Published in final edited form as:

Nat Biotechnol. 2006 September ; 24(9): 1112–1113. doi:10.1038/nbt0906-1112.

Data quality in genomics and microarrays

Hanlee Ji and

Department of Medicine, Division of Oncology

Ronald W. Davis

Department of Biochemistry and Department of Genetics at Stanford University School of Medicine,
269 Campus Drive, CCSR 1115, Stanford, California 94305-5151, USA.

Abstract

Objective quality control indices are needed to facilitate clinical implementation of DNA microarrays used in transcriptional profiling as well as other types of genomic analysis.

DNA microarrays are increasingly used for investigating gene expression in human diseases with the hope of identifying signatures that correlate with specific clinical outcomes. The discovery of these signatures offers the tantalizing possibility that they could be translated into fully fledged clinical diagnostic tests. Significant hurdles exist, however, in transitioning microarray technology and gene expression analysis into the complicated realm of the clinic. Namely, gene expression genomic data quality, a measure of its general reproducibility and ultimately, its true biological relevance, requires significant improvement¹. For example, comparing gene expression studies using different microarray formats is fraught with difficulty, even under circumstances in which the same type of tissue is analyzed². A recent prominent example illustrates a case where different clinical conclusions were derived from the same gene expression data set³.

Currently, few if any objective metrics or established quality control standards are used to evaluate the quality of microarray studies. Often, the assessment of microarray data quality requires running replicates and making intra-sample comparisons to determine reproducibility. Using replicate arrays is an expensive strategy and cannot be routinely applied where quantities of precious biological samples, such as tumor biopsies, are limited. The majority of clinically related studies do not have replicates, leaving genomic data purveyors little in the way of guidance to determine the overall quality of submitted microarray data. Two major efforts currently under way, however, offer an opportunity to improve genomic data quality for gene expression.

Looking at gene expression data quality

Several studies have addressed the issues of genomic data quality in the realm of gene expression analysis through comparison of different formats of microarrays^{4–8}. To date, the MicroArray Quality Control (MAQC) project—the first results of which are presented in this issue—and the External RNA Controls Consortium (ERCC) are the most comprehensive efforts in assessing and comparing gene expression data derived from common samples among different microarray platforms^{9,10}. Both projects are focused on the analysis of highly calibrated reference RNA pools with the potential for wide distribution to the research community. Analysis of the MAQC and ERCC RNA sets has resulted in extensive gene expression data sets with validation across many microarray platforms and systems (e.g., by

quantitative reverse transcription (qRT)-PCR)^{9,10}. The public release of these results should spawn new applications to evaluate gene expression data quality.

A vital part of the MAQC project has been the identification of common transcripts that are mutually represented among the various microarray platforms included in the analysis⁹. This aspect will enormously facilitate cross-platform comparisons of gene expression and open the door to robust meta-analysis studies in clinical gene expression studies.

The completion of these projects also provides an opportunity to advocate for the adoption of genomic data quality control processes into clinically oriented studies. It is critical that there be wide acceptance of some type of quality control standards at the planning stages of clinically oriented projects. Accurate and routinely reproducible data will improve the clinical validity of molecular signatures and speed transition into the clinical setting. For translational research, adoption of quality control will be faster if there is easy accessibility of quality standards to any size research group.

We anticipate that establishing quality control standards for genomic data will substantially reduce genomic analysis costs by eliminating the need for replicate experiments and improve the design and implementation of large translational studies involving hundreds if not thousands of patient samples. Another benefit is that genomic data quality standards will facilitate future technology development. When established standards exist, it is much easier to conduct proof-of-principle studies using new systems.

Moving beyond RNA

There is a general recognition that quality control standards for transcriptional profiling experiments are an absolute necessity given the complexities of working with RNA, the wide variety of methodologies and different microarray platforms. We suggest that it is equally important to establish such standards for all other microarray formats, including array comparative genomic hybridization (CGH) and genotyping. When microarrays are used, the analysis of DNA has major advantages over RNA in terms of its physical and biochemical properties. Even so, many of the inherent issues of microarray performance and reliability in analyzing RNA are just as relevant.

One could imagine a future consortium, similar to the MAQC and ERCC, developing a universal set of standardized DNA references with known genotypes and gene-copy alterations for use in high-throughput genotyping, sequencing and gene-copy microarray technologies. Many highly characterized DNA samples already exist, the larger hurdle being one of establishing a consensus about the samples to be included. A set of DNA references would enable quality control assessment, facilitate data set comparisons among different microarray platforms and provide a valuable resource for validating new genomic technologies.

Controls to assess genomic data quality

Numerous studies have identified sources of inter- and intralaboratory error and variability in microarray experimental results^{6,7,11}. They include variation in tissue processing, RNA extraction, inherent biological differences in normal tissue and microarray assay protocols¹¹. The MAQC's and ERCC's RNA pools of highly characterized transcripts could be incorporated into the microarray workflow process. For example, an individual site could analyze the RNA pools characterized by the MAQC project to make performance comparisons. Leveraging the MAQC data sets will prompt the development of methods to increase the confidence that differential expression of specific genes will be reproducible. For example, we and others (H.J. and R.W.D., unpublished data; Lin, G., He, G., Shi, L. & Zhong, S., personal

communication) are currently developing algorithmic methods that use the MAQC data set to account for interlaboratory variation in the discovery of differentially expressed genes.

Another strategy for monitoring genomic data quality would rely on highly characterized external controls or RNA pools at every step of a microarray experimental protocol¹². We and others have suggested the incorporation of a universal set of nucleic acid controls tailored to measure performance for individual steps of microarray analysis. Several RNA transcripts or pools, PCR products, oligonucleotides or other external ‘spiked’ controls would be added at every individual step of the microarray analysis process. For example, one could imagine multiple ‘spiked’ external control sequences that would directly measure the quality and subsequent level of degradation of nucleic acid extracted from processed tissues; other controls would assess an assay’s enzyme quality and some would be specific for the hybridization process. These external controls would be assessed via microarray hybridization. To facilitate the development of external controls, one could design synthetic sequences as probes to avoid problems of cross-hybridization and reduce the interfering aspects of nucleic acid secondary structure. Incorporating synthetic sequence probes and targets would be quite similar to the development of oligonucleotide barcodes in microarrays, which has proven to be quite robust¹³. Another application that would improve genomic data quality is the inclusion of universal external controls in different concentrations for normalization in individual microarray experiment. As the final step of a quality control assessment process, the formal report of quality control performance would be incorporated in the resulting data file output.

Building in a quality control assessment and an incorporated report of quality metrics would be enormously useful in a variety of settings (Fig. 1). We offer some hypothetical examples; a genomic data quality report would assist the individual researcher in measuring the performance of an experiment ‘on-the-fly,’ and provide journal editors with some external criteria to judge the quality of submitted data sets, and embedded quality metrics in genomic data reports would substantially quicken the complicated task of regulatory agency analysis and review. A universal set of quality control reagents for genomic data quality assessment also has the potential to decrease costs. Individual researchers could assess their genomic data quality at the very beginning of a project and avoid costly mistakes. Like its MAQC and ERCC predecessors, any future efforts would require general agreement and coordination among the research community, government agencies, microarray manufacturers and producers of biological reagents. We believe this is a realistic goal.

Conclusions

The completion of the MAQC and ERCC collaborative projects sets the foundation for future consortia working towards universal genomic data quality control standards. These projects herald a movement in the genomics community to improve the reliability of microarray technologies in both basic and clinical research. Perhaps our greatest aspiration is that through these efforts we will improve genomic data quality sufficiently to spur rapid development of the next generation of genomic diagnostics and thus have a positive impact on the provision of human healthcare.

References

1. Steinmetz LM, Davis RW. *Nat Rev Genet* 2004;5:190–201. [PubMed: 14970821]
2. Tan PK, et al. *Nucleic Acids Res* 2003;31:5676–5684. [PubMed: 14500831]
3. Tibshirani R. *N Engl J Med* 2005;352:1496–1497. [PubMed: 15814892]
4. Jarvinen AK, et al. *Genomics* 2004;83:1164–1168. [PubMed: 15177569]
5. Bammler T, et al. *Nat Methods* 2005;2:351–356. [PubMed: 15846362]
6. Irizarry RA, et al. *Nat Methods* 2005;2:345–350. [PubMed: 15846361]

7. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. *Nat Methods* 2005;2:337–344. [PubMed: 15846360]
8. Shi L, et al. *BMC Bioinformatics* 2005;6(Suppl 2):S12. [PubMed: 16026597]
9. Consortium MACQ. *Nat Biotechnol* 2006;24:1151–1161. [PubMed: 16964229]
10. Baker SC, et al. *Nat Methods* 2005;2:731–734. [PubMed: 16179916]
11. Cobb JP, et al. *Proc Natl Acad Sci USA* 2005;102:4801–4806. [PubMed: 15781863]
12. van Bakel H, Holstege FC. *EMBO Rep* 2004;5:964–969. [PubMed: 15459748]
13. Eason RG, et al. *Proc Natl Acad Sci USA* 2004;101:11046–11051. [PubMed: 15258289]

Figure 1.

DNA microarray analysis of human tissues involves multiple steps and protocols. As a result, these assays are susceptible to variance throughout the process. Improved methods are needed for monitoring experimental variation during this workflow.

(Pictures used with permission of Affymetrix.)