**Journal of Pathology**
*J Pathol* 2005; **205**: 236–247
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/path.1702

**Review Article**

# Hunting the primary: novel strategies for defining the origin of tumours

Jayne L Dennis[1] and Karin A Oien[2]*

[1]*Department of Cancer Medicine, Imperial College of Science, Technology and Medicine at Hammersmith Hospital, London, UK*
[2]*Cancer Research UK Centre for Oncology and Applied Pharmacology and University Department of Pathology, University of Glasgow, Glasgow, UK*

**Correspondence to:*
*Dr Karin A Oien, Cancer Research UK Centre for Oncology and Applied Pharmacology, Cancer Research UK Beatson Laboratories, University of Glasgow, Garscube Estate, Switchback Road, Glasgow G61 1BD, UK.*
*E-mail: k.oien@beatson.gla.ac.uk*

## Abstract

**In 1995, two methods of genome-wide expression profiling were first described: expression microarrays and serial analysis of gene expression (SAGE). In the subsequent 10 years, many hundreds of papers have been published describing the application of these technologies to a wide spectrum of biological and clinical questions. Common to all of this research is a basic process of data gathering and analysis. The techniques and statistical and bioinformatic tools involved in this process are reviewed. The processes of class discovery (using clustering and self-organizing maps), class prediction (weighted voting, *k* nearest neighbour, support vector machines, and artificial neural networks), target identification (fold change, discriminant analysis, and principal component analysis), and target validation (RT-PCR and tissue microarrays) are described. Finally, the diagnostic problem of adenocarcinomas that present as metastases of unknown origin is reviewed, and it is demonstrated how integration of expression profiling techniques promises to throw new light on this important clinical challenge.**


Received: 7 October 2004
Accepted: 17 October 2004

**Keywords: microarray; serial analysis of gene expression (SAGE); clustering; artificial neural network; principal component analysis; tissue microarray (TMA); immunohistochemistry; tumour marker; adenocarcinoma; unknown primary**

## Introduction

Tumours have traditionally been assessed by histopathology in order to predict their clinical behaviour [1]. Gross and microscopic examination is used to confirm the diagnosis and then to predict prognosis and guide therapy. The diagnosis of cancer is made according to structural, thus morphological, abnormalities at the levels of the overall tissue architecture and the individual component cells and their nuclei. The tumour is then classified according to the normal tissue from which the tumour originates. The commonest human cancers arise from epithelial cells and are named carcinomas, which fall into three broad groups: squamous carcinomas; adenocarcinomas, which arise from simple glandular epithelium; and carcinomas of solid organs, such as the liver, kidney, and endocrine organs. Both pathological and clinical criteria, including radiological examination, are used to determine the extent of tumour spread, that is, the cancer's stage. This classical anatomical and histological approach to the prediction of the behaviour of tumours has been developed and used successfully over the past 150 years or so. However, it has limitations: for any single patient, the prognosis given can often only be fairly broad, since some tumours may appear similar, yet behave very differently in individual patients.

It has long been recognized that the phenotypic abnormalities detected by pathologists are underpinned by changes in gene expression at the level of mRNA and ultimately protein [2]. The hope is that new assays based on the molecular changes underlying cancer may yield clinically useful information beyond that which can be provided by traditional histopathology: for example, enabling better prediction of the outcome and response to therapy in individual patients. For the past 30 years, protein expression has been exploited in tissue samples through immunohistochemistry, which is the standard method for confirmation of broad tumour types and for categorizing lymphomas [3,4]. Until recently, the identification of useful molecules was mainly on a candidate gene basis, because methodological constraints prevented the study of more than a few genes at any one time.

The 1990s saw the dawn of a new era: genome-wide expression profiling. Rapid advances in technology enabled the transition from investigating expressed genes individually to simultaneous assessment of almost all genes in a tissue [5]. This shift in scale has led to greater knowledge and increased understanding of cells and their biology and pathology [6]. For example, using gene expression profiling, it is possible to identify genes expressed in different cell types [7]; observe changes in gene expression in different

disease states [8]; correlate gene expression profiles with disease outcome and response to therapy [9]; and identify previously unrecognized and important subtypes within histologically homogeneous but clinically heterogeneous disease states [10].

These questions can be addressed by using similar processes of data gathering and analysis, outlined in Figure 1 and described herein. First, we describe methods of obtaining gene expression data. Generally, gene expression datasets are very large, the scale of which are beyond traditional statistical techniques familiar to medical and biological sciences. Therefore, new statistical techniques have been developed. These are discussed in terms of class discovery (defining groups and sub-groups of patients or samples), class prediction (predicting which group or sub-group a given patient or sample belongs in), and target identification (identifying genes of interest from thousands studied). Another consequence of genome-wide analysis has been the requirement for high-throughput validation methods [2], in which context tissue microarrays (TMAs) are described [11]. Finally, these technologies and techniques are integrated using, as an example, the clinical problem of adenocarcinomas that present as metastases of unknown primary site [12]. Novel strategies within the field of gene expression profiling may help to define the origin of these tumours.

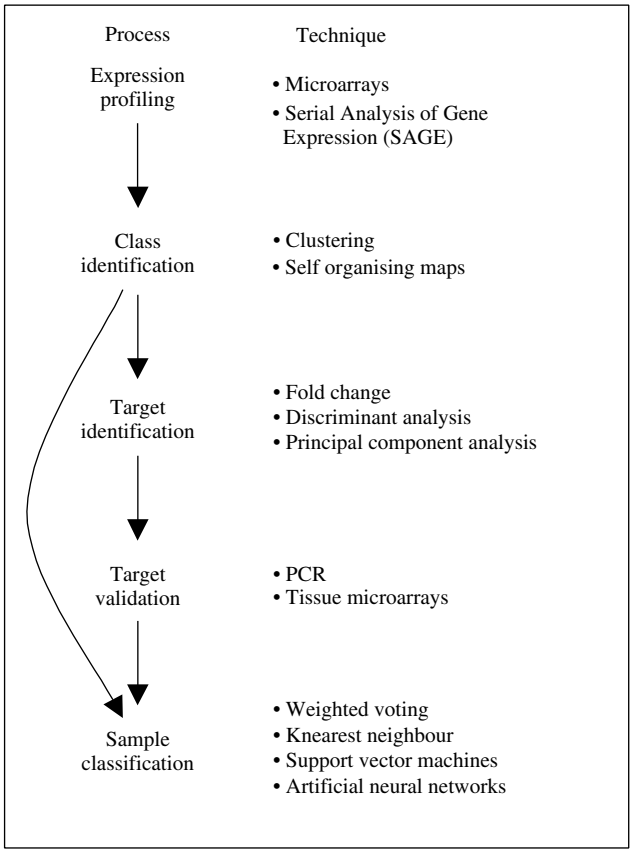


**Figure 1.** The data analysis process is similar in all gene expression profiling experiments. The tools and techniques used in each step of this process are described in the text

This review focuses on the analysis of the transcriptome (expressed genes at the mRNA level). Similar analyses may be undertaken using protein expression profiles, the proteome, which encompasses encoded proteins and their post-translational modifications [13,14]. For example, Petricoin *et al* described the identification of patients with ovarian carcinoma based on serum protein profiles [15]. Equally, technology is now emerging for large-scale analysis of DNA and its regulatory elements (the genome and epigenome) [16]. These non-transcriptional technologies are, however, beyond the scope of this discussion.

## Gene expression profiling technologies

Until recently, the capacity of gene expression profiling was limited by the methods available: techniques such as northern blotting only allowed the assessment of individual genes in relatively small sample numbers [17]. New technologies, however, have allowed genome-wide analysis of gene expression [16].

### Microarrays

Oligonucleotide and cDNA microarrays were first described by Schena *et al* [18] and have been extensively reviewed in 'The Chipping Forecast II' [19]. Microarrays are solid supports upon which several thousand gene-specific nucleic acids have been placed at defined locations, by either spotting or direct synthesis. cDNA microarrays are made by printing cDNAs onto the solid support, generally using a robotic arrayer; oligonucleotide microarrays, such as those available from Affymetrix, are made by direct synthesis of oligonucleotides onto the array surface by photolithographic masking techniques. The latter technique is easier to control, resulting in less variation between individual arrays. Early (micro)arrays were constructed on nylon membranes, although glass slides are commonly used now.

cDNA arrays enable users to compare the relative abundance of genes expressed in two tissues. RNA is isolated from each tissue and, during a reverse transcription process, is labelled with different fluorescent dyes, typically cyanine 3-dNTP (Cy3, green) and cyanine 5-dNTP (Cy5, red). The labelled samples are then mixed and hybridized onto the array. The relative abundance of each gene transcript is then determined from the ratio of red to green fluorescence on each spot: a red or green spot indicates a difference in the expression level between samples, whereas a yellow spot indicates no change [20]. To compare the gene expression patterns of a range of samples, a number of pair-wise comparisons are made, typically against a standard, or reference, made from a mix of several normal or control samples.

In contrast, oligonucleotide arrays do not require co-hybridization of samples. Instead, RNA extracted from one sample is labelled with biotin, hybridized

to the array, and stained using fluorescence-labelled streptavidin, which binds to biotin. Fluorescence is then detected using a laser scanner [20]. This process is repeated for all samples of interest. This technique generates an absolute measure of gene expression, unlike the cDNA array, which gives a relative expression level. Consequently, data and results from oligonucleotide arrays are easier to compare between experiments and research groups than are data from spotted arrays.

Increasingly, data from microarray experiments are deposited in public databases. This was made possible by the development of the guidelines on Minimum Information About a Microarray Experiment (MIAME), developed by the Microarray Gene Expression Data Society [21]. Currently, researchers are encouraged to submit data to one of three repositories: ArrayExpress at the European Bio-informatics Institute [22], Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information [23], or the Center for Information Biology Gene Expression Database (CIBEX) at the National Institute of Genetics, Japan [24,25]. Data can be analysed either with on-line tools at each database or by downloading data files.

## Serial analysis of gene expression (SAGE)

The SAGE method described by Velculescu *et al* generates quantitative profiles of gene expression [26,27]. Briefly, this is achieved through the generation of cDNA from mRNA transcripts and then isolation of short representative tags from cDNA. Sequence information from these tags allows identification of the source mRNA transcripts, while tag counts describe expression levels. Tag sequences and frequencies are collated into libraries representing samples of interest; each library typically contains upwards of 10 000 tags [28]. As SAGE tag numbers directly reflect the abundance of mRNA transcripts, SAGE libraries are highly accurate, quantitative, and comprehensive representations of the samples from which they are derived. The SAGE method is shown in further detail in Figure 2.

Since 1998, SAGE data have been deposited in the SAGEmap database within the Cancer Genome Anatomy Project (CGAP) at the NCBI (www.ncbi.nlm.nih.gov/SAGE) [29]. To date, in excess of nine million tags are held in this database, representing both normal and malignant human tissues, along with plant and animal specimens. Data in this valuable resource can be analysed either with on-line comparison tools or by downloading sequence data.

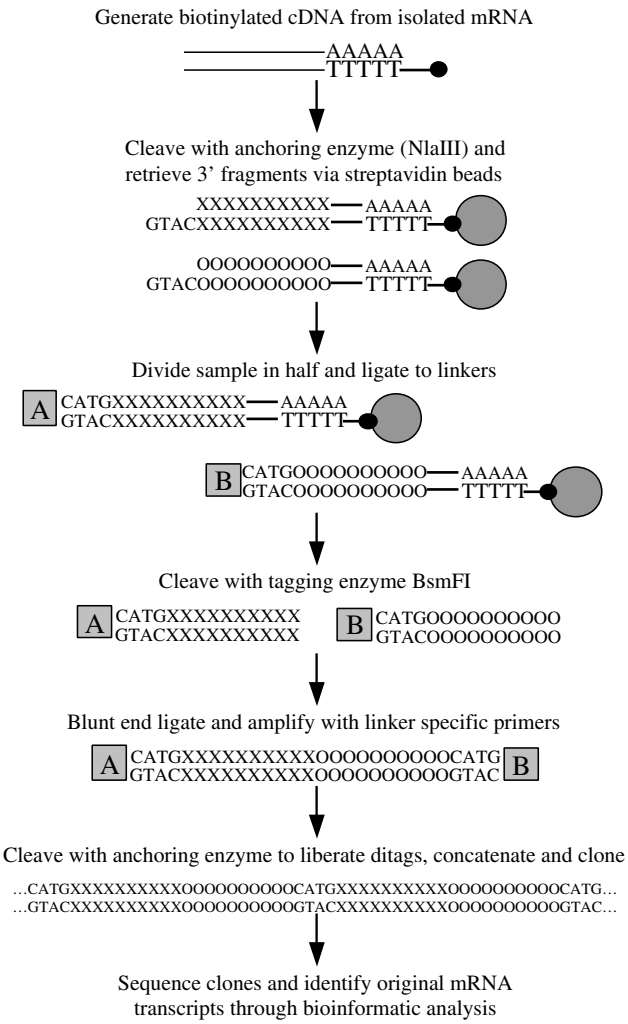


**Figure 2.** The SAGE method. First, mRNA isolated from the sample of interest is used to generate cDNA using biotinylated oligo(dT) primers. cDNA is then cleaved using a frequent cutting 'anchoring enzyme', typically NlaIII. The 3′ fragment is retrieved using streptavidin beads. The resulting fragment pool is split in half: fragments in each pool are ligated to one of two linkers via the overhang created by the anchoring enzyme. The linkers contain restriction sites for the tagging enzyme, BsmFI. The fragments are then cleaved with the tagging enzyme BsmFI, which, being a type IIS restriction enzyme, cleaves at a defined distance from its recognition site, generating 10 bp SAGE tags. The two pools of tags are then ligated, resulting in ditags of linker A-tag A-tag B-linker B. Ditags are PCR-amplified using primers for the linker sequences. Ditags are liberated from linkers using the initial anchoring enzyme NlaIII. Liberated ditags are then concatenated (joined up), cloned, and sequenced. The SAGE tags are extracted from the raw sequence data, counted and compiled into libraries, and then matched to the originating mRNA transcripts using sequence databases [26,28]

## Expressed sequence tags (ESTs)

EST libraries contain uncharacterized sequences of cDNA: the technology pre-dates microarrays and SAGE [30,31]. ESTs are generated through isolation and reverse transcription of RNA, typically by the use of oligo(dT) primers to select for poly-adenylated transcripts. The resulting cDNA libraries are ligated into plasmid vectors and transformed into competent bacteria. Isolation and sequencing of plasmid DNA from individual clones then generates sequence tags for each transcript. Construction of EST libraries in this way allows for the assessment of total mRNA populations in samples. The main strength of the technique is that libraries generated in different laboratories can

be compared, as abundance of individual transcripts is described as the percentage of transcripts in the whole population. The technology, however, is labour-intensive and despite the fall in sequencing costs, remains expensive compared with some techniques. The NCBI, however, has established a publicly available database of these partial, single-pass sequences, dbEST, which can be interrogated using an on-line bio-informatics tool, Digital Differential Display [32].

### Published literature

Although often over-looked, the published literature represents a valuable resource in modern gene expression profiling studies. It is increasingly common for software packages designed to assist expression profiling studies to include tools for mining literature databases such as MEDLINE at the National Library of Medicine. An example of such a tool is PubGene [33]. By entering the name of a gene, this tool allows users to investigate relationships between genes based on the incidence of co-citation in the literature. Alternatively, literature databases may be searched directly for gene expression studies in the tissue or disease of interest.

## Class discovery

Most gene expression profiling experiments fall into one of two categories: class discovery or class prediction. The former may involve, for example, the identification of new disease subtypes or correlation of gene expression profiles with recognized pathological classes. Conversely, class prediction involves predicting which known class a given sample belongs to. Class prediction is described later; here, two methods of class discovery are described.

### Clustering

Clustering techniques organize multivariate data (ie many variables; in this instance, genes or samples) into classes with similar patterns. Objects within one class are more similar to each other than to objects outside the class. The most commonly used clustering technique in the analysis of gene expression data is hierarchical clustering. Hierarchical clustering starts with all objects having their own, individual clusters and then the two objects (or clusters) most closely related are merged. This process is repeated until a single cluster remains. Relatedness between objects is determined by one of many algorithms which generate a distance metric describing the distance, or similarity, between objects. The product of hierarchical clustering is a tree structure, or dendrogram, as shown in Figure 3. Hierarchical clustering in gene expression profiling has proved useful in confirming the presence of known classes, for example the type of cell line [7]; identifying new classes, for example in lung adenocarcinoma [34]; and ascribing likely gene function to novel genes, for example identified in the response of fibroblasts to serum [35].

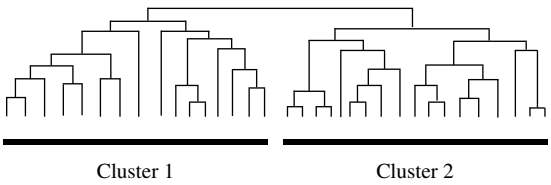


**Figure 3.** Clustering techniques organize data into classes with similar patterns, shown in this dendrogram. In this example, there are two major classes, or clusters. Objects within a cluster are more similar to each other than to objects outside it. The vertical length of branches in the tree represents the similarity between objects: the shorter the branch, the fewer changes or differences observed between objects

### Self-organizing maps (SOMs)

SOMs were first described by Tamayo and co-workers [36]. SOMs are constructed by selecting a number of nodes which are mapped into $k$-dimensional space. Nodes are initially mapped at random, then iteratively adjusted over 20 000–50 000 iterations. Each iteration involves randomly selecting a data point, then moving nodes in the direction of that data point. The distance over which nodes are moved is inversely related to their position with respect to the chosen data point: nodes closest to the data point are moved the most, while those furthest away are moved the least. This technique was applied to data gathered using oligonucleotide arrays from a myeloid leukaemia cell line over 24 h following treatment to induce differentiation [36]. Data representing 5223 genes and 1193 expressed sequence tags (ESTs) were filtered and organized onto a map of 12 clusters. One cluster contained 32 genes which were gradually induced over the time course, during which the cells acquired features of differentiation. Four of these genes were duplicates and two were ESTs for which no coding sequence was available. Of the remaining 26 genes in the cluster, 18 were known to be associated with haemopoietic differentiation, confirming the functional relevance of the clusters identified.

## Class prediction

Class prediction involves predicting which known class a given sample belongs to. Development of a classification algorithm involves analysis of data from samples of known class: the training set. The algorithm is then tested by applying it to a series of independent samples: the test set. A trade-off exists in the development of a classifier: it must classify as many of the training samples as possible, but this must not be at the expense of being over-fitted (or too rigid) to the training data. Here, four techniques used to develop classifiers are described: weighted voting, $k$-nearest neighbour, support vector machines, and artificial neural networks.

### Weighted voting

Weighted voting, like self-organizing maps, was first described by Golub *et al* for class prediction between acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) [36]. The average (mean) expression level of 6817 genes was determined by oligonucleotide arrays in a series of 11 AML and 27 ALL samples. Weighted voting techniques may be confused by the inclusion of too many variables; gene pre-selection is therefore required. A subset of 50 informative genes was selected whose expression correlated with the distinction between the two classes. Genes that better correlated with class distinction received a higher priority, or weight, than genes less well correlated with class. The expression levels of the informative genes were then determined in a series of test samples. Each gene voted samples into a class depending on the product of two factors: weight, as described above, and the expression of the gene in the test sample compared with its known mean expression in AML and ALL. The sample was classified if the sum of votes was greater than a predetermined threshold. Thirty-four independent samples were thus classified. In total, 29 of 34 samples were sub-divided by class with 100% accuracy.

### *k*-Nearest neighbour

*k*-Nearest neighbour classifiers are based on distances between pairs of samples, as determined by one of a number of algorithms (for example, Euclidean distance). To classify a test sample, its distance to all samples within the training set is determined. The $k$ training set samples closest to the test sample are thus identified. The test sample is then classified by majority vote: that is, the class that is most common among these $k$ neighbours. The number of neighbours $k$ is chosen by leave-one-out cross-validation performed on the training set. Each sample in the training set is removed from the dataset in turn and then its distance to all other training samples is calculated. The sample is classified according to its $k$ nearest neighbours for varying values of $k$, for example where $k$ is 1, 3 or 9. The classification for each training sample is then compared with the known class to produce the cross-validation error rate; the $k$ for which the cross-validation error rate is smallest is used on test samples [37]. Pre-selection of genes used to calculate distances may be required as this classifier may be confused by too many variables.

*k*-Nearest neighbour was used by Yeoh *et al* to assign ALL samples to classes of diagnostic importance [38]. Gene expression data were generated from 327 samples using oligonucleotide microarrays representing 12 600 genes. Hierarchical clustering of these data revealed previously recognized sub-groups of ALL, including cases with BCR-ABL chromosomal rearrangements, T-cell lineage leukaemia, and hyperdiploid karyotype. Supervised learning techniques, including *k*-nearest neighbour, were used to assign 112 samples to these sub-groups with 98% accuracy.

### Support vector machines (SVMs)

SVMs are a family of algorithms which plot samples in $n$-dimensional gene expression space. A hyperplane is then drawn to separate the samples into two classes. If no separating hyperplane exists, the samples are mapped into a higher-dimensional space where such a separator exists. This use of higher-dimensional space allows greater assessment of the data and generates a robust classifier. SVMs, however, are inherently binary, capable only of defining two classes [39]. This technique was introduced by Brown *et al* to predict gene function for 15 open reading frames of unknown function in the budding yeast *Saccharomyces cerevisiae* [40]. Data representing 2467 genes in 79 array hybridizations were used to recognize six previously described functional classes. SVMs were then used to predict the function of 15 unannotated genes. One such gene was predicted to be involved in respiration and was subsequently described as a subunit of the ATP synthase complex, confirming the prediction and supporting the use of SVMs in predicting gene function.

### Artificial neural networks

Artificial neural networks are computer-based algorithms modelled on the structure and behaviour of neurones in the human brain, as shown in Figure 4. Artificial neural networks can be trained to recognize and categorize complex patterns. An initial training dataset is used to set such parameters as the numbers of neurones used, the relative weight of each neurone, the threshold required for a neurone to fire, and the number of neural layers in the model. Once the algorithm is established, it is then validated on a second, test, dataset. The main strength of the technique is its ability to recognize complex patterns, and gene expression patterns are undeniably complex. There are, however, two main disadvantages of the technique. First, artificial neural networks may be over-fitted to the training data, ie generate rules that are too specific to the training data and not sufficiently general to allow classification of an independent dataset. Second, the network parameters are hidden, which prevents an understanding of how the samples have been classified.

Khan *et al* used artificial neural networks in the analysis of gene expression data from four types of 'small round blue cell tumour' [41]: Ewing's sarcoma, rhabdomyosarcoma, neuroblastoma, and Burkitt's lymphoma. Small round blue cell tumours share a rather similar histological appearance, but comprise different cancer types with different outcomes and therapies. Artificial neural networks were developed on a training set of 63 samples using expression data collected from cDNA microarrays. Models developed with these

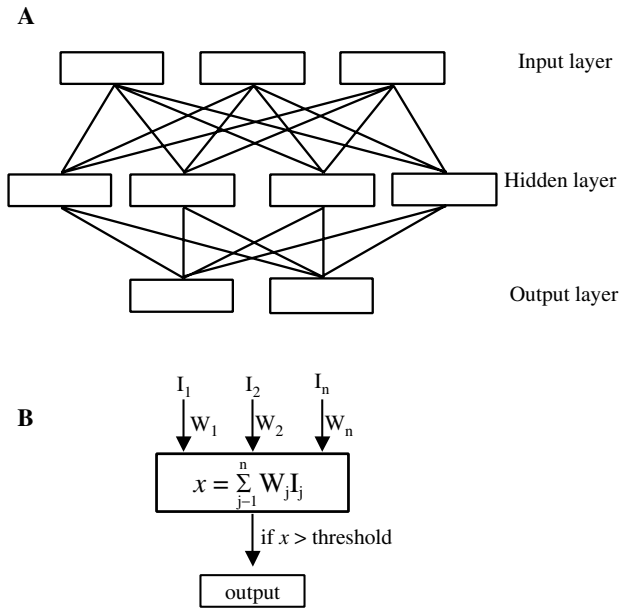


**Figure 4.** (A) The overall structure of an artificial neural network. The algorithm contains three layers, one of which is hidden. In this example, the input layer takes information from three variables. Data are then integrated in 'neurones' in the hidden layer, resulting in output to two output modules. These modules may take different forms, such as tumour classes or prediction of patient outcome. (B) Data integration in the hidden layer. Data are received from $n$ inputs, $I_{1-n}$. These inputs, however, are not given equal weight. Relative weights are defined by $W_{1-n}$. An output signal is generated if the sum of the input • weight products is greater than a pre-defined threshold. Input weights and threshold values are calculated by training the network to recognize classes in an initial set of samples. Weights and values are then refined by iterative processes such as leave-one-out cross-validation

data correctly classified 100% of tumours. The models were then applied to a blinded test set of 20 small round blue cell tumours: all samples were correctly classified.

## Target identification

The main advantage of genome-wide expression profiling technology is that it enables the simultaneous investigation of thousands of genes. However, it is often desirable to identify a small number of key genes (targets) whose expression correlates significantly with the phenomenon under investigation, or which are potentially causing the phenomenon. This subset of genes can then be used in a variety of ways, for example as therapeutic targets, markers of disease response or progression, or otherwise for the translation of profiling results into a clinical setting. Three main techniques used to identify targets are described below.

### Fold change analysis

The most basic question in the analysis of gene expression data is determining for which genes have the expression levels changed significantly. This is usually addressed by $k$-fold change: determining which genes' expression changed by a pre-determined cut-off, $k$. Microarray experiments commonly use a cut-off at two-fold change (for example, DeRisi *et al* [42]), although SAGE studies use a higher threshold, up to ten (for example, Nacht *et al* [43]). The significance of the observed fold change is then commonly determined by an appropriate statistical test. This approach, however, fails to identify small but reproducible changes in expression.

The significance of $k$-fold expression changes identified through SAGE is typically determined by Monte Carlo simulations. In the SAGE software, 100 000 simulations of the data are created [26]. The observed change in expression is then compared with these simulations and the relative probability of its occurrence calculated. Under the null hypothesis, the observed change should occur frequently in the simulations, creating a $p$-chance close to 1. A significant result, however, is a rare event with a small relative probability of occurring, and is therefore represented by a low $p$-chance value.

More recently, variants of common statistical tests to determine change in gene expression have been used. This is a two-stage process: calculating a test statistic and then determining the significance of this statistic. When comparing changes in expression between two groups, the $t$-test, or a variant thereof (often called $t$-like tests), is used. Determination of changes in expression between multiple groups commonly uses the ANOVA $F$ statistic. These approaches were used by Hedenfalk *et al* [44]. The main problem with these approaches, however, is that these statistical tests assume that the data are normally distributed with equal variance. This assumption may not, however, be valid in all experiments. In this situation, log transformation of data can help to normalize their distribution and equalize variance. This then allows meaningful biological comparisons to be made [45].

### Discriminant analysis

Discriminant analysis is a supervised learning technique that identifies relationships between qualitative variables or classes (for example, disease status) and quantitative predictor values (for example, gene expression) [46]. Thus, discriminant analysis allows the identification of genes that discriminate between disease states with statistical significance. Discriminant analysis was used by Spanakis and Brouty-Boye to differentiate between fibroblast subtypes [47]. Expression levels of 34 genes in 23 samples were determined by probing slot blots (RNA is spotted onto a membrane and then probed with gene-specific probes). Models were constructed using multivariate analysis and discriminant analysis to differentiate between eight fibroblast subtypes: 100% of samples were correctly classified with this model.

### Principal component analysis

Principal component analysis is a multi-dimensional scaling method that reduces data complexity. This is achieved through identification of key (principal) components in the data. The first principal component accounts for as much variability in the data as possible. Each succeeding component accounts for as much of the remaining variability as possible. For example, if a sample is profiled using a microarray containing spots for 10 000 genes, 10 000 data points will be produced. Imagine these data in multi-dimensional space, where the axis in each dimension shows the expression level of one gene: a cloud of 10 000 points will result. This cloud will not be hyperspherical. Instead, it will be irregular or extended in one direction. This is the axis of the first principal component. The axis of the second principal component is orthogonal to the first [20]. Schwartz *et al* used principal component analysis to visualize the differences in gene expression profiles between histological subtypes of ovarian carcinoma. They demonstrated that although mucinous and clear cell tumours could be distinguished from serous tumours, endometrioid tumours showed significant overlap in their gene expression profiles with other ovarian tumours [48].

## Target validation

Gene expression profiling of samples through such technologies as SAGE and microarrays has described many gene expression patterns and generated many hypotheses. These patterns and hypotheses require independent validation. Validation may be undertaken using either *in vitro* techniques on sample extracts or *in situ* methods on intact samples. Techniques such as northern and western blotting, performed on RNA and protein extracts respectively, can, however, analyse only a limited number of samples. A higher-throughput alternative is to perform reverse transcription-polymerase chain reaction (RT-PCR) on extracted RNA. *In situ* techniques include *in situ* hybridization and immunohistochemistry to assess RNA and protein, respectively [3,49]. The throughput of these techniques was radically improved with the introduction of tissue microarray technology [11].

### Reverse transcription-polymerase chain reaction (RT-PCR)

RT-PCR can be used to detect the presence of any transcribed gene, regardless of the amount of starting material or abundance of the transcript of interest [16]. A retroviral reverse transcriptase and primers, either oligo(dT) or random hexamers, are used to generate cDNA from RNA templates. The gene of interest is then PCR-amplified using sequence-specific primers. The product, whose size can be predicted from sequence information, is typically detected by gel electrophoresis. The main advantage of this technique is its sensitivity, although this also requires that samples must be free of genomic DNA or other contaminants. RT-PCR is relatively simple, fast, and cost-effective, especially when performed in a 96-well format. Results generated through RT-PCR are relative, although transcript quantitation is possible. This is achieved through the use of fluorescent probes to determine transcript abundance compared with standards of known concentrations. An internal control must be co-amplified to normalize against inter-sample variation.

### Tissue microarrays

Tissue microarray (TMA) technology enables the representation of hundreds of samples on a standard microscope slide [11]. Briefly, this is achieved by using hollow needles to array small cores (eg, 600 μm diameter) of paraffin-embedded tissue samples in a recipient wax block. Sections cut from the array can then be assessed by *in situ* hybridization or immunohistochemistry according to standard protocols [3,49]. Tissue arrays therefore enable high-throughput assessment of the presence and location of expressed genes, saving time, reagents, and clinical material.

Cancers are heterogeneous, and variability exists not only between tumours — hence the reason for studying many samples — but also within tumours. An important consideration when constructing tissue arrays, therefore, is capturing this intra-tumour variability. In their original report of TMA technology, Kononen *et al* constructed tissue arrays from a series of 645 primary breast cancers [11]. Arrays were used to investigate the amplification of three oncogenes (erbB2, c-myc, and cyclin D1) by fluorescence *in situ* hybridization (FISH). They reported amplification of these genes at rates comparable to previously published data, suggesting that data from tissue array technology are comparable to those from whole tissue sections. In another study, Camp *et al* compared staining in up to ten cores with results from whole tissue sections using antibodies against oestrogen receptor, progesterone receptor, and erbB2 [50]. Results from the three antibodies were similar and showed that staining in one core represented the whole section in an average of 92% cases. This figure increased to approximately 96%, 98%, 99%, 100%, and 100% for two, three, four, six, and ten cores, respectively. A trade-off exists, however, between using more cores to capture intra-tumour variability and receiving the large-scale benefits of tissue array technology. Clearly, the standard issues of immunohistochemistry quality assurance must be addressed with TMAs, as with whole sections [51].

## Defining the primary site of adenocarcinomas

Many questions remain in the biological and medical sciences. The advent of the expression profiling era

has renewed hope that some of these long unanswered questions can be addressed. One such problem is that of adenocarcinoma that presents as metastases of unknown origin.

## The clinical problem of adenocarcinoma of unknown origin

Most cancers present with the primary tumour, at its site of origin. Some 10–15% of cancers, however, present as metastases in solid organs, body cavities or lymph nodes [12,52]. Most of these secondary tumours are adenocarcinomas, for which the seven commonest primary sites are breast, ovary, prostate, stomach, pancreas, colon and lung [52]. The prognosis and therapy of patients are linked to the site of tumour origin, knowledge of which is becoming vital as more specific and effective chemotherapeutic regimens emerge [53]. Consequently, these sites, and others, are investigated by clinical examination, radiology, and serum tumour markers [12,54,55]. If no primary cancer is found, then the metastatic deposit is usually biopsied. Metastatic adenocarcinoma of unknown origin is a common clinical problem: it constitutes around 3% of all cancers and is thus one of the ten most common malignant diagnoses [12]. Its prognosis is poor, with a median survival of only 4 months [54].

## The pathological approach to adenocarcinoma of unknown origin

In patients with metastases of unknown origin, biopsy is performed in order to confirm the diagnosis of malignancy; to type the tumour and thus identify subsets of tumours (for example, lymphomas, germ cell tumours, and small cell and other neuroendocrine carcinomas) which are exquisitely chemosensitive; and increasingly, where the tumour is an adenocarcinoma, to predict the primary site in order to provide prognostic information, guide therapy, and inform the patient [53,56].

Metastatic adenocarcinomas can usually be diagnosed as such by their microscopic appearance. Since adenocarcinomas share a common derivation from glandular epithelium, morphology is of less help in the prediction of primary site; but with the addition of minimal clinical data, including gender and biopsy site, up to 55% of metastases can be correctly assigned to a site of origin [57]. Clearly, each type of glandular epithelium has a different biological function and therefore expresses specific genes associated with this differentiation; if this expression profile is maintained during carcinogenesis, and furthermore during metastasis, then these genes may be of diagnostic use in predicting the primary site.

The obvious method for their investigation is immunohistochemistry, which has already been used to address this diagnostic dilemma specifically [51,58–64]. Genes proposed as potential markers of the primary site have been identified from the literature and information on their immunohistochemical expression in the seven main adenocarcinomas retrieved [51,58,61–63,65–104]. The data are summarized in Figure 5, which lists CA125, CA19-9, carcinoembryonic antigen (CEA), CDX2, cytokeratins 7 and 20 (CK7 and CK20), oestrogen receptor (ER), gross cystic disease fluid protein 15 (GCDFP-15), mesothelin, pS2 (trefoil factor 1), prostate-specific antigen (PSA), prostatic acid phosphatase (PAP), surfactants A and B,

| | Percentage positivity of markers in adenocarcinomas from each site | | | | | | |
|---|---|---|---|---|---|---|---|
| **Marker** | **Breast** | **Ovary** | **Prostate** | **Stomach** | **Pancreas** | **Colon** | **Lung** |
| **CA125** | 13, 13, 13, 23, 24 | 61, 63, 74, 80, 91, 96 | 2, 2, 5 | 7, 11 | 48, 48 | 4, 4, 9, 10, 13, 18 | 20, 20, 20, 35 |
| **CA19-9** | 6, 11, 24, 45,62 | 41, 57, 77 | 3, 6 | 56, 71, 80, 89 | 57, 71, 85, 85, 86, 91, 94 | 26, 30, 41, 59, 71, 76, 79, 85 | 13, 30, 32, 48, 69 |
| **CEA** | 32, 37, 40, 57, 71 | 0, 0, 21, 37, 40 | 0, 14 | 67, 72, 75, 80 | 50, 81, 88, 92, 98 | 58, 86, 96, 98, 98, 99, 100, 100, 100 | 54, 63, 88, 90, 91 |
| **CDX2** | 0, 0 | 0, 0 | 1, 4 | 20, 70 | 0, 32, 54 | 90, 99 | 0, 0 |
| **CK 7** | 70, 89, 93, 93, 96 | 83, 89, 91, 100, 100 | 0, 5, 12, 29 | 38, 51, 60, 71 | 87, 92, 95 | 5, 6, 7, 9, 16, 23, 38 | 96, 100, 100 |
| **CK 20** | 0, 0, 0, 0, 4, 7, 19 | 0, 0, 4, 19 | 0, 0, 21 | 30, 41, 50, 51, 54, 68 | 0, 35, 39, 44, 62 | 65, 73, 84, 88, 92, 92, 93, 100 | 0, 8, 9, 10, 10 |
| **ER** | 32, 33, 58, 60, 63, 73 | 4, 12, 34, 50 | 10 | 0, 2 | 0, 0 | 0, 0, 0, 0, 13 | 0, 0, 3, 11 |
| **GCDFP15** | 33, 44, 47, 52, 62, 67, 72, 74, 77 | 0, 0, 0, 0, 4, 4 | 10, 15 | 0, 0, 3 | 0, 0, 0 | 0, 0, 0, 0, 0, 0, 0 | 0, 0, 0, 0, 4, 6 |
| **Mesothelin** | 0, 3, 6 | 95, 100 | 0, 0 | 10, 29 | 75, 86 | 4, 22 | 22, 24 |
| **pS2** | 27, 53, 73 | 36 | 0 | 60 | 77, 80 | 84 | 28 |
| **PSA** | 0 | 0, 0 | 86, 96 | 0 | 0 | 0 | 0, 9 |
| **PAP** | 0 | 2 | 97 | 0 | 0 | 0 | 0 |
| **Surf A** | 0, 0 | 0, 0 | 0 | 0 | 0 | 0, 0 | 48, 48 |
| **Surf B** | 0 | 0 | 0 | 0 | 0 | 0 | 52 |
| **TTF1** | 0, 0 | 0, 0 | 0, 0 | 0, 0 | 0, 0, 0 | 0, 0, 0 | 56, 66, 71, 74, 75, 75 |

**Figure 5.** Expression of potential markers of the primary site in the seven main adenocarcinomas. Assembled from references 51, 58, 61–63, and 65–104. The ovarian tumours are serous (or at least non-mucinous). The background shading is equivalent to the median of the percentage positivities obtained from the literature

and thyroid transcription factor 1 (TTF1). The utility of each marker relates to its specificity and sensitivity, and as the figure shows, results may vary between laboratories. Relatively few markers show truly site-specific expression: examples are PSA and TTF1. For this reason, pathologists are accustomed to using antibodies with lower specificities combined in a marker panel [58,61–63], and CK7 and CK20 are particularly useful in initial biopsy evaluation.

Only a few formal studies of the utility of such panels in predicting the primary site of adenocarcinomas have been performed. Brown *et al* studied metastatic adenocarcinomas from five sites (breast, ovary, upper gastro-intestinal tract, colon, and lung) and correctly predicted 66% using four markers (CEA, CA19-9, CA125, and BCA225) [63]. Lagendijk *et al* studied three sites (breast, ovary, and colon) and were successful in 80–90% with six markers (CK7, CK20, CA125, CEA, ER, and GCDFP-15) [58]. DeYoung and Wick studied over 2800 carcinomas from a range of sites and achieved 66% prediction with 14 markers including PSA, GCDFP-15, CEA, CA125, CA19-9, CK20, and ER [51]. Clearly, while these results are good, they are not perfect, and so the problem of adenocarcinoma classification is now being addressed by expression profilers.

## Expression profiling and molecular classification of adenocarcinomas

In our own study, we first used publicly available SAGE data and performed hierarchical clustering to demonstrate that the common adenocarcinomas cluster according to their site of origin [105]. This supports the hypothesis that there are similarities in the expression profiles between tumours of the same origin and differences between the primary sites. The clustered SAGE data also included two pairs of breast carcinoma samples derived from primary tumours and corresponding lymph node metastases. These four samples clustered together. This suggests that the metastases more closely resembled primary tumours of the same origin than primary adenocarcinomas from elsewhere, and others have reported a similar correlation [106,107]. This suggests that markers identified in primary tumours should be equally applicable to their metastatic counterparts.

Buckhaults *et al* extended this work with SAGE data and were able to construct a classifier to differentiate between adenocarcinomas of the breast, ovary, pancreas, and colon [106]. Hierarchical clustering was used to identify genes whose expression correlated with each site. Five genes [fatty acid binding protein 1, caeruloplasmin, MUC16 (encodes CA125), and the genes SLPI and PDEF] were assessed by quantitative RT-PCR in an independent set of tumours. Analysis of these data using self-organizing maps resulted in the construction of a classifier within which 81% of tumours were correctly classified.

Three studies have described the classification of adenocarcinomas using microarray data. Su *et al* used oligonucleotide arrays to profile the expression of 12 533 genes in 100 primary tumours from 11 classes [107]. These sites included the main adenocarcinoma primary sites as well as transitional, renal, hepatocellular, and lung squamous carcinomas. Hierarchical clustering of these data showed that some tumours grouped according to their anatomical sites, particularly those arising from the breast, prostate, pancreas, colon, liver, and kidney. A classification algorithm was then developed, similar to the $k$-nearest neighbour method described above and including a confidence score to estimate the strength of the prediction. Tumours classified above a confidence threshold were assigned a class. Empirically, prediction was based on a set of 110 genes, representing ten genes per tumour type. The 100 training samples were then classified by leave-one-out cross-validation: 94% were assigned to a class, of which 98% were correct. Of the six samples that did not pass the confidence threshold, five were correctly predicted, but with low confidence.

The classification scheme was then validated on an independent set of 75 tumours [107]. This tumour set included 63 primary and 12 metastatic tumours, although tumours of gastro-oesophageal, pancreatic, bladder, and kidney origin were either under-represented or not represented. Confident and accurate predictions were made for 87% of primary tumours and 75% of metastatic tumours. Of 11 unclassified cases, seven were assigned correctly, but with low confidence.

A second oligonucleotide array multi-class study was later published by Ramaswamy *et al* [108]. Like Su *et al*, Ramaswamy *et al* studied tumours arising from the breast, ovary, prostate, pancreas, colon, lung, bladder, and kidney. In addition, lymphoma, leukaemia, mesothelioma, central nervous system tumours, and endometrial adenocarcinoma were included. Data collected from 16 063 probe sets were analysed using hierarchical clustering and self-organizing maps. Results for both techniques were similar: leukaemias, lymphomas, mesotheliomas, and central nervous system tumours readily grouped together. Adenocarcinomas did not, however, cluster according to their primary site. A classification scheme, using all 16 063 genes, was constructed to assign samples to their type and site of origin. A support vector machine-based algorithm achieved correct classification in 78% of training set tumours. When applied to an independent test set of 54 tumours, 78% were again correctly classified.

Finally, Bloom *et al* used both oligonucleotide and spotted arrays to profile 78 adenocarcinomas from seven common sites [109]. They developed three artificial neural network classifiers: the first was derived from cDNA microarray data; the second was generated from oligonucleotide microarray-derived data; and data from both platforms were used for the third. These classifiers correctly classified 83%, 88%, and 85% of tumours, respectively, and required at least 400 genes for accuracy. The oligonucleotide and cDNA classifiers were further validated using a series of 50

metastatic tumours. These tumours arose in the kidney and the seven main adenocarcinoma sites and metastasized to solid organs such as the liver, lung, and CNS. Eighty-four per cent of these tumours were correctly classified.

### Clinical translation of research

The results from these large-scale profiling and classification studies are encouraging: they support the hypotheses that, at the level of gene expression, differences exist between the main sites of origin of adenocarcinomas, and that these differences can be used to assign likely primary sites to metastatic tumours. Despite the promise shown, we are unaware of clinical translation of the results in terms of microarrays themselves: such analyses currently remain purely a research tool, and their routine diagnostic use is not imminent, through cost, technical, and interpretative issues.

Gene expression profiling studies may, however, prove extremely useful in identifying new tumour markers. For example, we have identified 61 site-specific genes from SAGE data [105] and others have done the same [106]. Such markers may thereafter be validated using TMAs and immunohistochemistry, and antibody panels developed to be taken forward into diagnostic use on standard formalin-fixed, paraffin-embedded biopsy material. A proof-of-principle study was provided by Nishizuka *et al*, who used microarrays (and protein arrays) to identify candidates to distinguish ovarian from colonic adenocarcinomas [110]. The favoured genes were villin for colon and moesin for ovary; interestingly, the latter was present in the tumours' stromal component rather than the epithelium.

This example of metastatic adenocarcinoma of unknown origin is just one instance of how gene expression profiling technologies and bio-informatic and statistical analyses can be combined to powerful effect to address previously unanswered (and unanswerable) biological problems, and pathologists are ideally placed to perform the translational studies required to take the resulting diagnostic and therapeutic markers forward into routine clinical use.

### Acknowledgements

Thanks are due to Cancer Research UK for funding through a Clinician Scientist Fellowship (KO) and to the University of Glasgow for funding, in particular the Florence Houston Bequest to the Department of Medical Oncology for a PhD studentship (JD).

## References

1. Wyllie AD. Growth and neoplasia. In *Muir's Textbook of Pathology*, MacSween RNM, Whaley K (eds). Edward Arnold: London, 1992; 355–410.
2. Liotta L, Petricoin E. Molecular profiling of human cancer. *Nature Rev Genet* 2000; **1**: 48–56.
3. Dabbs DJ. *Diagnostic Immunohistochemistry*. Churchill Livingstone: Edinburgh, 2002.
4. Jaffe ES. Hematopathology: integration of morphologic features and biologic markers for diagnosis. *Mod Pathol* 1999; **12**: 109–115.
5. Emmert-Buck MR, Strausberg RL, Krizman DB, *et al*. Molecular profiling of clinical tissue specimens: feasibility and applications. *Am J Pathol* 2000; **156**: 1109–1115.
6. Alizadeh AA, Ross DT, Perou CM, van de Rijn M. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol* 2001; **195**: 41–52.
7. Ross DT, Scherf U, Eisen MB, *et al*. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet* 2000; **24**: 227–235.
8. Logsdon CD, Simeone DM, Binkley C, *et al*. Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. *Cancer Res* 2003; **63**: 2649–2657.
9. van't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**: 530–536.
10. Alizadeh AA, Eisen MB, Davis RE, *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**: 503–511.
11. Kononen J, Bubendorf L, Kallioniemi A, *et al*. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Med* 1998; **4**: 844–847.
12. Pavlidis N, Briasoulis E, Hainsworth J, Greco FA. Diagnostic and therapeutic management of cancer of an unknown primary. *Eur J Cancer* 2003; **39**: 1990–2005.
13. Patterson SD, Aebersold RH. Proteomics: the first decade and beyond. *Nature Genet* 2003; **33**: 311–323.
14. Rai AJ, Chan DW. Cancer proteomics: serum diagnostics for tumor marker discovery. *Ann N Y Acad Sci* 2004; **1022**: 286–294.
15. Petricoin EF, Ardekani AM, Hitt BA, *et al*. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002; **359**: 572–577.
16. Brown TA. *Genomes 2*. BIOS Scientific Publishers: Abingdon, 2002.
17. Sambrook J, Russell D. *Molecular Cloning: A Laboratory Manual* (3rd edn). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2000.
18. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**: 467–470.
19. Supplementary issue: Chipping forecast II. *Nature Genet* 2002; **32**: (Suppl): 461–552.
20. Knudsen S. *A Biologist's Guide to Analysis of DNA Microarray Data*. Wiley: New York, 2002.
21. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genet* 2001; **29**: 365–371.
22. Brazma A, Parkinson H, Sarkans U, *et al*. ArrayExpress — a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003; **31**: 68–71.
23. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**: 207–210.
24. Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y. CIBEX: center for information biology gene expression database. *C R Biol* 2003; **326**: 1079–1082.
25. Ball CA, Brazma A, Causton H, *et al*. Submission of microarray data to public repositories. *PLoS Biol* 2004; **2**: E317.
26. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995; **270**: 484–487.
27. Polyak K, Riggins GJ. Gene discovery using the serial analysis of gene expression technique: implications for cancer research. *J Clin Oncol* 2001; **19**: 2948–2958.
28. Oien KA. Serial analysis of gene expression. *Methods Mol Biol* 2003; **226**: 271–284.
29. Strausberg RL. The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. *J Pathol* 2001; **195**: 31–40.

30. Adams MD, Kelley JM, Gocayne JD, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991; **252**: 1651–1656.
31. Adams MD, Kerlavage AR, Fleischmann RD, *et al*. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995; **377**: (Suppl): 3–17.
32. Wheeler DL, Church DM, Edgar R, *et al*. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* 2004; **32**: D35–D40.
33. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet* 2001; **28**: 21–28.
34. Beer DG, Kardia SL, Huang CC, *et al*. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med* 2002; **8**: 816–824.
35. Iyer VR, Eisen MB, Ross DT, *et al*. The transcriptional program in the response of human fibroblasts to serum. *Science* 1999; **283**: 83–87.
36. Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–537.
37. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002; **3**: RESEARCH0036.
38. Yeoh EJ, Ross ME, Shurtleff SA, *et al*. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002; **1**: 133–143.
39. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nature Genet* 2002; **32**: (Suppl): 502–508.
40. Brown MP, Grundy WN, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 2000; **97**: 262–267.
41. Khan J, Wei JS, Ringner M, *et al*. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med* 2001; **7**: 673–679.
42. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997; **278**: 680–686.
43. Nacht M, Ferguson AT, Zhang W, *et al*. Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Res* 1999; **59**: 5464–5470.
44. Hedenfalk I, Duggan D, Chen Y, *et al*. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 2001; **344**: 539–548.
45. Quackenbush J. Microarray data normalization and transformation. *Nature Genet* 2002; **32**: (Suppl): 496–501.
46. Mendez MA, Hodar C, Vulpe C, Gonzalez M, Cambiazo V. Discriminant analysis to evaluate clustering of gene expression data. *FEBS Lett* 2002; **522**: 24–28.
47. Spanakis E, Brouty-Boye D. Discrimination of fibroblast subtypes by multivariate analysis of gene expression. *Int J Cancer* 1997; **71**: 402–409.
48. Schwartz DR, Kardia SL, Shedden KA, *et al*. Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. *Cancer Res* 2002; **62**: 4722–4729.
49. McNicol AM, Farquharson MA. *In situ* hybridization and its diagnostic applications in pathology. *J Pathol* 1997; **182**: 250–261.
50. Camp RL, Charette LA, Rimm DL. Validation of tissue microarray technology in breast carcinoma. *Lab Invest* 2000; **80**: 1943–1949.
51. DeYoung BR, Wick MR. Immunohistologic evaluation of metastatic carcinomas of unknown origin: an algorithmic approach. *Semin Diagn Pathol* 2000; **17**: 184–193.
52. Nystrom JS, Weiner JM, Heffelfinger-Juttner J, *et al*. Metastatic and histologic presentations in unknown primary cancer. *Semin Oncol* 1977; **4**: 53–58.
53. Mintzer DM, Warhol M, Martin AM, Greene G. Cancer of unknown primary: changing approaches. A multidisciplinary case presentation from the Joan Karnell Cancer Center of Pennsylvania Hospital. *Oncologist* 2004; **9**: 330–338.
54. Hillen HF. Unknown primary tumours. *Postgrad Med J* 2000; **76**: 690–693.
55. Varadhachary GR, Abbruzzese JL, Lenzi R. Diagnostic strategies for unknown primary cancer. *Cancer* 2004; **100**: 1776–1785.
56. Blaszyk H, Hartmann A, Bjornsson J. Cancer of unknown primary: clinicopathologic correlations. *APMIS* 2003; **111**: 1089–1094.
57. Sheahan K, O'Keane JC, Abramowitz A, *et al*. Metastatic adenocarcinoma of an unknown primary site. A comparison of the relative contributions of morphology, minimal essential clinical data and CEA immunostaining status. *Am J Clin Pathol* 1993; **99**: 729–735.
58. Lagendijk JH, Mullink H, van Diest PJ, Meijer GA, Meijer CJ. Immunohistochemical differentiation between primary adenocarcinomas of the ovary and ovarian metastases of colonic and breast origin. Comparison between a statistical and an intuitive approach. *J Clin Pathol* 1999; **52**: 283–290.
59. Pecciarini L, Giulia Cangi M, Doglioni C. Identifying the primary sites of metastatic carcinoma: the increasing role of immunohistochemistry. *Curr Diagn Pathol* 2001; **7**: 168–175.
60. Hammar SP. Metastatic adenocarcinoma of unknown primary origin. *Hum Pathol* 1998; **29**: 1393–1402.
61. Perry A, Parisi JE, Kurtin PJ. Metastatic adenocarcinoma to the brain: an immunohistochemical approach. *Hum Pathol* 1997; **28**: 938–943.
62. Lagendijk JH, Mullink H, Van Diest PJ, Meijer GA, Meijer CJ. Tracing the origin of adenocarcinomas with unknown primary using immunohistochemistry: differential diagnosis between colonic and ovarian carcinomas as primary sites. *Hum Pathol* 1998; **29**: 491–497.
63. Brown RW, Campagna LB, Dunn JK, Cagle PT. Immunohistochemical identification of tumor markers in metastatic adenocarcinoma. A diagnostic adjunct in the determination of primary site. *Am J Clin Pathol* 1997; **107**: 12–19.
64. Ellis IO, Hitchcock A. Tumour marker immunoreactivity in adenocarcinoma. *J Clin Pathol* 1988; **41**: 1064–1067.
65. Torenbeek R, Lagendijk JH, Van Diest PJ, *et al*. Value of a panel of antibodies to identify the primary origin of adenocarcinomas presenting as bladder carcinoma. *Histopathology* 1998; **32**: 20–27.
66. Tot T. Cytokeratins 20 and 7 as biomarkers: usefulness in discriminating primary from metastatic adenocarcinoma. *Eur J Cancer* 2002; **38**: 758–763.
67. Tot T. Adenocarcinomas metastatic to the liver: the value of cytokeratins 20 and 7 in the search for unknown primary tumors. *Cancer* 1999; **85**: 171–177.
68. Zamecnik J, Kodet R. Value of thyroid transcription factor-1 and surfactant apoprotein A in the differential diagnosis of pulmonary carcinomas: a study of 109 cases. *Virchows Arch* 2002; **440**: 353–361.
69. Werling RW, Yaziji H, Bacchi CE, Gown AM. CDX2, a highly sensitive and specific marker of adenocarcinomas of intestinal origin: an immunohistochemical survey of 476 primary and metastatic carcinomas. *Am J Surg Pathol* 2003; **27**: 303–310.
70. Moskaluk CA, Zhang H, Powell SM, *et al*. Cdx2 protein expression in normal and malignant human tissues: an immunohistochemical survey using tissue microarrays. *Mod Pathol* 2003; **16**: 913–919.
71. Ordonez NG. Application of mesothelin immunostaining in tumor diagnosis. *Am J Surg Pathol* 2003; **27**: 1418–1428.
72. Kaufmann O, Dietel M. Thyroid transcription factor-1 is the superior immunohistochemical marker for pulmonary adenocarcinomas and large cell carcinomas compared to surfactant proteins A and B. *Histopathology* 2000; **36**: 8–16.
73. Kaufmann O, Deidesheimer T, Muehlenberg M, Deicke P, Dietel M. Immunohistochemical differentiation of metastatic breast carcinomas from metastatic adenocarcinomas of other common primary sites. *Histopathology* 1996; **29**: 233–240.
74. Frierson HF Jr, Moskaluk CA, Powell SM, *et al*. Large-scale molecular and tissue microarray analysis of mesothelin

expression in common human carcinomas. *Hum Pathol* 2003; **34**: 605–609.
75. Wick MR, Lillemoe TJ, Copland GT, *et al*. Gross cystic disease fluid protein-15 as a marker for breast cancer: immunohistochemical analysis of 690 human neoplasms and comparison with alpha-lactalbumin. *Hum Pathol* 1989; **20**: 281–287.
76. Loy TS, Quesenberry JT, Sharp SC. Distribution of CA 125 in adenocarcinomas. An immunohistochemical study of 481 cases. *Am J Clin Pathol* 1992; **98**: 175–179.
77. Chu P, Wu E, Weiss LM. Cytokeratin 7 and cytokeratin 20 expression in epithelial neoplasms: a survey of 435 cases. *Mod Pathol* 2000; **13**: 962–972.
78. Park SY, Kim HS, Hong EK, Kim WH. Expression of cytokeratins 7 and 20 in primary carcinomas of the stomach and colorectum and their value in the differential diagnosis of metastatic carcinomas to the ovary. *Hum Pathol* 2002; **33**: 1078–1085.
79. Stenhouse G, Fyfe N, King G, Chapman A, Kerr KM. Thyroid transcription factor 1 in pulmonary adenocarcinoma. *J Clin Pathol* 2004; **57**: 383–387.
80. Machado JC, Nogueira AM, Carneiro F, Reis CA, Sobrinho-Simoes M. Gastric carcinoma exhibits distinct types of cell differentiation: an immunohistochemical study of trefoil peptides (TFF1 and TFF2) and mucins (MUC1, MUC2, MUC5AC, and MUC6). *J Pathol* 2000; **190**: 437–443.
81. Jobsis AC, De Vries GP, Meijer AE, Ploem JS. The immunohistochemical detection of prostatic acid phosphatase: its possibilities and limitations in tumour histochemistry. *Histochem J* 1981; **13**: 961–973.
82. Haines AM, Larkin SE, Richardson AP, Stirling RW, Heyderman E. A novel hybridoma antibody (PASE/4LJ) to human prostatic acid phosphatase suitable for immunohistochemistry. *Br J Cancer* 1989; **60**: 887–892.
83. Ohshio G, Ogawa K, Kudo H, *et al*. Immunohistochemical studies on the localization of cancer associated antigens DU-PAN-2 and CA19-9 in carcinomas of the digestive tract. *J Gastroenterol Hepatol* 1990; **5**: 25–31.
84. Taguchi T, Kijima H, Mitomi T, Osamura RY. Immunohistochemical study of colorectal adenocarcinomas and adenomas with antibodies against carcinoembryonic antigen (CEA), CA19-9, keratin, alpha-tubulin and secretory component (SC). *Gastroenterol Jpn* 1991; **26**: 294–302.
85. Yamaguchi K, Enjoji M, Tsuneyoshi M. Pancreatoduodenal carcinoma: a clinicopathologic study of 304 patients and immunohistochemical observation for CEA and CA19-9. *J Surg Oncol* 1991; **47**: 148–154.
86. Mizutani Y, Nakajima T, Morinaga S, *et al*. Immunohistochemical localization of pulmonary surfactant apoproteins in various lung tumors. Special reference to nonmucus producing lung adenocarcinomas. *Cancer* 1988; **61**: 532–537.
87. Nicholson AG, McCormick CJ, Shimosato Y, Butcher DN, Sheppard MN. The value of PE-10, a monoclonal antibody against pulmonary surfactant, in distinguishing primary and metastatic lung tumours. *Histopathology* 1995; **27**: 57–60.
88. Arends JW, Verstynen C, Bosman FT, Hilgers J, Steplewski Z. Distribution of monoclonal antibody-defined monosialoganglioside in normal and cancerous human tissues: an immunoperoxidase study. *Hybridoma* 1983; **2**: 219–229.
89. Magnani JL, Nilsson B, Brockhaus M, *et al*. A monoclonal antibody-defined antigen associated with gastrointestinal cancer is a ganglioside containing sialylated lacto-*N*-fucopentaose II. *J Biol Chem* 1982; **257**: 14 365–14 369.
90. Atkinson BF, Ernst CS, Herlyn M, *et al*. Gastrointestinal cancer-associated antigen in immunoperoxidase assay. *Cancer Res* 1982; **42**: 4820–4823.
91. Walker RA, Day SJ. Expression of the antigen detected by the monoclonal antibody CA 19.9 in human breast tissues. *Virchows Arch A Pathol Anat Histopathol* 1986; **409**: 375–383.
92. Neunteufel W, Breitenecker G. Tissue expression of CA 125 in benign and malignant lesions of ovary and Fallopian tube: a comparison with CA 19-9 and CEA. *Gynecol Oncol* 1989; **32**: 297–302.
93. Almeida R, Silva E, Santos-Silva F, *et al*. Expression of intestine-specific transcription factors, CDX1 and CDX2, in intestinal metaplasia and gastric carcinomas. *J Pathol* 2003; **199**: 36–40.
94. Satoh F, Umemura S, Osamura RY. Immunohistochemical analysis of GCDFP-15 and GCDFP-24 in mammary and non-mammary tissue. *Breast Cancer* 2000; **7**: 49–55.
95. Mazoujian G, Pinkus GS, Davis S, Haagensen DE Jr. Immunohistochemistry of a gross cystic disease fluid protein (GCDFP-15) of the breast. A marker of apocrine epithelium and breast carcinomas with apocrine features. *Am J Pathol* 1983; **110**: 105–112.
96. Miettinen M, Sarlomo-Rikala M. Expression of calretinin, thrombomodulin, keratin 5, and mesothelin in lung carcinomas of different types: an immunohistochemical analysis of 596 tumors in comparison with epithelioid mesotheliomas of the pleura. *Am J Surg Pathol* 2003; **27**: 150–158.
97. Argani P, Iacobuzio-Donahue C, Ryu B, *et al*. Mesothelin is overexpressed in the vast majority of ductal adenocarcinomas of the pancreas: identification of a new pancreatic cancer marker by serial analysis of gene expression (SAGE). *Clin Cancer Res* 2001; **7**: 3862–3868.
98. Leung WK, Yu J, Chan FK, *et al*. Expression of trefoil peptides (TFF1, TFF2, and TFF3) in gastric carcinomas, intestinal metaplasia, and non-neoplastic gastric tissues. *J Pathol* 2002; **197**: 582–588.
99. Tan D, Li Q, Deeb G, *et al*. Thyroid transcription factor-1 expression prevalence and its clinical implications in non-small cell lung cancer: a high-throughput tissue microarray and immunohistochemistry study. *Hum Pathol* 2003; **34**: 597–604.
100. Sagol O, Tuna B, Coker A, *et al*. Immunohistochemical detection of pS2 protein and heat shock protein-70 in pancreatic adenocarcinomas. Relationship with disease extent and patient survival. *Pathol Res Pract* 2002; **198**: 77–84.
101. Ahr A, Scharl A, Gohring UJ, Crombach G, Stoffl M. Immunohistochemical detection of pS2 protein in paraffin sections of breast carcinoma tissue. Comparison with results of an immunoradiometry assay. *Pathologe* 1995; **16**: 278–284.
102. Soubeyran I, Wafflart J, Bonichon F, *et al*. Immunohistochemical determination of pS2 in invasive breast carcinomas: a study on 942 cases. *Breast Cancer Res Treat* 1995; **34**: 119–128.
103. Pallud C, Le Doussal V, Pichon MF, *et al*. Immunohistochemistry of pS2 in normal human breast and in various histological forms of breast tumours. *Histopathology* 1993; **23**: 249–256.
104. Luqmani YA, Ryall G, Shousha S, Coombes RC. An immunohistochemical survey of pS2 expression in human epithelial cancers. *Int J Cancer* 1992; **50**: 302–304.
105. Dennis JL, Vass JK, Wit EC, Keith WN, Oien KA. Identification from public data of molecular markers of adenocarcinoma characteristic of the site of origin. *Cancer Res* 2002; **62**: 5999–6005.
106. Buckhaults P, Zhang Z, Chen YC, *et al*. Identifying tumor origin using a gene expression-based classification map. *Cancer Res* 2003; **63**: 4144–4149.
107. Su AI, Welsh JB, Sapinoso LM, *et al*. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* 2001; **61**: 7388–7393.
108. Ramaswamy S, Tamayo P, Rifkin R, *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 2001; **98**: 15 149–15 154.
109. Bloom G, Yang IV, Boulware D, *et al*. Multi-platform, multi-site, microarray-based human tumor classification. *Am J Pathol* 2004; **164**: 9–16.
110. Nishizuka S, Sing-Tsung C, Gwadry FG, *et al*. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic and tissue array profiling. *Cancer Res* 2003; **63**: 5243–5250.