central to all forms of Hirschsprung disease, although mutations in its coding regions have only been identified in 40% of linked families. Although the genes at the other two loci were not identified in this study, the authors suggest that they are modifiers of *RET* expression. The success of this and several other studies demonstrate that it is possible to identify susceptibility loci in complex genetic diseases using whole-genome-scanning approaches, especially in diseases where accurate phenotypic discrimination among subtypes is possible. (Gabriel, S. B. *et al.* [2002] *Nat. Genet.* DOI: 10.1038/ng868) *AP*

Localizing the proteome

The first proteome-scale analysis of protein localization within a eukaryote has been published. By epitope-tagging more than half of the *Saccharomyces cerevisiae* proteome, Kumar *et al.* localized 2744 yeast proteins to regions within the cell. They extrapolated their data using a Bayesian algorithm, and defined the location of all 6100 yeast proteins – what they term the yeast 'localizome'. Of the proteins in the localizome, about 1000 are transmembrane proteins, and the rest are soluble. Of the ~27% of proteins that are nuclear or nucleolar, more than a third are associated with chromosomal DNA. This study also provides empirically determined locations for 955 proteins that have no known function. The raw data can be accessed at http://ygac.med.yale.edu/. (Kumar, A. *et al.* [2002] *Genes Dev.* 16, 707–719) *NJ*

Therapeutic cloning to treat genetic disease

A report by Rideout *et al* provides a 'proof of principle' that therapeutic cloning combined with gene therapy is a feasible approach to treat genetic disease. Tail tip cells from Rag2 recombinase negative mice, which lack mature B and T cells, were used in nuclear transfer (NT) experiments to generate NT ES cells. These were repaired by homologous recombination, differentiated *in vitro* into hematopoietic stem cells (using novel methods described in an accompanying paper) and grafted into Rag2 mice.

Disappointingly, the treated mice did not regain immune function, and only limited immune function was obtained in mice mutated to remove NK cells (a source of cell-mediated rejection). This important work supports the possibility of using NT and gene therapy to treat genetic disorders, but shows that even genetically matched cells can 'face barriers' to effective transplantation following therapeutic cloning. (Rideout, W.M. III *et al.* [2002] *Cell* 109, 17–27) *NL*

Asthma-associated genes revealed by microarrays

A monkey model for asthma has been used to assess changes in lung gene expression patterns following exposure to an antigen. After inhaling the antigens, produced by *Ascaris suum*, allergic cynomolgus monkeys exhibit symptoms similar to asthma. Using microarrays, Zou and colleagues profiled the gene expression patterns in the lungs of these monkeys over a time course following antigen exposure. They also examined profiles of monkeys given interleukin-4 (IL-4) treatment. Of ~40 000 cDNAs on the microarray, 169 cDNAs encoded by 149 genes changed their expression by more than a factor of 2.5. A third of these genes were previously unknown. Downregulation predominated in the antigen-challenged monkeys, especially soon after the challenge. By contrast, many genes were upregulated in the IL-4-treated monkeys. Based on cluster analysis of expression patterns, there are at least five groups of genes. Several chemokine mediators are in one such cluster. (Zou, J. *et al.* [2002] *Genome Biol.* 3, research0020.1–0020.13) *NJ*

BLAST goes BLAT

James Kent at the University of California, Santa Cruz, has developed a new tool called BLAT (for 'BLAST-like alignment tool'). It is optimized for comparing mRNA or DNA sequences, as well as for protein alignments in vertebrate genomes. BLAT is 500 times faster for mRNA/DNA alignments and about 50 times faster for protein alignments compared with other programs that are used currently. A major difference between BLAST and BLAT is that BLAST breaks down the query sequence to search against the database, whereas BLAT keeps an index of the entire genomic or protein sequence in the memory and compares it with the query sequence. BLAT was designed to find DNA sequences with 95% identity over 40 nucleotides or proteins with 80% identity over 20 residues. Therefore, it is not recommended for sequences that are distantly related or for short sequence alignments. BLAT can be used through a web interface to search the human genome (http://genome.ucsc.edu) and is also available as a stand-alone application. High-speed alignment tools such as BLAT will be crucial for large-scale alignments, such as those required to keep genomic sequences updated in a timely fashion. (Kent, W.J. [2002] *Genome Res.* 12, 656–664) *AP*

Norman A. Johnson
njohnson@ent.umass.edu
Natasha Lane
n.lane@ic.ac.uk
Petros Ligoxygakis
P.Ligoxygakis@ibmc.u-strasbg.fr
Richard Morgan
rmorgan@sghms.ac.uk
Akhilesh Pandey
pandey@cebi.sdu.dk

Letter

Cell cycle analysis and microarrays

A recent profusion of microarray analysis of gene expression during the cell division cycle raises questions as to whether statistical problems with the data analysis have been considered. Two non-scientific examples – ESP and the Bible Code – illustrate the problem.

Individuals with extrasensory perception (ESP) were identified by their ability to guess pictograms in a deck of 25 cards. There were five different pictures so that random guessing would lead to an expectation of 20% correct. When thousands of individuals were tested, some guessed the cards correctly with a frequency greater than 20%. These individuals were proposed to have 'ESP-ability'. Upon re-testing, some of these individuals repeated above-average guessing, whereas others did not. With

0168-9525/02/$ – see front matter PII: S0168-9525(02)02694-X

further testing, the ESP phenomenon disappeared. The explanation was that extended testing destroyed the ESP phenomenon.

The fallacy in this experiment is clear. Among a large number of subjects there will be a spread in the scores – the classic bell-shaped curve. Some will do better – even much better – than 20%, and others will do worse. If the high scorers are re-tested, some of these high scorers will repeat, and some will not. Eventually all of those that had initial high scores will not repeat. They will be deemed to have lost the ESP-ability that they once had. Without a precise and pre-set definition of a good score defining ESP (e.g. why not 100% correct?), it is merely the result of statistical variation that some subjects appear to be endowed with paranormal ability.

The Bible Code [1] identified numerous prophecies embedded in the Bible by searching with a computer for words formed from letters separated by one letter, or two letters, or three letters, and so on. After numerous searches it was possible to find predictions of such events as both Kennedy assassinations, the Oklahoma City bombing, the election of Bill Clinton, and everything from World War II to Watergate, from the Holocaust to Hiroshima, from the Moon landing to the collision of a comet with Jupiter.

This is the ESP fallacy again. With so many letters in the Bible, merely by looking over different skip distances between letters one can find anything one wants. David Thomas, writing in the Skeptical Inquirer, performed the same search for occurrences of ‘Clinton’ in War and Peace (212 000 characters, 7.6 billion possible seven-letter equidistant sequences) and found 21 occurrences; the predicted number based on letter frequencies was 21. A search for Apollo yielded 129, with a predicted number of 128.1.

How do these phenomena relate to microarrays? Recently, Kerby Shedden (Statistics Dept, University of Michigan, USA) and I have reanalyzed the proposal that there are numerous genes exhibiting cell-cycle-specific expression patterns in animal cells [2]. Our analysis of the available data on the Internet showed that the results are consistent with statistical variation (i.e. random noise) [3]. Specifically, randomized data exhibit periodic patterns of similar or greater strength than the experimental data. This suggests that the apparent cycling in the expression measurements might arise from chance fluctuations. In addition, the degree of cyclicity and the timing of peak cyclicity in a given gene are not reproduced in two replicate experiments.

Comparison of these results with the ESP and Bible Code examples illustrates a major problem with this type of microarray data. When there are so many data points – thousands of genes over a number of different time points – it is possible, by statistical variation alone, to find results that fit various ‘successful’ patterns. When searching for sinusoidal variation over two successive cell cycles, patterns can occur merely by the accidental arrangement of experimental variation.

The absence of pre-set criteria for successful identification of cell-cycle-specific expression patterns is particularly evident in a widely cited analysis of yeast gene expression during the cell cycle [4]. Cycle-specific gene expression in yeast was identified by fitting the microarray data to a sine wave. Those patterns with a coefficient of correlation of 0.7 or better were proposed to be expressed in a cell-cycle-specific pattern. Of 104 ‘known’ cell-cycle-regulated genes, 94 were included in the group, a success rate of 92%. But this success was achieved by lowering the correlation threshold until a significant number of the known genes were included. This approach led to the inclusion of many genes that were not previously identified as cell-cycle specific or that, from biochemical considerations alone, would not be expected to be cell-cycle specific. The threshold could also be raised to eliminate patterns not expected to be cell-cycle specific. Then one could see which genes among the known cyclical genes are not included in the list.

A microarray analysis of gene expression in *Caulobacter* has produced the astounding result that 533 genes are expressed in a cell-cycle-specific manner [5,6]. The question that should be raised for this result is: are the patterns reproducible? Replication of microarray experiments is made difficult by the expense of the procedure. Because it appears that only one experiment was performed to obtain the *Caulobacter* data, it is difficult to determine whether the results are reproducible. The question raised here about the reproducibility of the *Caulobacter* results is made even more pertinent by the findings that two replicate experiments on gene expression during the division cycle of human cells [2] do not give reproducible results, either in degree of cyclicity or location of peak expression [3].

Precise criteria must be used before accepting claims for cyclical variation of genes during the division cycle when microarray data is being analyzed. Just as with ESP and the Bible Code, objectivity requires that the criteria for success should be set before the experiment is run. Experiments must also be reproducible. Until these criteria for a successful cell-cycle-expression experiment are in place, the microarray results on cell-cycle-specific gene expression should be considered with caution.

References

1 Drosnin, M. (1997) *The Bible Code*, Simon and Schuster

2 Cho, R.J. *et al*. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.* 27, 48–54

3 Shedden, K. and Cooper, S. (2002) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4379–4384

4 Spellman, P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297

5 D’Ari, R. (2001) Cycle-regulated genes and cell cycle regulation. *BioEssays* 23, 563–565

6 Laub, M.T. *et al*. (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science* 290, 2144–2148

Stephen Cooper
Dept of Microbiology and Immunology,
University of Michigan Medical School,
Ann Arbor MI 48109-0620, USA.
e-mail: cooper@umich.edu

0168-9525/02/$ – see front matter