Conservation of Breast Cancer Molecular Subtypes and Transcriptional Patterns of Tumor Progression Across Distinct Ethnic Populations

Kun Yu,[1] Chee How Lee,[1] Puay Hoon Tan,[3] and Patrick Tan[1,2]

[1]National Cancer Centre/[2]Defence Medical and Environmental Research Institute, and [3]Department of Pathology, Singapore General Hospital, Republic of Singapore

ABSTRACT

***Purpose:* Breast cancers can display distinct clinical characteristics in different ethnic populations. Previous studies involving European and United States patients have shown that breast tumors can be divided by their gene expression profiles into distinct "molecular subtypes." In this report, we surveyed a series of invasive and preinvasive breast tumors from Asian-Chinese patients to investigate whether similar subtypes could also be observed in this ethnic group.**

***Experimental Design and Results:* An analysis of expression profiles generated from 11 nonmalignant breast tissues, 17 ductal carcinomas *in situ* (DCIS) and 98 invasive carcinomas identified three broad molecular subtypes of breast [estrogen receptor (ER)+, ERBB2+ and ER−] in the Asian-Chinese population. These subtypes were highly similar to the "Luminal," "ERBB2+," and "Basal" molecular subtypes defined in previous studies, and the subtype-specific expression signatures were also observed in preinvasive DCIS tumors. By comparing the expression profiles of nonmalignant DCIS and invasive breast cancers for two subtypes (ER+ and ERBB2+), we identified several genes that were regulated in both a common and subtype-specific manner during the normal/DCIS and DCIS/invasive carcinoma transitions. Several of these genes were validated by comparison with another recently published similar, but not identical, study.**

***Conclusions:* Our results suggest that molecularly similar subtypes of breast cancer are indeed broadly conserved between Asian and Caucasian patients, and that these subtypes are already present at the preinvasive stage of carcinogenesis. To our knowledge, this study is among the first to directly compare the expression profiles of breast tumors across two different ethnic populations.**

INTRODUCTION

Breast cancer is a significant cause of worldwide morbidity and mortality in females (1). A major challenge in the diagnosis and treatment of breast cancer is its heterogeneity, because individual breast tumors can exhibit tremendous variations in clinical presentation, disease aggressiveness, and treatment response (2). Breast cancers can also display strikingly distinct clinical characteristics in different patient and ethnic populations (3). For example, in Caucasian populations, most breast cancers occur in postmenopausal women at a mean and median age of 60 and 61 years, respectively (4). In contrast, a bimodal pattern of incidence beginning at age 40 is observed in Asian populations, such as those in Singapore and Japan (ref. 1; see also Supplementary Information[4]). Besides differing in their incidence rates, breast cancers in Caucasian and Asian patients have also been reported to differ in terms of their risk factors and repertoires of molecular and chromosomal abnormalities (5–8).

Recently, several groups have found that breast tumors can be classified into different "molecular subtypes" based on their global expression profiles (9–14). Many of these subtypes were discernible by gene expression profiling but not by more conventional methodologies and were associated with distinct clinical outcomes, demonstrating the clinical utility of using gene expression information to develop a molecular taxonomy of cancer. A potential limitation of these studies, however, is that they were primarily based on United States and European patient populations. Because of the observed clinical differences in breast cancer between Asian-Chinese and Caucasian patients (described above), we investigated, in this report, whether similar molecular subtypes could also be observed in a predominantly Chinese population, or whether breast tumors between these different ethnic populations might also differ at the gene expression level. By surveying the expression profiles of preinvasive and invasive cancers obtained from predominantly Chinese patients, we found that many, but not all, of the molecular subtypes and their associated subtype-specific gene expression signatures were indeed conserved between Caucasian and Chinese patients, suggesting that the molecular subtypes defined using expression-based genomics are highly robust and may be population independent. We also found, for the first time, that the subtype-specific expression signatures were also present in

Received 1/15/04; revised 4/23/04; accepted 5/14/04.

Grant support: Funded by a grant from Agenica Research.

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 can be found at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org).

Requests for reprints: Patrick Tan, National Cancer Centre/Defence Medical and Environmental Research Institute, 11 Hospital Drive, Singapore 169610, Republic of Singapore. Phone: 65-6-436-8345; Fax: 65-6-226-5694; cmrtan@nccs.com.sg.

[4] Supplementary information is available at Internet address: http://clincancerres.aacrjournals.org.

preinvasive breast cancers [ductal carcinomas in-situ (DCIS)], indicating that these molecular subtypes can already be discerned even at the preinvasive stage of carcinogenesis. By comparing the expression data from normal tissue, DCIS, and invasive carcinomas (IDCs) belonging to two specific subtypes (ER+ and ERBB2+), we identified several genes that were regulated in both a common and subtype-specific manner during the normal/DCIS and DCIS/IDC transitions. Many of these genes were then validated by comparison with publicly available data from another recently published, related but not identical, study (15). The identities of these genes may prove useful in elucidating the molecular events regulating tumor progression in distinct molecular subtypes of breast cancer.

MATERIALS AND METHODS

Breast Tissues and Clinical Information. Human breast tissues were obtained from the (National Cancer Centre Tissue Repository, with approvals from the National Cancer Centre Repository and Ethics Committees. Samples were grossly dissected in the operating theater immediately after surgical excision and were flash-frozen in liquid nitrogen. Patients had not been treated with preoperative chemotherapy. For histological assessment of tumors and of axillary lymph nodes, formalin-fixed, paraffin-embedded tumor tissue was used to determine tumor subtype (WHO classification), histological grade, and lymphovascular invasion. Estrogen receptor (ER) status was determined by immunohistochemistry, with a positive result being ≥10% of carcinoma cells showing nuclear reactivity of at least +2 intensity. For ERBB2 immunohistochemistry, the Dako classification system was used with scores of 0 and 1+ considered negative and 2+ and 3+ considered positive. An indeterminate conclusion was made when benign breast epithelium was immunoreactive. Profiled invasive tumors contained at least 50% tumor content, whereas DCIS samples contained 20–30% (see Results). Confirmation of the DCIS status of samples was achieved by conventional H&E staining of archival samples, as well as direct cryosections of the sample that was processed for expression profiling. Four of the DCIS samples were pure DCIS, and the other samples were DCIS adjacent to invasive tumors. The clinical characteristics of the invasive and DCIS tumors (*e.g.,* tumor size, nodal status, histological grade and type, ER/progesterone receptor status, and ERBB2 status) are presented in the Supplementary Information.[4]

Sample Preparation and Microarray Hybridization. RNA was extracted from tissues using Trizol reagent (Invitrogen, Carlsbad, CA), purified through a Qiagen Spin Column (Qiagen; Valencia, CA), and processed for Affymetrix Genechip (Affymetrix Inc., Santa Clara, CA) hybridization according to the manufacturer's instructions. Hybridizations were performed using Affymetrix U133A Genechips. Annotations assigned to array probes are based on the Dec 2003 release available from the Affymetrix website.[5]

Data Processing and Analysis. Raw Genechip scans were quality controlled using a commercially available software (Genedata Refiner; Genedata, Basel, Switzerland) and were deposited into a central data storage facility. The expression data were filtered by (*a*) removing genes the expression of which was absent in all of the samples (*i.e.*, "A" calls), and (*b*) performing a $\log_2$ transformation, and were normalized by median centering all remaining genes for each sample. Average-linkage hierarchical clustering using a Pearson correlation metric was performed using CLUSTER software and were displayed by TREEVIEW (16). Wilcoxon tests, which are nonparametric alternative methodologies to conventional *t* tests, were used to identify genes whose expression levels were significantly different between two groups, using a 2-fold cutoff and *P* value of <0.01. Support Vector Machines (SVMs) are classification algorithms that define a discrimination surface in the used feature (gene) space that attempts to maximally separate classes of training data (17). Both Wilcoxon tests and SVM analyses were performed using Genedata Analyst software. Random permutation assays, performed to obtain estimates of potential false discovery error rates, were implemented as described in the Results. Selection of random gene sets was performed using a starting population of 7,288 genes, corresponding to the same gene set on which the normal *versus* DCIS, or DCIS/IDC Wilcoxon analysis was performed. The various gene sets are available for download at http://www.omniarray.com/DCIS.html. Kolmogorov-Smirnov analysis was performed as described in Lamb *et al.* (18). Briefly, we first used Significance Analysis of Microarrays (SAM; ref. 19) to identify the top genes exhibiting the strongest expression differences between the ER+ and ER− subtypes in our data set, and then calculated the distribution of these genes in the Stanford data set (comprising primarily Caucasian patients) using the Kolmogorov-Smirnov nonparametric rank statistic. In this analysis, the genes in the Stanford data set were ranked according to their signal to noise ratio (20) corresponding to the ER+/ER− class distinction. The significance of an observed distribution was estimated through a series of random permutations (see Supplementary Information for more details).[4]

RESULTS

Distinct Molecular Subtypes of Breast Cancer in an Asian-Chinese Patient Population. The major objective of this study was to perform a molecular survey of invasive and preinvasive breast tumors in our local predominantly Chinese patient population and to investigate whether they could be subdivided into distinct molecular subtypes of breast cancer as described in previous reports using Caucasian patient cohorts (9–11). First, using Affymetrix U133A Genechips, we generated expression profiles for 98 sporadic invasive breast tumors. After data normalization and preprocessing, we applied a SD filter to identify a set of 367 genes exhibiting a high degree of gene expression variation across the tumor series. The 367-gene set was then used to group the tumor expression profiles to one another on the basis of their overall similarity using unsupervised hierarchical clustering (16). As seen in Fig. 1*A*, the breast tumors self-segregated into three major subgroups, hereon referred to as ER+, ER−, and ERBB2+. A similar segregation pattern was also observed when the tumors were reclustered using principal components analysis (PCA), an independent

[5] Affymetrix website address: www.affymetrix.com.

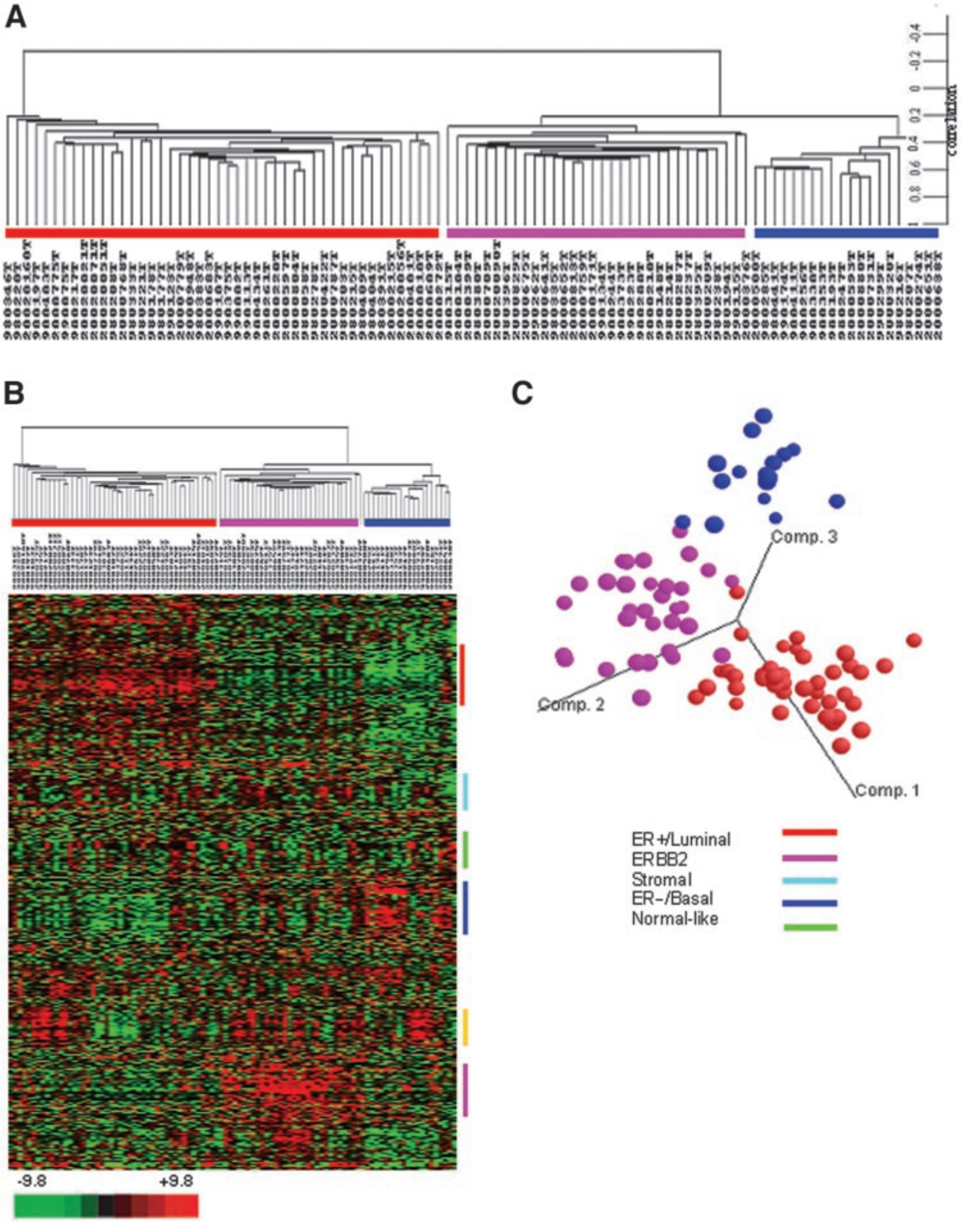

Fig. 1 Identification of molecular subtypes of breast cancer by unsupervised analysis. *A,* unsupervised hierarchical clustering of 98 invasive breast tumors using the top 376 genes exhibiting the highest variation in gene expression. *B,* gene clusters and examples of their members: *ER+/Luminal epithelial cluster* (*ERSR1, GATA3, TFF1, TFF3, STC2*), *ERBB2 cluster* (*ERBB2, maspin, SFRP1*), *Normal east/adipose-enriched cluster* (*FBP4, ADH1B*), *immune cluster* (*IGLJ3, IGHM, IGL*α). A complete list of the 376-gene set and the gene clusters are available for download[4] (see Materials and Methods). *C,* principal component (*Comp. 1, 2,* and *3*) analysis using the 376-gene set. Similar molecular groupings are observed as in *A.*

analytical technique (Fig. 1*C*). There was also good agreement between these molecular subgroups and conventional immunohistochemistry (Supplementary Information),[4] because all of the ER+ tumors identified by molecular profiling were also ER+ by conventional immunohistochemistry, and almost all (15 of 17) of the ER− tumors identified by molecular profiling were also ER− by immunohistochemistry. For the ERBB2+ subtype, 18 tumors for which ERBB2 immunohistochemistry had been performed were all ERBB2+ by immunohistochemistry as well.

ER+ Subtype. This subtype, similar to the "Luminal" subtype described in other studies (9–11), exhibited high levels of a gene expression signature containing the ER gene *ESR1,* and several estrogen-regulated genes, such as *STC2, TFF1, TFF3,* and *MYB* (Supplementary Information[4]). High expression levels of *GATA3* and *Annexin A9* were also observed (9–11). Interestingly, previous studies have also reported the presence of at least two subgroups of Luminal tumors, referred to as Luminal A and Luminal B/C, the latter being associated with decreased levels of the ER-related expression signature and a poorer clinical prognosis (10–11). However, the Luminal B/C subtypes could not be clearly discerned in our data set, because all of the tumors belonging to the ER+ subtype exhibited uniformly high expression levels of the ER related expression signature.

ER− Subtype. The ER− subtype, similar to the "basal" subtype described in other studies,[9-11] was characterized by high levels of markers of the basal mammary epithelia, such as keratin 5 and 17 and the serine proteinase inhibitor maspin, a

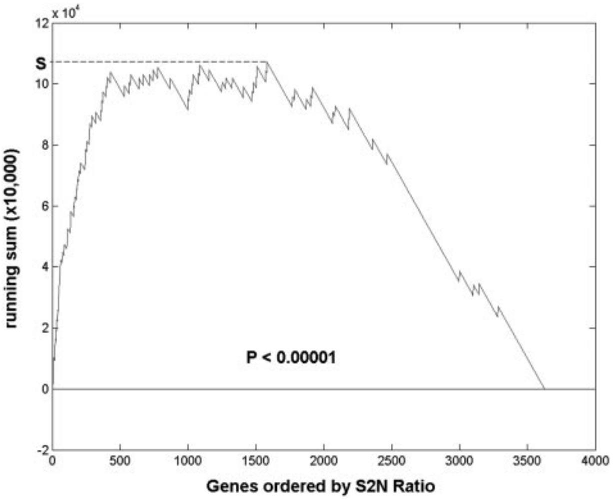

Fig. 2 Kolmogorov-Smirnov (KS) analysis to determine whether genes exhibiting strong differences in expression between the ER+ and ER− subtypes in our Asian data set also exhibited a similar behavior in the Caucasian data set. The *X* axis represents the list of genes from the Caucasian data set (10) ordered by their signal to noise ratio (20) according to their ER classification. The running sum (*Y* axis) of consecutive values of the Vector V (see ref. 18 for details) indicates the distribution of gene expression signatures derived from the Asian-Chinese population within the Caucasian data set, after removal of several well-known ER marker genes (*i.e., ESR1, GATA3, TFF1, TFF3, XBP1, MYB, IGFR1, MUC1, BCL2*). The statistic *S* (KS score) is the maximum score of the running sum. The statistical significance of statistic *S* is indicated by the *P*-value (see Supplementary Information[4] and ref. 18 for details).

tamoxifen-inducible gene expressed in an inverse fashion to ER (21). Notably, SFRP1, a modulator of Wnt signaling, was also highly expressed in this group, a finding also reported by others (11).

ERBB2+ Subtype. The ERBB2+ subtype was associated with high expression levels of the *ERBB2 receptor* and other genes physically linked to the 17q21 locus, such as *PNMT* (22) and *PPARBP*, suggesting the presence of DNA amplification. Interestingly, many of the ERBB2 tumors also exhibited low expression levels of the ER-related gene expression signature, which may reflect the consequences of cellular cross-talk between ERBB2 and ER signaling (23).

The observation that breast tumors from Asian-Chinese patients can also be segregated into distinct ER+, ER− and ERBB2+ subtypes suggests that these subtypes may be independent of specific ethnic population. However, it is still formally possible that the apparent similarities between the Asian and Caucasian subtypes might be limited to only a few well-known marker genes, such as ER, ERBB2, and certain keratins. To address this issue, we determined the statistical commonality of the subtype specific gene expression signatures between our Asian patient population and a previously published Caucasian cohort (10). Specifically, we used Kolmogorov-Smirnov analysis (18) to determine whether genes exhibiting strong differences in expression between the ER+ and ER− subtypes in the Asian data set also exhibited a similar behavior in the Caucasian data set. As shown in Fig. 2, even after removing several "well-known" marker genes, such as *ESR1, GATA3,* and *TFF3,* from the analysis (see Fig. 2 legend), the gene expression signatures associated with the ER+ subtype in our Asian population still exhibited a highly significant degree of commonality with tumors from the Caucasian cohort ($P < 1 \times 10^{-5}$). These results suggest that these subtypes are molecularly conserved between the ethnic groups not simply at the level of a few surface markers but also at the deeper cellular level of biological signaling pathways and transcriptional networks.

Subtype-Specific Expression Signatures Are Already Present in Preinvasive Cancers. We then investigated whether a series of preinvasive breast cancers (DCIS) could also be similarly divided into distinct molecular subtypes, because previous reports have primarily relied on analyzing the expression profiles of only invasive breast tumors. We identified 17 DCIS tissue samples and confirmed their preinvasive status using conventional H&E staining as well as frozen cryosections of the sample that was processed for expression profiling (Fig. 3*A–D*). Compared with the invasive tumor samples, the percentage of malignant cells in the DCIS samples was comparatively smaller (typically 20–30%). To assess whether the higher percentage of nonmalignant cells in the DCIS samples might compromise the resultant tumor expression profile, we compared the expression profiles of the DCIS samples with a series of nonmalignant breast tissue samples using a PCA (Fig. 3*E*). The DCIS profiles were distinctly different from the nonmalignant profiles, suggesting at a first approximation that the higher percentage of nonmalignant cells does not severely confound the overall tumor expression profile. Expression profiles of the DCIS samples were generated and compared with their invasive counterparts using the 367-gene set of Fig. 1*A* and unsupervised clustering. The DCIS samples intermingled with the invasive tumors across distinct categories, with eight DCIS samples segregating into the ER+ subtype, eight into the ERBB2+ subtype, and one into the ER− subtype (*vertical lines,* Fig. 3*F*). This segregation pattern was also confirmed using PCA (Supplementary Information).[4] This result indicates that DCIS tumors, despite being preinvasive, can already be segregated into distinct molecular subtypes, suggesting that the hallmark expression signatures defining the different subtypes may already be present at the preinvasive stage of carcinogenesis.

To further explore this possibility, we used a supervised learning algorithm, SVM (17), to classify the 98 invasive tumors by their specific subtype (the "training set"). Unlike the unsupervised clustering analyses in which a 367-gene set was used, this analysis used all of the genes on the U133A Genechip. The trained algorithm was then asked to classify the DCIS tumors (the blinded "test" set), with quantitative metrics being used to provide an index of the "confidence" or "certainty" of the classification. Sixteen of 17 DCIS samples were classified by the SVM algorithm as belonging to the same subtype as the unsupervised clustering analysis (Supplementary Information),[4] with only one sample (980173) being labeled as an "uncertain" or "no-call" classification. To rule out the possibility that the ability of the DCIS lesions to be separated into "distinctive molecular subtypes" might be due simply to their ERBB2 status, we repeated the SVM analysis on a truncated expression data set in which all of the genes located on the Chromosome 17q locus (where the *ERBB2* gene resides) were excluded. Despite the

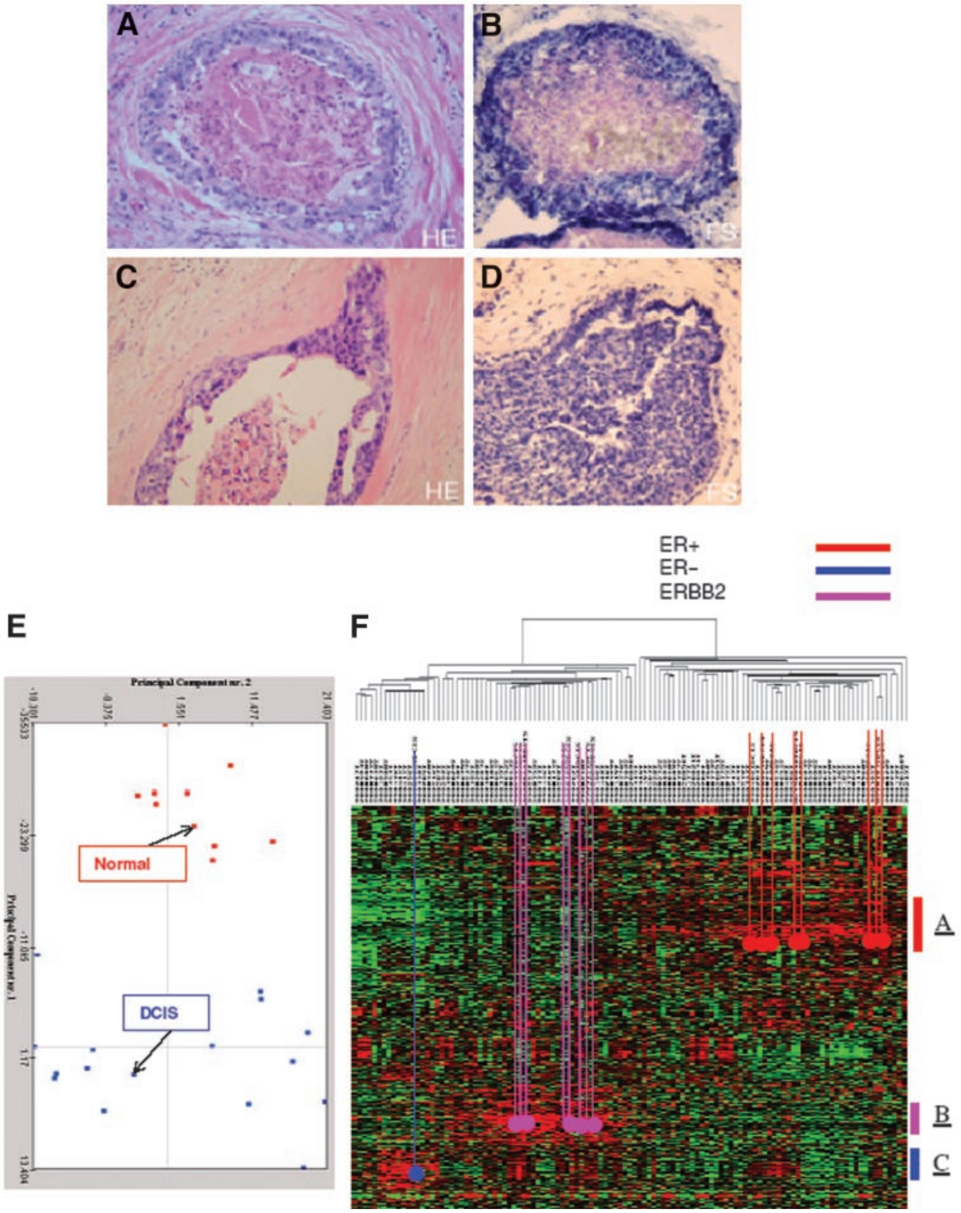

Fig. 3 *A–D*, examples of DCIS samples used in this study. Two samples are shown (*A*, *B*), and (*C*, *D*). The DCIS status of each sample was confirmed both by examination of paraffin-embedded H&E sections of samples (*A* and *C*, *HE*) and by examination of frozen cryosections (*B* and *D*, *FS*) of the actual sample that was processed for expression profiling. DCIS samples express the hallmark genes of invasive carcinoma subtypes. *E*, PCA analysis of normal tissue and DCIS tumors on the 367-gene set. *Red*, normal samples; *blue*, DCIS samples. *The vertical axis*, principal component 1; *the horizontal axis*, principal component 2. *F*, unsupervised clustering of DCIS and IDC tumors. *Lines*, DCIS samples. Eight of 17 DCIS samples cluster within the ERBB2+ group, 8 samples in the ER+ group, and 1 sample was in the ER− group. *Red*, ER+/Luminal-like; *pink*, ERBB2+; *blue*, ER−/Basal-like. *Colored bars to the right of the clustergram*: A, Luminal epithelial genes with ER; B, genes with ERBB2; C, Basal epithelial genes.

absence of the 17q genes, the SVM algorithm was nevertheless still able to clearly distinguish the DCIS samples into separate classes (Supplementary Information).[4] Furthermore, the gene expression profile data also confirmed that the ER+ DCIS samples expressed high levels of keratin 18 (a conventional luminal marker), whereas the single basal DCIS sample expressed high levels of keratin 17 (a conventional basal marker; Supplementary Information).[4] Taken collectively, these results suggest that DCIS tumors, similar to invasive tumors, can also be robustly divided into distinct molecular subtypes.

Identification of Common and Subtype-Specific Genes Involved in Tumor Progression. The observation that different molecular subtypes of breast cancer are associated with distinctive profiles of gene expression has led some investigators to propose that the former have arisen from distinct cells of origin (9). This hypothesis, if true, raises the possibility that the gene expression pathways controlling various aspects of tumor progression may be distinct between different subtypes. To identify sets of common and subtype-specific genes involved in tumor progression, we compared the expression profiles of normal tissue, DCIS, and invasive tumors belonging to either the ER+ or the ERBB2+ subtypes (We were unable to perform a similar analysis for the ER− tumors because only one DCIS tumor segregated within the ER− subtype, which is insufficient for analysis).

Using a nonparametric Wilcoxon test, we first identified genes that were significantly regulated during the transition from the nonmalignant tissue to DCIS for the ER+ and ERBB2+ subtypes. For the ER+ subtype, we identified 113 up-regulated genes and 310 down-regulated genes among non-

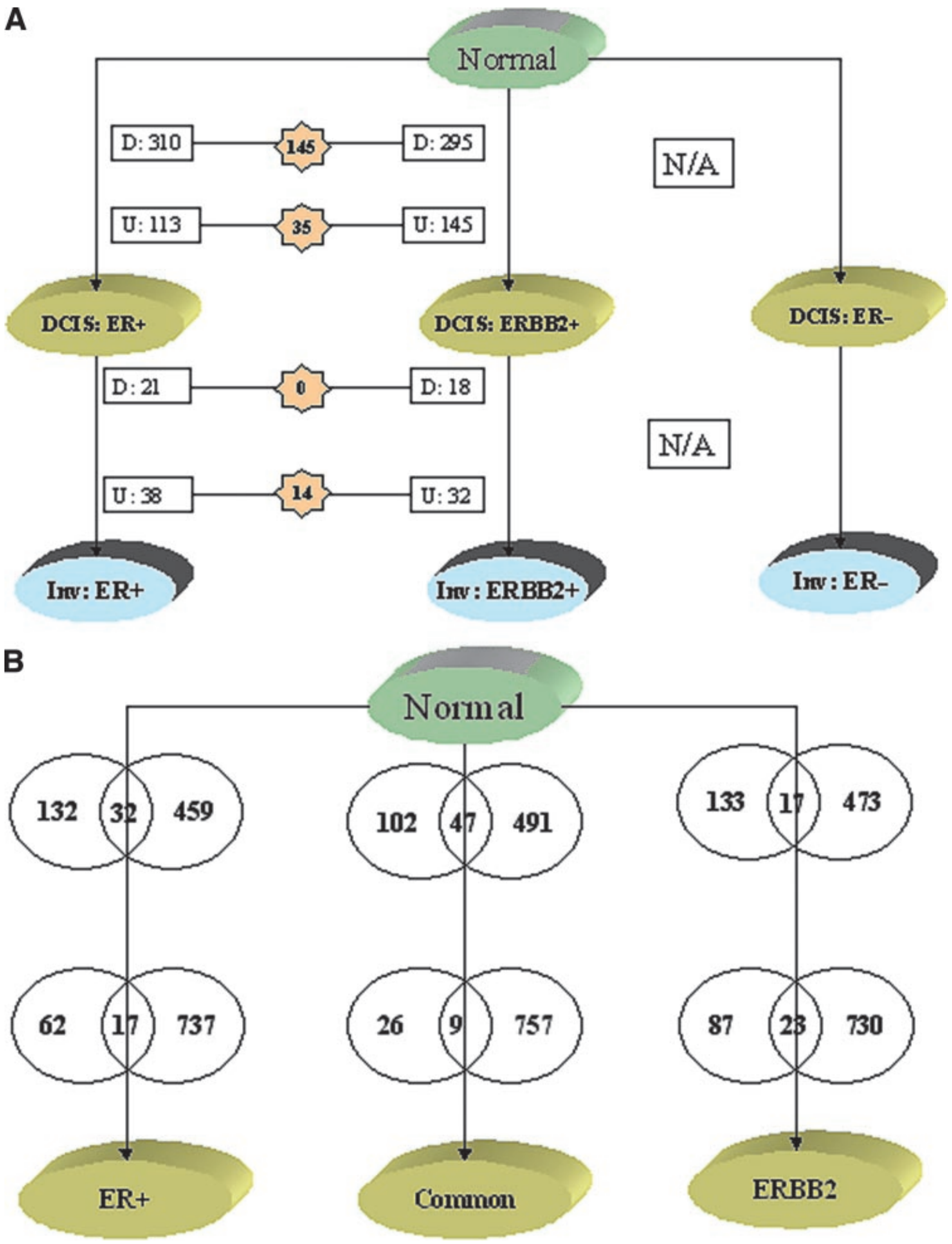

Fig. 4 *A,* summary of subtype-specific and commonly regulated genes for the ER+ and ERBB2+ molecular subtypes. *U,* up-regulated genes; *D,* down-regulated genes. For example, there are 113 genes up-regulated and 310 genes down-regulated during the normal→DCIS (ER+) transition. *Numbers in bold,* overlapping genes between the subtypes. (Inv, invasive; N/A, not available. *B,* a comparison of genes regulated during the normal/DCIS transition between our study and the study of Ma *et al.*, (15). *Numbers outside the intersecting areas,* genes that were found to not overlap between the studies; *numbers in the joint areas,* overlapping genes commonly identified in the both studies.

malignant breast tissues and ER+ DCIS ($P < 0.01$; 2-fold change cutoff). This analysis was then repeated for the ERBB2+ subtype, in which 145 up-regulated genes and 295 down-regulated genes were identified among normal breast tissues and ERBB2+ DCIS samples. A total of 180 genes (145 down-regulated and 35 up-regulated) were found in common between the ER+ and ERBB2+ subtypes. The results are summarized in Fig. 4*A*, and the full gene lists are provided in the Supplementary Information.[4]

A similar analysis was then performed for the transition from DCIS to invasive cancer for both molecular subtypes. We identified 59 genes as being significantly regulated (38 up-regulated and 21 down-regulated) for the ER+ subtype, and 50 genes for the ERBB2+ subtype (32 up-regulated and 18 down-regulated). This comparatively smaller number, compared with the normal-to-DCIS transitions, is consistent with previous findings that DCIS and invasive cancers are molecularly similar (15, 24–25). To confirm that this number was indeed significant, we performed a series of random permutation assays to estimate the potential rate of "false discoveries" (*i.e.*, the number of genes that would mistakenly be identified as being regulated because of random chance). For each of the subtypes (ER+ or ERBB2+), a total of 2000 randomly generated sets were created, each set comprising two groups of randomly selected invasive tumors of the same subtype, the numbers of tumors in each group being present in the same proportions as the actual

Table 1 Overlapping genes between this study and the study by Ma *et al.* compared with overlaps obtained by random permutation

	ER+	Common genes	*ERBB2+*
Down-regulated in normal/DCIS transition			
Random gene-set size	150*	150	150
No. of overlapping genes (99.9%)	11 (13†)	11 (13)	11 (13)
Actual experiment	32	47	17
Up-regulated in normal/DCIS transition			
Random gene-set size	79	35	110
No. of overlapping genes (99.9%)	9 (11)	6 (7)	11 (13)
Actual experiment	17	9	23

NOTE. See Results section (ref. 15; Ma *et al.*).

* For genes down-regulated in the normal/DCIS transition, all three types of samples (*ER+*, *ERBB2+*, and common genes) contained a similar number of genes (164, 149 and 150). Thus, a random-set size of 150 genes was used in all three types of samples.

† Numbers within parentheses indicate the maximum number of overlapping genes observed across 10,000 random sets.

DCIS-*versus*-invasive-tumor set. For each random set, the number of genes identified as being differentially regulated was determined, and this number was compared with the results obtained from the actual comparison. In 99.9% of the random sets, we found that substantially fewer number of genes were identified as being regulated compared with the actual DCIS/invasive comparison [ER+, 59 genes (actual DCIS *versus* invasive tumors) *versus* 22 genes (random permutation); ERBB2+, 50 *versus* 29 genes, respectively]. These results indicate that the transcriptional differences identified between DCIS and IDC samples are unlikely to have been entirely due to random chance ($P < 0.001$). Predictably, many of the genes up-regulated during the DCIS-to-invasive transition for both the ER+ and ERBB2+ subtypes included genes involved in wound healing/stromal remodeling (*COL1A1, COL5A1*; ref. 26), cellular adhesion (*OSF-2*; ref. 27), and endothelial growth (*nidogen 2, SPARC*; ref. 26) consistent with the presence of active tumor invasion and stromal remodeling.

Comparison of Common and Subtype-Specific Genes with an Independent Gene Expression Data Set. In a recent independent report (15), Ma *et al.*, using a combination of laser capture microscopy and cDNA microarrays, reported the identification of various sets of genes associated with human breast cancer progression. One unaddressed question, however, is the extent to which these genes are common or subtype specific, because the Ma *et al.* did not explicitly group their samples by their specific molecular subtype. To explore this question and to further validate our own findings, we filtered the Ma *et al.* data and derived from the original data set 767 up-regulated and 539 down-regulated genes associated with the transition from normal breast epithelia to DCIS (Supplementary Information).[4] We then compared our list of subtype-specific and commonly regulated genes for the normal-to-DCIS transition with this derived gene list. The overlaps between the two data sets are represented as Venn diagrams in Fig. 4*B*). We found that for all six comparisons (*i.e.,* ER+-specific; ERBB2+-specific; common to ER+ and ERBB2+; both up-regulated and down-regulated), there were significant overlaps between the genes found in our study and the gene set studied by Ma *et al.*, ranging from 11% (ERBB2+-specific down-regulated genes, 17 of 151) to 32% (commonly down-regulated genes, 47 genes of 149). To confirm that it would be highly unlikely for these overlaps to have occurred by random chance, we again performed a series of random permutation assays. Briefly, for each comparison, a total of 10,000 random gene sets were created, the total number of genes in each set being comparable with the sizes of the identified gene sets in Fig. 4*A*). As one example, for the normal-to-DCIS transition, a total of 149 genes were identified by Wilcoxon analysis as being commonly down-regulated in both the ER+ and the ERBB2+ subtypes. This set of 149 genes was then compared with 10,000 random gene sets in which each set contained 150 randomly selected genes. As shown in Table 1, the numbers of overlapping genes between our study and study by Ma *et al.* were consistently and significantly greater than overlaps created by 99.9% of all randomly generated gene sets ($P = 0.036$, *t* test for paired two samples). This suggests that the genes identified in our study (particularly those overlapping with the Ma *et al.* gene set) are likely to be of biological relevance, because they were independently identified in two different studies. Furthermore, through this comparison, we were also able to subdivide the original Ma *et al.* (2003) gene list into distinct sets of genes that were either common or pathway-specific. Table 2 lists the genes that were commonly regulated in both the ER+ and ERBB2+ subtypes, and that were commonly found in both data sets. A few of these genes and their potential implications for breast carcinogenesis are described in the Discussion.

DISCUSSION

In this study, we performed a molecular survey of invasive and preinvasive breast cancers from a predominantly Asian-Chinese patient population to investigate whether the various molecular subtypes of breast cancer, originally defined in Caucasian patient cohorts (9–11), could also be observed in an independent ethnic population. We performed this comparison because of the many reported clinical differences in breast cancer between Caucasian and Asian populations. For example, Caucasian and Asian-Chinese populations are known to differ in their absolute incidence rates of breast cancer and peak ages of incidence (Ref. 1; Supplementary Information[4]), their patterns of molecular abnormalities [*e.g.,* BRCA1 mutations, ERBB2 incidence, and chromosomal abnormalities (7, 8, 28, 29)], breast cancer risk factors [*e.g.,* IGFBP3 (6)], and hormonal physiology [*e.g.,* levels and profiles of estrogen metabolites (30)]. In addition, the specific environmental exposures encountered by these

Table 2 Genes that were commonly regulated for both ER+ and ERBB2+ molecular subtypes in the normal/DCIS transition

Probe_ID	Gene name	Gene symbol	UniGene
Down			
204294_at	*aminomethyltransferase* (*glycine cleavage system protein T*)	*AMT*	Hs.12
201012_at	*annexin A1*	*ANXA1*	Hs.78225
203324_s_at	*caveolin 2*	*CAV2*	Hs.139851
200985_s_at	*CD59 antigen p18–20* (antigen identified by monoclonal antibodies 16.3A5, EJ16, EJ30, EL32 and G344)	*CD59*	Hs.278573
221556_at	*CDC14 cell division cycle 14 homolog B* (*Saccharomyces cerevisiae*)	*CDC14B*	Hs.22116
202259_s_at	*hypothetical protein from BCRA2 region*	*CG005*	Hs.23518
203477_at	*collagen, type XV,* α*1*	*COL15A1*	Hs.83164
201289_at	*cysteine-rich, angiogenic inducer, 61*	*CYR61*	Hs.8867
201581_at	*hypothetical protein DJ971N18.2*	*DJ971N18.2*	Hs.169358
207761_s_at	*DKFZP586A0522 protein*	*DKFZP586A0522*	Hs.288771
203881_s_at	*dystrophin* (muscular dystrophy, Duchenne and Becker types)	*DMD*	Hs.169470
212730_at	*desmuslin*	*DMN*	Hs.10587
208370_s_at	*Down syndrome critical region gene 1*	*DSCR1*	Hs.184222
201540_at	*four and a half LIM domains 1*	*FHL1*	Hs.239069
218804_at	*hypothetical protein FLJ10261*	*FLJ10261*	Hs.26176
218823_s_at	*hypothetical protein FLJ20038*	*FLJ20038*	Hs.72071
203706_s_at	*frizzled homolog 7* (*Drosophila*)	*FZD7*	Hs.173859
209292_at	*inhibitor of DNA binding 4, dominant negative helix-loop-helix protein*	*ID4*	Hs.34853
208966_x_at	*interferon, gamma-inducible protein 16*	*IFI16*	Hs.155530
204773_at	*interleukin 11 receptor,* α	*IL11RA*	Hs.64310
206766_at	*integrin,* α10	*ITGA10*	Hs.158237
201656_at	*integrin,* α6	*ITGA6*	Hs.227730
201466_s_at	*v-jun sarcoma virus 17 oncogene homologue* (avian)	*JUN*	Hs.78465
209351_at	*keratin 14* (epidermolysis bullosa simplex, Dowling-Meara, Koebner)	*KRT14*	Hs.355214
202350_s_at	*matrilin 2*	*MATN2*	Hs.19368
209583_s_at	*antigen identified by monoclonal antibody MRC OX-2*	*MOX2*	Hs.79015
202431_s_at	*v-myc myelocytomatosis viral oncogene homolog* (avian)	*MYC*	Hs.79070
209272_at	*NGFI-A binding protein 1* (*EGR1 binding protein 1*)	*NAB1*	Hs.107474
209289_at	*nuclear factor I/B*	*NFIB*	Hs.33287
211671_s_at	*nuclear receptor subfamily 3, group C, member 1* (*glucocorticoid receptor*)	*NR3C1*	Hs.75772
218589_at	*purinergic receptor P2Y, G-protein coupled, 5*	*P2RY5*	Hs.123464
203131_at	*platelet-derived growth factor receptor,* α *polypeptide*	*PDGFRA*	Hs.74615
203097_s_at	*PDZ domain containing guanine nucleotide exchange factor* (*GEF*) *1*	*PDZGEF1*	Hs.373588
218319_at	*pellino homolog 1* (*Drosophila*)	*PELI1*	Hs.7886
207943_x_at	*pleiomorphic adenoma gene-like 1*	*PLAGL1*	Hs.75825
201578_at	*podocalyxin-like*	*PODXL*	Hs.16426
209147_s_at	*phosphatidic acid phosphatase type 2A*	*PPAP2A*	Hs.406043
215707_s_at	*prion protein* (*p27–30*) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia)	*PRNP*	Hs.74621
221523_s_at	*Ras-related GTP binding D*	*RRAGD*	Hs.238679
200937_s_at	*ribosomal protein L5*	*RPL5*	Hs.180946
202037_s_at	*secreted frizzled-related protein 1*	*SFRP1*	Hs.7306
200795_at	*SPARC-like 1* (*mast9, hevin*)	*SPARCL1*	Hs.75445
204955_at	*sushi-repeat-containing protein, X-linked*	*SRPX*	Hs.15154
216037_x_at	*transcription factor 7-like 2* (T-cell specific, HMG-box)	*TCF7L2*	Hs.348412
208944_at	*transforming growth factor,* β *receptor II* (*70/80 kDa*)	*TGFBR2*	Hs.82028
204731_at	*transforming growth factor,* β *receptor III* (*betaglycan, 300 kDa*)	*TGFBR3*	Hs.342874
213900_at	*Friedreich ataxia region gene X123*	*X123*	Hs.77889
UP			
204170_s_at	*CDC28 protein kinase regulatory subunit 2*	*CKS2*	Hs.83758
212057_at	*KIAA0182 protein*	*KIAA0182*	Hs.75909
210519_s_at	*NAD*(*P*)*H dehydrogenase, quinone 1*	*NQO1*	Hs.406515
208874_x_at	*protein phosphatase 2A, regulatory subunit B*′ (*PR 53*)	*PPP2R4*	Hs.400740
208734_x_at	*RAB2, member RAS oncogene family*	*RAB2*	Hs.78305
200832_s_at	*stearoyl-CoA desaturase* (δ*-9-desaturase*)	*SCD*	Hs.119597
209218_at	*squalene epoxidase*	*SQLE*	Hs.71465
201689_s_at	*tumor protein D52*	*TPD52*	Hs.2384
36936_at	*tissue specific transplantation antigen P35B*	*TSTA3*	Hs.404119

NOTE. See Fig. 4B. “Down” indicates down-regulation in the normal/DCIS transition.
Abbreviation: ID, identification.

two ethnic groups may also be distinct. Taken collectively, these clinical differences raise the possibility that breast tumors in the two ethnic groups may also be distinct at the gene expression level. This hypothesis is supported by several independent findings. First, many molecular differences have already been characterized between these different ethnic groups at the DNA or protein level, and it has been shown that distinct and reproducible gene expression differences can result from subtle and complex patterns of genetic variation (31–33). Second, a recent study, published while the manuscript for this article was in review, has reported the successful identification of gene expression differences in breast tumors between Caucasian and African-American patients (34), validating our basic hypothesis that significant gene expression differences can be identified in breast cancers from different racial groups.

Although the molecular subtypes in our patient population were independently derived using an unsupervised analysis, these subtypes were, for the most part, highly comparable with similar subtypes observed in other studies (*e.g.,* ER+ to Luminal, ER− to Basal; refs. 9–11). It should also be noted that these breast tumor expression data sets are distinct in multiple ways, including (*a*) choice of patient population, (*b*) sample handling protocols, (*c*) scoring pathologist, and (*d*) choice of array technology and probe sets (two-color *versus* single color). Despite these methodological differences, our results suggest that, at a first approximation, breast cancers between the ethnic groups are remarkably similar at the molecular level. We also note that we did observe a few possible differences, such as the absence of a "Luminal C" subtype in the Asian population. However, further work will have to be performed to determine whether these apparent differences are truly due to genetic or environmental differences between the ethnic groups or whether the differences are merely experimental artifacts due to the different array technology platforms used in the studies.

Besides invasive breast cancers, we also profiled a series of DCIS, which represent the earliest malignant breast lesion detectable by conventional histopathology. Although DCIS cancers have long been recognized as the major precursor to invasive breast cancer, some studies have also suggested that DCIS cancer may also be distinct from invasive cancers in certain respects. For example, retrospective reports have shown that the majority of low-nuclear-grade DCIS undergo a long clinical evolution to invasive cancer (35–37), which may indicate that additional genetic events must occur before they become invasive. We found that the gene expression profiles of DCIS cancers are highly similar to their invasive counterparts, and that they exhibited robust expression of many "hallmark" subtype-specific gene expression signatures. These findings suggest, for the first time, that the molecular subtypes of breast cancer can be discerned even at the preinvasive stage of carcinogenesis. Interestingly, of the 17 DCIS cancers that we profiled, 16 belonged to the ER+ or ERBB2+ molecular subtypes, with only one DCIS cancer belonging to the ER− subtype. Histopathological studies have shown that ERBB2+ cancers seem to be found much more often in DCIS compared with invasive cases (38), which is consistent with our gene expression profiling data. One possible hypothesis to explain this difference might be that tumors of the ER− variety may be associated with an extremely transient DCIS stage compared with ER+ or ERBB2+ tumors.

Finally, by integrating the expression profiles of normal breast tissue, DCIS, and IDCs belonging to the ER+ and ERBB2+ subtypes, we were able to define various sets of genes that were regulated in a common and subtype-specific manner during the normal/DCIS/IDC transitions. We then validated several of these genes by showing that they were also commonly identified in a separate, related but not identical, study. Of the various gene sets, perhaps the most interesting group comprised genes that were commonly regulated in both the ER+ and ERBB2+ tumorigenic pathways and that were found to overlap in both studies (Table 2). We speculate that an in-depth study of these individual genes may provide important insight into the pathogenesis of breast cancer. Indeed, many of the genes in this list could be associated with various cellular functions associated with carcinogenesis, such as cellular proliferation (*CDC14B*, *CKS2*, *JUN*, *MYC*, *PLAGL1*), biosynthesis and energy utilization (*RPL5*, *NQO1*), cell-to-cell communication (*ANXA1*, *ITGA6*, *ITGA10*), and cellular signaling (*NR3C1*, *PDGFRA*, *PPAP2A*, *PPP2R4*). One particularly interesting finding was that several genes exhibiting common regulation in both subtypes were involved in the modulation of the Wnt signaling pathway (*FZD7*, *TCF7L2*, and *SFRP1*). Previous studies have reported the involvement of the Wnt pathway in breast cancer (11, 39), and our data suggest that misregulation of this pathway may be required in both subtypes for breast carcinogenesis. If so, small molecules or compounds that selectively regulate the Wnt pathway may function as promising therapeutic or chemopreventive agents for breast cancer. Addressing this intriguing issue will be a promising area for future research efforts.

ACKNOWLEDGMENTS

The National Cancer Centre Tissue Repository is acknowledged for their contributions toward this project. K. M. Hui (National Cancer Centre, Singapore) is thanked for his generous and continual support and encouragement. We also acknowledge and thank Hong Ga Sze, Wee Siew Bok, and Wong Chow Yin (National Cancer Centre, Singapore) for providing surgical samples.

REFERENCES

1. Chia KS, Seow A, Lee HP, Shanmugaratnam K. Cancer incidence in Singapore, 1993–1997. In: Singapore cancer registry report 5. Singapore: Singapore Cancer Registry, 2000.

2. Tavassoli FA, Schnitt SJ. Pathology of the breast. New York: Elsevier; 1992.

3. Wiencke JK. Opinion: impact of race/ethnicity on molecular pathways in human cancer. Nat Rev Cancer 2004;4(1):79–84.

4. Giuliano AE. Breast. In: Tierney LM, McPhee SJ, Papadakis MA, editors. Current medical diagnosis and treatment, 37th ed. Stamford: Appleton and Lange; 1998. p. 666–90.

5. Ferlay J, Parkin DM, Pisani, P., editors. GLOBOCAN: cancer incidence and mortality worldwide. International Agency for Research on Cancer (IARC) Cancer Base 3. Lyon, France: IARC; 1998.

6. Yu H, Jin F, Shu XO, et al. Insulin-like growth factors and breast cancer risk in Chinese women. Cancer Epidemiol Biomark Prev 2002; 11:705–12.

7. Fung LF, Wong N, Tang N, et al. Genetic imbalances in pT2 breast cancers of southern Chinese women. Cancer Genet Cytogenet 2001;124: 56–61.

8. Ford CE, Tran D, Deng YM, Ta VT, Rawlinson WD, Lawson JS. Mouse mammary tumor virus-like gene sequences in breast tumors of Australian and Vietnamese women. Clin Cancer Res 2003;9:1118–20.

9. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumors. Nature (Lond) 2000;406:747–52.

10. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98:10869–74.

11. Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA 2003;100(18):10393–8.

12. Gruvberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res 2001;61:5979–84.

13. Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. N Engl J Med 2001;344:539–48.

14. Hedenfalk IA, Ringner M, Trent JM, Borg A. Gene expression in inherited breast cancer. Adv Cancer Res 2002;84:1–34.

15. Ma XJ, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci USA 2003; 100(10):5974–9.

16. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.

17. Vapnik V. Statistical learning theory. New York: Wiley; 1998.

18. Lamb J, Ramaswamy S, Ford HL, et al. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell 2003;114:323–34.

19. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 2001;98:5116–21.

20. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Wash DC) 1999;286:531–7.

21. Martin KJ, Kritzman BM, Price LM, et al. Linking gene expression patterns to therapeutic groups in breast cancer. Cancer Res 2000;60: 2232–8.

22. Dressman MA, Baras A, Malinowski R, et al. Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. Cancer Res 2003;63(9):2194–9.

23. Yu K, Lee CH, Tan PH, et al. Classifying the estrogen receptor status of breast cancers by expression profiles reveals a poor prognosis subpopulation exhibiting high expression of the ERBB2 receptor. Hum Mol Genet 2003;12(24):3245–58.

24. Porter DA, Krop IE, Nasser S, et al. A SAGE (serial analysis of gene expression) view of breast tumor progression. Cancer Res 2001; 61(15):5697–702.

25. Porter D, Lahti-Domenici J, Keshaviah A, et al. Molecular markers in ductal carcinoma in situ of the breast. Mol Cancer Res 2003;1(5): 362–75.

26. St Croix B, Rago C, Velculescu V, et al. Genes expressed in human tumor endothelium. Science (Wash DC); 2000;289:1197–202.

27. Takeshita S, Kikuno R, Tezuka K, Amann E. Osteoblast-specific factor 2: cloning of a putative bone adhesion protein with homology with the insect protein fasciclin I. Biochem J 1993;294(Pt 1):271–8.

28. Suter NM, Ray RM, Hu YW, et al. BRCA1 and BRCA2 mutations in women from Shanghai China. Cancer Epidemiol Biomark Prev 2004;13:181–9.

29. Choi DH, Shin DB, Lee MH, et al. A comparison of five immunohistochemical biomarkers and HER-2/neu gene amplification by fluorescence in situ hybridization in White and Korean patients with early-onset breast carcinoma. Cancer (Phila) 2003;98:1587–95.

30. Ursin G, Wilson M, Henderson BE, et al. Do urinary estrogen metabolites reflect the differences in breast cancer risk between Singapore Chinese and United States African-American and White Women? Cancer Res 2001;61:3326–3329.

31. Jin W, Riley R, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 2001;29:389–95.

32. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science (Wash DC) 2002; 296:752–5.

33. Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse, and man. Nature (Lond) 2003;422: 297–302.

34. Selaru FM, Yin J, Olaru A, et al. An unsupervised approach to identify molecular phenotypic components influencing breast cancer features. Cancer Res 2004;64:1584–8.

35. Page D, Dupont W, Rogers L, Landenberger M. Intraductal carcinoma of the breast: follow-up after biopsy only. Cancer (Phila) 1982; 49:751–78.

36. Betsill WLJ, Rosen PP, Lieberman PH, Robbins GF. Intraductal carcinoma. Long-term follow-up after treatment by biopsy alone. JAMA 1978;239:1863–7.

37. Rosen P, Braun D, Kinne D. The clinical significance of pre-invasive breast carcinoma. Cancer (Phila) 1980;46:919–25.

38. Barnes DM, Bartkova J, Champlejon RS, Gullick WJ, Smith PJ, Millis R. Overexpression of c-erbB2 oncoprotein: why does this occur more frequently in ductal carcinoma in situ than in invasive mammary carcinoma and is this of prognostic significance? Eur J Cancer 1992; 28:644–8.

39. Li Y, Welm B, Podsypanina K, et al. Evidence that transgenes encoding components of the Wnt signaling pathway preferentially induce mammary cancers from progenitor cells. Proc Natl Acad Sci USA 2003;100(26):15853–8.

American Association for Cancer Research

Clinical Cancer Research

Conservation of Breast Cancer Molecular Subtypes and Transcriptional Patterns of Tumor Progression Across Distinct Ethnic Populations

Kun Yu, Chee How Lee, Puay Hoon Tan, et al.

Clin Cancer Res 2004;10:5508-5517.