Systematic variation in gene expression patterns in human cancer cell lines

Douglas T. Ross[1], Uwe Scherf[5], Michael B. Eisen[2], Charles M. Perou[2], Christian Rees[2], Paul Spellman[2], Vishwanath Iyer[1], Stefanie S. Jeffrey[3], Matt Van de Rijn[4], Mark Waltham[5], Alexander Pergamenschikov[2], Jeffrey C.F. Lee[6], Deval Lashkari[7], Dari Shalon[6], Timothy G. Myers[8], John N. Weinstein[5], David Botstein[2] & Patrick O. Brown[1,9]

We used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 cell lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression patterns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of the expression patterns in the tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours *in vivo*.

Introduction

Cell lines derived from human tumours have been extensively used as experimental models of neoplastic disease. Although such cell lines differ from both normal and cancerous tissue, the inaccessibility of human tumours and normal tissue makes it likely that such cell lines will continue to be used as experimental models for the foreseeable future. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines (the NCI60) derived from tumours from a variety of tissues and organs[1–4]. The DTP has assessed many molecular features of the cells related to cancer and chemotherapeutic sensitivity, and has measured the sensitivities of these 60 cell lines to more than 70,000 different chemical compounds, including all common chemotherapeutics (http://dtp.nci.nih.gov). A previous analysis of these data revealed a connection between the pattern of activity of a drug and its method of action. In particular, there was a tendency for groups of drugs with similar patterns of activity to have related methods of action[3,5–7].

We used DNA microarrays to survey the variation in abundance of approximately 8,000 distinct human transcripts in these 60 cell lines. Because of the logical connection between the function of a gene and its pattern of expression, the correlation of gene expression patterns with the variation in the phenotype of the cell can begin the process by which the function of a gene can be inferred. Similarly, the patterns of expression of known genes can reveal novel phenotypic aspects of the cells and tissues studied[8–10]. Here we present an analysis of the observed patterns of gene expression and their relationship to phenotypic properties of the 60 cell lines. The accompanying report[11] explores the relationship between the gene expression patterns and the drug sensitivity profiles measured by the DTP. The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions *in vitro* and *in vivo*, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes[11–16]. The databases, plus tools for analysis and visualization of the data, are available (http://genome-www.stanford.edu/nci60 and http://discover.nci.nih.gov).

Results

We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides[17,18]. The cDNAs included approximately 8,000 different genes: approximately 3,700 represented previously characterized human proteins, an additional 1,900 had homologues in other organisms and the remaining 2,400 were identified only by ESTs. Due to ambiguity of the identity of the cDNA clones used in these studies, we estimated that approximately 80% of the genes in these experiments were correctly identified. The identities of approximately 3,000 cDNAs

Departments of [1]Biochemistry, [2]Genetics, [3]Surgery and [4]Pathology, Stanford University School of Medicine, Stanford, California, USA. [5]Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [6]Incyte Pharmaceuticals, Fremont, California, USA. [7]Genometrix Inc., The Woodlands, Texas, USA. [8]Information Technology Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA. [9]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu) or J.N.W. (e-mail: Weinstein@dtpax2.ncifcrf.gov).

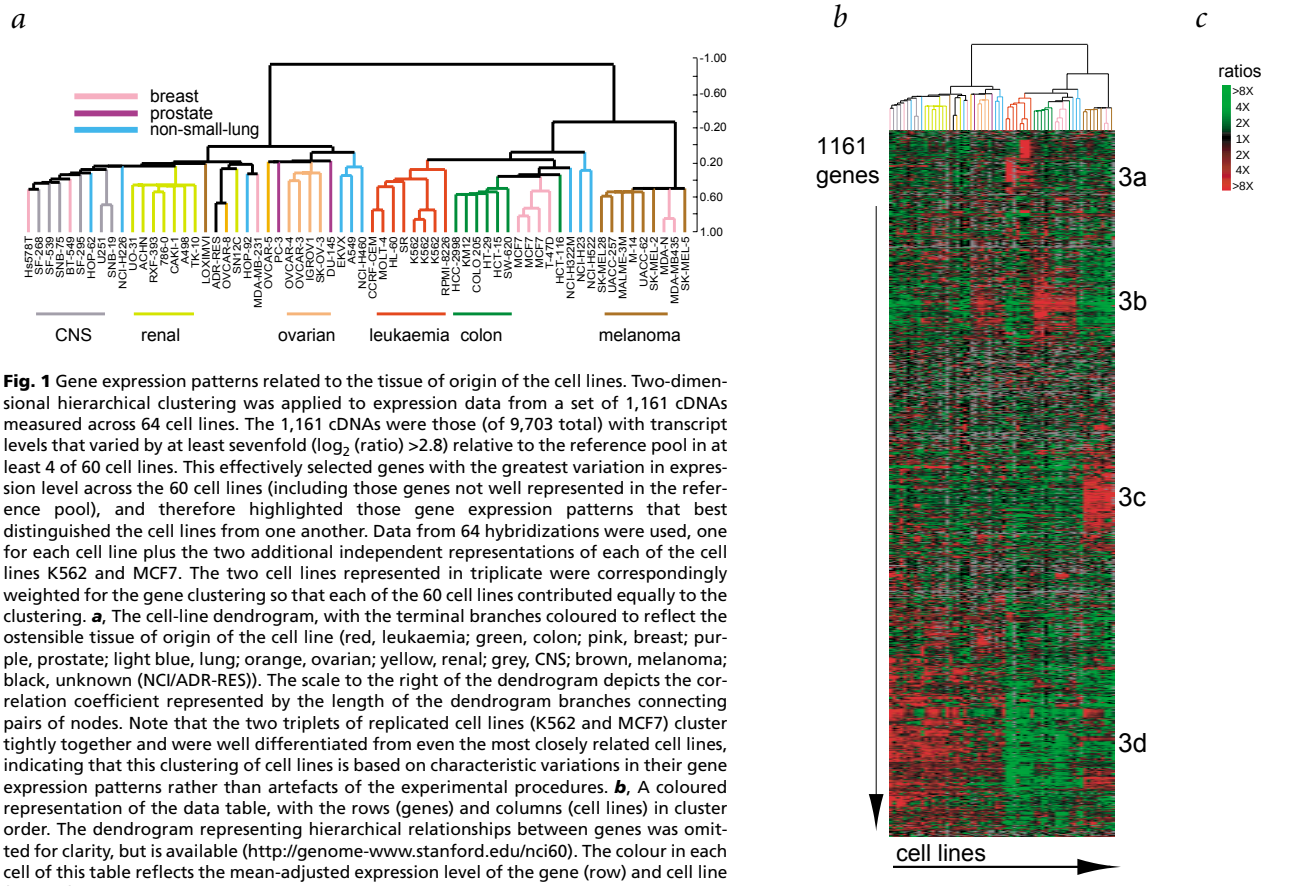

Fig. 1 Gene expression patterns related to the tissue of origin of the cell lines. Two-dimensional hierarchical clustering was applied to expression data from a set of 1,161 cDNAs measured across 64 cell lines. The 1,161 cDNAs were those (of 9,703 total) with transcript levels that varied by at least sevenfold ($\log_2$ (ratio) >2.8) relative to the reference pool in at least 4 of 60 cell lines. This effectively selected genes with the greatest variation in expression level across the 60 cell lines (including those genes not well represented in the reference pool), and therefore highlighted those gene expression patterns that best distinguished the cell lines from one another. Data from 64 hybridizations were used, one for each cell line plus the two additional independent representations of each of the cell lines K562 and MCF7. The two cell lines represented in triplicate were correspondingly weighted for the gene clustering so that each of the 60 cell lines contributed equally to the clustering. ***a***, The cell-line dendrogram, with the terminal branches coloured to reflect the ostensible tissue of origin of the cell line (red, leukaemia; green, colon; pink, breast; purple, prostate; light blue, lung; orange, ovarian; yellow, renal; grey, CNS; brown, melanoma; black, unknown (NCI/ADR-RES)). The scale to the right of the dendrogram depicts the correlation coefficient represented by the length of the dendrogram branches connecting pairs of nodes. Note that the two triplets of replicated cell lines (K562 and MCF7) cluster tightly together and were well differentiated from even the most closely related cell lines, indicating that this clustering of cell lines is based on characteristic variations in their gene expression patterns rather than artefacts of the experimental procedures. ***b***, A coloured representation of the data table, with the rows (genes) and columns (cell lines) in cluster order. The dendrogram representing hierarchical relationships between genes was omitted for clarity, but is available (http://genome-www.stanford.edu/nci60). The colour in each cell of this table reflects the mean-adjusted expression level of the gene (row) and cell line (column). The colour scale used to represent the expression ratios is shown. The labels '3a–3d' in (*b*) refer to the clusters of genes shown in detail in Fig. 3.

from these experiments have been sequence-verified, including all of those referred to here by name.

Each hybridization compared Cy5-labelled cDNA reverse transcribed from mRNA isolated from one of the cell lines with Cy3-labelled cDNA reverse transcribed from a reference mRNA sample. This reference sample, used in all hybridizations, was prepared by combining an equal mixture of mRNA from 12 of the cell lines (chosen to maximize diversity in gene expression as determined primarily from two-dimensional gel studies[2]). By comparing cDNA from each cell line with a common reference, variation in gene expression across the 60 cell lines could be inferred from the observed variation in the normalized Cy5/Cy3 ratios across the hybridizations.

To assess the contribution of artefactual sources of variation in the experimentally measured expression patterns, K562 and MCF7 cell lines were each grown in three independent cultures, and the entire process was carried out independently on mRNA extracted from each culture. The variance in the triplicate fluorescence ratio measurements approached a minimum when the fluorescence signal was greater than approximately 0.4% of the measurable total signal dynamic range above background in either channel of the hybridization. We selected the subset of spots for which significant signal was present in both the numerator and denominator of the ratios by this criterion to identify the best-measured spots. The pair-wise correlation coefficients for the triplicates of the set of genes that passed this quality control level (6,992 spots included for the MCF7 samples and 6,161 spots for K562) ranged from 0.83 to 0.92 (for graphs and details, see http://genome-www.stanford.edu/nci60).

To make the orderly features in the data more apparent, we used a hierarchical clustering algorithm[19,20] and a pseudo-colour visualization matrix[3,21]. The object of the clustering was to group cell lines with similar repertoires of expressed genes and to group genes whose expression level varied among the 60 cell lines in a similar manner. Clustering was performed twice using different subsets of genes to assess the robustness of the analysis. In one case (Fig. 1), we concentrated on those genes that showed the most variation in expression among the 60 cell lines (1,167 total). A second analysis (Fig. 2) included all spots that were thought to be well measured in the reference set (6,831 spots).

Gene expression patterns related to the histologic origins of the cell lines

The most notable property of the clustered data was that cell lines with common presumptive tissues of origin grouped together (Figs 1*a* and 2). Cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Many of these coherent cell line clusters were distinguished by the specific expression of characteristic groups of genes (Fig. 3*a–d*). For example, a cluster of approximately 90 genes was highly expressed in the melanoma-derived lines (Fig. 3*c*). This set was enriched for genes with known roles in melanocyte biology, including tyrosinase and dopachrome tautomerase (TYR and DCT; two subunits of an enzyme complex involved in melanin synthesis[22]), MART1 (MLANA; which is being investigated as a target for immunotherapy of melanoma[23]) and S100-β (S100B; which has been used as an antigenic marker in the diagnosis of

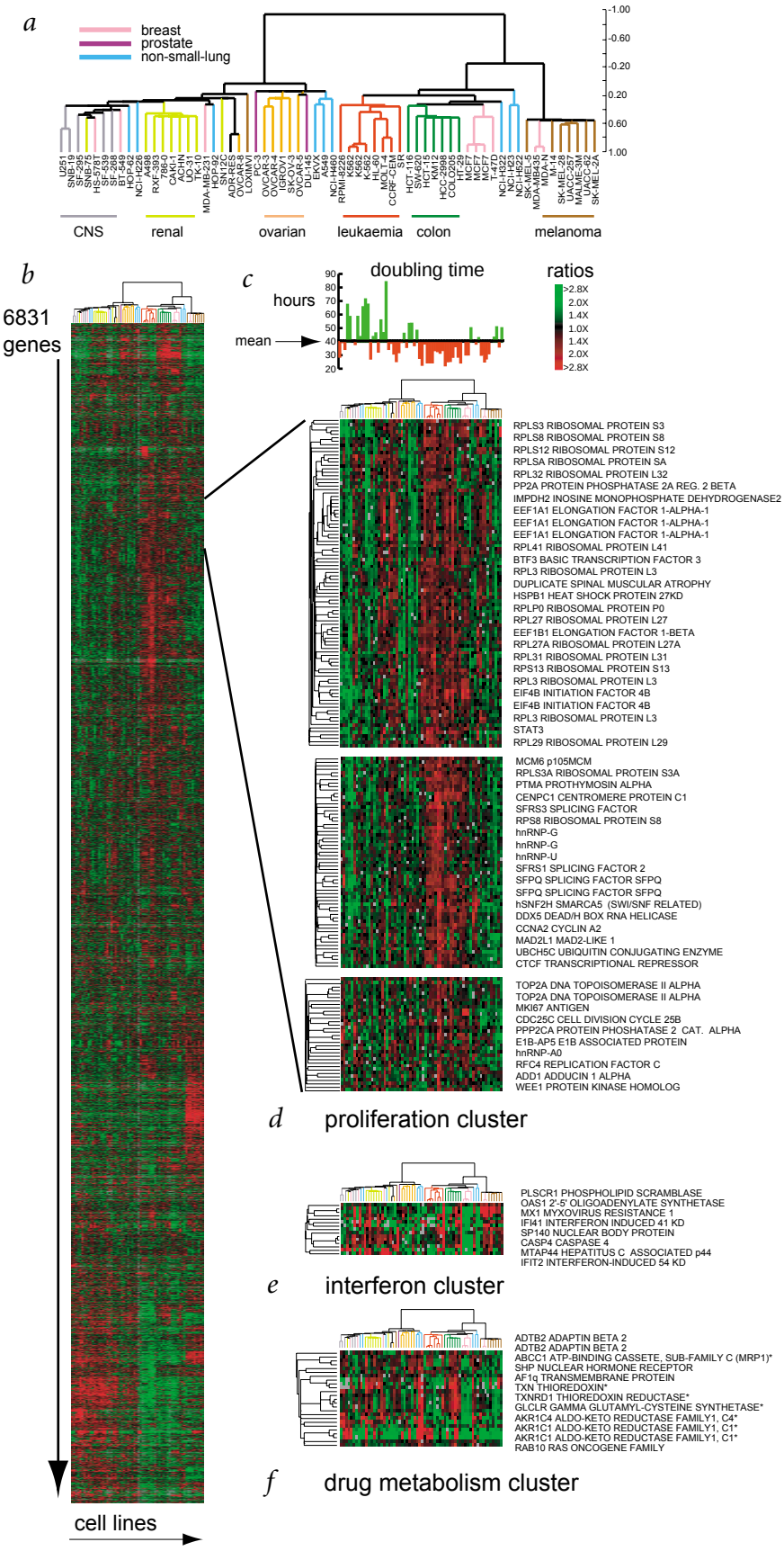

Fig. 2 Gene expression patterns related to other cell-line phenotypes. ***a***, We applied two-dimensional hierarchical clustering to expression data from a set of 6,831 cDNAs measured across the 64 cell lines. The 6,831 cDNAs were those with a minimum fluorescence signal intensity of approximately 0.4% of the dynamic range above background in the reference channel in each of the six hybridizations used to establish reproducibility. This effectively selected those spots that provided the most reliable ratio measurements and therefore identified a subset of genes useful for exploring patterns comprised of those whose variation in expression across the 60 cell lines was of moderate magnitude. ***b***, Cluster-ordered data table. ***c***, Doubling time of cell lines. Cell lines are given in cluster order. Values are plotted relative to the mean. Doubling times greater than the mean are shown in green, those with doubling time less than the mean are shown in red. ***d***, Three related gene clusters that were enriched for genes whose expression level variation was correlated with cell line proliferation rate. Each of the three gene clusters (clustered solely on the basis of their expression patterns) showed enrichment for sets of genes involved in distinct functional categories (for example, ribosomal genes versus genes involved in pre-RNA splicing). ***e***, Gene cluster in which all characterized and sequence-verified cDNAs encode genes known to be regulated by interferons. ***f***, Gene cluster enriched for genes that have been implicated in drug metabolism (indicated by asterisks). A further property of the gene clustering evident here and in Fig. 2 is the strong tendency for redundant representations of the same gene to cluster immediately adjacent to one another, even within larger groups of genes with very similar expression patterns. In addition to illustrating the reproducibility and consistency of the measurements, and providing independent confirmation of many of our measurements, this property also demonstrates that these, and probably all, genes have nearly unique patterns of variation across the 60 cell lines. If this were not the case, and multiple genes had identical patterns of variation, we would not expect to be able to distinguish, by clustering on the basis of expression variation, duplicate copies of individual genes from the other genes with identical expression patterns.

melanoma). LOXIMVI, the seventh line designated as melanoma in the NCI60, did not show this characteristic pattern. Although isolated from a patient with melanoma, LOXIMVI has previously been noted to lack melanin and other markers useful for identification of melanoma cells[1].

Paradoxically, two related cell lines (MDA-MB435 and MDA-N), which were derived from a single patient with breast cancer and have been conventionally regarded as breast cancer cell lines, shared expression of the genes associated with melanoma. MDA-MB435 was isolated from a pleural effusion in a patient with metastatic ductal adenocarcinoma of the breast[24,25]. It remains possible that the origin of the cell line was a breast cancer, and that its gene expression pattern is related to the neuroendocrine features of some breast cancers[26]. But our results suggest that this cell line may have originated from a melanoma, raising the possibility that the patient had a co-existing occult melanoma.

The higher-level organization of the cell-line tree—in which groups span cell lines from different tissue types—also reflected shared biological properties of the tissues from which the cell lines were derived. The carcinoma-derived cell lines were divided into major branches that separated those that expressed genes characteristic of epithelial cells from those that expressed genes more typical of stromal cells. A cluster of genes is shown (Fig. 3*b*) that is most strongly expressed in cell lines derived from colon carcinomas, six of seven ovarian-derived cell lines and the two breast cancer lines positive for the oestrogen receptor. The named genes in this cluster have been implicated in several aspects of epithelial cell biology[27]. The cluster was enriched for genes whose products are known to localize to the basolateral membrane of epithelial cells, including those encoding components of adherens complexes (for example, desmoplakin (DSP), periplakin (PPL) and plakoglobin (JUP)), an epithelial-expressed cell-cell adhesion molecule (M4S1) and a sodium/hydrogen ion exchanger[28–31] (SLC9A1). It also contained genes that encode putative transcriptional regulators of epithelial morphogenesis, a human homologue of a *Drosophila melanogaster* epithelial-expressed tumour suppressor (LLGL1) and a homeobox gene thought to control calcium-mediated adherence in epithelial cells[32,33] (MSX2).

In contrast, a separate, major branch of the cell-line dendrogram (Fig. 1*a*) included all glioblastoma-derived cell lines, all renal-cell-carcinoma–derived cell lines and the remaining carcinoma-derived lines. The characteristic set of genes expressed in this cluster included many whose products are involved in stromal cell functions (Fig. 3*d*). Indeed, the two cell lines originally described as 'sarcoma-like' in appearance (Hs578T, breast carcinosarcoma, and SF539, gliosarcoma) expressed most of these genes[34,35]. Although no single gene was uniformly characteristic of this cluster, each cell line showed a distinctive pattern of expression of genes encoding proteins with roles in synthesis or modification of the extracellular matrix (for example, caldesmon (CALD1), cathepsins, thrombospondin (THBS), lysyl oxidase (LOX) and collagen subtypes). Although the ovarian and most non-small-cell-lung–derived carcinomas expressed genes characteristic of both epithelial cells and stromal cells, they probably clustered with the CNS and renal cell carcinomas in this analysis because genes characteristically expressed in stromal cells were more abundantly represented in this gene set.

Physiological variation reflected in gene expression patterns

A cluster diagram of 6,831 genes (Fig. 2) is useful for exploring clusters of genes whose variation in mRNA levels was not obviously attributable to cell or tissue type. We identified some gene clusters that were enriched for genes involved in specific cellular processes; the variation in their expression levels may reflect corresponding differences in activity of these processes in the cell lines. For example, a cluster of 1,159 genes (Fig. 2*a*) included many whose products are necessary for progression through the cell cycle (such as CCNA1, MCM106 and MAD2L1), RNA processing and translation machinery (such as RNA helicases, hnRNPs and translation elongation factors) and traditional pathologic markers used to identify proliferating cells (MKI67). Within this large cluster were smaller clusters enriched for genes with more specialized roles. One cluster was highly enriched for numerous ribosomal genes, whereas another was more enriched for genes encoding RNA-splicing factors. The variation in expression of these ribosomal genes was significantly correlated with variation in the cell doubling time (correlation coefficient of 0.54), supporting the notion that the genes in this cluster were regulated in relation to cell proliferation rate or growth rate in these cell lines.

In a smaller gene cluster (Fig. 2*d*), all of the named genes were previously known to be regulated by interferons[13,36]. Additional groups of interferon-regulated genes showed distinct patterns of expression (data not shown), suggesting that the NCI60 cell lines exhibited variation in activity of interferon-response pathways, which was reflected in gene expression patterns[36].

Another cluster (Fig. 2*e*) contained several genes encoding proteins with possible interrelated roles in drug metabolism, including glutamate-cysteine ligase (GLCLC, the enzyme responsible for the rate limiting step of glutathione synthesis), thioredoxin (TXN) and thioredoxin reductase (TXNRD1; enzymes involved in regulating redox state in cells), and MRP1 (a drug transporter known to efficiently transport glutathione-conjugated compounds[37]). The elevated expression of this set of genes in a subset of these cell lines may reflect selection for resistance to chemotherapeutics.

Cell lines facilitate interpretation of gene expression patterns in complex clinical samples

Like many other types of cancer, tumours of the breast typically have a complex histological organization, with connective tissue and leukocytic infiltrates interwoven with tumour cells. To explore the possibility that variation in gene expression in the tumour cell lines might provide a framework for interpreting the expression patterns in tumour specimens, we compared RNA isolated from two breast cancer biopsy samples, a sample of normal breast tissue and the NCI60 cell lines derived from breast cancers (excluding MDA-MB-435 and MDA-N) and leukaemias (Fig. 4). This clustering highlighted features of the gene expression pattern shared between the cancer specimens and individual cell lines derived from breast cancers and leukaemias.

The genes encoding keratin 8 (KRT8) and keratin 19 (KRT19), as well as most of the other 'epithelial' genes defined in the complete NCI60 cell line cluster, were expressed in both of the biopsy samples and the two breast-derived cell lines, MCF-7 and T47D, expressing the oestrogen receptor, suggesting that these transcripts originated in tumour cells with features similar to those of luminal epithelial cells (Fig. 5*a*). Expression of a set of genes characteristic of stromal cells, including collagen genes (*COL3A1*, *COL5A1* and *COL6A1*) and smooth muscle cell markers (*TAGLN*), was a feature shared by the tumour sample and the stromal-like cell lines Hs578T and BT549 (Fig. 5*b*). This feature of the expression pattern seen in the tumour samples is likely to be due to the stromal component of the tumour. The tumours also shared expression of a set of genes (Fig. 5*c*) with the multiple myeloma cell line (RPMI-8226), notably including immunoglobulin genes, consistent with the presence of B cells in the tumour (this was confirmed by staining with anti-

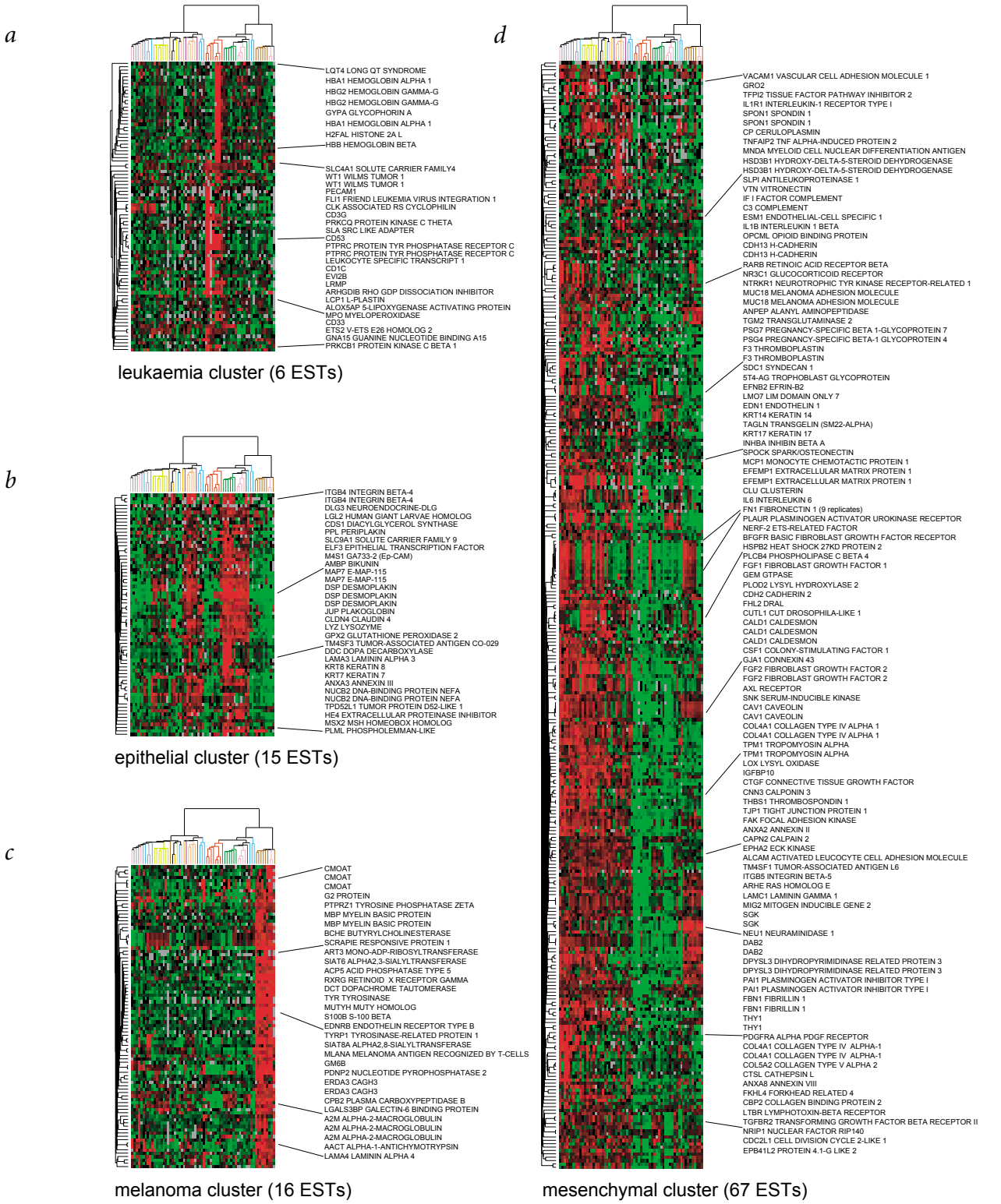

Fig. 3 Gene clusters related to tissue characteristics in the cell lines. Enlargements of the regions of the cluster diagram in Fig. 1 showing gene clusters enriched for genes expressed in cell lines of ostensibly similar origins. ***a***, Cluster of genes highly expressed in the leukaemia-derived cell lines. Two sub-clusters distinguish genes that were expressed in most leukaemia-derived lines from those expressed exclusively in the eryroblastoid line, K562 (note that the triplicate hybridizations cluster together). ***b***, Cluster of genes highly expressed in all colon (7/7) cell lines and all breast-derived cell lines positive for the oestrogen receptor (2/2). This set of genes was also moderately expressed in most ovarian lines (5/6) and some non-small-cell-lung (4/6) lines, but was expressed at a lower level in all renal-cancer–derived lines. ***c***, Cluster of genes highly expressed in most melanoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB435 and MDA-N). ***d***, Cluster of genes highly expressed in all glioblastoma (6/6) lines and most lines derived from renal-cell carcinoma (7/8), and more moderately expressed in a subset of carcinoma-derived lines. In all panels, names are shown only for all known genes whose identities were independently re-verified by sequencing. The number of sequence-validated ESTs within the cluster is indicated below the cluster in parentheses. The position of gene names in the adjacent list only approximates their position in the cluster diagram as indicated by the lines connecting the colour chart with the gene list. Complete cluster images with all gene names and accession numbers are available (http://genome-www.stanford.edu/nci60).

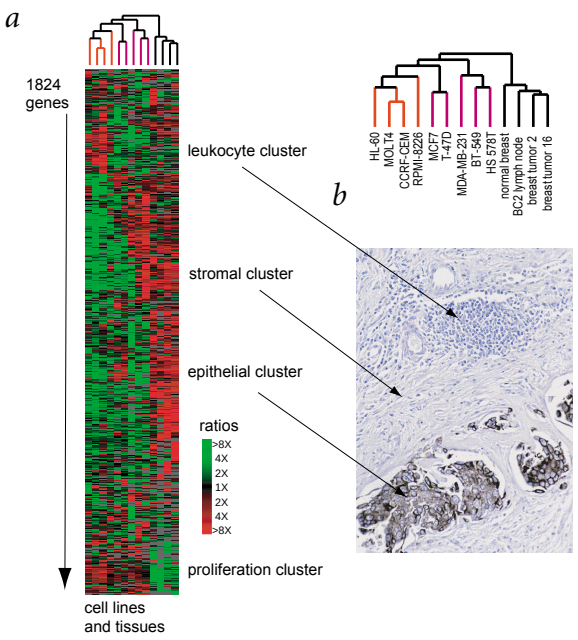

Fig. 4 Comparison of the gene expression patterns in clinical breast cancer specimens and cultured breast cancer and leukaemia cell lines. ***a***, Two-dimensional hierarchical clustering applied to gene expression data for two breast cancer specimens, a lymph node metastasis from one patient, normal breast and the NCI60 breast and leukaemia-derived cell lines. The gene expression data from tissue specimens was clustered along with expression data from a subset of the NCI60 cell lines to explore whether features of expression patterns observed in specific lines could be identified in the tissue samples. Labels indicate gene clusters (shown in detail in Fig. 5) that may be related to specific cellular components of the tumour specimens. ***b***, Breast cancer specimen 16 stained with anti-keratin antibodies, showing the complex mix of cell types characteristically found in breast tumours. The arrows highlight the different cellular components of this tissue specimen that were distinguished by the gene expression cluster analysis (Fig. 5).

immunoglobulin antibodies; data not shown). Therefore, distinct sets of genes with co-varying expression among the samples (Fig. 4, arrow) appear to represent distinct cell types that can be distinguished in breast cancer tissue. A fourth cluster of genes, more highly expressed in all of the cell lines than in any of the clinical specimens, was enriched for genes present in the 'proliferation' cluster described above (Fig. 5*d*). The variation in expression of these genes likely paralleled the difference in proliferation rate between the rapidly cycling cultured cell lines and the much more slowly dividing cells in tissues.

Discussion

Newly available genomics tools allowed us to explore variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumour tissues. We used a simple cluster analysis to identify the prominent features in the gene expression patterns that appeared to reflect 'molecular signatures' of the tissue from which the cells originated. The histological characteristics of the cell lines that dominated the clustering were pervasive enough that similar relationships were revealed when alternative subsets of genes were selected for analysis. Additional features of the expression pattern may be related to variation in physiological attributes such as proliferation rate and activity of interferon-response pathways.

The properties of the tumour-derived cell lines in this study have presumably all been shaped by selection for resistance to host defences and chemotherapeutics and for rapid proliferation in the tissue culture environment of synthetic growth media, fetal bovine serum and a polystyrene substratum. But the primary identifiable factor accounting for variation in gene expression patterns among these 60 cell lines was the identity of the tissue from which each cell line was ostensibly derived. For most of the cell lines we examined, neither physiological nor experimental adaptation for growth in culture was sufficient to overwrite the gene expression programs established during differentiation *in vivo*. Nevertheless, the prominence of mesenchymal features in the cell lines isolated from glioblastomas and carcinomas may reflect a selection for the relative ease of establishment of cell lines expressing stromal characteristics, perhaps combined with physiological adaptation to tissue culture conditions[38–40].

Biological themes linking genes with related expression patterns may be inferred in many cases from the shared attributes of known genes within the clusters. Uncharacterized cDNAs are likely to encode proteins that have roles similar to those of the known gene products with which they appear to be co-regulated. Still, for several clusters of genes, we were unable to discern a common theme linking the identified members of the cluster. Further exploration of their variation in expression under more diverse conditions and more comprehensive investigation of the physiology of the NCI60 cells may provide insight[10]. The relationship of the gene expression patterns to the drug sensitivity patterns measured by the DTP is an example of linking variation in gene expression with more subtle and diverse phenotypic variation[11].

The patterns of gene expression measured in the NCI60 cell lines provide a framework that helps to distinguish the cells that express specific sets of genes in the histologically complex breast cancer specimens[41]. Although it is now feasible to analyse gene expression in micro-dissected tumour specimens[42,43], this observation suggests that it will be possible to explore and interpret some of the biology of clinical tumour samples by sampling them intact. As is useful in conventional morphological pathology, one might be able to observe interactions between a tumour and its microenvironment in this way. These relationships will be clarified by suitable analysis of gene expression patterns from intact as well as dissected tumours[12,14,15,41].

Methods

cDNA clones. We obtained the 9,703 human cDNA clones (Research Genetics) used in these experiments as bacterial colonies in 96-well microtitre plates[9]. Approximately 8,000 distinct Unigene clusters (representing nominally unique genes) were represented in this set of clones. All genes identified here by name represent clones whose identities were confirmed by re-sequencing, or by the criteria that two or more independent cDNA clones ostensibly representing the same gene had nearly identical gene expression patterns. A single-pass 3′ sequence re-verification was attempted for every clone after re-streaking for single colonies. For a subset of genes for which quality 3′ sequence was not obtained, we attempted to confirm identities by 5′ sequencing. Of the subset of clones selected for 5′ sequence verification on the basis of an interesting pattern of expression (888 total), 331 were correctly identified, 57, incorrectly identified, and 500, indeterminate (poor quality sequence). We estimated that 15%–20% of array elements contained DNA representing more than one clone per well. So far, the identities of ~3,000 clones have been verified. The full list of clones used and their nominal identities are available (gene names preceded by the designation "SID#" (Stanford Identification) represent clones whose identities have not yet been verified; http://genome-www.stanford.edu:8000/nci60).

Production of cDNA microarrays. The arrays used in this experiment were produced at Synteni Inc. (now Incyte Pharmaceuticals). Each insert was amplified from a bacterial colony by sampling 1 μl of bacterial media and performing PCR amplification of the insert using consensus primers for the three plasmids represented in the clone set (5′–TTGTAAAACGACG GCCAGTG–3′, 5′–CACACAGGAAACAGCTATG–3′). Each PCR product

(100 µl) was purified by gel exclusion, concentrated and resuspended in 3×SSC (10 µl). The PCR products were then printed on treated glass microscope slides using a robot with four printing tips. Detailed protocols for assembling and operating a microarray printer, and printing and experimental application of DNA microarrays are available (http://cmgm.stanford.edu/pbrown).

Preparation of mRNA and reference pool. Cell lines were grown from NCI DTP frozen stocks in RPMI-1640 supplemented with phenol red, glutamine (2 mM) and 5% fetal calf serum. To minimize the contribution of variations in culture conditions or cell density to differential gene expression, we grew each cell line to 80% confluence and isolated mRNA 24 h after transfer to fresh medium. The time between removal from the incubator and lysis of the cells in RNA stabilization buffer was minimized (<1 min). Cells were lysed in buffer containing guanidium isothiocyanate and total RNA was purified with the RNeasy purification kit (Qiagen). We purified mRNA as needed using a poly(A) purification kit (Oligotex, Qiagen) according to the manufacturer's instructions. Denaturing agarose gel electrophoresis assessed the integrity and relative contamination of mRNA with ribosomal RNA.

The breast tumours were surgically excised from patients and rapidly transported to the pathology laboratory, where samples for microarray analysis were quickly frozen in liquid nitrogen and stored at –80 °C until use. A frozen tumour specimen was removed from the freezer, cut into small pieces (~50–100 mg each), immediately placed into 10–12 ml of Trizol reagent (Gibco-BRL) and homogenized using a PowerGen 125 Tissue Homogenizer (Fisher Scientific), starting at 5,000 r.p.m. and gradually increasing to ~20,000 r.p.m. over a period of 30–60 s. We processed the Trizol/tumour homogenate as described in the Trizol protocol, including an initial step to remove fat. Once total RNA was obtained, we isolated mRNA with a FastTrack 2.0 kit (Invitrogen) using the manufacturer's protocol for isolating mRNA starting from total RNA. The normal breast samples were obtained from Clontech.

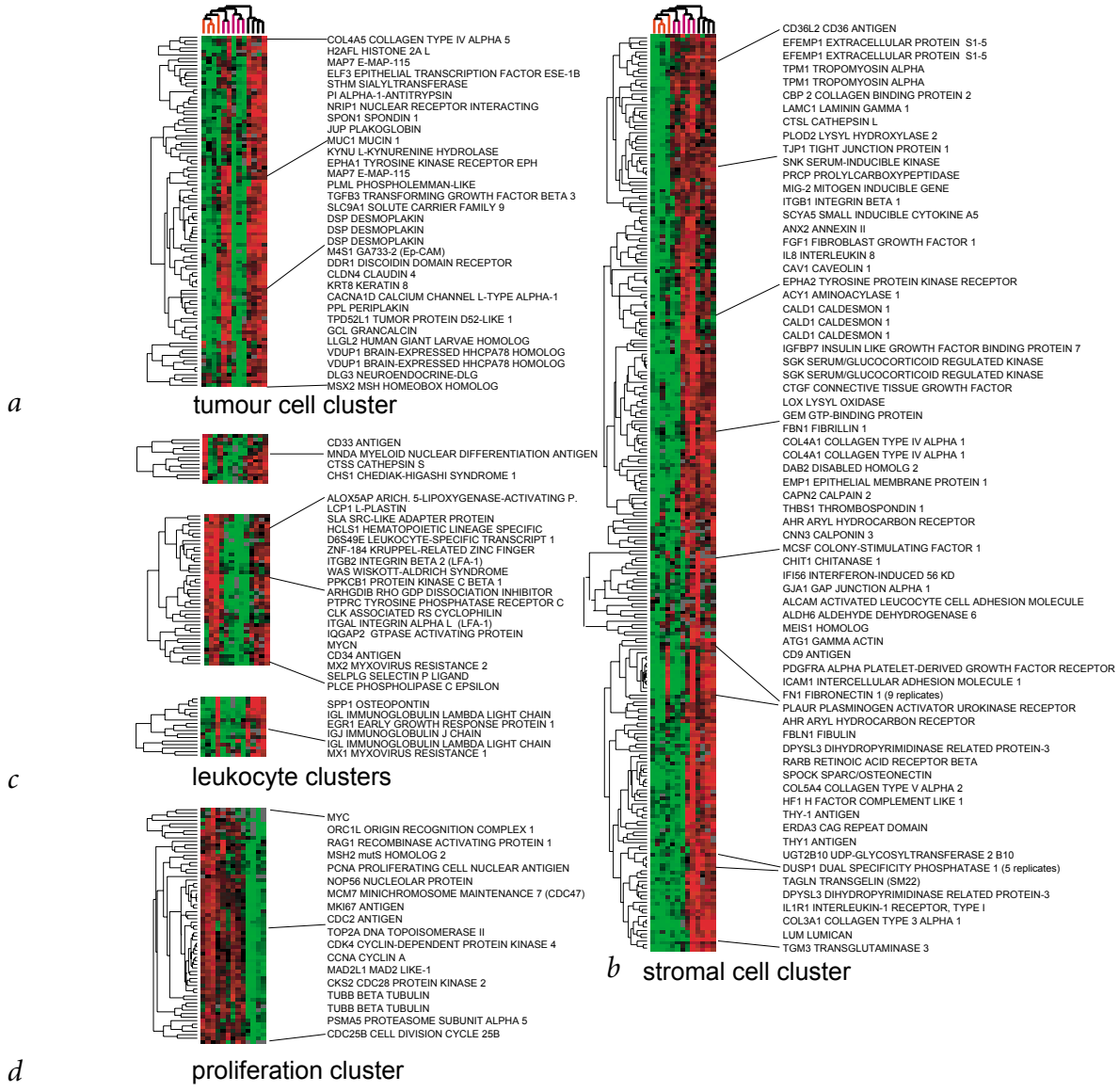

Fig. 5 Histologic features of breast cancer biopsies can be recognized and parsed based on gene expression patterns. Enlargements of the regions of the cluster diagram in Fig. 4 showing gene clusters enriched for genes expressed in different cell types in the breast cancer specimens, as distinguished by clustering with the cultured cell lines. ***a***, A cluster including many genes characteristic of epithelial cells expressed in cell lines (T47D and MCF7) derived from breast cancer positive for the oestrogen receptor and tumours. ***b***, Genes expressed in cell lines derived from breast cancer with stromal cell characteristics (Hs578T and BT549) and tumour specimens. Expression of these genes in the tumour samples may reflect the presence of myofibroblasts in the cancer specimen stroma. ***c***, Genes expressed in leukocyte-derived cell lines, showing common leukocyte, and separate 'myeloid' and 'B-cell', gene clusters. ***d***, Genes that were relatively highly expressed in all cell lines compared with the tumour specimens and normal breast. The higher expression of this set of genes involved in cell cycle transit in the cell lines is likely to reflect the higher proliferative rate of cells cultured in the presence of serum compared with the average proliferation rate of cells in the biopsied tissue.

We combined mRNA from the following cells in equal quantities to make the reference pool: HL-60 (acute myeloid leukaemia) and K562 (chronic myeloid leukaemia); NCI-H226 (non-small-cell-lung); COLO 205 (colon); SNB-19 (central nervous system); LOX-IMVI (melanoma); OVCAR-3 and OVCAR–4 (ovarian); CAKI-1 (renal); PC-3 (prostate); and MCF7 and Hs578T (breast). The criterion for selection of the cell lines in the reference are described in detail in the accompanying manuscript[12].

Doubling-time calculations. We calculated doubling times based on routine NCI60 cell line compound screening data; and they reflect the doubling times for cells inoculated into 96-well plates at the screening inoculation densities and grown in RPMI 1640 medium supplemented with 5% fetal bovine serum for 48 h. We measured cell populations using sulforhodamine B optical density measurement assay. The doubling time constant k was calculated using the equation: $N/No = e^{kt}$, where No is optical density for control (untreated) cells at time zero, *N* is optical density for control cells after 48-h incubation, and t is 48 h. The same equation was then used with the derived k to calculate the doubling time t by setting N/No = 2. For a given cell line, we obtained No and N values by averaging optical densities (N>6,000) obtained for each cell line for a year's screening. Data and experimental details are available (http://dtp.nci.nih.gov).

Preparation and hybridization of fluorescent labelled cDNA. For each comparative array hybridization, labelled cDNA was synthesized by reverse transcription from test cell mRNA in the presence of Cy5-dUTP, and from the reference mRNA with Cy3-dUTP, using the Superscript II reverse-transcription kit (Gibco-BRL). For each reverse transcription reaction, mRNA (2 μg) was mixed with an anchored oligo-dT (d-20T-d(AGC)) primer (4 μg) in a total volume of 15 μl, heated to 70 °C for 10 min and cooled on ice. To this sample, we added an unlabelled nucleotide pool (0.6 μl; 25 mM each dATP, dCTP, dGTP, and 15 mM dTTP), either Cy3 or Cy5 conjugated dUTP (3 μl; 1 mM; Amersham), 5×first-strand buffer (6 μl; 250 mM Tris-HCL, pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 0.1 M DTT (3 μl) and 2 μl of Superscript II reverse transcriptase (200 μ/μl). After a 2-h incubation at 42 °C, the RNA was degraded by adding 1 N NaOH (1.5 μl) and incubating at 70 °C for 10 min. The mixture was neutralized by adding of 1 N HCL (1.5 μl), and the volume brought to 500 μl with TE (10 mM Tris, 1 mM EDTA). We added Cot1 human DNA (20 μg; Gibco-BRL), and purified the probe by centrifugation in a Centricon-30 micro-concentrator (Amicon). The two separate probes were combined, brought to a volume of 500 μl, and concentrated again to a volume of less than 7 μl. We added 10 μg/μl poly(A) RNA (1 μl; Sigma) and tRNA (10 μg/μl; Gibco-BRL) were added, and adjusted the volume to 9.5 μl with distilled water. For final probe preparation, 20×SSC (2.1 μl; 1.5 M NaCl, 150 mM NaCitrate, pH 8.0) and 10% SDS (0.35 μl) were added to a total final volume of 12 μl. The probes were denatured by heating for 2 min at 100 °C, incubated at 37 °C for 20–30 min, and placed on the array under a 22 mm×22 mm glass coverslip. We incubated slides overnight at 65 °C for 14–18 h in a custom slide chamber with humidity maintained by a small reservoir of 3×SSC. Arrays were washed by submersion and agitation for 2–5 min in 2×SSC with 0.1% SDS, followed by 1×SSC and then 0.1×SSC. The arrays were "spun dry" by centrifugation for 2 min in a slide-rack in a Beckman GS-6 tabletop centrifuge in Microplus carriers at 650 r.p.m. for 2 min.

Array quantitation and data processing. Following hybridization, arrays were scanned using a laser-scanning microscope (ref. 17; http://cmgm.stanford.edu/pbrown). Separate images were acquired for Cy3 and Cy5. We carried out data reduction with the program ScanAlyze (M.B.E., available at http://rana.stanford.edu/software). Each spot was defined by manual positioning of a grid of circles over the array image. For each fluorescent image, the average pixel intensity within each circle was determined, and a local background was computed for each spot equal to the median pixel intensity in a square of 40 pixels in width and height centred on the spot centre, excluding all pixels within any defined spots. Net signal was determined by subtraction of this local background from the average intensity for each spot. Spots deemed unsuitable for accurate quantitation because of array artefacts were manually flagged and excluded from further analysis. Data files generated by ScanAlyze were entered into a custom database that maintains web-accessible files. Signal intensities between the two fluorescent images were normalized by applying a uniform scale factor to all intensities measured for the Cy5 channel. The normalization factor was chosen so that the mean log(Cy3/Cy5) for a subset of spots that achieved a minimum quality parameter (approximately 6,000 spots) was 0. This effectively defined the signal-intensity-weighted 'average' spot on each array to have a Cy3/Cy5 ratio of 1.0.

Cluster analysis. We extracted tables (rows of genes, columns of individual microarray hybridizations) of normalized fluorescence ratios from the database. Various selection criteria, discussed in relation to each data set, were applied to select subsets of genes from the 9,703 cDNA elements on the arrays. Before clustering and display, the logarithm of the measured fluorescence ratios for each gene were centred by subtracting the arithmetic mean of all ratios measured for that gene. The centring makes all subsequent analyses independent of the amount of each gene's mRNA in the reference pool.

We applied a hierarchical clustering algorithm separately to the cell lines and genes using the Pearson correlation coefficient as the measure of similarity and average linkage clustering[3,19–21]. The results of this process are two dendrograms (trees), one for the cell lines and one for the genes, in which very similar elements are connected by short branches, and longer branches join elements with diminishing degrees of similarity. For visual display the rows and columns in the initial data table were reordered to conform to the structures of the dendrograms obtained from the cluster analysis. Each cell in the cluster-ordered data table was replaced by a graded colour (pure red through black to pure green), representing the mean-adjusted ratio value in the cell. Gene labels in cluster diagrams are displayed here only for genes that were represented in the microarray by sequence-verified cDNAs. A complete software implementation of this process is available (http://rana.stanford.edu/software), as well as all clustering results (http://genome-www.stanford.edu/nci60).

Acknowledgements

We thank members of the Brown and Botstein labs for helpful discussions. This work was supported by the Howard Hughes Medical Institute and a grant from the National Cancer Institute (CA 077097). The work of U.S. and J.N.W. was supported in part by a grant from the National Cancer Institute Breast Cancer Think Tank. D.T.R is a Walter and Idun Berry Fellow. M.B.E. is an Alfred P. Sloan Foundation Fellow in Computational Molecular Biology. C.M.P. is a SmithKline Beecham Pharmaceuticals Fellow of the Life Science Research Foundation. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

Received 20 July 1999; accepted 13 January 2000.

1. Stinson, S.F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
2. Myers, T.G. *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18**, 647–653 (1997).
3. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
4. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
5. Paull, K.D. *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl Cancer Inst.* **81**, 1088–1092 (1989).
6. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
7. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* **86**, 1853–1859 (1994).
8. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
9. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
10. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21** (suppl.), 33–37 (1999).
11. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
12. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
13. Der, S.D., Zhou, A., Williams, B.R. & Silverman, R.H. Identification of genes differentially regulated by interferon-α, -β or -γ or using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
14. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
15. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
16. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
17. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
18. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
19. Sokal, R.R. & Sneath, P.H.A. *Principles of Numerical Taxonomy* (W.H. Freeman, San Francisco, 1963).
20. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
22. del Marmol, V. & Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* **381**, 165–168 (1996).
23. Kawakami, Y. *et al.* The use of melanosomal proteins in the immunotherapy of melanoma. *J. Immunother.* **21**, 237–246 (1998).
24. Cailleau, R., Olive, M. & Cruciger, Q.V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* **14**, 911–915 (1978).
25. Brinkley, B.R. *et al.* Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. *Cancer Res.* **40**, 3118–3129 (1980).
26. Nesland, J.M., Holm, R., Johannessen, J.V. & Gould, V.E. Neuroendocrine differentiation in breast lesions. *Pathol. Res. Pract.* **183**, 214–221 (1988).
27. Davies, J.A. & Garrod, D.R. Molecular aspects of the epithelial phenotype. *Bioessays* **19**, 699–704 (1997).
28. Garrod, D., Chidgey, M. & North, A. Desmosomes: differentiation, development, dynamics and disease. *Curr. Opin. Cell Biol.* **8**, 670–678 (1996).
29. Cowin, P. & Burke, B. Cytoskeleton-membrane interactions. *Curr. Opin. Cell Biol.* **8**, 56–65 (1996); erratum: **8**, 244 (1996).
30. Litvinov, S.V. *et al.* Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *J. Cell Biol.* **139**, 1337–1348 (1997).
31. Helmle-Kolb, C. *et al.* Na/H exchange activities in NHE1-transfected OK-cells: cell polarity and regulation. *Pflugers Arch.* **425**, 34–40 (1993); erratum: **427**, 387 (1994).
32. Manfruelli, P., Arquier, N., Hanratty, W.P. & Semeriva, M. The tumor suppressor gene, lethal(2)giant larvae (1(2)g1), is required for cell shape change of epithelial cells during Drosophila development. *Development* **122**, 2283–2294 (1996).
33. Lincecum, J.M., Fannon, A., Song, K., Wang, Y. & Sassoon, D.A. Msh homeobox genes regulate cadherin-mediated cell adhesion and cell-cell sorting. *J. Cell Biochem.* **70**, 22–28 (1998).
34. Hackett, A.J. *et al.* Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines. *J. Natl Cancer Inst.* **58**, 1795–1806 (1977).
35. Rutka, J.T. *et al.* Establishment and characterization of a cell line from a human gliosarcoma. *Cancer Res.* **46**, 5893–5902 (1986).
36. Nguyen, H., Hiscott, J. & Pitha, P.M. The growing family of interferon regulatory factors. *Cytokine Growth Factor Rev.* **8**, 293–312 (1997).
37. Moscow, J.A., Schneider, E., Ivy, S.P. & Cowan, K.H. Multidrug resistance. *Cancer Chemother. Biol. Response Modif.* **17**, 139–177 (1997).
38. Smith, H.S. & Hackett, A.J. The use of cultured human mammary epithelial cells in defining malignant progression. *Ann. N Y Acad. Sci.* **464**, 288–300 (1986).
39. Rutka, J.T. *et al.* Establishment and characterization of five cell lines derived from human malignant gliomas. *Acta Neuropathol.* **75**, 92–103 (1987).
40. Ronnov-Jessen, L., Petersen, O.W. & Bissell, M.J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* **76**, 69–125 (1996).
41. Perou, C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
42. Bonner, R.F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
43. Sgroi, D.C. *et al.* In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**, 5656–5661 (1999).