confer a number of benefits. "As annotation shifts, people's bets on which sequence is useful for a gene can prove wrong," says Blume. "This design philosophy of multiple sequence probes per gene provides a buffer that single sequences cannot." It also protects against unexpected glitches at the hybridization stage. In addition, Affymetrix offers broadened coverage through its genomic-tiling arrays and, more recently, new exon arrays that allow users to assemble detailed, genome-wide exon-usage profiles for human, mouse and rat studies.

Probing for answers

Illumina also uses redundancy to maintain experimental quality control in its bead-based arrays. "Each of our arrays has about 30 replicates of each probe," says Baker. "Because these 30 measurements are spread randomly across the chip, we don't have to worry about little things like smudges on the array — any outlier measurements get removed." These arrays also benefit from the combination of high probe density with the inclusion of multiple arrays on a given chip. This allows users to simultaneously profile a number of samples — up to 96 parallel arrays — in an 'array of arrays' format.

Agilent Technologies of Santa Clara, California, touts the use of long probes — 60mers, compared with Affymetrix's 25mers — as an advantage for enhancing sensitivity to low-abundance transcripts, typically a weakness for microarray platforms. Agilent's instruments incorporate a multiple-scan approach, further extending the sensitivity of detection. "You can look at a broader range of transcripts and still get linearity with regard to the signal recorded," explains Kevin Meldrum, director of genomics marketing. Agilent has also incorporated proprietary 'spike-in' controls into its platform, which allow monitoring of experimental quality.

An efficient and cost-effective production process gives NimbleGen Systems of Madison, Wisconsin, particular flexibility in the generation of its arrays. These combine a maskless photolithography method with a proprietary chemical process for efficient and accurate *in situ* synthesis of high-density probe arrays. The company's latest generation chips contain more than 2 million probes. NimbleGen also favours the use of long, typically 60mer, probes. "We are the only company that combines long oligomers with high density," says vice-president of business development Emile Nuwaysir. NimbleGen's rapid production process also allows it to continually update its probe sequences to align with the latest genome-annotation data. Affymetrix is currently taking advantage of this process for the production of NimbleGen-manufactured NimbleExpress custom GeneChips.

A relatively recent entrant into the gene expression array field, Applied Biosystems of Foster City, California, has used years of experience in genomic work — and access to the proprietary genome databases of Celera Genomics, based in Rockville, Maryland — to good advantage in the design of its oligonucleotide arrays. "We've basically front-loaded all of the bioinformatics work," says staff scientist Chris Streck. "We do all the curation and annotation of these particular genes, and we make sure we have the most comprehensive and complete view of the genome to begin with." Applied Biosystems also benefits from a chemiluminescence-based approach to detection, with considerably reduced background noise relative to standard fluorescent systems.

The number crunch

However, high-quality samples and high-tech instrumentation alone won't save the microarray experiment. Some of the most fundamental challenges lie in gleaning biological significance from mounds of data and designing experiments with a statistically sound foundation.

David Allison, a biostatistician at the University of Alabama at Birmingham, remembers the early days of microarray work with horror. "The sample sizes were way too small, unjustified statements were made, and the analyses were primitive," he says. Fortunately, he adds, "the field recognized this, and a lot of people started weighing in with their own methods".

According to Irizarry, an important first step for good analysis is the effective pre-processing of raw data, using algorithms that accurately convert spot fluorescence to gene-expression estimates. "Changing those algorithms can make a difference," he says, "and you can turn an experiment that looks so-so into something that looks powerful and precise." Irizarry has also called attention to the importance of data normalization, and designed an online tool, Affycomp II, which allows users to benchmark their normalization methods using 'known'

SHARE AND SHARE ALIKE

Many working with microarrays now recognize that one way uncertainty about experimental findings can be dispelled is by being more transparent about methodology and data. This realization has transformed the field. For instance, after some initial resistance, almost every major commercial vendor has made the sequences and annotations of their probes publicly available — to the considerable benefit of the community as a whole.

This awareness has also manifest itself in the drive to develop shared resources for pooling experimental data and systems for clearly defining how these data were obtained. A leading force in this regard is the Microarray Gene Expression Data (MGED) Society, which put forward a proposal in 2001 for experimental annotation standards known as minimum information about a microarray experiment (MIAME), designed to record key details about factors such as sample preparation and experimental design. These standards were embraced by many, and several leading journals, including *Cell*, *The Lancet* and *Nature*, demand MIAME compliance from all microarray research submissions. However, some aspects of MIAME have proved problematic.

"I think almost all academic biologists embrace the concept of openly sharing data," says Catherine Ball of Stanford University in California, the current president of the MGED Society. "But embracing the process and actually taking part are very different, and it can be difficult to fully annotate your data." According to Gavin Sherlock, also of Stanford and MGED, part of the problem was MAGE-ML (microarray and gene expression markup language), the XML-based language initially developed for MIAME data recording. "Nobody can look at it, nobody can read it, nobody can edit it," he says. "It's very difficult to use." This is reflected in the uploading of data to public databases, another process strongly advocated by MGED.

The ArrayExpress database of the European Bioinformatics Institute in Cambridge, UK, is strictly MIAME-compliant, and receives considerably fewer submissions than the non-MIAME-compliant Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information in Rockville, Maryland. MGED is now poised to release a considerably simpler format for data submission, and Ball is hopeful that this, along with other user-friendly software tools, will make a difference.

Catherine Ball believes simpler software tools could encourage better MIAME compliance.

But, fundamentally, compliance comes down to the effort scientists can and will put in. All of the MicroArray Quality Control project's data are being deposited into both GEO and ArrayExpress, and although this has proved an onerous task, Leming Shi of the US Food and Drug Administration sees clear rewards in the effort. "Depositing the data may be a painful process, but we have to do it for the sake of the community," he says. "The more information we have in the future, the better." M.E.