Calls for New Reporting Standards, Quality Control in Microarrays

By Rabiya S. Tuma

Concerns about the quality of microarray data in three Duke University cancer trials have simmered for several years. Then in July, after evidence surfaced regarding alleged inconsistencies in reporting of credentials by one of the Duke researchers, the issue came to a boil: More than 30 bioinformatics and biostatistics specialists sent a letter to National Cancer Institute director Harold Varmus, M.D., and others expressing concern about the reproducibility of the data and asking that the trials be suspended. Within 3 days, they were.

Now, using the momentum from that incident, the group of biostatisticians and bioinformatics specialists is calling for improved reproducibility and transparency in genomics experiments. Writing under the designation reproducible research, a small group of biostatistics and bioinformatics specialists is calling for improved reproducibility and transparency in genomics experiments. Writing under the designation reproducible research, a small group of biostatistics and bioinformatics specialists is calling for improved reproducibility and transparency in genomics experiments. Writing under the designation reproducible research, a small group of biostatistics and bioinformatics specialists is calling for improved reproducibility and transparency in genomics experiments. Writing under the designation reproducible research,

For Duke’s microarray data, for example, Baggerly and his colleague Kevin R. Coombes, Ph.D., professor and deputy chair of bioinformatics and computational biology at M. D. Anderson, tried to reproduce the studies, only to find that the method sections in the published articles were inadequate. Baggerly estimates that they spent 1,500 hours trying to re-create what was done. Instead of reproducing the outcomes that Anil Potti, M.D., and colleagues, reported, the Houston-based team uncovered many mistakes. For instance, they found that the Duke researchers had shifted the gene names relative to the samples—an off-by-one error—so that the name no longer accurately identified the sample.

To ensure that independent scientists don’t have to go through that sort of backtracking, which Baggerly calls forensic bioinformatics, the Scientists for Reproducible Research Working Group wants journal editors to require study authors to supply more detailed methods. Specifically, they say that authors should provide access to the entire data set that the experiments were based on, as well as information on the source of the data, including database accession numbers or appropriate URLs. They also want access to the software code used for the analysis and any instructions needed to run it; step-by-step information of any analytic or data manipulation steps performed by hand and thus not included in the code; and any detailed prespecified research protocols.

Moreover, if the study authors decline to furnish any of this information, the group wants them to have to explain that decision. “We actually think that might result in a lot more people supplying the data and code because, quite frankly, most people don’t want to have to explain why they are not supplying data and code,” Baggerly said. Also, if study authors decline to provide the software code because they want to protect future intellectual property rights, readers may view the work more critically, as they might if an author had a financial conflict of interest.

Both Code and Data

Access to the software code would help with reproducibility because it is an exact record of what was done. “Simple mistakes are here,” Baggerly said. “They are not going to go away. And it is not just us. It is not just Duke. If there is an off-by-one error, and you have given me the code, I can recognize that, and I can fix it. But if you don’t tell me what the code is, then the odds are pretty good that I won’t even notice it.”

The code would help, but it won’t ensure reproducibility on its own, according to Rafael A. Irizarry, Ph.D., professor of biostatistics at Johns Hopkins Bloomberg School of Public Health in Baltimore, who was a key instigator of the letter to Varmus and the follow-up working group proposal. He said that many scientists keep track of their samples or data by using point-and-click software, such as spreadsheet programs. Then no record of the scientists’ procedures exists unless they write down every manipulation step by step.

Although the request for software code and raw data might look burdensome, it may be...
a sensible thing to do and may not be that much extra work, according to David Ransohoff, M.D., a professor in the department of medicine and epidemiology at the University of North Carolina at Chapel Hill, who has written about biomarker discovery and reproducibility. “I think there are many labs that already keep careful track of their own analyses in such a way that simply putting them on the Web would not be that much effort,” he said. “In other words, you can make the case that if you make a paper trail that many good scientists would simply want to do on their own, then it is no huge burden to stick it on the Web for others to look at.”

He also noted that the whole point of having a method section in scientific reports is to allow others to reproduce, check, or build on what the authors have done. From that point of view, the working group’s proposal is really about reestablishing fundamental approaches in science. He said that many published reports look convincing, but the only way to know whether it is a house of cards or a house of bricks is to examine the insides. “If you are a reviewer, you can judge whether or not it is strong,” Ransohoff said. “And if you’re an investigator, you can reproduce it on your own. That is the fundamental principle. It is that simple.”

MAQC-II: Testing Reliability
The organizers of the MAQC-II study also confronted nearly all the issues that the Working Group raised. The study, which started in 2006, was designed to test the reproducibility and reliability of predictive models developed from microarrays. “We were hearing different voices from the field with some people saying microarrays were reliable and others saying they were just noise discovery and nothing valid,” said Leming Shi, Ph.D., of the National Center for Toxicological Research at the U.S. Food and Drug Administration in Jefferson, Ark., who led the project. “So we set out to engage the community to have a check on what is really going on in the community,” said Shi, who is also a member of the Scientists for Reproducible Research Working Group.

Thirty-six research groups participated in the project and developed predictive models based on six microarray data sets and 13 endpoints, including blinded positive and negative controls. The project sought to predict lung and liver toxicity in rodent models and clinical endpoints such as disease progression in breast cancer, multiple myeloma, or neuroblastoma patients. One of the primary findings was a surprise for Shi. “Given the same data set, you can see models developed by different teams yielded quite dramatic differences in terms of prediction performance,” he said.

However, many of the top-performing teams obtained similar results. Shi noted that these groups tended to come from industry. The bottom 25% of the teams came from academic institutions and had members, such as graduate students, with relatively little experience, he said. Shi thinks the trends they saw in MAQC-II reflect what is happening in the broader scientific community, including poor reproducibility of many studies in the literature. “Even in this project, when we asked teams what they did, it was a challenge for some teams to tell us. It took several months to dig out the details of exactly what they did,” he said. “I think this is something the scientific community should be thinking about: how to more easily replicate results.”

Blinded Data Set
From Baggerly’s point of view, efforts like MAQC-II are important because they can highlight the limits of the technology, even when it is done right. For example, the MAQC-II researchers found that the endpoint affected the predictive ability of the models. That is, some endpoints were more amenable to the approach than others, regardless of the expertise of the team developing the test.

One key aspect of the MAQC-II study was that the teams had access to data on which to build and internally validate their model, but then their models were frozen and tested on a completely blinded data set. “We imposed this very strictly in MAQC-II, and the [external] validation set was not seen by anyone in the model development process,” Shi said. “You can only test it once. If you go back and change your model and predict again, then the validation set is no longer blinded. The unfortunate reality is, in the scientific community, we often lose the blinded data set.” Instead of reserving a data set for a final independent test of the model, researchers often use an iterative approach—developing a model, testing it, and then modifying it—which compromises the independence of the validation test.

Irizarry thinks the MAQC-II study used the right approach when it required that the models be frozen and tested on a blinded data set. “There should be no other way than this to assess whether prediction algorithms are working,” he said. “You should have a data set saved that nobody ever sees.” In fact, he thinks that if the FDA used that level of test to validate predictive models, “it would protect us from bad algorithms.”

Shi said that the MAQC-II researchers have put together hundreds of pages of documentation from the project, only a fraction of which was reflected in the recent Nature Biotechnology publication. “We really think the composition of the MAQC-II mimics the larger scientific community,” he said. “So we are in discussions about how to take these results further . . . so some of the errors that we had—and that the community has already had—can be avoided in the future.” One possibility is to give the FDA information for a draft guidance on predictive model development and validation, but thus far the MAQC-II participants have not reached a consensus on exactly what to include.

For Ransohoff, though, the issues that MAQC-II and the Scientists for Reproducible Research Working Group addressed have much broader implications than microarrays or high-throughput experiments. “This is a broad and deep story. . . . It is not just about genomics. It is about how we write and report and test science,” he said. “It is about the whole field of science and how we either build it with bricks or with things that aren’t bricks.”