Omnibus or ArrayExpress databases clearly separated from the peer-reviewed published data sets.

We can add some further suggestions. Jafari and Azuaje (2006) surveyed publications for common statistical analyses, tracking how various factors in experimental design are typically reported, for example, sample size and statistical power calculations, normalization methods, one- or two-sided *t*-tests, missing values, and homogeneity of variances. A similar literature survey of typical quality assessment practices and reporting would be extremely useful. Large government-funded projects, such as the Cancer Genome Atlas project, might serve as examples of good quality reporting. It would be interesting to see and compare all of the unpublished data from the various Cancer Genome Characterization Centers, in addition to the published data. It also would be interesting to see the internal quality assessments made by the National Cancer Institute or National Center for Biotechnology Information. We also can expect microarray manufacturers to increasingly add quality control and normalization features to their arrays and advertise their advantages. As these quality assessments are built into the arrays and processing software, the information will be incorporated into the routine reporting of experimental results.

Allison, Cui, Page, and Sabripour (2006), in their recent review of microarray data analysis methods, pointedly noted that, “the usefulness of most QC measures is unsubstantiated and no specific QC method has been embraced by the community.” Progress in this area will be made by detailed comparisons of various proposed methods, indicating the range of useful applications of each method and their relative strengths and weaknesses. Beyond that, further research is needed into the theoretical foundations of quality assessment of high-dimensional data sets and the interrelationships among experimental design, microarray design, data normalization, various statistical analyses, and quality control methods. We look forward to continuing our own research in these areas and also following the progress of others. Finally, we anticipate that the most interesting results will come from a deep synthesis of currently disparate investigations.

ACKNOWLEDGMENTS

The authors thank Mike Wigler, Kenny Ye, Boris Yamrom, and Mike Ronemus for helpful discussions. This work was supported by the Simons Foundation.

ADDITIONAL REFERENCES

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni Jr., J. F., Freimer, N. B., Gerhard, D. S., Gunter, C., Guttmacher, A. E., Guyer, M. S., Harris, E. L., Hoh, J., Hoover, R., Kong, C. A., Merikangas, K. R., Morton, C. C., Palmer, L. J., Phimister, E. G., Rice, J. P., Roberts, J., Rotimi, C., Tucker, M. A., Vogan, K. J., Wacholder, S., Wijsman, E. M., Winn, D. M., and Collins, F. S. (2007), “Replicating Genotype–Phenotype Associations,” *Nature*, 447, 655–660.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N. E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V., Yoon, S., Wigler, M., Ye, K., Børresen-Dale, A.-L., Naume, B., Schlicting, E., Norton, L., Hägerström, T., Skoog, L., Auer, G., Månér, S., Lundin, P., and Zetterberg, A. (2006), “Novel Patterns of Genome Rearrangement and Their Association With Survival in Breast Cancer,” *Genome Research*, 16, 1465–1479.

Jafari, P., and Azuaje, F. (2006), “An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors,” *BMC Medical Informatics and Decision Making*, 6.

Kendall, J., Liu, Q., Bakleh, A., Krasnitz, A., Nguyen, K. C. Q., Lakshmi, B., Gerald, W. L., Powers, S., and Mu, D. (2007), “Oncogenic Cooperation and Coamplification of Developmental Transcription Factor Genes in Lung Cancer,” *Proceedings of the National Academy of Sciences*, 104, 16663–16668.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C. Q., Powers, S., Ye, K. Q., Olshen, A., Venkatraman, E., Norton, L., and Wigler, M. (2003), “Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation,” *Genome Research*, 13, 2291–2305.

MAQC Consortium (2006), “The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements,” *Nature Biotechnology*, 24, 1151–1161.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K., and Wigler, M. (2007), “Strong Association of de novo Copy Number Mutations With Autism,” *Science*, 316, 445–449.

Wellcome Trust Case Control Consortium (2007), “Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls,” *Nature*, 447, 661–678.

Comment

Darlene R. Goldstein

Ecole Polytechnique Fédérale de Lausanne
Institut de Mathématiques and
Swiss Institute of Bioinformatics
CH-1015 Lausanne, Switzerland

The authors provide an exceptional contribution to the growing literature on quality assessment of microarrays. Their article emphasizes the frequently overlooked importance of quality assessment in the analysis of microarray data and supplies investigators with important new quantitative tools to carry it out. It is also a substantial benefit to the community that the authors provide software implementing these methods to the BioConductor project. This ready availability should allow NUSE

© 2008 American Statistical Association
TECHNOMETRICS, AUGUST 2008, VOL. 50, NO. 3
DOI 10.1198/004017008000000370

plots and quality landscapes to become standard tools for short oligonucleotide microarray quality assessment.

It is clear that even within the microarray community, there is no consensus on the meaning of "quality," nor any agreed upon standards prescribing quality assessment protocol. Anomaly detection is a sound first step in data analysis, yet it is insufficiently recognized that different quality measures are aimed at different aspects of quality. The measures proposed by Affymetrix concentrate mostly on experimental noise in the hybridizations and on the integrity of the sample RNA, whereas the RMA-based measures focus on outliers at the level of measured expression (Jones et al. 2006). The issue becomes the identification of anomalies that are of interest. The determination of which artifacts matter most will depend greatly on the measure used to quantify expression and also on the analysis aims. Where inference is based on robust gene expression measures such as RMA, then such characteristics as streaks or small bright areas on the chip image should have little effect on subsequent inference and thus should not be grounds for chip exclusion (as suggested by some guidelines).

I now comment on some of the issues brought to light in the figures and the discussion section.

Sample Quality versus Hybridization Quality (fig. D). The fact that chip 12 is an outlier in both the A and B chips suggests a problem in the sample rather than a hybridization artifact. One of the first questions that investigators ask on QC failure is whether they should attempt to rehybridize the sample. Thus separation into hybridization artifacts and sample problems is a useful feature of the quality assessment tools presented here.

There is a distinction between large but localized (and removable) flaws and smaller but more pervasive flaws. Patterns in the quality landscape, such as those shown in figure J, are more suggestive of technical/experimental problems and indicate that a rehybridization has a good chance of succeeding. A more uniform distribution of outliers across the whole of the chip would tend to signal problems with the sample, in which case rehybridization would not be advised.

Spatial statistics also may be of use in automatic detection of a nonuniform distribution of the weights across the chip. At a more crude level, the chip could be divided into a set of regions and a chi-squared goodness-of-fit test for uniformity applied.

Experimental Design Considerations (fig. F) In addition to the dependency between quality measures and hybridization date, these figures also show a large degree of confounding between date and mutant type. Any comparisons between mutants hybridized on different days thus are likely to be very misleading. In principle, it should be easy to avoid this problem in the first place by hybridizing different mutants on the same day. This "importance of a well-designed experiment" also is emphasized in the conclusion section.

For practical reasons, hybridizations often are batched by condition (e.g., mutant type), but the pitfalls of this practice are obvious, and biologists need to be made more aware of the risks involved in such a strategy. Many biologists are more used to thinking of "design" as the sample size in each group, which time points to study, and so on, and do not realize that, for instance, the assignment of samples to hybridization days also is an important aspect of design. These graphs might be shown to biologists planning an experiment as a forceful argument against this type of design.

Statisticians also would do well to recognize that there often is a trade-off between optimal design and robust design. A design that is "optimal" according to a statistical criterion may be useless if the experiment cannot be successfully executed by biologists. The practicalities of sample acquisition and storage, along with the risks of mislabeling or cross-contamination, are among the challenges facing experimentalists. Such difficulties argue for a design that is simple enough to be properly executed. Thus a balance must be struck that avoids as much as possible this kind of confounding while not being too complicated logistically to carry out without errors.

Unreliable Chips versus Outlier Chips (figs. G, H, and I). I have a quibble with the terminology used in several figures, including figure G—the authors seem to be equating "different" (or "outlier") with "lower quality." Instead, maybe what is being signaled in the plots in figure G is the fact that the normalization algorithm has failed to remove the hybridization day artifact, not that the second replicate data are "bad" and should be removed. My interpretation of these plots is that it seems that a more appropriate means of normalizing across the replicates is needed. A more detailed example follows.

In their comments on figure H, the authors state that "[the NUSE and RLE boxplots] show systematically much better quality in Lab M than in Lab I." Similar comparative remarks are made with respect to figure I. Again, what are essentially systematic differences in model fit are interpreted as "better quality," a value judgment that I am not sure is justified.

Would the authors propose discarding the Lab I chips (and thus half of what must be a considerable sum of money) because they are of worse quality? What does quality look like within a site? It is not clear that because the model fits better to the Lab M chips, the Lab I chips should be considered "bad." Agreed, the PM plots show that the Lab I chips seem to be rather brighter than typical, but the Lab M chips, in contrast, seem somewhat dimmer than might be expected. The model may be capturing relatively unimportant aspects that are just different between laboratories.

The following example demonstrates that within-study quality can appear acceptable, but when combining the chips and then assessing quality, some of the chips become outliers. One study uses a subset of the original St. Jude's Hyperdip 50 HG U95Av2 chips; the other uses a subset of chips from an ovarian cancer study (Lancaster et al. 2006). Unlike the Pritzker two-lab example, these studies are completely unrelated; however, it does not seem unreasonable to use them to illustrate the point.

Figure 1 presents boxplots of PM, NUSE, and RLE for each study analyzed separately. A notable difference between the two studies is that for the Hyperdip 50 chips, the PM intensities are both higher and more compressed. Within the study, there do not appear to be any outlier chips. For the separate studies, all chips would be judged to be of "acceptable quality."

Figure 2 shows boxplots of the same measures derived from a mixture of chips from both studies. All chips are used as input to the probe-level model, from which subsequent quality measures are derived. Several of the Hyperdip 50 chips are now NUSE outliers and thus might be considered "low quality"; in addition, the Hyperdip 50 chips in general have larger RLE. But nothing about the chips themselves has changed—rather, the measures have changed based on new relationships between

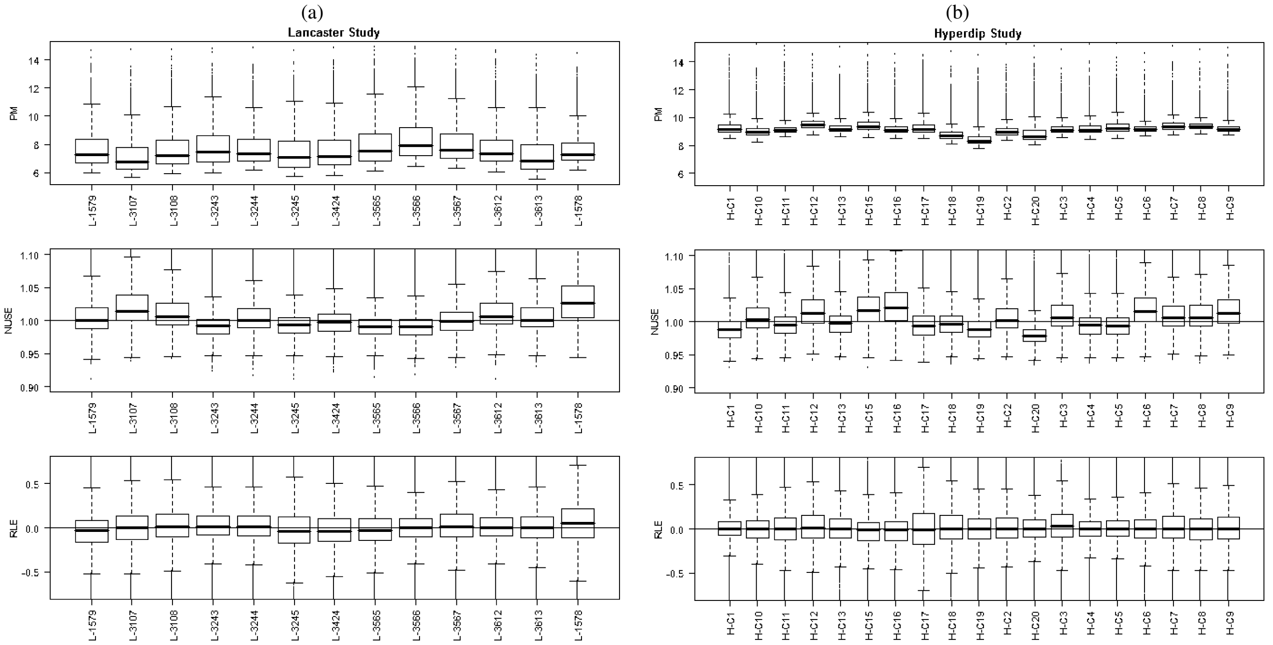

Figure 1. Boxplots of within-study PM, NUSE, and RLE quality measures for Lancaster (a) and Hyperdip (b) chips. No chips are obvious outliers.

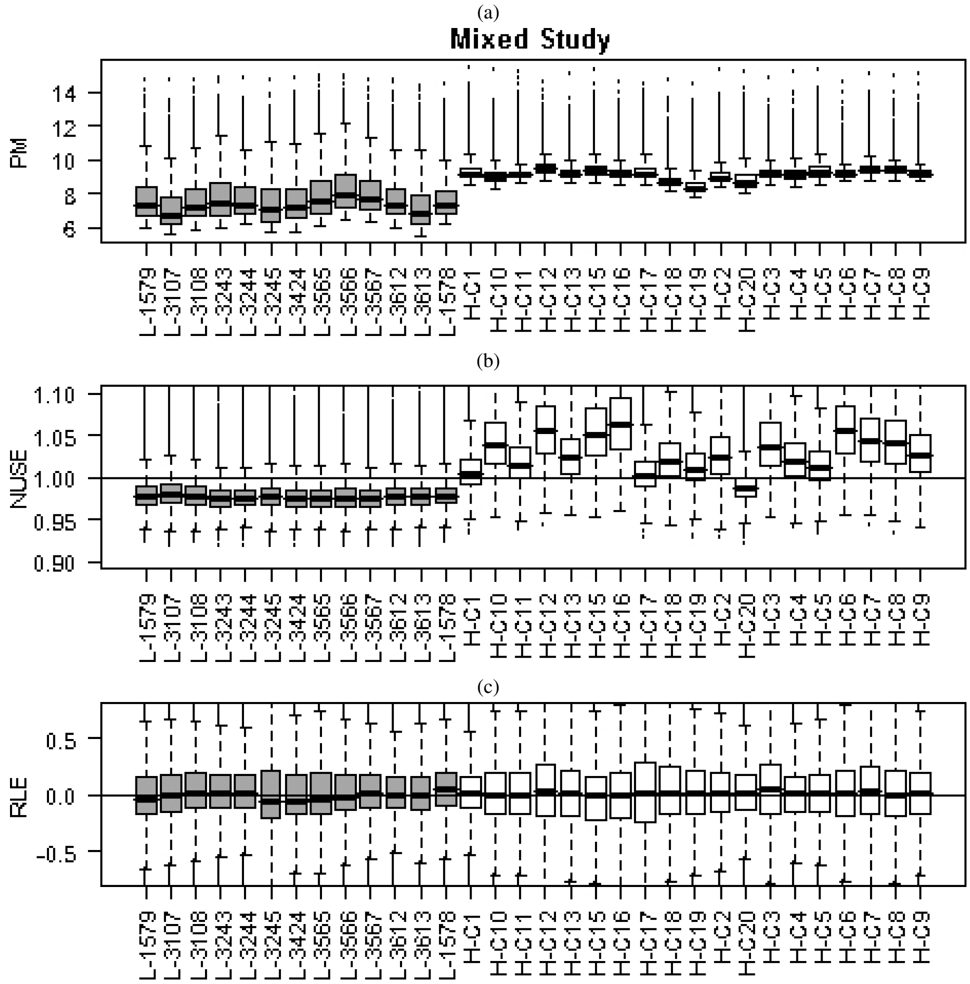

Figure 2. Boxplots of PM (a), NUSE (b), and RLE (c) quality measures for the combined set of Lancaster (gray) and Hyperdip chips. Several Hyperdip chips are now apparent NUSE outliers, and H–C17 is a possible RLE outlier.

different chips. These two aspects should be somehow separated: chips that are of low quality (i.e., unreliable) and chips that are outliers (which may just be different rather than “bad”).

An alternative interpretation is that the Pritzer figures highlight the need for a site effect that the normalization algorithm has failed to provide, because it is apparent that measurements of the same sample between the two laboratories are not at all comparable. It is often thought that carrying out RMA on a combined set of chips will remove such artifacts, but in general, batch variation of some sort still persists (Goldstein et al. 2009). This finding has great implications for combining data, either from multisite studies or across independent studies. Some type of additional adjustment (e.g., random effect modeling, within-site centering and/or scaling, use of relative measures) is required to make the data combinable.

Artifacts in Quality Landscapes (fig. J). Here a different meaning of the term “poor quality” is used. These figures show several types of artifacts, but most of these are not apparent in the numerical summaries introduced in the article and thus seem unlikely to be problematic if included for downstream analysis. In preprocessing sets containing these chips, I found that J5 is not a NUSE outlier but has a noticeably larger IQR(RLE), and that J6 and J7 are both NUSE and RLE outliers. But J1, J2, J3, J4, and J8 are neither NUSE nor RLE outliers. Clearly, the hybridizations are not perfect, but are they truly “poor quality,” and should they be removed or rehybridized? Rehybridization is not without problems either, because a different hybridization date is potentially a new artifact.

Use of Quality Information. In the discussion section, the authors state that “it remains an open question how to use this kind of assessment beyond the detection and classification of quality problems.” Some of the possibilities here include recommendations for when it may be advisable to rehybridize a failed sample, guidelines on when to remove chips, and examination of strategies involving selective downweighting or removal of probesets from downstream analysis. Even for the “triangle” chip (fig. J7), most of the area of the chip is not affected. Thus one might still be able to use the information by weighting expression measures in downstream analysis (e.g., by estimated variances or effective number of probes).

Robust Downstream Analyses. In the conclusion section, the authors also propose that “when there is uncertainty about whether or not to include a chip… we can do both analyses and compare the results.” In light of their reliance on robust methods throughout, it is curious that they stop short of recommending a robust approach to the downstream analyses as well. In smaller experiments (e.g., three treatment vs. three control subjects), robust methods of analysis of preprocessed data may not be an option. But for larger experiments, in the case where the outliers are not themselves of direct interest, outlier accommodation through robust procedures should be considered.

It is also worth noting that even chips that are of “good quality” overall are likely to have some probesets for which the expression measure is “bad quality.” Robust methods of analysis at the single-gene level would seem to provide a straightforward way to handle this situation.

Standards for Quality Assessment. Many statisticians have made a call for systematic quality experiments, because such studies will provide valuable empirical information that can aid the diagnosis of problems and can be used for quality decisions. But such studies are an extravagance for most experimental laboratories, because chips are not cheap enough to use for this purpose. A concerted effort is needed to obtain sufficient funding, as well as to coordinate the planning and execution of such studies.

I wholeheartedly agree with the authors’ statement that standards must be developed by the community. Care must be taken to ensure that the recommended quality standards are sufficiently dependable and suitable for long-term use, so that practices of questionable utility do not become ingrained. One large-scale attempt at developing standards is the MAQC project, described at *http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/*. A fundamentally sound quality assessment framework is needed for standards development, and the exemplary approach that the authors detail in this article represents an important step in the proper direction.

ADDITIONAL REFERENCES

Goldstein, D. R., Delorenzi, M., Luthi-Carter, R., and Sengstag, T. (2009), “Meta-Analysis of Microarray Studies,” in *Meta-Analysis and Combining Information in Genetics*, eds. R. Guerra and D. R. Goldstein, Boca Raton, FL: Chapman & Hall/CRC Press.

Lancaster, J. M., Dressman, H. K., Clarke, J. P., Sayer, R. A., Martino, M. A., Cragun, J. M., Henriott, A. H., Gray, J., Sutphen, R., Elahi, A., Whitaker, R. S., West, M., Marks, J. R., Nevins, J. R., and Berchuck, A. (2006), “Identification of Genes Associated With Ovarian Cancer Metastasis Using Microarray Expression Analysis,” *International Journal of Gynecological Cancer*, 16, 1733–1745.

Rejoinder

Julia Brettschneider, François Collin,
Benjamin M. Bolstad, and Terence P. Speed

We are most grateful to the editors of *Technometrics* for providing such an excellent platform to discuss quality assessment and control for microarray data. This opportunity helps us to bring the statistical challenges that have emerged with the development of new genomic technologies to the close attention of leading applied statisticians. Furthermore, we are most grateful to the editors for inviting so many fine statisticians to comment on our article. We are impressed by the

TECHNOMETRICS, AUGUST 2008, VOL. 50, NO. 3
DOI 10.1198/004017008000000389