Molecular portraits and the family tree of cancer

Christine H. Chung[1], Philip S. Bernard[2] & Charles M. Perou[3]

doi:10.1038/ng1038

The twenty-first century heralds a new era for the biological sciences and medicine. The tools of our time are allowing us to analyze complex genomes more comprehensively than ever before. A principal technology contributing to this explosion of information is the DNA microarray, which enables us to study genome-wide expression patterns in complex biological systems. Although the potential of microarrays is yet to be fully realized, these tools have shown great promise in deciphering complex diseases such as cancer. The early results are painting a detailed portrait of cancer that illustrates the individuality of each tumor and allows familial relationships to be recognized through the identification of cell types sharing common expression patterns.

Cancer is a heterogeneous disease in most respects, including its cellularity, different genetic alterations and diverse clinical behaviors. Many analytical methods have been used to study human tumors and to classify samples into homogeneous groups that can predict clinical behavior. DNA microarrays have made significant contributions to this field by detecting similarities and differences among tumors through the simultaneous analysis of the expression of thousands of genes.

Gene expression data are often referred to as 'signatures' or 'portraits' because most tumors show expression patterns that are as unique and recognizable as paintings by Da Vinci or Pollock. Coupled with statistical analysis, DNA microarrays have allowed investigators to develop expression-based classifications for many types of cancer, including breast[1–6], brain[7,8], ovary[9–11], lung[12–14], colon[15–17], kidney[18], prostate[19–22], gastric[23], leukemia[24–26] and lymphoma[27–29]. Metaphorically speaking, the faces of those in the 'malignant family' may seem different from one another, but they all have features that are common to their family and that differentiate them from members of the 'benign family'. Some functional classes of genes are invariably altered when normal cells transform to malignant, including genes involved in cell-cycle control, adhesion and motility, apoptosis and angiogenesis[30]. Thus, despite the morphological and molecular heterogeneity among different cancer types, there are common threads that allow members to be recognized as branches of the same family tree.

A main challenge to the study and treatment of cancer is resolving the tumor heterogeneity that exists both between and within tumors. By light microscopy, the cellular complexity of a tumor can be visually dissected through differences in the appearance of malignant and non-malignant cells. By microarray, the make-up of complex tissue samples can be resolved as dominant patterns of gene expression representing the origin and function of different cell types. For example, solid tumors can be molecularly dissected into epithelial cells, infiltrating lymphocytes, adipose cells and surrounding stromal cells[1,31]. But microarray analysis can do more than differentiate a mixture of cell types and can often resolve levels of heterogeneity that are not apparent by eye. Because the clinical behavior of tumors cannot be accounted for completely by morphology, it is the hope of medicine that a molecular taxonomy based on 'signature' profiles will provide a more accurate prognosis and prediction of response to therapy.

The analysis of microarray data obtained from tumor samples is extremely complex. The analytical issues are reviewed in this issue by Churchill (pages 490–495)[32] and Slonim (pages 502–508)[33], and here we focus instead on the biological and clinical implications of profiling studies of human tumors.

Experimental design and analysis overview

For this review, we need to discuss two general statistical approaches for tumor classification. The first is 'supervised' analysis, in which one searches for genes whose expression patterns correlate with an external parameter. The most commonly used 'supervising' parameters are clinical features such as survival, presence of metastases and response to therapy. Many statistical metrics have been used successfully in 'supervised' analyses, including the standard *t*-test and signal-to-noise ratios[8,19,27].

Algorithms such as weighted-voting, *k*-nearest-neighbor classifiers, support vector machines and artificial neural networks can be applied to the set of genes selected using one of these metrics to build models capable of predicting the class of a particular sample. To test the robustness of classification, these methods are often coupled with a leave-one-out cross-validation analysis[3,8,19,24], in which one of the samples from the original 'training' set is withheld and a class prediction is made on the withheld sample[3,34].

The second approach is 'unsupervised' analysis, in which no external feature is used to guide the analysis process. Instead, the data are used to search for patterns without any *a priori* expectation concerning the number or type of groups that are present.

[1] *Division of Hematology/Oncology, University of North Carolina at Chapel Hill, CB #7305, 3009 Old Clinic Building, Chapel Hill, North Carolina 27599, USA.*
[2] *Department of Pathology, University of Utah, Huntsman Cancer Institute, 2000 Circle of Hope, Salt Lake City, Utah 84112-5550, USA.*
[3] *Department of Genetics, and Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Lineberger Comprehensive Cancer Center, CB #7295, Chapel Hill, North Carolina 27599, USA. Correspondence should be addressed to C.M.P. (e-mail: cperou@med.unc.edu).*

Bob Crimi

Fig. 1 Molecular portraits gallery.

The most common 'unsupervised' analysis method is hierarchical cluster analysis[35]. Each analytical method has its own strengths and weaknesses, and because classifications tend not to be mutually exclusive, most investigators base the significance of their microarray findings on more than one analysis.

Although both methods can analyze thousands of expressed genes, minimization of a discriminatory gene list can ease the biological interpretation and facilitate use in a clinical test. Several methods have been used for gene selection, such as correlation metrics or *t*-tests coupled with permutation testing[3,24,36]. Other methods work by selecting a gene list that gives rise to the highest prediction accuracy during leave-one-out cross-validation[3,19,24,27] or nearest 'centriods' analysis[37]. For complete validation, gene lists should be tested on a second 'test' set of samples that were not used to derive the discriminatory gene list.

Molecular portraits of individuality

In the cancer family, as in every family, each individual is unique and is a product of its genetics and environment. The concept of each tumor as an individual was proposed during the classification of breast carcinomas, because repeated samplings of the same tumor, either before and after chemotherapy or as a tumor–metastasis pair, were found to have much more similarity to each other than to any other tumor tested[1]. This important biological result showed that the technical aspects of microarrays are sound and led to the term 'molecular portrait' to describe the unique gene expression profile of a given tumor (Fig. 1). The individuality of single tumors as judged by hierarchical clustering analysis and other correlation analyses has now been demonstrated in breast, lung, liver and diffuse large B-cell lymphomas[1,13,14,29,38].

How much of a tumor's individuality is due to a person's own genetic make-up, and how much is due to the stochastic events that led to tumor formation? This question has been addressed by Chen *et al.*[38], who examined spatially separated hepatocellular carcinoma foci taken from the livers of six individuals. In three individuals, the two foci analyzed were 'clustered' adjacent to each other on a terminal branch of the hierarchical cluster dendrogram, indicating a very close relationship. In one individual, however, the two foci analyzed showed different hepatitis B virus (HBV) integration sites and were as different from one another as were any two randomly selected samples.

In the other two individuals, two of the three foci analyzed clustered together, whereas the third clustered apart. Although all of the foci in each individual shared common chromosomal alterations including the same HBV integration site, the orphan foci always showed aberrations that were not present in the other two. For example, the unpaired focus in one patient was positive for p53 immunostaining, whereas the other two were negative. These data imply that each independently arising tumor has a distinct profile and that clonally related tumors in the same individual can show different features owing to divergent histories.

Portraits of cell types

In addition to their individuality, tumors also show many similarities that can be ascribed to their cell type of origin. Microarrays are assisting in delineating the cell type specific branches of the cancer family tree by providing a tumor genealogy. For example, soft-tissue tumors such as synovial sarcomas, gastrointestinal stromal tumors (GISTs), neural tumors and a subset of leiomyosarcomas show markedly different patterns of gene expression[39]. For tumors such as leiomyosarcomas, where the cell type of origin is known, these gene expression patterns correlate well with the cell type of origin.

Tumors of the central nervous system (CNS) have also been classified according to their cell type of origin. Pomeroy *et al.*[8] assayed 99 samples of CNS tumors and defined a molecular descriptor using principal component analysis and 50 genes (Fig. 2). This descriptor successfully distinguished medulloblastomas from other histologically similar brain tumors (35 of 42 classified correctly; Fig. 2*a*). Their data also implicated cerebellar granule cells as the origin of medulloblastomas, and provided support for the idea that gliomas are probably derived from non-neural cells (such as oligodendrocytes).

Perhaps one of the greatest challenges in cancer medicine is to recognize the 'black sheep of the family'; that is, those tumors in the family that are clinically distinct but blend in by histology. For example, infiltrating breast carcinomas appear fairly homogenous in histological assessments but are heterogeneous in clinical behavior. In a study of mostly infiltrating ductal breast carcinomas (85%), at least five distinct tumor subtypes were identified by gene expression profiling[1,2]. This classification by hierarchical clustering divided the samples into two large groups: those positive for estrogen receptor (ER) and those negative for ER[2]. There were at least two subtypes in the large ER-positive group and three subtypes in the ER-negative group: one containing all of the normal breast samples, a second containing tumors distinguished by a high expression of HER2, and a third showing features of breast basal epithelial cells. These data suggest that ductal carcinomas are derived from two distinct types of cell (basal and luminal). In addition, these molecular subtypes were shown to predict overall survival and relapse-free survival times[2].

Much work has been carried out to classify lung tumors using microarrays[12–14,40]. Two separate groups using different microarray platforms (Affymetrix and cDNA microarrays) have defined similar patterns of expression that accurately recapitulated the four histological subtypes of lung cancer[13,14]. These data could mean that each subtype is derived from a distinct progenitor cell, with the progenitor of small-cell carcinomas showing neuroendocrine features, the progenitor of large-cell carcinomas showing mesenchymal features (possibly representing a cell type that has undergone an epithelial-to-mesenchymal

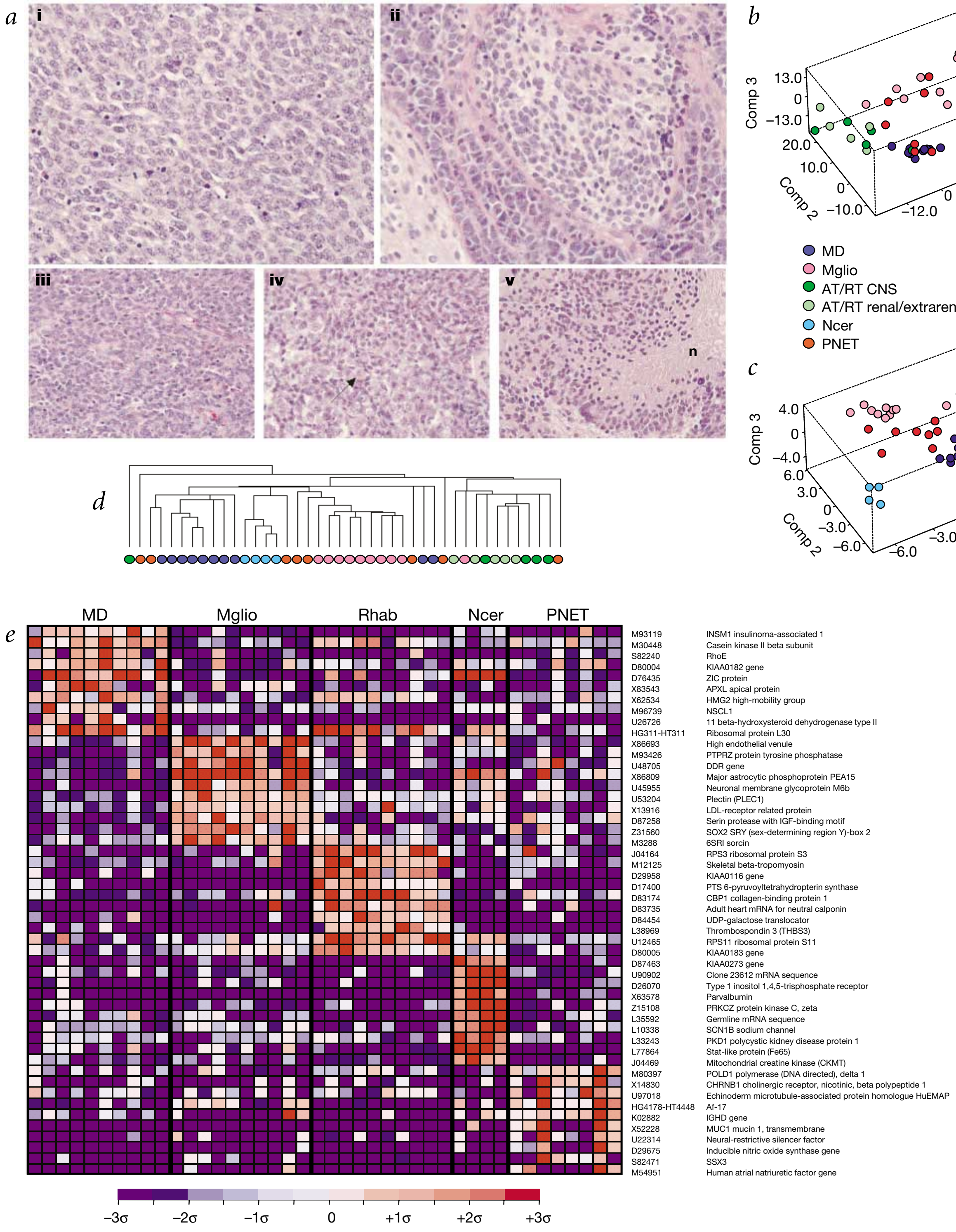

Fig. 2 Classification of histologically similar CNS tumors. ***a***, Representative photomicrographs of embryonal and non-embryonal tumors. ***i***, Classic medulloblastoma; ***ii***, desmoplastic medulloblastoma; ***iii***, supratentorial primitive neuroectodermal tumor (PNET); ***iv***, atypical teratoid/rhabdoid tumor (AT/RT; arrow indicates rhabdoid cell morphology); and ***v***, glioblastoma with pseudopalisading necrosis (n). Original magnification, ×400. ***b***, Principal component analysis (PCA) of tumor samples using all genes exhibiting variation across the data set. The axes represent the three linear combinations of genes that account for most of the variance in the original data set. MD, medulloblastoma; Mglio, malignant glioma; Ncer, normal cerebella. ***c***, PCA using 50 genes selected by signal-to-noise metric to be most highly associated with each tumor type (the top 10 for each tumor are listed in e). ***d***, Clustering of tumor samples by hierarchical clustering using all genes exhibiting variation across the data set. ***e***, Signal-to-noise rankings of genes comparing each tumor type to all other types combined. For each gene, red indicates a high level of expression relative to the mean; blue indicates a low level of expression relative to the mean. Rhab, rhabdoid. The standard deviation (*σ*) from the mean is indicated. Figure reproduced, with permission, from ref. 8.

transition), and the progenitor of squamous carcinomas showing basal epithelial features.

Adenocarcinomas also showed a distinct profile and are potentially derived from yet another unique progenitor cell. But Garber *et al.*[13] identified significant diversity in expression patterns in adenocarcinomas and defined three subtypes that predicted favorable, intermediate and poor outcomes. Two of these three adenocarcinoma subtypes were observed by Bhattacharjee *et al.*[14], and Beer *et al.*[12] also identified prognostic groups within adenocarcinomas using survival as supervision.

Studies on renal cell carcinomas[18] and prostate tumors[19,22] have also identified patterns of gene expression with prognostic importance, and it is likely that other types of tumor not profiled as yet will yield subtypes of clinical importance.

Portraits of development

Tumors are thought to arise at specific stages of cellular development, and their phenotypic properties are thought to be dictated by the stage of developmental arrest. Microarray analysis has provided insight into the developmental stages of many hematological malignancies. Staudt and colleagues[28,29] generated a molecular portrait of the stages of development present in diffuse large B-cell lymphomas (DLBCLs). From the gene expression patterns of hundreds of DLBCL samples, they identified three subtypes: a germinal center B cell–like (GCBL) subtype, an activated B cell–like (ABL) subtype, and a third subtype (T3) lacking high expression of either the GCBL- or the ABL-defining genes[28].

The GCBL subtype showed high expression of genes characteristic of normal germinal center B cells and was associated with relatively good outcomes. By contrast, the ABL subtype expressed genes characteristic of activated peripheral blood B cells (a later stage in B-cell development) and showed a worse clinical outcome. Using survival as supervision, Staudt and colleagues[28] identified 16 genes that could be used for accurate predictions of outcome. This predictor was shown to be an independent prognostic indicator in a multivariate analysis when compared with the international prognostic index—the standard clinical prognosticator.

Shipp *et al.*[27] carried out a similar study of DLBCLs using Affymetrix microarrays, from which they developed a 13-gene predictor using outcome as a supervising parameter[27]. Individuals with DLBCL were divided into two groups: those with cured disease (n = 32), and those with fatal or refractory disease (n = 26). Kaplan–Meier survival curves and log-rank tests indicated that individuals in the 'cured' group had significantly better long-term survival than those in the 'fatal/refractory' group (5-year overall survival, 70% versus 12%). Shipp *et al.*[27] also tested their cohort with the predictive gene set identified by Staudt and colleagues[29] and could reproduce the classifications according to stage of development; however, the GCBL and ABL subtypes did not differ in clinical outcome in their cohort. There are several possible explanations for this difference, including the use of different cohorts and partially overlapping gene sets. As with all microarray studies so far, the true significance of these classifications will only be shown through testing several large cohorts of homogenously treated individuals using the same gene sets and methodologies.

To understand the gene expression changes resulting from specific genetic abnormalities involving transcription factors such as HOX11, TAL1 and LYL1, Ferrando *et al.*[26] analyzed 59 cases of T-cell acute lymphoblastic leukemia (ALL). They found that the oncogenic transcription factors were often expressed aberrantly even in the absence of chromosomal abnormalities. Genes that correlated with the expression of HOX11, TAL1 and LYL1 were selected using a nearest-neighbor analysis with each subtype showing gene expression features indicative of arrest at specific stages of thymocyte development. The LYL1 subtype represented a pro-T cell–arrested signature, the HOX11 subtype represented an early cortical thymocyte signature, and the TAL1 subtype represented a late cortical thymocyte signature. These subtypes predicted survival and showed, as for DLBCLs, that the stage of cellular development is important in determining clinical behavior.

Portraits of metastasis

Through screening programs, individuals with cancer are often identified at an early stage of disease. In general, tumors detected in their early stages have good clinical outcomes; however, many individuals deviate from this course and have recurrences. Unfortunately, there are no markers that can predict accurately when or if a tumor will recur, which leaves physicians with the difficulty of choosing the best treatment options. To assure a positive outcome, therefore, individuals with early stage cancer often receive intensive treatment with toxic therapies that have their own associated morbidity and mortality.

To address this common issue, van 't Veer *et al.*[3] used microarrays to analyze a cohort of individuals with breast cancer who presented with tumors smaller than 5 cm, had no lymph node metastases and were aged under 55 years. These individuals were treated by modified radical mastectomy or breast-conserving therapy with axillary lymph node dissection and radiation therapy. Most individuals did not receive any additional treatment and were followed annually for at least 5 years. This cohort therefore provided a setting in which the natural history of tumors could be followed without the confounding feature of different responses to systemic therapy.

van 't Veer *et al.*[3] profiled the primary tumors and carried out a supervised analysis in which the individuals were separated into two groups: those who developed metastases in less than 5 years (bad prognosis), and those who were metastasis-free for longer than 5 years (good prognosis). They identified 231 genes that correlated with this parameter. Through further statistical analysis, a discriminatory set of 70 genes was identified that showed 81% accuracy when tested in a leave-one-out cross-validation analysis on the training set, and 89% accuracy on a test set. This group of markers was found to be a statistically significant predictor in a multivariate analysis, showing that a test based on gene expression can add value to the current list of clinical tests.

MacDonald *et al.*[7] identified a set of genes that could differentiate between medulloblastomas with and without metastases[7]. This predictor was validated by a leave-one-out cross-validation approach and correctly predicted the metastasis status of 72% of their tumors. The metastasis-associated portrait strongly implicated the platelet-derived growth factor receptor (PDGFR) and the RAS/MAPK pathway. Thus, interventions directed against these pathways may be effective against metastasis-positive medulloblastomas.

An important ramification of these two studies is that gene expression features present in primary tumors at the time of presentation can predict the course of the disease[41]. This implies that properties such as the propensity to metastasize or treatment response might be diagnosed from analysis of the primary tumor and used as the basis for clinical decision-making. For example, if a individual presents with a medium-sized tumor with a 'good prognosis' signature, the best treatment may be surgical resection of the primary tumor with no chemotherapy but with close follow-up of the individual. Conversely, the best treatment for a small tumor (1–2 cm) with a 'poor prognosis' signature may be combined treatment modalities, including surgery, radiation therapy and chemotherapy.

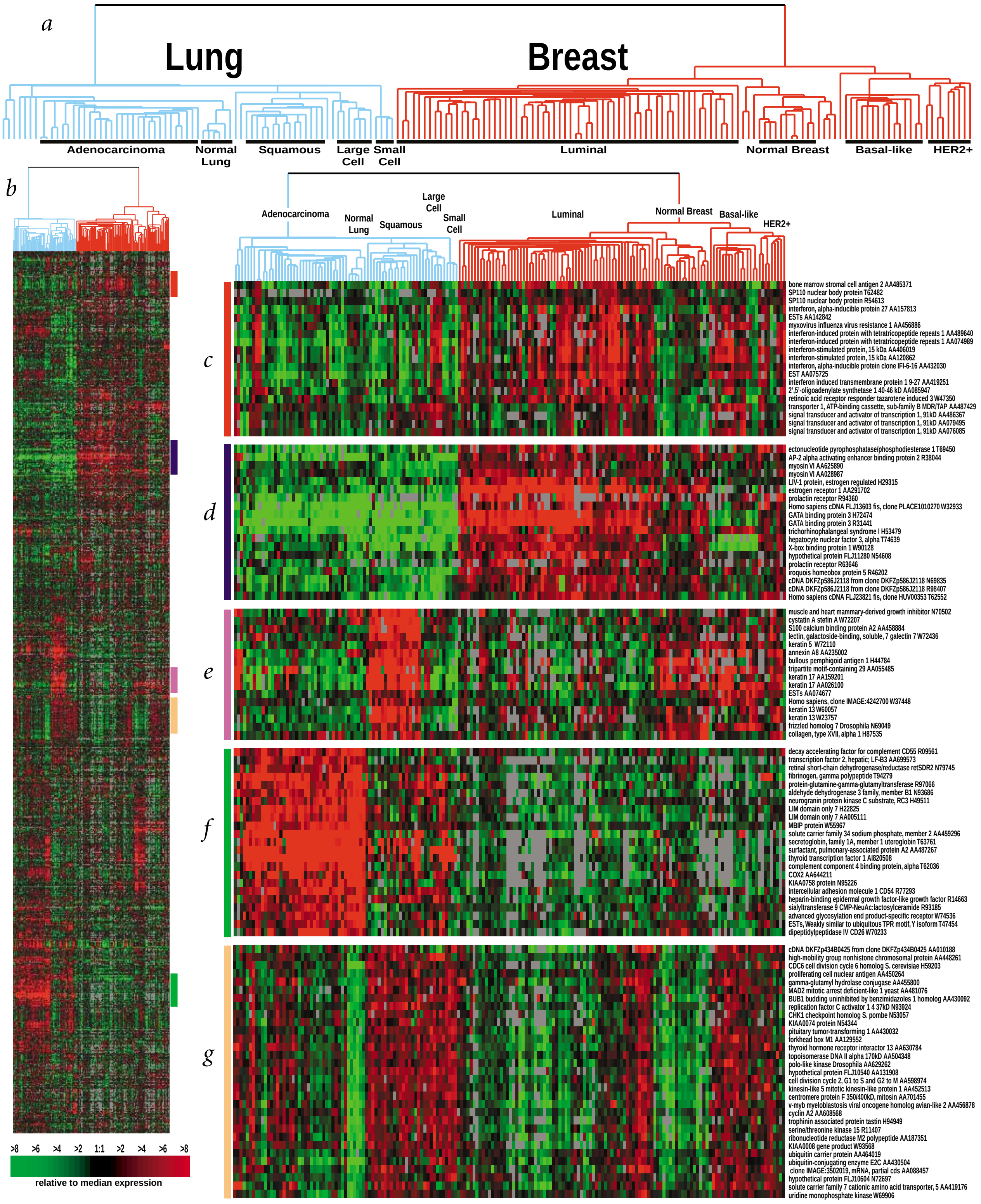

Fig. 3 Combined hierarchical cluster analysis of breast and lung carcinomas. A gene list comprising the nonredundant combination of 'intrinsic' gene sets[1,13] was used in a hierarchical clustering analysis of the publicly available breast and lung carcinoma data sets taken from the Stanford Microarray Database. ***a***, Experimental sample-associated cluster dendrogram. ***b***, Complete cluster diagram. ***c***, Interferon-regulated gene set. ***d***, Breast luminal cell profile. ***e***, Basal epithelial cell profile. ***f***, Lung adenocarcinoma-enriched profile. ***g***, Proliferation gene set. The complete breast and lung cluster diagram is shown in Web Fig. A online.

Portraits of therapeutic response

Another problem for clinicians is predicting who will respond to therapy. Most clinical markers are 'prognostic' (that is, they predict outcome), but markers that are 'predictive' of therapeutic response are more useful. Only a few predictive markers are used routinely in cancer medicine. In breast cancer, for example, the presence of the ER predicts response to tamoxifen[42]. The ER-positive breast tumor subtype has a distinct microarray expression profile (Fig. 3*d*)[1–4,6].

Microarrays have also been used to identify gene expression patterns that predict response to therapy in Philadelphia chromosome–positive ALL (Ph+ALL)[25]. Individuals with Ph+ALL have been shown to respond to STI571 (Gleevec), a tyrosine kinase inhibitor of ABL1 that is effective in treating chronic myeloid leukemia (CML)[43]. Although the response to STI571 is much lower in Ph+ALL than in CML (59% versus 95–100%), Hofmann *et al.*[25] profiled 25 bone marrow samples from 19 individuals with Ph+ALL and identified 95 genes that could predict response to STI571 before treatment. STI571 has now been shown to be a more general inhibitor of tyrosine kinases and may be effective in tumors expressing PDGFR (gliomas and prostate) and c-KIT (GISTs)[44]. Not surprisingly, c-KIT is part of the GIST expression portrait[39].

These predictive markers will enable tumors to be treated on the basis of the presence of biological pathways that contribute to malignant transformation. Certainly, many future challenges lie ahead as markers that predict resistance (rather than response) are identified.

Portraits of mutation

Inherited germline mutations in some genes predispose individuals to developing tumors in specific tissues. For example, even though the *BRCA1* protein is expressed widely, women with a germline *BRCA1* mutation most frequently develop breast and/or ovarian carcinomas. Although we do not understand the molecular basis of this tissue-specific tumorigenesis, tumors that arise from these predisposing mutations may have unique molecular portraits. Indeed, germline carriers of *BRCA1* mutations develop breast or ovarian tumors whose gene expression patterns can be differentiated from those of most sporadic tumors[3,5,9]. In some cases, sporadic breast tumors with a *BRCA1* mutant expression profile were identified and later shown to contain a methylated *BRCA1* promoter[5]. An interesting ethical point is that an individual could be identified to have a germline *BRCA1* mutation through an expression-based test, which would have important genetic implications for that individual and her offspring.

Common mechanisms in the development of ALL are chromosomal translocations or intrachromosomal rearrangements. ALL can be divided into six subtypes, B-cell, pre-B, early pre-B, T-cell and acute mixed-lineage (AML) leukemia, using lineage-specific features of lymphoblasts. This group of diseases is frequently classified on the basis of underlying chromosomal abnormalities and the clinical significance that is associated with these abnormalities. In a cohort of 327 bone marrow samples analyzed by Yeoh *et al.*[24], the six main ALL subtypes with prognostic significance were identified with an accuracy of 96%. In at least four samples, the expression-based classifications were more accurate than were current molecular methods based on polymerase chain reaction with reverse transcription (RT–PCR). Unlike in DLBCLs, the identified profiles in ALL do not represent a specific stage of differentiation, but instead reflect the presence of specific genetic abnormalities. Notably, Yeoh *et al.*[24] also identified a distinct TEL-AML1 subtype, which was distinguished by 20 genes that could predict, with 100% accuracy, those individuals who would progress to develop a secondary AML. These findings imply that some individuals with ALL have a genetic predisposition to develop secondary AML.

Multiclass tumor prediction

Many of the main branches of the cancer family tree represent functionally distinct types of cells and, as expected, each specialized type of cell expresses a unique set of genes needed for its function[31]. These dominant patterns of cell type expression are evident in any microarray study in which samples derived from different tissues are compared directly[34,45,46], or in studies of histologically complex tumor samples analyzed using unsupervised methods[1,21,29,38]. Several microarray studies have been designed to define tumor- and tissue-specific portraits[34,45,46], in which data sets containing between 10 and 19 different tissues have been analyzed using the site of sample origin as the supervising parameter.

Ramaswamy *et al.*[34] analyzed 144 primary tumors representing 14 common tumor types. They carried out a step-by-step analysis to define a set of genes that identified each of the 14 tissues separately. In essence, a portrait of each tumor type was painted and represented by a specific set of genes. When a new sample was analyzed, its portrait was compared with each of the 14 portraits, and it was assigned to the class it most resembled. In a test set of 54 primary tumor samples, 6 of 8 metastasis samples were classified correctly by this method (78% accuracy across the 54 samples). Notably, many of the samples that were not correctly classified were histologically described as poorly differentiated adenocarcinomas (6 of 20 correctly classified). Ramaswamy *et al.*[34] interpreted this result to suggest that poorly differentiated adenocarcinomas may represent a distinct class rather than being related to a differentiated cell–type class that simply lacks expression of a few genes.

Su *et al.*[45] also developed a multiclass prediction for 10 tumor types that showed 85% accuracy on test set predictions and 75% accuracy on metastasis sample predictions. As two different laboratories have shown high accuracy in predicting the site of origin of metastatic samples, multiclass predictions may have practical applications in classifying tumors of unknown origin. If the data of Ramaswamy *et al.*[34] are correct, however, this potential clinical test may need thousands of genes to perform well.

Common patterns of gene expression

As many different tumors have now been profiled, we can examine whether any patterns of expression are common among tumors derived from different tissues. To illustrate some of the most common patterns, we present a combined hierarchical clustering analysis of breast[1,2] and lung[13] tumor data sets taken from the public domain (Fig. 3 and Web Fig. A online) and analyzed using a combined breast and lung 'intrinsic' gene set and hierarchical clustering analysis[1]. The most striking and most common gene expression pattern is the 'proliferation' cluster[47]. This gene set contains genes involved in regulating the cell cycle and genes that encode structural protein components required for DNA replication and chromosome dynamics (Fig. 3*g*). The proliferation signature is correlated with cellular growth rates *in vitro*[31] and has been identified in breast[1], lung[13,14], ovary[10], prostate[20], liver[38], gastric tumors[23], gliomas[48] and lymphomas[28]. The identification of this cluster *in vivo* probably represents how rapidly a given tumor is growing and, as might be expected, the high expression of these genes has been found to indicate a poor prognosis. In addition, studies of synchronized cultures of HeLa cells have shown that the tumor-defined 'proliferation' genes are regulated by the cell cycle[49]. Somatic mutations of the *TP53* gene are also correlated with high expression of the proliferation signature[2,38].

The second most common pattern is an interferon-responsive gene set (Fig. 3*c*)[1,13,29,31,38,47]. Although the biological significance of this set of genes is not understood in terms of its effects on tumor formation or growth, this set represents an example of a 'functional cluster' in which a transcription factor (STAT1) and its target genes are present in the same expression cluster.

A third feature of our analysis is that the breast basal-like tumors (and some HER2-positive breast tumors) share a similar pattern of expression with lung squamous carcinomas (Fig. 3*e*). This shared pattern is characterized by the high expression of genes that are normally enriched in basal epithelial cells—a gene set that has been identified independently in the two tumor data sets[1,13]. This integrated analysis suggests that breast basal-like tumors share significant characteristics with lung squamous carcinomas. Both types of tumor are also highly proliferative (Fig. 3*e*,*g*).

Clinical application of genomic analyses

Emerging technological advancements in genomics are allowing us to extract larger amounts of information from smaller and smaller samplings. In the future, single-cell analyses may allow us to study cancer heterogeneity and the effects of complex cellular interactions precisely[50]. As genomic and proteomic tools continue to develop, it will be essential to integrate information from several analyses into a common framework to develop a more complete understanding of cancer biology.

Although the application of microarrays to tumor biology is only beginning, the studies that have already classified many types of tumor are ready for the next step; that is, to use a molecular taxonomy in clinical medicine to improve cancer treatment. But many challenges, not all scientific, need to be met before this step is realized. For example, what type of genomic assays will be offered and on what platforms? What are the legal issues surrounding 'multianalyte'-based tests? Will patent laws make such clinical assays cost prohibitive?

In many ways, the immediate step is to validate the first step by determining whether the same clinically significant molecular classifications can be observed using different cohorts and different genomic platforms. Next, no matter what markers are found and what platform is used, the importance of the molecular taxonomy will have to be shown in large-scale clinical trials. This is a necessary step before molecular classifications can be used to make medical decisions about giving or withholding therapy. Because there is no substitute biomarker for overall survival, the hard truth about the validation of these genomic classifications is that it may take long-term prospective trials to determine whether a new molecular taxonomy is equal to, or better than, the standard methods. Although there is no substitute for randomizing individuals to homogenous treatment arms in a prospective study, genomic assays that can be done on archived tissue samples from completed clinical trials could hasten the process of validation because patient materials and outcomes are already available.

Ironically, microarrays have been very valuable in identifying genes with importance to cancer medicine but may not be practical for the clinical setting. In routine surgical pathology, specimens are formalin-fixed and paraffin-embedded to preserve tissue histology, which makes the recovery of intact RNA for microarray analysis difficult. In addition, microarrays were developed for the large-scale scanning of genomes to find genes that are quantitatively altered in biological systems. But once subsets of genes are identified as clinically significant, microarrays may be superfluous for routine testing. As discussed above, 'supervised' analyses can effectively reduce the number of genes necessary for class discrimination from thousands to between six and seventy (these studies have also shown that 1 or 2 genes do not perform well at classification)[8,19,24,28,37].This facilitates the use of other techniques such as real-time quantitative PCR, which conserves samples and is faster, easier and less expensive than DNA microarrays. We think that the best results from tumor profiling studies are yet to come. It is uncertain whether microarrays will be used for routine clinical diagnostics, but microarrays will undoubtedly continue to be used in research to find molecular portraits of clinical significance.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank J. Parker for comments and apologize to the researchers whose work was not discussed owing to length considerations. This work was supported by grants from the National Cancer Institute and the National Institute of Environmental Health Sciences.

1. Perou, C.M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
2. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
3. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
4. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA* **98**, 11462–11467 (2001).
5. Hedenfalk, I. *et al.* Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**, 539–548 (2001).
6. Gruvberger, S. *et al.* Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**, 5979–5984 (2001).
7. MacDonald, T.J. *et al.* Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nature Genet.* **29**, 143–152 (2001).
8. Pomeroy, S.L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
9. Jazaeri, A.A. *et al.* Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. *J. Natl Cancer Inst.* **94**, 990–1000 (2002).
10. Welsh, J.B. *et al.* Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl Acad. Sci. USA* **98**, 1176–1181 (2001).
11. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
12. Beer, D.G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.* **8**, 816–824 (2002).
13. Garber, M.E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA* **98**, 13784–13789 (2001).
14. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA* **98**, 13790–13795 (2001).
15. Zou, T.T. *et al.* Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene* **21**, 4855–4862 (2002).
16. Lin, Y.M. *et al.* Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. *Oncogene* **21**, 4120–4128 (2002).
17. 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).
18. Takahashi, M. *et al.* Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. *Proc. Natl Acad. Sci. USA* **98**, 9754–9759 (2001).
19. Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209 (2002).
20. LaTulippe, E. *et al.* Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res.* **62**, 4499–4506 (2002).
21. Welsh, J.B. *et al.* Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.* **61**, 5974–5978 (2001).
22. Dhanasekaran, S.M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).
23. Hippo, Y. *et al.* Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res.* **62**, 233–240 (2002).
24. Yeoh, E.J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143 (2002).
25. Hofmann, W.K. *et al.* Relation between resistance of Philadelphia-chromosome-positive acute lymphoblastic leukaemia to the tyrosine kinase inhibitor STI571 and gene-expression profiles: a gene-expression study. *Lancet* **359**, 481–486 (2002).
26. Ferrando, A.A. *et al.* Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**, 75–87 (2002).
27. Shipp, M.A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Med.* **8**, 68–74 (2002).
28. Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1947 (2002).

29. Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
30. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
31. Ross, D.T. *et al*. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
32. Slonim, D.K. From patterns to pathways: gene expression data analysis comes of age. *Nature Genet.* **32**, 502–508 (2002).
33. Churchill, G.A. Fundamentals of experimental design for cDNA microarrays. *Nature Genet.* **32**, 490–495 (2002).
34. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
35. 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).
36. Tusher, V., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
37. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 6567–6572 (2002).
38. Chen, X. *et al.* Gene expression patterns in human liver cancers. *Mol. Biol. Cell* **13**, 1929–1939 (2002).
39. Nielsen, T.O. *et al.* Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* **359**, 1301–1307 (2002).
40. Wigle, D.A. *et al.* Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.* **62**, 3005–3008 (2002).
41. Bernards, R. & Weinberg, R.A. A progression puzzle. *Nature* **418**, 823 (2002).
42. Fisher, B. *et al.* A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. *N. Engl. J. Med.* **320**, 479–484 (1989).
43. Druker, B.J. *et al.* Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med.* **344**, 1038–1042 (2001).
44. Dematteo, R.P., Heinrich, M.C., El-Rifai, W.M. & Demetri, G. Clinical management of gastrointestinal stromal tumors: before and after STI-571. *Hum. Pathol.* **33**, 466–477 (2002).
45. Su, A.I. *et al.* Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* **61**, 7388–7393 (2001).
46. Hsiao, L.L. *et al.* A compendium of gene expression in normal human tissues. *Physiol. Genom.* **7**, 97–104 (2001).
47. 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).
48. Rickman, D.S. *et al.* Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res.* **61**, 6885–6891 (2001).
49. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
50. Levsky, J.M., Shenoy, S.M., Pezo, R.C. & Singer, R.H. Single-cell gene expression profiling. *Science* **297**, 836–840 (2002).