A high-resolution integrated analysis of genetic and expression profiles of breast cancer cell lines

Alan Mackay, Narinder Tamber, Kerry Fenwick, Marjan Iravani, Anita Grigoriadis, Tim Dexter, Christopher J. Lord, Jorge S. Reis-Filho, Alan Ashworth

Breast Cancer Res Treat (2009) 118:481–498
DOI 10.1007/s10549-008-0296-7

PRECLINICAL STUDY

A high-resolution integrated analysis of genetic and expression profiles of breast cancer cell lines

Alan Mackay · Narinder Tamber · Kerry Fenwick · Marjan Iravani · Anita Grigoriadis · Tim Dexter · Christopher J. Lord · Jorge S. Reis-Filho · Alan Ashworth

Received: 23 December 2008 / Accepted: 23 December 2008 / Published online: 24 January 2009

Abstract Tumour cell lines derived from breast cancer patients constitute one of the cornerstones of breast cancer research. To characterise breast cancer cell lines at the genetic level, we have developed a full tiling path bacterial artificial chromosome (BAC) array collection for comparative genomic hybridisation (aCGH). This aCGH BAC collection covers 98% of the entire human genome at a resolution of 40–60 kbp. We have used this platform alongside an in-house produced 17 K cDNA microarray set to characterise the genetic and transcriptomic profiles of 24 breast cancer cell lines, as well as cell types derived from non-diseased breast. We demonstrate that breast cancer cell lines have genomic and transcriptomic features that recapitulate those of primary breast cancers and can be reliably subclassified into basal-like and luminal subgroups. By overlaying aCGH and transcriptomic data, we have identified 753 genes whose expression correlate with copy number; this list comprised numerous oncogenes recurrently amplified and overexpressed in breast cancer (e.g., *HER2*, *MYC*, *CCND1* and *AURKA*). Finally, we demonstrate that although breast cancer cell lines have genomic features usually found in grade III breast cancers (i.e., gains of 1q, 8q and 20q), basal-like and luminal cell lines are characterised by distinct genomic aberrations.

Keywords Breast cancer · Cell lines · aCGH

Electronic supplementary material The online version of this article (doi:10.1007/s10549-008-0296-7) contains supplementary material, which is available to authorized users.

A. Mackay · N. Tamber · K. Fenwick · M. Iravani · A. Grigoriadis · T. Dexter · C. J. Lord (✉) · J. S. Reis-Filho · A. Ashworth
The Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK
e-mail: lordc@icr.ac.uk

A. Ashworth
e-mail: Alan.Ashworth@icr.ac.uk

Present Address:
A. Grigoriadis
The Breakthrough Breast Cancer Research Unit, King's College London School of Medicine, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK

Introduction

Microarray-based RNA expression profiling has been extensively used as a tool to help unravel the complexity of breast cancer [1–6]. These expression profiling studies have contributed to a working model of breast tumour taxonomy that suggests the existence of at least four distinct subgroups that have both biological and clinical significance. These include: (1) two groups of luminal tumours (luminal A and luminal non-A), which are characterised by the expression of ERα and ERα-regulated genes, (2) normal-like cancers, which are poorly characterised but have expression profiles similar to those of normal breast and fibroadenomas, (3) HER2 tumours, which display overexpression of HER2 and of other genes mapping to the 17q12–q21 region of the genome, and (4) basal-like cancers, which in general lack ERα and HER2 expression but express a significant number of genes usually associated with normal breast basal/myoepithelial cells [3–5, 7, 8].

Tumour cell lines derived from breast cancer patients constitute one of the cornerstones of breast cancer research [9]. Given our improving understanding of the heterogeneity of breast cancer, it is thus self evident that a thorough molecular characterisation of genetic and transcriptomic

profiles of breast cancer cell lines is required to best identify which cell lines can be effectively used as models to study specific subgroups of breast cancer. Recent profiling efforts have in fact demonstrated that the genomic and transcriptomic profiles of commonly used cell lines do, to some extent, recapitulate those of primary breast cancers [4]. Moreover, molecular profiling can also inform the selection of appropriate cell lines to use in order to experimentally model particular breast cancer subsets [4]. Here, we extend these profiling studies and integrate transcriptomic profiling with high-resolution genomic profiling, performed using a tiling-path (TP) array comparative genomic hybridisation (CGH) platform.

Materials and methods

Cell lines

The following cell lines were used in this study: (1) 24 breast cancer cell lines—BT-474, CAL51, DU4475, GI101, HBL100, HMT3552, Hs578T, MCF7, MDA-MB-134, MDA-MB-157, MDA-MB-175, MDA-MB-231, MDA-MB-361, MDA-MB-453, MDA-MB-468, MDA-MB-469, PMC42, SKBR3, SKBR5, SKBR7, T47D, ZR75.1, ZR75.30 and UACC3199, (2) HB4a, a cell line derived from immortalisation of normal breast luminal epithelial cells [10], (3) two non-tumourigeneic breast epithelial cell lines (MCF10A and MCF12A), (4) human mammary epithelial cells (HMECs), (5) pools of immunomagnetically sorted normal primary breast cells representing normal luminal cells [11], myoepithelial cells [11] intra-lobular fibroblasts [12] and endothelial cells [12].

Cell culture, DNA and RNA extraction

Cell lines were obtained from ATCC and cultured under standard conditions [13] using DMEM: F12 supplemented with 10% (v/v) foetal calf serum, 100 U/ml penicillin, 100 U/ml streptomycin and 5 mM glutamine (Gibco), unless otherwise stated. HMECs were maintained in MEBM serum-free medium (Clonetics), supplemented with EGF, hydrocortisone, insulin, and bovine pituitary extract using Singlequot reagent packs from Clonetics. MCF10A and MCF12A were maintained in DMEM/Ham's F-12 medium supplemented with 5% (v/v) horse serum, 100 U/ml penicillin, 100 U/ml streptomycin, 10 μg/ml insulin, 20 ng/ml epidermal growth factor, 0.5 μg/ml hydrocortisone, and 100 ng/ml cholera toxin. HB4a cells were maintained in RPMI-1640 medium (Gibco) supplemented with 10% (v/v) foetal calf serum, 5 μg/ml hydrocortisone, 5 μg/ml insulin, 100 ng/ml cholera toxin, 20 ng/ml recombinant human epidermal growth factor, 100 U/ml penicillin, 100 U/ml streptomycin and 5 mM glutamine (Gibco). Primary endothelial cells were maintained in EBM-2 (SingleQuot Media BulletKit, Cambrex) supplemented with 2% (v/v) foetal calf serum, 100 U/ml penicillin, 100 U/ml streptomycin and 5 mM glutamine (Gibco).

For DNA extraction, cells were washed with PBS, trypsinised and cell pellets prepared by centrifugation. DNA was extracted from cell pellets using the DNeasy tissue kit (Qiagen) according to the manufacturer's instructions. RNA was prepared from sub-confluent cell monolayers by direct lysis in Trizol (Invitrogen) followed by RNA extraction according to the manufacturer's instructions. In all cases media was changed 24 h prior to DNA or RNA extraction.

Microarray platforms

The 32 K human bacterial artificial chromosome (BAC) re-array collection covers 98% of the entire human genome at a resolution of 40–60 kb. To produce full tiling (TP) path arrays for microarray-based comparative genomic hybridisation (aCGH) analysis, the 32 K BAC re-array collection (CHORI) was obtained as DNA and amplified using Genomiphi v2 (GE Healthcare), according to the manufacturer's instructions [14]. Amplification was confirmed by electrophoresing an aliquot of amplified material on E-gel-96 gels (Invitrogen). Amplified material was resuspended in DMSO-based microarray spotting buffer before spotting onto APS coated cDNA slides (Corning UltraGaps), using a QArray2 microarray spotter (Genetix).

The 17 K cDNA array was produced by PCR amplification of a sequence validated IMAGE clone collection kindly donated by the CRUK microarray facility (ICR Sutton, UK). The PCR products for the expression array were purified and spotted as described for BAC products., but using a Lucidea microarray spotter (GE Healthcare).

The production of each array and the annotation of the two collections are both recorded in array definition files as submitted to ArrayExpress (E-TABM-593).

Microarray-based comparative genomic hybridisation (aCGH) and cDNA microarray hybridisations

For array CGH, 1 μg DNA from each cell line was labelled with either Cy3 or Cy5 using a modified BioPrime labelling reaction (InVitrogen) [15, 16]. Labelling, hybridisations and washes were performed as previously described [17, 18].

For expression microarrays, 1 μg of total RNA was T7 amplified by in vitro transcription using the amino allyl message Amp kit (Ambion), according to the manufacturer's instructions amplified RNA was then coupled to either Cy3 or Cy5 and hybridised against a reference RNA

in separate dye-swap hybridisations [19]. Reference RNA comprised 33 RNA samples representing the majority of the cell lines used in the study and cell lines derived from purified non-tumour cell types as previously described [19]. Arrays were hybridised in a hybridisation buffer (GE Healthcare) at 42°C overnight and successively washed at 65°C in 2× SSC, 0.1% SDS and 0.1× SSC, 0.1% SDS, prior to scanning.

Both aCGH and cDNA arrays were scanned on an Axon 4000B microarray scanner and TIFF images were analysed using GenePix 5.1 software.

Array data analysis—expression profiling

All primary data were pre-processed and analysed using an in-house R script (BACE.R) in R version 2.8.0 (package available on request). Data were loaded from gpr files without background subtraction and corrected for spatial and dye bias by normalising using printTiploess based upon loessFit for each block on each array using the *limma* package. Probes were annotated based upon Unigene build 206 and the hg18/NCBI36 build of the human genome. Probes with levels of expression in the lowest 20% of all genes in more than 80% of the samples were removed. Probes with more than 30% missing values across the experiment were removed and the remaining missing values were imputed using *k* nearest neighbour. For replicated probes on the 17 K array, the probe with the greatest median absolute deviation across the study was selected. This left a final expression dataset of 8,121 genes for further analysis. Normalised expression data are available in Supplementary Table S1.

Assignment of molecular subgroups

The Hu et al. [3] dataset was used as a training set to predict basal and luminal phenotypes using the prediction analysis of microarrays package (PAM), as previously described [3, 20]. Briefly, 24,000 probes from Hu et al. [3] were filtered to remove probes flagged as absent in more than 10% of the samples. Probes with the greatest variation across samples were then selected (those with an inter-quartile range of log ratio values of more than 0.65). The average value for each gene was calculated from probes annotated with the same Unigene accession number. This left 4,005 genes for further analysis. The Hu et al. [3] dataset was combined with the 17 K expression dataset on the basis of Unigene accession number. Expression data were then centred and missing values were imputed using a *k* nearest neighbour algorithm. This left 1,645 genes for PAM prediction analysis.

A PAM predictor of genes intersecting with cell line expression profiles was constructed which correctly classified 100% of luminal and basal tumours in Hu et al. [3]. This predictor was trained on tumour data and cross-validated before being applied to the cell lines and primary cultures in this study.

Genes specific to luminal and basal-like cell lines were identified using the statistical analysis of microarrays (SAM) package. Significant genes were identified with a local false discovery rate of less than 5%.

Array data analysis—aCGH

aCGH data were loaded from gpr files without background subtraction and normalised to remove spatial and dye bias within each block on the array using loessFit (printTipLoess) from the limma package. aCGH data were mapped to the genome using build hg18 NCBI36. As BAC clones in the tiling path platform overlap by approximately 30%, extreme outliers in aCGH could be removed, i.e., BACs whose log2 ratios differed from their immediate genomic neighbours by more than the twice median absolute deviation of all BACs across the genome. Dye swap pairs were combined and BAC clones whose values were missing in >40% of samples were removed from subsequent analysis. Missing values were imputed from the median of BACs with neighbouring genomic positions.

aCGH data were smoothed across each chromosome using adaptive weight smoothing (aws) with a maximal bandwidth set to the length of the chromosome in each case. Smoothed, normalised aCGH data are available in Supplementary Table S2.

Thresholds for gain and loss in aCGH were estimated as previously described [17, 18, 21, 22]. Smoothed log2 ratio values less than −0.08 were categorised as losses (i.e., heterozygous deletions), those greater than 0.08 as gains, and those in between as unchanged. Amplifications were defined as smoothed log2 ratio values greater than 0.4. Thresholds for gains and losses were established as the measurement of three times the standard deviation of all BACs across the autosomes in male:female hybridisations. These figures were further validated by comparison with interphase fluorescence in situ hybridisation (FISH) data generated using probes for genes at different chromosomal locations [23]. aCGH states were based only on aws values of three or more contiguous BACs which were consistently above and below relevant thresholds. For analyses based on categorical data, aws-smoothed aCGH ratios were assigned as five categorical states −2 (suspected deletion), −1 (loss), 0 (no change), 1 (gain) and 2 (amplification) across the genome.

Data processing and analysis were carried out in R 2.8.0 (http://www.r-project.org/) and BioConductor 2.3 (http://www.bioconductor.org/), making extensive use of modified versions of the packages *aCGH*, *marray*, and *aws* in particular. Thresholded data for each clone were also used for

categorical analysis, using a Fisher's exact test adjusted for multiple testing using the stepdown permutation procedure maxT, thus providing strong control of the family-wise type I error rate (FWER) [17, 18, 21, 22].

Classification of cell lines according to genomic architecture

Genome architecture patterns were essentially determined as described by Hicks et al. [24]. Cases were considered of 'simplex' pattern if their genomic profiles were characterised by broad segments of duplication and deletion, usually comprising entire chromosomes or chromosomal arms. Complex patterns included 'sawtooth' and 'firestorm' [24]. Cases with a 'sawtooth' profile were characterised by many narrow segments of duplication and deletion, often alternating and affecting most if not all chromosomes. Although most of the genome displayed low level gains or losses, amplifications are rarely found in cases with this pattern. Cases were considered of 'firestorm' pattern if they resembled the 'simplex' type except that the profiles contained at least one localised region of clustered, relatively narrow peaks of amplification, with each cluster confined to a single chromosomal arm or chromosome [24].

To reduce the subjectivity in determining whether a case pertained to the 'simplex' or 'complex' group, we employed the "Firestorm Index" essentially as described by Hicks et al. [24], given that the resolution of our platform is comparable to that of the ROMA assay described in [24] (i.e., 50 vs. 35 kb, respectively). Discontinuities above and below thresholds of 0.08 and 0.4 were used to assign breakpoints in the assessment of an F score [24]. The F score is represented by the sum over all breakpoints of the Z score for each breakpoint. Z score is calculated as 2 divided by the sum of the distance to the next breakpoint (or the end of the chromosome) to the left and the distance to the next breakpoint (or the end of the chromosome) to the right. We have also tested the 0.1 threshold as described by Hicks et al. [24]. Results of the "Firestorm Index" are reported in Table 1. Cases were considered of 'complex' pattern if $F > 1$. As stressed by Hicks et al., using this algorithm, both 'sawtooth' and 'firestorm' patterns achieve high F-values. After determining whether a case displayed a simplex or complex pattern, differentiation between 'sawtooth' and 'firestorm' was performed by visual inspection of the genome plots by three of the authors, independently. In all cases, a perfect agreement between the observers was achieved.

Identification of genes whose expression correlates with copy number changes

To identify genes whose expression correlated with genetic copy number changes, aws-smoothed aCGH data were used to assign the median aCGH states for each of the 8,121 genes in the gene expression dataset using the median values for all BACs which overlap with the genomic positions of each gene. This resulted in a 1:1 matrix of expression data and aCGH values used in correlations. Pearson correlations and Spearman correlations were performed between cDNA

Table 1 Phenotypes of the 24 breast cancer cell lines, non-tumourigenic breast epithelial cell lines, immortalised normal luminal epithelial cells and immunomagnetically purified primary cells used in this study

Cell line	Phenotype	Firestorm Index	ER	HER2	EGFR
Breast cancer cell line phenotypes					
BT474	Luminal	Firestorm	+	AMP	N
CAL51	Basal	NC	−	N	N
DU4475	Luminal	Sawtooth	−	N	N
GI101	Basal	Sawtooth	+	N	N
HB4a	Basal	Firestorm	−	N	N
HBL100	Basal	Sawtooth	−	N	N
HMEC	Luminal	NC	−	N	N
HMT3552	Basal	Firestorm	−	N	N
Hs578T	Luminal	Firestorm	−	N	N
MCF10A	Basal	Simplex	−	N	N
MCF12A	Basal	Simplex	−	N	N
MCF7	Luminal	Firestorm	+	N	N
MDA-MB-134	Luminal	Firestorm	+	N	N
MDA-MB-157	Basal	Sawtooth	−	N	N
MDA-MB-175	Luminal	Firestorm	+	N	N
MDA-MB-231	Basal	Sawtooth	−	N	N
MDA-MB-361	Luminal	Firestorm	+	AMP	N
MDA-MB-453	Luminal	Firestorm	−	AMP	N
MDA-MB-468	Basal	Firestorm	−	N	AMP
MDA-MB-469	Luminal	Sawtooth	−	N	N
PMC42	Basal	Sawtooth	−	N	N
SKBR3	Luminal	Firestorm	−	AMP	N
SKBR5	Luminal	Firestorm	−	N	AMP
SKBR7	Basal	Firestorm	−	N	N
T47D	Luminal	Sawtooth	+	N	N
UACC3199	Basal	Sawtooth	−	N	N
ZR75.1	Luminal	Firestorm	+	N	N
ZR75.30	Luminal	Firestorm	+	AMP	N
Normal cells					
ENDO	Basal	ND	−	N	N
FIB	Basal	ND	−	N	N
LUM	Luminal	ND	−	N	N
MYO	Basal	ND	−	N	N

Phenotypes were ascribed by a PAM predictor based upon 1,645 genes intersecting between this study and that of Hu et al. [3] (aCGH Firestorm index was calculated as described in the "Material and methods" based upon the algorithm of Hicks et al. [24]. ER status was ascribed based upon published data for each line and the median microarray-based expression of ESR1. *HER2* and *EGFR* are based upon FISH-validated amplification (data not shown). NC, not calculated; ND, not done

expression log2 ratios and median aws-smoothed ratios derived from aCGH analysis for each gene. *P* values for each test were adjusted with Benjamini and Hochberg multiple *P*-value adjustment [25].

Regions of recurrent amplification were identified as contiguous regions of three or more clones that carried amplification in at least two cell lines. To test for genes which were overexpressed when amplified in these regions Wilcoxon rank sum tests were performed on gene expression measurements within genomic regions harbouring amplifications in two or more cell lines using categorical aCGH states as the grouping variable. For each gene, the gene expression values were compared in those cell lines that showed a median aws value above each threshold (e.g., amplified) and those with aws values below each threshold (e.g., un-amplified). *P*-values for the Wilcoxon rank-sum test were adjusted within each region of copy number change using Benjamini and Hochberg multiple *P*-value adjustment [25].

For regions of recurrent amplification, matched heatmaps were created by retrieving gene expression values and corresponding median-overlay aCGH states for each gene. Genes were ordered according to chromosomal location and cases were separated into those that harbour amplifications in the region and those which do not. Within these groups the samples were ordered based upon the sum of expression values within the region.

Characterisation of copy number changes associated with specific molecular subgroups

After assigning each cell line to a specific molecular subgroup using the PAM predictor described above, the frequency of copy number gains, losses and amplifications in cell lines pertaining to each molecular subgroup was defined based on the categorical aCGH data. To determine the copy number aberrations significantly associated with each molecular subgroup, a multi-Fisher's exact test using an adjustment for multiple-testing using the step-down permutation procedure maxT, providing strong control of the family-wise type I error rate (FWER) [17, 18].

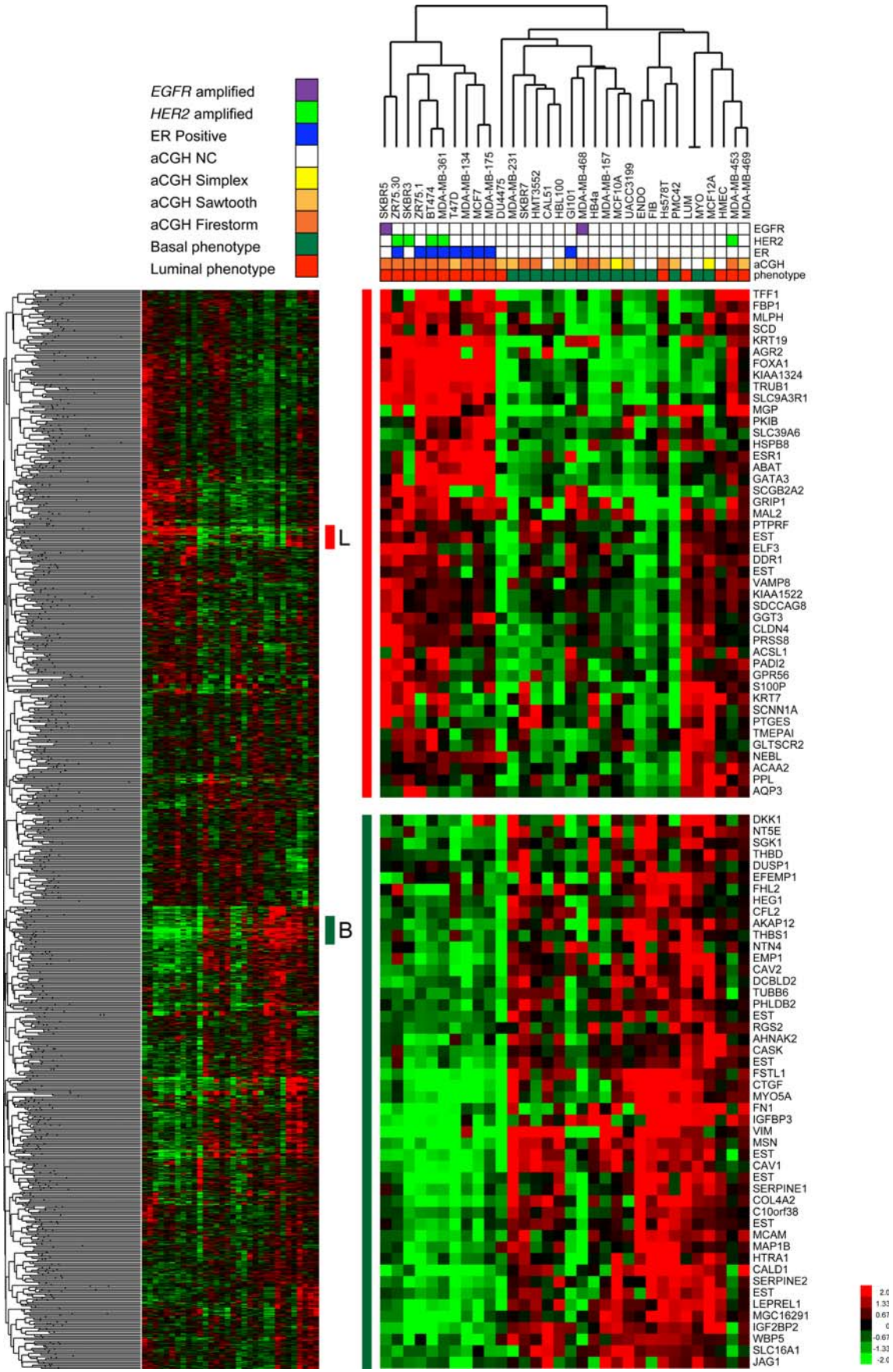
Results

Unsupervised clustering analysis of gene expression

We generated transcriptomic profiles of 24 breast cancer cell lines and also cell types found in the non-diseased breast using a cDNA array. We sought to determine whether breast cancer cell lines would recapitulate at the transcriptional level the molecular subgroups of breast cancer identified in previous class discovery studies. Unsupervised hierarchical clustering analysis using the expression values of 8,121 genes revealed two main groups (Fig. 1). Group 1 was characterised by cell lines with epithelioid morphology that express genes associated with luminal epithelial cells of normal breast. Interestingly, this group contained four cell lines that have been shown to harbour *HER2* gene amplification and ERBB2 over-expression (BT-474, MDA-MB-361, ZR-75.30 and SKBR3). In fact, group 1 is remarkably similar to that described by Neve et al. [4] as a "luminal" group. Group 2 was comprised of cell lines with variable morphology, which expressed genes usually found in basal/myoepithelial cells of the normal breast and primary basal-like breast cancers [5, 6, 26]. It should be noted that group 2 comprised two separate subgroups, the first encompassing normal endothelial cells, fibroblasts, luminal epithelial cells, myoepithelial cells, HMECs and MCF12A, the *HER2* amplified cell line MDA-MB-453 and the carcinosarcoma/metaplastic breast cancer cell line Hs578T. The second subgroup within group 2 comprised a diverse group of cell lines. All but three cell lines within this group (i.e., GI101, MDA-MB-468 and HB4a) were characterised by high-level expression of genes expressed at high levels in primary basal-like breast cancers (see Table 2). The consistency of these clusters was further corroborated when a class predictor for luminal and basal-like profiles derived from the Hu et al. [3] dataset was applied to our profiles. The PAM predictor was able to correctly identify luminal and basal-like phenotypes in the populations of pooled normal luminal epithelial and myoepithelial cells used. Normal human mammary fibroblasts and endothelial cells were classified as of basal-like phenotype. The results of this predictor were in agreement with those of previously published analyses of luminal and basal-like phenotypes [1, 4, 27] for all cell lines studied with the exception of Hs578T. Predicted cell line phenotypes are shown in Table 1 and in Supplementary Table S3.

Luminal/basal gene expression

To determine the genes significantly associated with luminal and basal-like subgroups of breast cancer cell lines identified with our PAM predictor, we performed significance analysis of microarrays (SAM), which identified 242 luminal-specific genes and 61 basal-specific genes with a local false discovery rate less than 5%. The top 20 luminal and basal-specific genes are shown in Table 2. The expression profile of luminal cells is dominated by *ESR1* (ERα) expression and that of a large number of genes that are associated with oestrogen response in breast cancer including *FOXA1*, *TFF1*, *AGR2* and *GATA3* [19]. Basal specific genes in breast cancer cell lines included two isoforms of caveolin (*CAV1* and *CAV2*), as well as *VIM*, *LY6K* and *PRNP* as previously described [28–30]. These genes correlate well with our previous reports of oestrogen


EGFR amplified
HER2 amplified
ER Positive
aCGH NC
aCGH Simplex
aCGH Sawtooth
aCGH Firestorm
Basal phenotype
Luminal phenotype
SKBR5
ZR75.30
SKBR3
ZR75.1
BT474
MDA-MB-361
T47D
MDA-MB-134
MCF7
MDA-MB-175
DU4475
MDA-MB-231
SKBR7
HMT3552
CAL51
HBL100
GI101
MDA-MB-468
HB4a
MDA-MB-157
MCF10A
UACC3199
ENDO
FIB
Hs578T
PMC42
LUM
MYO
MCF12A
HMEC
MDA-MB-453
MDA-MB-469
EGFR
HER2
ER
aCGH
phenotype
L
B
TFF1
FBP1
MLPH
SCD
KRT19
AGR2
FOXA1
KIAA1324
TRUB1
SLC9A3R1
MGP
PKIB
SLC39A6
HSPB8
ESR1
ABAT
GATA3
SCGB2A2
GRIP1
MAL2
PTPRF
EST
ELF3
DDR1
EST
VAMP8
KIAA1522
SDCCAG8
GGT3
CLDN4
PRSS8
ACSL1
PADI2
GPR56
S100P
KRT7
SCNN1A
PTGES
TMEPAI
GLTSCR2
NEBL
ACAA2
PPL
AQP3
DKK1
NT5E
SGK1
THBD
DUSP1
EFEMP1
FHL2
HEG1
CFL2
AKAP12
THBS1
NTN4
EMP1
CAV2
DCBLD2
TUBB6
PHLDB2
EST
RGS2
AHNAK2
CASK
EST
FSTL1
CTGF
MYO5A
FN1
IGFBP3
VIM
MSN
EST
CAV1
EST
SERPINE1
COL4A2
C10orf38
EST
MCAM
MAP1B
HTRA1
CALD1
SERPINE2
EST
LEPREL1
MGC16291
IGF2BP2
WBP5
SLC16A1
JAG1

◀**Fig. 1** Heatmap of unsupervised clustering using expression data of breast cancer cell lines and normal breast cell types. A heatmap of the unsupervised clustering of the 31 samples using 2,337 of the most variable genes is shown. Log2 expression values were clustered with a Wards algorithm based upon Euclidean distance. The entire heatmap is shown in miniature on the left. A luminal cluster containing the genes *ESR1* and *TFF1* is shown in the *panel marked L*. A basal cluster containing the genes *VIM* and *CAV1* is shown in the *panel marked B*. Phenotypes of each sample are shown below the dendrogram in the *top panel*. Basal and Luminal phenotype shown is based upon PAM prediction. aCGH genomic profiles shown are NC (no change), Simplex, Sawtooth, Firestorm. ER status is shown as positive/negative. *HER2* and *EGFR* status are shown as normal and amplified

responsive genes in primary breast cancers [19] and with genes associated with the basal-like phenotype of breast epithelial cells [26] and primary basal-like breast cancers [3]. The complete list of significant genes differentially expressed in luminal and basal-like cells are shown in Supplementary Tables S4A and S4B.

Genomic profiling of breast cancer cell lines

We have recently demonstrated that the genomic profiles as defined by Hicks et al. [24] are associated with the molecular subgroups of breast cancer: whilst basal-like breast cancers preferentially display 'sawtooth' profiles, *HER2* amplified breast cancers are preferentially of 'firestorm' pattern [31]. By applying Hicks et al. [24] algorithm to breast cancer cell lines, we observed that in a way akin to human HER2 positive breast cancers [18], *HER2* amplified cell lines consistently displayed a firestorm pattern. Basal-like cell lines and luminal cell lines were either of sawtooth or firestorm profiles. As expected, HMECs and MCF10A cells displayed a simplex profile. Interestingly, CAL51, which harbours DNA mismatch repair (MMR) deficiency and is known to have microsatellite instability [32], displayed minimal changes in the genome, corroborating the hypothesis that cancers/cell lines with MMR deficiency usually have simple karyotypes [33]. No correlation between ER expression and genomic architecture was observed in the cell lines analysed. Hierarchical clustering analysis of cell lines based upon aws-smoothed aCGH values across the genome identified three main clusters, which did not correlate with HER2 status, ER status, molecular subgroup or genetic pattern (Fig. 2).

Identification of genes whose expression correlates with copy number

To determine the contribution of copy number aberrations to the transcriptomic characteristics of breast cancer cell lines, we sought to identify the genes whose expression significantly correlates with copy number. By correlating the aws-smoothed log2 aCGH ratios (copy number states) with expression levels using Pearson's correlation, we identified 753 genes whose expression significantly positively correlates with copy number changes (adjusted P-value <0.05, Table 3; Supplementary Table S5). This gene list generated by this analysis was enriched for genes mapping to chromosomes 17 (10.36%), 8 (9.69%), 11 (9.56%), 1 (9.03%), 3 (7.57%) and 6 (7.04%), and, as expected, one of the genes that displayed the strongest correlation was *ERBB2* (*HER2*). Previous analysis of primary breast cancers and breast cancer cell lines has demonstrated that *HER2* mRNA levels are strongly correlated with copy number and that protein overexpression is underpinned by gene amplification in >90% of cases [2, 4, 16, 34]. This analysis also identified genes that have been shown to be either amplicon drivers or are required for the survival of cells with copy number gains/amplification of their respective loci such as *CCND1*, *MYC*, *FGFR1*, *AURKA* [2, 4, 34, 35]. Furthermore, genes whose expression levels correlate with copy number and are amplified in breast cancer cell lines could be identified in a systematic fashion. For example, *CCNG2* was shown to be amplified in BT-474 cells and was shown to have expression levels determined by gene copy number. *CDK4* mRNA levels were also correlated with copy number in ZR-75.1 cells and *TSFM* is co-amplified with *CDK4* in the same amplicon in ZR-75.1 and displays expression levels that correlate with copy number. *PRKAR1A*, a protein kinase, was shown to be amplified in six cell lines (i.e., MCF7, MDA-MB-157, MDA-MB-361, MDA-MB-453, SKBR3, ZR75.1) and its expression levels correlated with gene copy number. The *STK3* gene was amplified in six cell lines (Hs578T, MCF12A, MDA-MB-134, MDA-MB-157, MDA-MB-175, UACC3199), and its expression levels correlated with gene copy number.

In the present study, six known genes and two ESTs displayed a significant inverse correlation between expression and copy number. Out of these transcripts, ESTs, for NAD(P) dependent steroid dehydrogenase-like (*NSDHL*), eukaryotic translation initiation factor 4E family member 3 (*EIF4E3*) and Regulatory factor X 1 (*RFX1*) showed more frequent deletions than gains and/or amplification, suggesting that these genes were overexpressed when one copy was deleted/dysfunctional, a feature often observed in known tumour suppressor genes mutated in cancer (e.g., *TP53* whose expression levels are increased in the presence of inactivating gene mutations [36, 37], or p16 and p14 overexpression in the presence of intragenic deletions or single point mutations [38]).

Identification of likeliest amplicon drivers of recurrent amplicons in breast cancer

Recent studies have demonstrated that amplicons often have more than one driver [16, 31, 39–41]. Previous studies

Table 2 Results of SAM analysis to identify genes with significantly differing expression between luminal and basal breast cancer cell lines

Gene ID	Description	Fold	*Q* value	lFDR	Unigene
Genes expressed at significantly higher levels in Luminal-like cell lines					
FOXA1	Forkhead box A1	6.89	0.00	0.06	Hs.163484
FBP1	Fructose-1,6-bisphosphatase 1	6.80	0.00	0.05	Hs.494496
TRUB1	TruB pseudouridine (psi) synthase homolog 1	6.78	0.00	0.06	Hs.21187
TFF1	Trefoil factor 1	6.72	0.00	0.00	Hs.162807
AGR2	Anterior gradient homolog 2	6.52	0.00	0.00	Hs.530009
KRT19	Keratin 19	6.34	0.00	0.00	Hs.654568
KIAA1324	KIAA1324	5.89	0.00	0.06	Hs.708190
SLC9A3R1	Solute carrier family 9, member 3 regulator 1	5.64	0.00	0.06	Hs.699203
GRIP1	Glutamate receptor interacting protein 1	4.25	0.70	2.10	Hs.505946
MLPH	Melanophilin	4.24	0.00	0.00	Hs.102406
GATA3	GATA binding protein 3	4.12	0.00	0.00	Hs.524134
MAL2	Mal, T-cell differentiation protein 2	3.80	0.70	2.74	Hs.201083
ESR1	Estrogen receptor 1	3.72	0.00	0.00	Hs.208124
KCNIP4	Kv channel interacting protein 4	3.65	0.00	0.00	Hs.655705
FAM113A	Family with sequence similarity 113A	3.57	0.00	0.00	Hs.29341
PIP	Prolactin-induced protein	3.57	0.00	0.00	Hs.99949
DNAJA4	DnaJ (Hsp40) homolog A4	3.34	0.00	0.00	Hs.513053
ABAT	4-Aminobutyrate aminotransferase	3.32	0.00	0.00	Hs.336768
AZGP1	Alpha-2-glycoprotein 1, zinc-binding	3.26	0.70	1.92	Hs.546239
TSPAN13	Tetraspanin 13	3.24	0.00	0.00	Hs.364544
GPR160	G protein-coupled receptor 160	3.23	0.00	0.03	Hs.231320
PKIB	Protein kinase (cAMP-dependent) inhibitor beta	3.13	0.00	0.00	Hs.486354
PADI2	Peptidyl arginine deiminase, type II	3.04	0.00	0.00	Hs.33455
EST	Transcribed locus	2.95	0.00	0.00	Hs.656379
SYT7	Synaptotagmin VII	2.95	0.00	0.00	Hs.502730
C16orf72	Chromosome 16 open reading frame 72	2.87	0.00	0.00	Hs.221497
CLU	Clusterin	2.81	0.00	0.00	Hs.436657
C9orf152	Chromosome 9 open reading frame 152	2.81	0.00	0.00	Hs.125608
BCAS1	Breast carcinoma amplified sequence 1	2.78	0.00	0.00	Hs.400556
LFNG	LFNG *O*-fucosylpeptide 3-beta-*N*-acetylglucosaminyltransferase	2.74	0.00	0.00	Hs.159142
Genes expressed at significantly higher levels in basal-like cell lines					
VIM	Vimentin	0.10	0.00	0.00	Hs.642813
EST	Transcribed locus	0.16	0.00	0.15	Hs.657308
CAV1	Caveolin 1	0.17	0.00	0.44	Hs.74034
MSN	Moesin	0.18	0.00	0.23	Hs.87752
FSTL1	Follistatin-like 1	0.19	1.24	0.89	Hs.269512
CXCL1	Chemokine (C–X–C motif) ligand 1	0.25	1.24	1.31	Hs.789
IGF2BP2	Insulin-like growth factor 2 binding protein 2	0.25	0.00	0.00	Hs.35354
SERPINE2	Serpin peptidase inhibitor, clade E 2	0.30	1.78	2.64	Hs.38449
CAV2	Caveolin 2	0.35	0.70	0.01	Hs.212332
SLC16A1	Solute carrier family 16, member 1	0.35	0.00	0.00	Hs.75231
SENP8	SUMO/sentrin specific peptidase family 8	0.37	0.00	0.00	Hs.513002
WBP5	WW domain binding protein 5	0.37	1.78	3.14	Hs.533287
LY6K	Lymphocyte antigen 6 complex, locus K	0.37	1.78	2.53	Hs.69517
TPM2	Tropomyosin 2 (beta)	0.38	2.68	4.84	Hs.300772
COTL1	Coactosin-like 1	0.40	1.78	2.36	Hs.289092
CRIM1	Cysteine rich transmembrane BMP regulator 1	0.41	1.24	1.45	Hs.699247

Table 2 continued

Gene ID	Description	Fold	*Q* value	lFDR	Unigene
PRNP	Prion protein (p27–30)	0.41	0.70	0.58	Hs.472010
PTRF	Polymerase I and transcript release factor	0.41	0.70	0.02	Hs.437191
BTG3	BTG family, member 3	0.41	0.70	0.29	Hs.473420
SRPX	Sushi-repeat-containing protein, X-linked	0.42	2.68	4.26	Hs.15154
RAI14	Retinoic acid induced 14	0.43	0.00	0.00	Hs.431400
CTNNAL1	Catenin alpha-like 1	0.46	0.70	0.72	Hs.58488
ZCCHC3	Zinc finger, CCHC domain containing 3	0.46	0.00	0.00	Hs.28608
CASP1	Caspase 1	0.46	1.78	3.19	Hs.2490
ERRFI1	ERBB receptor feedback inhibitor 1	0.48	1.78	2.81	Hs.605445
EMP3	Epithelial membrane protein 3	0.49	0.70	0.09	Hs.9999
CXADR	Coxsackie virus and adenovirus receptor	0.49	0.70	0.38	Hs.705503
CTSC	Cathepsin C	0.50	2.68	4.39	Hs.128065
LOC728449	LOC728449 mRNA	0.50	2.68	3.95	Hs.463110
CKAP2	Cytoskeleton associated protein 2	0.50	1.24	1.09	Hs.444028

Significant genes were identified as those with a local false discovery rate (lFDR) of less than 5%. Shown here are the 30 genes with greatest fold change increase in expression in either luminal (Table 2) or basal (Table 2) cell lines. The complete list of 242 significant luminal genes and 61 significant basal genes are available in Supplementary Tables S4A and S4B

have demonstrated that expression of genes consistently overexpressed when amplified may be required for the survival of cells harbouring amplification of their genetic loci [31, 39, 42, 43]. To identify the likeliest drivers of the amplicons 8p11.2–p12, 8q24, 11q13.3, 17q12–q21 and 20q13.3, we first retrieved the genes mapping to these amplicons, performed a Pearson's correlation to determine those whose expression correlate with copy number and then identified all genes that were overexpressed when amplified within each region using a Wilcoxon sign rank test (Fig. 3). Based on this rationale, we demonstrate that (1) on 8p11.2–p12, the likely amplicon drivers are *ZNF703*, *WHSC1L1*, *FGFR1*, *TM2D2* and *MYST3*; (2) on 8q24, apart from *MYC*, *C8orf76*, *C8orf32*, *TRMT12*, *RNF139*, *NDUFB9*, *SQLE*, *KIAA0196*, *TRIB1*, *NDRG1*, *ZFAT*, PTP4A3, *BAI1*, *LY6E*, *KIAA1833* and *ADCK5* are potential additional amplicon drivers; (3) on 11q13.2–q14.3, amplicon drivers include *SAPS3*, *FADD*, *PPFIA1*, *CTTN*, *POLD3*, *RNF169*, *PKRIR*, *RSF1*, *C11orf67*, *NDUFC2*, *CREBZF* and *CTSC*; (4) on 17q12–q21, apart from *HER2*, *MLLT6*, *CISD3*, *PSMB3*, *LASP1*, *LOC90110*, *FBXL20*, *ACACA*, *CASC3*, *MED1* and *MED24* may also be considered putative additional drivers of the amplicon. Finally, on 20q13.3–20q13.33, *STAU1*, *PTGIS*, *B4GALT5*, *ZNF313*, *UBE2V1*, *DPM1*, *PFDN4*, *AURKA*, *RAE1*, *RAB22A*, *GNAS*, *TH1L* and *TAF4* would fulfil the minimum requirement to be tested as possible drivers. Most importantly, this approach circumvents limitations of previous analysis, as it takes into account (1) the possible existence of multiple cores in a single amplicon and (2) the hypothesis that each amplicon contains more than one amplicon driver. It should be noted, however, that *STARD3* and *GRB7*, two of the most commonly overexpressed genes in the *HER2* amplicon (17q12–q21) in previous studies [4, 44], failed to display a significant correlation between amplification and overexpression, owing to the suboptimal performance of the respective cDNA probes. Similar analyses were performed for other recurrent amplicons in these cell lines and genes that may be considered as potential amplicon drivers for each amplicon are described in Supplementary Table S6.

Identification of copy number changes specific to luminal and basal-like cell lines

Previous studies have demonstrated that basal-like and luminal primary breast cancers not only have distinct genomic profiles, but also harbour specific genetic aberrations. A multi-Fisher's comparison of aCGH revealed that luminal cell lines significantly more frequently harboured gains on 7q, 8q and 12q and losses on 1p, 2q, 3q, 4q, 16q and 20q, whereas basal-like cell lines significantly more frequently displayed gains on 2p, 9p and 19q and losses on 3q, 4q and 12q (Fig. 4).

We next determined the high level gains/amplifications significantly associated with basal-like and luminal cell lines. Luminal cell lines more frequently harboured amplifications on 1q and 17q12. Basal-like cell lines did not harbour any high-level amplifications not observed in luminal-like cell lines (Fig. 5). The complete list of regions

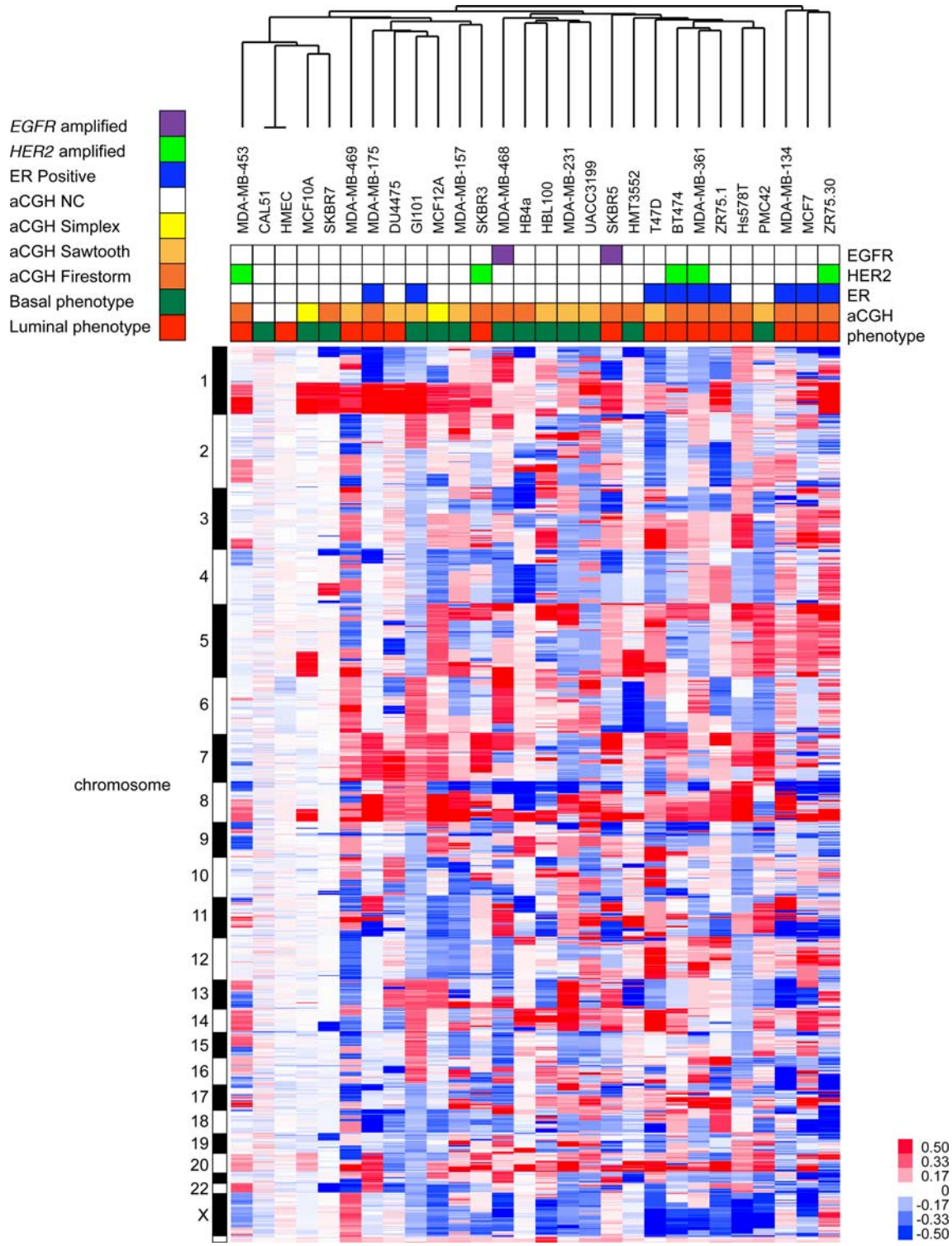


Fig. 2 Heatmap of unsupervised clustering using the aCGH data of breast cancer cell lines. aws-smoothed log ratios from 31,544 BACs were used to cluster the cell lines using a Wards algorithm based upon Euclidean distance. Chromosomes are indicated with a *bar* on the left. Phenotypes of each sample are shown below the dendrogram in the *top panel*. Basal and Luminal phenotype shown is based upon PAM prediction. aCGH genomic profiles shown are NC (no change), Simplex, Sawtooth, Firestorm. ER status is shown as positive/negative. *HER2* and *EGFR* status are shown as normal and amplified

Table 3 The results of an analysis to identify genes whose expression is correlated with copy number: correlation between expression values and smoothed aCGH log2 ratios

Symbol	Description	Chrom	Start	Cytoband	Cor	Unigene
ECD	Ecdysoneless homolog	10	74.56	10q22.1	0.896	Hs.631822
FBXL20	F-box and leucine-rich repeat protein 20	17	34.67	17q12	0.895	Hs.462946
SDF2	Stromal cell-derived factor 2	17	24.00	17q11.2	0.869	Hs.514036
METTL6	Methyltransferase like 6	3	15.43	3p24.3	0.858	Hs.149487
C8orf32	Chromosome 8 open reading frame 32	8	124.52	8q24.13	0.859	Hs.18029
CISD3	CDGSH iron sulphur domain 3	17	34.14	17q12	0.857	Hs.462923
PSMB3	Proteasome (macropain) subunit, beta type, 3	17	34.16	17q12	0.860	Hs.82793
ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2	17	35.13	17q12	0.859	Hs.446352
AASDHPPT	Aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase	11	105.47	11q22.3	0.854	Hs.524009
MED24	Mediator complex subunit 24	17	35.43	17q21.1	0.853	Hs.462983
KIAA0100	KIAA0100	17	23.97	17q11.2	0.849	Hs.591189
KLHL9	Kelch-like 9 (*Drosophila*)	9	21.32	9p21.3	0.848	Hs.522029
PPFIA1	Protein tyrosine phosphatase, interacting protein, alpha 1	11	69.91	11q13.3	0.842	Hs.530749
C3orf31	Chromosome 3 open reading frame 31	3	11.81	3p25.2	0.840	Hs.475472
CEP192	Centrosomal protein 192 kDa	18	13.10	18p11.21	0.837	Hs.100914
WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1	8	38.29	8p12	0.832	Hs.700599
UNC119	Unc-119 homolog (*C. elegans*)	17	23.90	17q11.2	0.833	Hs.410455
TRPC4AP	Transient receptor potential cation channel, subfamily C, member 4 associated protein	20	33.05	20q11.22	0.832	Hs.168073
DDX24	DEAD (Asp–Glu–Ala–Asp) box polypeptide 24	14	93.59	14q32.13	0.830	Hs.510328
RNF169	Ring finger protein 169	11	74.23	11q13.4	0.829	Hs.556037
C13orf1	Chromosome 13 open reading frame 1	13	49.39	13q14.3	0.826	Hs.44235
POLDIP2	Polymerase (DNA-directed), delta interacting protein 2	17	23.70	17q11.2	0.825	Hs.241543
LASP1	LIM and SH3 protein 1	17	34.33	17q12	0.825	Hs.548018
ZFYVE20	Zinc finger, FYVE domain containing 20	3	15.09	3p24.3	0.822	Hs.475565
GLT8D1	Glycosyltransferase 8 domain containing 1	3	52.70	3p21.1	0.822	Hs.297304
PHB	Prohibitin	17	44.84	17q21.33	0.820	Hs.514303
NDUFB9	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9	8	125.62	8q24.13	0.818	Hs.15977
EST	Transcribed locus	3	13.52	3p25.1	0.807	Hs.404802
ASH2L	Ash2 (absent, small, or homeotic)-like (*Drosophila*)	8	38.11	8p12	0.807	Hs.521530
MRPS28	Mitochondrial ribosomal protein S28	8	81.08	8q21.13	0.805	Hs.521124

About 753 genes were identified which had significant adjusted *P* values in a Pearson correlation (Cor) between expression values and smoothed aCGH log2 ratios for each gene. Table 3 shows the 30 most significant genes. The complete list of significantly correlated genes is available in Supplementary Table S5

identified in the multi-Fisher's comparison is listed in Supplementary Table S7 and all recurrent contiguous aCGH aberrations are listed in Supplementary Table S8.

Discussion

Here we provide the most comprehensive characterisation to date, of the genomic profiles of 24 breast cancer cell lines using a high resolution aCGH platform. The 32 K tiling path BAC array employed in this study has been shown to be as robust as and to have a similar resolution to oligonucleotide arrays. Data generated with this platform not only provided a detailed characterisation of the genomic profiles of these cell lines, but also allowed for a direct integration of genomic and transcriptomic data.

Our results demonstrate that breast cancer cell lines have genomic and transcriptomic features that recapitulate those of primary breast cancers [2, 4, 34, 35, 45]. First, in accordance with previous studies [1, 4], we show that breast cancer cell lines can be subclassified in basal-like and luminal subgroups and that the genes associated with each subgroup are remarkably similar to those significantly expressed in primary basal-like and luminal breast cancers, respectively [3, 5, 6]. Interestingly, unlike in primary breast cancers, the molecular subgroups luminal A, normal breast-like and HER2 could not be reliably identified. Luminal A cancers are usually of low histological grade,

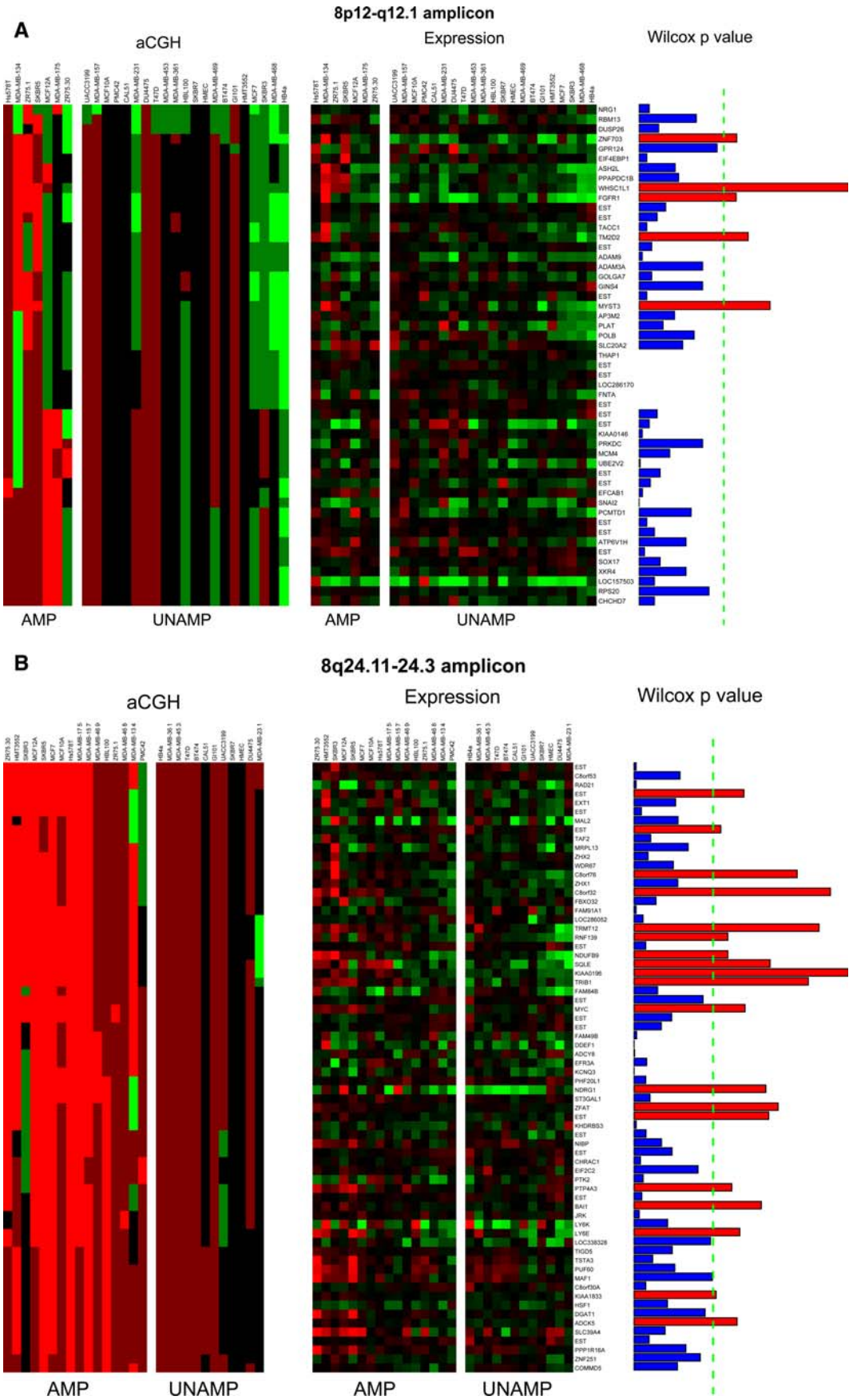
strong ERα expressers and rarely metastasise, whereas breast cancer cell lines harbour genomic features consistent with high histological grade (e.g., gains of 1q, 8q and 20q) [31], are usually derived from ERα negative samples and metastatic deposits [4]. Normal breast-like has been shown to be an unstable subgroup and there is evidence to suggest that samples pertaining to this subgroup are enriched for stromal cells [46, 47]. Second, although numerous genetic changes are prevalent in all subgroups of breast cancer cell lines (e.g., gains of 1q, 8q and 20q), each subgroup seems to be characterised by a rather specific constellation of genetic changes (Figs. 4, 5). Interestingly, the regions identified as gained, lost or amplified in basal-like and luminal cell lines are similar to those reported to be changed in primary breast cancers of similar phenotype. Interestingly, losses of the whole long arm of chromosome 16, which are usually found in low grade and luminal breast cancers [31], were significantly more prevalent in luminal cell lines. Here we also demonstrated that in a way akin to primary breast cancers, *HER2* amplified cell lines preferentially display a 'firestorm' genetic profile [18, 31] and that high level amplifications are not common events in basal-like cell lines [2, 4, 27].

Owing to the high resolution and near-complete genomic coverage of this platform, we were able to directly integrate genomic and transcriptmic data. Our analysis revealed that 753 genes displayed expression levels that correlated with copy numbers, suggesting that to some extent the expression of these genes may be affected by gene dosage. These genes preferentially mapped to six chromosomes 17, 8, 11, 1, 3 and 6. This analysis identified numerous genes previously shown to be involved in breast cancer development and progression. It should be noted *ESR1* expression levels did not correlate with copy number and no amplification of *ESR1* gene was found, corroborating previous studies by our group [23] and others [48–50] and calling into question the results by Holst et al. [51], who described *ESR1* gene amplification in >20% of all breast cancers.

Tumour suppressor genes known to be deleted in breast cancers failed to show a correlation between expression and copy number. This is not surprising, as homozygous deletions are uncommon events in breast cancer and haploinsufficiency is not a common phenomenon. Furthermore, tumour suppressor genes whose inactivation is mediated though a combination of deletions and missense mutations are not uncommonly overexpressed at mRNA and/or protein levels. For instance, *TP53* missense mutations have been shown to be associated with increased levels of protein expression [37]. Allelic imbalance of *RB1* gene has been shown to be associated with increased mRNA expression [52] and p16 and p14 overexpression in the presence of intragenic deletions or single point mutations [38]. It could be hypothesised that some genes that are heterozygously deleted an overexpressed, may be potential tumour suppressor genes where the second hit is a missense mutation leading to mRNA overexpression possibly due to dysfunctional negative feedback loops. Here, we identified three protein coding genes that displayed a significant inverse correlation between expression and copy number and showed more frequent deletions than gains and/or amplification: *NSDHL*, *EIF4E3* and *RFX1*. Interestingly, germline mutations of *NSDHL* cause congenital hemidysplasia ichthyosiform eithroderma and limb defect (CHILD) syndrome [53]. *RFX1* has been implicated in transcriptional downregulation of the proto-oncogene c-myc and shown to be epigenetically silenced in human glioma cell lines and tissues [54]. No mutations in these genes have been found on COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/). It should be noted, however, that in COSMIC the number of samples analysed may not be sufficient for ruling out the presence of mutations in these genes in distinct phenotypes of breast cancer.

Finally, the integrated genomic-transcriptomic analysis of amplicons in breast cancer cell lines demonstrates that the amplicons found in these cell lines are similar to those described in primary breast cancer. We [31, 42, 55] and others [39, 43] have previously demonstrated that genes significantly overexpressed when amplified are likely amplicon drivers, as the expression of these genes is required for the survival of cells harbouring amplification of their genomic region. Our integrated analysis has not only identified genes previously shown to be potential drivers of amplicons 8p11.2 (*FGFR1*), 8q24 (*MYC*), 11q13 (*CTTN*, *FADD*), 17q12 (*ERBB2*) and 20q13 (*AURKA*), but also identified a list of 269 that should be investigated as potential therapeutic targets in the amplicons on chromosomes 1q, 3q, 6p, 8q, 13q, 14q, 17q, 19q and 20q (Table 4; Supplementary Table S6). Further functional studies to determine which genes are required for the survival of cells harbouring amplification of their genomic regions (i.e., the likeliest amplicon drivers) are perhaps warranted [56].

Fig. 3 Matched heatmaps of expression and aCGH within regions of recurrent amplification in breast cancer cell lines. Matched heatmaps are shown for recurrently amplified regions 8p12–8q12.1 (**a**) 8q24.11–8q24.3 (**b**) 17q12–21.1 (**c**) and 20q13.13–20q13.33 (**d**). Additional amplicons are shown in Supplementary Figs. 1–9. For each amplicon, RefSeq genes within the amplified region are recovered and median aCGH values and states are assigned. Samples are separated into those harbouring an amplification within the region and those that do not. Expression and aCGH values are depicted in two matching heatmaps (aCGH states on the left and expression values on the right) in which the genes are ordered according to their chromosomal position and the cell lines ordered according to the sum of their aCGH values. Bar plots show the result of a Wilcoxon rank sum test for each gene using the aCGH states at that point as the grouping variable to test expression values. *Bars in red* show unadjusted *P*-values of less than 0.05 ▶


A
8p12-q12.1 amplicon
aCGH
Expression
Wilcox p value
AMP
UNAMP
AMP
UNAMP
B
8q24.11-24.3 amplicon
aCGH
Expression
Wilcox p value
AMP
UNAMP
AMP
UNAMP

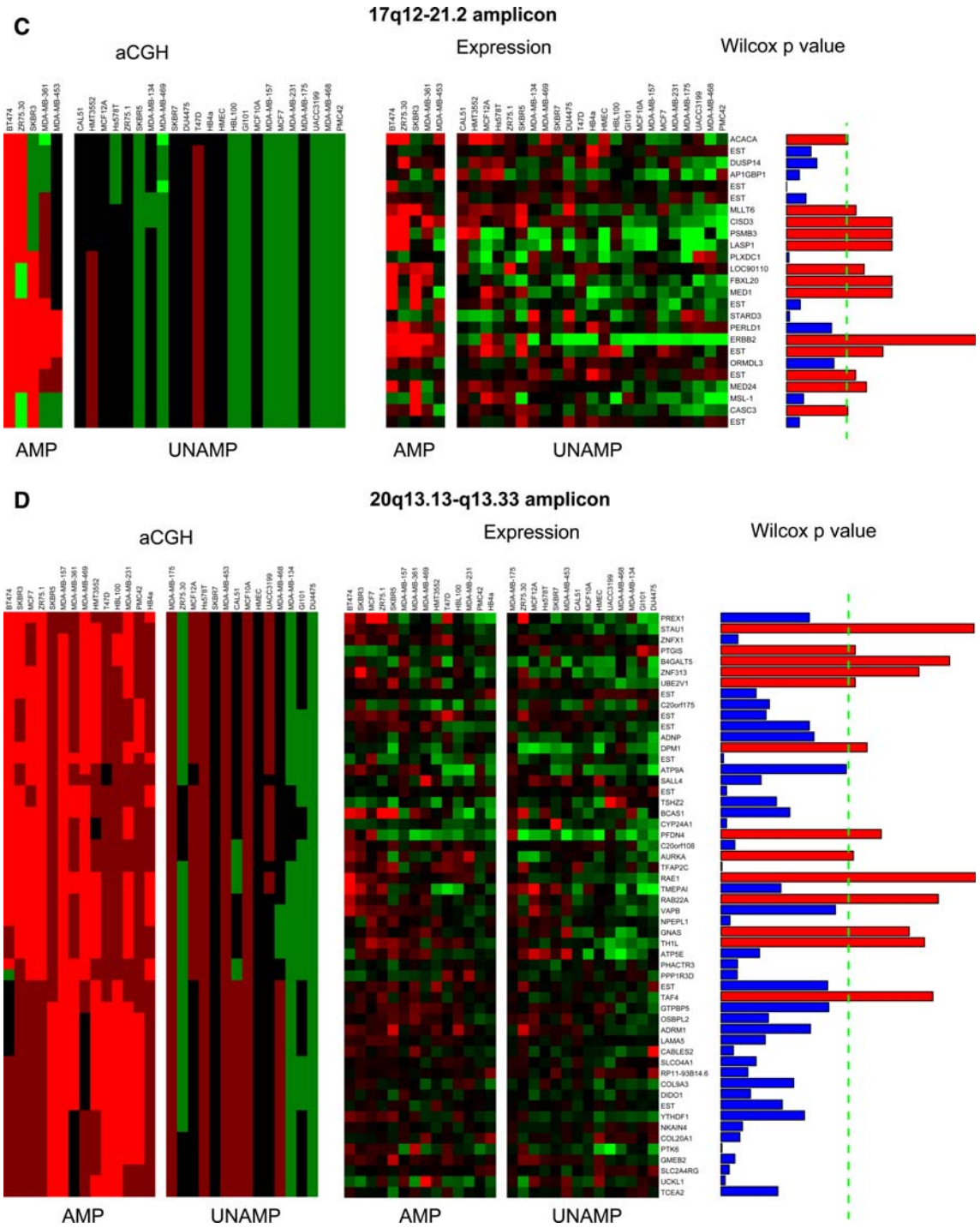


Fig. 3 continued

The high resolution of the aCGH platform employed here has confirmed the structural complexity of amplicons found in breast cancer cell lines and primary breast cancers [40, 41, 57]. For instance, in agreement with previous studies [40, 41], on 8p11.2–q12.1, 4 smallest regions of amplification (SRAs) were identified, whereas on 20q13, at least three distinct SRAs were found. It should be noted, however, that the results of our study suggest that the SRA of each amplicon in breast cancer cell lines is larger than that found in human cancers (e.g., the 17q12–q21 SRA in this study spanned the region from 32.33 to 35.69 Mb, whereas in a previous study using the same platform, we have demonstrated that the 17q12–q21 SRA in primary breast cancers spans 414 kb) [18]. This is likely to reflect

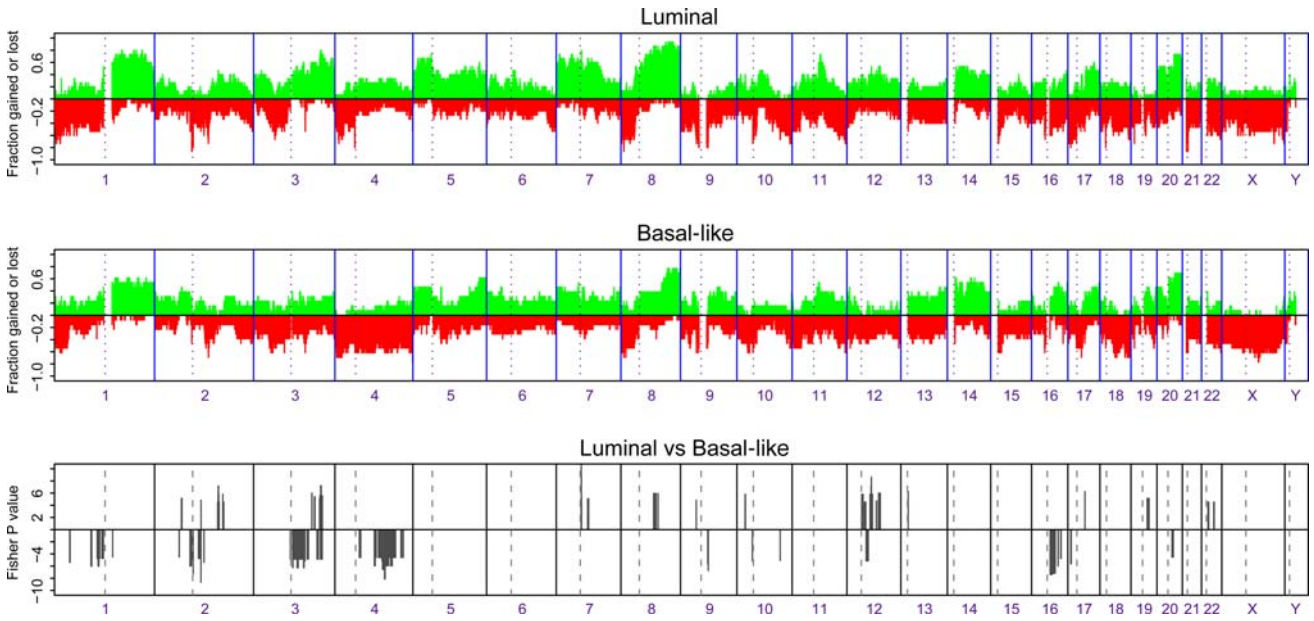


Fig. 4 Frequency of gains and losses described by aCGH profiling of breast cancer cell lines. Contiguous regions of gain and loss identified with aws-smoothed log ratios greater than ±0.08 from zero are plotted for *lines* with predicted luminal and basal phenotypes. The results of a Fishers test to compare counts of gain and loss between the two groups are plotted in the *bottom panel* as log(10) of the *P* values for all BACs with unadjusted *P* values less than 0.05 for this comparison

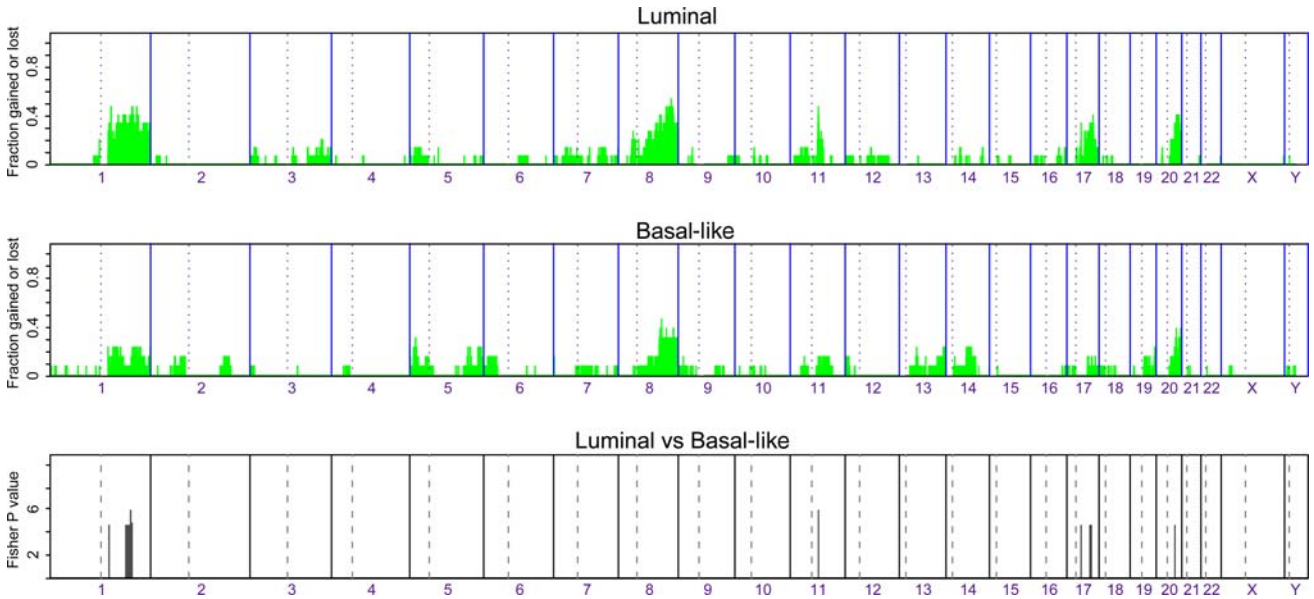


Fig. 5 Frequency of amplifications and deletions described by aCGH profiling of breast cancer cell lines. Contiguous regions of amplification and deletion identified with aws-smoothed log ratios greater than ±0.4 from zero are plotted for *lines* with predicted luminal and basal phenotypes. The results of a Fishers test to compare counts of amplification and deletion between the two groups are plotted in the *bottom panel* as log(10) of the *P* values for all BACs with unadjusted *P* values less than 0.05 for this comparison

the effect of analysing a larger number of samples and emphasises that to define SRAs, both the resolution of the platform and the number of samples are of paramount importance [58]. Hence, further studies of larger numbers of cell lines of each molecular subtype are warranted to determine the amplicons associated with each molecular subtype. Furthermore, additional samples are required to determine whether the SRAs of amplicons in cell lines are similar to those of primary breast cancers and whether amplicons are stable in tumours pertaining to distinct molecular subgroups.

In conclusion, we have characterised the genomic and transcriptomic profiles of 24 breast cancer cell lines and demonstrated that these cell lines have molecular features

Table 4 The results of an analysis to identify genes whose expression is correlated with copy number: overexpression of genes within amplified regions

Symbol	Description	Chrom	Start	Cytoband	*P* value	Unigene
CTTN	Cortactin	11	69.96	11q13.3	0.002	Hs.632133
ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2	17	35.13	17q12	0.004	Hs.446352
COIL	Coilin	17	52.37	17q22	0.006	Hs.532795
COX11	Cytochrome *c* oxidase assembly protein	17	50.39	17q22	0.006	Hs.591171
PTRH2	Peptidyl-tRNA hydrolase 2	17	55.13	17q23.1	0.006	Hs.12677
RNF169	Ring finger protein 169	11	74.23	11q13.4	0.006	Hs.556037
PPFIA1	Protein tyrosine phosphatase, interacting protein, alpha 1	11	69.91	11q13.3	0.010	Hs.530749
PRKAR1A	Protein kinase, cAMP-dependent, regulatory, type I, alpha	17	64.04	17q24.2	0.010	Hs.280342
NOL11	Nucleolar protein 11	17	63.16	17q24.2	0.011	Hs.463936
C17orf58	Chromosome 17 ORF 58	17	63.42	17q24.2	0.011	Hs.90790
C11orf67	Chromosome 11 ORF 67	11	77.23	11q14.1	0.015	Hs.503357
WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1	8	38.29	8p12	0.015	Hs.700599
EST	Transcribed locus	11	67.68	11q13.2	0.020	Hs.503001
SLC35B1	Solute carrier family 35, member B1	17	45.14	17q21.33	0.024	Hs.154073
ANKRD40	Ankyrin repeat domain 40	17	46.13	17q21.33	0.024	Hs.463426
C17orf71	Chromosome 17 ORF 71	17	54.65	17q22	0.024	Hs.7296
JARID1A	Jumonji, AT rich interactive domain 1A	12	0.26	12p13.33	0.025	Hs.654806
SSH3	Slingshot homolog 3 (*Drosophila*)	11	66.84	11q13.1	0.026	Hs.29173
ANKRD13D	Ankyrin repeat domain 13 family, member D	11	66.82	11q13.1	0.026	Hs.438673
UNC119	Unc-119 homolog	17	23.90	17q11.2	0.026	Hs.410455
PIGS	Phosphatidylinositol glycan anchor biosynthesis, class S	17	23.90	17q11.2	0.026	Hs.462550
KIAA0100	KIAA0100	17	23.97	17q11.2	0.026	Hs.591189
TLCD1	TLC domain containing 1	17	24.08	17q11.2	0.026	Hs.705716
TRAF4	TNF receptor-associated factor 4	17	24.10	17q11.2	0.026	Hs.8375
EST	Transcribed locus	17	56.10	17q23.2	0.031	Hs.286073
PCID2	PCI domain containing 2	13	112.88	13q34	0.031	Hs.508769
MAPK1IP1L	Mitogen-activated protein kinase 1 interacting protein 1-like	14	54.60	14q22.3	0.032	Hs.594338
PHB	Prohibitin	17	44.84	17q21.33	0.033	Hs.514303
TRIM37	Tripartite motif-containing 37	17	54.43	17q22	0.033	Hs.579079
EST	Transcribed locus	17	54.64	17q22	0.033	Hs.634149

About 269 genes were significant when tested for overexpression within amplified regions using a Wilcoxon rank sum test to compare the expression values in cell lines harbouring an amplification covering each gene with those that do not. Table 4 shows the top 30 genes significantly amplified when overexpressed listing the adjusted *P* value for the each gene. The complete list of genes overexpressed when amplified is available in Supplementary Table S6

that closely recapitulate those of primary breast cancer. Given that the heterogeneity found in primary breast cancer is mirrored by breast cancer cell lines and that cell lines pertaining to distinct molecular subgroups have different genetic aberrations, our results suggest that specific cell lines harbouring the correct phenotype and genotype should be employed for in vitro and in vivo modelling of subgroups of breast cancer. In fact, our results constitute a useful resource for other groups to select the most appropriate cell line models. Finally, we demonstrate that amplifications in breast cancer cells are complex and amplicons may harbour more than one driver.

Acknowledgments We thank Breakthrough Breast Cancer and Cancer Research UK for their continued support of this work.

References

1. Charafe-Jauffret E, Ginestier C, Monville F et al (2006) Gene expression profiling of breast cell lines identifies potential new basal markers. Oncogene 25:2273–2284. doi:10.1038/sj.onc.1209254
2. Chin K, DeVries S, Fridlyand J et al (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 10:529–541. doi:10.1016/j.ccr.2006.10.009

3. Hu Z, Fan C, Oh DS et al (2006) The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7:96. doi:10.1186/1471-2164-7-96
4. Neve RM, Chin K, Fridlyand J et al (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 10:515–527. doi:10.1016/j.ccr.2006.10.008
5. Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. Nature 406:747–752. doi:10.1038/35021093
6. Sorlie T, Tibshirani R, Parker J et al (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100:8418–8423. doi:10.1073/pnas.0932692100
7. Brenton JD, Carey LA, Ahmed AA et al (2005) Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol 23:7350–7360. doi:10.1200/JCO.2005.03.3845
8. Rakha EA, Reis-Filho JS, Ellis IO (2008) Basal-like breast cancer: a critical review. J Clin Oncol 26:2568–2581. doi:10.1200/JCO.2007.13.1748
9. Lacroix M, Leclercq G (2004) Relevance of breast cancer cell lines as models for breast tumours: an update. Breast Cancer Res Treat 83:249–289. doi:10.1023/B:BREA.0000014042.54925.cc
10. Stamps AC, Davies SC, Burman J et al (1994) Analysis of proviral integration in human mammary epithelial cell lines immortalized by retroviral infection with a temperature-sensitive SV40 T-antigen construct. Int J Cancer 57:865–874. doi:10.1002/ijc.2910570616
11. Clarke C, Titley J, Davies S et al (1994) An immunomagnetic separation method using superparamagnetic (MACS) beads for large-scale purification of human mammary luminal and myoepithelial cells. Epithelial Cell Biol 3:38–46
12. O'Hare MJ, Bond J, Clarke C et al (2001) Conditional immortalization of freshly isolated human mammary fibroblasts and endothelial cells. Proc Natl Acad Sci USA 98:646–651. doi:10.1073/pnas.98.2.646
13. Freshney RI (2005) Culture of animal cells: a manual of basic technique, 5th edn. Wiley-Liss, New York
14. Arriola E, Lambros MB, Jones C et al (2007) Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. Lab Invest 87:75–83. doi:10.1038/labinvest.3700495
15. Reis-Filho JS, Simpson PT, Jones C et al (2005) Pleomorphic lobular carcinoma of the breast: role of comprehensive molecular pathology in characterization of an entity. J Pathol 207:1–13. doi:10.1002/path.1806
16. Arriola E, Marchio C, Tan DS et al (2008) Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines. Lab Invest 88:491–503. doi:10.1038/labinvest.2008.19
17. Marchio C, Iravani M, Natrajan R et al (2008) Genomic and immunophenotypical characterization of pure micropapillary carcinomas of the breast. J Pathol 215:398–410. doi:10.1002/path.2368
18. Marchio C, Natrajan R, Shiu K et al (2008) The genomic profile of HER2-amplified breast cancers: the influence of ER status. J Pathol 216:399–407. doi:10.1002/path.2423
19. Mackay A, Urruticoechea A, Dixon JM et al (2007) Molecular response to aromatase inhibitor treatment in primary breast cancer. Breast Cancer Res 9:R37. doi:10.1186/bcr1732
20. Tibshirani R, Hastie T, Narasimhan B et al (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99:6567–6572. doi:10.1073/pnas.082099299
21. Pierga JY, Reis-Filho JS, Cleator SJ et al (2007) Microarray-based comparative genomic hybridisation of breast cancer patients receiving neoadjuvant chemotherapy. Br J Cancer 96:341–351. doi:10.1038/sj.bjc.6603483
22. Natrajan R, Little SE, Sodha N et al (2007) Analysis by array CGH of genomic changes associated with the progression or relapse of Wilms' tumour. J Pathol 211:52–59. doi:10.1002/path.2087
23. Reis-Filho JS, Drury S, Lambros MB et al (2008) ESR1 gene amplification in breast cancer: a common phenomenon? Nat Genet 40:809–810. doi:10.1038/ng0708-809b (author reply 810–802)
24. Hicks J, Krasnitz A, Lakshmi B et al (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Res 16:1465–1479. doi:10.1101/gr.5460106
25. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc [Ser A] 57:289–300
26. Jones C, Mackay A, Grigoriadis A et al (2004) Expression profiling of purified normal human luminal and myoepithelial breast cells: identification of novel prognostic markers for breast cancer. Cancer Res 64:3037–3045. doi:10.1158/0008-5472.CAN-03-2028
27. Jonsson G, Staaf J, Olsson E et al (2007) High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. Genes Chromosomes Cancer 46:543–558. doi:10.1002/gcc.20438
28. Savage K, Lambros MB, Robertson D et al (2007) Caveolin 1 is overexpressed and amplified in a subset of basal-like and metaplastic breast carcinomas: a morphologic, ultrastructural, immunohistochemical, and in situ hybridization analysis. Clin Cancer Res 13:90–101. doi:10.1158/1078-0432.CCR-06-1371
29. Savage K, Leung S, Todd SK et al (2008) Distribution and significance of caveolin 2 expression in normal breast and invasive breast cancer: an immunofluorescence and immunohistochemical analysis. Breast Cancer Res Treat 110:245–256. doi:10.1007/s10549-007-9718-1
30. Weigelt B, Kreike B, Reis-Filho JS (2008) Metaplastic breast carcinomas are basal-like breast cancers: a genomic profiling analysis. Breast Cancer Res Treat. doi:10.1007/s10549-008-0197-9
31. Natrajan R, Lambros MB, Rodrigues Pinilla SM, et al (2008) Tiling path genomic profiling of grade III invasive ductal breast cancers. Clin Cancer Res (in press)
32. Seitz S, Wassmuth P, Plaschke J et al (2003) Identification of microsatellite instability and mismatch repair gene mutations in breast cancer cell lines. Genes Chromosomes Cancer 37:29–35. doi:10.1002/gcc.10196
33. Schlegel J, Stumm G, Scherthan H et al (1995) Comparative genomic in situ hybridization of colon carcinomas with replication error. Cancer Res 55:6002–6005
34. Chin SF, Teschendorff AE, Marioni JC et al (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol 8:R215. doi:10.1186/gb-2007-8-10-r215
35. Adelaide J, Finetti P, Bekhouche I et al (2007) Integrated profiling of basal and luminal breast cancers. Cancer Res 67:11565–11575. doi:10.1158/0008-5472.CAN-07-2536
36. Bartek J, Bartkova J, Vojtesek B et al (1991) Aberrant expression of the p53 oncoprotein is a common feature of a wide spectrum of human malignancies. Oncogene 6:1699–1703
37. Hsu HC, Tseng HJ, Lai PL et al (1993) Expression of p53 gene in 184 unifocal hepatocellular carcinomas: association with tumor growth and invasiveness. Cancer Res 53:4691–4694
38. Lang JC, Borchers J, Danahey D et al (2002) Mutational status of overexpressed p16 in head and neck cancer: evidence for germline mutation of p16/p14ARF. Int J Oncol 21:401–408
39. Bernard-Pierrot I, Gruel N, Stransky N et al (2008) Characterization of the recurrent 8p11–12 amplicon identifies PPAPDC1B, a phosphatase protein, as a new therapeutic target in breast cancer. Cancer Res 68:7165–7175. doi:10.1158/0008-5472.CAN-08-1360
40. Gelsi-Boyer V, Orsetti B, Cervera N et al (2005) Comprehensive profiling of 8p11–12 amplification in breast cancer. Mol Cancer Res 3:655–667. doi:10.1158/1541-7786.MCR-05-0128

41. Ginestier C, Cervera N, Finetti P et al (2006) Prognosis and gene expression profiling of 20q13-amplified breast cancers. Clin Cancer Res 12:4533–4544. doi:10.1158/1078-0432.CCR-05-2339
42. Reis-Filho JS, Simpson PT, Turner NC et al (2006) FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas. Clin Cancer Res 12:6652–6662. doi:10.1158/1078-0432.CCR-06-1164
43. Cheng KW, Lahad JP, Kuo WL et al (2004) The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers. Nat Med 10:1251–1256. doi:10.1038/nm1125
44. Orsetti B, Nugoli M, Cervera N et al (2004) Genomic and expression profiling of chromosome 17 in breast cancer reveals complex patterns of alterations and novel candidate genes. Cancer Res 64:6453–6460. doi:10.1158/0008-5472.CAN-04-0756
45. Bergamaschi A, Kim YH, Wang P et al (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer 45:1033–1040. doi:10.1002/gcc.20366
46. Farmer P, Bonnefoi H, Becette V et al (2005) Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 24:4660–4671. doi:10.1038/sj.onc.1208561
47. Rouzier R, Perou CM, Symmans WF et al (2005) Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res 11:5678–5685. doi:10.1158/1078-0432.CCR-04-2421
48. Brown LA, Hoog J, Chin SF et al (2008) ESR1 gene amplification in breast cancer: a common phenomenon? Nat Genet 40:806–807. doi:10.1038/ng0708-806 (author reply 810–802)
49. Horlings HM, Bergamaschi A, Nordgard SH et al (2008) ESR1 gene amplification in breast cancer: a common phenomenon? Nat Genet 40:807–808. doi:10.1038/ng0708-807 (author reply 810–802)
50. Vincent-Salomon A, Raynal V, Lucchesi C et al (2008) ESR1 gene amplification in breast cancer: a common phenomenon? Nat Genet 40:809. doi:10.1038/ng0708-809a (author reply 810–802)
51. Holst F, Stahl PR, Ruiz C et al (2007) Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. Nat Genet 39:655–660. doi:10.1038/ng2006
52. Lai PS, Cheah PY, Kadam P et al (2006) Overexpression of RB1 transcript is significantly correlated with 13q14 allelic imbalance in colorectal carcinomas. Int J Cancer 119:1061–1066. doi:10.1002/ijc.21945
53. Konig A, Happle R, Bornholdt D et al (2000) Mutations in the NSDHL gene, encoding a 3beta-hydroxysteroid dehydrogenase, cause CHILD syndrome. Am J Med Genet 90:339–346. doi:10.1002/(SICI)1096-8628(20000214)90:4<339::AID-AJMG15>3.0.CO;2-5
54. Ohashi Y, Ueda M, Kawase T et al (2004) Identification of an epigenetically silenced gene, RFX1, in human glioma cells using restriction landmark genomic scanning. Oncogene 23:7772–7779. doi:10.1038/sj.onc.1208058
55. Reis-Filho JS, Pinheiro C, Lambros MB et al (2006) EGFR amplification and lack of activating mutations in metaplastic breast carcinomas. J Pathol 209:445–453. doi:10.1002/path.2004
56. Iorns E, Lord CJ, Turner N et al (2007) Utilizing RNA interference to enhance cancer drug discovery. Nat Rev Drug Discov 6:556–568. doi:10.1038/nrd2355
57. Campbell PJ, Stephens PJ, Pleasance ED et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40:722–729. doi:10.1038/ng.128
58. Albertson DG, Snijders AM, Fridlyand J et al (2006) Genomic analysis of tumors by array comparative genomic hybridization: more is better. Cancer Res 66:3955–3956. doi:10.1158/0008-5472.CAN-05-3611 (author reply 3956)