Journal of Pathology
J Pathol 2001; **195**: 312–320.
DOI: 10.1002/path.955

Original Paper

Molecular classification of breast cancer patients by gene expression profiling

André Ahr[1]*, Uwe Holtrich[1], Christine Solbach[1], Anton Scharl[2], Klaus Strebhardt[1], Thomas Karn[1] and Manfred Kaufmann[1]

[1] *Department of Obstetrics and Gynecology, J.W. Goethe-University, Theodor-Stern-Kai 7, D-60590 Frankfurt, Germany*
[2] *Department of Obstetrics and Gynecology, Klinikum St. Marien, D-92224 Amberg, Germany*

*Correspondence to:
A. Ahr, Department of Obstetrics and Gynecology, J. W. Goethe-University, Theodor-Stern-Kai 7, D-60590 Frankfurt, Germany.
E-mail: ahr@em.uni-frankfurt.de

Received: 4 January 2001
Accepted: 3 May 2001
Published online: 25 July 2001

Abstract

For many tumuors, pathological subclasses exist which have to be further defined by genetic markers to improve therapy and follow-up strategies. In this study, cDNA array analyses of breast cancers have been performed to classify tumuors into categories based on expression patterns. Comparing purified normal ductal epithelial cells and corresponding tumour tissues, the expression of only a small fraction of genes was found to be significantly changed. A subset of genes repeatedly found to be differentially expressed in breast cancers was subsequently employed to perform a classification of 82 normal and malignant breast specimens by cluster analysis. This analysis identifies a subgroup of transcriptionally related tumours, designated class A, which can be further subdivided into A1 and A2. Correlation with classical clinicopathological parameters revealed that subgroup A1 was characterized by a high number of node-positive tumours (14 of 16). In this subgroup there was a disproportionate number of patients who had already developed distant metastases at the time of diagnosis (25% in this subgroup, compared with 5% among the rest of the samples). Taken together, the use of these differentially expressed marker genes in conjunction with sample clustering algorithms provides a novel molecular classification of breast cancer specimens, which facilitates the identification of patients with a higher risk of recurrence.

Keywords: breast cancer; DNA array; tumour classification; cluster analysis; differentially expressed genes; diagnosis

Introduction

Breast cancer is a major cause of death among women in the age group of 35–55 years. Despite important advances in therapy, still more than half of the affected patients suffer from relapses [1]. This is in part due to the highly heterogenous nature of this disease; the various pathological breast cancer subclasses have markedly different clinical courses and treatment responses. Thus, breast cancer subclasses have to be further defined by genetic markers to improve therapy and follow-up strategies.

Although little is known about the genetic events implicated in tumour development and progression, common hallmarks of cancer cells include oncogene activation and loss of tumour suppressor gene function, as well as karyotypic mutations [for a recent review see Reference 2]. These mutations induce complex changes in cellular gene expression, which in concert define the biological tumour phenotype. No close, consistent correlation has been found between the expression of any given single gene and the clinical behaviour of breast tumours. Recently, however, novel array hybridization techniques based on cDNA or oligonucleotides have enabled the parallel expression profiling of several thousand genes, providing a powerful tool for characterizing complex cellular transcriptional activities [3–16]. At present, one major aim is to use DNA arrays as a tool to understand and classify tumours into categories based on shared gene expression patterns. It is anticipated that global determination of cellular transcriptional activity will identify gene expression signatures that predict the clinical behaviour of tumuors.

In the present study, we applied low and high density cDNA array analyses to identify differentially expressed genes and to evaluate transcriptional diversity among human breast cancers. The detected differentially expressed transcripts include several genes known from previous studies, as well as previously unrecognized transcripts. We show that class discovery analysis based on our gene expression profiling of 82 specimens identifies four main sample groups. A correlation of the cluster data with classical clinicopathological parameters revealed that one subgroup was characterized by a remarkably high number of node-positive tumours and a disproportionate number of patients who had already developed distant metastases at the time of diagnosis. These M1 patients compared 25% of this subgroup, compared with 5% of the rest of the samples. These cluster analysis data may

Abbreviations: FAM: 6-carboxy-fluorescein-succinimidylester; TAMRA: 6-carboxy-tetramethyl-rhodamine-succinimidylester.

help to define patients with an early onset of disease progression, providing a first step towards improved patient-adapted therapy.

Materials and methods

Tissue samples

All tissue samples were obtained from patients undergoing surgical resection between June 1997 and June 1999 at the Department of Obstetrics and Gynecology of the J. W. Goethe University (Frankfurt). Specimens included ductal and lobular carcinomas of different tumour size (T1–T4), lymph node status (N0–1), grade (G1–G3), hormone receptor status (ER/PR positive and negative) and distant metastases (M0–1). Normal tissue samples were obtained from patients undergoing surgical breast reduction. The cluster analysis was performed on a sample group of 9 benign and 73 malignant breast specimens (7 M1, 66 M0, 41 N1, 26 N0, 6 NX) as well as several additional samples, such as cell lines and lymph node metastases, which were not considered in the statistical evaluations. Only pathologically verified data on lymph node status were considered for calculations.

Epithelial cell purification from human mammary gland

Mammary ductal epithelial cells were isolated from sections of breast tissue by two rounds of immunomagnetic purging, using the monoclonal antibody HEA125 as described [17]. Briefly, tissue was mechanically disintegrated (MediMachine, DAKO, Hamburg, Germany) and the cell suspension was subsequently incubated with mAb HEA125 for 1 hour at 4°C. Cells were then washed with PBS/EDTA, incubated with a magnetic bead-coupled goat anti-mouse IgG antibody and passed through a MACS (magnetic cell separation system) separator column (Miltenyi Biotec, Bergisch Gladbach, Germany). The purity of the isolated epithelial cells was $>90\%$ as judged microscopically.

RNA isolation and cDNA array hybridization

Total RNA from human primary mammary carcinomas was isolated by the guanidinium isothiocyanate method [18] in combination with affinity purification (RNeasy, Qiagen, Hilden, Germany). Radiolabelling of the nucleic acid was performed by reverse transcription of 5 μg total RNA according to the protocols of the Atlas Array Blots from Clontech (Palo Alto, CA) using the reagents provided in these kits (MMLV reverse transcriptase) and [α-^{32}P]-dATP. The probes were added to the ExpressHyb (Clontech) hybridization solutions at a concentration of 1×10^6 dpm/ml. Hybridizations were performed overnight and blots were subsequently washed under high stringency conditions ($0.1\times$SSC, 0.5% SDS, 68°C). After autoradiography, films were analysed by densitometric scanning (Personal Densitometer, Molecular Dynamics, Sunnyvale, CA). Raw data were processed using imaging software (ImageQuant, Molecular Dynamics, Sunnyvale, CA), transferred to spreadsheet programs and normalized by calibration markers. Low density arrays (Human Cancer Blot, 588 genes) were obtained from Clontech (Palo Alto, CA). High density arrays (Human GDA 1.3 containing 45 000 cDNA clones) were supplied by GenomeSystemsInc (St. Louis, MO).

Real-time PCR analysis

Real-time PCR analyses were performed using the ABI 7700 Sequence Detection System (PE-Applied Biosystems, Foster City, CA). cDNAs for all PCRs were generated by random primed reverse transcription (ProSTAR cDNA-synthesis kit, Stratagene, La Jolla, CA). PCR reactions were performed according to the manufacturer's protocols (PE-Applied Biosystems, Foster City, CA). VIC-fluorophore labelled GAPDH TaqMan probes served as internal quantification markers in multiplex PCR reactions. Each quantitation was reproduced three times and normalized by GAPDH, actin and 18S rRNA standards.

Cluster analysis and class prediction

Differentially expressed genes recurrently observed in multiple array analyses of primary breast cancers were used to screen a panel of 94 specimens by real-time PCR assays (TaqMan); 15 of the marker genes most varying in expression among samples were picked from low density arrays, as well as eight genes from high density arrays. In addition, 11 genes were included whose role in breast cancer has already been described or which are useful as surrogate markers for proliferation (MKi67, PoloLikeKinase), IFN inducible genes (STAT1), stromal cells (DDR2) and vascularization (VEGFR). Prior to cluster analysis, expression data were log-transformed and 10-times median centred for each sample. Samples were grouped according to these normalized expression data by average linkage clustering, using the Pearson correlation as implemented in the program CLUSTER [19]. Calculated relative similarities were subsequently graphically displayed using the TreeView program [19]. The output of this program is an unrooted tree, where lengths of the horizontal branches represent similarity distances of the expression profiles (1 – Pearson correlation coefficient).

To validate class distinctions identified by this cluster analysis and to test their consistency, the method of class prediction as proposed by Golub *et al.* [7] was used. In brief, a prediction strength (*PS*) for the assignment of a sample to a specific class is calculated by $PS=(V_{\mathrm{win}}-V_{\mathrm{lose}})/(V_{\mathrm{win}}+V_{\mathrm{lose}})$, where V is the absolute value of the vote total for the respective classes ('win' or 'lose'). A weighted vote v_i in favour of a class for each gene i is calculated by $v_i=P_i\cdot(x_i-(\mu_1+\mu_2)/2)$, where x_i represents the expression level of the gene i in the sample, μ and σ represent the mean and standard deviation respectively of the gene's expression in the two classes and the correlation

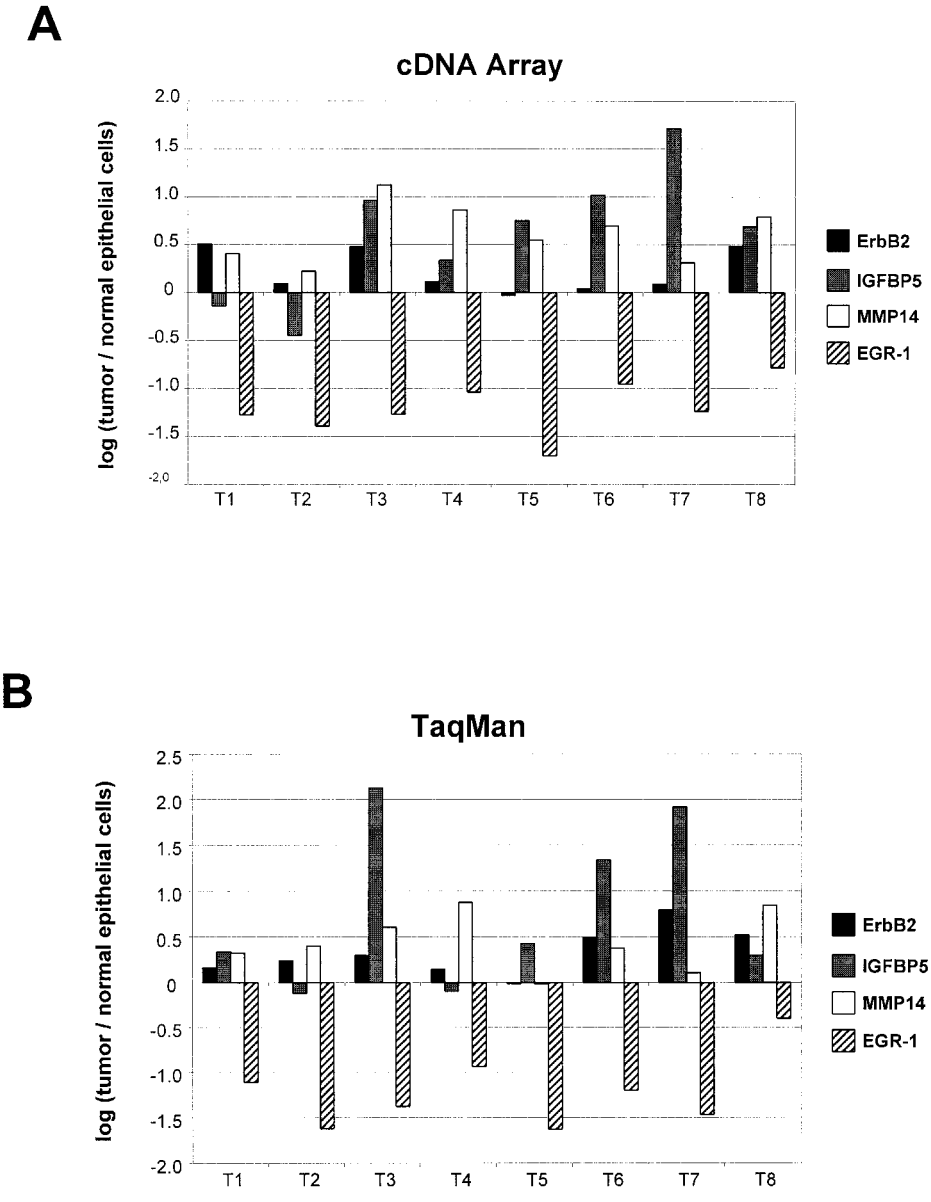


Figure 1. Comparison of cDNA array and RT-PCR results. (A) Ratios of expression levels detected by cDNA array hybridization from eight different tumour samples (T1–T8) compared with normal epithelial cells are given in log units on the *y* axis for four different genes (see squares on the right for gene identities). (B) The same samples and genes as in (A) were analysed by real-time RT-PCR. Expression level ratios of tumour versus normal cells are displayed in log units on the *y* axis

metric P is defined by $P=(\mu_1-\mu_2)/(\sigma_1+\sigma_2)$. Positive values for v_i represent votes for the winning class and negative value votes for the losing class. For cross-validation, one sample was withheld, the p-value of each gene was calculated for the rest of the samples in the respective class and the prediction strength (PS) for the withheld sample was computed [for details see Reference 7].

Results

Detection of differentially expressed genes in human primary mammary carcinomas

To collect differentially expressed genes as markers for a molecular tumour classification, cDNA array analyses of human mammary carcinomas were performed. Pathological cases used for mRNA expression analysis encompassed 15 ductal and 2 lobular carcinomas of different grading (G1–G3), tumour size (T1–T4) and lymph node status. These samples were hybridized to low density arrays (Atlas Cancer Array containing 588 genes). In addition, four tumour samples (three ductal and one lobular carcinoma, G1–G2, N0 and N1) were analysed by using high density arrays (GenomeSystems GDA 1.3 filters containing 45 000 genes). A pool of RNAs from antibody purified normal mammary gland epithelial cells obtained from two healthy donors were subsequently hybridized to high and low density arrays, which served as a reference for all later performed comparisons. Antibody purified cells were chosen since total benign tissue is composed of a

Table 1. Differentially expressed genes in mammary carcinomas versus normal epithelial cells

Fold difference of expression	Percentage of genes altered*
≥5×	6.4%
≥10×	2.7%
≥15×	1.5%
≥20×	0.9%

*Tumour RNA as well as RNA from antibody-purified mammary epithelial cells were hybridized to cDNA arrays containing 588 different genes. Differences in gene expression were determined as a ratio between signals of tumour and normal cells. The percentage of altered genes was calculated for different cut-off values in fold expression. The data represent mean values of seven analyses.

mixture of different cell types, while tumours represent enriched transformed epithelial cell populations. After hybridization, autoradiographs were scanned densitometrically, normalized and differences in relative gene expression were determined as signal ratio of tumour versus normal mammary gland epithelial cells. The use of this common reference allows the comparison of relative expression levels across all our samples. Array hybridization results were validated by real-time PCR analysis of 15 differentially expressed genes. As shown by an example of four genes (Figure 1), quantification with both methods gave similar results, although it should be noted that the range of detection is more dynamic for PCR analysis. In general, measurement of gene expression by low and high density array hybridization yielded levels over background for more than half of the analysed genes. We observed subtle differences (2∼3-fold) in gene expression for about one-third of the analysed genes when RNA from different parts of the same tissue sample were compared. Thus, to define a cut-off value for the comparison of different samples, only changes greater than five-fold were considered in further analyses. Comparing the expression profiles of mammary carcinomas versus normal epithelial cells by low density array analysis (588 genes), approximately 6.4% of the genes were found to be more than five-fold altered; only 2.7% differed when the cut off value was increased to ten-fold (Table 1). Less than 1% of the genes showed more than a 20-fold alteration in expression, indicating that between individual tumours the number of strong expression differences is smaller than 2%. These values are in agreement with data from high density arrays: in two comparisons, we identified about 100 genes with differences in expression of more than 20-fold, corresponding to 0.5% of clones with hybridization signals over background (data not shown). To estimate the transcriptional diversity among mammary carcinomas, the cumulative number of detected differentially expressed genes was plotted against the number of analysed samples. As depicted in Figure 2, the data set generates an asymptotic curve with a typical saturation plateau. From this plot it can be calculated that about 20 mammary tumour samples are sufficient to detect most of the differentially expressed genes.

Molecular tumour classification by sample clustering

Gene expression data generated by array hybridization can be used to group tumour samples into clusters that reflect their biological properties [19–21]. A number of genes repeatedly found to be differentially expressed in array analyses were applied in TaqMan assays to perform a molecular tumour classification of 82 normal and malignant breast specimens (see Materials and methods for details). As shown in Figure 3A, cluster analysis of the expression data identifies four main sample groups (indicated by I–IV in the figure). The largest group (III) splits into two sub-branches containing one population of transcriptionally related tumour samples (designated as class A), as well as a less homologous tumour class B. Note the shorter branch distances between tumours of class A compared with class B, indicating the higher degree of transcriptional

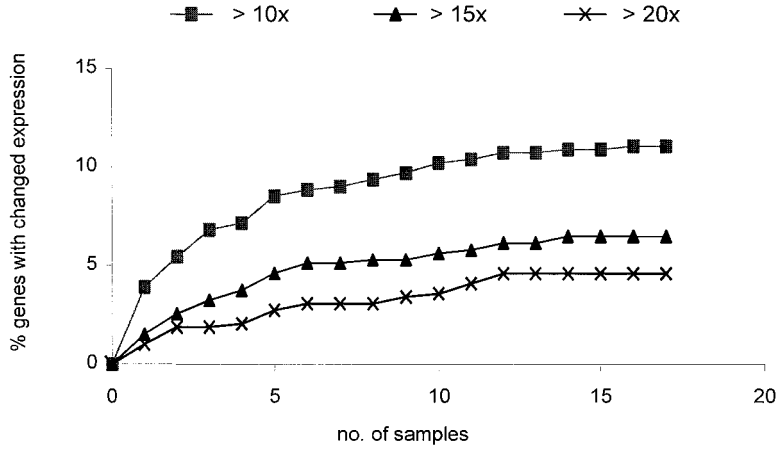
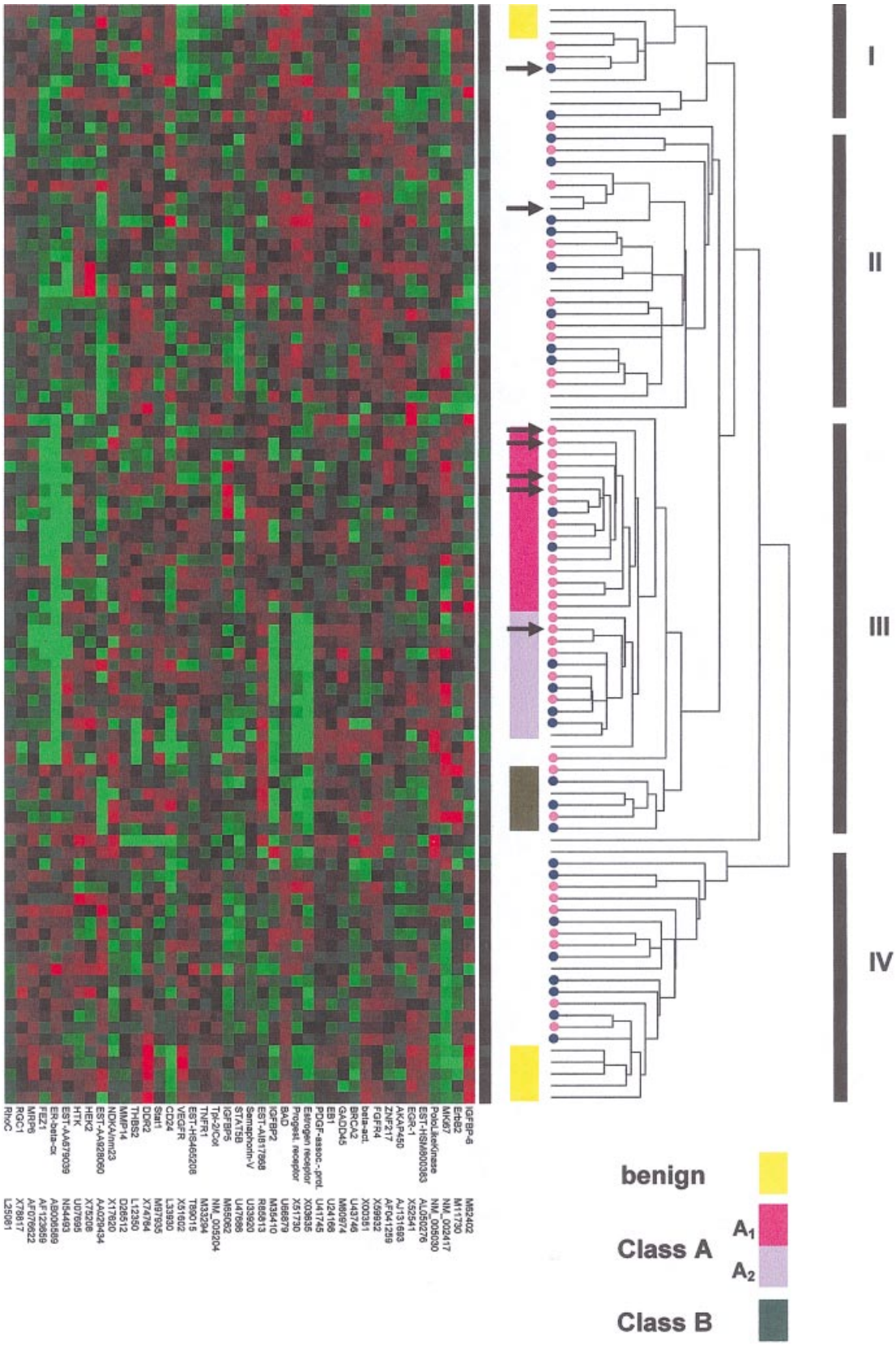


Figure 2. Analysis of the transcriptional heterogeneity of mammary carcinomas. Several randomly selected mammary carcinomas were analysed by array hybridization and the cumulative number of detected differences in gene expression (*y* axis) was plotted against the number of analysed samples (*x* axis). Data from low density arrays (588 genes) are shown for three different cut-off values in fold expression (see respective symbols on the top)



similarity. A more detailed inspection of class A tumours revealed a further splitting into two subpopulations (A1 and A2). This subdivision is also visible at the transcriptional level of single genes by an exceptionally low expression of the oestrogen and progesterone receptor, as well as the proapoptotic BAD gene and IGFBP2 in subpopulation A2 compared with A1. However, one of the most common hallmarks of class A tumours, the striking downregulation of the recently described ER-β-cx gene [22], is perfectly retained in

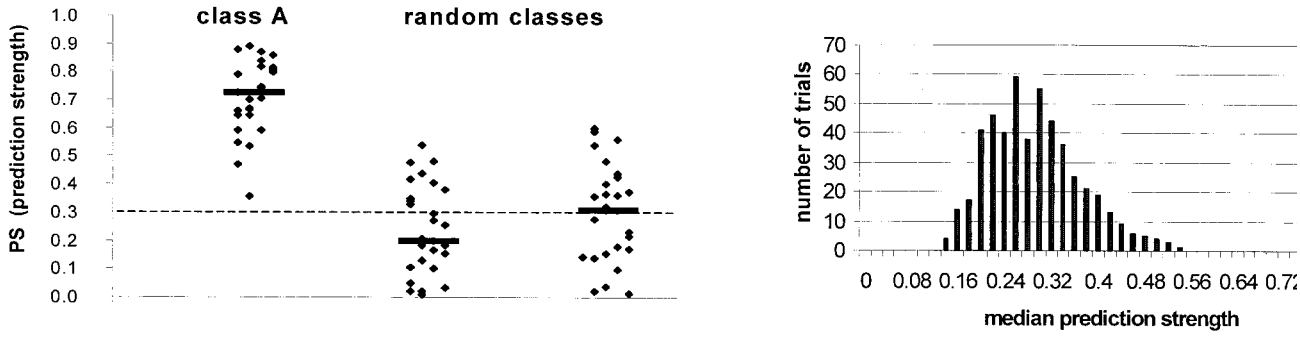


Figure 3. Class discovery of primary breast cancers by cluster analysis. (A) Cluster analysis of expression profiles. Shown is a schematical representation of gene expression patterns across all specimens. Red indicates expression levels above median, green below. The corresponding unrooted tree, where branch lengths represent similarity distances (1 – Pearson correlation coefficient) of samples as judged by their expression patterns, is depicted on the right. The four main sample groups (I–IV) are indicated by vertical bars on the right. Note the clustering of samples in classes A1 and A2 (magenta and violet coloured bars, respectively). Pathologically verified node-positive patients are represented by magenta coloured dots, node-negative patients by dark blue dots. The percentage of node-positive patients in class A1 is 88%, in contrast to 53% across the remaining malignant samples. The accumulation of patients with distant metastases at time of diagnosis in this subgroup is highlighted by arrows. A second group of mammary tumours was identified in class B, which consists exclusively of oestrogen receptor negative specimens. (B) Validation of sample clustering by class prediction. The method of class prediction as proposed by Golub *et al.* [7] was used to validate class distinctions. The scatterplots on the left show the distribution of prediction strength (*PS*) scores, which measure the assignment of a sample to a given class. The first plot shows the prediction strength values observed in cross-validation of samples belonging to class A compared with all malignant samples not belonging to class A or B (median = 0.73, represented by a vertical bar). The remaining plots show the distribution of predictors corresponding to two randomly generated classes (median 0.21 and 0.31, respectively). A total of 500 such random class distinctions using the same data were analysed to evaluate the statistical significance of the class A distinction. The histogram on the right shows the distribution of median *PS* values obtained. The highest median *PS* value observed was 0.52, on one occasion among these permutations. The distinction between subclasses A1 and A2 with a median *PS* value of 0.54 seems therefore to be less significant than the class A identification, but for most samples even this distinction is above a threshold *PS* of 0.3 as suggested by Golub *et al.* [7]

each sample of both subpopulations. We used cross-validation analysis [7] to verify the consistency of sample distribution, yielding prediction strengths significantly higher than would be expected for random class distinctions (Figure 3B).

We next correlated the cluster data with classical clinicopathological parameters (TNM state, histological subtype and grade) to elucidate relationships between gene expression profiles and the biological behaviour of the analysed tumours. While no correlation was detectable between cluster data and tumour size, grade or histological subtype, a striking enrichment of node-positive tumours (88% relative to 61% overall) was observed in subgroup A1. Interestingly, in this subgroup we also found an accumulation of samples from patients who had already developed distant metastases at the time of diagnosis (Figure 3A, marked by arrows). The percentage of these M1 patients among the node-positive samples was determined to be 29% within this subgroup, in contrast to 11% of the node-positive samples outside subgroup A1. Overall, 25% of all samples within subgroup A1 are M1, in contrast to only 5% among the rest of the malignant samples. In conclusion, this subgroup contained a disproportionate number of breast cancers which already showed peripheral tumour cell dissemination, associating these patients with a higher risk of disease recurrence.

Discussion

The analysis of pathological changes in gene expression can contribute to the understanding of disease mechanisms, the improvement of diagnosis and the identification of novel therapeutic targets. New technical advances and the completion of several sequencing projects enabled the production of high density DNA arrays, providing ideal tools to analyse the complex transcriptional changes accompanying cellular transformation.

The major aim of our study was the identification of differentially expressed genes in breast cancer which can subsequently be employed as markers for a molecular characterization of tumour samples. First, systematic expression analyses were performed to give hints about the transcriptional diversity and the number of tissue samples which have to be analysed to detect the bulk of differentially expressed genes among breast cancers. A comparative expression analysis between normal mammary ductal epithelium and primary mammary carcinomas, based on array hybridization results, revealed that most genes were expressed at roughly equal levels. These findings are similar to SAGE (serial analysis of gene expression [23]) analysis data provided by Zhang *et al.* [24], who described approximately 2% of transcriptionally altered genes, comparing normal colon epithelium and

primary colon cancer. We could show that an expression analysis of approximately 20 tumours should be sufficient to detect most of the transcriptionally altered genes. According to these data, we estimate that our analyses using high density 45 000 clone arrays have so far detected about 10–20% of the differentially expressed genes in breast cancer.

The histomorphological and clinical parameters in use today seem not to be sufficient to discriminate some subtypes of breast cancer with a markedly different clinical course and response to therapy. It can be expected that analyses of tumour expression profiles will allow a precise definition of the cellular *status quo*, allowing this gap to be filled. An intriguing possibility for the interpretation of global cellular expression data is provided by gene and sample clustering algorithms [19–21]. Based on transcriptional similarities, a molecular classification of pathological tissue specimens is performed, providing the opportunity to detect novel prognostic or predictive markers in each subgroup [7,11]. A molecular classification of tumour samples can be achieved using either unsupervised methods like hierarchical clustering [19], *k*-means clustering [20] or 'SOMs' (self organizing maps) [21], as well as supervised methods like 'SVMs' (support vector machines) [25]. We chose hierarchical clustering, which has already been successfully used to classify tumour samples [6,11]. Alizadeh *et al.* [11] were able to identify formerly unknown types of B-cell lymphoma with distinct clinical behaviour by using hierachical clustering of expression data. Golub *et al.* [7] used SOMs on DNA array data to differentiate subtypes of acute leukaemia. By using 50 'informative genes' they succeeded in discriminating the different treatment-requiring forms of ALL and AML, as well as classifying new subtypes.

In our sample group, a hierarchical clustering programme identified one cluster of mammary carcinomas which consisted disproportionately of node-positive tumours, predicting an unfavourable outcome for these patients. Interestingly, in this subgroup we also detected an accumulation of samples from patients who had already developed distant metastases at the time of diagnosis. The total percentage of these M1 patients was determined to be 25% in this subgroup, compared with 5% among the rest of the samples. Carcinomas in this cluster seem therefore to share biological properties which allow an early peripheral dissemination of viable tumour cells. Thus, our actual set of differentially expressed marker genes may be useful to define cancer patients with a higher risk of disease recurrence.

Although a hallmark of patients from subgroup A1 is a positive lymph node status, two breast cancer specimens from node-negative patients were found in this cluster branch. It is not yet clear if this is due to an incorrect classification by pathological examination, or if further markers are required to define this subgroup more precisely. We are currently analysing breast cancers with long term follow-up, to determine the exact predictive and prognostic value of these marker genes and we are checking the inclusion of additional informative genes.

Acknowledgements

We thank Silke Deckert for technical assistance. This work was supported by a grant from the Deutsche Krebshilfe (10-1478-Ka2).

References

1. Ries LAG, Eisner MP, Kosary CL, *et al.* (eds). *SEER Cancer Statistics Review, 1973–1997*. National Cancer Institute: Bethesda, MD, 2000.
2. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000; **100**: 57–70.
3. 'The Chipping Forecast'. *Nat Genet* 1999; **21** (Suppl).
4. Granjeaud S, Bertucci F, Jordan BR. Expression profiling: DNA arrays in many guises. *Bioessays* 1999; **21**(9): 781–790.
5. Alon U, Barkai N, Notterman DA, *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 U S A* 1999; **96**: 6745–6750.
6. Perou C, Jeffrey SS, van de Rijn M, *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* 1999; **96**: 9212–9217.
7. Golub TR, Slonim DK, Tamayo P, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999 **286**(5439): 531–537.
8. Backert S, Gelos M, Kobalz U, *et al.* Differential gene expression in colon carcinoma cells and tissues detected with a cDNA array. *Int J Cancer* 1999; **82**: 868–874.
9. Moch H, Schraml P, Bubendorf L. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol* 1999; **154**: 981–986.
10. Emmert-Buck MR, Strausberg RL, Krizman DB, *et al.* Molecular profiling of clinical tissue specimens: feasibility and applications. *Am J Pathol* 2000; **156**: 1109–1115.
11. Alizadeh AA, Eisen MB, Davis RE, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403**: 503–511.
12. Fambrough D, McClure K, Kazlauskas A, Lander ES. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* 1999; **97**(6): 727–741.
13. Anbazhagan R, Tihan T, Bornman DM, *et al.* Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Res* 1999; **59**(20): 5119–5122.
14. Harkin DP, Bean JM, Miklos D, *et al.* Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* 1999; **97**: 575–586.
15. Martin KJ, Kritzman BM, Price LM, *et al.* Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res* 2000; **60**(8): 2232–2238.
16. Sgroi DC, Teng S, Robinson G, LeVangie R, Hudson JR Jr, Elkahloun AG. *In vivo* gene expression profile analysis of human breast cancer progression. *Cancer Res* 1999; **59**(22): 5656–5661.
17. Miltenyi S, Müller W, Weichel W, Radbruch A. High gradient magnetic cell separation with MACS. *Cytometry* 1990; **11**: 231–238.
18. Chirgwin JM, Przybyla AE, McDonald RJ, Rutter WJ. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 1979; **18**: 5294–5299.
19. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; **95**: 14863–14868.

20. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999; **22**: 281–285.
21. Tamayo P, Slonim D, Misirov J, *et al*. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999; **96**: 2907–2912.
22. Ogawa S, Inoue S, Watanabe T, *et al*. Molecular cloning and characterization of human estrogen receptor betacx: a potential inhibitor of estrogen action in human. *Nucl Acids Res* 1998; **26**(15): 3505–3512.
23. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995; **270**: 484–487.
24. Zhang L, Zhou W, Velculescu VE, *et al*. Gene expression profiles in normal and cancer cells. *Science* 1997; **276**: 1268–1272.
25. Brown MP, Grundy WN, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 2000; **97**(1): 262–267.

Supplementary information: nucleotide sequences of primers and probes

Marker	Acc.-No.	Primer/Probe	Sequence (5′–3′)
AKAP450	AJ131693	Upper	GAGCAGGGCTTCTCTGTGGAAC
AKAP450	AJ131693	Lower	CACAAAACTGCAACCATTTACC
AKAP450	AJ131693	Probe	FAM-CAACCACAGCAGATGACTCAGTT-TAMRA
BAD	U66879	Upper	GCACAGCAACGCAGATGCGGC
BAD	U66879	Lower	AACTTCCGATCCCACCAGGAC
BAD	U66879	Probe	FAM-CTCCAGCTGGACGCGAGTCTTCC-TAMRA
beta-act.	X00351	Upper	CCATCATGAAGTGTGACGTGGAC
beta-act.	X00351	Lower	TGGTGGTGCCGCCAGACAGCAC
beta-act.	X00351	Probe	FAM-CCGCAAAGACCTGTACGCCAAC-TAMRA
BRCA2	U43746	Upper	CAAGATGGTGCAGAGCTTTATG
BRCA2	U43746	Lower	TCTTCACTGAAATAACCCTCAAG
BRCA2	U43746	Probe	FAM-CAGTGAAGAATGCAGCAGACCCAGC-TAMRA
CD24	L33930	Upper	ACTAATGCCACCACCAAGGCGGC
CD24	L33930	Lower	TGCAGAAGAGAGAGTGAGACCAC
CD24	L33930	Probe	FAM-CCTGCAGTCAACAGCCAGTCTC-TAMRA
DDR2	X74764	Upper	CCATTGTAGCCAGATTTGTCC
DDR2	X74764	Lower	GCTCCACTCTCATACACACATTC
DDR2	X74764	Probe	FAM-CATTCCAGTCACCGACCACTCC-TAMRA
EB1	U24166	Upper	ATTGTCAGTTTATGGACATGC
EB1	U24166	Lower	CTAGCTTAGCTTGGAATTTCAC
EB1	U24166	Probe	FAM-CCCTGGCTCCATTGCCTTGAA-TAMRA
EGR-1	X52541	Upper	GTGCCGCATCTGCATGCGCAAC
EGR-1	X52541	Lower	GCTTTTCGCCTGTGTGGGTGCGG
EGR-1	X52541	Probe	FAM-CAGCCGCAGCGACCACCTCACC-TAMRA
ErbB2	M11730	Upper	AATGAGGACTTGGGCCCAGC
ErbB2	M11730	Lower	CAGATACTCCTCAGCATCCACCAGG
ErbB2	M11730	Probe	TAM-CAGCACCTTCTACCGCTCACTGC-TAMRA
ER-beta cx	AB006589	Upper	GTAGACAGCCACCATGAATATCC
ER-beta cx	AB006589	Lower	CTGGGAATGCTGTAATTCATC
ER-beta cx	AB006589	Probe	FAM-CCATGACATTCTATAGCCCTGC-TAMRA
EST-AA679039	N54493	Upper	CCTCTCTAGGGAGGAGAAAGG
EST-AA679039	N54493	Lower	CCCAGCCAGTCCTGCTCTGTG
EST-AA679039	N54493	Probe	FAM-CTAGCTACAGTCACCAGCAGGACC-TAMRA
EST-AA928060	AA029434	Upper	GTAACCATAATGTCAACATAACC
EST-AA928060	AA029434	Lower	GGAAAACGCACGCACTTAGGC
EST-AA928060	AA029434	Probe	FAM-CCTAACGGAACAGGAGATCGCC-TAMRA
EST-AI817868	R85813	Upper	GGCCATTGTGTCAATGGCTCAG
EST-AI817868	R85813	Lower	GCTGAATCGAACATTCCAATCC
EST-AI817868	R85813	Probe	FAM-CTTCAAGATCTTCGCTGGAACC-TAMRA
EST-HS465208	T89015	Upper	AATTATCTAATAGGTTGGCAC
EST-HS465208	T89015	Lower	AGGACAATAGAGAGCTTCACC
EST-HS465208	T89015	Probe	FAM-CATGAGCCCCTGTTCTCATTCTGC-TAMRA
EST-HSM800383	AL050276	Upper	GACAAACTGCTGCTTGGCTAC
EST-HSM800383	AL050276	Lower	AGTCAATGAGCTTTTGCACTGAC
EST-HSM800383	AL050276	Probe	FAM-CATCGAGATCCCGTCGGTGGTGTC-TAMRA
Estrogen receptor	X03635	Upper	CAAGGAGACTCGCTACTGTGC
Estrogen receptor	X03635	Lower	GCCCTCACAGGACCAGACTCC
Estrogen receptor	X03635	Probe	FAM-CAATGACTATGCTTGACCGTACC-TAMRA
FEZ1	AF123659	Upper	TGGCCATGTACCAGCGGAACC
FEZ1	AF123659	Lower	CCGGCGCTGTCCCCACGTGC
FEZ1	AF123659	Probe	FAM-CCTGGAGAAGGCCCTGCAGCAGC-TAMRA
FGFR4	X59932	Upper	GTGGCCAAGGTCAGCGACTTTG
FGFR4	X59932	Lower	TGACTGGCAGCTTGCCCGTGTC
FGFR4	X59932	Probe	FAM-CACCAAGGAGGCGTCCAGCACC-TAMRA

GADD45	M60974	Upper	CGAGGACGACGACAGAGATGTGGC
GADD45	M60974	Lower	ATGTCGTTCTCGCAGCAAAACGC
GADD45	M60974	Probe	FAM-CAGATCCACTTCACCCTGATCC-TAMRA
HEK2	X75208	Upper	AGGCTGCCCCGTCTGAAGTGC
HEK2	X75208	Lower	AGGATAGGGTGAGGCTGCTGC
HEK2	X75208	Probe	FAM-CACACTACGCCTGCACAGCAGCTC-TAMRA
HTK	U07695	Upper	CATCGCCTCGGGCATGCGGTAC
HTK	U07695	Lower	GATGTTGCGAGCAGCCAGGTC
HTK	U07695	Probe	FAM-CCGAGATGAGCTACGTCCACC-TAMRA
IGFBP2	M34510	Upper	CTGCACATCCCCAACTGTGAC
IGFBP2	M34510	Lower	GCCCGTTCAGAGACATCTTGC
IGFBP2	M34510	Probe	FAM-CATGGCCTGTACAACCTCAAAC-TAMRA
IGFBP5	M65062	Upper	TACTCCCCCAAGATCTTCCGGCC
IGFBP5	M65062	Lower	TTCTGCGGTCCTTCTTCACTGC
IGFBP5	M65062	Probe	FAM-CCCGCATCTCCGAGCTGAAGGC-TAMRA
ICFBP6	M62402	Upper	TGGGCCCATGCCGTAGACATC
ICFBP6	M62402	Lower	TGTTTGAGCCCCTCGGTAGAC
ICFBP6	M62402	Probe	FAM-CAGTGCTGCAGCAACTCCAGAC-TAMRA
MMKi67	NM_002417	Upper	AGACTTGGCTGGCTTGAAAGAGC
MMKi67	NM_002417	Lower	GTGTTTTCTCGTGAGTCGTGGGC
MMKi67	NM_002417	Probe	FAM-CCAGACACCAGTATGCACTGA-TAMRA
MMP14	D26512	Upper	TGCCGAGGGCTTCCATGGCGAC
MMP14	D26512	Lower	GCCCTGGGAAGTAGGCATGG
MMP14	D26512	Probe	FAM-AGCCGCCCTCACCATCGAAGGGC-TAMRA
MRP6	AF076622	Upper	CCCATGTACCTCTGGGTCCTTG
MRP6	AF076622	Lower	ACCATCTTGGCTTTGAAGAGTG
MRP6	AF076622	Probe	FAM-TCCACAGGTAGCCCCGGCCATGGT-TAMRA
NDKA/nm23	X17620	Upper	ACCCTGCAGACTCCAAGCCTG
NDKA/nm23	X17620	Lower	ATGTATAATGTTCCTGCCAAC
NDKA/nm23	X17620	Probe	FAM-CCATCCGTGGAGACTTCTGCATAC-TAMRA
PDGF-assoc.-prot.	U41745	Upper	GGAAGACAGAGCAAGCCAAGGC
PDGF-assoc.-prot.	U41745	Lower	TTCCGGGCAGCCTCCTCCCGC
PDGF-assoc.-prot.	U41745	Probe	FAM-CCTGGCCCGGCTGGCCATCATCC-TAMRA
PoloLikeKinase	NM_005030	Upper	GATACTACCTACGGCAAATTGTGC
PoloLikeKinase	NM_005030	Lower	AGGTTGCCCAGCTTGAGGTCTC
PoloLikeKinase	NM_005030	Probe	FAM-CTGCCAGTACCTGCACCGAAACC-TAMRA
Progest receptor	X51730	Upper	ACCACGGTGATGGATTTCATCC
Progest receptor	X51730	Lower	AGCAGCTGCCGAGTGCGGGCTGC
Progest receptor	X51730	Probe	FAM-CCTATCCTGCCTCTCAATCACGCC-TAMRA
RGC1	X78817	Upper	GGGCAGGTGCTCCGGAGCTAC
RGC1	X78817	Lower	GGCTGCCCAGGCCCTGCACTTG
RGC1	X78817	Probe	FAM-CGCTGAGAGCCGCACCCAAGCC-TAMRA
RhoC	L25081	Upper	GACACAGCAGGGCAGGAAGAC
RhoC	L25081	Lower	ACATGAGGATGACATCAGTGTC
RhoC	L25081	Probe	FAM-CGACTGCGGCCTCTCTCCTACCC-TAMRA
Semaphorin-V	U33920	Upper	TCCCGTGCACTGCAGCTCAGCGATC
Semaphorin-V	U33920	Lower	GACGACGTGCTTAAAGTTGTTC
Semaphorin-V	U33920	Probe	FAM-CCTCTACTCCTGCACAGCCAC-TAMRA
Stat1	M97935	Upper	CATTCAGAGCTCGTTTGTGGTG
Stat1	M97935	Lower	CTTCAAGACCAGCGGCCTCTG
Stat1	M97935	Probe	FAM-CAGCCCTGCATGCCAACGCACC-TAMRA
STAT5B	U47686	Upper	CTCTCCAGCTGGAAGCCTTGC
STAT5B	U47686	Lower	GTCGCAGCTCCTCAAACGTCTG
STAT5B	U47686	Probe	FAM-CATGTCCCAGAAACACCTCCAGATC-TAMRA
THBS2	L12350	Upper	CAACCTCAATCTGGTCTGCGC
THBS2	L12350	Lower	TGGCAGATGGGGGCAGTTATC
THBS2	L12350	Probe	FAM-CAACGCCACCTACCACTGCATC-TAMRA
TNFR1	M33294	Upper	AGATTGAGAATGTTAAGGGCAC
TNFR1	M33294	Lower	GGCAAAGACCAAAGAAAATGAC
TNFR1	M33294	Probe	FAM-CTCAGGCACCACAGTGCTGTTGC-TAMRA
Tpl-2/Cot	NM_005204	Upper	ACCCGCCCAGAGAGGATCAGC
Tpl-2/Cot	NM_005204	Lower	CTCAGCAGCCTCTTGCGCTCC
Tpl-2/Cot	NM_005204	Probe	FAM-CTGTACGAGTCTGGACTCTGCC-TAMRA
VEGFR	X51602	Upper	CACATGACTGAAGGAAGGGAGC
VEGFR	X51602	Lower	TAAAGTAACAGTGATGTTAGG
VEGFR	X51602	Probe	FAM-CGTCATTCCCTGCCGGGTTAC-TAMRA
ZNF217	AF041259	Upper	GATGTTACTCCTCCTCCGGATG
ZNF217	AF041259	Lower	CACACTTGGCCTGTATCTGCA
ZNF217	AF041259	Probe	FAM-AAAGAGAAGCAAACGGAGACCGCAGC-TAMRA