Of fish and chips

Gavin Sherlock

Since the introduction of microarray technology into the biologist's arsenal, there have been concerns about the reproducibility of experimental results obtained using different microarray platforms. In this issue, three articles address this point, and show that with carefully designed and controlled experiments using standardized protocols and data analyses, reproducibility across platforms is much better than previously shown.

When reading the summary statement of a grant proposal, the last words you typically want to read are 'fishing expedition', as this is usually the kiss of death for the hard work you put into your application. However, microarray studies have often had minimally formulated hypotheses (some genes will go up, some will go down) and can involve a large amount of fishing, casting the net far and wide. The question is, when we've sifted our genes through the net, are they worth keeping or should we be throwing them back? On pages 337, 345 and 351 of this issue, three separate studies[1–3] tackle the issue of the reproducibility of microarray experiments, both across platforms and between laboratories. In each case, by using standardized protocols, they demonstrate that reproducibility is better than previously thought, though there are still some nagging discrepancies.

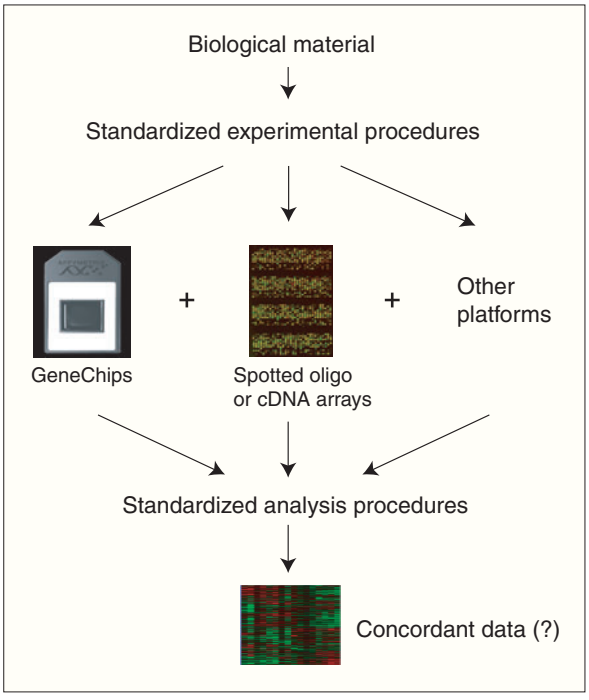

Figure 1 | The use of standardized laboratory procedures for the preparation of labeled nucleic acid, and standardized methodologies for data normalization, filtering and analysis can result in concordant data, both across microarray platforms and between different laboratories.

Although some early microarray studies by different laboratories, using different microarray platforms to study the same question, validated each other's findings in a broad sense (for example, the work of Cho *et al.*[4] and Spellman *et al.*[5]), research specifically directed toward the question of reproducibility painted a much bleaker picture. Kuo *et al.*[6] looked at microarray data for the NCI60 cell lines, generated in their lab using the Affymetrix platform and by Ross *et al.*[7] using the spotted cDNA platform. Comparing both the ratios and the spot intensities from the cDNA platform to the Affymetrix Average Difference metric, they found very poor correlation between the two datasets, and suggested that probe-specific factors influence measurements differently in the two platforms. It should have been entirely expected that these particular metrics would not be correlated because even for the same transcript we would expect different hybridization properties (and thus signals) for the probes on the different microarrays. They concluded, however, that the prognosis for the integration of gene expression measurements across platforms was poor. It is also important to note that the Affymetrix and cDNA array experiments were performed entirely separately, in different labs. In contrast, Tan *et al.*[8] recently investigated reproducibility of microarray data by hybridizing identical RNA preparations to three different commercial array platforms, Affymetrix GeneChips, Agilent cDNA arrays and Amersham Codelink arrays. They found little overlap between the lists of genes that showed significant gene expression changes across the different platforms (the opposite of what was found by Yauk *et al.*[9]). Tan *et al.*, who are extensively cited in a recent report in *Science*[10], provide a strongly negative picture of the reproducibility across microarray platforms and hence the reliability of microarray data in general. It is not yet clear what impact this study has had on researchers' attitudes toward microarrays.

In the three articles presented in this issue, the picture presented is much more positive, though certainly not yet perfect.

Department of Genetics, 300 Pasteur Drive, Stanford University Medical School, Stanford, California 94305-5120, USA.
e-mail: sherlock@genome.stanford.edu

Larkin *et al.*[1], in a carefully controlled study, used both Affymetrix GeneChips and spotted cDNA arrays to examine a clearly defined biological question: the gene expression changes in mouse heart in response to treatment with angiotensin II (used as a model for hypertension). Using standardized procedures for all steps of the experiment, including the data analysis pipeline, they show, using two-factor analysis of variance (ANOVA, which is used to quantify the sources of variation in the experiment), that for 5,853 genes that were measured by both platforms, in the vast majority of cases (88%) the microarray platform used had no significant effect on the expression levels observed.

Both Irizarry *et al.*[2] and Bammler *et al.*[3] take this a step further, by making comparisons not only across platforms but also across laboratories. In the case of Irizarry *et al.*, a consortium of ten labs compared data generated from Affymetrix GeneChips, spotted cDNA arrays and spotted long oligonucleotide arrays using identical RNA samples. They determined that there were sometimes large differences between laboratories, even with the same platform, but that data from the best-performing labs (where data are reproducible within the lab) agree with each other rather well. This suggests that it is not an inherent problem of the technology *per se*, but rather with the use of the technology, such that data can be reproducible across both labs and platforms when good technique is employed. This point is illustrated further by Bammler *et al.*[3]. Their initial experiments, in seven laboratories, showed that although reproducibility for a platform within a single laboratory was good, reproducibility between platforms and across laboratories was generally poor. However, the implementation of standardized protocols for all aspects of the study, both experimental and computational—such as RNA labeling, hybridization, data acquisition and data normalization—dramatically increased reproducibility. Clearly, this demonstrates that making the full raw data available (including the original images) will make it easier to combine data from different platforms.

Yet just because standardized protocols are adopted, does that mean that the data generated using them are correct, or do the protocols simply introduce some overwhelming bias that makes all the data look similar? Larkin *et al.*[1] used quantitative RT-PCR (qRT-PCR) as a gold standard to validate their findings. Through quadruplicate PCR runs, they showed that for 10 randomly chosen genes out of all those that showed concordant measurements between the platforms, the qRT-PCR validated the expression data. However, for 11 genes for which the measurements differed between platforms, only 1 of the qRT-PCR reactions gave robust confirmation of the data obtained with one platform as compared with the other, whereas for the other 10, neither platform was validated—instead the qRT-PCR reactions yielded a third expression profile. It is possible that for the 11% of genes that were discordant between platforms, the two platforms may have measured different variants of the same gene. Larkin *et al.* found that for discordant genes, the probes from the different platforms were significantly more likely not to map to each other than was true for concordant genes. In other words, although the probes intended to assay the same gene, for genes with discordant measurements, the Affymetrix probes were less likely to be contained within the expressed sequence tag (EST) sequences for that gene on the cDNA array. Furthermore, the alignment of multiple EST sequences to the genome in the region containing the array probes generally suggested the existence of multiple splice variants in regions where there was commonly only a single annotated gene structure in the EnsEMBL database. This lends further weight to the hypothesis that discordant measurements may be the result of measuring different splice variants of the same gene.

The three papers in this issue provide a cautionary tale for microarray research, but also a reason for optimism as compared with earlier studies. They demonstrate that it is possible to perform microarray experiments that are reproducible between labs and across platforms, provided standard methodologies are adopted for best performance. In the case of discordant genes, it appears that the discrepancies may in large part be due to the two array types measuring different variants of the same gene. This analysis required access to the sequences on the commercial array, which are readily available for the Affymetrix platform, though that is certainly not true for all commercial vendors. In addition, for cDNA arrays, typically only the ends of the cloned sequences (ESTs) are available, so there is some uncertainty as to what intervening exonic sequences are present in the clone. It is likely that as we gain a better understanding of the transcripts encoded by genomes, and laboratories start to predominantly use short or long oligonucleotides instead of cDNA microarrays, we will have a better understanding of whether probes on different platforms are really assaying the same thing, and if so, whether they are doing it reproducibly. These studies highlight the need for array manufacturers to be more open about the probes on their arrays so that researchers can better understand what was measured, and also the need for authors to accurately record how they performed their experiments, by providing fully MIAME-compliant annotation[11] as well as the complete raw data generated by the experiment—up to, and possibly including, the scanned images of the microarrays. They also indicate that when a group of genes has been deemed to be important in a microarray experiment, validation, using a different technology, of at least a subset of the data is advisable. Though we may still be fishing, adopting standardized procedures should improve the catch.

1. Larkin, E., Frank, B., Gavras, H., Sultana, R. & Quackenbush, J. *Nat. Methods* **2**, 337–343 (2005).
2. Irizarry, R.A. *et al. Nat. Methods* **2**, 345–349 (2005).
3. Bammler, T. *et al. Nat. Methods* 2, 351–356 (2005).
4. Cho, R.J. *et al. Mol. Cell* 2, 65–73 (1998).
5. Spellman, P.T. *et al. Mol. Biol. Cell* **9**, 3273–3297 (1998).
6. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. & Kohane, I.S. *Bioinformatics* **18**, 405–412 (2002).
7. Ross, D.T. *et al. Nat. Genet.* **24**, 227–235 (2000).
8. Tan, P.K. *et al. Nucleic Acids Res* **31**, 5676–5684 (2003).
9. Yauk, C.L., Berndt, M.L., Williams, A. & Douglas, G.R. *Nucleic Acids Res.* **32**, e124 (2004).
10. Marshall, E. *Science* **306**, 630–631 (2004).
11. Brazma, A. *et al. Nat. Genet.* **29**, 365–371 (2001).