Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis

Mads Thomassen, Qihua Tan, Torben A. Kruse

Breast Cancer Res Treat (2009) 113:239–249
DOI 10.1007/s10549-008-9927-2

PRECLINICAL STUDY

Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis

Mads Thomassen · Qihua Tan · Torben A. Kruse

Received: 27 January 2008 / Accepted: 28 January 2008 / Published online: 22 February 2008

Abstract Breast cancer cells exhibit complex karyotypic alterations causing deregulation of numerous genes. Some of these genes are probably causal for cancer formation and local growth whereas others are causal for the various steps of metastasis. In a fraction of tumors deregulation of the same genes might be caused by epigenetic modulations, point mutations or the influence of other genes. We have investigated the relation of gene expression and chromosomal position, using eight datasets including more than 1200 breast tumors, to identify chromosomal regions and candidate genes possibly causal for breast cancer metastasis. By use of "Gene Set Enrichment Analysis" we have ranked chromosomal regions according to their relation to metastasis. Overrepresentation analysis identified regions with increased expression for chromosome 1q41–42, 8q24, 12q14, 16q22, 16q24, 17q12–21.2, 17q21–23, 17q25, 20q11, and 20q13 among metastasizing tumors and reduced gene expression at 1p31–21, 8p22–21, and 14q24. By analysis of genes with extremely imbalanced expression in these regions we identified DIRAS3 at 1p31, PSD3, LPL, EPHX2 at 8p21–22, and FOS at 14q24 as candidate metastasis suppressor genes. Potential metastasis promoting genes includes RECQL4 at 8q24, PRMT7 at 16q22, GINS2 at 16q24, and AURKA at 20q13.

Keywords Metastasis · Distant metastasis · Metastasis genes · Causal genes · Breast cancer · Somatic mutations · Copy number · Microarray · Gene expression profiling

Introduction

Breast cancer is the most common cancer among women and the leading cause of cancer related death. Metastasis is the main cause of death of the disease. Metastasis is believed to progress in a multi-step fashion including mutations in several genes. Somatic mutations inactivating or amplifying genes in tumors are often large genomic gains or losses. These aberrations can be identified with laborious techniques like Southern blotting and comparative genome hybridization (CGH). By array based CGH, large-scale experiments may potentially be performed with high throughput. However, large dataset with clinical outcome are not yet available for breast cancer, in CGH-studies.

Gene expression profiling has been used for classification of cancer outcome in several studies [1–9] showing that the overall gene expression pattern in primary tumors is associated with clinical outcome. However, from these studies it has not been possible to pinpoint the causal gene expression changes. It is our hypothesis that some of the

Authors' contributions M. Thommassen and T. A. Kruse designed the study, Q. Tan developed methods for statistical analysis and M. Thommassen performed data analysis.

Electronic supplementary material The online version of this article (doi:10.1007/s10549-008-9927-2) contains supplementary material, which is available to authorized users.

M. Thomassen (✉) · Q. Tan · T. A. Kruse
Department of Biochemistry, Pharmacology, and Genetics, Odense University Hospital and Human Microarray Centre (HUMAC), University of Southern Denmark, Odense, Denmark
e-mail: mads.thomassen@ouh.regionsyddanmark.dk

T. A. Kruse
e-mail: torben.kruse@ouh.regionsyddanmark.dk

Q. Tan
Institute of Public Health, University of Southern Denmark, Odense, Denmark
e-mail: qihua.tan@ouh.regionsyddanmark.dk

differentially expressed genes are causal for metastasis and that their changed expression level is due to somatic mutations whereas the changed expression of other genes is due to the direct or indirect influence of the causal genes. We expect that mutations causing metastasis are present in primary tumors and that they are not the characteristics of rare cells but of bulk of tumor. Furthermore, we anticipate that a fraction of the somatic mutations are large rearrangements, amplifications or deletions and that such regional allelic imbalance will influence not only the causal genes but most genes from this region. By analyzing the expression level over chromosomal regions our aim is to identify regions potentially harboring one or more genes with a causal effect on metastasis. For this analysis we have used two methods: gene set overrepresentation analysis of chromosomal regions and sliding mean analysis for validation and fine mapping. We expect some of the causal genes to be mutated by other mechanisms in some tumors. Our second aim is to identify such candidate causal genes by comparing the differential expression of individual genes in the identified regions.

We have investigated expression profiles along the chromosomes offering the possibility to observe combined effects of large somatic rearrangements, point mutations, epigenetic changes and gene regulation involved in metastasis. Chromosomal gains and losses are identified by summarizing effects from many genes of larger regions. Furthermore, single candidate genes causal for metastasis of the cancer cells harboring gains and losses are proposed because these genes display additional imbalanced transcription compared to surrounding genes indicating that expression of these genes are changes by several mechanisms.

Materials and methods

Datasets

Eight publicly available datasets were included in the analysis. These studies are performed with different platforms, different populations etc. as depicted in Table 1. The outcome differs in that local and regional recurrences are included in some studies. However, non-metastatic relapse constitute a minority of clinical cohorts. There may be a small overlap in the samples in the different dataset, e.g., samples from Uppsala in Sotiriou 2006 and Uppsala datasets, but the total number of different tumor samples is at least 1200.

The normalizations performed in the studies were retained because the authors found these methods optimal for the datasets and because gene set enrichment analysis was performed separately in each dataset.

Gene set enrichment analysis

GSEA v 2.0 [10] was used with positional gene sets delimited by cytobands downloaded from the Molecular Signature Database (MSigDb). The program ranks genes according to a signal to noise ratio defined as $(\mu_A - \mu_B)/(\sigma_A + \sigma_B)$, where μ is the mean and σ the standard deviation for the two classes A and B (metastasis and non-metastasis). When several probes recognized the same gene, median expression values was calculated using the "collapse to gene set" function. Gene sets represented by less than 15 genes in a dataset were excluded except for the Sotiriou 2003 dataset where this threshold was set to 10 genes because of the low number of genes on that chip.

The output is an enrichment score, describing the imbalance of gene expression in each gene set between metastasizing and non-metastasizing tumors. The enrichment score is normalized according to size of the gene sets. Then, gene sets were ranked according to the normalized enrichment score with gene sets upregulated in metastasizing tumors at the top and downregulated gene sets at the bottom.

Gene set enrichment meta-analysis

The ranked lists of gene sets from the eight datasets were integrated and the initial number of 386 positional gene sets in MSigDb was reduced to 103 gene sets passing the threshold (10 or 15 genes) in all datasets. For each dataset each gene set was assigned a ranking value from 1 to 103. The mean ranking value was calculated across the datasets and finally the gene sets were ranked according to this value. The significance of obtaining a certain mean ranking value was estimated by simulating random drawing of eight ranking values 10^6 times and calculating the mean each time. This calculation of *P*-value and estimation of false discovery rate (FDR) was performed in R environment (http://cran.r-project.org/). Gene sets with FDR values below 0.01 were considered significant.

To examine transcriptional phenomena covering more than 1 cytoband, 79 neighboring regions, i.e., two chromosomal consecutive gene sets, were examined by simulating the drawing of 16 ranking numbers (this time from 1–79) and comparing the mean of these with observed values. The 16 drawings correspond to the observation of 2 gene sets in eight datasets. The *P*-values were calculated and adjusted by FDR similar to the above-described method. The ranking numbers in this analysis (1–79) differs from ranking numbers in the analysis of individual gene sets (1–103) because a limited number of intrachromosomal haplotypes of neighboring regions can be generated.

Table 1 Characteristics of patients and platforms in included studies

Data	Chip	# probes (K)	Patients, country, nodal status[a]	Outcome[b]	Adjuvant systemic treatment[c]
HUMAC [1]	Spotted oligonucleotides	29	$n = 60$, DK N−, low-malignant	Metastasis	nil
Huang [2]	Affymetrix 95av2	12	$n = 52$, Taiwan N+	Relapse	ct
Sotiriou 2003 [3]	Spotted cDNA	7.6	$n = 99$, UK N+/N−	Relapse	et, ct
Sotiriou 2006 [4]	Affymetrix HG-133A	22	$n = 179$ S (Uppsala), UK N+/N−	dm	et
Rotterdam [5]	Affymetrix HG-133A	22	n = 286, NL N−	dm	nil
Amsterdam [6]	Rosetta	25	$n = 295$, NL N+/N−	dm	nil, ct, et
Uppsala [7]	Affymetrix HG133A + B	44	$n = 236$, S (Uppsala) N+/N−	Death from breast cancer	nil, ct, et
Stockholm [8]	Affymetrix HG-133A + B	44	$n = 159$, S (Stockholm) N+/N−	Relapse	nil, ct, et

[a] n, number of patients included; N+, positive nodal status; N−, negative nodal status; DK, Denmark; UK, United Kingdom; NL, the Netherlands; S, Sweden

[b] dm, distant metastasis

[c] ct, chemotherapy, et: endocrine therapy

Refining regions and candidate genes

To narrow down regions with possible metastatic impact identified by the gene set enrichment meta-analysis, ratios of differential expression (RDE) where calculated for each gene as RDE = μm/μn where μm and μn are the mean expression values of samples with poor and good outcome respectively. Three of the datasets, HUMAC, Amsterdam and Sotiriou are in log scale and inverse logarithm was calculated before calculation of RDE values. RDE values where scaled to obtain more comparable values between the datasets. Scaling compensate for very different amplitudes in the datasets due to different background correction and normalization methods. This was done as described by Yang et al. [11]. In brief, RDE was calculated for all genes in a given dataset. In each dataset RDE's were scaled so that median deviation from 1 reached the geometric mean of these medians in all eight datasets.

Annotation files with exact chromosomal positions were downloaded for commercial chips. Spotted chips were annotated using Gene Bank accessions and annotation files provided by NCBI (http://www.ncbi.nlm.nih.gov). Datasets where integrated using Microsoft Access with gene symbol as identifier. To make the scaled RDE values more easily interpretable, sliding means where calculated over 100 rows in the integrated file and plotted against the gene number in chromosomal order. Because of several missing values in the integrated file, resulting from different feature numbers and gene symbol annotation on the chips, the sliding mean spanned from average 36 genes on the low density chip used by Sotiriou 2003 till average 73 genes on the high density Affymetrix chip sets used in the Uppsala and Stockholm datasets. Finally, the sliding mean curves were used to pinpoint core regions with concordant tendencies of differentially expression in the majority of datasets.

To identify candidate genes that might be causal for metastasis we selected genes with imbalanced expression (RDE) in addition to the general tendency in the region determined as the sliding means at this position. Average difference in RDE for a gene i over up to eight datasets was calculated as $dRDE_i$ = mean ($RDE_i - sm_i$) where sm is sliding mean for gene i. *P*-values were calculated by Students *t*-test. The following criteria were used to select additionally regulated genes in core regions: $|dRDE| > 0.15$ and $P < 0.05$. The cut off value of |dRDE| corresponds to approximately 5 times mean observed RDE in core regions.

These *P*-values are based on up to eight observations, i.e., RDÉs and sliding means for the eight datasets. To take advantage of the large number of tumors in each dataset, *P*-values were also calculated for each gene within each original dataset by Students *t*-test.

Results

Gene set enrichment analysis

Data from more than 1200 breast cancer patients were collected (Table 1). Gene set enrichment analysis only identified few significant regions within each dataset (data not shown). However, by performing meta-analysis of gene sets ranked by normalized enrichment score, several gene sets turned out to have low ranking number in the majority of datasets indicating upregulation of gene sets in

metastasizing tumors compared to non-metastasizing tumors. Similarly, gene sets with a high mean ranking value indicated low expression in metastasizing tumors compared to non-metastasizing tumors (Table 2). Low false discovery rates indicated several of these gene sets to be significantly differentially expressed: 8q24, 16q24, 20q11, and 20q13, were significantly upregulated and 8p21 was significantly downregulated.

Several consecutive cytogenetic regions had similar mean ranking values in this meta-analysis, e.g., 20q11 and 20q13 at the top of Table 2 indicating that a larger region of chromosome 20q is gained in metastasizing tumors. To address this question, pairs of neighboring regions were analyzed resulting in identification of further large regions significantly related to metastasis (Table 3). Most of the gene set identified in the single gene set analysis were extended to significant neighboring regions except 8p21 that was not significant in the neighboring region analysis. On the other hand 1q32–42, 17q23–25, and 12q12–13 that were not significant in the single gene set analysis were significant in the neighboring region analysis. The borderline significant region 1p31 was extended to a large region at chromosome 1p (1p32–13) significantly downregulated in metastasizing tumors.

Refining differentially expressed regions and candidate genes

To narrow down the regions, sliding mean plots were generated for chromosome arms or entire chromosomes containing regions displaying differential expression in the above neighboring region and single gene set analysis. Core regions were refined and genes within these regions fulfilling predefined criteria for additional imbalanced expression were selected as candidate genes (Table 4). *P*-values calculated within each dataset for these genes are shown in supplementary Table 1.

The large downregulated region on chromosome 1p32–13 observed in Table 3 was sustained by minima of several datasets in that region in the sliding mean plot, and a core region at 1p31–21 was resolved (supplementary Fig. 1a). One gene, *DIRAS3*, met the selection criteria as candidate metastasis suppressor gene (supplementary Fig. 1b). Furthermore, *TGFBR3* have recently been proposed as candidate gene on 1p [12], supported by the present data, although not fulfilling candidate gene selection criteria (supplementary Fig. 1b).

The next region, in chromosomal order, identified in the gene set enrichment meta-analysis is 1q32–42, that by sliding mean analysis can be refined to a core region of upregulated genes at 1q41–42 (supplementary Fig. 2) but no single gene met selection criteria for causal candidate genes.

Chromosome 8 exhibit up and downregulated regions in gene set enrichment meta-analysis summarized in one sliding mean plot for entire chromosome 8 (Fig. 1a). A core minima-peak region of 8p22–21 was selected. In this region, three genes fulfill criteria for additionally downregulated: *PSD3*, *LPL*, and *EPHX2* (Fig. 1b). The upregulated region in the far distal end of the q arm diverges because the sliding mean is calculated with fewer genes for the last 50 genes (Fig. 1a). However, the ratio plot points clearly at *RECQL4* fulfilling metastasis candidate gene criteria (Fig. 1c). This gene is upregulated in 6 of 7 datasets, with *P*-values below 0.05 in four of these datasets (supplementary Table 1).

At 12q a core region is not so obvious, but at 12q14 3–4 datasets shows a local maximum in a region surrounding a previously proposed candidate gene *MDM2* (Supplementary Fig. 3). However, this gene is only slightly upregulated and does, like all other gene in the region, not fulfill selection criteria.

Decreased expression at 14q11–24 is observed in the neighboring region analysis (Table 3) but the core region can be limited to a part of 14q24 containing the additionally downregulated candidate gene *FOS* (supplementary Fig. 4).

Table 2 Gene set enrichment meta-analysis of chromosomal regions differentially expressed between metastasizing and non-metastasizing tumors

Gene set	Amsterdam	HUMAC	Huang	Rotterdam	Sotiriou 2003	Sotiriou 2006	Uppsala	Stockholm	Mean	*P*-value	fdr
Upregulated											
CHR20Q13	6	30	30	3	3	2	9	11	11.8	4.0E−06	2.1E−04
CHR20Q11	4	41	25	9	7	5	1	3	11.9	4.0E−06	2.1E−04
CHR16Q24	5	59	5	17	18	12	2	1	14.9	3.5E−05	1.2E−03
CHR8Q24	10	66	43	6	16	3	6	10	20.0	3.9E−04	1.0E−02
CHR16Q22	7	62	9	15	59	24	4	6	23.3	1.6E−03	3.2E−02
Downregulated											
CHR1P31	102	63	102	47	70	97	100	101	85.3	2.3E−04	1.2E−02
CHR8P21	103	92	70	102	57	94	103	103	90.5	6.0E−06	6.2E−04

The ranking numbers indicate the ranking of each gene set out of the 103 gene sets in each dataset and the mean ranking number indicate the ranking in the meta-analysis. Only seven significant or borderline significant out of a total of 103 regions are shown

Table 3 Effect of thyroxine on nodular thyroid disease recurrence

AUTHORS	LOCATION	LEVEL OF EVIDENCE	TOTAL NO. OF PTS.	LOST TO F/U OR EXCLUDED	DOSAGE OF THYROXINE	MEAN F/U (years)	# of pts with and without thyroxine	% recurrence with thyroxine	% recurrence without thyroxine	p value
Anderson et al. [53]	England	retrospective	218	33 (15%)	100 mcg/day	10.3	171 / 14	9 (5%)	6 (41%)	p=0.003
Berghout et al. [12]	Netherlands	retrospective	146	33 (23%)	N/A	7.5	11 / 102	1 (9%)	19 (19%)	p=NS
Berglund et al. [19]	Sweden	retrospective	287	26 (9%)	N/A	8.0	75 / 186	6 (8%)	20 (11%)	p=NS
Bistrup et al. [13]	Denmark	randomized, prospective	100	31 (31%)	100 mcg/day	9.0	27 / 42	5 (19%)	11 (26%)	p=NS
Geerdsen and Frølund [24]	Denmark	randomized, prospective	29	0 (0%)	200 mcg/day	1.5	17 / 12	0 (0%)	0 (0%)	p=NS
Hedman et al. [52]	Sweden	retrospective	178	72 (40%)	100 to 150 mcg/day	15.0	58 / 37	8 (14%)	5 (14%)	p=NS
Hegedus et al. [10]	Denmark	randomized, prospective	110	0 (0%)	150 mcg/day	1.0	52 / 58	2 (4%)	1 (2%)	p=NS
Ibis et al. [54]	Turkey	retrospective	206	N/A	at least 100 mcg/day	7.8	58 / 148	23 (40%)	116 (78%)	p<0.001
Miccoli et al. [20]	Italy	randomized, prospective	60	0 (0%)	100 mcg/day for half and 2.2–3 mcg/kg/day to rest	3.0	32 substitutive dosages / 28 with suppressive dosages	25 (78%) with substititutive and 6 (21%) with suppressive	No pts without thyroxine	p<0.005
Persson et al. [51]	Sweden	retrospective	211	4 (2%)	100 mcg/day	5.0	168 / 29	10 (6%)	1 (3.4%)	p=NS
Röjdmark and Järhult [55]	Sweden	retrospective	36	0 (0%)	50–200 mcg/day	30.0	11 / 32	5 (45%)	13 (41%)	p=NS

Table 4 Candidate metastasis suppressor and promoting genes

Gene symbol	Cytoband	dRDE	*P*-value	GO biological process
DIRAS3	1p31	−0.22	0.004	GTPase activity; regulation of cyclin-dependent protein kinase activity
PSD3	8p22	−0.15	0.032	ARF protein signal transduction
LPL	8p22	−0.17	0.002	Circulation; fatty acid metabolism; lipid catabolism; posttranslational membrane targeting
EPHX2	8p21	−0.18	0.009	Aromatic compound metabolism; calcium ion homeostasis; drug metabolism; inflammatory response; oxygen and reactive oxygen species metabolism; positive regulation of vasodilation; regulation of blood pressure; response to toxin; xenobiotic metabolism
RECQL4	8q24.3	0.25	0.038	DNA repair; development; positive regulation of cell proliferation; pigmentation
FOS	14q24.3	−0.15	0.003	DNA methylation; inflammatory response; regulation of cell cycle; regulation of transcription from Pol II promoter
PRMT7	16q22	0.28	0.004	Histone methylation; peptidyl-arginine methylation; regulation of protein binding
GINS2	16q24.1	0.35	0.011	DNA strand elongation during DNA replication
AURKA	20q13.2–13.3	0.15	0.004	Mitotic cell cycle

16q is consistently upregulated in the majority of datasets. Local maxima are observed at 16q22 and 16q24 containing additionally upregulated candidate genes *PRMT7* and *GINS2*, respectively (supplementary Fig. 5).

17q23–25 display increased expression in gene set enrichment meta-analysis. Sliding mean plot of entire 17q identifies two core peaks within this region and in addition a peak at the *ERBB2* locus (supplementary Fig. 6a). However, no genes fulfill criteria for additional upregulation.

Two core peak regions are identified from the sliding mean plot of chromosome 20q: 20q11 and 20q13 and last mentioned region contains an additionally upregulated candidate gene *AURKA* (supplementary Fig. 7).

Discussion

Regions of differential expression

We have used meta-analysis of tumor gene expression data to identify several chromosomal regions associated with metastasis of breast cancer. The results indicate that regional copy number imbalance is linked with metastasis and is reflected in overall gene expression of the region, in agreement with our hypothesis. Many studies have compared allelic imbalance in tumor tissue compared to normal tissue. However, studies of the association of allelic imbalance to disease outcome are sparse as discussed in the following.

At chromosome 1, LOH at 1p31 has been associated with survival [13] supported by our findings of 1p31–21 downregulated in metastasizing breast tumors. At 1q an early LOH study found association to 1q21 [14] and CGH has been used to demonstrate prognostic disadvantage of simultaneously gain of 1q and 8q [15]. Our analysis point at the distal end of the chromosome 1q arm and refines the region to 1q41–42.

The previous evidence for prognostic association to 8p is weak: By use of LOH analysis, Morikawa et al., demonstrated borderline significance of 8p imbalance when 18p was retained [16]. However, in poor outcome tumors, our results demonstrate a very concordant downregulation of gene expression of the eight dataset, and the region is refined to 8p22–21 (Fig. 1a). The prognostic disadvantage of having 8q amplification in breast cancer has previously been demonstrated by CGH analysis [17] and several focused analyses of the major candidate gene *MYC*, e.g., with FISH based tissue array have supported this [18]. Our analysis exhibit amplification of entire 8q, but the strongest signal is observed at the distal end of the chromosome arm (Fig. 1a).

Amplification of 12q14–22 has been linked with poor outcome by CGH analysis [19], in agreement with our results, and expression of the major candidate gene MDM2 has been linked with poor prognosis in many cancers other than breast cancer [20]. At 14q, loss of heterozygosity is observed in breast cancer [21], and prognostic advantage of 14q31 loss has been reported in one study [22]. Contradictory to this, our results points at 14q24 and indicates poor prognosis when gene expression is decreased (supplementary Fig. 4a). Loss of 16q is one of the most common observed copy number aberrations in breast cancer [23] and has previously been associated with good prognosis in three studies using LOH [24], gene expression profiling [25], or aCGH [26].

The *ERBB2* locus at 17q is not significant in the gene set enrichment meta-analysis. The reason for this is probably

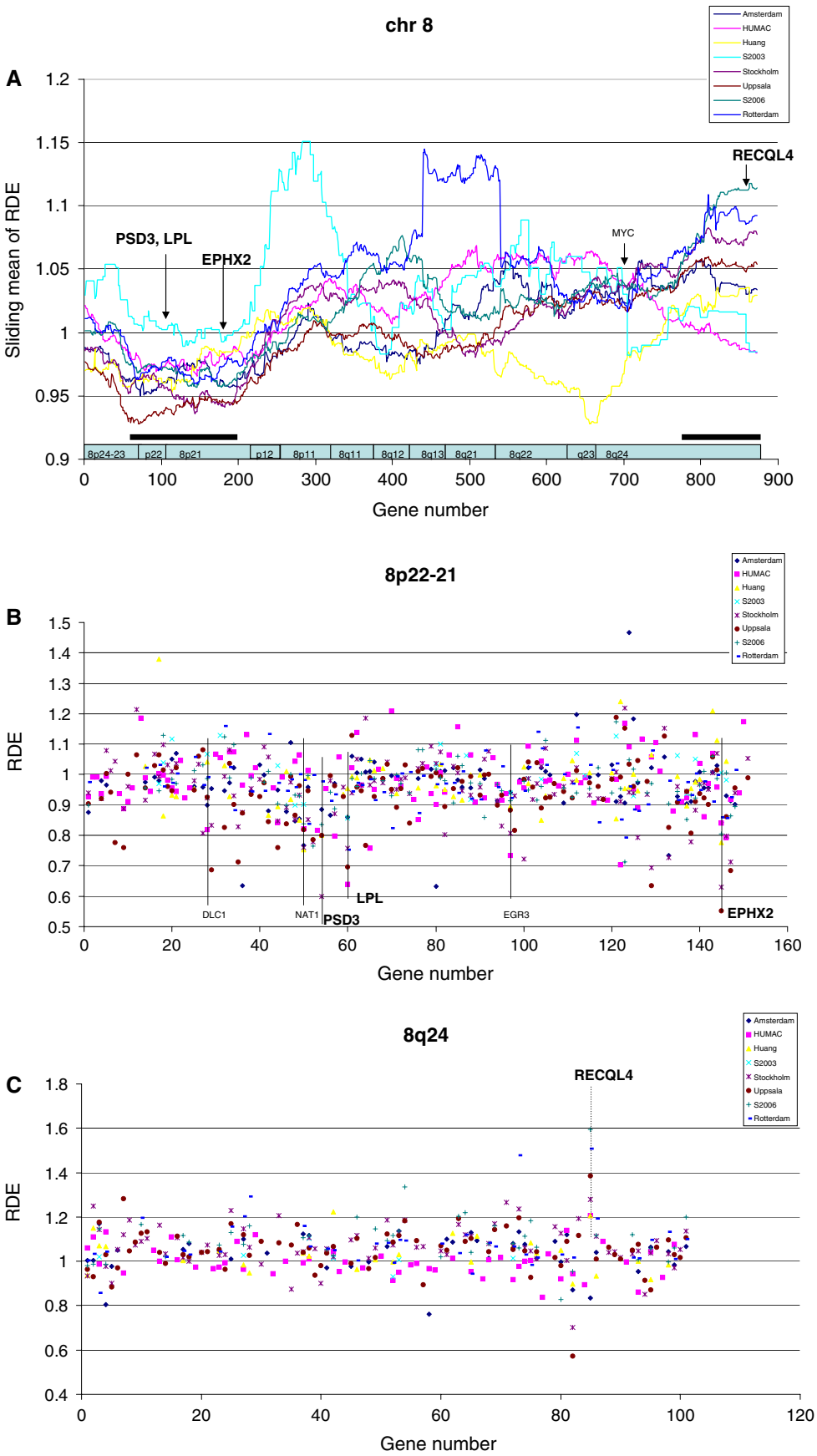

Fig. 1 Refining of differentially expressed regions on chromosome 8. (**a**) Sliding mean plot of entire chromosome 8. The cytogenetic bands corresponding to gene sets in the GSEA analysis are indicated in the blue bar, and the core downregulated region at 8p and upregulated region at 8q are highlighted with black bars. (**b**) Scaled RDE's for all genes in the 8p core region. The positions of candidate genes fulfilling criteria for additional regulation are indicated with bold whereas other obvious genes, however not fulfilling criteria have normal type. (**c**) Scaled RDE for genes in the 8q core region

that the amplicon is of limited size and is shared between 17q12 and 17q21 gene sets. However, *ERBB2* amplification is clear in the sliding mean analysis (supplementary Fig. 6a). The 17q23-q25 neighboring region can be divided into two regions, 17q21–23 and 17q25, by sliding mean analysis. The 17q21–23 region coincides with a core amplicon identified in breast tumors compared to normal tissue [27].

20q13 amplification has been associated with poor prognosis [28] and has been validated in several studies [29] while 20q11 gain has not previously been reported to have prognostic meaning independent of 20q13.

In summary our approach supports several previous findings. Several of the regions have been found in only one or few studies and in general a limited number of patients have been included. This study serves as an independent validation with an independent method of these findings using a large patient group. To our knowledge this is the first report of prognostic significance of expression imbalance at 1q41–42, 14q24 and 17q21.33–23.

A strength of our analysis is that regions can be subdivided into separate sub regions. 16q is split into 16q22 and 16q24 with a local minimum in-between (supplementary Fig. 5). Similarly, 17q is subdivided into three regions where amplification associates with metastasis: 17q12–21 (ERBB2 locus), 17q22, and 17q25. Amplification of 20q is also comprised of two peaks located at 20q11 and 20q13 respectively (supplementary Fig. 7a).

Only one previous study performed by Wennmalm et al. has used genome-wide gene expression data to identify positional effects in metastasis of breast cancer [25]. The major finding in that study is negative association of 16q expression and survival. Furthermore, they found negative association of 20q and positive association of 1p and 8p22–p21 and survival, respectively, all in agreement with our findings. However, they also identified positive association of 13q, 9p, 9q22 and survival, not observed in our results. On the other hand we have observed association of allelic imbalance at 8q24, 12q14–15, 14q24, 17q12–21.2, and 17q21.33–23 not identified in their analysis. This may seem peculiarly because both studies use analysis of expression of gene sets limited by cytogenetic band positions and the data they used is also included in our meta-analysis. However, the analysis differs fundamentally because they only included 200–500 genes most differentially expressed between outcome groups and used Fishers exact test to calculate significance of overrepresentation in the gene sets. We used gene set enrichment analysis and included all genes to rank the gene sets according to prognostic significance. However, not all gene sets were included in the analysis because the number of genes in small cytogenetic regions was not sufficient in all dataset to perform the analysis which may explain the absence of regions identified by others. Another main difference from the study by Wenmalm et al. is that we performed a meta-analysis enabling us to perform sliding mean analysis to narrow down regions of differential expression with concordance in the eight datasets included.

The inclusion of different outcome, i.e., metastasis and local recurrence in our study may potentially bias the results; however, local recurrence constitute a minor fractions of recurrences compared to distant metastasis. Furthermore, the classification of lymph node positive patients without recurrence as non-metastasis may be controversial. This may bias the results towards the metastatic mechanisms following primary spread to lymph node. Treatment response may also influence-identified regions, however majority of patients did not receive adjuvant treatment and the fraction of patients responding to treatment is low.

A different pattern of differential expression of HUMAC dataset is observed compared to the other dataset, i.e., ranking numbers in Tables 2 and 3 that differs from the general pattern in the other dataset. The tumors included in the HUMAC dataset are all node negative, estrogen receptor positive, and low malignant and the results might indicate that different mutational patterns characterize these tumors. Further molecular characterization of low-malignant cancers is required to verify this.

Candidate genes

The sliding mean analysis is used to narrow down core regions of differential expression with concordance in the eight datasets included. Furthermore, inspection of differential expression ratios for single genes in identified core regions with predefined selection criteria has allowed identification of single genes that may be causal genes in metastasis. These genes having expressional imbalance, in addition to a general change in the region may be explained by mutations in the genes, epigenetic changes like promoter methylation, gene regulation mediated by metastatic pathways etc. In the following these candidate genes are discussed in more detail.

In core region 1p31–21, *DIRAS3* is additionally downregulated. *DIRAS3* is member of the *RAS* super family of protooncogenes and negatively regulates the transcription of *cyclin D1* a central regulator of cell cycle. *DIRAS3* is often imprinted and LOH in breast and ovarian tumors most often affect the non-imprinted allele [30]. Furthermore, methylation of the *DIRAS3* promoter has been demonstrated to predict poor survival in breast cancer [31]. This illustrates that different mechanisms, in this case allelic loss and methylation, can deactivate the same gene and support the viability of our method to detect these genes. The finding of 2 consecutive obvious downregulated genes, *GADD45A* and *GNG12*, right beside *DIRAS3* in the 1p core region, although

not fulfilling our criteria, may indicate a micro segmental deletion within a larger region often lost in breast tumors. However, the promoter of Growth Arrest- and DNA Damage-inducible gene *GADD45A* is often methylated in breast cancer [32]. Another potential metastasis suppressor gene, *TGFBR3*, at 1p is proposed by Dong et al. [12] who recently linked decreased expression of this gene with poor prognosis. This is supported by lowered expression of *TGFBR3* in 6 of 8 datasets (supplementary Fig. 1b), although the gene does not fulfill criteria for additional regulation. Noteworthy, TGFB3, the ligand for TGFBR3 at 14q core region is also borderline additionally downregulated (supplementary Fig. 4b).

Although a very clear peak is observed in the sliding mean plot for 1q, no genes fulfill criteria for additional upregulation in this region (supplementary Fig. 2b). The reason for this can be that the causal gene amplified in this region is not regulated by other mechanisms, hence no additional effect is observed at gene expression level.

The gene set defined by cytoband 8p21 appears significantly downregulated in the GSEA meta-analysis of single regions but not in the neighboring region analysis indicating limited size of this segment which is supported by the sliding mean plot (Fig. 1a). Three additionally regulated genes *PSD3*, *LPL* and *EPHX2* are identified in the region (Fig. 1b). In agreement with our results, low expression of ADP-ribosylation factor *PSD3* has previously been linked with poor prognosis in ovarian carcinomas [33]. Interestingly elevated expression of *LPL* is a marker of poor prognosis in chronic lymphocytic leukemia conflicting with the present result for breasts cancer [34]. *EPHX2* is a cytosolic epoxide hydrolase that has not been linked to cancer but its diverse functions in metabolism of possible carcinogens might prevent progression and metastasis (Table 4). Strikingly germline mutations in both *LPL* and *EPHX2* are involved in familial hypercholesterolemia in agreement with pathway analysis of present datasets demonstrating downregulation of lipid metabolism in metastasizing tumors (results will be published separately). The abundance of candidate genes at 8p may indicate that several genes affect metastasis.

At 8q, *MYC* is a major candidate gene amplified in several cancers [17]. This is supported by general trend of upregulation at 8q22–24 (Fig. 1a). However, the present data point towards the distal end of the chromosome. At this locus the helicase *RECQL4* gene is additionally upregulated (Fig. 1c). This gene is mutated in Rothmund–Thompson syndrome, a rare hereditary dermatosis with high incidence of osteogenic sarcoma. Furthermore, the highly conserved helicase gene family members *WRN* and *BLM* predispose for Werner and Bloom syndromes characterized by high incidence of sarcomas and multiple cancers respectively. *RECQL4* has been reported to be amplified and overexpressed in colorectal, and cervical cancer [35, 36], but the prognostic significance has not been reported.

Amplification of the *MDM2* locus at 12q has been associated with poor prognosis in several cancers [20]. In breast cancer expression of MDM2 at protein level is related to survival [37]. Our results clearly demonstrate a relation between amplification of 12q and poor prognosis in breast cancer although no genes fulfill criteria for additional regulation.

At 14q the additionally downregulated transcription factor FOS is member of a family of oncogenes that together with JUN constitutes transcription factor AP-1 and regulates the prominent cell cycle regulators cyclin D1 and Rb (reviewed by [38]. In agreement with present finding, *FOS* transcription is strongly downregulated in breast cancer cell lines with metastatic potential compared to non-metastatic cells, while another gene family member *FRA-1* is upregulated in metastasizing cells [39].

CDH1 at 16q22 has been reported as the major candidate gene and is often methylated [40] but do not fulfill our selection criteria. However, *PRMT7* coding for an arginine methyltransferase, with unknown relation to cancer prognosis, is additionally upregulated (supplementary Fig. 5b). The additionally upregulated candidate gene at 16q24, *GINS2*, is essential for initiation of for replication of DNA [41] making it a relevant metastasis candidate gene.

No genes at 17q12–21, 17q21–23, and 17q25 are meeting our criteria for additionally upregulation. At 20q the sliding mean plot identifies two regions 20q11 and 20q13 upregulated in metastasizing breast tumors (supplementary Fig. 7a). No additionally regulated genes are identified at 20q11. 20q13 is intensively studied and previously proposed prognostic candidate genes includes *NCOA3*, *ZNF217*, and *AURKA* [28, 42] supported by concordant upregulation of these genes in majority of datasets (supplementary Fig. 7c) and *AURKA* meets our selection criteria for additionally upregulation. *AURKA* is a cell cycle-regulated kinase that appears to be involved in microtubule formation and stabilization at the spindle pole during chromosome segregation. Overexpression of *AURKA* has been associated with poor prognosis in ovary and several other cancers [43] and is correlated to nuclear grade in breast carcinomas [44].

In summary several candidate genes are identified as possible cause of metastasis in regions with copy number aberrations in metastasizing tumors. The main function of the identified genes is cell cycle and secondly DNA replication and repair (*DIRAS3*, *FOS*, *GINS2*, *AURKA*, and *RECQL4*, Table 4) in agreement with several reports of these as major metastatic pathways [45]. Some regions with differential expression do not contain obvious candidate genes despite significant general tendencies in the

region (1q41–42 and 12q14–21, 17q12–21 17q21–23, 17q25, and 20q11). This may be explained by no additional mechanisms of regulation of causal genes. Strikingly, the clinically used prognostic marker, *ERBB2* at 17q12–21 is not obviously upregulated compared to near surrounding genes. However, this is supported by lacking reports of alternative regulatory mechanisms and considerable number of neighboring transcript often amplified [46] making distinct regulation of this gene compared to surrounding genes unlikely.

On the other hand, 8p22–21 regions display three additionally expressed genes, making it difficult to suggest the causal gene. The reason may be that several of these genes affect metastasis and concordant loss promotes metastasis. Alternatively some of the genes with additional differential expression may be secondary targets for other primary events.

Most studies of allelic imbalance and candidate genes have compared cancer tissue and normal tissue. These tumor suppressor genes and oncogenes lost or gained resulting in selection advantage during tumorigenesis may not necessarily be the same genes causal for metastasis. An example is 16q where loss leads to good prognosis. This may be explained by one tumor suppressor gene lost in tumorigenesis and at the same time a metastasis-promoting gene is lost resulting in good prognosis [24]. In some cases there are concordance between tumorigenic genes and metastasis promoting genes, e.g., *RECQL4* and *AURKA* amplified in several cancers, compared to normal tissue and related to metastasis according to our results.

Acknowledgments All the researchers that have generated gene expression data that we have included in the analysis are acknowledged for allowing us to use their data.

References

1. Thomassen M, Tan Q, Eiriksdottir F, Bak M, Cold S, Kruse TA (2007) Prediction of metastasis from low-malignant breast cancer by gene expression profiling. Int J Cancer 120:1070–1075
2. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF et al (2003) Gene expression predictors of breast cancer outcomes. Lancet 361:1590–1596
3. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A et al (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A 100:10393–10398
4. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J et al (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262–272
5. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679
6. van de Vijver, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009
7. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A et al (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 102:13550–13555
8. Calza S, Hall P, Auer G, Bjohle J, Klaar S, Kronenwett U et al (2006) Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Res 8:R34
9. Tan Q, Thomassen M, Kruse TA (2007) Feature selection for predicting tumor metastases in microarray experiments using paired design. Cancer Inform 2:133–138
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550
11. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J et al (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30:e15
12. Dong M, How T, Kirkbride KC, Gordon KJ, Lee JD, Hempel N et al (2007) The type III TGF-beta receptor suppresses breast cancer progression. J Clin Invest 117:206–217
13. Ragnarsson G, Eiriksdottir G, Johannsdottir JT, Jonasson JG, Egilsson V, Ingvarsson S: (1999) Loss of heterozygosity at chromosome 1p in different solid human tumours: association with survival. Br J Cancer 79:1468–1474
14. Borg A, Zhang QX, Olsson H, Wenngren E (1992) Chromosome 1 alterations in breast cancer: allelic loss on 1p and 1q is related to lymphogenic metastases and poor prognosis. Genes Chromosomes Cancer 5:311–320
15. Janssen EA, Baak JP, Guervos MA, van Diest PJ, Jiwa M, Hermsen MA (2003) In lymph node-negative invasive breast carcinomas, specific chromosomal aberrations are strongly associated with high mitotic activity and predict outcome more accurately than grade, tumour diameter, and oestrogen receptor. J Pathol 201:555–561
16. Morikawa A, Williams TY, Dirix L, Colpaert C, Goodman M, Lyles RH et al (2005) Allelic imbalances of chromosomes 8p and 18q and their roles in distant relapse of early stage, node-negative breast cancer. Breast Cancer Res 7:R1051–R1057
17. Isola JJ, Kallioniemi OP, Chu LW, Fuqua SA, Hilsenbeck SG, Osborne CK et al (1995) Genetic aberrations detected by comparative genomic hybridization predict outcome in node-negative breast cancer. Am J Pathol 147:905–911
18. Al-Kuraya K, Schraml P, Torhorst J, Tapia C, Zaharieva B, Novotny H et al (2004) Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer Res 64: 8534–8540
19. Karlsson E, Danielsson A, Delle U, Olsson B, Karlsson P, Helou K (2007) Chromosomal changes associated with clinical outcome in lymph node-negative breast cancer. Cancer Genet Cytogenet 172:139–146
20. Rayburn E, Zhang R, He J, Wang H (2005) MDM2 and human malignancies: expression, clinical pathology, prognostic markers, and implications for chemotherapy. Curr Cancer Drug Targets 5:27–41
21. Wang ZC, Lin M, Wei LJ, Li C, Miron A, Lodeiro G et al (2004) Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. Cancer Res 64:64–71
22. Martin MD, Fischbach K, Osborne CK, Mohsin SK, Allred DC, O'Connell P (2001) Loss of heterozygosity events impeding breast cancer metastasis contain the MTA1 gene. Cancer Res 61:3578–3580

23. Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, Leibu E et al (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Res 16: 1465–1479
24. Hansen LL, Yilmaz M, Overgaard J, Andersen J, Kruse TA (1998) Allelic loss of 16q23.2–24.2 is an independent marker of good prognosis in primary breast cancer. Cancer Res 58: 2166–2169
25. Wennmalm K, Calza S, Ploner A, Hall P, Bjohle J, Klaar S et al (2007) Gene expression in 16q is associated with survival and differs between Sorlie breast cancer subtypes. Genes Chromosomes Cancer 46:87–97
26. Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Segraves R et al (2006) Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer 6:96
27. Parssinen J, Kuukasjarvi T, Karhu R, Kallioniemi A (2007) High-level amplification at 17q23 leads to coordinated overexpression of multiple adjacent genes in breast cancer. Br J Cancer
28. Letessier A, Sircoulomb F, Ginestier C, Cervera N, Monville F, Gelsi-Boyer V et al (2006) Frequency, prognostic impact, and subtype association of 8p12, 8q24, 11q13, 12p13, 17q12, and 20q13 amplifications in breast cancers. BMC Cancer 6:245
29. Chin SF, Wang Y, Thorne NP, Teschendorff AE, Pinder SE, Vias M et al (2007) Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. Oncogene 26:1959–1970
30. Yu Y, Xu F, Peng H, Fang X, Zhao S, Li Y et al (1999) NOEY2 (ARHI), an imprinted putative tumor suppressor gene in ovarian and breast carcinomas. Proc Natl Acad Sci U S A 96:214–219
31. Widschwendter M, Siegmund KD, Muller HM, Fiegl H, Marth C, Muller-Holzner E et al (2004) Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. Cancer Res 64:3807–3813
32. Wang W, Huper G, Guo Y, Murphy SK, Olson JA Jr, Marks JR (2005) Analysis of methylation-sensitive transcriptome identifies GADD45a as a frequently methylated gene in breast cancer. Oncogene 24:2705–2714
33. Pils D, Horak P, Gleiss A, Sax C, Fabjani G, Moebus VJ et al (2005) Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma: N33 and EFA6R have a potential impact on overall survival. Cancer 104:2417–2429
34. Heintel D, Kienle D, Shehata M, Krober A, Kroemer E, Schwarzinger I et al (2005) High expression of lipoprotein lipase in poor risk B-cell chronic lymphocytic leukemia. Leukemia 19:1216–1223
35. Buffart TE, Coffa J, Hermsen MA, Carvalho B, van dS Jr, Ylstra B et al (2005) DNA copy number changes at 8q11–24 in metastasized colorectal cancer. Cell Oncol 27:57–65
36. Narayan G, Bourdon V, Chaganti S, rias-Pulido H, Nandula SV, Rao PH et al (2007) Gene dosage alterations revealed by cDNA microarray analysis in cervical cancer: identification of candidate amplified and overexpressed genes. Genes Chromosomes Cancer 46:373–384
37. Turbin DA, Cheang MC, Bajdik CD, Gelmon KA, Yorida E, De LA et al (2006) MDM2 protein expression is a negative prognostic marker in breast carcinoma. Mod Pathol 19:69–74
38. Milde-Langosch K (2005) The Fos family of transcription factors and their role in tumourigenesis. Eur J Cancer 41:2449–2461
39. Kustikova O, Kramerov D, Grigorian M, Berezin V, Bock E, Lukanidin E et al (1998) Fra-1 induces morphological transformation and increases in vitro invasiveness and motility of epithelioid adenocarcinoma cells. Mol Cell Biol 18:7095–7105
40. Caldeira JR, Prando EC, Quevedo FC, Neto FA, Rainho CA, Rogatto SR (2006) CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. BMC Cancer 6:48
41. Seki T, Akita M, Kamimura Y, Muramatsu S, Araki H, Sugino A (2006) GINS is a DNA polymerase epsilon accessory factor during chromosomal DNA replication in budding yeast. J Biol Chem 281:21422–21432
42. Ginestier C, Cervera N, Finetti P, Esteyries S, Esterni B, Adelaide J et al (2006) Prognosis and gene expression profiling of 20q13-amplified breast cancers. Clin Cancer Res 12:4533–4544
43. Landen CN Jr, Lin YG, Immaneni A, Deavers MT, Merritt WM, Spannuth WA et al (2007) Overexpression of the centrosomal protein Aurora-A kinase is associated with poor prognosis in epithelial ovarian cancer patients. Clin Cancer Res 13:4098–4104
44. Royce ME, Xia W, Sahin AA, Katayama H, Johnston DA, Hortobagyi G et al (2004) STK15/Aurora-A expression in primary breast tumors is correlated with nuclear grade but not with prognosis. Cancer 100:12–19
45. Thomassen M, Tan Q, Eiriksdottir F, Bak M, Cold S, Kruse TA (2007) Comparison of gene sets for expression profiling: prediction of metastasis from low-malignant breast cancer. Clin Cancer Res 13:5355–5360
46. Kauraniemi P, Kuukasjarvi T, Sauter G, Kallioniemi A (2003) Amplification of a 280-kilobase core region at the ERBB2 locus leads to activation of two hypothetical proteins in breast cancer. Am J Pathol 163:1979–1984