** RESEARCH0046.
5. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Irizarry RA, Liu J, Maier DS, Miller M, Petersen K, Quackenbush J, Sherlock G, Stoeckert CJ, Jr., White J, Whetzel PL, Wymore F, Parkinson H, Sarkans U, Ball CA, Brazma A: **A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB.** *BMC Bioinformatics* 2006, **7:**489.
6. Landemaine T, Jackson A, Bellahcene A, Rucci N, Sin S, Abad BM, Sierra A, Boudinet A, Guinebretiere JM, Ricevuto E, Nogues C, Briffod M, Bieche I, Cherel P, Garcia T, Castronovo V, Teti A, Lidereau R, Driouch K: **A six-gene signature predicting breast cancer lung metastasis.** *Cancer Res* 2008, **68:**6092-6099.
7. Culhane A, Schwarzl T, Sultana R, Picard KC, Picard SC, Lu TH, Franklin KR, French SJ, Papenhausen G, Correll M, Quackenbush J: **GeneSigDB - a curated database of gene expression signatures.** *Nucl Acids Res* 2009, in press.
8. Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, Austin CP, Berglund L, Bobrow M, Bountra C, Brookes AJ, Cambon-Thomsen A, Carter NP, Chisholm RL, Contreras JL, Cooke RM, Crosby WL, Dewar K, Durbin R, Dyke SO, Ecker JR, El Emam K, Feuk L, Gabriel SB, Gallacher J, Gelbart WM, Granell A*, et al.*: **Prepublication data sharing.** *Nature* 2009, **461:**168-170.
9. Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, Einhorn D, Tocchini-Valentini G, Hrabe de Angelis M, Rosenthal N: **Post-publication sharing of data and tools.** *Nature* 2009, **461:**171-173.

Published: 25 November 2009
doi:10.1186/gm111