Prediction of high-risk patients by genome-wide copy number alterations from remaining cancer after neoadjuvant chemotherapy and surgery

SUN YOUNG RHA[1,2,3,4,5], HEI CHEUL JEUNG[1,2,3,5], MIN YOUNG SEO[1,2,3], SANG CHEOL KIM[1,2,3,4], WOO ICK YANG[1,4,6], YONG WHA MOON[1,2,5] and HYUN CHEOL CHUNG[1,2,3,4,5]

[1]Cancer Metastasis Research Center, [2]Yonsei Cancer Center, [3]Yonsei Cancer Research Institute, [4]Brain Korea 21 Project for Medical Sciences, Departments of [5]Internal Medicine and [6]Pathology, Yonsei University College of Medicine, Seoul, Korea

Abstract. In breast cancer, changes of gene copy number were analyzed by cDNA microarray-based comparative genome hybridization using post-treatment archived tissues. Genomic DNA was extracted from 45 surgical specimens after chemotherapy. Informative genes were selected by t-test and were statistically validated by prediction analysis using support vector machine in R package. Fluorescence *in situ* hybridization and quantitative PCR were performed for validation. We developed three clinical models: comparing good vs. poor prognosis (I), comparing good vs. poor prognosis among poor responders (II) and among good responders (III). 158 gene set (I) differentiated high and low risk of relapse group with 92% accuracy. 51 gene set (II) differentiated good and poor prognosis subgroups among poor responders with 99.9% accuracy. 32 gene set (III) differentiated good and poor prognosis subgroups among good responders with 96% accuracy. This approach has potential applications in the identification of high risk of recurrence after neoadjuvant chemotherapy and surgery.

Introduction

Several techniques have been proved to be useful for detection of genetic alterations. Karyotype analysis requires living cells and is difficult to interpret because of the low number of good metaphase spreads obtained (1). Conventional comparative genomic hybridization (CGH) is difficult to interpret because of repetitive sequences in genomes. The regions covered by bacterial artificial chromosome (BAC) array-CGH are too large and poorly defined to provide specific genetic information (2,3). To understand the complicated molecular characteristics of cancer more precisely, it is necessary to identify individual genes with altered copy number, which may result in changes in gene products (4,5). The cDNA microarray technique is an innovative method for integrated analyses of the genome by both expression profiling and copy number change (4-6). Thus, cDNA microarray-based comparative genomic hybridization (microarray-CGH) is capable of detecting and mapping genome-wide amplifications and deletions with high resolution power among current CGH technologies (4-6). UV exposure history predicted by microarray-CGH suggests that genomic DNA copy patterns can reveal cancer biology that is not detectable by standard approaches (7).

In locally advanced breast cancer (LABC), especially in stage III cancer, it is difficult to predict the clinical outcome after neoadjuvant chemotherapy and definite local control with surgery and radiotherapy by current clinical, pathological and biological factors (8,9). Pathological complete response (pCR) is currently considered to be the best, although imperfect, early marker for cure (8,10). If there is no pCR, even if maximal tumor shrinkage has been induced with neoadjuvant chemotherapy, the patient prognosis does not correlate with the grade of tumor shrinkage. Patients with extensive tumor shrinkage can relapse within 5 years after surgery and *vice versa*. Therefore, like osteosarcoma therapeutics strategy, additional treatment after surgery and radiotherapy may be required in high-risk patients. The point is whether we can use the same chemotherapy regimen after surgery as neoadjuvant regimen or we need to change the regimen based on the markers presented in the remaining cancer after neoadjuvant chemotherapy. Even well-studied biomarkers such as HER2, p53 and topoisomerase IIα have given inconsistent clinical results so far (11-13). Since the prognosis of breast cancer differs considerably based on genetic background (14-16), detailed knowledge of genetic changes on a genome-wide basis after neoadjuvant chemotherapy may facilitate the establishment of individualized treatment.

The initial aim of this study was to show that archived paraffin blocks from LABC that contained remaining tumor

---

Correspondence to: Professor Hyun Cheol Chung, Cancer Metastasis Research Center, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul 120-752, Korea
E-mail: unchung8@yuhs.ac

after neoadjuvant chemotherapy could be used for microarray-CGH. The hypothesis was that multi-gene aberrations detected in these samples could be formalized into a predictor of clinical outcome after surgery or a guideline in determining post-operative chemotherapy regimen.

Materials and methods

Patients and treatment scheme. Paraffin blocks from 45 LABC patients of clinical stage IIIA (T3N2) and IIIB who were treated at Yonsei Cancer Center, Yonsei University College of Medicine from June 1991 to June 2001 were used for this study.

LABC was defined as follows: tumor ≥5 cm with metastasis to the ipsilateral axillary nodes, tumor with direct extension to the chest wall or skin, or tumor of any size with fixed ipsilateral axillary nodes. Neoadjuvant chemotherapy with infusional 5-fluorouracil, adriamycin and cyclo-phosphamide (iFAC) was administered according to the following schedule: 1000 mg/m^2 5-FU continuous infusion on days 1 to 3, 40 mg/m^2 doxorubicin and 600 mg/m^2 cyclo-phosphamide bolus by i.v. injection on day 1. The treatment was repeated every 3 weeks. When maximum tumor response was reached, as determined by no change in tumor size for 2 consecutive treatment cycles, resectability was assessed by an oncologic surgeon. After surgery, adjuvant chemotherapy with iFAC was administered up to a maximum of 12 cycles including the neoadjuvant chemotherapy. Radiotherapy was performed 5.5 weeks after surgery with a dose of 50.4 Gy.

Clinical drug response was evaluated on the basis of residual tumor size determined by ultrasonography and mammography after treatment. A reduction in tumor size of >70% was arbitrarily defined as 'good responder' and a reduction of <70% as 'poor responder'. Tumors treated in this study were large and a regression of at least 70% after chemotherapy would represent a clinically important response. The goal of this study was to identify high-risk patients among those with tumor remaining after neoadjuvant chemotherapy. Two pCR cases were excluded from prognostic gene set selection after pre-operative chemotherapy because these patients had no residual tumors for genomic evaluation. Thus, to predict the prognosis of 43 cases with residual tumors in the breast, recurrence within 5 years after surgery was defined as 'poor prognosis' and a disease-free state after 5 years of surgery as 'good prognosis'.

Tissue samples and DNA preparation. Paraffin blocks were sectioned by 10 μm for genomic DNA extraction. Pathologist confirmed the tumor areas by hematoxylin and eosin (H&E) staining and tumor tissues were obtained after microdissection. Genomic DNA was prepared from frozen normal placenta (reference samples) and paraffin-embedded tumor tissue (test samples) by phenol/chloroform/isoamylalcohol extraction, as previously described (17).

cDNA microarray-based CGH (microarray-CGH). The 17,664 cDNA microarrays (CMRC-GenomicTree, Korea) used in this study included 15,723 unique genes. Genes were mapped for chromosomal location by SOURCE (http://genome-www5.stanford.edu/cgi-bin/source/source

Search) and DAVID (http://apps1.niaid.nih.gov/david/). Microarray-CGH was performed using 6 μg of extracted genomic DNA. Genomic DNA from normal placenta and tumor tissue was labeled with Cy3-dCTP and Cy5-dCTP respectively, and mixed with human Cot-1 DNA, poly (dA)-poly (dT) oligonucleotides and yeast tRNA. After boiling at 100°C for 90 sec and pre-annealing at 37°C for 30 min, the labeled DNA mixture was applied to a cDNA microarray. Microarray slides were scanned using a GenePix 4000B scanner (Axon Instruments, Foster City, CA). Fluorescence signals were calculated by subtracting the background intensity from the spot intensity by using GenePix Pro 4.1 software (Axon Instruments).

Analysis of microarray-CGH data
Raw data preprocessing and normalization. Genes with missing values in >20% of the experiments were removed and missing values were adjusted by kNN-based imputation. To correct for differences caused by sample labeling efficiencies and spatial effects of the slide, 'within-print-tip normalization' with median correction was applied to the log$_2$-transformed data. After data preprocessing, a total of 12,492 genes were used for data analysis.

Gene selection. Gene aberrations were determined based on a previous study comparing X chromosomal changes between female and male using normal placenta and lymphocytes as a reference (18). Genes with log$_2$ (Cy5/Cy3) >0.3 or <-0.3 were considered low-level alterations of gene copy number and were defined as 'gain or loss'. Genes with log$_2$ (Cy5/Cy3) >0.58 or <-0.58 were considered high-level alterations and defined as 'amplification or deletion'.

The t-test was used to identify prognostic and predictive genes. We developed three models: comparing good vs. poor prognosis i), comparing good vs. poor prognosis only among poor responders ii) and comparing good vs. poor prognosis only among good responders iii). Prognostic genes were selected from post-chemotherapy surgery specimens and predictive genes were selected from pre-treatment biopsy samples. Two-way hierarchical cluster analysis was performed using the Pearson Correlation (Silicon Genetics, Redwood City, CA, USA) on gene sets selected for each p-value.

Prediction analysis. R 1.9.0 software was used for the prediction analysis (19). To validate the discrimination power and prediction accuracy of selected prognostic and predictive genes, leave-one-out cross-validation (LOOCV) or prediction analysis was performed using support vector machine linear (SVML) with genes ranked to their p-values in the t-test. For statistical validation, p≤0.01 was used to select a similar number of genes in each set. The 43 samples were randomly divided into training and test sets in a 2:1 ratio and this process was repeated ten times.

Gene annotation and mapping for chromosomal location. Functional annotation was performed with GeneSpring 6 (Silicon Genetics). Analyzed genes were mapped on each chromosome by CGH analyzer and map viewer program of S-PLUS (Cancer Metastasis Research Center, Yonsei University College of Medicine, Korea).

Table I. Characteristics of patients (n=45).

Age	Median (range)	45 (32-70)
Menopause	Pre-menopause	28 (62%)
	Post-menopause	17 (38%)
Initial T stage	T2	7 (16%)
	T3	28 (62%)
	T4	10 (22%)
Initial N stage	N0	3 (7%)
	N1	23 (51%)
	N2	18 (40%)
	N3	1 (2%)
Initial stage	IIIA	31 (69%)
	IIIB	14 (31%)
Pathology	Infiltrating ductal cancer	43 (96%)
	Infiltrating lobular cancer	2 (4%)

Fluorescence in situ hybridization. FISH targeting the HER-2/ neu gene was performed according to the manufacturer's recommendation (Vysis PathVysion® HER-2 DNA Probe kit, Vysis Inc, IL, USA). The HER-2 probe was labeled with the Spectrum Orange fluorophore and covered ~190 kb (17q11.2-q12) containing the entire erbB-2 gene, while the CEP 17 control probe was labeled with Spectrum Green fluorophore and covered ~5.4 kb (CEP17, 17p11.1-q11.1) containing α-satellite sequence of chromosome 17. DAPI I counter stain (Vysis Inc.) was applied to the specimen to allow visualization of the nuclei. A ratio of erbB-2/CEP-17 >2 or at least 5 copies of the erbB-2 gene was considered genetic amplification.

Validation of copy number changes by real-time polymerase chain reaction (RT-PCR). QuantiTect SYBR-Green PCR master mix (Qiagen, Valencia, CA) was used to measure the relative copy number changes. The reaction consisted of a total volume of 30 μl, containing 10 μl of 2X PCR master mix, 70 ng of the genomic DNA and final concentrations of 0.3 μM oligonucleotide primers as was suggested by the manufacturer. PCR was performed at 95˚C for 15 min in order to activate the HotstarTaq® DNA polymerase and then for 40 cycles of amplification at 95˚C for 15 sec, 58˚C for 1 min, 72˚C for 1 min on a PRISM® 7700 Sequence Detection System (Applied Biosystems, Foster City, CA). Reactions were repeated at least twice for each gene and average threshhold cycle (Ct) was obtained. For parallel comparison between RT-PCR and microarray-CGH, the Ct of each gene was normalized by subtracting the Ct of *Actin* (ΔCt) from same tissue and was compared to the $\log_2$ (Cy5/Cy3) of target gene divided by $\log_2$ (Cy5/Cy3) of actin from microarray-CGH. Dissociation curves were drawn based on multicomponent data to show single amplification product was generated from each reaction. The oligonucleotide primers used for the amplification were as follows: ACTG2 (*Actin*, γ 2, *smooth muscle*, *enteric*, Exon 8) forward (F) (5'-GGAGTCCGCTGG AATTCATGAGACAA-3'), ACTG2 reverse (R) (5'-TGATC TCCTTCTGCATCCTGTCAGCAA-3'), PGM2 (*Phospho-*

glucomutase 2, Exon 9) F (5'-GAGGGTGTTTTCAGGCA ATG-3'), PGM2 R (5'-TAAGGCAATGGCCCGCAAGA-3'), DMD (*Dystrophin*, Exon 76) F (5'-CCAGGCAGAGGCCAA AGTGA-3') DMD R (5'-CCGAAGTTTGACTGCCAACCA-3') and TROAP (*Troponin associated protein*, Exon 13) F (5'-CCT GTCTTCCAGAGGAGTGC-3'), TROAP R (5'-CTCCGGT ATCTCAGGCTCAC-3').

Prediction of response to neoadjuvant chemotherapy with biopsy specimens. Chemo-response prediction with genes that were selected was performed, using Prediction Analysis of Microarray (PAM) software (http://www-stat.stanford.edu/ ~tibs/PAM). PAM adopts a variation of shrunken-centroid classification, with an automated gene selection step integrated into the algorithm. We determined the optimal Δ value from a train set and applied it to an independent test set of patients.

Results

Clinical characteristics and treatment results. The clinical response rate following neoadjuvant chemotherapy was 87%. The median follow-up duration was 51 months (range 7-122 months). Among 45 patients, 27 patients (60%) relapsed within 5 years after surgery. Five-year disease-free survival and overall survival rates were 32.9 and 52.7%, respectively. In 12 patients, pre-treatment biopsy specimens were paired to resected specimens after neoadjuvant chemotherapy. An additional 5 specimens were obtained from the recurred sites. Hence, a total of 62 microarray-CGHs from 45 LABC patients were performed. The characteristics of the patients are summarized in Table I. The scheme for microarray-CGH data analysis is illustrated in Fig. 1A.

Global genetic alterations after unsupervised clustering. From 12,492 filtered genes, 9,269 genes were selected that changed over |0.58| in at least one out of 62 cases. Selected genes were used in unsupervised cluster analysis and the global pattern of gene copy number alterations was visualized as a Treeview dendrogram (Fig. 1B). Although the 62 tissue samples were separated into four broad subclasses, these subgroups did not reflect clinical features such as stage, pathology, response to chemotherapy and recurrence. Nevertheless, 8 out of 12 paired specimens (pre-treatment biopsy and post-chemotherapy and surgery samples) clustered together and another 2 paired specimens showed homology, suggesting a similar genetic pattern in tissues from the same patient irrespective of chemotherapy. Paired biopsy and surgery specimens clustered tightly in one patient with mucinous tumor type and in a second patient with a squamous type. In addition, three loco-regional relapsed specimens (chest wall and axillary lymph node) out of five relapsed specimens were located close to the primary tumor under the same branch, which also reflected similar intra-individual genetic pattern (Fig. 1B).

Selection of prognostic genes after neoadjuvant chemotherapy Prognostic gene set I. Gene copy number alterations were compared between good and poor prognosis groups from 43 samples (excluding the 2 cases of pCR) after neoadjuvant chemotherapy. The t-test with nominal p-values of 0.001,
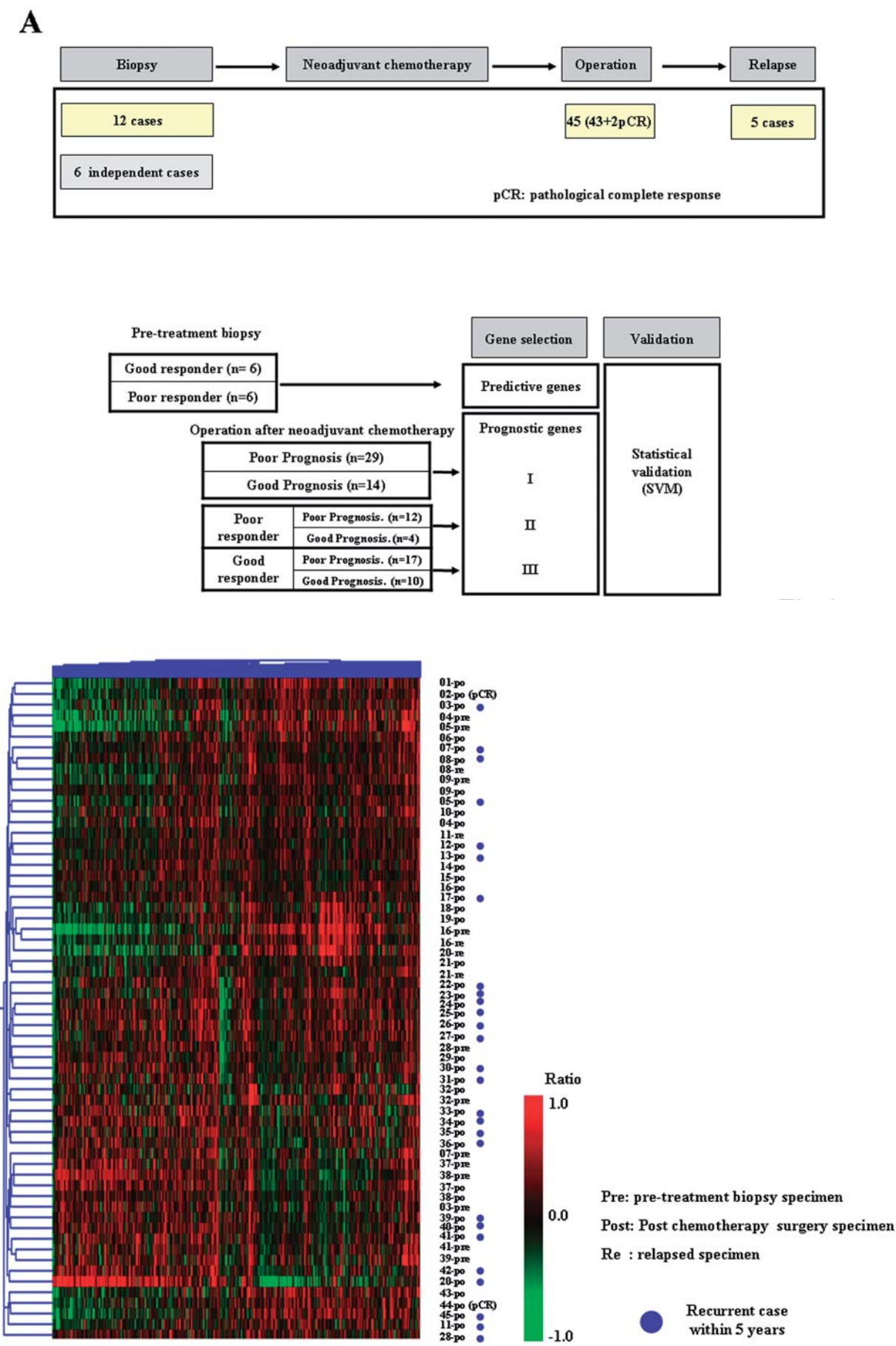
A

B

Figure 1. Study schemes and unsupervised clustering for global pattern. (A) Study design based on clinical status of the patients and microarray-CGH data analysis algorism. (B) Unsupervised cluster analysis with selected 9,269 genes for global pattern of gene copy number alterations using Treeview dendrogram. Pathological CR cases are displayed with the sample. Blue circle indicates the patient recurred within 5 years after the treatment. 'Pre' and 'po' denotes before and after the neoadjuvant chemotherapy, respectively. 'Re' denotes relapse of cancer.
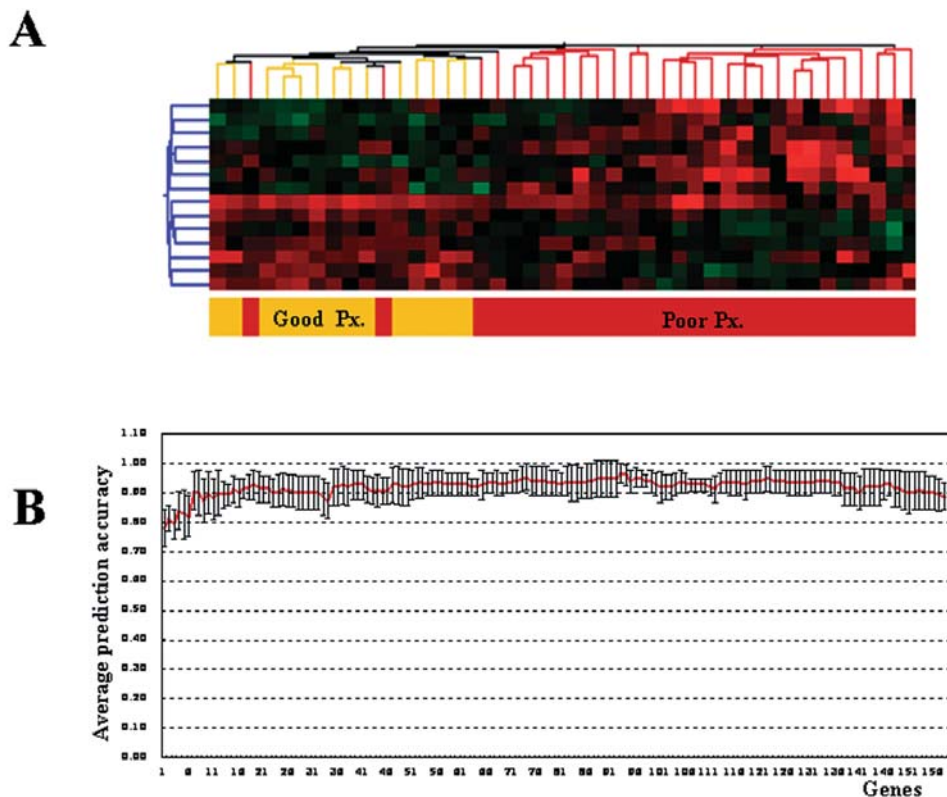
Figure 2. Prognostic gene set I from surgical specimen after neoadjuvant chemotherapy. (A) Two-way hierarchical clustering with selected 158 genes resulting in effective separation of the good and poor prognosis groups. (B) Ninety-two percent of average accuracy in recurrence prediction with the selected 158 genes.

0.005, 0.01, and 0.05 selected sets of 14, 81, 158 and 779 genes, respectively. Two-way hierarchical clustering performed with each gene set showed that genetic patterns differentiated the good and poor prognosis groups effectively. For example, 39 of 43 cases were correctly classified into the good or poor prognosis group by the copy number patterns of 14 genes (Fig. 2A). The prognostic performance of the 158 selected genes with $p < 0.01$ was evaluated by randomly dividing the 43 samples into training and testing samples. Repeated re-grouping of training and testing sets was performed with 29 recurrent and 14 nonrecurrent cases. The entire training and testing processes were repeated using SVM algorithms in R package. After the reiterative sequential addition of one gene to 158 genes, the average recurrence prediction accuracy was determined to be 92% (Fig. 2B).

When we compared the prognostic significance of gene set I to the survival after surgery with known clinical factors, clinical or pathological tumor response and microarray-CGH-based prediction by gene set I were the strongest prognostic factors. Gene set I was found to be an independent predictive factor by multivariate analysis (Table II).

Prognostic gene set II and III. To find the subgroups of good prognosis among poor responders and poor prognosis among good responders, the 43 patients were divided into good and poor responder groups based on the pathological tumor response after iFAC treatment. Prognostic gene set II, which differentiates good and poor prognosis subgroups among poor responders, and prognostic gene set III, which differentiates good and poor prognosis subgroups among good responders, were selected by t-test with nominal p-values of

0.001, 0.005, 0.01 and 0.05, respectively. Two-way hierarchical clustering with 51 selected genes for set II (Fig. 3A) and 32 genes for set III (Fig. 3B) showed clear genetic patterns that differentiated the poor and good prognostic subgroups in each set.

Using SVM analysis adding one gene to 104 genes for set II and 55 genes of set III, binary classification was performed to validate the gene sets II and III. The overall predictive accuracy was 99.9 and 96% with gene sets II (Fig. 3C) and III (Fig. 3D), respectively.

Validation of cDNA-based microarray-CGH. The microarray-CGH technique enables direct comparison with FISH data on a gene-by-gene basis throughout the genome. Two clones of the ERBB2 gene, AW057736 and AA446928, were spotted on the cDNA microarray used in this study (Fig. 4A). FISH analysis targeting the ERBB2 gene (190 kb, 17q11.2-q12) was performed for 34 patients and the amplification of ERBB2 compared to control CEP was evaluated (Fig. 4B). The data from microarray-CGH and FISH were compared. Among the 34 patients, the amplification signal in FISH and the gene copy number gain in microarray-CGH correlated in 8 cases. However, in 12 cases the copy number gain of clones AW057736 (10 cases) and AA446928 (2 cases) shown by microarray-CGH was under-estimated in the FISH study. Six cases of gene loss in microarray-CGH showed neither signal by FISH.

We compared the values between microarray-CGH and quantitative PCR with 3 other genes (DMD, PGM2, TROAP). The two experimental values correlated well in these patients (Fig. 4C).

Table II. Comparison of prognostic factors for survival.

|	Disease-free survival		Overall survival	
	Univariate	Multivariate	Univariate	Multivariate
Age (<50 vs. >50)	0.471	-	0.457	-
Menopause	0.499	-	0.792	-
Estrogen receptor	0.461	-	0.272	-
Progesterone receptor	0.895	-	0.523	-
cerbB-2	0.660	-	0.527	-
Tumor grade (1, 2 vs. 3)	0.530	-	0.885	-
Prechemotherapy stage (IIIA vs. IIIB)	0.378	-	0.748	-
Pathological stage (I, II vs. III)	0.289	-	0.933	-
Clinical response (CR, PR vs. SD, PD)	0.02	-	0.063	-
Pathological response (CR, PR vs SD, PD)	0.323	-	0.03	-
RDI (<80 vs. >80)	0.639	-	0.366	-
Microarray-CGH (good vs. poor)	<0.0000	0.004	0.0001	0.015
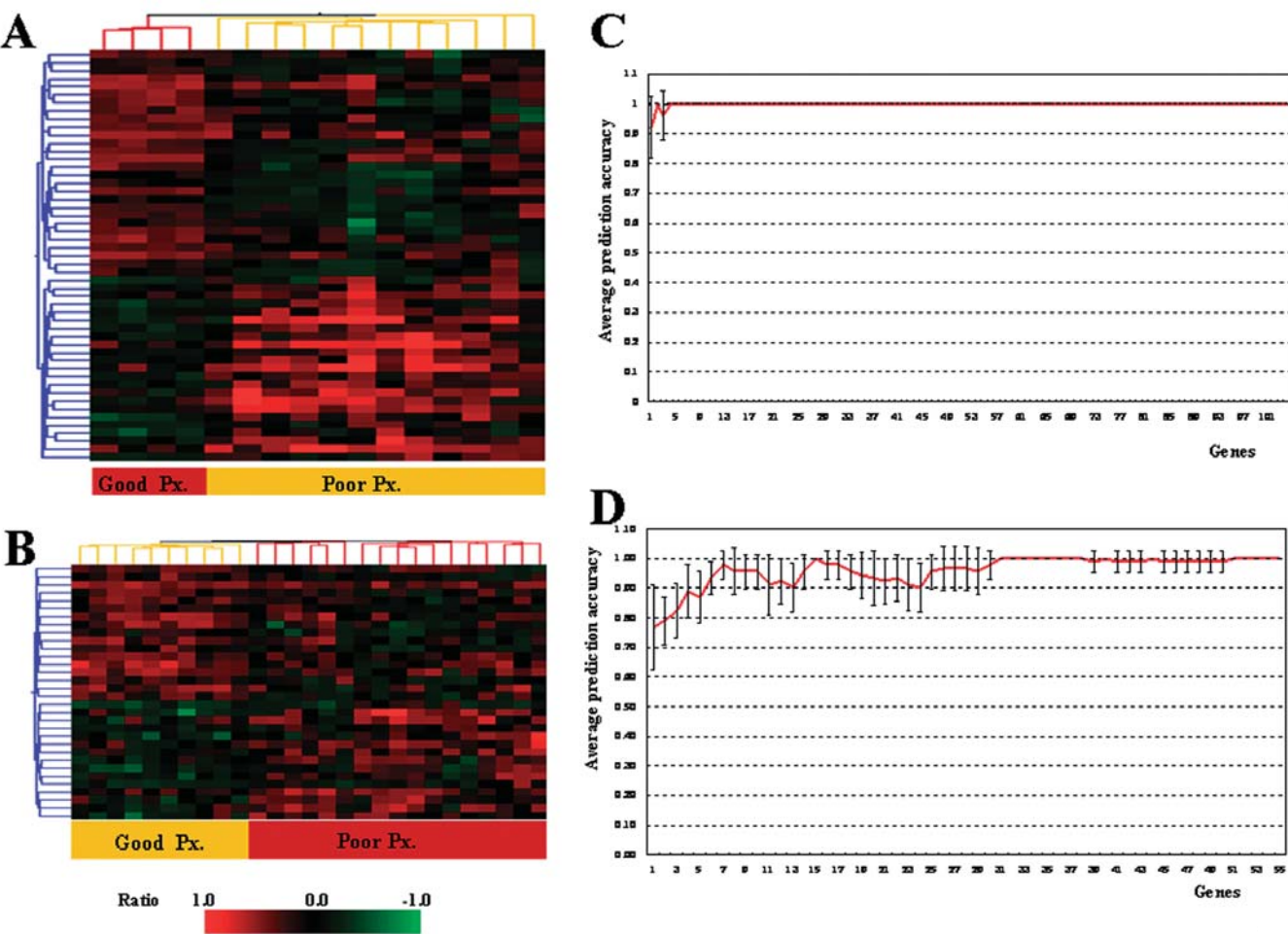
Figure 3. Prognostic gene set II and III from surgical specimen after neoadjuvant chemotherapy. (A) Two-way hierarchical clustering with gene set II for poor-responders showing different prognosis with different genetic patterns. (B) Two-way hierarchical clustering with gene set III for good-responders showing different prognosis with different genetic patterns. (C) Average accuracy (99.9%) in relapse prediction with the selected 104 genes. (D) Average accuracy (96.2%) in relapse prediction with the selected 55 genes.
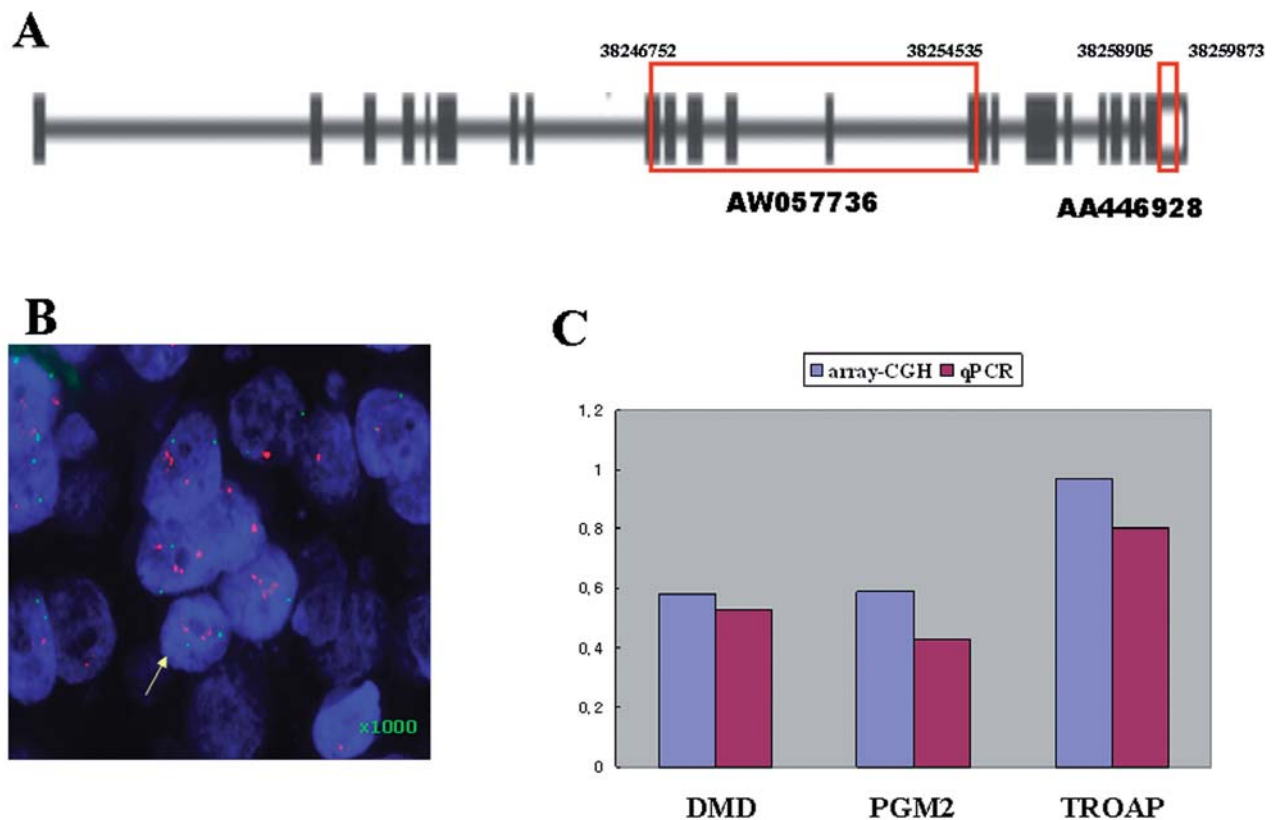
Figure 4. Validation of microarray-CGH with FISH of ERBB2 and quantitative PCR in other selected genes. (A) Location of 2 clones of the ERBB2 gene, AW057736 and AA446928, spotted on the cDNA microarray. (B) FISH result showing target gene (red, cerbB-2) and reference gene (blue). (C) Comparison of quantitative PCR and copy number change by microarray-CGH with 3 separate genes in independent patients.

Selection of predictive genes from biopsy specimens before neoadjuvant chemotherapy. Based on the tumor response to neoadjuvant chemotherapy, 12 biopsy tissues were divided into 6 cases of good responders (chemosensitive) and 6 cases of poor responders (chemoresistant). Predictive genes were selected as those with differences in gene copy number between good and poor responders. On hundred and eighty-four predictive genes associated with drug response were selected (p<0.05), the majority of which were related to apoptosis and cell cycle regulation, such as DAD1 and CASP2.

Two-way hierarchical clustering was performed with the 184 genes. Good and poor responder groups were correctly classified by the 184 genes (Fig. 5A). Leave-one-out cross-validation was used to assess the significance of the selected 184 genes using SVML. Fig. 5B shows the results of the validation for the 12 biopsy specimens. The average accuracy of classification was 99.8% and the accuracy was 100% when 4 to 184 genes were used for prediction in SVML.

There were 6 independent patients who were not involved in the gene selection process. The responses to neoadjuvant chemotherapy of those 6 patients were 5 PR and 1 SD. PAM analysis with predictive gene set showed that one SD patient was predicted as indeterminate prediction as expected and 4 of 5 PR patients were predicted as good responders. One PR patient was miss-classified as a poor responder, resulting in the classification accuracy of 83.3% (Fig. 5C). We also compared 4 selected genes (EPO, Hox4, DAD1, CASP2) with quantitative PCR between responders and non-responders.

There were differences in these 4 genes between the two groups (Fig. 5D).

Discussion

Previous studies showed that genome-wide expression profiles could predict clinical outcome in breast cancer (15,16,20-29). An important question addressed in these studies is whether there is a relationship between changes in gene copy number and treatment response or prognosis. In multiple myeloma, minimal common regions for recurrence and distinct genomic subtypes were discovered by high-resolution genomic profiles (30). To our knowledge, this is the first study of genome-wide alterations of gene copy number using cDNA microarray (microarray-CGH) using paraffin blocks of surgical specimens as a prognostic profile after neoadjuvant chemotherapy in locally advanced breast cancer of stage III.

Initial analyses used unsupervised hierarchical clustering to study global genetic patterns. In most cases, pre- and post-therapy specimens or relapsed tumors from the same patient clustered together. Similar results were obtained from RNA expression profiling studies in breast cancer after pre-operative chemotherapy (14,23,25,26). This reflects rather low number of copy change responded to chemotherapy than individual variance from another patients, which can be the mainstay of individualized medicine.

Subsequently, three prognostic gene sets corresponding to various clinical situations were selected. There are two potential problems with this approach (31). The first issue is
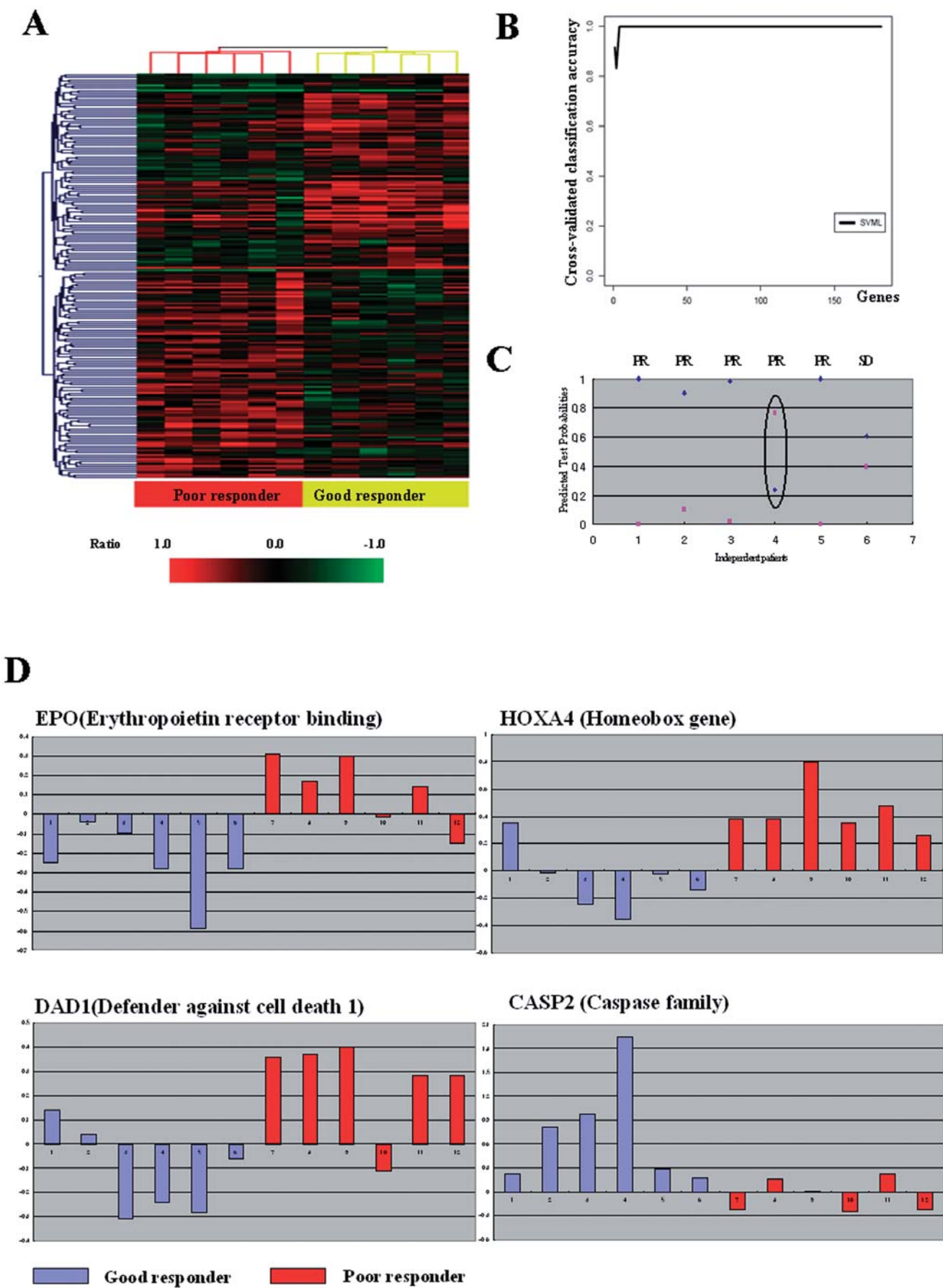
Figure 5. Neoadjuvant chemotherapy response predictive gene set from biopsy. (A) Two-way hierarchical clustering with the selected 184 genes. (B) Leave-one-out cross-validation with 12 biopsy specimens to assess the significance of the selected 184 genes using SVML (support vector machine linear) method. (C) Response prediction by PAM assay using predictive gene set in 6 independent patients. Circle represents one miss-predicted patient. (D) Comparison of quantitative PCR with 4 different genes between responders and non-responders.

the phenomenon of increasing number of 'false positives' with increasing numbers of hypotheses tested, known as the 'multiple comparison problem'. To overcome this, we used adjusted p-value by permutation from which we generated a random distribution of prediction error rates even if no gene was selected. The second issue with this approach is that standard parametric t-test assumes a normal distribution of the data. Before analysis, we confirmed that the microarray-CGH data were closer to the normal distribution than RNA expression profiling data (data not shown). The selected informative gene lists can change with the increment of the training set size. In addition, it is well known that the rank order of genes is especially susceptible to change from one study to another. The gene sets selected in the present study do not share many genes with a previous expression profiling study (14-16,20-29). This may be due to heterogeneity of the patients in the previous study relative to that of the cohort reported here, use of different drugs or platforms. Interestingly, there was little overlap among selected gene sets in our study.

For data validation, LOOCV analysis was initially performed with prognostic gene set II and III, giving an overall accuracy of 99.7 and 96%. For statistical validation of the selected gene sets by t-test, prediction analysis was performed by repeatedly grouping patients randomly into training and test sets. In all cases, the predictive accuracies were >90% for each gene set. The gene sets were also tested by other algorithms including k-nearest neighbor, neural network and linear discrimination analysis in R package (data now shown). All the data showed similar results to SVML. In this study, various attempts were made to use the most parsimonious model to minimize potential errors from over fitting, since the same samples were used for both gene selection and prediction, which could increase the chance of bias in prediction.

We developed three clinical models: comparing good vs. poor prognosis i), comparing good vs. poor prognosis only among poor responders ii) and comparing good vs. poor prognosis only among good responders iii). In view of the accepted adjuvant treatment in patients with residual tumor after neoadjuvant chemotherapy and surgery (32), it would be useful to identify patients with tumors at a high risk of recurrence for treatment with a non cross-resistant adjuvant regimen as we do in osteosarcoma. When we compared the gene set I for the survival prediction after surgery with known clinical factors, it was found to be an independent factor for the prediction. Based on the clinical features of patients in this study, our data suggest that molecular prediction based on microarray-CGH may aid clinical decisions. In this study, patients were treated with the same regimen before and after surgery. The point is whether we can use the same chemotherapy regimen after surgery as neoadjuvant regimen or we need to change the regimen based on our markers presented in the remained cancer after neoadjuvant chemotherapy. This point needs further prospective clinical trial for the validation as a predictive biomarker after surgery.

The present study was not designed to discover specific genes related to particular drugs, but rather to identify patterns of gene copy number alterations that could be used as a prognostic marker. As a result, our gene set excluded some genes with low genetic alteration, some of which may have been biologically interesting. Although these values are likely to be slightly biased, gene copy number alterations could be used in a prognostic role in future biomarker discovery studies.

For validation, FISH was performed with a 190 kb region of HER-2/neu/c-ERBB2 and GRB7 (H53703) genes. Amplification of HER-2/neu/c-ERBB2 in FISH correlated well with microarray-CGH data in both HER-2/neu/c-ERBB2 and GRB7. We also confirmed that GRB7 was in the predictive gene sets and was amplified in poor responders. As many cancer-related genes exist around HER-2/neu/c-ERBB2 and complex genetic aberrations exist in breast cancer (11,33), it can be assumed that alteration of one gene at the DNA level is insufficient to account for response or recurrence, compared with profiles of related gene sets (34,35).

In conclusion, the present study demonstrates that sets of prognostic genes can be selected from paraffin blocks of LABC using microarray-CGH. Such gene sets may be valuable for the prediction of clinical outcome. Determination of prognostic alterations in gene copy number from tissues after chemotherapy may be of importance with respect to prediction of patients at high risk of recurrence and for treatment regimen selection following neoadjuvant chemotherapy and surgery.

Acknowledgements

References

1. Monni O, Hyman E, Mousses S, Barlund M, Kallioniemi A and Kallioniemi OP: From chromosomal alterations to target genes for therapy: integrating cytogenetic and functional genomic views of the breast cancer genome. Semin Cancer Biol 11: 395-401, 2001.
2. Alnertson DG, Ylstra B, Segraves R, et al: Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. Nat Genet 25: 144-146, 2000.
3. Collins C, Volik S, Kowbel D, et al: Comprehensive genome sequence analysis of a breast cancer amplicon. Genome Res 11: 1034-1042, 2001.
4. 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 USA 99: 12963-12968, 2002.
5. Hyman E, Kauraniemi P, Hautaniemi S, et al: Impact of DNA amplification on gene expression patterns in breast cancer. Cancer Res 62: 6240-6245, 2002.
6. Pollack JR, Perou CM, Alizadeh AA, et al: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23: 41-46, 2001.
7. O'Hagan RC, Brennan CW, Strahs A, et al: Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. Cancer Res 63: 5352-5356, 2003.
8. Buchholz TA, Hunt KK, Whitman GJ, Sahin AA and Hortobagyi GN: Neoadjuvant chemotherapy for breast carcinoma: multi-disciplinary considerations of benefits and risks. Cancer 98: 1150-1160, 2003.
9. Faneyte IF, Schrama JG, Peterse JL, Remijnse PL, Rodenhuis S and van de Vijver MJ: Breast cancer response to neoadjuvant chemotherapy: predictive markers and relation with outcome. Br J Cancer 88: 406-412, 2003.

10. Steger GG, Galid A, Gnant M, *et al*: ABCSG-14, Pathologic complete response with six compared with three cycles of neoadjuvant epirubicin plus docetaxel and granulocyte colony-stimulating factor in operable breast cancer: Results of ABCSG-14. J Clin Oncol 25: 2012-2018, 2007.

11. Coon JS, Marcus E, Gupta-Burt S, *et al*: Amplification and overexpression of topoisomerase II alpha predict response to anthracycline-based therapy in locally advanced breast cancer. Clin Cancer Res 8: 1061-1067, 2002.

12. Chang J, Powles TJ, Allred DC, *et al*: Biologic markers as predictors of clinical outcome from systemic therapy for primary operable breast cancer. J Clin Oncol 17: 3058-3063, 1999.

13. Slamon D, Eiermann W and Robert N: Phase III randomized trial comparing doxorubicin and cyclophosphamide followed by docetaxel (AC→T) with doxorubicin and cyclophosphamide followed by docetaxel and trastuzumab (AC→TH) with docetaxel, carboplatin and trastuzumab (TCH) in HER2 positive early breast cancer patients: BCIRG 006 study. Breast Cancer Res Treat 94 (Suppl 1): S5, 2005.

14. Chang JC, Wooten EC, Tsimelzon A, *et al*: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 362: 362-369, 2003.

15. van 't Veer LJ, Dai H, van de Vijver MJ, *et al*: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536, 2002.

16. van de Vijver MJ, He YD, van't Veer LJ, *et al*: A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009, 2002.

17. Sambrook J, Fritsch EF and Maniatis T: Molecular cloning, a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp21-26, 1989.

18. Park CH, Keong HJ, Choi YH, *et al*: Systematic analysis of cDNA microarray-based CGH. Int J Mol Med 17: 261-267, 2006.

19. Venables WN and Ripley BD: Modern Applied Statistics with S. 4th ed. Springer, ISBN 0-387-95457-0, 2002.

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

21. Paik SM, Shak S, Tang G, *et al*: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817-2826, 2004.

22. Sotiriou C, Powles TJ, Dowsett M, *et al*: Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. Breast Cancer Res 4: R3, 2002.

23. Buchholz TA, Stivers DN, Stec J, *et al*: Global gene expression changes during neoadjuvant chemotherapy for human breast cancer. Cancer J 8: 461-468, 2002.

24. Modlich O, Prisack HB, Munnes M, Audretsch W and Bojar H: Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease. Clin Cancer Res 10: 6418-6431, 2004.

25. Ayers M, Symmans WF, Stec J, *et al*: Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 22: 2284-2293, 2004.

26. Huang E, Cheng SH, Dressman H, *et al*: Gene expression predictors of breast cancer outcomes. Lancet 361: 1590-1596, 2003.

27. van't Veer LJ, Dai H, van de Vijver MJ, *et al*: Expression profiling predicts outcome in breast cancer. Breast Cancer Res 5: 57-58, 2003.

28. Lonning PE, Sorlie T, Perou CM, Brown PO, Botstein D and Borresen-Dale AL: Microarrays in primary breast cancer - lessons from chemotherapy studies. Endocr Relat Cancer 8: 259-263, 2001.

29. Iwao-Koizumi K, Matoba R, Ueno N, *et al*: Prediction of docetaxel response in human breast cancer by gene expression profiling. J Clin Oncol 23: 422-431, 2005.

30. Carrasco DR, Tonon G, Huang Y, *et al*: High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. Cancer Cell 9: 313-325, 2006.

31. Westfall PH and Young SS: Resampling-based multiple testing. Wiley, New York, p72, 1993.

32. Estevez LG and Gradishar WJ: Evidence-based use of neoadjuvant taxane in operable and inoperable breast cancer. Clin Cancer Res 10: 3249-3261, 2004.

33. Luoh SW, Venkatesan N and Tripathi R: Overexpression of the amplified Pip4k2beta gene from 17q11-12 in breast cancer cells confers proliferation advantage. Oncogene 23: 1354-1363, 2004.

34. Fridlyand J, Snijders AM, Ylstra B, *et al*: Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer 6: 96, 2006.

35. McSherry EA, McGoldrick A, Kay EW, Hopkins AM, Gallagher WM and Dervan PA: Formalin-fixed paraffin-embedded clinical tissues show spurious copy nymber changes in array-CGH profiles. Clin Genet 72: 441-447, 2007.