LEADING ARTICLE

Expression profiling: fishing and chips

I P Hall

Expression profiling approaches are potentially valuable in helping to define novel gene products which may be important in disease pathogenesis or treatment response. However, there are many pitfalls which need to be considered in the design of experiments of this kind and in considering interpretation of data from such studies. Some of these are discussed.

This issue of *Gut* contains an article by Diosdado and colleagues¹ on the use of expression profiling approaches to study the pathogenesis of coeliac disease [see page 944]. Such approaches are potentially valuable in helping to define novel gene products which may be important in disease pathogenesis or treatment response.² However, there are many pitfalls which need to be considered in the design of experiments of this kind and in considering the interpretation of data from such studies. Some of these are discussed below.

Expression profiling can be performed at the genomic level, the protein level (proteomics), or at the signalling pathway (sometimes called signalomics or transducomics) or metabolite level (metabolomics). In each case the principle of the approach is the same. In essence, the investigator takes a cell type or tissue of interest under two conditions (for example, disease/non-diseased; following treatment/placebo) and makes a comparison of the expression pattern of the gene products in order to identify targets which are up or downregulated. The methodologies are most advanced for genomic expression profiling. Both microarrays on glass slides or filters and custom built chips are widely available commercially, each containing probes to identify the pattern of expression of numbers of genes varying from a few hundred to the vast majority of those in the human genome. Increasingly, academic departments are setting up inhouse facilities to take advantage of the availability of the clone sets in the public domain through sources such as the Medical Research Council. Attempts to produce protein chips, for example using panels of monoclonal antibodies, have so far been less successful and proteomic profiling has in general been performed using two dimensional gel electrophoresis. This requires identification techniques dependent on mass spectrometry. This approach is less amenable to high throughput studies but has the obvious advantage that one is looking at protein levels rather than RNA levels.

In considering a series of experiments using expression profiling, it is absolutely critical that

Gut 2004;53:917–918. doi: 10.1136/gut.2003.026971

investigators ensure they use a rigorous study design. The simplest experiments consist of taking homogenous cell populations following a specific treatment (for example, stimulation with a single cytokine) and then, following RNA extraction, making a comparison of gene expression in the two populations of cells. Even this simple paradigm needs to be thought through carefully. For example, at what time point should one look for changes? Gene expression may change at an early point following stimulation (for example, within the first hour) or at a later time point. Therefore, one would probably wish to have samples at a minimum of two time points as well as a baseline. The next question is how many replicates one should do in these experiments for each condition. Usually, a minimum of three independently handled replicates would be considered essential to enable at least some assessment of reproducibility to be made. When one factors in more time points and more concentrations of the stimulus, the number of samples to be handled rapidly increases. Typically, most experiments will therefore generate 20–40 samples at a minimum.

“Expression profiling can be performed at the genomic level, the protein level, or at the signalling pathway or metabolite level”

Even though the unit cost of microarrays and gene chips has come down markedly over recent years, such experiments will prove relatively expensive. In theory, one can pool samples to minimise the number of chips used but this ignores the potential problem of variable hybridisation. Therefore, ideally, each sample should be hybridised with a single array or chip: this will then produce a data set where change in gene expression for each of the replicates under a different condition can be assessed independently and allow an estimate of confidence intervals for the degree of change in expression to be made.

The situation is more complex when tissue rather than homogenous cell populations are being considered. Obviously, if one compares tissue from a disease versus a non-disease state (for example, inflamed bowel versus non-inflamed bowel in patients with ulcerative colitis) the constituent cell types in the biopsies may well be very different. This will by itself produce marked differences in the pattern of gene expression when the samples are compared. This sort of approach has recently been used extensively in the cancer field to compare expression profiles of different tumours in order to provide prognostic information^{3–5} or to give an

Correspondence to: Professor I P Hall, Division of Therapeutics and Molecular Medicine, D Floor, South Block, Queen's Medical Centre, Nottingham NG7 2UH, UK; ian.hall@nottingham.ac.uk

Accepted for publication 15 September 2003

indication of relapse rates or sensitivity to different chemotherapeutic agents.⁶ However, to date, there have been relatively few attempts to compare gene expression profiling approaches with traditional approaches to assess the histological grade of a tumour, and hence the real clinical value of this approach remains to be fully established. One interesting corollary of these approaches is the observation that tumours from different sites may actually be quite similar in terms of their expression patterns. Currently, our approach to treatment of malignancy is usually based on tumour site of origin coupled with histology: this may not prove logical in the future if it is clear that gene expression patterns provide a better signature for treatment response (irrespective of site of origin) than traditional approaches.

"Perhaps the biggest problem of expression profiling approaches is dealing with the vast volume of data generated"

Perhaps the biggest problem of expression profiling approaches is dealing with the vast volume of data generated. A typical experiment using, for example, a genechip probing for 10 000 different gene products will generate (if performed in triplicate) 30 000 data points per sampling point. If five time points are compared with and without treatment with a stimulus at a single concentration, this results in 300 000 data points. Handling these volumes of data is a major logistical exercise: in general, bioinformatic support in the academic sector at least is rudimentary. To deal with the bioinformatic issues, a number of software programs have been developed to try and prioritise which gene products are worth pursuing. The most frequently used approach involves hierarchical clustering.⁷ The fundamental paradox is of course that expression profiling approaches will tell you about genes which you may already have considered to be important in the disease of interest and a whole series of genes which you will either have not considered to be relevant or (more likely) for which very little functional information is available. There is little point in using expression profiling approaches to tell an investigator that their favourite cytokine has changed its expression profile when they could simply have gone and measured it in the first place. The most interesting targets are therefore those for which, in general, least information is available as these may prove to be truly novel disease related genes. Prioritising which of these targets to pursue is a key issue which has not been fully resolved. Some approaches which are being attempted include combining expression data with linkage based approaches to identify genes which might be important in disease initiation, and using analyses based on gene networks (for example, components in a signalling cascade) to look for coordinated gene expression for pathways which may be important in disease pathogenesis.⁸

"There are many potential sources of error in expression profiling approaches"

Finally, there are many potential sources of error in expression profiling approaches. Some of these have been dealt with above but it is also worth remembering that inevitably, given the number of data points being generated in a typical experiment, at least a proportion of these will represent "genuine" false positives. Before chasing after one's favourite target, some validation is required: how far one goes to verify a target depends to some extent on the reliability of the initial expression profiling data. Typically, quantitative reverse transcription-polymerase chain reaction

approaches or (if an antibody is available) measurement of protein directly are the methods used. The other obvious source of error is in the reliability of the arrays and chips themselves. Obviously, using an array or chip to define expression profiles is dependent on the relevant probes for the genes of interest being specific and sensitive. Significant error rates have been reported in the past for some of the available tools for expression profiling (for example, due to cross contamination of the clone sets used to generate arrays): quality control is therefore a critical issue. This may prove a particular challenge in the academic sector where many universities are attempting to establish inhouse arrays.

"Expression profiling approaches are here to stay"

Expression profiling approaches are here to stay. Increasingly, journals and meetings will be receiving studies containing these kind of data and there is no doubt that potentially these approaches will be of enormous value in identifying novel genes and pathways important in disease processes. A search on PubMed (www.ncbi.nlm.nih.gov/entrez) performed while writing this article using the search term "expression profiling" gave 6850 hits. However, in assessing studies of this nature it is important to ensure that the points considered above are adequately addressed. Finally, there is a major challenge in presenting these data to the general scientific community in a user friendly way. Many journals have developed online supplements to allow such data to be deposited. There are also moves to develop centralised repositories of expression profiling data sets which might in themselves be useful resources within which investigators can go fishing. These resources are also now being linked to the literature base on gene function.⁹ In an attempt to help standardise data presentation for studies of this nature, a set of guidelines has been laid down by the International Microarray Gene Expression Data society called the Minimum Information About a Microarray Experiment (MIAME). Some journals (for example, *Nature*) have adopted these guidelines and require authors to provide a complete checklist against MIAME guidelines.¹⁰ Increasing standardisation of the way such data are presented can only help research in the long run.

ACKNOWLEDGEMENTS

Work in the author's laboratory is funded by the National Asthma Campaign, Medical Research Council, the Wellcome Trust, and Biotechnology and Biological Sciences Research Council.

REFERENCES

- 1 **Diosdado B**, Wapenaar MC, Franke L, *et al.* A microarray screen for novel candidate genes in coeliac disease pathogenesis. *Gut* 2004;**53**:944–51.
- 2 **Lindpaintner K**. The impact of pharmacogenetics and pharmacogenomics on drug discovery. *Nat Rev* 2002;**1**:463–75.
- 3 **Alizadeh AA**, Eisen MB, Davis RE, *et al.* Distinct patterns of B-cell lymphoma identified by gene expression profiling. *Nature* 2000;**403**:503–11.
- 4 **Yanagisawa K**, Shyr Y, Xu BJ, *et al.* Proteomic pattern of tumour subsets in non-small-cell lung cancer. *Lancet* 2003;**362**:433–9.
- 5 **Bittner M**, Meltzer P, Chen Y, *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000;**406**:536–40.
- 6 **Ishibashi Y**, Hanyu N, Nakada K, *et al.* Profiling gene expression ratios of paired cancerous and normal tissue predicts relapse of esophageal squamous cell carcinoma. *Cancer Res* 2003;**63**:5159–64.
- 7 **Eisen MB**, Spellman PT, Brown PU, *et al.* Cluster analysis and display of genome wide expression patterns. *Proc Natl Acad Sci USA* 1998;**75**:14863–8.
- 8 **Hughes TR**, Marton MJ, Jones AR, *et al.* Functional discovery via a compendium of expression profiles. *Cell* 2000;**102**:109–26.
- 9 **Jensen TK**, Laegreid A, Komorowski J, *et al.* A literature network of human genes for high throughput analysis of gene expression. *Nat Genet* 2001;**28**:21–8.
- 10 **Tilstone C**. DNA microarrays: vital statistics. *Nature* 2003;**424**:610–12.