Opinion

Microarrays and molecular markers for tumor classification

Brian Z Ring and Douglas T Ross

Address: Applied Genomics Inc., 525 Del Rey Ave #B, Sunnyvale, CA 94085, USA.

Correspondence: Douglas T Ross. E-mail dross@applied-genomics.com

Published: 29 April 2002

Genome ***Biology*** 2002, **3(5)**:comment2005.1–2005.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/5/comment/2005

Abstract

Human cancers have traditionally been classified according to their tissue of origin, histological characteristics and, to some extent, molecular markers. Clinical studies have associated different tumor classes with differences in prognosis and in response to therapy. Measurement of the expression of thousands of genes in hundreds of cancer specimens has begun to reveal novel molecularly defined subclasses of tumor; some of these classes appear to predict clinical behavior, while others may define tumor types that are ripe for directed development of therapeutics. Unfortunately, at present, differences between studies of similar tumor types can be as striking as their similarities.

The diagnosis and classification of human cancers currently relies upon microscopic examination of tissue supplemented by findings from ancillary studies such as immunohistochemistry. Pathologic classification combined with clinical information can be used to differentiate distinct tumor classes that differ in both prognosis and response to therapy. The diversity of clinical behaviors exhibited by tumors within recognized classes suggests, however, that biologically distinct subtypes remain to be identified. Progress in the systematic identification and validation of novel clinical classes has been hampered both by the technical limitation of the number of markers that can be examined efficiently, and by the difficulty in comparing numerous small-scale studies that use different reagents and different sample sets.

The recent completion of the first draft of the human genome sequence has raised hopes that a more accurate classification of human neoplasia will emerge that relies on a better characterization of the patterns of mutation and expression of genes in tumors. The most striking progress has been made using microarray technology, which measures gene expression for tens of thousands of genes in simple overnight experiments. The expectation is that this genomic-scale measurement of gene expression in thousands of clinical specimens will reveal a detailed molecular classification of malignant tumors and will allow more reliable prediction of clinical behavior, better stratification of patients, and the development of novel therapeutics targeted to the distinguishing characteristics of different tumor classes.

It should be noted that microarrays are not the optimal technology for the measurement of expression of individual genes; rather, their utility is in the identification of patterns of coordinately expressed genes. Although any single measurement out of the tens of thousands of measurements on a single hybridization array may be problematic, the patterns of gene expression represented by large sets of genes have proven highly reproducible when compared between many related samples. The power of the microarray tool-kit thus lies in the identification and interpretation of patterns. The danger is that artifacts can be systematic, and the interpretation of patterns can be fraught with error. In this article, we briefly review the methods used to obtain and analyze tumor microarray data, and the types of conclusions that can be drawn, as well as considering in more detail the insights from several recent studies of breast tumors and lymphomas.

The search for meaning

The primary goals of large-scale gene-expression studies include, firstly, discovering the common patterns of variation of genes across measured experimental samples and, secondly,

extrapolating from the particular genes that comprise these patterns to understand function or to identify potential therapeutic targets. In order to study novel and clinically relevant features of malignancies, analyses of microarray data have therefore largely focused on two broad analytical goals. The first is the discovery of novel biologically significant features; this requires the correlation of patterns of gene expression with various biological characteristics of clinical samples. The second is the development of clinical prognostic tools, which requires the identification of 'predictor' genes - a pattern of gene expression that predicts clinical outcome - and the verification of their utility in independent patient groups.

To a large degree, biochemical pathways, responses to environmental stimuli, and other variations in physiology are governed by the coordinated regulation of large sets of genes. Cluster analysis, which identifies genes that have similar expression patterns, allows the dominant gene-expression patterns in a dataset to drive the separation of clinical samples into groups on the basis of overall similarity in expression pattern, without allowing experimenter-bias to influence the outcome. One of the most widely used clustering algorithms relies on agglomerative hierarchical clustering, which involves the determination of the pair-wise distance measurements between all genes in a set of experiments, and subsequent agglomeration of clustered pairs into larger clusters, again on the basis of distance [1]. Patterns can be visualized as dendrograms (hierarchical tree diagrams) that depict relationships between genes and samples, and as pseudo-color tables that allow exploration of the underlying data (see Figure 1). It is worth noting that similar methods are used when analyzing data acquired using either of the two most popular microarray platforms that are in use at present: 'home-made' DNA arrays of the type pioneered at Stanford University, and oligonucleotide arrays of the type manufactured by Affymetrix Inc.

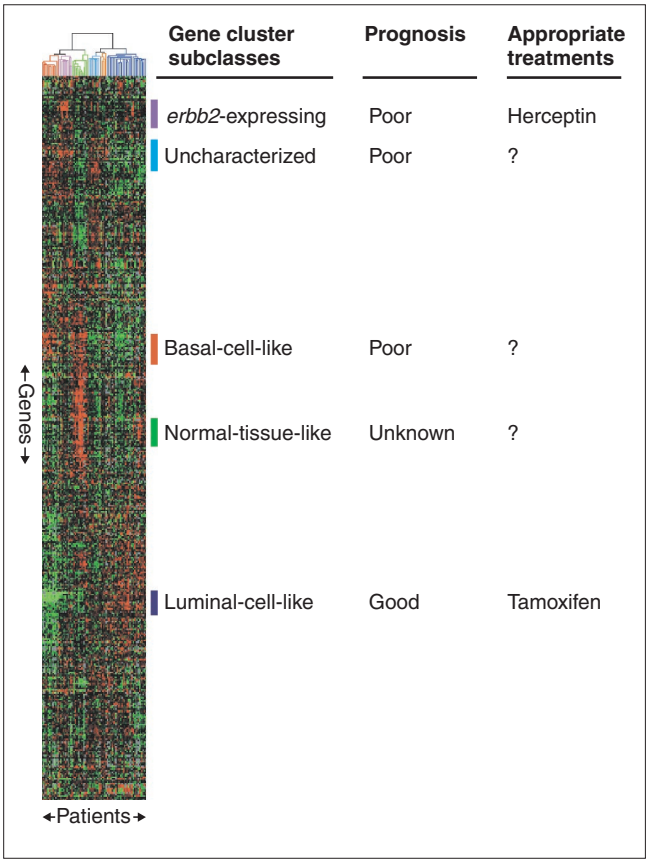

Figure 1
Gene-expression patterns of 85 different breast cancer specimens for the 456-gene 'intrinsic gene list' identified by Sorlie *et al.* [17], depicted as a pseudo-color hierarchical cluster-diagram. Highlighted areas depict sets of genes whose expression has been inferred to distinguish classes of breast cancer as determined by cluster analysis. The luminal class and *erbb2* class are candidates for treatment with tamoxifen and herceptin, respectively. Other identified classes may be useful for the identification of novel markers of prognosis and for the identification of targets for rational drug design. Figure adapted from [17].

The discovery of novel tumor classes by cluster analysis has the problem that the relationships revealed are highly dependent upon the set of genes chosen for analysis. 'Agnostic' approaches are most often used: these select gene sets on the basis of data quality and/or select for unbiased features of datasets. Although these methods are useful for distinguishing novel tumor classes, they also tend to be dominated by shared characteristics of the samples distinguished by large sets of genes. The result is that physiological attributes of specimens that could in theory be distinguished by variations in the expression of a small set of genes (for example, drug resistance) can in fact be lost in clustering patterns that are dominated by more pervasive aspects of the tumor's biology (such as differentiated versus undifferentiated cell types). Because medical genomic research has the more specific goal of associating gene-expression patterns with clinically relevant knowledge about patient samples, such as tissue of origin, tumor grade, treatment susceptibility, or eventual outcome, a number of supervised techniques for analyzing gene-expression data have been proposed, including weighted voting [2], support vector machines (SVM) [3], and statistical analysis of microarrays (SAM) [4]. These methods differ in underlying algorithm, but all aim to identify genes that are significantly associated in their pattern of expression with measured attributes of the samples (such as outcome data, pathological classification, or proliferative index). One current problem with these techniques is that the present lack of standardized public experimental datasets means there is a lack of 'training' data with which to guide the methods. Most reported results therefore lack testing on independent data sets, but they nevertheless serve as important initial steps towards the identification of candidate markers of tumor class, for subsequent verification in larger independent studies.

Tumor analysis

Human cancer is currently classified according to a standard set of clinical parameters and a limited array of molecular and immunological markers. Here, we consider in more detail how microarray data analysis is contributing to comparisons of normal tissues with tumor cells, analyses of tissue of origin, and, in particular, to the discrimination of novel classes of tumor and their prognostic significance.

Comparing tumors and normal tissues

There was great initial interest in the use of gene-expression microarrays to identify gene-expression differences in a simple comparison between tumor and normal cells. Although it might seem logical that this would be an efficient means of determining tumor-specific markers, perhaps related to oncogenesis, the analysis has in fact proven to be more complex. Large-scale gene-expression studies have demonstrated that dominant gene-expression patterns in tumors in fact reflect the overall biology of the samples and reveal little about the critical differences that contribute to the oncogenic physiology to the cells. Tumors and normal tissues differ dramatically in the proportion of different types of tissue present (for example, stroma, lymphocytic, or epithelial tissue), in the overall fraction of cells that are actively dividing, and in the relationships of tissue components to one another; a simple comparison of tumor versus normal largely yields gene-expression differences that can be interpreted in the context of these self-evident tissue differences. For example, the variation in the expression of a set of genes enriched for those regulated in relation to the cell cycle, such as genes that encode DNA repair proteins, cyclins, and PCNA (proliferating cell nuclear antigen) , often reflects the overall differences in proliferative index of the samples [5,6]. Similar sets of genes have been identified as being differentially expressed when comparing many different types of sample that vary in overall proliferation rate, and these differences have often been misinterpreted as reflecting functional disregulation of sets of genes that have previously been implicated in oncogenic processes. Although this may be an interesting result in and of itself, it does not further the goal of identifying relevant tumor markers, given that there are already adequate immunohistochemical indicators of proliferation.

Identifying the tissue of origin

Although the clinical finding of a cancer of unknown origin is a relatively rare occurrence, appropriate treatment does depend on ascertaining the tissue of origin of a patient's tumor. It is therefore obviously useful to identify markers that could aid in determining the origin of tumors. Two recent studies [7,8] describe surveying a panel of tumors of disparate tissue origins, using Affymetrix oligonucleotide arrays, with the goal of identifying genes characteristic of each tumor type. Each of the studies used support vector machines on an initial training set of tumor samples to determine the genes that best discriminated tumor classes from one another, and then evaluated the derived predictors on a separate test set of tumors. Su *et al.* [7] correctly identified the tissue of origin of 87% of a test set of primary tumors, and were also able to classify 75% of a test set of metastatic carcinomas. Ramaswamy *et al.* [8] reported a 78% success rate with their test set of tumors but interestingly reported a significantly lower success rate (30%) with tumors previously classified as poorly differentiated (for which the tissue-of-origin features may be less prominent). Both studies reported that the predominant classifier genes are expressed in a tissue-specific manner in the tissue from which the tumors were presumed to be derived. Furthermore, each study demonstrated that tumor metastases were in general classified correctly by microarray analysis, demonstrating that tissue-of-origin identity of a malignancy is preserved through the process of metastatic development. A significant difference between the two studies was the number of genes found to be optimal for distinguishing the tumor types. Su *et al.* [7] found that 10 genes for each tumor type most accurately predicted the origin of the tumor samples, while Ramaswamy *et al.* [8] found the greatest classification accuracy when the predictor used all the tested (> 16,000) probes, with a significant decrease in predictive powers as the number of genes used fell below 50 per type. It is difficult to determine whether this difference is due to the more iterative statistical analysis of Su *et al.* [7], differences in the tumor samples examined, or technical differences in the datasets. Nevertheless, the results from the two studies are very encouraging for the clinical objective of being able to use gene-expression patterns to identify the tissue-origins of tumors that are identified initially as metastases.

Tumor subclassification

The most compelling impact of gene-expression profiling on the care of cancer patients comes from its promise in distinguishing biologically and clinically distinct classes of cancer. The utility of subclassifying cancers lies in the promise that it will be useful for distinguishing which patients could best benefit from particular patient-care algorithms, or that it can serve as the basis for development of novel therapies targeted to the genes that distinguish each tumor class. A large number of recent studies have aimed to identify gene-expression patterns that predict the clinical outcome of different cancer types [9-20]. We focus here on recent studies of breast carcinomas and lymphomas, as these have been the subject of comparable studies using similar samples but different array and analysis technologies.

The current methods used to assess prognosis for individual breast cancer patients are primarily based on clinical parameters such as the size of tumor at diagnosis and the presence or absence of local and distant metastases. In addition, estrogen-receptor expression and over-expression or amplification of the *erbb2* gene is routinely assessed in order to stratify patients into those who might benefit from therapy with tamoxifen (estrogen-receptor antagonist) or

Herceptin© (Trastuzumab; anti-Erbb2 antibody), respectively. There is still marked variability in patients' clinical course, however: current diagnostic and stratification algorithms clearly do not adequately detect the biological and clinical heterogeneity of breast cancer. Given that breast cancer is now being detected much earlier in its natural history, as a result of better radiographic detection and clinical surveillance, great potential benefit could come from improved markers of tumor prognosis.

A study last year by Sorlie *et al.* [17] used cDNA microarrays and hierarchical clustering analysis to explore whether breast-cancer patient samples could be grouped on the basis of distinguishing patterns of gene expression. In order to objectively identify a set of genes useful for the classification of patients, they exploited the unusual opportunity afforded by a clinical study in which two independent tumor biopsy specimens were sampled from each patient at two time points separated by sixteen weeks of chemotherapy. They selected the subset of genes that varied least between the two independent samplings of each patient tumor, working on the logic that those genes that remained relatively consistent over time would be involved in 'intrinsic properties' of each tumor, as opposed to other variables such as sampling error. Hierarchical clustering analysis of 78 tumors using this 'intrinsic gene list' revealed a dendrogram of branching patterns that distinguished five, and possibly six, classes of tumor (see Figure 1). Two of the classes were distinguished by gene sets that included genes already known to be useful for identifying clinically significant classes of breast tumors, namely the estrogen-receptor-expressing and *erbb2*-overexpressing classes. In addition, classes of patients could be identified whose tumors expressed gene sets that did not correlate with any commonly recognized tumor-class distinctions. For example, a distinct subset of the tumors expressed genes reminiscent of basal epithelial cells of the normal breast's lactation ducts; this feature of some breast cancers had been recognized previously on the basis of cytokeratin staining patterns [21] but had not generally been thought to be clinically significant. When patients were grouped according to which terminal dendrogram branch their tumor fell on, statistically significant differences in clinical outcome were apparent between classes. As expected, patients classified into the *erbb2* subclass had poor overall survival, whereas patients classified in the 'luminal' estrogen-receptor-expressing subclass had much better overall survival. In addition, those patients whose tumors expressed features reminiscent of breast basal epithelial cells had long-term survival rates as poor as those classified in the *erbb2* class; a subgroup of patients within the luminal class also had relatively poor overall survival. This study therefore confirmed that gene-expression-based classification of tumors by cluster analysis can not only identify patients who are most likely to benefit from current treatments but also identify tumor types with poor prognosis that are not specifically targeted by current therapies. The set of genes that distinguished these novel classes are candidates not only for biomarkers of the stratification of patients but also for providing targets for the development of novel therapeutics.

Gene-expression analysis of patient cohorts for whom long-term clinical follow-up is available provides an opportunity to search directly for correlates with outcome, independent of tumor classification. The study by Sorlie *et al.* [17] thus used supervised analysis to identify genes associated with outcome amongst all patient samples, regardless of the tumor subclass identified by cluster analysis. They found that 264 cDNA spots on the array were significantly associated with patient survival. In a separate study on a similar patient cohort, van't Veer *et al.* [20] used oligonucleotide arrays to compare gene-expression patterns of breast cancers from 44 patients who were free from metastases five years after initial treatment with patterns from tumors from 34 patients who developed metastases within five years of treatment. Of a set of approximately 5,000 genes (of the 25,000 measured) that varied significantly in expression level across their sample set, van't Veer *et al.* [20] identified 231 genes associated with metastasis; they then optimized this gene set to identify 70 genes that best predicted poor clinical outcome. Expression levels of this set of genes were then used to classify an independent set of patients according to predicted outcome, and this correctly classified seventeen of nineteen patients. Our own (unpublished) comparison of the two studies [17,20] revealed only fifteen genes in common between the two lists of prognosticator genes. The difference between the gene lists may in part be due to the different clinical endpoints measured (survival versus metastasis), or to technical differences in experimentation, including the gene sets analyzed. Both groups did identify genes involved in transit through the cell cycle as common amongst those that conveyed a poor prognosis, however. In the cohort studied by van't Veer *et al.* [20], the classification of patients according to their gene-expression patterns was superior, for identifying poor outcomes, to that found when the patients were classified according to the application of the conventional algorithm that relies on clinical parameters. These results hold great promise for the development of improved markers for the prediction of clinical course, even in patients for whom disease is detected prior to the development of clinical indicators of poor prognosis.

Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin's lymphoma and is treated primarily with conventional chemotherapy. As in breast cancer, clinical prognostic indices are primarily used in DLBCL for the identification of subsets of patients who are likely to have a poor response to treatment. There is still marked heterogeneity of clinical course within prognostic classes, however, so it is likely that there is greater variability in tumor types than is sampled by clinical parameters. In one of the earliest

examples of class discrimination by microarray Alizadeh *et al.* [9] used cDNA arrays to characterize gene-expression profiles in a diverse set of lymphoid malignancies including 43 DLBCLs, as well as numerous samples of both transformed and normal lymphoid cells manipulated in culture. Hierarchical clustering analysis demonstrated that the DLBCL cases could be classified by gene-expression patterns into subclasses that could be interpreted as suggesting that the tumor cells were immortalized at different stages of B-cell maturation. One class expressed genes that have been previously shown to be expressed exclusively in germinal center B cells, for example, while a second-class expressed genes characteristic of activated B cells. A follow-up study demonstrated that the somatic mutation of antibody genes had occurred in all of the cases classified as 'germinal-center-cell-like' and was absent or rare in the 'activated-B-cell-like' cases [22]. Kaplan-Meier survival-curve analysis demonstrated that the two morphologically indistinguishable classes had dramatically different overall survival characteristics; now, the differential expression of hundreds of genes has identified a large set of candidate markers that might discriminate between them.

In a similar recent study, Shipp *et al.* [16] used Affymetrix arrays to measure gene-expression patterns in 58 cases of DLBCL, including 32 patients who were subsequently apparently 'cured' of disease and 26 for whom the disease was fatal or refractory to treatment. Shipp *et al.* [16] used a supervised-learning classification approach to identify sets of genes that could predict outcome, and they found that a set of 13 genes could optimally predict outcome; this set included some genes that had previously been associated with DLBCL outcome. A comparison of the two studies [9,16] reveals rather poor concordance between the identified prognosticators, however. Only three of the best prognosticator genes identified by Shipp *et al.* [16] were also found by Alizadeh *et al.* [9]. The pattern of expression of these genes in the datasets of Alizadeh *et al.* [9] did in fact predict poor outcome for this independent patient cohort. Conversely, the Affymetrix dataset [16] included 90 of the genes implicated in distinguishing germinal-center from activated B cells in the cDNA array dataset of Alizadeh *et al.* [9]. Although a subset of these 90 genes measured by the Affymetrix approach appeared to subdivide the tumors into subsets in a manner similar to the division in the study by Alizadeh *et al.* [9], the classes as distinguished by hierarchical clustering in this experiment did not show statistically different overall survival characteristics. But a large proportion of these 90 genes were measured with low confidence in the second set of experiments [16] (as assessed by the Affymetrix software), raising the possibility that these genes were not measured with adequate precision to be used as classifiers in this experiment. It is possible that much heterogeneity has yet to be discovered in DLBCL and these studies have uncovered only the beginning of candidate biomarkers that may distinguish clinically distinct DLBCL subtypes.

Open questions

The studies discussed above, involving breast cancer and lymphoma, demonstrate the two fundamentally different approaches that are most often used in the identification of genes useful for prognostication and stratification of cancer. Analysis of patient cohorts by hierarchical clustering assumes that biologically distinct tumor types can be identified by their characteristic patterns of expressed genes and that some of these classes will prove clinically distinct. In contrast, the direct search for correlates with clinical outcome assumes that among the heterogeneity within currently identified tumor classes there will be sets of genes that are highly correlated with survival, independent of tumor class. One theme to emerge from the latter approach has been the identification of proliferation markers, which were already well known to be predictive of aggressive tumor types in many different cancers. Perhaps the most instructive insights will come from much larger studies, where correlates with clinical outcome can be identified in patient samples for which the contribution of tumor subtype can be factored out by multivariate analysis.

In the few comparable pairs of experiments that have been published, researchers are identifying markedly different numbers of genes as optimal for identifying tumor classes, and different genes that appear useful for prognostication in similar patient cohorts. There are many possible explanations for this divergence. Firstly, some disparity is expected until the entire human transcriptome has been defined and is represented in a microarray format. Currently, differences in the technologies and in the gene sets analyzed make it difficult simply to compare and contrast independent datasets generated in different labs. Secondly, the population size of current published gene-expression datasets (usually involving less than 100 patients) may be inadequate for tumor types that exhibit great complexity. Perhaps most importantly, manipulation of microarray data has yet to be standardized; differences in chip-to-chip normalization, filtering for data quality, and the analytical methods used to identify genes of interest can all lead to varying results. Our (unpublished) survey of the available data suggests that, probably because of the difficulty in performing microarray experiments, there is inadequate filtering for poor measurements in many experiments, and this is likely to be a significant contributor to the disparity in the results of separate experiments.

Gene-expression microarrays are cumbersome tools for the analysis of sparse clinical material. Nevertheless, the studies reviewed here, as well as numerous others involving other tumor types, are rapidly identifying candidate genes and gene sets that appear useful for the development of a novel molecular-based classification of human cancers. It remains unclear what number of genes will be needed to reliably identify all clinically distinct tumor classes. Some clinically relevant tumor classes may prove to be easily distinguished by immunohistochemical reagents directed at

key markers identified in gene-expression studies. Other classes may require measurement of the expression of numerous markers that can most conveniently be analyzed by quantitative measurement of RNA. The evolution of clinical practice awaits the exploration of these nascent tools for tumor classification in large prospective and retrospective studies where their true utility in the management of patients can be discovered.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
3. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97:**262-267.
4. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
5. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
6. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al.*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
7. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Jr, Hampton GM: **Molecular classification of human carcinomas by use of gene expression signatures.** *Cancer Res* 2001, **61:**7388-7393.
8. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al.*: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci USA* 2001, **98:**15149-15154.
9. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
10. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ: **MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.** *Nat Genet* 2002, **30:**41-47.
11. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, *et al.*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98:**13790-13795.
12. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer.** *Nature* 2001, **412:**822-826.
13. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, *et al.*: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344:**539-548.
14. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, *et al.*: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415:**436-442.
15. Rickman DS, Bobek MP, Misek DE, Kuick R, Blaivas M, Kurnit DM, Taylor J, Hanash SM: **Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis.** *Cancer Res* 2001, **61:**6885-6891.
16. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, *et al.*: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8:**68-74.
17. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, *et al.*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
18. Stratowa C, Loffler G, Lichter P, Stilgenbauer S, Haberl P, Schweifer N, Dohner H, Wilgenbus KK: **cDNA microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential new prognostic markers involved in lymphocyte trafficking.** *Int J Cancer* 2001, **91:**474-480.
19. Takahashi M, Rhodes DR, Furge KA, Kanayama H, Kagawa S, Haab BB, Teh BT: **Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification.** *Proc Natl Acad Sci USA* 2001, **98:**9754-9759.
20. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, *et al.*: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.
21. Wetzels RH, Kuijpers HJ, Lane EB, Leigh IM, Troyanovsky SM, Holland R, van Haelst UJ, Ramaekers FC: **Basal cell-specific and hyperproliferation-related keratins in human breast cancer.** *Am J Pathol* 1991, **138:**751-763.
22. Lossos IS, Alizadeh AA, Eisen MB, Chan WC, Brown PO, Botstein D, Staudt LM, Levy R: **Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas.** *Proc Natl Acad Sci USA* 2000 **97:**10209-10213.