Hum Genet (2002) 111 : 411–420
DOI 10.1007/s00439-002-0804-5

ORIGINAL INVESTIGATION

T.-K. Jenssen · W. P. Kuo · T. Stokke · E. Hovig

Associations between gene expressions in breast cancer and patient survival

Received: 19 April 2002 / Accepted: 1 July 2002 / Published online: 23 August 2002

Abstract We analyzed associations between gene expression in breast cancer and patient survival for 8024 genes from a previously published microarray data set. Analysis of survival, by using the logrank test, was performed automatically for each gene. After correcting for multiple testing, we identified 95 genes whose expression was significantly associated with patient survival. The independent prognostic value of the genes ranking the highest in univariate analysis, together with clinical parameters, was assessed by Cox multivariate regression anaysis. The *P*-values from these logrank tests were also mapped to chromosomal positions and compared with previously reported amplicon regions. We used PubGene web tools to identify groups of genes that had co-occurred in the literature and whose expression patterns were associated with survival. Our analyses demonstrate the comprehensiveness of the microarray technology with respect to measuring gene expression and indicate that the technology may be used to screen for potential clinical markers.

T.-K. Jenssen (✉)
Department of Computer and Information Science,
Norwegian University of Science and Technology,
NO-7491, Trondheim, Norway
e-mail: tkj@idi.ntnu.no, Tel.: +47-22-935416

T.-K. Jenssen · E. Hovig
Department of Tumor Biology, The Norwegian Radium Hospital,
Montebello, Oslo, Norway

T. Stokke
Department of Biophysics, The Norwegian Radium Hospital,
Montebello, Oslo, Norway

W.P. Kuo
Children's Hospital Informatics Program
and Division of Endocrinology, Department of Medicine,
Children's Hospital, Boston, MA 02115, USA

W.P. Kuo
Decision Systems Group, Brigham and Women's Hospital,
Harvard Medical School, Boston, MA 02115, USA

W.P. Kuo
Division of Health Sciences and Technology,
Harvard University and Massachusetts Institute of Technology,
Cambridge, MA 02139, USA

Introduction

High-throughput methods to measure gene expression, such as DNA microarrays and serial analysis of gene expression (DeRisi et al. 1997; Velculescu et al. 1995), have been established to characterize tumors. Relatively comprehensive profiles of mRNA levels can be obtained and used to discriminate cancer cells from normal cells (Alon et al. 1999) and to provide sub-classes of tumor types (Alizadeh et al. 2000; Dhanasekaran et al. 2001; Golub et al. 1999; Hedenfalk et al. 2001; Khan et al. 2001; Perou et al. 2000). Sub-classes may be associated with outcome and may be used to predict prognosis (Sorlie et al. 2001; van 't Veer et al. 2002). Comparative genomic hybridization (CGH) and other methods for determining variation in genotype have similar applications in providing outcome predictors (Jain et al. 2001) and indicating candidate genes in carcinogenesis (Daigo et al. 2001; Pinkel et al. 1998; Pollack et al. 1999).

The possibility of measuring gene expression simultaneously for many thousands of genes represents a challenge in terms of analysis and interpretation. One aim is the identification of genes whose expression levels are associated with outcome. In addition to clinical use as prognostic markers, such genes may be central factors in the mechanisms causing the observable variation in outcome. A number of approaches have been proposed and employed in ranking genes according to predictive strength in univariate contexts (Dhanasekaran et al. 2001; Tusher et al. 2001). Important genes with independent prognostic value can be sought by estimating their relative importance in multivariate classification settings (Khan et al. 2001) through the employment of strategies similar to the variable selection strategies used in regression analyses. Survival analyses include information regarding the time elapsed from diagnosis to the manifestation of an event of interest, such as death or metastasis. Many previous survival time studies of gene expression have used clustering methods based on "global" expression profiles to group patients (Alizadeh et al. 2000; Garber et al. 2001; Sorlie et al. 2001) and subsequently estimated the significance of these groups by the logrank test (Altman 1991). Despite its applicability to groups defined by the expression of single genes (Dhanasekaran et al. 2001), to our knowledge, the logrank test has not been used previously on a large scale to estimate significance on a per gene basis from microarray data.

We wanted to investigate, directly and individually, each gene with respect to the possible significance of its expression for survival. We carried out a systematic analysis of association between the expression of 8024 genes and patient survival by using data from a previously published microarray data set of breast tumours (Sorlie et al. 2001). We linked the genes, represented by their expressed sequence tag (EST) clone probes used on the microarrays, to their respective chromosomal positions. By comparison with publicly available CGH data, we were able to locate genes whose expression was highly associated with survival in a number of previously reported amplicons, thus relating these amplicons to survival. To aid the interpretation of our results, we used PubGene webtools (Jenssen et al. 2001).

Materials and methods

Nomenclature

Gene symbols used in this article follow the recommendations of the HUGO Gene Nomenclature Committee (Povey et al. 2001).

Gene expression data and preprocessing

cDNA microarray data and clinical information, including survival status and survival time in months, were downloaded from Sorlie et al. (2001). Survival data were available for 76 patients. Four of these patients were excluded, because the microarray data were from samples obtained after treatment. We used only before treatment data and extracted gene expression ratios from 72 raw data files (as had been exported by Sorlie et al. 2001, from ScanAlyze).

Data extraction was slightly complicated by the microarray analyses having been carried out with microarray slides from six different print-batches with slightly different clone sets. Since most of the arrays were from a batch named "svcc8k", we used the print layout from this batch as the "master" and identified spots on arrays from the other batches having the same clones as spots on the svcc8k arrays. Some of the clones had been printed in different multiples on the various array batches. This ambiguity was resolved in the following way. Let spots i_1, ..., i_n, be the number of spots for a given clone on the svcc8k arrays; let n' be the minimum number of times that the clone had been spotted on any of the arrays in all array batches. Data were then extracted from the svcc8k arrays from spots $i_{1'}$, ..., $i_{n'}$, and for any of the other arrays, we extracted data from spots j_1, ..., $j_{n'}$, where j_1, ..., j_m are the spots on the other array containing the same clone; m≥n'. Data across arrays were then matched such that, for a given clone, data from spot i_1 was matched with data from j_1, i_2 with j_2, and so on up to n'. This procedure ensured that the total data set had the same number of observations for each clone regardless of which array batch that had been used for a given patient.

The total number of "spots" that could be traced in this way was 8024. Thus, the total size of our gene expression dataset was 8024 rows (clones/genes) and 72 columns (patients). From this data set, we excluded all measurements from spots that had been manually flagged or where the estimated background-corrected signal intensity was non-positive in either channel.

Survival analyses

Logrank tests

The logrank test was implemented in Matlab (MathWorks, Natick, Mass., USA; http://www.mathworks.com). Correctness of the implementation was verified by using a commercially available implementation in SPSS (SPSS, Chicago, Ill., USA; http://www.spss.com). We wrote a Matlab function to run the implemented logrank test in batch-mode for all genes in the data set. In addition, we wrote a Matlab function implementing an automated search for the most significant grouping (also called discretization or binning) of a factor variable with respect to a given set of survival data. Briefly, the search tries all possible groupings and returns the one giving the most significant association with survival for a given variable. The automated search was essentially exhaustive, but when performing the analyses, we constrained the search space in two ways: (1) we only allowed the program to try two or three groups; (2) we required the smallest group in a given discretization to have a certain percentage of patients. In this study, for two and three groups, 10%and 8%, respectively, were used. The first constraint made the program only look for categorizations of gene expression into a dichotomous (low=1 and high=2) or trichotomous (low=1, medium=2, and high=3) scale. The second constraint made the program avoid reporting highly unbalanced groupings as being significant, as they would not be considered clinically useful.

We performed the logrank test for all 8024 rows in the gene expression dataset and for the clinical variables provided as supplementary information by Sorlie et al. (2001). Survival status and survival times were obtained from the same source. As identifiers for the genes, we used the first available gene symbol, UniGene cluster ID, or GenBank accession number. Following Sorlie et al. (2001), we encoded only survival status "Dead of Disease" (DOD) as an event. Survival status "Dead of other Causes" (DOC) was treated as a censored observation. Missing values were handled by simply omitting, from the analysis of a gene, those patients corresponding to arrays where the gene had not been measured. In this screening analysis, we required the smallest groups to contain at least 10% or 8% of the remaining patients when the expression levels were grouped into two or three groups, respectively. As a correction for multiple testing, we used the Bonferroni adjustment to find the critical *P*-value required for simultaneous significance; the corrected *P*-value threshold was α/8024, or 6.23E-06 for α=0.05.

Kaplan-Meier plots were created in SPSS.

Cox regression

To select gene expression patterns to use as variables in multivariate analyses, we constrained the discretizaton procedure to dichotomize the log-ratios and further required the smallest group to contain at least 30% of the patients. This resulted in 12 gene expression patterns meeting the criterion of simultaneous significance. There were two copies of GATA3, corresponding to two distinct spots on the arrays. We used the copy determined as most significantly associated. We also excluded four ESTs from consideration in Cox regression analyses. Four of the six clinical parameters had significant prognostic value in univariate analysis (ER-protein and TP53-type, tumour and grade). Thus, seven gene expression patterns were categorized as described, and four clinical variables were then analyzed by Cox regression analysis. We first estimated relative risks for these variables by univariate analyses and then used forward and backward variable selection to obtain a multivariate model. All Cox regression analyses were carried out with SPSS.

Chromosome data

We downloaded chromosome positions for chromosome-mapped ESTs from the Human Genome Working Draft (August 6, 2001 freeze; http://genome.ucsc.edu; UCSC). *P*-values from the logrank analyses were obtained for all IMAGE clones in the data set and mapped onto their respective positions on all 24 chromosomes by using the information from UCSC. We also downloaded chromosome positions for all mapped cytobands from the same version of the Human Genome Draft. Plots with chromosome-mapped negative logarithms (base 10) of *P*-values and cytobands were created in Matlab.

CGH comparisons

We downloaded previously reported amplicons in breast cancer cell lines from the online supplementary information from Daigo et al. (2001) and mapped the maximum negative logarithms (base 10) of *P*-values of the genes contained within the chromosomal bands indicated for the amplicons (Daigo et al. 2001), and one subband to either side of the region, allowing for mapping uncertainty of genes and amplicons.

PubGene analysis

The PubGene Gene Expression Analysis tool (http://www.pubgene.org/) was used to search for literature gene-networks dominated by survival-significant genes. Instead of gene expression ratios, we used negative logarithms (base 10) of *P*-values as input for the PubGene tool. A tab-delimited text-file with two columns, one with official gene symbols and one with negative logarithms, was prepared and submitted to the online analysis. To score gene neighborhoods, we used the parameter settings "Score-depth" =3, "Neighborhood size" =25, and "Up-regulation" as score criteria. In this case, "Up-regulation" corresponded to high negative logarithms, i.e., low *P*-values. For "Calculation scheme", we used both "By individual gene" and "By gene associations".

Results

Genes ranked by optimized significance to survival

We performed the logrank test on all samples based on a scheme often employed in Northern analysis, i.e., grouping patient samples into either low, median, or high expressor categories for a given gene. With this approach, optimal *P*-values for all genes were obtained and ranked. We found 95 genes whose expression was simultaneously significantly associated with survival, by using a correction for multiple testing (see Table 1). Kaplan-Meier plots for GRO1 and SPTBN2, the highest ranked gene and the gene ranked as number 95, respectively, are shown in Fig. 1. Both genes provide a satisfactory separation of patients with good outcome from those with poorer outcome in terms of survival status and duration. Of the 8024 genes, as many as 278 genes had a *P*-value of 1E-04 or lower. The complete list of genes with corresponding *P*-values is available as supplementary information.

Clinical association

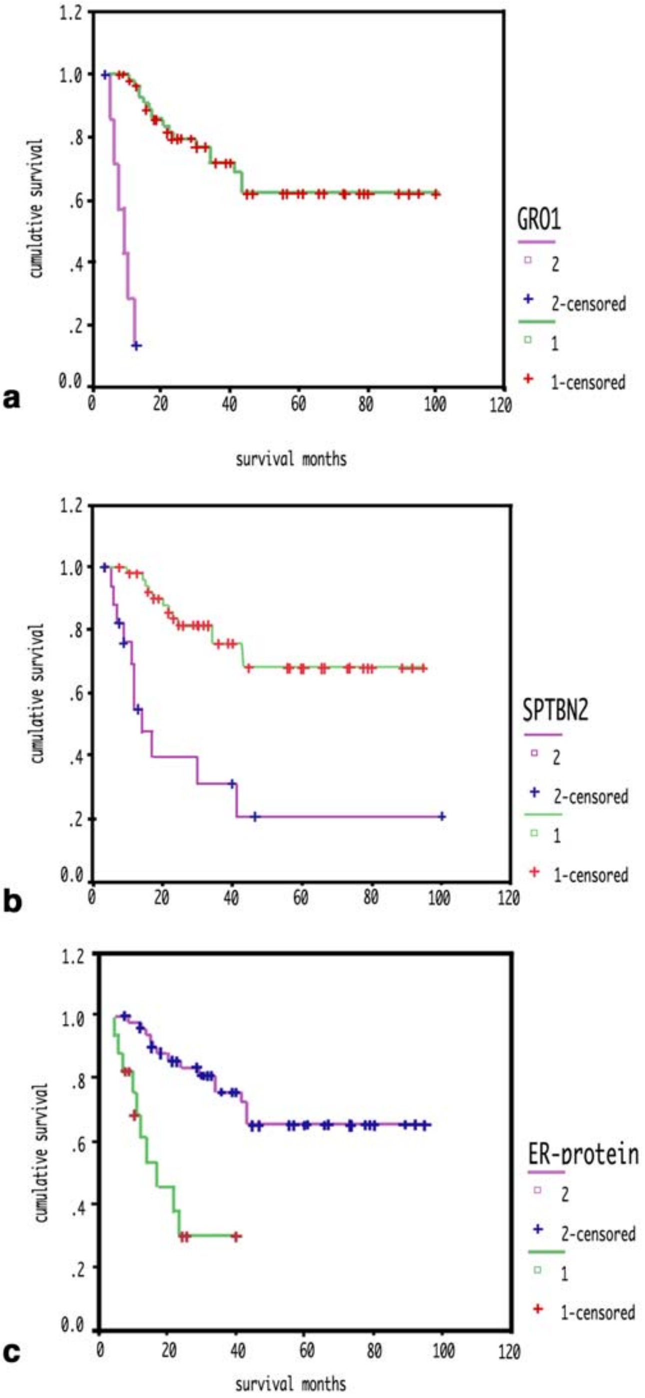

Fig. 1 Kaplan-Meier plots. Survival functions for patient groups based on gene expression of GRO1 (**a**) and SPBTN2 (**b**) show that both genes had gene expressions that could distinguish patients with significantly different prognoses. A Kaplan-Meier plot for the clinical variable ER-protein status is shown in (**c**)

As the first step in investigating the clinical association of the 95 highest ranked genes, we determined the relative values of gene expression associated with a good prognosis, for each of these genes (Table 1). We examined the biomedical literature to evaluate the extent of association to cancer previously reported for these genes. A summary of the genes in this highest scoring list with respect to association to cancer and suggested cancer prognosis mark-

Table 1 List of 95 genes having a *P*-value significant at α=0.05, when corrected for multiple testing. The total number of genes analyzed with the logrank test was 8024. Accession number and gene name (or Unigene cluster), when available, are shown for each gene expression pattern

Gene	Accession no.	*P*-value	Good[a]	Number of groups[b]
GRO1	W42723	4.4 E-16	Low	2
GATA3	R31441	3.8 E–14	High	3
DPP6	W96197	2.0 E-12	High	3
KNSL5	AA452513	1.8 E-11	Low	3
ACADSB	H95792	2.7 E-11	High	2
MGC11352	AA128362	5.2 E-11	Low	3
CDK8	R59697	7.9 E-11	Low	3
TNA	W73889	1.3 E-10	High	3
BTN3A3	AA478585	5.0 E-10	Low	3
Hs.8236	AA029948	7.1 E-10	High	2
Hs.27475	W56526	9.4 E-10	High	2
Hs.287827	R53330	1.1 E-09	Low	2
Hs.349177	N35341	1.2 E-09	High	3
SCAND1	W69094	3.9 E-09	Low	2
IGFBP2	H78560	6.7 E-09	High	3
GW128	H62527	6.9 E-09	High	3
GPX1	AA485362	9.3 E-09	High	3
RPS6KB2	AA284234	1.2 E-08	Low	3
MSE55	H73234	1.4 E-08	Low	3
IFNGR1	H11482	1.5 E-08	Low	3
GATA3	H72474	1.9 E-08	High	3
TMSB10	AA486085	1.9 E-08	Low	3
POLYDOM	R33004	2.0 E-08	High	3
HMGIY	AA448261	3.2 E-08	Low	2
TRPC1	AA017132	4.2 E-08	High	2
SLC7A5	AA419176	4.5 E-08	Low	2
ACK1	AA427891	4.9 E-08	High	3
CCNE1	T54121	8.4 E-08	Low	2
ZYX (FLT1)	AA058828	8.7 E-08	High	2
S100P	AA053016	8.8 E-08	Low	2
KIAA0212	AA630346	9.6 E-08	High	3
Hs.250535	AA428477	1.3 E-07	High	2
CSNK2A2	AA054996	1.4 E-07	Low	2
RARRES3	W47350	1.4 E-07	High	3
MGC2747	R91577	1.4 E-07	High	2
RNPC1 (HSRNASEB)	AA459363	1.4 E-07	Low	2
CPA4	AA016234	1.5 E-07	Low	2
H105E3	AA436425	1.6 E-07	Low	2
OIP2	N50079	2.0 E-07	Low	3
SLPI	AA026192	2.9 E-07	Low	2
KIAA0290	AA400186	3.8 E-07	Low	3
ERO1L	AA186803	4.5 E-07	Low	3
MPI	AA482198	4.7 E-07	High	3
MYBL2	AA456878	5.1 E-07	Low	2
LOC51312	H40448	5.3 E-07	Low	2
E2-EPF	AA464019	5.6 E-07	Low	3
VCY	AA406064	5.8 E-07	Low	2
ESR1	AA291702	5.8 E-07	High	3
PPP1R14C(NY-BR-81)	R18901	6.4 E-07	Low	2
SCYB14	W72294	7.4 E-07	High	3
MYB	N49284	9.5 E-07	High	2
PDE7A	H65033	1.1 E-06	Low	2
Hs.34054	R86669	1.1 E-06	High	3
ID3	AA482119	1.1 E-06	Low	2
Hs.167598	AA453470	1.2 E-06	High	2
CRR9	H84443	1.3 E-06	Low	2
PECI	AA620556	1.3 E-06	High	3
BCL2	W61100	1.4 E-06	High	3
KIF3C	AA446908	1.5 E-06	Low	2
MPP2	R60019	1.6 E-06	High	2
RNPC1 (HSRNASEB)	AA459363	1.7 E-06	Low	2
ETFB	T62040	1.8 E-06	Low	2
MAD4	AA447515	2.2 E-06	High	3
KYNU	H87471	2.2 E-06	Low	2
Hs.293737	H94897	2.2 E-06	High	3
Hs.273483	AA620802	2.4 E-06	High	2
RNF30 (MURF)	AA428229	2.4 E-06	High	2
POV1	T64312	2.4 E-06	High	3
FLJ20568	AA454563	2.5 E-06	High	3
FLJ11795	AA459693	2.5 E-06	Low	3
Hs.117078	AA436564	2.6 E-06	High	3
Unknown	T60075	2.6 E-06	Low	2
KRT5	W72110	2.7 E-06	Low	2
Hs.71331	AA130595	2.7 E-06	Low	2
PLAGL1	AA463204	2.8 E-06	Low	2
KIAA0040	AA465478	2.8 E-06	High	2
HERC3	AA282253	3.1 E-06	High	2
SNRPA1	AA122272	3.4 E-06	Low	2
KPNA6	AA009595	3.4 E-06	High	3
EXTL3	AA431408	3.6 E-06	High	2
TUBA1	AA180742	3.8 E-06	Low	2
HEAB	AA700336	4.1 E-06	Low	2
GDI2	R92806	4.2 E-06	Low	3
RIL	AA045672	4.3 E-06	Low	3
TUBB5	H37989	4.5 E-06	Low	3
CNN3	AA043227	4.5 E-06	Low	2
Hs.22483	W93688	4.6 E-06	High	2
FLJ11320	T74039	4.9 E-06	Low	2
KIAA0948	AA676387	5.0 E-06	Low	2
MRPS12	R23752	5.0 E-06	Low	2
YWHAG	AA432085	5.1 E-06	Low	2
KIAA0173	AA682815	5.5 E-06	Low	2
HGS	N20338	5.5 E-06	Low	2
HTPAP	T48411	6.0 E-06	High	3
SPTBN2	H28127	6.0 E-06	Low	2

[a] Good prognosis is associated with each category of relative expression level for the given gene

[b] Number of groups used to derive the resulting *P*-value, as determined by automated search

ers are given in Table 2. Briefly, 50% of the genes in this list were identified as previously described in the literature as associated with cancer, whereas 18% could be associated with the use or suggested use as prognostic markers for cancer. According to the logrank analyses, the most significantly associated expression pattern was for

Table 2 Summary statitics for the detectable association between the biomedical literature and the list of genes given in Table 1. All results are based on manual literature search in Medline (http://www.medlineplus.gov/)

Association	Number	% of informative
Proven or possible prognostic marker in cancer	13	18.8
Implicated with cancer changes	22	31.9
No detectable link	34	49.3
EST or not characterized	26	37.7
Sum, any association to cancer	35	50.7
Sum, after EST exclusion	69	100.0
Sum total	95	

the chemokine GRO1, with a P-value of 4.44E-16. This protein has previously been reported to be associated with metastasis in a murine SCC model (Loukinova et al. 2000). Among the genes linked to prognostic use in the list, GATA3 occurred twice as a consequence of appearing several times on the array with different clones. As might have been expected, ESR1, KRT5, and tetranectin were also identified.

Cox regression analysis based on expression patterns for seven genes and four clinical variables resulted in a multivariate model with four variables: the clinical variable ER-protein and the three gene expression patterns of EXTL3, ID3, and YWHAG. These variables are independent in the sense of simultaneously contributing to explaining outcome in the analyzed patient panel. Estimates of relative risk and significance are shown in Table 3, which also includes results from the univariate Cox regression analyses of these variables.

Comparison of chromosome plots with CGH in detecting "survival amplicons"

Some previously identified breast cancer amplicons were analyzed. We first examined the chromosome 17q12-q21 ERBB2 amplicon region (van de Vijver et al. 1987), previously investigated by cDNA microarrays in breast cancer cell lines (Kauraniemi et al. 2001) and breast cancer patients (Bertucci et al. 2000); a plot of chromosome 17 with P-values for mapped ESTs is shown in Fig. 2a. Within the amplicon region as defined by amplicon mapping by expression arrays (Kauraniemi et al. 2001), the highest ranked candidate gene with respect to survival association is an uncharacterized gene, Hs.28893, with a P-value of 8.02E-4, whereas the ERBB2 has a P-value of 2.72E-2 and GRB7 a P-value of 2.27E-2. The Hs.28893 gene has a central location within the amplicon. A poor prognosis was associated with a low expression of all of these genes.

Amplification of 8q24 is a predictor of poor prognosis, in contrast to reports for the 17q12-q21 amplicon (Jain et al. 2001). There is one candidate in this region, NDRG1, with a P-value of 2.56E-5 (Fig. 2b). NDRG1 is a cytoplasmic protein involved in stress responses, hormone responses, cell growth, and differentiation. Another gene known to be involved in many cancer types, MYC, is also present in this cytoband, but this gene correlates less well, with a P-value of 7.05E-2. Several other genes in this region displayed correlation to survival, albeit at smaller significance, including Hs.260644 (2.20E-4), MAPK13 (1.09E-4), and HT002 (4.26E-4). High expression of NDRG1, MAPK13, and HT002 predicted a poor prognosis, this was also the tendency for MYC.

To correlate general breast cancer CGH information with cDNA microarray-derived survival values, we downloaded a digitally available breast cancer CGH amplicon set in breast cancer cell lines (Daigo et al. 2001) in order to identify the maximum log P-values obtained from the logrank list for these amplicons (data not shown). The amplicon list contained 42 non-overlapping regions. Within 14 (33%) of these, we found one or more genes

Table 3 Cox regression analysis. Because of non-overlapping cases with missing-values, there were 56 patients available for multivariate analysis. Age, node, and metastatis status were excluded from consideration in multivariate analysis because of their weak association to survival in univariate analysis. We used a discretized version of age, obtained by splitting patients into two groups, corresponding to younger or older than 40 years, respectively. Variables in *bold* were input to the multivariate analysis

Variable	Cases	Univariate analysis			Multivariate analysis		
		Exp(B)	95% CI	P-value	Exp(B)	95% CI	P-value
Age	72	1.3	0.39–4.4	0.67			
Tumor	71	1.9	1.0–3.5	0.050			0.14
Node	72	1.1	0.63–1.8	0.82			
Grade	72	2.7	1.3–5.6	0.007			0.45
Met	71	1.7	0.23–13	0.61			
TP53 variant	65	3.4	1.4–8.2	0.008			0.51
ER-protein	70	0.18	0.073–0.42	<0.001	0.13	0.04–0.42	0.001
FLT1	72	0.13	0.055–0.31	<0.001			0.41
GATA3(1)	72	0.14	0.060–0.32	<0.001			0.54
CCNE1	64	8.4	3.3–22	<0.001			0.62
SLC7A5	72	6.0	2.6–14	<0.001			0.73
EXTL3	72	0.17	0.073–0.40	<0.001	0.27	0.083–0.86	0.026
YWHAG	72	6.3	2.6–15	<0.001	3.4	1.03–11	0.044
ID3	72	5.9	2.5–14	<0.001	9.4	3.0–29	<0.001

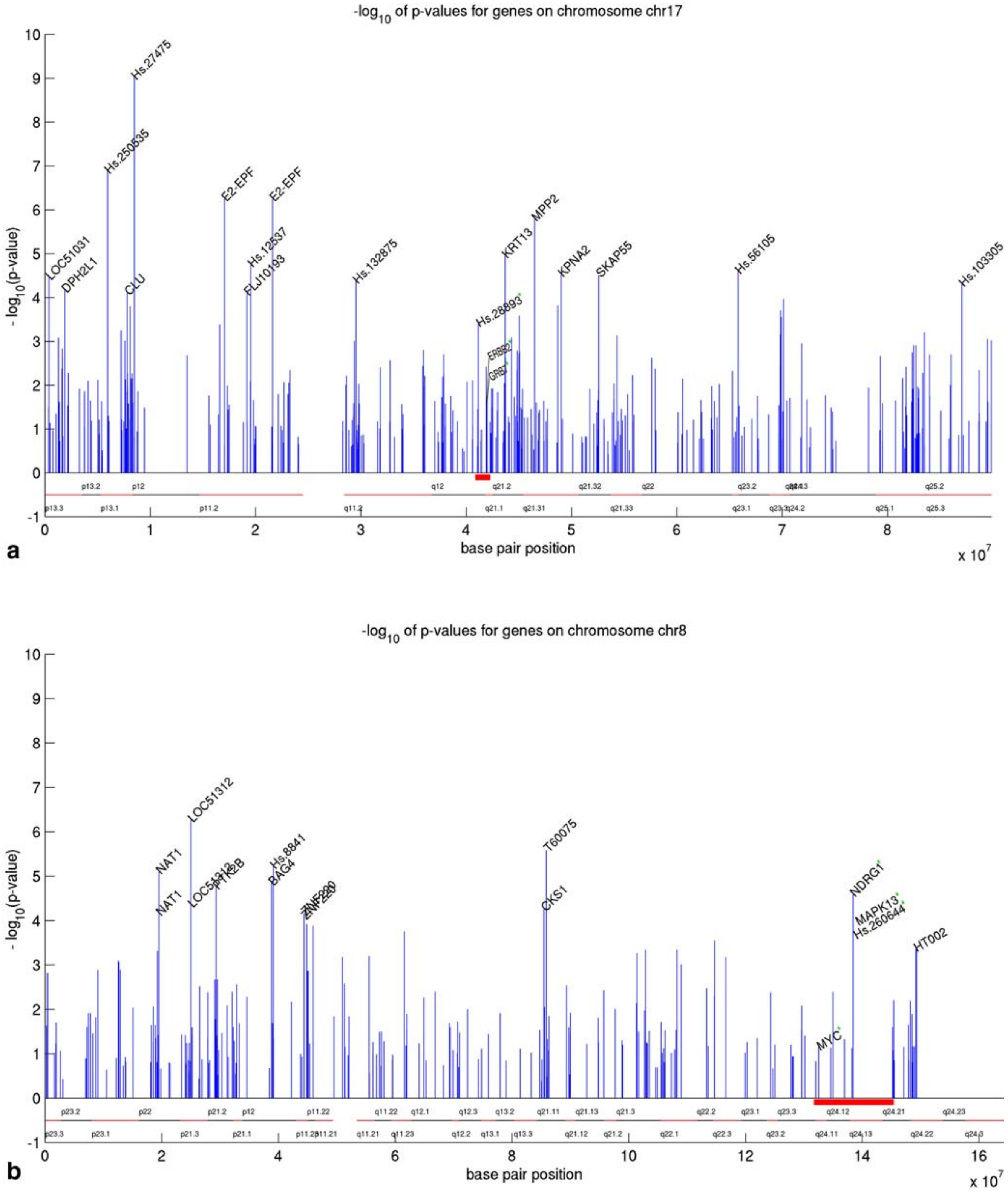

Fig. 2a, b Chromosome plots with log-transformed *P*-values from logrank tests. Each *bar* is located at the start of the EST used to detect expression of a given gene. The *height* of the *bar* is proportional to the log-transformed *P*-value. Cytobands (to scale) are shown *bottom*. Cytobands with positive staining are given in *red* and bands with negative staining are colored in *black*. *Thick horizontal red lines* Amplicons, *green asterisks* genes discussed in the text

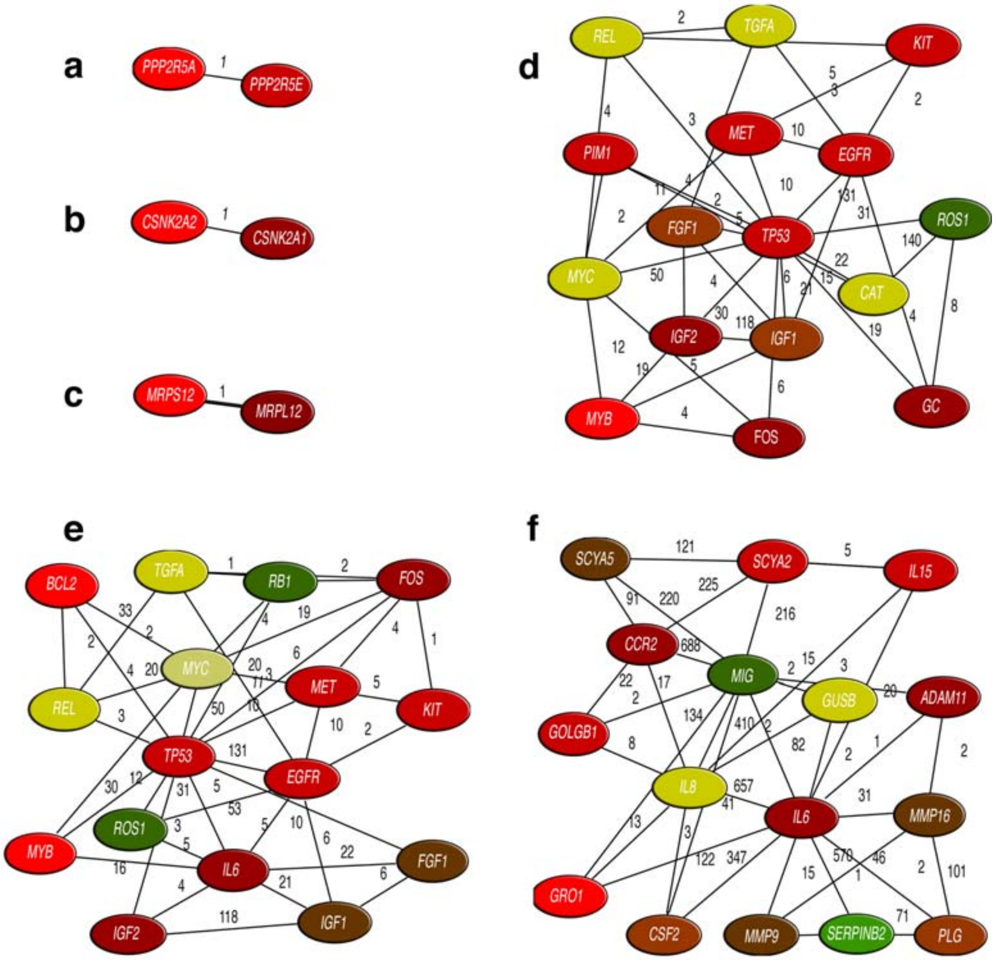

Fig. 3 Literature gene networks from PubGene analyses. Literature networks were ranked by scores based on log-transformed *P*-values from log-rank tests. The three top-scoring networks from search by gene-values contained two genes each (**a–c**). The three top-scoring networks from search by gene associations were the literature neighborhoods of PIM1 (**d**), REL (**e**), and GRO1 (**f**). *Red* Low *P*-values, *green* high *P*-values. The *degree of redness* is on a linear scale in log-space. *Numbers* Co-occurrence counts from the literature

from our list with the 95 most significant expression patterns. When each amplicon region was extended also to include the two cytobands immediately before or after it, the number of regions containing one or more of genes with the most significant expression patterns increased to 23 (55%).

PubGene analysis of log-transformed *P*-values

We have previously developed a method for literature mining, viz., a set of tools collectively termed PubGene (Jenssen et al. 2001). We submitted log-transformed *P*-values for 5449 unique gene symbols to examine the network for informative clusters of genes with high local neighborhoods contributing to low *P*-values (see above).

Interestingly, in an analysis for high-scoring networks based on direct gene scores, two of the most high-scoring gene clusters only consisted of two members where both were regulatory subunits of the same protein. One was phosphatase 2 (PPP2R5A and PPP2R5E; Fig. 3a). These subunits had separate *P*-values of 1.34E-5 and 1.17E-3, respectively; a poor prognosis was associated with high expression in both cases. The other was casein kinase 2 (CSNK2A1, CSNK2A2; both genes representing subunits of casein kinase 2 known to be upregulated in many human cancers; Fig. 3b). High expression was associated with a poor prognosis for both subunits of casein kinase 2. Moreover, two ribosomal proteins (MRPS12, MRPL12; high expression associated with poor prognosis in both cases) are also found in a small cluster (Fig. 3c). Of these, MRPL12 has been associated with colon cancer (Marty et al. 1997). Two other small networks are also highly ranked, one including KNSL5 (ABL1, KDR, MET, PLK, KNSL5) and one including SCAND1 (MYB, CRAT, ZNF42, SCAND1).

The top-scoring gene when we performed an analysis based on gene association, PIM1 (Fig. 3d), has previously been reported as being associated with survival for prostate cancer (Dhanasekaran et al. 2001). We found high expression of PIM1 to be associated with poor prognosis. The REL gene, a part of the NF-kappaB protein that regulates genes controlling cell proliferation, survival, and transformation ranked in the second cluster (Romieu-Mourez et al. 2001; Fig. 3e). The third ranking cluster featured GRO1 as the central gene, i.e., the highest ranked gene (Fig. 3f). The cluster consisted of 25 genes (MCP, PTK2, TIE, CALR, SCYA2, PLG, MMP9, EGFR, ROS1, REN, SERPINB2, SCYA5, CCR2, IL6, MPO, IL8, CSF1, CSF2, MMP16, GUSB, ADAM11, IL15, GOLGB1, MIG, in addition to GRO1), indicating necrosis, transcription, and chemotaxis as the main ontology terms.

Discussion

We have developed a new set of tools for analyzing associations between gene expression and survival in studies where the expression of a large number of genes is assessed. A routine was made for the analysis of the prognostic value of a large number of parameters employing the logrank test. During categorization of the parameters, the sizes of the (two or three) groups were varied to maximize the prognostic value. The parameters with the highest prognostic values could be entered into multivariate Cox regression analysis together with, for example, clinical parameters for assessment of independency. The negative logarithms of the *P*-values from the logrank tests were further plotted according to the chromosomal location of the genes, yielding a "prognostic value karyotype". Several of the genes, the expression of which had high prognostic value, were found in regions frequently amplified in breast tumors. Finally, the negative logarithms of the *P*-values of all the scored genes on the array were fed into our literature cluster analysis program (Jenssen et al. 2001). The program returned clusters of genes with prognostic value, pointing to pathways (consisting of coregulated genes) associated with survival.

In this study, we chose to optimize for prognostic value based on the commonly used discretization in three relative expression groups (or two wherever this resulted in better *P*-values). Other choices, such as discretization in two groups on either side of the median value could have been applied. We chose the former in order to maximize the number of candidate genes, with a conservative Bonferroni adjustment to correct the *P*-value threshold for significance. This adjustment is approximately exact when the variables are independent but tend to be overly conservative when they are not. In effect, the test should result in few false positives but is likely to give a number of false negatives. We did not consider it necessary to correct for multiple testing with respect to the different categorizations tried for each gene, as this procedure was only used to determine the thresholds for the expression levels.

The genes having expression patterns with significant association to survival all represent highly relevant candidates for examination as prognostic markers, as evaluated in univariate analysis. The resultant proteins apparently serve varied functions in cellular processes. Two of these are members of the heat-shock protein 40 family of proteins, containing the DNAJ domain, some are involved in fatty acid metabolism, and others are involved in tubulin-related or inflammatory processes. Both up- and down-regulated genes relative to prognostic indication have been identified, and both types are associated with their involvement in cancer or as suggested markers as indicated in Table 2. About half of the genes previously described in the literature could easily be identified as having been associated with cancer, underscoring the relevance of the approach. In this report, we have not assessed the underlying quality of the data analyzed, and the number of patients is also relatively small, preventing an extensive survey of all relevant genes as candidate markers.

After a suitable reduction of the number of genes with prognostic value, the independency of these parameters and that of several clinical parameters with prognostic value could be assessed with Cox multivariate regression analysis. Because the number of patients included in the material analyzed here was low, only a limited number could be included in regression analysis. We limited the number of genes included by requiring that expression patterns remained significant when the categorization groups were no smaller than 30%. When the variables were checked for correlations (using Kendall's Tau-b), we noted that TP53-type was significantly ($P<0.001$) correlated to the expression of YWHAG, one of the variables included in the model. The four gene expression patterns that were not included all showed significant ($P<0.001$) correlation with the ER-protein assay. The expression of YWHAG was, however, also significantly ($P\leq0.001$) correlated both to ER-protein and the expression of EXTL3, partly explaining why it was the least significant variable in the model. No significant correlations were found between the other variables included.

High-throughput analytical methods for amplifications and deletions in cancer by using array-CGH mapping with genomic arrays or cDNA arrays have recently been developed (Lucito et al. 2000). At the RNA level, cDNA arrays may also reveal amplifications, but over-expression is not necessarily restricted to the amplified gene. These methods give much higher resolution than traditional karyotyping and CGH. To demonstrate the usefulness of applying the logrank results in identifying candidate prognostic genes in frequent cancer amplicons, we examined the well-known ERBB2 amplicon. We did not find any genes in this amplicon whose over-expression could be associated with poor prognosis. This was in agreement with previous CGH results (Jain et al. 2001), where the presence of the 17q12-q21 amplicon was not found to result in a poor prognosis. However, other reports have found ERBB2 amplification and/or high ERBB2 protein expression to be associated with poor prognosis (Slamon et al. 1987; Wright et al. 1989). The amplification of chromosome 8q24 (by CGH) was previously reported to be a predictor of poor prognosis (Jain et al. 2001). We identified NDRG1 at 8q24 as resulting in poor prognosis when up-regulated. This gene is therefore a very interesting candidate gene for further examination. Another gene in the same chromosome band, MYC, is frequently amplified in breast cancer. However, MYC could not explain the poor prognosis of the patients with 8q24 amplification by CGH. The lack of prognostic value for MYC expression indicates that MYC is not the important gene leading to selection for amplification of 8q24 during carcinogenesis. Alternative candidates obviously include the NDRG1 gene.

The PubGene analysis permits a novel type of analysis of survival associations with the biomedical literature. Subjecting log-transformed *P*-values to a PubGene analysis is carried out essentially to identify smaller regions of a large network as being heavily affected, compared with all other areas of a global literature-based network. This analysis may be focused on very tight literature clusters by evaluating direct links or, in a broader context, by using gene associations as the scoring parameter. When analyzing for tight clusters, two proteins (protein phosphatase 2 and casein kinase 2) have been identified by their subunit contributions. This clearly demonstrates the relevance of the approach, as both subunits would by themselves fall outside of the 95 genes most highly associated with survival. Protein phosphatase 2 is suggested to be an endogenous regulator of inflammatory cell signalling (Shanley et al. 2001). Casein kinase 2 is upregulated in many human cancers (Landesman-Bollag et al. 2001), and a poor prognosis has also been observed with increased expression for both subunits. Recently, primary human breast cancer specimens that displayed aberrant constitutive expression of NF-kappaB/REL were found to exhibit increased casein kinase 2 and/or IKK kinase activity (Romieu-Mourez et al. 2001), indicating the important oncogenic role for casein kinase 2 in breast cancer. When using the alternate approach of identifying larger, less well-defined networks, several relevant genes were identified.

The data set used in this study has previously been analyzed by Sorlie et al. (2001) by the so-called SAM method in conjunction with hierarchical clustering; with SAM, the authors identified a list of 264 genes with expression associated to survival (SAM264 gene list). Although, Sorlie et al. (2001) used a different patient subset for their survival analyses, we have noticed a significant overlap between the SAM264 gene list and the genes that we have found to be the most associated with survival. We have found 29 SAM genes in our list, representing a 10-fold increase over the number expected by chance (P<1E-10). Recently, van't Veer et al. (2002) published a list of 231 genes correlated to patient outcome. In this case, patient outcome was defined as the occurrence of metastasis, and among the 174 genes (of the 231 possible) that were also in our data set, we found only FLT1 and PECI in the list with the 95 most significant from our analysis.

In conclusion, the genes suggested by this analysis as having prognostic value may serve as candidates for more detailed examination by using a larger number of clinical samples and verification at the protein level. We believe this study shows that genome-wide screens for survival markers at the mRNA level are both feasible and sensible. Moreover, our approach can be applied to other clinical studies, including other cancer types and other outcomes as events in the survival time analyses.

Acknowledgements We thank T. Sorlie et al. and Y. Daigo et al. for making their data publicly available and thus enabling our study. We thank Pål Møller for comments and discussion. This work was supported in part by the Norwegian Cancer Association (DNK).

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) 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

Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London

Bertucci F, Houlgatte R, Benziane A, Granjeaud S, Adelaide J, Tagett R, Loriod B, Jacquemier J, Viens P, Jordan B, Birnbaum D, Nguyen C (2000) Gene expression profiling of primary breast carcinomas using arrays of candidate genes. Hum Mol Genet 9:2981–2991

Daigo Y, Chin SF, Gorringe KL, Bobrow LG, Ponder BA, Pharoah PD, Caldas C (2001) Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon profiling of breast cancers : a new approach for the molecular analysis of paraffin-embedded cancer tissue. Am J Pathol 158:1623–1631

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686

Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412:822–826

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, Rijn M van de, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I (2001) Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci USA 98:13784–13789

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J (2001) Gene-expression profiles in hereditary breast cancer. N Engl J Med 344:539–548

Jain AN, Chin K, Borresen-Dale AL, Erikstein BK, Eynstein Lonning P, Kaaresen R, Gray JW (2001) Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival. Proc Natl Acad Sci USA 98:7952–7957

Jenssen TK, Laegreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28:21–28

Kauraniemi P, Barlund M, Monni O, Kallioniemi A (2001) New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. Cancer Res 61:8235–8240

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7:673–679

Landesman-Bollag E, Romieu-Mourez R, Song DH, Sonenshein GE, Cardiff RD, Seldin DC (2001) Protein kinase CK2 in mammary gland tumorigenesis. Oncogene 20:3247–3257

Loukinova E, Dong G, Enamorado-Ayalya I, Thomas GR, Chen Z, Schreiber H, Van Waes C (2000) Growth regulated oncogene-alpha expression by murine squamous cell carcinoma promotes tumor growth, metastasis, leukocyte infiltration and angiogenesis by a host CXC receptor-2 dependent mechanism. Oncogene 19:3477–3486

Lucito R, West J, Reiner A, Alexander J, Esposito D, Mishra B, Powers S, Norton L, Wigler M (2000) Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res 10:1726–1736

Marty L, Taviaux S, Fort P (1997) Expression and human chromosomal localization to 17q25 of the growth-regulated gene encoding the mitochondrial ribosomal protein MRPL12. Genomics 41:453–457

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. Nature 406:747–752

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23:41–46

Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H (2001) The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 109:678-680

Romieu-Mourez R, Landesman-Bollag E, Seldin DC, Traish AM, Mercurio F, Sonenshein GE (2001) Roles of IKK kinases and protein kinase CK2 in activation of nuclear factor-kappaB in breast cancer. Cancer Res 61:3810–3818

Shanley TP, Vasi N, Denenberg A, Wong HR (2001) The serine/threonine phosphatase, PP2A: endogenous regulator of inflammatory cell signaling. J Immunol 166:966–972

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235:177–182

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn M van de, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98:10869–10874

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116–5121

Veer LJ van't, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487

Vijver M van de, Bersselaar R van de, Devilee P, Cornelisse C, Peterse J, Nusse R (1987) Amplification of the neu (c-erbB-2) oncogene in human mammmary tumors is relatively frequent and is often accompanied by amplification of the linked c-erbA oncogene. Mol Cell Biol 7:2019–2023

Wright C, Angus B, Nicholson S, Sainsbury JR, Cairns J, Gullick WJ, Kelly P, Harris AL, Horne CH (1989) Expression of c-erbB-2 oncoprotein: a prognostic indicator in human breast cancer. Cancer Res 49:2087–2090