Oncogene (2007) 26, 1959–1970

www.nature.com/onc

ONCOGENOMICS

Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers

S-F Chin[1], Y Wang[1], NP Thorne[1], AE Teschendorff[1], SE Pinder[1,2], M Vias[1], A Naderi[1], I Roberts[3], NL Barbosa-Morais[1,4], MJ Garcia[1], NG Iyer[1,6], T Kranjac[1], JFR Robertson[5], S Aparicio[1,7], S Tavaré[1], I Ellis[5], JD Brenton[1] and C Caldas[1]

[1]*Cancer Genomics Program, Department of Oncology, Hutchison/MRC Research Centre, University of Cambridge, Cambridge, UK;* [2]*Department of Histopathology, Addenbrookes Hospital, Cambridge, UK;* [3]*Cancer Cell Unit, Hutchison/MRC Research Centre, Cambridge, UK;* [4]*Institute of Molecular Medicine, Lisbon, Portugal;* [5]*Departments of Histopathology and Surgery, Breast Unit, Nottingham City Hospital NHS Trust and University of Nottingham, UK*

We analysed 148 primary breast cancers using BAC-arrays containing 287 clones representing cancer-related gene/loci to obtain genomic molecular portraits. Gains were detected in 136 tumors (91.9%) and losses in 123 tumors (83.1%). Eight tumors (5.4%) did not have any genomic aberrations in the 281 clones analysed. Common (more than 15% of the samples) gains were observed at 8q11–qtel, 1q21–qtel, 17q11–q12 and 11q13, whereas common losses were observed at 16q12–qtel, 11ptel–p15.5, 1p36–ptel, 17p11.2–p12 and 8ptel–p22. Patients with tumors registering either less than 5% (median value) or less than 11% (third quartile) total copy number changes had a better overall survival (log-rank test: $P = 0.0417$ and $P = 0.0375$, respectively). Unsupervised hierarchical clustering based on copy number changes identified four clusters. Women with tumors from the cluster with amplification of three regions containing known breast oncogenes (11q13, 17q12 and 20q13) had a worse prognosis. The good prognosis group (Nottingham Prognostic Index (NPI) $\leqslant 3.4$) tumors had frequent loss of 16q24–qtel. Genes significantly associated with estrogen receptor (ER), Grade and NPI were used to build *k*-nearest neighbor (KNN) classifiers that predicted ER, Grade and NPI status in the test set with an average misclassification rate of 24.7, 25.7 and 35.7%, respectively. These data raise the prospect of generating a molecular taxonomy of breast cancer based on copy number profiling using tumor DNA, which may be more generally applicable than expression microarray analysis.
Oncogene (2007) **26,** 1959–1970. doi:10.1038/sj.onc.1209985; published online 25 September 2006

Keywords: breast cancer; array-CGH; amplification; deletion

Correspondence: Professor C Caldas, Oncology, Hutchison/MRC Research Centre, University of Cambridge, Hills Road, Cambridge, Cambridgeshire CB2 2XZ, UK.
E-mail: cc234@cam.ac.uk
[6]Current address: Singapore General Hospital, Outram Road, Singapore.
[7]Current address: BC Cancer Research Center, 675 West 10th Avenue, Vancouver, BC, V5L 1L3.
Received 29 March 2006; revised 11 July 2006; accepted 25 July 2006; published online 25 September 2006

Introduction

Breast cancer is the most common malignancy in women when skin cancers are excluded. Morphology and surgical staging are used to derive prognostic classifiers such as the Nottingham Prognostic Index (NPI), widely used in the UK and Europe (Elston and Ellis, 1991), but subject to limitations such as observer variability (Gilchrist *et al*., 1985). Large-scale expression analyses using either complementary DNA (cDNA) or oligonucleotide arrays have been utilized to obtain gene signatures and there is emerging evidence that this approach can be used to classify breast tumors into different prognostic groups. Expression signatures correlated with estrogen receptor (ER), and ERBB2 status have also been described (Perou *et al*., 2000; Sorlie *et al*., 2001; van't Veer *et al*., 2002). Arrays can also be used to measure DNA copy number alterations. Several studies using such arrays have now shown that DNA copy number changes correlate with gene expression (Pollack *et al*., 2002; Heidenblad *et al*., 2005; Kim *et al*., 2005). Moreover, nonrandom physical clusters of genes with correlated expression in invasive ductal breast cancers have been described and some of these gene clusters coincide with loci found to be frequently altered at the copy number level in breast cancers (Reyal *et al*., 2005). This suggests the potential use of copy number profiling as an alternative to, or in combination with, expression analysis to subtype breast cancers.

Cytogenetic methods, including conventional comparative genomic hybridization (CGH), have revealed chromosomal regions that are frequently altered in breast tumors (Kallioniemi *et al*., 1994; Tirkkonen *et al*., 1998). Some of these regions contain known tumor suppressor genes (TSG), for example, *TP53* and oncogenes, for example, *ERBB2*, but target genes for other regions have yet to be identified. Conventional CGH is limited by a 5–10 Mb resolution (Forozan *et al*., 1997), is time consuming and requires karyotyping expertise. Array-based CGH combines the resolution of fluorescent *in situ* hybridization (FISH) (i.e. gene level) with the whole genome screening capacity of conventional CGH, thereby allowing the analysis of

DNA copy number alterations in clinical samples (Pinkel *et al*., 1998).

In this report, we used a commercially available array containing 287 clones printed in triplicate to profile 148 primary breast cancers from a cohort of patients with a median age of 58 years, 68% being postmenopausal and ER positive, and with median clinical follow up of 11 years.

Results

Overview of genomic changes in breast tumors

We found that 136 breast tumors (91.9%) had at least one gain and 123 tumors (83.1%) had at least one loss. Only eight tumors (5.4%) did not have any copy number aberrations. The median number of gains per tumor was eight clones (range: 0–60, s.d. = 13.7) and the median number of losses was six clones (range: 0–66, s.d. = 15.9).

An overall assessment of total copy number aberrations in each tumor was determined by calculating the percentage of clones out of the 281 analysed with alterations. In this assessment, the tumors were classified into three groups, designated low (<5% changes), intermediate (5–11% changes) and high (>11% changes; Supplementary Figure 1), where 5 and 11% are the median and third quartile of percentage of copy number aberrations across tumors. The majority of the tumors (55.4%) belonged to the low group and 22.3% were classified into the high group. This contrasted with what we observed in cell lines where the median percentage of copy number changes was 24% and more than 56% (27/48) had greater than 24% of changes (data not shown).

We used the Fisher's exact test to compare the morphological phenotypes of the groups of tumors based on the number of copy number aberrations (Table 1). A correlation with histological grade was found when comparing the group with low copy number changes (<5%) vs all others ($P = 0.028$).

Kaplan–Meier survival analysis showed that the overall survival of the patients in the high copy number change group was worse than that for the remaining cases (hazard ratio (HR) = 1.83 with 95% confidence interval (CI): 0.94–3.56; log rank $P = 0.0375$). Survival of the combined group of high and intermediate changes was also worse (HR = 1.95 with 95% CI: 1.01–3.73; log rank $P = 0.047$) (Figure 1; top row). Multivariate analysis using a Cox proportional hazards model and stepwise regression was performed including the variables grade, size, stage, ER status and copy number aberrations (percentage). Copy number aberrations (percentage), treated both as a continuous and dichotomous variable, was an independent predictor of overall survival, along with grade and stage (Supplementary Table 1A). For dichotomized analysis, cutoffs at 5% (median) and at 11% (third quartile) gave significant *P*-values ($P = 0.042$ and $P = 0.003$, respectively) for copy number aberrations in the multivariate Cox proportional hazards model. This result was not affected by whether patients received adjuvant endocrine therapy (results not shown).

Table 1 Clinical information of the tumor groups segregated by total genomic changes

Clinical parameters	*Tumor subgroups*		P-*value*
	< 5%	*>5%*	*Fisher's exact test*
Samples	82 (55.4%)	66 (44.6%)	
NPI score			
⩽3.4	40 (48.8%)	23 (34.8%)	0.097
>3.4	42 (51.2%)	43 (65.2%)	
Grade			
I+II	57 (69.5%)	34 (51.5%)	**0.028**
III	25 (30.5%)	32 (48.5%)	
Distant metastasis			
No	66 (80%)	48 (72.7%)	0.33
Yes	16 (20%)	18 (27.3%)	
ER			
Negative	27 (33%)	18 (27.3%)	0.48
Positive	53 (64.6%)	47 (71.2%)	
NI	2 (2.4%)	1 (1.5%)	

Abbreviations: ER, estrogen receptor; NI, no information and cases excluded from analysis; NPI, Nottingham Prognostic Index.

In total, 3876 copy number changes were detected: 1866 (48%) gains and 2020 (52%) losses (Supplementary Excel File 2). For each clone, we calculated the number of tumors for which there was gain or loss and summarized it as a percentage of the 148 tumors studied (Figure 2a). To highlight the most frequently altered regions, we set a threshold of 15% and this identified chromosomal loci with known oncogenes and TSG. This was not surprising as the array we used contains mostly genes that have been previously shown to be altered in cancers. Chromosomal regions (more than one clone showing gain/loss as highlighted in Figure 2b) that were most commonly gained were 8q11–qtel (seven clones, range 11–33%), 1q21–qtel (seven clones, range 7–36%), 17q11–q12 (four clones, range 11–18%) and 11q13 (five clones, range 8–15%). The most common losses were found at 16q12–qtel (six clones, range 7–35%), 11p15.5–ptel (three clones, range 16–19%), 1p36–ptel (three clones, range 4–18%), 17p11.2–p12 (four clones, range 9–19%) and 8p22–ptel (five clones, range 8–18%).

We identified novel aberrations at the telomeric ends of chromosome 1p and 1q, 5p, 11p and 16q. The clones that were most frequently deleted mapped to 16qtel (43.9%) (identified by marker stSG30213) and to 16q24.3 (43.2%; containing *FANCA*). The clones most frequently gained were SHGC-18290 (31.8%) at 1qtel and *EXT1* (31.8%) at 8q24. The only clone that did not register any changes was *PDGFRA*, localized on 4q12.

Amplification (fluorescence ratios (FR) ⩾1.6; $\log_2$ FR ⩾0.67) of at least one of the known breast cancer oncogenes was observed in 55 tumors: *ERBB2* (24/148,

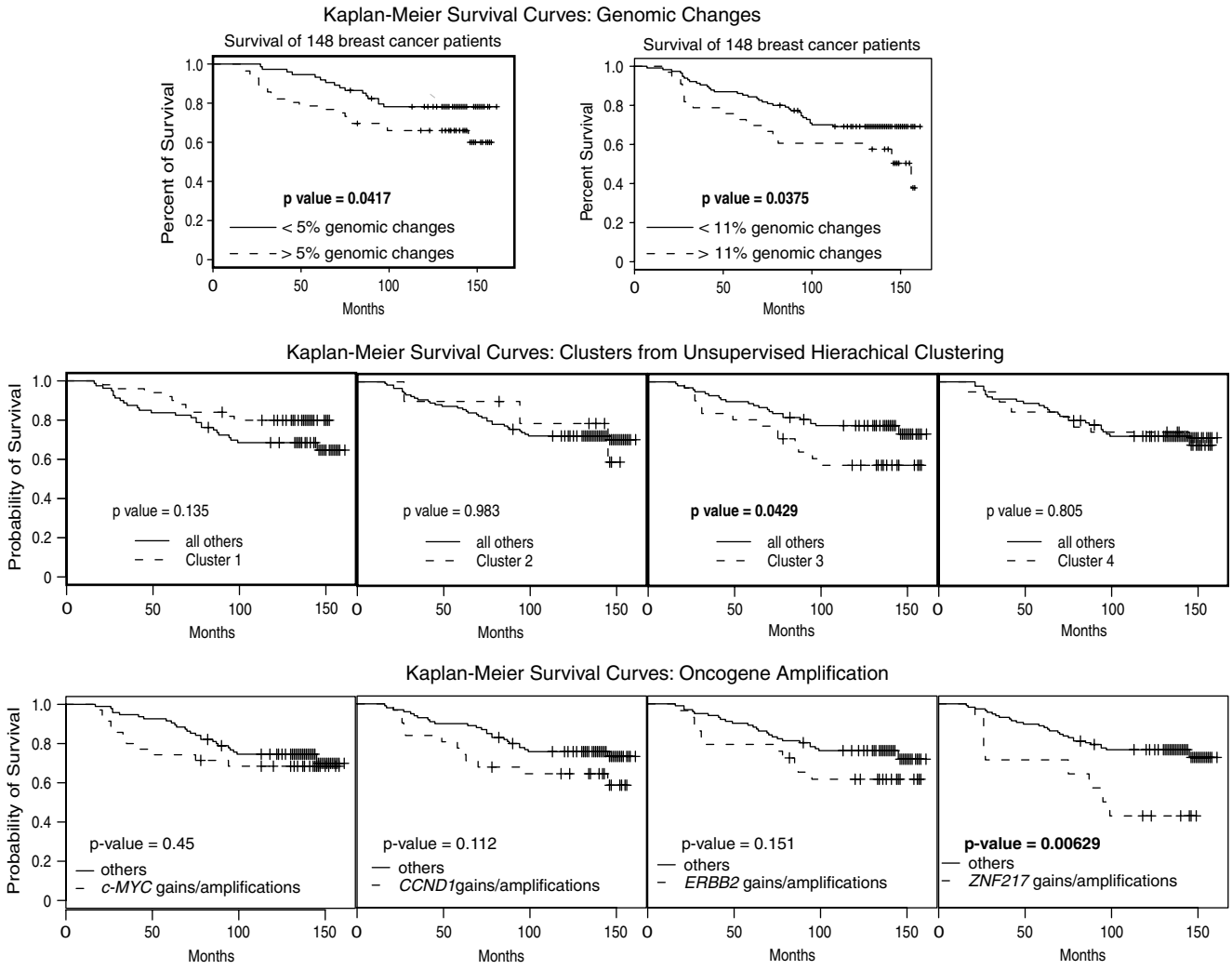


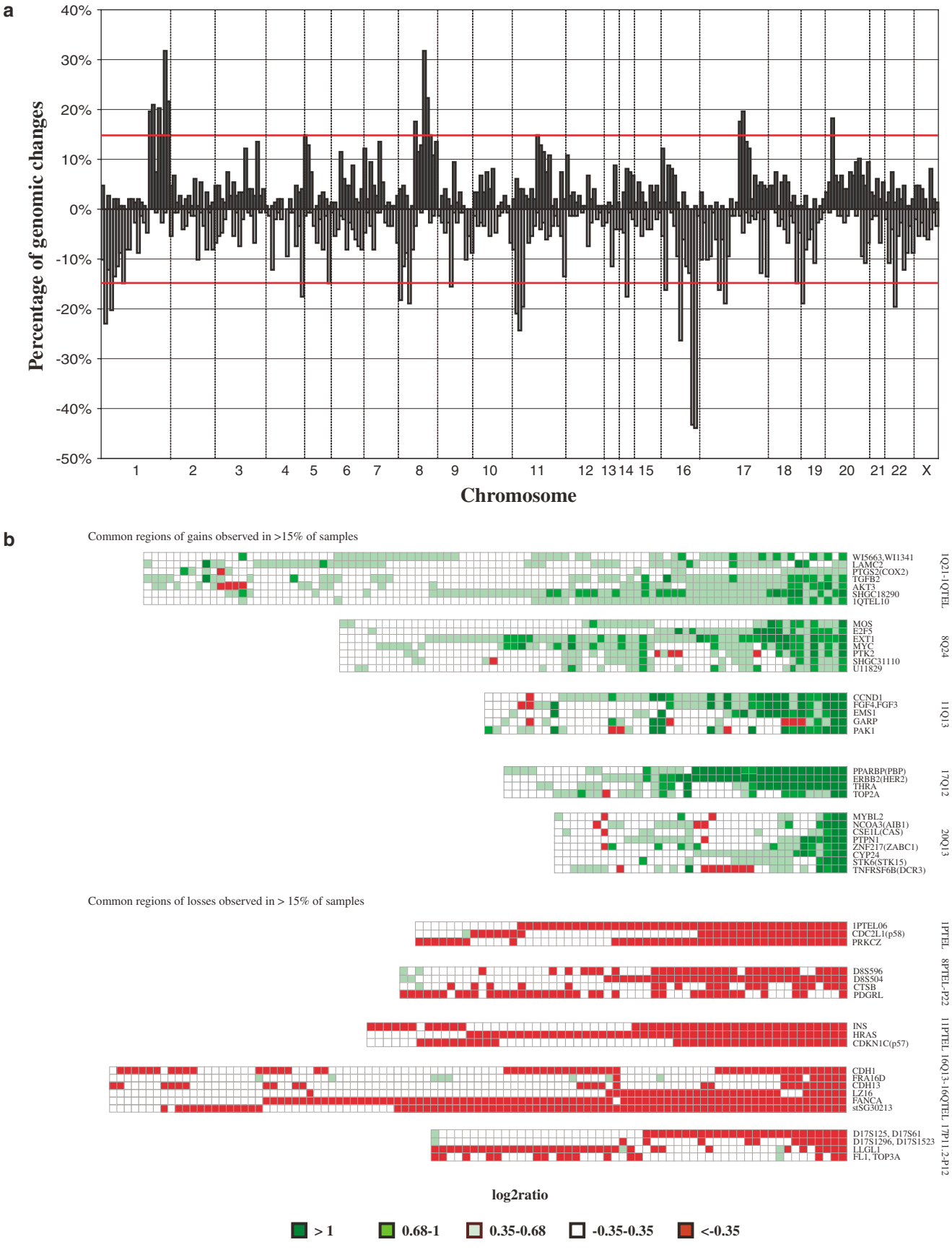
Figure 1 Kaplan–Meier plots generated for different genomic parameters. Top panels: Overall survival of 148 breast tumors based on percentage of genomic changes (left, tumors with less than 5% changes vs remainder and right, tumors with less than 11% changes vs remainder). Middle panels: overall survival analysis of 148 tumors based on the four clusters generated from unsupervised clustering (refer to Figure 3). Bottom panels: overall survival analysis of 148 tumors based on the gains/amplifications in 8q24, 11q13, 17q12 and 20q13.

16.2%), *CCDN1* (17/148, 11.5%), *C-MYC* (16/148, 10.8%) and *ZNF217* (10/148, 6.8%). These tumors tended to have poor prognostic features: grade 2/3 (52/55) and high NPI (46/55). Co-amplification of two or more of these oncogenes was uncommon: 13 tumors (8.8%) showed co-amplification of two (most common *ERBB2* and *ZNF217* in four cases), two tumors (1.3%) showed co-amplification of three and no tumors had simultaneous amplification of all four. We also noted that for all these genes, there were several cases with high level amplification (FR $\geqslant$3.0; $\log_2$ FR $\geqslant$1.58).

Other genes that registered high level amplifications (FR > 3), similar to those observed with known breast cancer oncogenes, in at least one case were: *EGFR* (7p12), *CDK4* (12q13), *FGFR2* (10q26), *HTR1B* (6q13), *DMBT1* (10q26), *AKT1* (14q32), *MYB* (6q22), *MDR1* (7q21), *GLI* (12q13), *YES1* (18p11), *HRAS* (11p15), *TK1* (17q23.2-q25.3) and *DCC* (18q21.3). The precise mapping and gene content of all the high level amplifications, as well as other copy number changes detected, is beyond the scope of this study given the low density of the array used.

Comparison with a published array-CGH data set

Nessling *et al*. (2005) reported a study using an array with 422 clones to analyse 31 formalin-fixed, paraffin-embedded advanced breast cancers with lymph node involvement. We identified 50 loci that were similar in both studies (clones were matched by gene content or genomic location based on Genome Assembly NCBI35) and performed a *t*-test to check if the differences in the observed percentage of gains/losses were statistically different. Of these clones, the difference observed in 22 (44%) was not statistically significant. Significant differences were found in 28 (56%) clones, where 25 showed gains and three showed losses.


a
Percentage of genomic changes
40%
30%
20%
10%
0%
-10%
-20%
-30%
-40%
-50%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome
b
Common regions of gains observed in >15% of samples
WI5663,WI1341
LAMC2
PTGS2(COX2)
TGFB2
AKT3
SHGC18290
1QTEL10
1Q21-1QTEL
MOS
E2F5
EXT1
MYC
PTK2
SHGC31110
U11829
8Q24
CCND1
FGF4,FGF3
EMS1
GARP
PAK1
11Q13
PPARBP(PBP)
ERBB2(HER2)
THRA
TOP2A
17Q12
MYBL2
NCOA3(AIB1)
CSE1L(CAS)
PTPN1
ZNF217(ZABC1)
CYP24
STK6(STK15)
TNFRSF6B(DCR3)
20Q13
Common regions of losses observed in > 15% of samples
1PTEL06
CDC2L1(p58)
PRKCZ
1PTEL
D8S596
D8S504
CTSB
PDGRL
8PTEL-P22
INS
HRAS
CDKN1C(p57)
11PTEL
CDH1
FRA16D
CDH13
LZ16
FANCA
stSG30213
16Q13-16QTEL
D17S125, D17S61
D17S1296, D17S1523
LLGL1
FL1, TOP3A
17P11.2-P12
log2ratio
> 1
0.68-1
0.35-0.68
-0.35-0.35
<-0.35

The clones that showed significant differences may reveal genes/loci that are associated with progression, especially when more than one clone within the same chromosomal band was found to be statistically different, suggesting that this difference did not occur by chance alone. Using this criterion, the most striking loci with higher frequency of gains in advanced tumors were 12q12–q15 with five clones (*WNT1*, *CDK2*, *GLI*, *CDK4* and *MDM2*), 8q24 with two clones (*c-MYC* and *PTK2*), 17q12 with two clones (*PPARBP* and *ERBB2*) and 20q13 with two clones (*NCOA3* and *MYBL2*). There was only one locus, 11q22–23, with two clones (*ATM* and *MLL*) registering more losses in advanced tumors.

Unsupervised hierarchical clustering based on profiling of copy number changes reveals molecular portraits of breast cancer

We binned the copy number changes into discrete $\log_2$-transformed FR intervals to represent the gene copy number changes (binning criteria: >1 = high level amplifications; 0.68–1.0 = amplifications; 0.35–0.68 = gains; -0.35 to 0.35 = no changes; <-0.35 = loss) as discrete events instead of as continuous, and used these binned $\log_2$ values to generate a 'pseudo'-heat map in essence similar to the heat maps used in expression studies. The tumors were grouped using unsupervised hierarchical clustering based on the similarity of the copy number changes, whereas the 281 clones were ordered according to their cytogenetic location (Figure 3).

Four main clusters were identified, with reasonable reproducibility (robustness indices: 45% for cluster 1, 73% for cluster 2, 71% for cluster 3 and 55% for cluster 4).

Cluster 1 (red) encompassed 57 tumors, of which 28 cases belonged to the low copy number change group (28/57). The most common aberrations in tumors belonging to this cluster were: loss of at least one clone on 16q (37/57) and gain of at least one clone on 1q (29/57). These tumors were mostly ER positive, grades1/2, in the low NPI group ($\leqslant 3.4$) and most had no distant metastases (76%). Within cluster 1, two smaller subclusters were defined: subcluster A (6/57), showing 8p loss, 8q gains, 16p gains and 16q loss; and subcluster B (13/57) with a relatively larger number of changes. The overall survival of patients in both subclusters was not different from the rest ($P=0.745$).

Cluster 2 (magenta) was the smallest in number (10) but these tumors had the highest number of clones registering gains/losses (median = 80.5). In addition, these tumors had quite distinct copy number changes: 7p gain, 9p loss, 11p gain, 20p gain and Xp loss. There were also specific loci changes: loss of *BMI1* (10p13), gains of *GLI* (12q13), D13S25 (13q14), *AKT1* (14q32), *YES1* (18p11) and *DCC* (18q21). Surprisingly, given the number of chromosomal aberrations, these tumors tended to be of lower grade (1/2) and ER positive but had mixed NPI.

Cluster 3 (green) includes 35 tumors, most with amplification in one of three chromosomal regions known to harbor breast cancer oncogenes: 11q13, 17q12 and 20q13. The most interesting observation within cluster 3 was that tumors tended to be grouped in subclusters of mutually exclusive amplification at each of the three loci: subcluster C (7/35), with amplification of 11q13 (*CCDN1*, *FGF4*, *EMS1*, *GARP* and *PAK1*); subcluster D (7/35), with amplification of 20q12–qtel (*NCOA3*, *MYBL2*, *CSE1L*, *PTPN1*, *STK6*, *ZNF217* and *CYP24*) and subcluster E (21/35), with amplification of 17q12–21 (*PPARBP*, *ERBB2*, *THRA* and *TOP2A*). Survival analysis showed that patients in cluster 3 have worst overall survival compared to all the other patients (HR = 1.98 with 95% CI: 1.01–3.9; log-rank $P=0.0429$; Figure 1; middle row). This poor outcome was not surprising as most of these tumors were of histological grade 3 (69%) and 34% of the patients developed distant metastases during follow-up. To access the relationship between clinical variables such as grade, size, stage and ER status and cluster 3 (with gains in 11q13, 17q12 and 20q13), a multivariate Cox proportional hazards model (Supplementary Table 1B) with stepwise regression and interaction terms between variables was used. Cluster 3 was found to be a predictor of overall survival ($P=0.02$) together with grade ($P=0.0001$) and stage ($P=0.001$), but there was a significant interaction between membership in cluster 3 and tumor grade ($P=0.02$). Cluster 3 is, therefore, not a truly independent prognostic variable.

Most tumors with 17q12 amplification (subcluster D) did not appear to have many other major genomic alterations: 47% (18/38) had less than 5% of total copy number changes. In this subcluster, tumors were mostly ER negative with poorer prognostic features (NPI > 3.4 and grade2/3) and there was a higher proportion of premenopausal women (41%). The survival of these patients tended to be shorter, but this was not statistically significant ($P=0.151$; Figure 1, bottom row). Tumors with amplification of either 11q13 or 20q12–q13 genes were grouped in two separate subclusters within cluster 3. Both subclusters contained tumors with more than 5% total copy number changes. Tumors in both subclusters appeared to have similar features: there was a predominance of ER positive and high grade tumors, but only cases with 20q13 amplification had significantly shorter overall survival (HR = 2.85 with 95% CI: 1.30–6.24; log rank $P=0.00629$; Figure 1; bottom row). However, Cox regression multivariate analysis showed that 20q13 amplification was not an independent predictor of overall survival, when tumor

Figure 2 (**a**) Global view of copy number gain and loss in 148 breast tumors. The frequency of changes (gain ■ plotted above 0 and loss ■ plotted below 0) observed is represented as bars at each of the genomic positions of the BAC clones. A threshold line of 15% was set to identify chromosomal regions with common changes. (**b**) Heat map of the clones within chromosomal regions registering changes in more than 15% of the samples.

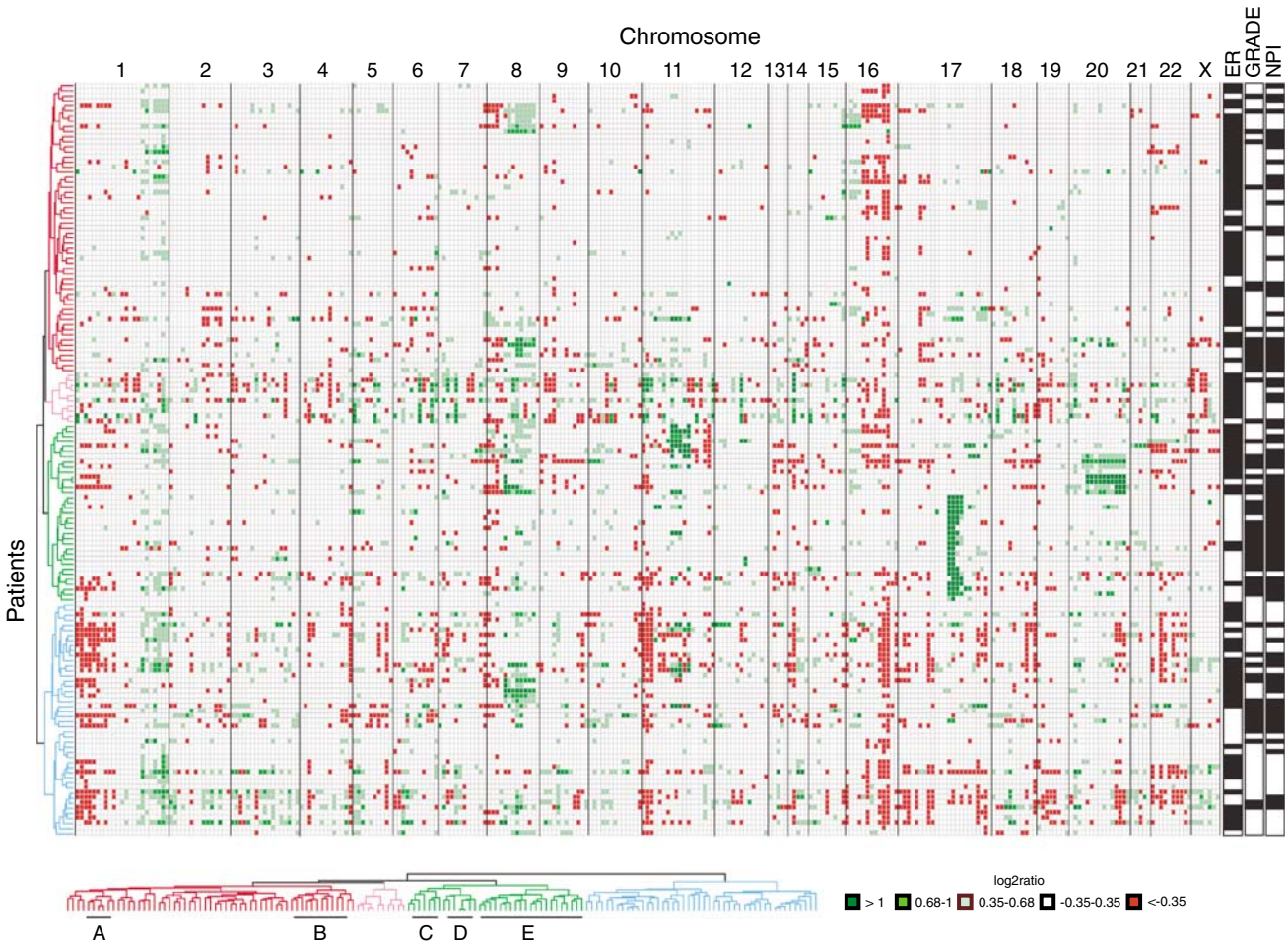


Figure 3 Unsupervised hierarchical clustering of 148 breast tumors. Each column represents a clone and each row a tumor. A dendrogram shows the clustering of tumors based on their genomic aberrations. An enlarged version of the dendrogram is depicted at the bottom left with five groups (**a**–**e**) highlighted. On the right, the bars represent the clinical features of each tumor: ER negative □, ER positive ■, Grade 1 + 2 □, 3 ■, NPI low □, NPI high ■. The heat map represents the fluorescence ratios (test/reference) on a $\log_2$ scale as shown at the bottom bar (red, loss; green, gain; white, no change).

grade, stage, size and ER status were also considered (Supplementary Table 1C).

In contrast to the other three oncogenes present in cluster 3, tumors with gains in the fourth known breast cancer oncogenic locus, 8q24 (*EXT*, *C-MYC* and *PTK2*), do not belong to a single, discreet cluster but are present in all four main clusters. Gains in 8q24 do not occur as an isolated event and indeed, most of these tumors (40/55; 73%) have more than 5% genomic changes.

Cluster 4 (blue) contained 46 tumors. Common alterations in this cluster were gains on 1q and losses on 1p, 11p, 16q, 17p and chromosome 22. There was no relationship with ER status, grade and NPI.

To analyse the relationship between gain/loss of individual clones in relation to the total number of changes in each tumor, we ranked the tumors according to their total genomic changes (Supplementary Figure 1). The main findings in tumors with less than 1% copy number changes were gain of 1q and loss of 16q. This suggests these two alterations occur together as an early event in tumor development, assuming a correlation between progression and increase in genetic aberrations. Alternatively, these tumors are different from more aggressive tumors and develop along a different genomic pathway. These tumors were mostly ER positive and were classified in the good NPI group. In tumors with total copy number changes between 1 and 5%, amplification of 17q12 and gains in 8q24 occurred in addition to changes on 1q (gains) and 16q (losses). In the remaining tumors with more than 5% copy number changes, 1p loss, 8p loss, 11p loss, 11q13 gains and 20q13 gains were common. Interestingly, certain changes were only observed in tumors with >20% genomic changes: gains on 5p, 14qtel, 20p and chromosome X; and losses on 6p, 17p, chromosome 19, 20qtel and chromosome 22.

Supervised classification using genes filtered by significance analysis

An important application of microarray technology in cancer is class prediction where classifiers can be

constructed that may reliably indicate subtype, invasiveness potential, expected progression and the best treatment strategy. To build our classifiers, we identified genes that were most closely associated with clinical phenotypes. Based on a modified two-sample *t*-test for each gene, 10000 permutations of each clinical grouping of tumors were performed to test the significance of the observed t-statistic. Only genes that have a false discovery rate (FDR) of less than 10% (q-value < 0.1) were selected. We found a minimum of 10 genes fulfilled the criteria defined in the significance analysis for each of three clinical variables. These were Grade ($1+2$ vs 3) with 61 genes (Figure 4), ER (positive vs negative) with 36 genes and NPI (NPI $\leqslant 3.4$ vs NPI > 3.4) with 31 genes (data not shown). Only eight genes were common for the three clinical variables and these were from only two chromosomal loci: 16q22–qtel (*CDH1*, *CDH13*, *LZ16*, *FANCA* and 16qtel013) and 17q12–21 (*PPARBP*, *ERBB2* and *THRA*). Surprisingly, genes on the 11q13 amplicon did not feature among those found to be significantly different statistically.

Using the genes selected from the significance analysis and leave-one-out cross-validation on the training set, we built KNN classifiers for predicting ER, histological grade and NPI on test samples. The average misclassification rates (over multiple runs) for ER, Grade and NPI on test samples were 24.7, 25.7 and 35.7%, respectively, in comparison to 18.3, 21.2 and 30.1% on the training set (through the leave-one-out cross-validation).

Discussion

In this study we used array-CGH to evaluate the copy number changes in 148 well-characterized breast cancers with a minimum of 5 years follow-up. We also examined the associations between genomic alterations and clinical phenotype of the tumors. Recently, several array CGH studies in breast cancer have been published (Naylor *et al*., 2005; Fridlyand *et al*., 2006; Stange *et al*., 2006), but our study is the largest breast cancer set studied to date. The main limitation of our study was the lower genomic density of the arrays used as compared to the other studies. Despite this limitation, we observed a similar pattern of genomic changes occurring in at least 5% of our samples with the previously published conventional and array-CGH studies (Kallioniemi *et al*., 1994; Tirkkonen *et al*., 1998; Roylance *et al*., 1999; Cingoz *et al*., 2003; Loo

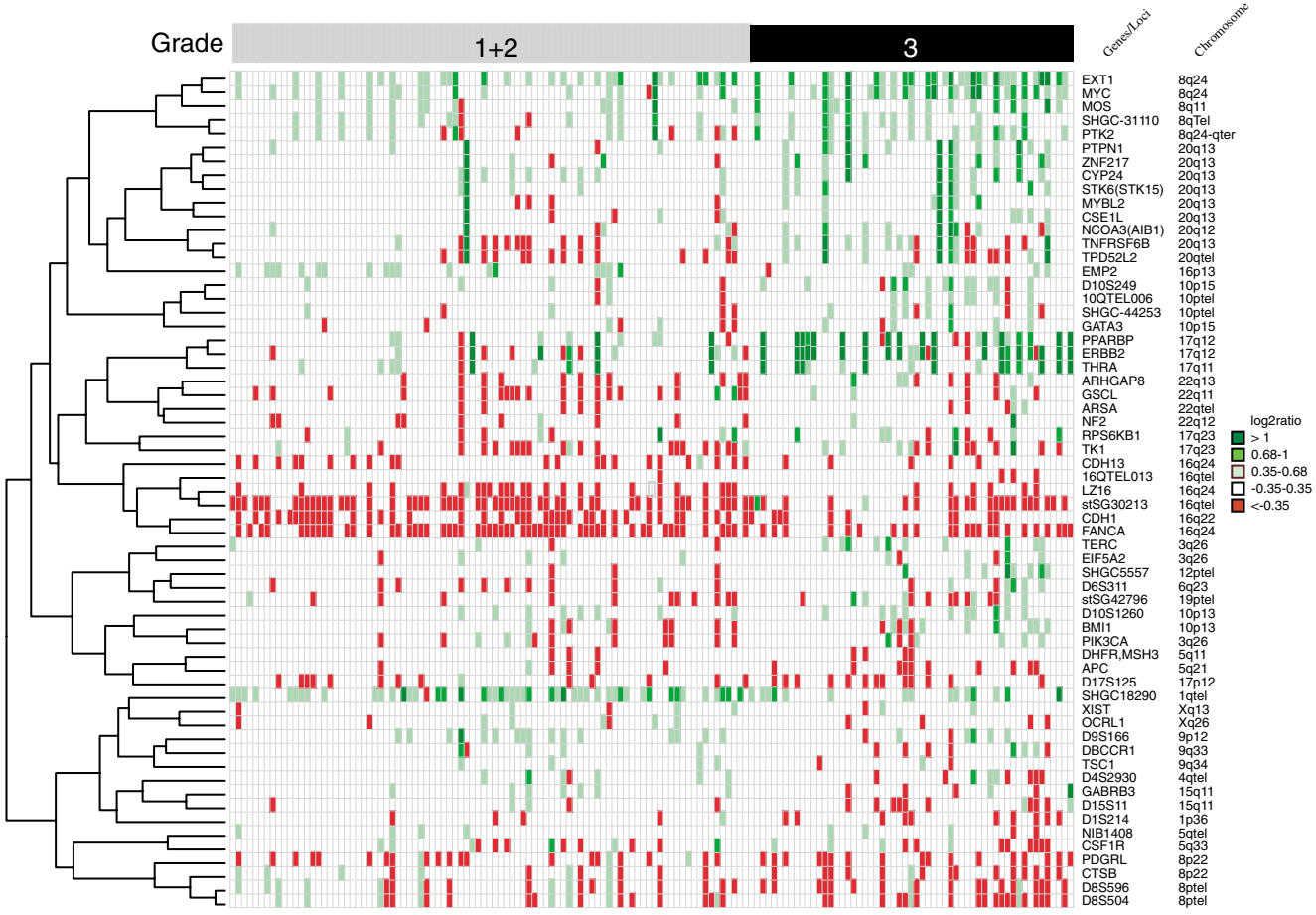


Figure 4 Hierarchical clustering of tumors using the 61 genes from the grade *k*-nearest neighbor classifier. Fluorescence ratios are depicted as in Figure 3.

et al., 2004; Naylor *et al*., 2005; Fridlyand *et al*., 2006; Stange *et al*., 2006). In addition, the low resolution of the platform did not hinder our ability to generate important novel observations about patterns of genomic alterations in human breast cancers. Studies that have looked at a greater number of samples used Southern blotting (Chin *et al*., 2001; Maass *et al*., 2002) and FISH (Al-Kuraya *et al*., 2004) but analysed significantly fewer genes. It is perhaps more important that the patient cohort used in our study is more representative of the usual clinical presentation of breast cancer (mostly postmenopausal, ER +, small size and lymph node negative) and was of a sufficient size to allow for meaningful correlations with clinical and pathological characteristics. This sharply contrasts with the studies that used higher density arrays, which included small patient numbers (Stange *et al*.) or very selected patient cohorts with clinically very aggressive tumors (for example in Fridlyand *et al*., 30/55 patients died during follow-up and 34/55 were lymph node positive and in Naylor *et al*., 33/46 were large/lymph node positive and 42% Her2 +), precluding the type of robust correlative analysis that we were able to do.

Aneuploidy is a frequent event in breast cancers (Wenger *et al*., 1993) and reflects genomic instability. In our cohort, we found about one-half of the cases had less than 5% genomic changes and this group had better clinical outcome. We also observed that the concurrent gain of 1q and loss of 16q occurred in tumors with a relatively stable genome, suggesting that this may be an early event. This would correlate with cytogenetically described changes as translocation involving both chromosomes 1 and 16 and isochromosome 1q (Tsarouha *et al*., 1999). These tumors in our series had good prognostic features and were usually ER positive, as previously described (Rennstam *et al*., 2003; Loo *et al*., 2004; Fridlyand *et al*., 2006).

In cases with higher number of genomic changes, gains in 8q24 and 17q12 were evident although these rarely occurred together. Other alterations such as losses on 1p and gains in known oncogenic regions were also observed. There were changes that were only observed when the total genomic changes were > 20%, suggesting these are events, which only occur in an unstable genome. These findings argue against a linear progression of accumulation of genomic aberrations. Tumors with different amounts and combinations of genomic aberrations probably originated from different cell populations (for example, luminal vs basal) and progress by accumulation of distinct genomic events at different rates of accumulation reflecting underlying genomic instability and clonal selection.

Gains of 1q, 8q and loss of 16q were the most frequent alterations occurring in more than 30% tumors. At 1q, the clone most frequently gained was RP1-407H12, containing the marker SHGC-18290 and several olfactory receptor genes. On chromosome 8q, the clone most frequently gained was surprisingly not *c-MYC* but *EXT1*. The distance between both genes is approximately 10 Mb, suggesting two separate amplicons. *EIF3S3*, located approximately 1 Mb proximal to *EXT1* but not present on our arrays, has been suggested to be the true target of these amplifications since it was consistently overexpressed in breast and prostate cancers (Nupponen *et al*., 2000). There appear to be two separate regions of loss on 16q: one around *CDH1* (which encodes E-cadherin) on 16q22 and another spanning 16q24–tel. The clone on 16q24–qtel most frequently lost was RP11-133L7, which contains marker stSG30213, *AFG3L1*, *MC1R*, *TUBB3*, *KIAA1049*, *SPIRE2* and the growth-arrest-specific protein 8 (*GAS8*) gene. *AFG3L1* is a human homolog of family gene 3-like 1, which encodes a mitochondrial ATP-dependent zinc metalloprotease and recently has been suggested to be a target of estrogen (Wang *et al*., 2004). Variants of the melanocortin-1 receptor gene (*MC1R*) have been associated with melanoma risk and progression (Landi *et al*., 2005). The GAS protein family is thought to be modulators of cell cycle progression and survival but no association with breast carcinogenesis has yet been established. Another clone on 16q that was frequently lost was RP11-368I7, which contains several genes namely *FANCA*, *ZFP276*, *RPL13*, *CDK10*, *SPG7*, *DPEP1* and *CPNE7*. The role of the Fanconi anemia (FA) proteins in breast cancer has been actively investigated in recent years. FA proteins, including FANCA, are required in a DNA damage response pathway and current data indicate that FA is a central node in a complex nuclear and cytoplasmic network of tumor suppressor and genome stability pathways (Bogliolo *et al*., 2002). Owing to the low resolution of the arrays used here, we recognize that the identity of the true targets of some of these copy number changes is yet to be firmly determined.

Besides gains and losses involving genes previously implicated in breast cancer, we identified novel loci. The short arm of chromosome 5 showed gains at 5p15 and losses at 5ptel. Gains on 5p12–14 have been observed but no target genes have yet been identified (Tirkkonen *et al*., 1998; Korsching *et al*., 2004). Two clones containing markers D5S23 and D5S2064, the cytogenetically identified critical region for the *Cri-du-chat* syndrome, were frequently gained. We noticed frequent loss of a single clone of the telomeric end of chromosome 5 containing marker C84C11/T3. The loss of 5ptel in breast cancer has not been reported previously. We found frequent gains of a single clone at 20p, CTC-334G22, which contains the *JAG1* gene that functions as a ligand for multiple Notch receptors and is involved in the mediation of Notch signaling. This may mediate estradiol-induced angiogenesis (Soares *et al*., 2004) and is frequently observed in epithelial-to-mesenchymal transitions in advanced carcinogenesis (Zavadil *et al*., 2004). Other reports have identified low frequency gains of 20p in breast cancer (James *et al*., 1997; Gunther *et al*., 2001). Again identifying the true target of these events will require the use of higher density arrays.

When our data were compared with a publicly available array-CGH study of advanced breast cancers, which used custom arrays comprising 422 mapped genomic sequences (Nessling *et al*., 2005), 56% of the 50 clones that were 'genomic overlaps' had significantly

different copy number changes, suggesting that these gene/loci are involved in breast cancer progression and aggressiveness. We also compared the frequency of copy number gain at selected loci that have been previously studied using FISH on tissue microarrays. For one of these studies (Al-Kuraya *et al*., 2004), we found no difference for *ERBB2* (16.2 vs 17.3%; $P=0.735$) and *EGFR* (2.7 vs 0.8%; $P=0.154$), a higher frequency of *c-MYC* gains (10.8 vs 5.3%; $P=0.036$) and a lower frequency of *CCDN1* gains (11.5 vs 20.1%; $P=0.002$). In comparison with a previous study from our group (Callagy *et al*., 2005), we found differences in amplification of *ERBB2* (27 vs 16.2%), *TOP2A* (14 vs 6.1%), *EMS1* (26 vs 9.5%) and *CCNE1* (6 vs 0.7%). These differences are most likely due to underlying biases from patient selection. The series analysed here is most representative of the demographics of commonly occurring breast cancers and therefore, the results are more generally applicable.

Unsupervised hierarchical clustering based on gene expression data has been used to characterize the molecular portraits of breast tumors, with consistent observation of distinct molecular subtypes associated with clinical and pathological features of the tumors (Perou *et al*., 2000; Sorlie *et al*., 2001; van't Veer *et al*., 2002). We used the same clustering algorithm to group tumors on the basis of the pattern of gene copy number changes along the genome. This array-CGH-based molecular portrait revealed four main clusters, which differed for some of their clinical features, but only the cluster with amplification of known oncogenes had worse survival. Notable features of these portraits were:

1. the amplicons on 11q13, 17q12 and 20q13 were usually mutually exclusive;
2. amplification on 17q12 and loss on 16q22–qtel were also mutually exclusive;
3. amplification of 8q24 region can occur with amplification of 11q13 or 20q13 but usually not with 17q12;
4. the 17q12 amplicon occurs in tumors with few copy number changes;
5. ER-negative tumors are predominantly 17q12 amplified, whereas ER-positive tumors are amplified for 8q, have gains in 1q, losses in 1p and 16q;
6. amplification of 20q13 occurs predominantly in ER-positive tumors with bad prognosis.

There has been widespread use of gene expression data to predict patient survival or other features (for example histological grade or ER status) in breast cancers. Using supervised approaches, we were able to identify a number of genes that could be used to build a classifier for three clinical variables only: ER, Grade and NPI. The performance of these three classifiers was not optimal but showed promise: average error rates of less than 36% in the test set. To date, no array-CGH study has attempted to produce a supervised classifier based on genome-wide copy number changes. A previous conventional CGH study produced a classifier that separated tumors based on cytokeratin-14 (CK14) status with an error rate of 24% (Jones *et al*., 2004). We attribute our high error rate to the small number of clones on the array and predict that higher resolution array data will produce classifiers that may be more robust than those derived from gene expression profiling (Glinsky *et al*., 2004). We were unable to produce a similar classifier for survival but this was not surprising given the small number of events in our cohort. Future studies besides using higher density arrays will need to include larger numbers of patients if we are to find signatures predictive of survival using array-CGH. Nevertheless, we believe the data presented here, together with previous reports in smaller cohorts, show the value of DNA-based profiling as a tool for molecular classification of breast cancer and shows the potential for predicting patient survival.

Materials and methods

Tumor and control samples

The well-characterized cohort of primary breast cancers (see Table 2) was obtained with appropriate ethical approval from the Nottingham Tenovus Primary Breast Carcinoma Series. All 148 samples were primary operable invasive breast carcinomas collected between 1990 and 1996. Tumor tissue specimens were stored at $-80^{\circ}C$ until use. Whole tissue sections were used for DNA extraction. Tumor cellularity (average 52%; range: 20–100%) was evaluated by determining the percentage of the surface area occupied by cancer cells in hematoxylin–eosin-stained sections. In most cases, the non-tumor components were either fat or fibrous tissue and therefore, the percentage of tumor nuclei was higher than estimated cellularity. Nevertheless, only 10% (15/148) of cases had tumor cellularity lower than 40%, which is the threshold for detecting single copy changes (Hodgson *et al*., 2001). Furthermore, there was no correlation between degree of copy number changes detected and cellularity (data not shown), supporting the robustness of the findings reported here. Cell lines with known chromosome copy number changes (trisomy 7, 13, 18, 1X, 2X, 3X) and peripheral blood white cells from 13 normal females (with ethical approval) were used as control samples.

Genomic DNA isolation and labeling

Tumor DNA was prepared from twenty 30-μm-thick frozen sections using homogenization in Trizol (Invitrogen, Paisley, UK) following manufacturer's instructions. Control DNA was prepared from cell pellets using standard proteinase K/sodium dodecyl sulfate (SDS) method.

Labeling of DNA was carried out using the Genosensor Random Prime kit (Vysis, Downer's Grove, Chicago, USA) and prepared for hybridization following the manufacturer's instructions. Briefly, 100 ng of DNA was digested with DNase to generate fragments of approximately 100–300 base pairs and labeled using random primed labeling with Cyanine 3-dUTP (tumors and cell lines) or Cyanine 5-dUTP (pooled normal female reference DNA provided by Vysis).

Microarray hybridization and data capture

The Genosensor arrays used (Vysis, Downer's Grove, USA) contain 287 target clone DNA (P1, PAC or BAC clones) representing loci previously shown to be important in cancer or involve in congenital syndromes and telomeres (Supplementary Excel File 1) and printed in triplicate. Target clones were identified either from human genome mapping

Table 2 Demographics of patients used in this study

Parameter	*Numbers*
No of patients	148
Median age, years (range)	58 (35–70)
Survival/months, median (range)	134 (7–161)
Tumor size, cm (range)	1.8 (0.1–4.5)
Patient status (%)	
Alive	93 (63)
Dead from breast cancer	37 (25)
Dead from other causes	15 (10)
Lost to follow up	3 (2)
Menopause status (%)	
Pre	48 (32)
Post	100 (68)
Pathologic grade (%)	
I	34 (23)
II	57 (39)
III	57 (39)
Lymph node status (%)	
Negative	103 (70)
Positive	45 (30)
Recurrence (%)	
No	100 (68)
Yes	48 (32)
Distant metastases (%)	
No	114 (77)
Yes	34 (23)
Vascular invasion (%)	
No	102 (69)
Yes	46 (31)
ER status (%)	
Positive	100 (68)
Negative	45 (30)
No Information	3 (2)
Nottingham Prognostic Index (%)	
$\leqslant 3.4$	63 (43)
>3.4	85 (57)

Abbreviation: ER, estrogen receptor.

information or by screening clone libraries with specific probes and/or PCR primers. The identity of each clone was confirmed with clone-specific PCR primers and by FISH where clones were assessed for their ability to produce only the expected number of signals on normal specimens at the expected cytogenetic location. The average resolution of the array is 10 Mb, at certain loci the resolution is higher than 1 Mb and the lowest resolution is 63 Mb.

The labeled DNA probes (normal and tumor) were mixed with microarray hybridization buffer containing human Cot-1 DNA. The probes were denatured at 80°C for 10 min followed by incubation at 37°C for an hour, and then hybridized onto the array for 72 h at 37°C in a humidified chamber. Following hybridization, slides were washed sequentially in 50% formamide/2 × SSC at 45°C (3 × 10 min), 1 × SSC at room temperature (4 × 5 min), and a brief rinse in distilled water. The slides were counterstained with 20 μl 4′6 diamidinophenylindole (DAPI) IV solution (Vysis). Each hybridized slide was captured using the Genosensor 300 microarray scanner (Vysis, USA).

Data normalization and validation of array platform

The GenoSensor software (Vysis, USA) segments each target using the DAPI image plane. Mean intensities were measured from the CY3 and CY5 image planes, background was subtracted, a mean ratio of green/red signal was determined, and the ratios were normalized. The normalized ratio for each target was calculated relative to the modal DNA copy number, and the statistical significance of each change was reported as a *P*-value (Piper *et al*., 2002). A *P*-value of <0.01 indicated a significant difference between the copy numbers of a target and the modal clones. Using the default settings of the Vysis Genosensor, individual clones were only accepted if the coefficient of variation between the triplicate spots was less than 10 and the average correlation coefficient $\geqslant 0.9$. To rescue data from clones that were rejected, the triplicate spots were inspected and if only a single of the three spots was inadequate (usually due to fluorescence artifacts) that spot was manually removed and an average of the two remaining spots was used. Cell lines with known copy number gains and losses in several chromosomes were used to establish mean $\log_2$-transformed FR for single copy gain and loss: 0.38 (FR 1.35) and -0.42 (FR 0.75), respectively. Data from 13 hybridizations using DNA from normal female subjects showed that 99.8% of the spots fell within the boundaries for no copy number changes. Using results of FISH for *Cyclin D1* (*CCDN1*) in six breast cancer cell lines, we determined the correlation between absolute copy number and the $\log_2$ ratio obtained by array-CGH. Correlation was excellent ($R^2 = 0.95$) for the range analysed (2–6 copies by FISH; $\log_2$ FR of 0.0–1.04). Extensive validation of the array platform including dye swap experiments has been previously reported by us (Daigo *et al*., 2001; Callagy *et al*., 2005). In addition, copy number changes (gain up to six copies, loss of one copy and homozygous deletions) have been validated using FISH in both primary tumor touch prints and in metaphases from cell lines (data not shown).

Data analysis

The $\log_2$-transformed array-CGH data generated for 281 clones in 148 breast tumors were collated on a 281 × 148 data matrix. Data from six clones were not used: two from chromosome Y, two from X/Y and two clones (*DMD* and D21S341) for which the clones were absent in the majority of the arrays used due to printing problems. This data matrix was used with the clinical and pathological information for unsupervised clustering, supervised classification, selection of significant genes associated with clinical outcomes, and survival analysis.

Unsupervised classification. Unsupervised analysis was carried out using the complete linkage-clustering algorithm with a distance metric of one minus Pearson's correlation coefficient. To assess the reproducibility of individual clusters (or in other words, the stability of the observed clusters in the background of experimental noise), reproducibility measures such as the *R*-index (robustness) and the *D*-index (discrepancy) were used (McShane *et al*., 2002). Survival analysis of patients in the different clusters obtained by unsupervised clustering was carried out using Kaplan–Meier (KM) estimation, Cox proportional hazards regression and log-rank tests.

Supervised classification. First the association between copy number changes and binary clinical variables (ER±; NPI $<3.4/>3.4$; grade 1 vs grade 2/3 (G1-2/3); dead/alive) was tested using the analysis described as follows. First, each clone was tested for difference between the two states using a modified two-sample t-statistic (Smyth *et al*., 2003). Then, the

observed t-statistic was tested for significance against the null hypothesis of no difference between the two statuses using a permutation method (10,000 random permutations were done) to assign a *P*-value to the observed t-statistics. The *P*-values were then transformed into *q*-values based on Storey and Tibshirani, 2003 (2003). A *q*-value cutoff of 0.1 (equivalent to a *false discovery rate* (FDR) of 10%) was used to select genes significantly associated with clinical variables. The *q*-value is a measure of significance in terms of FDR rather than the FPR (*false positive rate*).

The *k* nearest neighbor (KNN) classifier was implemented using the genes selected as a result of the significance analysis. The148 breast tumors were randomly divided into a training set which included 80% of the tumors (118 samples), and a test set with 20% of the tumors (30 samples). The optimal number (*k*) of nearest neighbors used in the classifier was determined in such a way that for a given value of *k*, we calculated the prediction error rate by the leave-one-out cross-validation on the training set. The optimal *k*, which gives the lowest prediction error rate, was chosen. It ranged from 1 to 6 in our case. The optimal number of genes used to build KNN classifier was determined likewise by the above procedure, which ranges from 30 to 60 in our study. The KNN classifiers were then used to predict clinical variables such as ER, Grade and NPI on the test set.

Abbreviations

NPI, Nottingham Prognostic Index; ER, estrogen receptor; HR, hazard ratio.

Acknowledgements

We thank Vysis for providing all the reagents and Teresa Ruffalo for technical help**.** The work was funded by Cancer Research UK. NLB-M was supported by Fundação para a Ciência e a Tecnologia, Portugal (Fellowship SFRH/BD/2914/2000).

References

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.

Bogliolo M, Cabre O, Callen E, Castillo V, Creus A, Marcos R *et al*. (2002). The Fanconi anaemia genome stability and tumour suppressor network. *Mutagenesis* **17**: 529–538.

Callagy G, Pharoah P, Chin SF, Sangan T, Daigo Y, Jackson L *et al*. (2005). Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. *J Pathol* **205**: 388–396.

Chin SF, Wang Q, Puisieux A, Caldas C. (2001). Absence of rearrangements in the BRCA2 gene in human cancers. *Br J Cancer* **84**: 193–195.

Cingoz S, Altungoz O, Canda T, Saydam S, Aksakoglu G, Sakizli M. (2003). DNA copy number changes detected by comparative genomic hybridization and their association with clinicopathologic parameters in breast tumors. *Cancer Genet Cytogenet* **145**: 108–114.

Daigo Y, Chin SF, Gorringe KL, Bobrow LG, Ponder BA, Pharoah PD *et al*. (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.

Elston CW, Ellis IO. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**: 403–410.

Forozan F, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP. (1997). Genome screening by comparative genomic hybridization. *Trends Genet* **13**: 405–409.

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.

Gilchrist KW, Kalish L, Gould VE, Hirschl S, Imbriglia JE, Levy WM *et al*. (1985). Interobserver reproducibility of histopathological features in stage II breast cancer. An ECOG study. *Breast Cancer Res Treat* **5**: 3–10.

Glinsky GV, Higashiyama T, Glinskii AB. (2004). Classification of human breast cancer using gene expression profiling as a component of the survival predictor algorithm. *Clin Cancer Res* **10**: 2272–2283.

Gunther K, Merkelbach-Bruse S, Amo-Takyi BK, Handt S, Schroder W, Tietze L. (2001). Differences in genetic alterations between primary lobular and ductal breast cancers detected by comparative genomic hybridization. *J Pathol* **193**: 40–47.

Heidenblad M, Lindgren D, Veltman JA, Jonson T, Mahlamaki EH, Gorunova L *et al*. (2005). Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene* **24**: 1794–1801.

Hodgson G, Hager JH, Volik S, Hariono S, Wernick M, Moore D *et al*. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat Genet* **29**: 459–464.

James LA, Mitchell EL, Menasce L, Varley JM. (1997). Comparative genomic hybridisation of ductal carcinoma *in situ* of the breast: identification of regions of DNA amplification and deletion in common with invasive breast carcinoma. *Oncogene* **14**: 1059–1065.

Jones C, Ford E, Gillett C, Ryder K, Merrett S, Reis-Filho JS *et al*. (2004). Molecular cytogenetic identification of subgroups of grade III invasive ductal breast carcinomas with different clinical outcomes. *Clin Cancer Res* **10**: 5988–5997.

Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L *et al*. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci USA* **91**: 2156–2160.

Kim YH, Girard L, Giacomini CP, Wang P, Hernandez-Boussard T, Tibshirani R *et al*. (2005). Combined microarray analysis of small cell lung cancer reveals altered apoptotic balance and distinct expression signatures of MYC family gene amplification. *Oncogene* **25**: 130–138.

Korsching E, Packeisen J, Helms MW, Kersting C, Voss R, van Diest PJ *et al*. (2004). Deciphering a subgroup of breast carcinomas with putative progression of grade during carcinogenesis revealed by comparative genomic hybridisation (CGH) and immunohistochemistry. *Br J Cancer* **90**: 1422–1428.

Landi MT, Kanetsky PA, Tsang S, Gold B, Munroe D, Rebbeck T *et al*. (2005). MC1R, ASIP, and DNA repair in

sporadic and familial melanoma in a Mediterranean population. *J Natl Cancer Inst* **97**: 998–1007.

Loo LW, Grove DI, Williams EM, Neal CL, Cousens LA, Schubert EL *et al*. (2004). Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res* **64**: 8541–8549.

Maass N, Rosel F, Schem C, Hitomi J, Jonat W, Nagasaki K. (2002). Amplification of the BCAS2 gene at chromosome 1p13.3–21 in human primary breast cancer. *Cancer Lett* **185**: 219–223.

McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**: 1462–1469.

Naylor TL, Greshock J, Wang Y, Colligon T, Yu QC, Clemmer V *et al*. (2005). High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res* **7**: R1186–R1198.

Nessling M, Richter K, Schwaenen C, Roerig P, Wrobel G, Wessendorf S *et al*. (2005). Candidate genes in breast cancer revealed by microarray-based comparative genomic hybridization of archived tissue. *Cancer Res* **65**: 439–447.

Nupponen NN, Isola J, Visakorpi T. (2000). Mapping the amplification of EIF3S3 in breast and prostate cancer. *Genes Chromosomes Cancer* **28**: 203–210.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA *et al*. (2000). Molecular portraits of human breast tumours. *Nature* **406**: 747–752.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D *et al*. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**: 207–211.

Piper J, Stegenga S, Pestova E, Marble H, Lucas M, Wilber K *et al*. (2002). An objective method for detecting copy number change in CGH microarray experiments. *Proceedings of the Third Euroconference on Quantitative Molecular Cytogenetics. Rosenon, Stockholm, Sweden* 4–6 July 2002; Rosenon, Stockholm, Sweden, pp 109–114.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE *et al*. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* **99**: 12963–12968.

Rennstam K, Ahlstedt-Soini M, Baldetorp B, Bendahl PO, Borg A, Karhu R *et al*. (2003). Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization. *Cancer Res* **63**: 8861–8868.

Reyal F, Stransky N, Bernard-Pierrot I, Vincent-Salomon A, de Rycke Y, Elvin P *et al*. (2005). Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes – a study of 130 invasive ductal breast carcinomas. *Cancer Res* **65**: 1376–1383.

Roylance R, Gorman P, Harris W, Liebmann R, Barnes D, Hanby A *et al*. (1999). Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. *Cancer Res* **59**: 1433–1436.

Smyth GK, Yang YH, Speed T. (2003). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* **224**: 111–136.

Soares R, Balogh G, Guo S, Gartner F, Russo J, Schmitt F. (2004). Evidence for the notch signaling pathway on the role of estrogen in angiogenesis. *Mol Endocrinol* **18**: 2333–2343.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H *et al*. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**: 10869–10874.

Stange DE, Radlwimmer B, Schubert F, Traub F, Pich A, Toedt G *et al*. (2006). High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer. *Clin Cancer Res.* **12**: 345–352.

Storey JD, Tibshirani R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**: 9440–9445.

Tirkkonen M, Tanner M, Karhu R, Kallioniemi A, Isola J, Kallioniemi OP. (1998). Molecular cytogenetics of primary breast cancer by CGH. *Genes Chromosomes Cancer* **21**: 177–184.

Tsarouha H, Pandis N, Bardi G, Teixeira MR, Andersen JA, Heim S. (1999). Karyotypic evolution in breast carcinomas with i(1)(q10) and der(1;16)(q10;p10) as the primary chromosome abnormality. *Cancer Genet Cytogenet* **113**: 156–161.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M *et al*. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530–536.

Wang DY, Fulthorpe R, Liss SN, Edwards EA. (2004). Identification of estrogen-responsive genes by complementary deoxyribonucleic acid microarray and characterization of a novel early estrogen-induced gene: EEIG1. *Mol Endocrinol* **18**: 402–411.

Wenger CR, Beardslee S, Owens MA, Pounds G, Oldaker T, Vendely P *et al*. (1993). DNA ploidy, S-phase, and steroid receptors in more than 127,000 breast cancer patients. *Breast Cancer Res Treat* **28**: 9–20.

Zavadil J, Cermak L, Soto-Nieves N, Bottinger EP. (2004). Integration of TGF-beta/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition. *EMBO J* **23**: 1155–1165.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc).