IN THE NEWS

Have you come far?

In 5 years time we could possess the most detailed genetic map so far of the history of human migrations. This is the ambitious plan of a privately financed US$4OM project, recently launched by the National Geographic Society and computer giant IBM. The project aims to collect blood samples from 100,000 indigenous people throughout the world, analyse them and try to determine their geographical origins (*The Indian Express*, 18 April 2005). To get the kind of sampling they need they have invited people from across the world to participate — as well as doing field research among hundreds of indigenous groups, the project is selling $99 cheek-swabbing kits (*USA Today*, 17 April 2005) for which the donors are given information on the migratory histories of their ancestors (kits are available at http://www3.nationalgeographic.com/genographic). Data are anonymous and will not be used for medical or political ends, assures the project director, Spencer Wells.

Not everyone is queuing up for the kit, however, as mistrust is brewing in various corners. Ethnic minorities are already boycotting the project — and even IBM computers — as they fear the project will be used to diminish their rights. Indigenous populations, burnt by previous encounters with scientists, including the Human Genome Diversity Project, are wary (*ABC Science Online*, Australia, 25 April 2005). And scientists themselves have been wondering whether they will have access to the samples, and under what terms.

But one happy customer is already selling the idea: a Navajo from Arizona was thrilled to bits when he learned from Wells that his genetic origins could be traced to Mongolia: "It's always been something that was in me, and finally I was able to say 'yeah'", he said. (*iAfrica.com*, 28 April 2005)

Tanita Casci

GENE EXPRESSION

Do microarrays match up?

Disparities between microarray data from different groups working on similar samples has made many question the validity of this widely adopted technology. Although the 'minimal information about a microarray experiment' (MIAME) guidelines set standards for the publication of microarray data, they do not address experimental reproducibility. As gene-expression data rapidly accumulate in the public domain, three papers in *Nature Methods* provide a timely investigation into the reproducibility of microarray data and suggest that, with appropriate caution, such data can be used with confidence.

One of the main issues when comparing microarray data is consideration of the metrics generated by different technology platforms. There is a tremendous choice of platforms available and much diversity in protocols for sample preparation, imaging and analysis. Furthermore, whereas some groups report the absolute level of expression of a particular gene, others compare the relative transcription of genes. This makes meaningful comparisons of gene-expression data from different sources challenging.

The three papers investigate different aspects of microarray reproducibility. Larkin *et al.* directly compared the performance of two microarray platforms — an in-house-developed two-colour cDNA array and a commercial oligonucleotide array — in a study of the effects of chronic and acute exposure of angiotensin II on cardiac gene expression in mice. Irizarry *et al.* studied the impact of inter-laboratory variation by providing a consortium of ten laboratories with an identical RNA sample processed according to individual laboratory protocols, and then comparing the results obtained from three widely used microarray platforms. Finally, the Toxicogenomics Research Consortium (TRC) used in-house and commercial microarrays with identical RNA samples to assess the variability caused by sample handling, imaging and data analysis.

The studies show that results from different platforms are remarkably consistent. Larkin *et al.* report that most genes had similar expression patterns, but that the relative amplitude of expression was greater according to the commercial array. Some genes had divergent expression patterns between platforms, but principal-components analysis clustered these genes by experimental treatment rather than platform. Mapping probes from both arrays

BIOINFORMATICS

Gene predictions — filling in the worm holes

The ultimate goal of genome-annotation programs is to correctly predict the sequence of every gene in a given organism. *Caenorhabditis elegans* has led the way, and Wei *et al.* now report an adaptation of the TWINSCAN gene-prediction program, with which they have discovered 1,119 new *C. elegans* genes.

Although the *C. elegans* genome sequence has been available since 1998, there are still thousands of genes without cDNA or EST evidence. Therefore, several gene-prediction programs were developed and optimized specifically for worms. Wei *et al.* used these resourses and compared the available data with their results using the TWINSCAN algorithm, which was originally developed to annotate the human genome. The advantage of their method lies in the fact that it combines the probabilistic Hidden Markov Model approach with information derived from the alignment of the target genome (*C. elegans*) to a second genome, known as the informant (*Caenorhabditis briggsae*).

Using information from the entire *C. elegans* genome, they predicted 2,891 open reading frames (ORFs) that do not overlap with existing WormBase annotations. The authors then tested 265 of these predicted ORFs through amplification and cloning procedures, and finally confirmed 146 novel gene predictions — 55% of those targeted. The genes were poorly conserved between *C. elegans* and *C. briggsae*; this is a reflection of the strength of this strategy for gene identification because poorly conserved genes are difficult to predict.

Why is this approach so successful? The authors claim that the models the program uses for GC–AG splice sites and intron-length distribution, together with the *C. briggsae* alignment, are the major advances contributing to the accuracy of TWINSCAN's *C. elegans* predictions.

to the genome revealed that the two platforms interrogated different sequences for these divergent genes; Larkin *et al.* suggest that the presence of poorly or non-annotated splice variants might explain this inconsistency.

Considerable variation between laboratories using identical RNA samples was identified by both Irizarry *et al.* and the TRC study, although the TRC study showed that reproducibility improved markedly after standardizing protocols for RNA labelling, hybridization, array processing, data acquisition and normalization.

All three papers agree that using a standard procedure to normalize data relative to controls provides a more meaningful value and eliminates technical variability caused by probe and target molecules. Moreover, the TRC study showed that the use of gene-ontology nodes to analyse groups of genes *in lieu* of direct gene-by-gene comparison identified significant biological themes even with low levels of correlation between data from different platforms and laboratories.

Despite some disagreement, the authors reach a common consensus that standardization of experimental and analytical procedures is warranted. These studies should boost confidence that robust and reproducible results can be obtained using microarrays.

Joanna Owens, Associate Editor,
Nature Reviews Drug Discovery

References and links

ORIGINAL RESEARCH PAPERS Larkin, J. E. *et al.* Independence and reproducibility across microarray platforms. *Nature Methods* **2**, 337–343 (2005) | Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2**, 345–349 (2005) | Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods* **2**, 351–356 (2005)

The total number of real genes in *C. elegans* is going to change as a result of this study — although its sequence is among the best annotated. This method is applicable to other model organisms, such as *Arabidopsis thaliana*, which is likely to contain more than 1,000 unannotated genes and thousands more that are misannotated. Because this computational approach is the first one to achieve 60% sensitivity in the exact prediction of proteins in a multicellular organism, the future for the correct annotation of other genomes is bright.

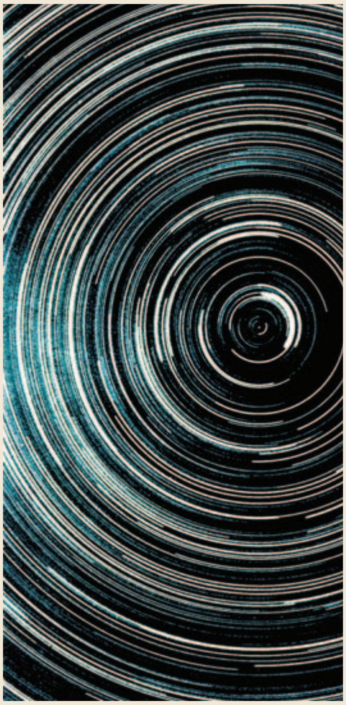
Ekat Kritikou

References and links

ORIGINAL RESEARCH PAPER Wei, C. *et al.* Cloning in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* **15**, 577–582 (2005)

WEB SITES

David Brent's web page: http://www.cs.wustl.edu/~brent



IN BRIEF

AGEING

Analysis of long-lived *C. elegans daf-2* mutants using serial analysis of gene expression.

Halaschek-Wiener, J. *et al. Genome Res.* 18 April 2005 (doi:10.1101/gr.3274805)

This is the first study to use serial analysis of gene expression (SAGE) to understand gene-expression patterns involved in the ageing process. By comparing control and long-lived (*daf-2* mutant) worms the authors identified whole gene families that were differentially regulated between the two groups. As long-lived worms showed a 'hypo-metabolic' state in early life, the authors speculate that the apparent metabolic repression contributes substantially to the observed longevity.

HUMAN DISEASE

A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk.

Sproat Emison, E. *et al. Nature* **434**, 857–863 (2005)

Hirschsprung disease (HSCR) is a complex, non-Mendelian disorder that has been linked to mutations in the coding sequence of the *RET* receptor tyrosine kinase. The authors used family-based association studies combined with comparative genomics analysis of *RET* sequences from several organisms to further the molecular understanding of this multifactorial disorder. Using this new approach, they show that the most common HSCR-associated mutation in *RET* is non-coding, has low penetrance and has sex-dependent effects.

GENE EXPRESSION

Special feature: Gene regulatory networks

Proc. Natl Acad. Sci. USA **102**, 5 April 2005

How do gene-regulatory networks control animal development? And what are the current approaches used to dissect those networks? A recent issue of *PNAS* addressed these questions in a special feature that contains commentaries and research articles. Understanding why, when and where genes are specifically expressed are the key issues that scientists are trying to tackle using different models — from nematodes and flies to sea urchins, frogs and mammals. Advanced technologies are also discussed, including a combination of DNA microarrays and bioinformatics that promises to accelerate regulatory-network studies.

HUMAN EVOLUTION

A scan for positively selected genes in the genomes of humans and chimpanzees.

Nielson, R. *et al. PLoS Biol.* **3**, e170 (2005)

Humans and chimpanzees have undergone pronounced changes in anatomy and cognitive ability in the 5 million years since their divergence. Nielson *et al.* compared the sequences of 13,731 annotated human genes to their orthologues in chimpanzees. They found that those genes with the strongest signatures of positive selection encode proteins that are involved in immunity, sensory perception, spermatogenesis and, surprisingly, tumour suppression and apoptosis. Unexpectedly, they found no evidence of positive selection on those genes that are maximally expressed in the brain.