Molecular Characterization of Breast Cancer with High-Resolution Oligonucleotide Comparative Genomic Hybridization Array

Fabrice Andre,[1,2] Bastien Job,[3] Philippe Dessen,[4,6] Attila Tordai,[7] Stefan Michiels,[5] Cornelia Liedtke,[8] Catherine Richon,[3] Kai Yan,[7] Bailang Wang,[7] Gilles Vassal,[1] Suzette Delaloge,[1,2] Gabriel N. Hortobagyi,[8] W. Fraser Symmans,[9] Vladimir Lazar,[3] and Lajos Pusztai[8]

Abstract

Purpose: We used high-resolution oligonucleotide comparative genomic hybridization (CGH) arrays and matching gene expression array data to identify dysregulated genes and to classify breast cancers according to gene copy number anomalies.

Experimental Design: DNA was extracted from 106 pretreatment fine needle aspirations of stage II-III breast cancers that received preoperative chemotherapy. CGH was done using Agilent Human 4 × 44K arrays. Gene expression data generated with Affymetrix U133A gene chips was also available on 103 patients. All *P* values were adjusted for multiple comparisons.

Results: The average number of copy number abnormalities in individual tumors was 76 (range 1-318). Eleven and 37 distinct minimal common regions were gained or lost in >20% of samples, respectively. Several potential therapeutic targets were identified, including *FGFR1* that showed high-level amplification in 10% of cases. Close correlation between DNA copy number and mRNA expression levels was detected. Nonnegative matrix factorization (NMF) clustering of DNA copy number aberrations revealed three distinct molecular classes in this data set. NMF class I was characterized by a high rate of triple-negative cancers (64%) and gains of 6p21. *VEGFA, E2F3*, and *NOTCH4* were also gained in 29% to 34% of triple-negative tumors. A gain of *ERBB2* gene was observed in 52% of NMF class II and class III was characterized by a high rate of estrogen receptor – positive tumors (73%) and a low rate of pathologic complete response to preoperative chemotherapy (3%).

Conclusion: The present study identified dysregulated genes that could classify breast cancer and may represent novel therapeutic targets for molecular subsets of cancers.

Authors' Affiliations: [1]Translational Research Unit, UPRES03535, Paris Sud University; [2]Breast Cancer Unit, Department of Medicine; [3]Functional Genomic Unit; [4]FRE 2939 Centre National de la Recherche Scientifique; and [5]Department of Biostatistics and Epidemiology, Institut Gustave Roussy, Villejuif, France; [6]Paris Sud University, Orsay, France; [7]National Blood Transfusion Service, Molecular Diagnostics Laboratory, Budapest, Hungary; and Departments of [8]Breast Medical Oncology and [9]Pathology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas

Received 7/11/08; revised 9/9/08; accepted 9/13/08.

Grant support: National Cancer Institute grant RO1-CA106290, the Breast Cancer Research Foundation, and the Goodwin Foundation (L. Pusztai); the Nellie B. Connally Breast Cancer Research Fund (G.N. Hortobagyi); Fondation de France, Fondation Lilly, Ligue Contre le Cancer, Paris, a career development award from American Society of Clinical Oncology, and the operation "support researchers" from Institut Gustave Roussy (F. Andre); a grant from the Deutsche Forschungsgemeinschaft, Germany (C. Liedtke); and the Hungarian American Enterprise Scholarship Fund (A. Tordai).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Requests for reprints: Fabrice Andre, Breast Cancer Unit, Institut Gustave Roussy, 39 Rue C. Desmoulins, 94805 Villejuif, France. Phone: 33-14-2114371; Fax: 33-14-2115274; E-mail: fandre@igr.fr.

doi:10.1158/1078-0432.CCR-08-1791

Genomic instability leads to DNA copy number abnormalities in cancer cells (1). Patterns of DNA copy number alterations may define distinct subsets of breast cancers and frequent alterations in particular chromosomal locations can draw attention to genes that are functionally important in carcinogenesis. Identification of these genes could allow the characterization of new oncogenic pathways and lead to the discovery of new therapeutic targets. DNA copy number changes may also serve as molecular markers of prognosis and response to therapy. Several studies examined whole genome copy number alterations in breast cancer (2–8). Most of these studies used metaphase comparative genomic hybridization (CGH) with bacterial artificial chromosomes (BAC) that have relatively low resolution (2). These studies established that gains in chromosomes 1q, 8q, 17q, 20q and losses in 5q, 6q, 8p are common in breast cancer (2). In addition, these studies also suggested that estrogen receptor (ER)–negative cancers frequently harbor losses in 5q and gains in 6p compared with hormone receptor–positive cancers (2). However, these earlier CGH technologies do not allow for fine mapping of copy number alterations at the individual gene level due to their low resolution; BAC arrays typically scan the genome at 1 to 2 Mb intervals.

Translational Relevance

The present study characterizes gene copy number anomalies in breast cancer. It presents several potential implications for cancer medicine. Based on a comparison with fluorescence *in situ* hybridization, it first shows that comparative genomic hybridization (CGH) array is a robust technology to assess gene copy number anomalies. Then, we identify two sets of gene copy number anomalies that could lead to the development of targeted therapy or subgroup analyses in breast cancer. The *FGFR1* gene was found to be amplified (log 2 ratio >1.5) in 10% of overall samples. Gene copy number anomalies were found to be enriched in triple-negative (ER-/PR-/Her2-) disease, including *VEGFA, NOTCH4*, and *E2F3* gains. Finally, unsupervised clustering identifies three CGH array classes of breast cancer. Further larger studies will determine whether such classification could complement the current gene expression based molecular classification for patient stratification in clinical trials.

In the current study, we used high-resolution genome-wide oligonucleotide CGH arrays (Agilent Human 4 × 44K CGH Array; ref. 9). These arrays scan the genome on average at 70-kb intervals. The goals of this study were to catalogue high-resolution DNA copy number alterations in breast cancer and correlate these genomic anomalies with mRNA expression and clinical characteristics, including response to preoperative chemotherapy. We also did consensus clustering of breast cancer samples based on DNA copy number anomalies to examine if new molecular classes of breast cancer could be discerned.

Patients and Methods

Patient selection and clinical characteristics. This study included 106 patients who participated in a pharmacogenomic predictive marker discovery study at the University of Texas M. D. Anderson Cancer Center. During this research, patients with newly diagnosed stage I-III breast cancer were asked to undergo pretreatment fine needle aspiration of the cancer using a 23- or 25-gauge needle before any therapy. Cells from two to three passes were collected into vials containing 1 mL of RNAlater solution (Ambion) and stored at -80°C. We previously established that fine needle aspirations contain predominantly neoplastic cells (80-95%) and some infiltrating leukocytes (5-15%) but are devoid of stromal elements (10). This is consistent with the general pathology literature and allows for RNA yield to serve as a quality control metric for neoplastic cellularity. Poor RNA yield indicates lack of sufficient amount of neoplastic cells in the biopsy. We also reported previously that gene expression data generated from fine needle aspirations captures the molecular characteristics of the invasive cancer, including molecular class (11). All patients received preoperative chemotherapy. Clinical characteristics and treatments are summarized in Supplementary Table S1.

All patients underwent modified radical mastectomy or lumpectomy and sentinel node dissection after completion of chemotherapy. Pathologic response was determined at the time of surgery by microscopic examination of the excised tumor and lymph nodes. Pathologic complete response was defined as no residual invasive cancer in the breast or lymph nodes. Pathologic complete response was observed in 19 of 103 evaluable patients (18%). Tumor grade was assessed by modified Black's nuclear grading. Estrogen receptor (ER) and progesterone receptor (PR) expression status were assessed by immunohistochemistry (Novocastra Laboratories Ltd.; cutoff for positivity: 10% positive tumor cells) and HER-2 status was assessed by either fluorescence *in situ* hybridization (FISH, PathVision kit, Vysis) or immunohistochemistry as part of routine clinical care. HER-2 positivity (HER-2+) was defined as either *HER-2* gene amplification on FISH analysis (>2.0 gene copy number ratio of HER-2 and centromere 17 probes) or 3+ signal on immunohistochemistry evaluation. This study was approved by the institutional review boards of the M. D. Anderson Cancer Center, and all patients signed an informed consent for voluntary participation.

DNA extraction and CGH array profiling. RNA was extracted first from the samples using the Qiagen RNEasy Mini Kit followed by DNA extraction from the flow-through phase of the RNA extraction procedure. The DNA extraction was done with the Qiagen DNA extraction kit. DNA concentration and purity were determined by using the NanoDrop 100ND-1000 Spectrometer.

DNA was hybridized to 4 × 44K whole-genome Agilent arrays (G4426A) at the Genomic Unit of the Institute Gustave Roussy, Villejuif, France. For each sample, 500 ng of DNA were fragmented by a double enzymatic digestion (*Alu*I + *Rsa*I) and checked with LabOnChip (2100 Bioanalyzer System, Agilent Technologies) before labeling and hybridization. Tumor DNA and control DNA from Promega (Human Genomic DNA Female G1521) were labeled by random priming with CY5-dCTPs and CY3-dCTP, respectively, and hybridized at 65°C for 17 h. The chips were scanned on an Agilent G2565BA DNA Microarray Scanner and image analysis was done using the Feature-Extraction V9.1.3 software (Agilent Technologies). Feature-Extraction was used for the fluorescence signal acquisition from the scans. Normalization was done using the ranking-mode method available in the Feature-Extraction V9.1.3 software, with default value for any parameter. Raw copy number ratio data were transferred to the CGH Analytics v3.4.40 software for further analysis. Raw data have been submitted to the Array Express database[10] with the accession number E-TABM-584.

Gene expression analysis. Gene expression profiling was done with Affymetrix U133A GeneChips following standard operating procedures as described previously (12). Gene expression data was normalized using dChip V1.3 software[11] to a single reference array.[12] Complete transcriptional profile data were available on 103 cases; the mRNA was not sufficient in quality or quantity for transcriptional analysis in three cases.

Data analysis. The ADM-2 algorithm of CGH Analytics v3.4.40 software (Agilent) was used to identify DNA copy number anomalies at the probe level (13). A low-level copy number gain was defined as a log 2 ratio >0.25 and a copy number loss was defined as a log 2 ratio <-0.25. A high-level gain or amplification was defined as a log 2 ratio >1.5. Minimum common regions (MCR) were defined as chromosome regions that show maximal overlapping aberrations across multiple samples and were defined using STAC v1.2 (14). Probe-level measurement MCRs do not include all genes that are altered within a given aberrant region in a particular tumor but define the recurrent abnormalities that span the region (15). DNA copy number anomalies were plotted by the aCGH software package v1.10.0[13] using the R statistical language.[14] Probe-level aberrations were used as variables for consensus clustering of tumor samples by the nonnegative matrix factorization method (NMF; ref. 16). The NMF consensus package for GenePattern v3.0 was used to generate NMF-derived molecular clusters. We examined the robustness of the resulting clusters by calculating the cophenetic correlation coefficient of the cluster assignment (17). In addition, we used Euclidean distance and the Ward method (18) to perform an unsupervised hierarchical clustering on the 106 samples.

[10] http://www.ebi.ac.uk/arrayexpress
[11] http://dChip.org
[12] http://bioinformatics.mdanderson.org/pubdata.html
[13] http://bioconductor.org/packages/2.0/bioc/vignettes/aCGH/inst/doc/aCGH.pdf
[14] http://www.r-project.org/

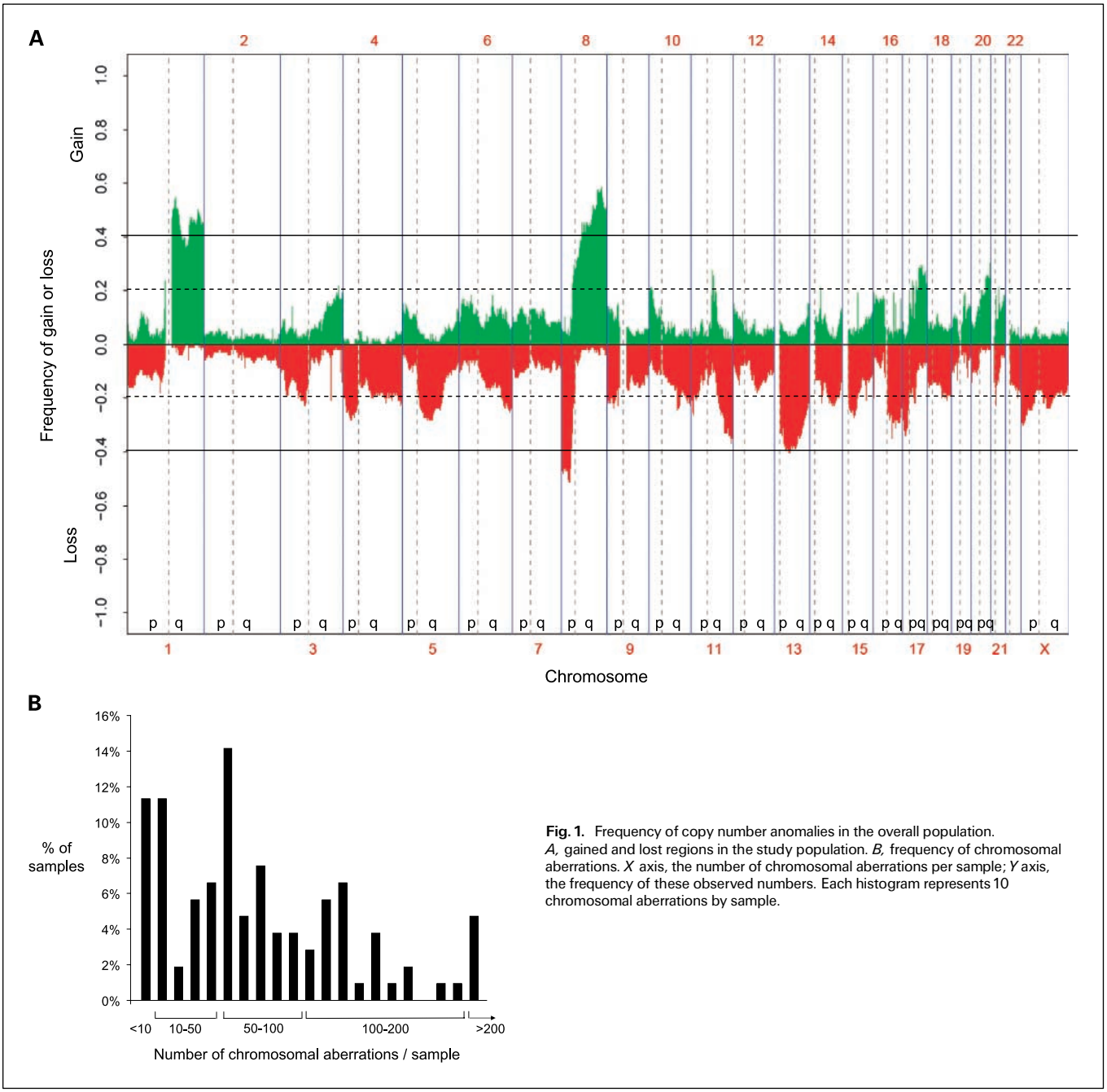

Fig. 1. Frequency of copy number anomalies in the overall population. *A,* gained and lost regions in the study population. *B,* frequency of chromosomal aberrations. *X* axis, the number of chromosomal aberrations per sample; *Y* axis, the frequency of these observed numbers. Each histogram represents 10 chromosomal aberrations by sample.

To correlate gene expression and copy number anomalies, a *t* test was performed. This test evaluated the distributions of the probe intensity of U133A chip according to the copy number anomalies of the matched probe on array CGH. ACE-it tool was used for doing such correlations (19).

We calculated the false discovery rate (FDR) using the procedure described by Benjamini and Hochberg (20). The *t* test was considered statistically significant when the FDR-adjusted *P* value for a *t* test was <0.05. All genomic positions are defined in this article according the University of California Santa Cruz Human version hg18 (March 2006).[15] The use of the single predictor (21) for five classes on the Affymetrix gene expression set has been made after mapping Agilent probes of the 306 gene set with Affymetrix HG-U133A probe sets of the corresponding gene. This mapping leads to a list of 276 probe sets (table not shown).

Results

Correlation between HER-2 FISH and HER-2 CGH array results. As a quality assessment measure, we examined if the CGH arrays correctly identified the known HER-2 amplification region in patients who were HER-2 – amplified by routine FISH analysis. Results are reported in Supplementary Fig. S1. HER-2 was amplified in 11 of the 68 samples for whom FISH results were available. In 8 of these 11 cases (73%), the log 2 (ratio) for the probe set [A14_P114826] corresponding to the *HER-2* gene

[15] http://genome.ucsc.edu

was >1.5, which has met our criteria for amplification based on CGH result. The sensitivity and specificity of this CGH cutoff were 0.73 and 1, respectively. The overall accuracy was 96%. Low-level gains (0.25 < log 2 ratio < 1.5) were apparent in two of the remaining three HER-2 FISH – amplified cases (log 2 ratio for the CGH probes: 0.69, 0.91). These observations suggest that 4 × 44K CGH array is a reliable technology to identify gene copy number anomalies on fine needle aspiration samples.

Chromosomal aberrations in the patient population. We examined the frequency of gained and lost DNA regions (Fig. 1A). The three most frequent gains (log 2 ratio >0.25) were observed in 8q at 116.7 and 127.5 Mb (58% of all cases) and in 1q at 153 Mb seen in 55% of cases. The two most frequently lost regions were 8p at 24.2 Mb and 13q at 47.8 Mb, seen in 51% and 41% of cases, respectively. The average number of chromosomal aberrations for a single tumor was 76 (range 1-318 in individual tumors; Fig. 1B). These results indicate that there is great individual variability in the number of chromosomal aberrations in breast cancer. Cancers with a low number of chromosomal aberrations ($n \leq 40$) included fewer high-grade tumors (38% versus 62%, $P = 0.03$) compared with tumors with high number (>40) of chromosomal aberrations.

Minimum common regions and highly amplified regions in the overall population. To identify genes with frequent abnormalities in multiple samples, MCRs were identified. MCR represents the maximal overlapping zone across samples within an abnormal (gained/lost) chromosomal region. Eleven MCRs were gained in ≥20% of cases and the three most common MCRs showed gain in 56% of all cases. Supplementary Table S2 lists these MCRs and the known genes that reside within these DNA segments. Several of these genes (*MYC*, *FOXA1*, *FGF3*, *FGF4*) were previously linked to breast cancer biology and others have functions that make them plausible candidates for therapeutic target discovery. Table 1 presents the 20 genes that showed amplification (log 2 ratio >1.5) in at least three or more consecutive probe sets in at least 10 samples. These genes were located in two amplicons (8p11-12 and 17q11-21). This included two trans-membrane targetable tyrosine kinases, i.e., ERBB2 and FGFR1. Other less frequently gained drug targets included PAK1 ($n = 4$), MYC ($n = 3$), EGFR ($n = 3$), IGFR1 ($n = 2$), and TERT ($n = 2$).

Next, we examined genes that were frequently lost. Thirty-seven MCRs were lost in >20% of samples (Supplementary Table S3). The single most frequently lost MCR was located in 11q and observed in 33% of cases. Several of the frequently lost genes are of interest for their known mechanism of action and include *BRCA1*, *STAT3*, *STAT5A*, *STAT5B*, and *MAPT* genes. An intriguing observation was the frequent (17%) loss of an MCR on 17q12 that encodes for eight chemokines (*CCL3*, *CCL4*, *CCL5*, *CCL14*, *CCL15*, *CCL16*, *CCL18*, *CCL23*) that are involved in the homing of immune cells to the tumor site.

Correlation between mRNA expressions results and DNA copy number changes. We identified 3,883 Affymetrix U133A probes (3,007 genes) whose expression correlated with DNA copy number alteration (adjusted $P < 0.05$; Supplementary Table S3). This indicates that copy number alterations lead to changes in gene expression levels that may have functional consequences. Of the 20 genes with the most frequent high-level gain shown on Table 1, 15 also showed significant mRNA overexpression compared with the cohort of cancers who had no amplification at this DNA region. Gene expression level according to the presence of gene amplification is reported in Fig. 2 for the 20 frequently amplified genes.

These data suggest that most DNA copy number alteration detected by CGH arrays also result in alterations in transcript levels.

Molecular classification of breast cancer based on DNA copy number changes. Clustering techniques are a commonly used

Table 1. Genes with high-level gains (i.e., amplification) in at least 10 cases

Gene	Locus	No. samples with gene amplification	Probe*	mRNA level (mean log 2 intensity)		*P*
				Amplified	Nonamplified	
PROSC	8p11.2	12	216519_s_at	7.4	7.4	0.87
GPR124	8p12	12	211814_at	7.8	7.7	0.45
ADRB3	8p12	12	206812_at	7.7	7.7	0.82
RAB11FIP1	8p11.22	12	219681_s_at	9.9	7.7	1.5E-5
BRF2	8p12	12	218955_at	9.4	8.5	4.9E-4
ASH2L	8p11.2	11	209517_s_at	10.3	8.8	7.7E-5
ERLIN2	8p11.2	11	221543_s_at	10.2	7.6	1.5E-6
WHSC1L1	8p11.2	11	221248_s_at	8.0	7.0	3.9E-4
STAR	8p11.2	11	204548_at	6.8	6.7	0.25
EIF4EBP1	8p12	11	221539_at	10.4	8.6	2E-6
BAG4	8p12	11	219624_at	7.0	6.0	5.6E-4
LSM1	8p11.2	11	203534_at	11.6	9.3	1.4E-6
DDHD2	8p12	11	212690_at	9.8	7.3	1.4E-6
FGFR1	8p11.2	10	211535_s_at	9.8	7.8	1.06E-4
ERBB2	17q21.1	10	216836_s_at	12.4	8.9	6.4E-8
TCAP	17q12	10	205766_at	7.5	7.2	0.073
PERLD1	17q12	10	55616_at	10.7	8.4	4.4E-6
PNMT	17q21	10	206593_at	9.2	7.8	0.01
STARD3	17q11	10	202991_at	9.4	8.1	2.7E-4
GRB7	17q21.1	10	210761_s_at	10.2	8.3	1.6E-6

*Probe on Affymetrix U133A 2.0.

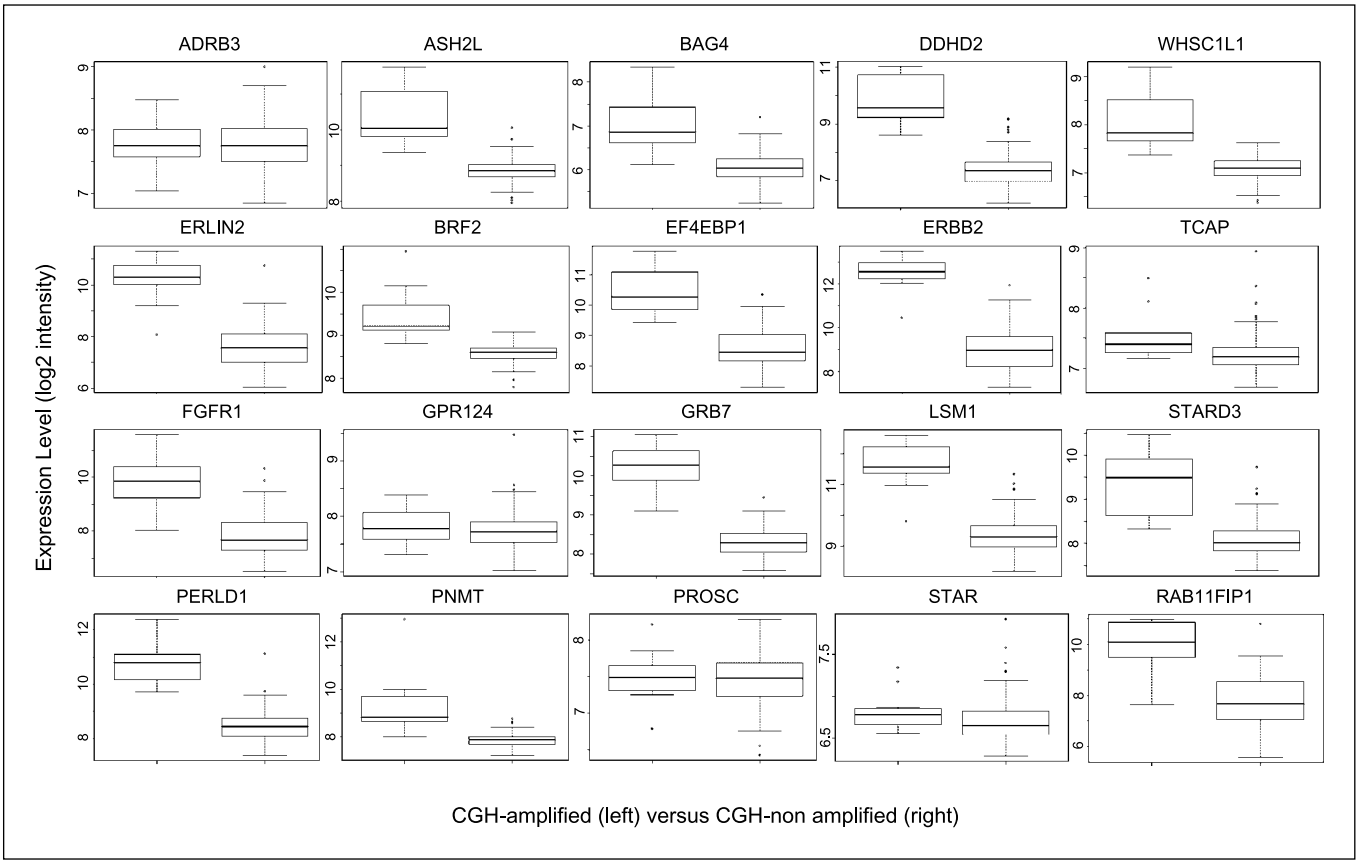

Fig. 2. Gene expression levels according to amplification in the 20 most frequently amplified genes. The mean expression levels are reported for the populations with amplified (*left*) and nonamplified (*right*) gene. The 20 genes that present amplification in 10 samples or more (Table 3) are reported. *Y* axis, the log 2 intensity of the matched probe detected on U133A 2.0. Mean values and *P* values are reported in Table 1.

method to discover and display large-scale molecular similarities across samples. Application of this method to transcriptional profiling data has led to the discovery of novel molecular classes of breast cancer (22). We applied the NMF clustering method to the CGH data. To assess how many clusters are robust, we calculated cophenetic correlation coefficients. The results indicated three CGH clusters when all probes were used for NMF clustering (Fig. 3A). The mean numbers of chromosomal aberrations were 81, 95, and 48 for class I, II, and III, suggesting that NMF-based clustering was not directly driven by the number of aberrations. The heat maps are reported in Fig. 3B. The three clusters correlated to some extent with known clinical variables (Table 2). NMF class I cluster included mostly triple-negative tumors (64%). NMF class II included the most of the HER2-overexpressing tumors (29%) and NMF class III was characterized by higher rate of ER+ (73%) and HER-2–negative tumors (97%). As predicted based on these clinical associations, class III tumors had the lowest rate of pathologic complete response (3%) compared with class II (21%) and class I (30%) cancers. Chromosomal aberrations that distinguish the three different CGH classes of breast cancer are presented in Supplementary Table S5. Four of the 12 regions that were gained in NMF class I were located between 6p21 and 6p23, and 10 of 15 regions that were lost in this class were located between 15q14 and 15q22. In class II, 11 of the 16 regions that were gained were located within 17q11 and 17q24, which correspond to the amplicons that include HER-2. High-level copy number gain (amplification) was present in 29% of the class II cases, and an additional 23% had low-level gains of this gene. Class III was characterized in gains in 1q 22-31 and frequent losses in 16q. We then determined the differentially expressed genes between the three CGH classes; 330 probe sets were different between the clusters (adjusted P value <0.001; Supplementary Table S6). In addition to NMF clustering, we performed an unsupervised hierarchical clustering. Results are reported in Fig. 4. Unsupervised hierarchical clustering suggested three different clusters: A, B, and C, respectively. The mean numbers of chromosomal aberrations were 65 (range 8-282), 123 (range 47-318), and 11 (range 1-86) in clusters A, B, and C, respectively ($P < 0.01$, t test). This suggested that the number of chromosomal aberrations was the prominent determinant of hierarchical clustering. Accordingly, cluster B was characterized by a higher rate of grade 3 (72% for B versus 40% and 44% for A and C, $P = 0.02$, χ^2 test) and triple-negative tumors (51% for B versus 21 and 33% for A and C, $P = 0.002$, χ^2 test).

Next, we compared gene expression and DNA copy number–based molecular classifications. The molecular classifier reported by Hu et al. (21) was used to assign molecular class to each tumor based on gene expression profiles. Using this classifier, 14 (16%), 32 (37%), 26 (30%), and 15 (17%) samples were classified as luminal B, luminal A, basal-like, and HER2, respectively. As shown in Fig. 3B, basal-like tumors were more frequently NMF class I (77%). Fifty-three percent of

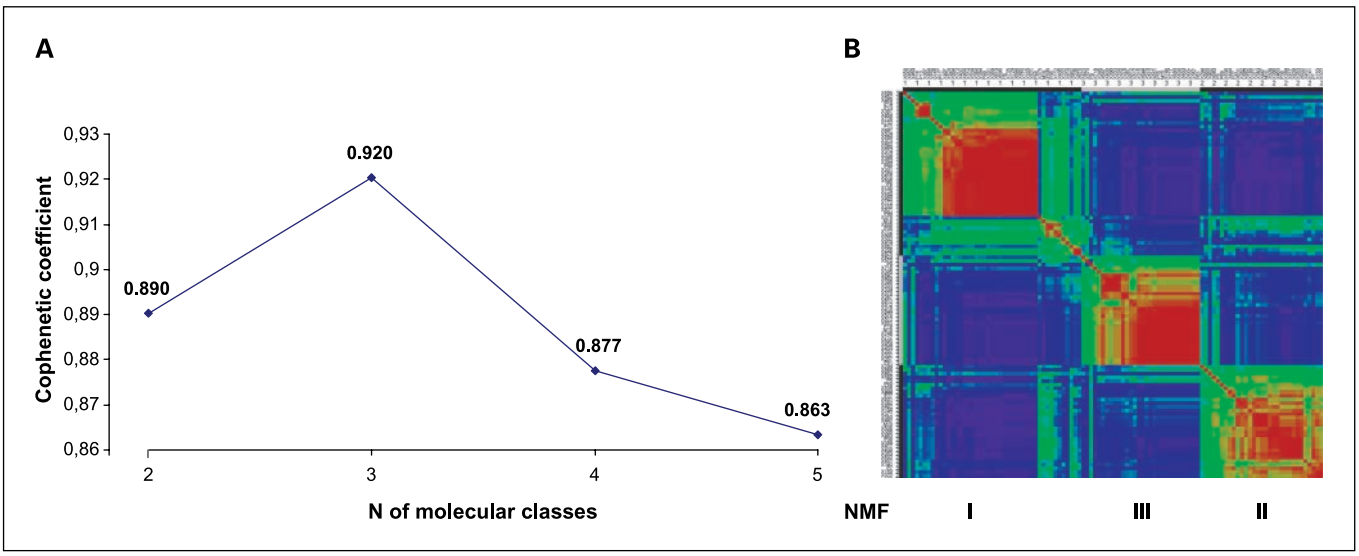

Fig. 3. Unsupervised classification by NMF clustering. *A,* cophenetic correlation according to number of classes. The cophenetic correlation was used to define the optimal number of clusters. *Y* axis, the cophenetic correlation; *X* axis, the number of classes. *B,* optimal NMF clustering and clinical characteristics. Heat map, the level of concordancy among two samples. Each pixel represents the level of concordancy between two CGH array profiles. Red, a high level of concordancy; blue, a low level of concordancy.

luminal A cancers were in NMF class III and 67% of HER2 tumors were in NMF-class II. When gene expression–based classification was compared with the CGH classes obtained by hierarchical clustering, 57%, 62%, and 60% of luminal B, luminal A, and HER2 tumors were classified as cluster A. Fifty percent (n = 13 of 26) of the basal-like breast cancer were classified as cluster B. These results show partial overlap between the gene expression– and CGH-based molecular classes.

DNA copy number alterations that are specific for triple-negative breast cancer. Estrogen, progesterone, and HER-2 receptor–negative ("triple-negative") breast cancers represent a distinct clinical entity. Identification of new therapeutic targets for this group of breast cancers is important because current treatment options are limited to chemotherapy and bevacizumab. Because NMF-cluster I included most of the triple-negative breast cancers, we hypothesized that specific DNA alterations might characterize this breast cancer subtype.

One region was significantly different between triple-negative (n = 35) and non–triple-negative (n = 69) tumors. This region is located in 6p21.2-6p12 (Supplementary Fig. S2) and was the predominant region to distinguish NMF class I to class II and III. This region is gained in 34% of triple-negative breast cancer and 10% of non–triple-negative tumors. To better define potential specific targets in this subset of tumors, we also analyzed probe level gains and losses. The list of genes with frequently altered copy number (FDR <0.05) between triple-negative and non–triple-negative tumors is reported in Table 3. The frequently gained 6p21-25 regions included more than 200 genes. Several were identified as potential major targets in this subset. Probes corresponding to the *E2F3* and *VEGFA* genes were gained in 29% and 34% of triple negative tumors, compared with 10% and 6% in non–triple-negative tumors. Other genes of potential interest were located in this region, including *DEK, NOTCH4*, and a cluster of heat shock proteins (HSPA1A, HSPA1L, HSP90AB1).

In addition, in the FDR range between 5% and 10%, we observed an enrichment of EGFR gains (29%) and PTEN losses (34%) in triple-negative breast cancer compared with non–triple-negative tumors. As reported in Supplementary Table S4,

Table 2. Clinical characteristics according to the classes

	NMF class			*P*
	I	III	II	
n	45	30	31	
Grade 3	24 (59%)	11 (41%)	12 (46%)	0.31
ER+	14 (31%)	22 (73%)	24 (77%)	<0.001
Her2+	1 (2%)	1 (3%)	9 (29%)	<0.001
Triple negative	28 (64%)	5 (17%)	2 (6%)	<0.001
Molecular class (gene expression array)				
Luminal B (n = 14)	4 (29%)	1 (7%)	9 (64%)	0.01
Luminal A (n = 32)	9 (28%)	17 (53%)	6 (19%)	<0.001
Basal-like (n = 26)	20 (77%)	4 (15%)	2 (8%)	<0.001
Her2 (n = 15)	3 (20%)	2 (13%)	10 (67%)	0.004

NOTE: Molecular subtypes were assigned to each sample according to a classifier derived from Hu et al. (19).

the level of mRNA expression significantly correlated to the copy number anomalies for E2F3, VEGFA, EGFR, and PTEN.

DNA copy number alterations that distinguish highly chemotherapy sensitive tumors from the rest. We also examined if we could find DNA copy number changes that are significantly associated with pathologic complete response to chemotherapy, indicating an extremely chemotherapy-sensitive subset of tumors. First, we compared DNA copy number changes

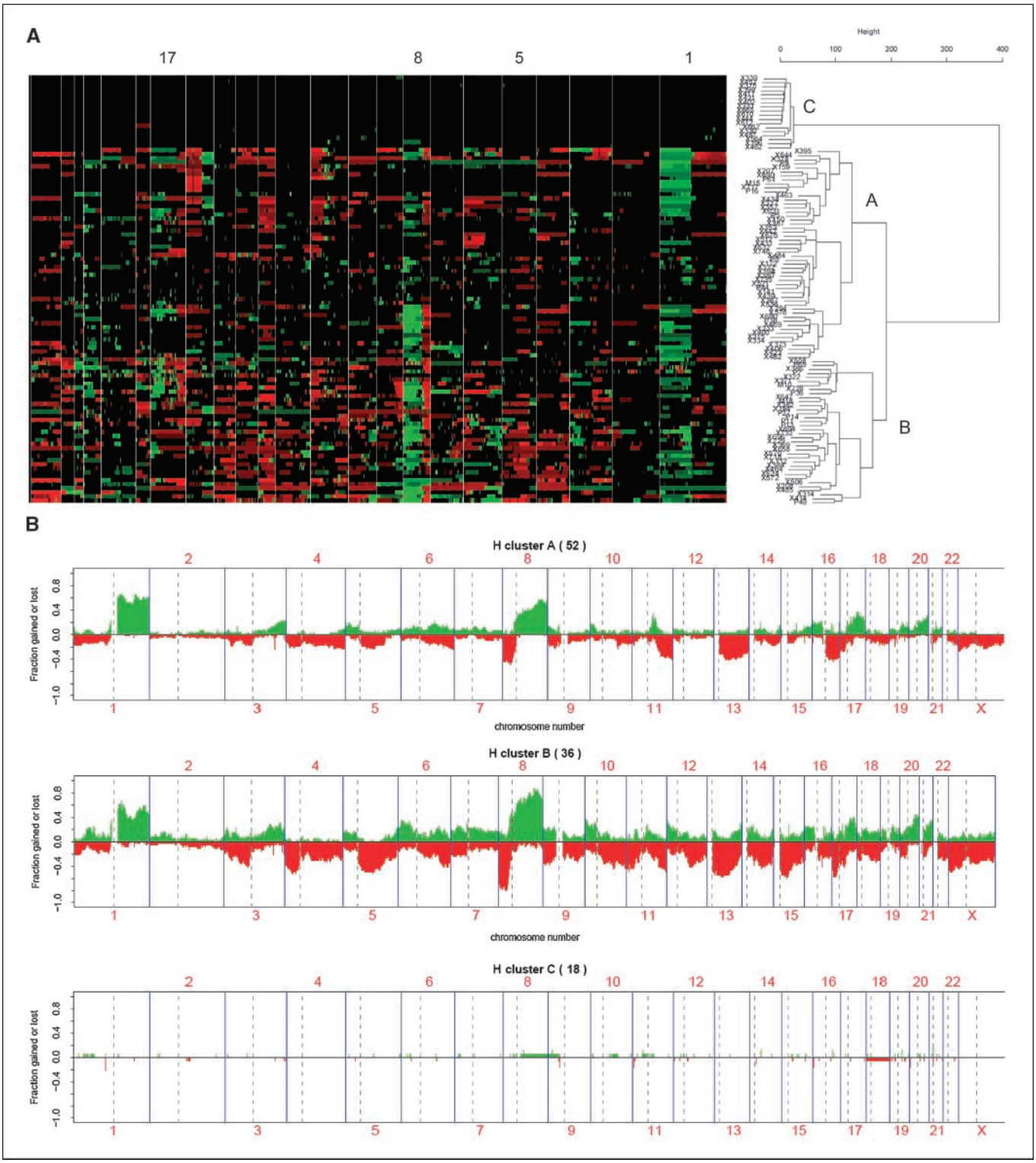

Fig. 4. Unsupervised hierarchical clustering. *A,* heat map and dendrogram of unsupervised hierarchical clustering of the 106 samples. Euclidean distance and the Ward method were used to generate clustering. *B,* the frequency of DNA copy number anomalies according to hierarchical clusters.

Table 3. Chromosomal regions harboring a different DNA copy number in triple-negative tumors compared with other tumor subtypes

Location	Triple negative (n = 35)	Triple negative tumors (n = 69)	P	FDR	Genes
	Frequency gain (%)/ lost (%)	Frequency gain/lost			
1q32.1	29/0	54/0	7E-4	4.9E-2	*SYT2*
5q11.2	3/37	4/10	5E-4	4.6E-2	*ITGA1, PELO*
5q13.2-5q14.1	0/43	3/20	5E-4	4.6E-2	*CRHBP, F2R, F2RL1, F2RL2, HEXB, HMGCR, ENC1, COL4A3BP, TINP1, IQGAP2, SV2C, KIAA0888, GCNT4, POLK, AGGF1, RGNEF, ZBED3, GFM2, C5orf37, S100Z, ANKRD31, LOC441086, LOC441087*
5q14.2-5q14.3	0/43	3/20	8E-4	4.9E-2	*HAPLN1, VCAN, XRCC4, EDIL3, TMEM167, NBPF22P*
5q21.1	0/46	3/17	2E-4	4.2E-2	*CHD1, ST8SIA4, RGMB, TMEM157*
5q21.1-5q21.2	0/43	1/17	3E-4	4.2E-2	*PAM, HISPPD1, GIN1, NUDT12, C5orf30, SLCO6A1, LOC285708, SLCO4C1*
6p25.1	20/6	4/10	8E-4	4.9E-2	*LY86, RP3-398D13.1*
6p24.1	23/0	9/9	8E-4	4.9E-2	—
6p22.3-5p22.2	29/0	10/9	1E-4	2.2E-2	*E2F3, ID4, PRL, SOX4, TPMT, DEK, CDKAL1, MBOAT1, HDGFL1, AOF1, RNF144B, NHLRC1, FLJ22536, LOC729177*
6p22.2	26/0	6/9	4E-4	4.2E-2	*SLC17A1*
6p21.33	26/0	6/7	7E-4	4.9E-2	*HCG9*
6p21.33	26/0	6/7	4E-4	4.2E-2	*ABCF1, GNL1, HLA-E, PPP1R10, DHX16, MDC1, NRM, MRPS18B, TRIM39, RPP21, C6orf134, PRR3, KIAA1949, TUBB, C6orf136*
6p21.33	26/0	4/7	4E-4	4.2E-2	*ATP6V1G2, CDSN, HLA-B, HLA-C, LTA, LTB, MICB, NFKBIL1, POU5F1, TCF19, TNF, BAT2, BAT1, LST1, HCP5, C6orf15, CCHCR1, PSORS1C1, PSORS1C2, HCG27, NCR3, MCCD1*
6p21.33-6p21.32	26/0	6/9	4E-4	4.2E-2	*CLIC1, CSNK2B, HSPA1A, HSPA1L, MSH5, NEU1, VARS, BAT4, BAT5, DDAH2, C6orf48, APOM, LSM2, C6orf47, LY6G5B, LY6G6D, C6orf27, C6orf25, C6orf21, C6orf26*
6p21.32	26/0	6/7	6E-4	4.6E-2	*AGER, CFB, C2, C4A, C4B, CREBL1, CYP21A2, DOM3Z, RNF5, SKIV2L, TNXB, RDBP, STK19, PPT2, AGPAT1, EHMT2, FKBPL, PRRT1, EGFL8, ZBTB12*
6p21.32	26/0	6/7	6E-4	4.6E-2	*HLA-DRA, NOTCH4, PBX2, C6orf10, BTNL2, GPSM3*
6p21.31	29/0	7/7	2E-4	4.2E-2	*KCTD20*
6p21.2	31/0	7/9	4E-4	4.2E-2	*RNF8, KIAA0082, TBC1D22B, ZFAND3, BTBD9, C6orf129, MDGA1, FLJ45825*
6p21.2	29/0	6/9	1E-4	2.2E-2	*DNAH8, GLO1, GLP1R, KCNK5, C6orf64, BTBD9*
6p21.2-6p12.3	34/3	6/10	1E-4	2.2E-2	*(105), BYSL, RUNX2, CCND3, CDC5L, SLC29A1, HSP90AB1, MEP1A, NFKBIE, NFYA, PGC, POLH, PPP2R5D, PTK7, PRPH2, SRF, TBCC, VEGFA, TFEB, SUPT3H, NCR2, RCAN2, GNMT, TNFRSF21, PRICKLE4, XPO5, TTBK1, ABCC10, FOXP4, PTCRA, HCRP1*
9p13.3	31/3	6/16	1E-4	2.2E-2	*C9orf23*
11q14.3	11/6	1/33	1E-4	2.2E-2	*CHORDC1*
11q22.3	6/9	0/38	4E-4	4.2E-2	*GRIA4, ICEBERG*
11q22.3	6/9	0/39	3E-4	4.2E-2	*GRIA4, GUCY1A2, AASDHPPT, KIAA1826, KBTBD3*
11q23.3	14/11	0/38	2E-4	4.2E-2	*C1QTNF5*
15q13.3-15q14	0/40	6/17	6E-4	4.6E-2	*CHRM5, CHRNA7, RYR3, SCG5, ARHGAP11A, SLC12A6, DKFZP434L187, GREM1, TMEM85, C15orf24, AVEN, C15orf29, FAM7A3, FAM7A1, PGBD4, LOC283701, FMN1, LOC440268, C15orf45*

(*Continued on the following page*)

Table 3. Chromosomal regions harboring a different DNA copy number in triple-negative tumors compared with other tumor subtypes (Cont'd)

Location	Triple negative (n = 35)	Triple negative tumors (n = 69)	P	FDR	Genes
	Frequency gain (%)/ lost (%)	Frequency gain/lost			
15q14	0/37	6/16	8E-4	4.9E-2	*AGPAT7*
15q14	0/37	6/16	8E-4	4.9E-2	*GOLGA8A, GJD2, GOLGA8B*
15q14	0/37	6/16	8E64	4.9E-2	*AQR, ZNF770, ATPBD4, hCG_1787519, LOC723972*
15q14	0/40	3/16	6E-4	4.6E-2	*C15orf41*
15q22.2	0/29	9/9	8E-4	4.9E-2	*TLN2, MGC15885*
15q22.2-15q22.31	0/29	10/9	6E-4	4.6E-2	*CA12, PPIB, SNX1, TPM1, HERC1, USP3, DAPK2, RPS27L, RAB8B, CSNK1G1, SNX22, APH1B, FAM96A, LACTB, LOC145842, FBXL22*
16q23.2	9/9	4/39	7E-4	4.9E-2	—
16q23.3	9/6	4/38	3E-4	4.2E-2	*OSGIN1, EFCBP2, LOC146167*
16q24.1	6/3	4/36	3E-4	4.2E-2	*TMEM148, LOC727710*
19q13.12	31/0	6/3	7E-4	4.9E-2	*ZFP30, ZNF571, ZNF607, ZNF573, ZNF781, ZNF540, ZNF793, DKFZp761D1918*
22q11.21	17/9	0/17	7E-4	4.9E-2	*HIRA, UFD1L, CDC45L, MRPL40, LOC128977*
22q11.21	20/9	0/17	3E-4	4.2E-2	*ARVCF, COMT, GP1BB, SEPT5, RANBP1, TBX1, TXNRD2, HTF9C, ZDHHC8, DGCR8, GNB1L, RTN4R, C22orf29, DGCR6L, KIAA1652, C22orf25, FLJ32575, LOC284865, LOC440792*

between tumors that achieved pathologic complete response (n = 20) and those that did not (n = 83) regardless of the type of treatment that they received. Supplementary Fig. S3 shows the overall genome-wide gains and losses for the two different response groups. Nine regions showed significant difference (P < 0.005) in copy number alterations; however, the corresponding false discovery rates were 82%; therefore, our confidence in this observation is limited.

It is possible that different molecular alterations determine extreme sensitivity to different therapies; therefore, we also examined the FAC alone (n = 35) and FAC-paclitaxel regimens separately (n = 64). Supplementary Fig. S4 shows the overall genome-wide gains and losses for the two different pathologic response groups in the taxane-treated patients. Fifteen DNA regions showed significant difference (P < 0.005) in copy number alterations between highly T/FAC sensitive (i.e., cases with pathologic complete response) and less sensitive tumors (Supplementary Table S7). Many of these associations may be spurious because the FDR corresponding to this P value cutoff was equal to 59%; however, several of the observed genomic abnormalities seem biologically plausible. For example, 11q22 was gained in 19% of sensitive tumors and lost in 29% of refractory tumors and encodes for a series of caspases that play an important role in executing apoptosis (CASP1, CASP4, CASP5, CASP12; ref. 23). Regions in 3p25 and 3p27 were gained in 47% of highly taxane-sensitive tumors compared with 6% to 9% of tumors with residual cancer and these regions code for components of the Rho/Ras signaling pathway (RAP2B, SSR3, GPR171, GPR87, GPR149) that has been suggested to mediate paclitaxel-induced apoptosis (24). These observations will require further mechanistic validation in the laboratory.

Discussion

In the present study, we examined DNA copy number alterations in 106 invasive cancers using 4 × 44K Agilent arrays. Our CGH analysis correctly identified clinical HER-2 amplification status with 96% to 99% accuracy, depending on the cutoff value used to determine HER-2 status by our CGH data. In addition, we report that 3,007 genes presented a significant correlation between copy number and mRNA levels. Interestingly, a high number of molecular targets (HER-2, EGFR, PTEN, VEGFA, AURKA, CHEK, AKT...) currently under investigation presented such correlation. Altogether, these two data suggest that (a) the occurrence of a DNA gain and/or lost leads to an unregulated overexpression or underexpression of mRNA and (b) high-resolution CGH array is a reliable technology to assess such copy number anomalies. CGH array, as complement to other technologies, could be an interesting approach in clinical practice to identify some subsets of patients who are candidates to targeted therapies.

We could correlate gene expression and copy number anomalies in 3,007 genes. This finding suggests that gene copy number variation is a crucial mechanism to drive gene expression in breast cancer. We consider each of these genes highly attractive for further functional characterization and it is likely that several of these represent promising future drug targets. For example, the region between 8q24.13 and 8q24.22 was gained in 58% of samples and it codes for c-MYC, an important oncogene, among many other genes (25, 26). Another region between 8p11 and 8p12 showed high level of amplification in around 10% of samples and it codes for potential therapeutic targets, including *FGFR1* and several other genes (*LSM1, BAG4, GPR124, STARD3, WHSC1L1, EIF4EBP1*)

that are involved in oncogenesis (27–29). This region has previously been reported to be amplified in 10% to 15% of breast cancer by others as well (27). Interestingly, *FGFR1* was overexpressed in tumors that harbor *FGFR1* gene amplification. These data suggest that there is a strong rationale to further evaluate FGFR1 inhibitors (27) in this subset of breast adenocarcinoma.

Unsupervised clustering of breast cancer based on gene expression profile has allowed some advances in the molecular classification of breast cancer. Nevertheless, there is a need to improve such classification using other molecular technologies including gene sequencing, CGH arrays, and proteomics. These other approaches could lead to classify tumors based on oncogenic molecular events, in addition to define them based on stem cell origin (basal- or luminal-like tumors). Unsupervised clustering of DNA copy number changes using NMF or hierarchical clustering yielded three molecular groups each. These groups differed in their clinical features and in the presumed molecular pathways that were affected by the DNA copy number alterations. Interestingly, both NMF-based classification and hierarchical clustering partially overlap with the ER, PR, and HER2 classification. Several other studies have identified clusters of copy number anomalies. Fridlyand et al. (4) have reported three NMF classes. The first class was enriched in 1q gain/16 q losses; the second class was enriched in complex copy number anomalies and included mainly basal-like breast cancer; and the third class included 8q gains. Farabegoli et al. (30) also reported a cluster of 1q gain/16q loss that was enriched in ER-positive/low-grade tumors. The copy number anomaly–based classifications generated in the present study were consistent with these previous findings. 1q gains and 16 q losses were indeed observed with a high frequency in the NMF group III, a group enriched in luminal A breast cancer. In addition, the molecular group that contains complex copy number anomaly alterations (class A) was observed more frequently in basal-like breast cancer. Because the present study was done based on high-resolution CGH arrays, several additional copy number anomalies were identified that drive a CGH array–based classification. As an illustration, 6p21 gains were predominantly found in the NMF group I. Altogether, these data suggested that the CGH array could identify subsets of patients with different outcomes and biological behavior and could contribute to a future molecular taxonomy of breast cancer. Further larger studies will aim at determining whether such CGH array–based molecular classification could add to the current tools for breast cancer classification.

The most striking finding was the enrichment of triple-negative tumors in the NMF cluster I, although NMF clustering was not based on expression data. This led us to hypothesize that this specific tumor subset could present some specific copy number changes that could be potentially targeted in this population, but not in other. We identify at least three genes (*VEGF*, *EGFR*, *PTEN*) that harbor a high frequency of copy number anomalies, correlated with gene expression, and could be targeted in triple-negative breast cancer in clinical trials. The most striking finding was the observation that a probe matching *VEGFA* gene is gained in 34% of triple-negative tumors. Finally, although there is no drug that targets this protein, *E2F3* (31) is a candidate gene to play a major role in a subset of triple-negative breast cancer.

One of the goals of this study was to correlate DNA copy number changes and response to therapy. We could not identify DNA copy number anomalies that are strongly associated with chemotherapy efficacy. We have previously reported on gene expression–based predictors of response to therapy using a subset of these specimens (12). We noted that because of the close association between chemotherapy sensitivity and clinical phenotype including ER status and tumor grade, it is relatively easy to define a robust gene expression pattern that is associated with chemotherapy response. However, this gene signature is driven by the large-scale gene expression differences that distinguish ER-negative from ER-positive tumors and low-grade from high-grade cancers (32). The current data set is too small to adjust for these phenotypic confounders and therefore separate analysis of the gene expression data to define chemotherapy response predictors was not done. However, a phenotype-adjusted analysis to define chemotherapy response predictors was done on a larger data set that include a subset of these samples (n = 233, HER-2 normal breast cancers) and preliminary results were presented (33).

In summary, we report a large number of previously unrecognized DNA copy number changes in breast cancer that alter the transcriptional activity of genes, including many that has not previously been linked to breast cancer biology. The lists of these genes represent a potentially fertile ground for further mechanistic research and could include several future therapeutic targets. In addition, we confirm that *FGFR1* and other genes located in the 8p11-12 amplicon are potentially major events in a subset of breast cancer. Finally, we propose three molecular classes based on copy number anomalies, whose complementarity with the DNA array–based classification will be determined in the future, and we identified some targetable molecular alteration that are enriched in triple-negative tumors, opening some therapeutic perspectives in this subset of cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Davies JJ, Wilson IM, Lam WL. Array CGH technologies and their applications to cancer genomes. Chromosome Res 2005;13:237–48.
2. Van Beers EH, Nederlof PM. Array-CGH and breast cancer. Breast Cancer Res 2006;8:210.
3. Bergamaschi A, Kim YH, Wang P, et al. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer 2006;45:1033–40.
4. Fridlyand J, Snijders AM, Ylstra B, et al. Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer 2006;6:96.
5. Pollack JR, Sorlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A 2002;99:12963–8.
6. Climent J, Dimitrow P, Fridlyand J, et al. Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. Cancer Res 2007;67:818–26.
7. Cingoz S, Altungoz O, Canda T, Saydam S, Aksakoglu G, Sakizli M. DNA copy number changes detected by comparative genomic hybridization and their association with clinicopathologic parameters in breast tumors. Cancer Genet Cytogenet 2003;145:108–14.
8. Chin K, DeVries S, Fridlyand J, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 2006;10:529–41.

9. Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! Or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res 2006;34:445–50.

10. Symmans WF, Ayers M, Clark EA, et al. Total RNA yield and microarray gene expression profiles from fine needle aspiration and core needle biopsy samples of breast cancer. Cancer 2003;97:2960–71.

11. Rouzier R, Perou CM, Symmans WF, et al. Different molecular subtypes of breast cancer respond differently to preoperative chemotherapy. Clin Cancer Res 2005;11:5678–85.

12. Hess KR, Anderson K, Symmans W, et al. Pharmacogenomic predictor of sensitivity to preoperative paclitaxel and 5-fluorouracil, doxorubicin, cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 2006;24:4236–44.

13. Lipson D, Ben-Dor A, Dehan E, Yakhini Z. Joint Analysis of DNA Copy Numbers and Gene Expression Levels. Algorithms Bioinformatics 2004;3240:135–46.

14. Diskin SJ, Eck T, Greshock J, et al. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Res 2006;16:1149–58.

15. Rouveirol C, Stransky N, Hupe P, et al. Computation of recurrent minimal genomic alterations from array-CGH data. Bioinformatics 2006;22:849–56.

16. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004;101:4164–9.

17. Sneath P, Sokal R. Numerical taxonomy. Nature 1962;193:855–60.

18. Hartigan JA. Clustering. Annu Rev Biophys Bioeng 1973;2:81–101.

19. Van Wieringen WN, Belien JAM, Vosse SJ, Achame EM, Ylstra B. ACE-it: A tool for genome-wide integration of gene dosage and RNA expression data. Bioinformatics 2006;22:1919–20.

20. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.

21. Hu Z, Fan C, Oh DS, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 2006;7:96.

22. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.

23. Tamm I, Schriever F, Dorken B. Apoptosis: implications of basic research for clinical oncology. Lancet Oncol 2001;2:33–42.

24. Liu Ax, Cerniglia GJ, Bernhard EJ, Prendergast GC. RhoB is required to mediate apoptosis in neoplastically transformed cells after DNA damage. Proc Natl Acad Sci U S A 2001;98:6192–7.

25. Nass SJ, Dickson RB. Defining a role for c-Myc in breast tumorigenesis. Breast Cancer Res Treat 1997;44:1–22.

26. Corzo C, Corominas JM, Tusquets I, et al. The MYC oncogene in breast cancer progression: from benign epithelium to invasive carcinoma. Cancer Genet Cytogenet 2006;165:151–6.

27. Reis-Filho JS, Simpson PT, Turner NC, et al. FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas. Clin Cancer Res 2006;12:6652–62.

28. Yang ZQ, Streicher KL, Ray ME, Abrams J, Ethier SP. Multiple interacting oncogenes on the 8p11–12 amplicon in human breast cancer. Cancer Res 2006;66:11632–43.

29. Bjarnadottir TK, Fredriksson R, Hoglund PJ, Gloriam DE, Lagerstrom MC, Schioth HB. The human and mouse repertoire of the adhesion family of G protein-coupled receptors. Genomics 2004;84:23–33.

30. Farabegoli F, Hermsen MA, Ceccarelli C, et al. Simultaneous chromosome 1q gain and 16q loss is associated with steroid receptor presence and low proliferation in breast carcinoma. Mod Pathol 2004;17:449–55.

31. Ginsberg D. E2F3—a novel repressor of the ARF/p53 pathway. Dev Cell 2004;6:742–3.

32. Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. Clin Cancer Res 2007;13:6080–6.

33. Hatzis C, Symmans WF, Lin F, et al. Genomic predictors of pathologic response to preoperative chemotherapy for triple negative and ER-positive/HER-2-negative breast cancers. ASCO Annual Meeting Proceedings. J Clin Oncol 2008;26:23.

Clinical Cancer Research

Molecular Characterization of Breast Cancer with High-Resolution Oligonucleotide Comparative Genomic Hybridization Array

Fabrice Andre, Bastien Job, Philippe Dessen, et al.

Clin Cancer Res 2009;15:441-451.