Meeting report

# MGED comes of age

Helen C Causton and Laurence Game

Address: CSC/IC Microarray Centre, Imperial College, Hammersmith Campus, DuCane Road, London W12 ONN, UK.

Correspondence: Helen C Causton. E-mail: helen.causton@csc.mrc.ac.uk

Published: 13 November 2003

*Genome* ***Biology*** 2003, **4:**351

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/12/351



A report on the Sixth International Meeting of the Microarray Gene Expression Data Society ('MGED6'), Aix-en-Provence, France, 6-8 September 2003.

The sixth Microarray Gene Expression Data Society meeting brought together computational and life scientists to discuss microarray data management and analysis, developments in microarray technology and functional genomics, and findings resulting from the use of microarrays. The meeting, and MGED itself, have come a long way since 1999 when Alvis Brazma and Alan Robinson (European Bioinformatics Institute (EBI), Cambridge, UK) canvassed community opinion to determine whether there was interest in establishing standards for microarray databases. Four years on, standards have been put in place, data are being generated on a large scale at a number of centers worldwide, and a wealth of tools and databases have been developed.

## Data management

Many of the topics covered at MGED6 are not specific to microarrays, but instead are pertinent to most high-throughput functional genomics technologies. They have been addressed by MGED because microarray technology was one of the first to reach maturity and the benefits of data sharing were appreciated at an early stage. The principal challenges in dealing with large amounts of data relate to data analysis and the establishment of commonly accepted standards for data storage and description. How should the data be described and what should be included in the description, so that others in the field can either repeat the experiment, or determine whether the conclusions are supported by the underlying data? These challenges surfaced several years ago as microarray data began to accumulate. It became clear that not only was there great value in mining across very large datasets, but that the data could be considered a phenotype, a way of describing biological systems that transcends species- and subject-specific boundaries.

In response to these challenges, MGED established four working groups with the remit of establishing standards for data management (the MAGE group), the minimum information that must be recorded to describe a microarray experiment (MIAME), the terms used to describe an experiment (Ontology) and the methods used for data analysis (Normalization). To date, the working groups have rolled out first versions of the MAGE object model and markup language, the MIAME guidelines for data annotation and the MGED ontology; these are being actively used by the community of data generators. The majority of talks at MGED6 either related to initiatives by one of the four working groups or provided examples of how microarray data are being used, particularly in a clinical setting. What is striking is the extent to which the basic concepts underlying MAGE and MIAME, in particular, are being extended by other groups to cover domain-specific knowledge in a variety of fields. Talks related to MIAME included ways in which the concept is being extended or modified to cover subjects as diverse as chromatin immunoprecipitation 'on a chip', toxicogenomics, tissue arrays, environmental biology, single-nucleotide polymorphism (SNP) analysis and proteomics. Abstracts of talks, lists of speakers, and the MGED6 handbook are freely available online [http://tagc.univ-mrs.fr/mged6].

All four MGED initiatives have been designed to support the transfer of data into the public domain on publication (in line with the accepted standards for non-high-throughput data) and there are now three public repositories that accept data in this format. Two of these repositories were represented at the meeting; Yoshio Tateno (National Institute of Genetics, Japan) described CIBEX at the DNA Data Bank of Japan, and Helen Parkinson (EBI) described ArrayExpress, the microarray data repository at the EBI. ArrayExpress currently has information on 1,200 hybridizations (arrays) and

submissions from 50 groups. Two thirds of the submissions have come from 50 studies via data-transfer pipelines and one third from small-scale users who have largely been prompted to transfer their data into the public domain by journals as a condition of publication. ArrayExpress now has facilities for submitting data prior to publication, and restricted access can be arranged for reviewers. Representatives from *Nature* (Chris Gunter, Washington, USA) and *The Lancet* (Virginia Barbour, London, UK) gave the perspective of journals dealing with microarray data submissions and the way in which the journals are working closely with the EBI. Although there have been a few problems, the overall message was that the standards are working and researchers submitting to journals and public data repositories are making every effort to comply with them.

Tutorials by Catherine Ball (Stanford University, USA), Jason Gonçalves (Iobion, Toronto, Canada) and H.C.C. on behalf of the Normalization working group discussed - for the first time at an MGED meeting - the analysis problems that are particularly relevant to data acquired using Affymetrix chips, in recognition of the fact that approximately half of the attendees were generating, managing or analyzing data from the Affymetrix platform. The working group has not yet proposed standards for describing data transformations used in analysis, although the Bioconductor project [http://www.bioconductor.org], described by Rob Gentleman (Harvard University, Boston, USA) and Sandrine Dudoit (University of California, Berkeley, USA) may offer a way forward. Bioconductor is an open-source project that currently contains 30 packages for microarray data analysis and uses the language 'R'. Gentleman described the concept of 'compendia' - self-contained objects that contain code, text about what the code does, data on which the code operates and information on conventions. Compendia could be used to describe the formulae used for data transformation, the data on which the transformation is effected and a description. A group that wanted to transform its data in the same way as another group could simply use the same compendium with its own data, or run a similar analysis in which some of the parameters are changed.

One obstacle to using Bioconductor for analyzing large datasets has been the lack of resources for porting the data from MAGE-ML into the required format. A solution was presented by Joke Allemeersch (University of Leuven, Belgium), who has created a package that extracts cDNA microarray data (other array types are not included in the current version) from a MAGE document and maps it to Bioconductor objects for further analysis.

## Biological insights

Many diseases are associated with quantitative and qualitative changes in plasma proteins, and plasma represents the largest and most accessible subset of the human proteome: it contains thousands of distinct proteins, including glycoproteins, tissue proteins and immunoglobulins, with an extraordinary dynamic range of more than 10 orders of magnitude. Leigh Anderson (Plasma Proteome Institute, Washington, USA) described the potential of the plasma proteome for disease diagnosis and therapeutic monitoring. Apart from the need to identify co-varying set of proteins and the conditions with which they are associated, the major challenge resides in the development of proteomics technologies to identify and quantify potential markers. Classical proteomics tools, such as two-dimensional gel electrophoresis, mass spectrometry and antibody microarrays, typically detect proteins over a dynamic range approximately $10^7$ times smaller than the variation in individual protein abundance found in plasma. New technologies are emerging, however, which should bridge the gap between discovery of polypeptide abundance patterns and their use as robust diagnostic biomarkers. Anderson emphasized the need to put sufficient disease biomarkers into the public domain, to avoid a bottleneck in the development of diagnostic arrays.

Tony Lee (Whitehead Institute, Cambridge, USA) and Dave Vetrie (Sanger Centre, Cambridge, UK) both presented genome-wide protein-location analysis data (the method for obtaining these data is also known as chromatin immunoprecipitation on a chip - or ChIP on a chip). This technique, which permits protein-DNA interactions to be characterized, was used by Richard Young's group at the Whitehead Institute to identify the target genes of 290 known protein transcriptional regulators in the baker's yeast, *Saccharomyces cerevisiae*. Over 5,000 interactions were identified and most gene promoters were found to be bound by several protein regulators. This information was used to build elegant networks of interactions between genes and regulators. 'Network motifs' - distinct ways in which genes and proteins regulate each other, such as feedback, autoregulatory, feedforward and multicomponent loops - were then assembled into larger networks by combining genome-wide location and expression data. The challenge in using this technique in higher eukaryotes is not only the size of the genome, but also the quality and accuracy of the annotation, the availability of promoter or intergenic region microarrays, and the vast number and heterogeneity of tissues and cell types. Vetrie has successfully tackled many of these technical challenges and talked about the high-resolution genomic arrays his group has developed: these either cover specific regions of interest and can be used for both expression and location analysis or are designed for comparative genome hybridization. These microarrays can be used to detect single exon deletions (at a resolution of 100-500 base-pairs) with 99.9% accuracy. A number of diseases are associated with single exon deletions and Vetrie's group intends to develop arrays for their diagnosis.

In many ways this was a landmark meeting. A group that began with the modest aim of sharing data and developing

standards for databases has grown into a body that has captured the imagination of the functional genomics community and represents a continuing 'work in progress'. Most research groups are still not able to share high-throughput functional genomics data in any meaningful way, but the good news is that we are well on the way to making this a reality and can begin to carry out more sophisticated analysis. Progress has been fast: as one attendee remarked, "A year ago we were talking about MAGE-ML tools; now we have them." Standards are in place and widely accepted, and there is a profusion of data-analysis software. By next year we should be able to see the impact of these tools in action.