Molecular classification of primary breast tumors possessing distinct prognostic properties

Kyoko Iwao¹², Ryo Matoba¹, Noriko Ueno¹, Akiko Ando², Yasuo Miyoshi², Kenichi Matsubara¹, Shinzaburo Noguchi² and Kikuya Kato¹,*

¹Taisho Laboratory of Functional Genomics, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0101, Japan and ²Department of Surgical Oncology, Osaka University Medical School, 2-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

Received October 30, 2001; Revised and Accepted November 12, 2001

The natural progression of breast cancer differs greatly between patients; the precise prediction of this disease course will improve the efficacy of therapeutics. Gene expression profiling may elucidate the undiscovered biological variations between seemingly similar cancers, leading to a new cancer classification system valuable in accurate diagnosis. The expression levels of 2412 genes, derived from 98 cancer samples, were precisely recorded by a high throughput RT–PCR technique, adapter-tagged competitive PCR. Subsequent cluster analysis revealed a molecular profile, correlating with estrogen receptor levels and the presence of lymph node metastases. We analyzed 301 cancer samples for the expression patterns of 21 genes critical in this categorization. The classification of the samples into three major groups was verified utilizing principal component analysis. This molecular classification system correlated significantly with early recurrence, independent of lymph node status. This malignant potential is associated with the expression levels of a group of genes, which comprise a set of candidates potentially useful in diagnostic prediction. These genes and the associated control mechanisms may also be effective therapeutic targets.

INTRODUCTION

The molecular characterization of cancers analyzes individual genes, including oncogenes and tumor suppressor genes. Gene expression profiling, a large-scale analysis of gene expression, has created new possibilities for the molecular characterization of cancer (1–3); exploration of marker genes associated with particular tumor characteristics is made possible. In addition, cancer is now understood as the coordinated action of a group of genes, not single mutations. cDNA microarray analysis of diffuse large B-cell lymphoma has facilitated the prediction of new tumor classes exhibiting distinct prognoses (1). Similar studies are expected to identify new cancer subclasses possessing unique characteristics, directly related to clinical applications. This method of investigation should be expanded to solid tumors of the breast, bladder and gastrointestinal tract, diseases with higher occurrence and greater social impact.

However, unlike leukemias and lymphomas, solid tumors are a mixture of various cell types. The composition of cell types varies between individual cancers. Current technology does not allow the purification of large quantities of each cell type from limited tissue samples. In addition, the characters of solid tumors are heavily influenced by the surrounding cellular environment. These circumstances make the analysis of solid tumors more technically demanding.

To circumvent these difficulties, we analyzed total RNA purified from whole tissue. In addition, we used high throughput RT–PCR [adapter-tagged competitive PCR (ATAC–PCR)] (4,5), instead of conventional cDNA microarrays (6,7). This technique detects trivial changes in gene expression among a mixture of RNA derived from various cell types, including the cancer cells.

Currently, breast cancer is one of the major causes of death in women. Biochemical studies are the most advanced among solid tumors, levels of estrogen receptor (ER), for example, delineate clinically distinct groups. However, the prediction of lymph node metastasis and recurrence is the most critical determination in the treatment of breast cancer. To date, none of the existing diagnostic methods is satisfactory (8–11). In this report, we describe the identification of a group of genes, which classify breast cancer into categories possessing distinct prognostic properties. These genes are strong candidates for the improvement of diagnostic prediction as well as possible therapeutic targets.

RESULTS

Library-sequencing and ATAC–PCR assay

Gene expression patterns in breast cancer cells were surveyed by preliminary EST sequencing. This analysis used a 3′ end-directed cDNA library representing the mixture of RNAs purified from 12 breast cancer tissues (12). A total of 7572 EST clones were sequenced; the 2379 genes of greatest abundance were selected, prioritizing known genes. We designed...
PCF primers for a total of 2412 genes, including additional 33 genes involved in breast cancer. This gene set exclusively included genes expressed in breast cancer. This specificity provides an advantage over the more universal sets, such as those selected from the UniGene database, which include genes not detected in breast cancer. We analyzed tissue samples derived from multiple sites, including 98 female breast cancers and 10 normal samples. The subsequent ATAC–PCR analysis yielded a data matrix consisting of 2412 genes × 108 tissue samples (see supplementary material).

Following data processing, we performed hierarchical cluster analysis (13) to phenotypically classify the various tissue samples. However, use of the 21 genes did not yield meaningful groupings; this analysis could not separate normal and malignant tissues. This result is not surprising, however, as the source tissues were mixtures of both carcinomas and normal cell types.

Selection of genes
Our analysis aims to classify individual breast cancer tissues into biologically distinct groups, related to clinical responses and characteristics (1–3). For analysis, we selected those genes whose expression patterns were highly variable among the tumor samples compared with that among normal samples; these genes may reflect variations among tumors with differing prognostic courses. We performed hierarchical cluster analysis on several sets of such genes, selected by the criteria described in the methods. Utilizing a variety of data sets, this analysis yielded clusters that correlated well with ER levels, consisting of specimens positive for ER expression, ER-negative samples and those containing both ER-positive and -negative samples. The results of a cluster analysis utilizing 152 genes revealed that the normal samples are grouped together; therefore, the selected genes are expressed similarly in normal samples (Fig. 1). This cluster model also correlates gene expression with lymph node metastasis; clusters 7–8 possess fewer lymph node metastases than the others. Additional clinical parameters, such as histological grade and tumor size, did not correlate with the clusters.

This cluster model is characterized by two groups of expressed genes. The 'ER+' gene grouping correlates with ER levels. 'ER–' genes specify clusters lacking ER; these genes are primarily immunoglobulins, suggesting B-cell infiltration into these tissues.

Utilizing the above cluster model, we then identified genes that may serve to classify cancer samples. As excessive numbers of genes could obscure the characteristics of the clusters (14), we utilized only a small number of genes. Testing a limited number of genes is also advantageous for accurate, expeditious diagnoses. To obtain genes characterizing clusters 7–8, we selected genes for which the averages of the ER+ and ER– groups were markedly different. Cluster analysis of the selected 21 genes yielded a classification scheme (Fig. 1) that correlated gene expression with both ER expression levels and the presence of lymph node metastasis.

Validation of the classification
These results were validated by an additional set of 203 samples. Like the original 98 samples, these samples were divided into a group having more lymph node metastasis positive samples (46%) and a group having less such samples (32%). Details of the validation study are described in the supplementary document. The classification scheme described above was consistent with the result of a cluster analysis of 301 breast cancer samples including the original 98 and additional 203 samples (Fig. 2). The samples were classified into three major groups (designated as 1, 2 and 3), of which the first two were subdivided into subcategories a and b. The genes analyzed were grouped into two categories, A and B. The major grouping of samples was created on the basis of group A expression levels; subgrouping relied on group B gene expression. Group A and B likely correspond to ‘ER+' and ‘ER–' genes, respectively (Table 1). The expression of Carbonic anhydrase XII (CA XII), appearing in both group A and ‘ER+' genes, has been linked to tumor invasiveness (15). Immunoglobulin genes comprised the majority of the group B, ‘ER–' genes.

To confirm the independence of lymph node metastasis status on the expression of ER in the tumors, we analyzed gene expression patterns in 175 node-negative breast cancers. These samples were classified into the three groups (data supplied in the supplementary material).

Principal component analysis
The validity of this cluster analysis was verified utilizing principal component analysis (PCA) (16). This technique classifies each cancer sample into a smaller dimensional space than represented by the original 21 dimensions. The first three principal components of the three-dimensional scatter plot (Fig. 3) capture 43.9% of the expression level variation for the 21 genes. The members of each group are closely located by PCA, confirming the cluster analysis. However, as the demarcation of each group in PCA is not clear, cluster analysis is better suited for diagnostic purposes. This resulting classification is similar to that proposed by previous microarray experiments; the microarray classification system, however, could not identify group 2, consisting of both ER+ and ER– samples, or subdivide the larger groups into more specific subgroups (6).

Correlation of the clinical parameters
Correlation of the clinical parameters with the classification system demonstrates significant differences between the various groups with respect to lymph node status and histological grade (Table 2). Metastases appear less frequently in group 1 than in groups 2 and 3. Surgical operation dates for the 119 samples examined dated back to 1996, with available records of clinical course, including early recurrence and subsequent death. Early recurrence occurred more frequently in groups 2 and 3 than in group 1. In addition, no death resulting from early recurrence is observed in group 1, in contrast to eight and two deaths in groups 2 and 3, respectively. Furthermore, we examined the correlation of the molecular grouping with erbB2 protein overexpression, a risk factor of breast cancer malignancy (17). The levels of erbB2 overexpression in tumors were significantly higher in group 2 than in either group 1 or 3 (Fig. 2 and Table 2).

In 110 patients with primary invasive breast cancers, we performed Cox regression analysis (18) to evaluate clinical parameters as possible risk factors for early recurrence, which would distinguish groups 1, and 2 and 3 (Table 3). Univariate
analysis revealed that our molecular groupings provided the highest risk assessment. Four-parameter (molecular grouping, lymph node status, ER status and histological grade) multivariate analysis indicated that molecular grouping and lymph node status were independent risk factors.

DISCUSSION

Our results clearly demonstrate that group A genes are linked to the malignant potential of breast cancers; low expression levels of these genes correlate with a poor prognosis. Many of these genes are known to be involved in cancer-related functions. CAXII (15) influences tumor invasiveness, conferring a poor prognosis. Rab11a is a member of the RAS oncogene superfamily (19). α1-antichymotrypsin (ACT) (20) complexes with prostate-specific antigen (PSA), a prognostic indicator in node-negative breast cancer (21). Expression levels of cytochrome c oxidase subunit VIc (COSVIc) increase in prostate cancer (22). Stanniocalcin 2 gene was included in the list of genes induced by 17β estradiol (23). These group A genes are likely to be governed by similar transcriptional control; activation of the controlling transcriptional machinery may effectively alter group 2 and 3 expression patterns to become a group 1-type cancer. This result suggests that the human X-box binding protein 1 (hXBP-1) (24) transcription factor may be a target for further studies.

Figure 1. Cluster analysis of the selected 152 genes from 108 experimental samples (98 breast cancers and 10 normal mammary tissues). Genes (vertical axis) and samples (horizontal axis) were grouped by individual gene expression patterns. The top line represents the truncation of clustering at the level of 12 clusters. Estrogen receptor level (ER) and lymph node metastasis (LN) for each case are designated by bars according to the following color scheme. The blue bars indicate normal breast tissue samples; green bars designate ER positive cancers; red bars indicate ER negative samples. The green bars indicate lymph node metastasis-negative samples; yellow bars symbolize node metastasis-positive (n = 1–3) samples; red bars indicate node positive cancers (n ≥ 4). ‘ER+’ genes designates those genes likely to be highly expressed in ER+ cancers; ‘ER−’ genes demonstrate high expression in ER− cancers.
The present work used ATAC–PCR, an advanced form of quantitative PCR. The data obtained by ATAC–PCR provides more quantitative results than those resulting from microarray analysis. The data obtained by ATAC–PCR suggested that expression levels of the majority of genes were too variable for accurate cancer classification. Only those genes with marked differences were selected for classification. This variation results from the varying fractions of stromal tissue contained within the tumor samples. The observed expression levels are the sum of the tumor and stromal tissues, respectively, which is affected by the relative quantities of the two tissues. This problem may be overcome by use of laser capture microdissection (25), purifying cancer and surrounding tissue separately.

Another advantage of the current approach is the ease with which this established and versatile technique can be adapted for extended studies. Real-time PCR, such as the TaqMan assay, is characterized by the ability to perform precise quantitation and analyze large sample numbers. As ATAC–PCR and real-time PCR share many features in common, adaptation of this assay to a large number of samples is possible. In addition, as the analysis of only 21 genes is necessary for categorization,
verification is accessible to many laboratories. We stress that quantitative PCR is more advantageous than the use of microarrays, a recently developed technique. First, experiments can be easily controlled using adequate standards. Secondly, the assay requires a far smaller amount of RNA. Thirdly, the PCR assay of selected genes costs much less than DNA microarrays which are likely to require at least several hundred spotted genes for diagnosis (6).

It should be noted that molecular grouping is determined by the behavior of a group of genes, not by the behavior of single genes. In particular, group 2 is likely to be determined by a sum of expression levels of 18 genes. Expression levels of single genes cannot construct grouping correlated with prognosis. This may explain why the quest for molecular diagnostic markers so far has not been so successful.

Accurate prediction of lymph node metastases and recurrence is pivotal in the treatment of breast cancer. Existing diagnostic indicators are not currently effective (8–11). Gene expression is more likely to reflect intrinsic characters of cancer tissues, and may be more advantageous than previously established methods. A new prognostic indicator, obtained by the above classification, should complement current diagnoses, upon verification of these results in larger clinical studies. Such a method could more accurately predict

Table 1. Information of selected 21 genes for data normalization

<table>
<thead>
<tr>
<th>GS (ID) number</th>
<th>GenBank accession no.</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6601 X60111</td>
<td>Motility related protein1 (hMRP-1)</td>
</tr>
<tr>
<td>2</td>
<td>2632 X03473</td>
<td>Histone H1 (0)</td>
</tr>
<tr>
<td>3</td>
<td>7288 AB021868</td>
<td>Homo sapiens PIAS3 mRNA for protein inhibitor of activated STAT3</td>
</tr>
<tr>
<td>4</td>
<td>1957 D14034</td>
<td>Zinc-α2-glycoprotein</td>
</tr>
<tr>
<td>5</td>
<td>7264 BC003559</td>
<td>α-1-Antichymotrypsin</td>
</tr>
<tr>
<td>6</td>
<td>7012 AL133281</td>
<td>EST</td>
</tr>
<tr>
<td>7</td>
<td>6770 AA865260</td>
<td>Stanniocalcin 2</td>
</tr>
<tr>
<td>8</td>
<td>6784 AJ293618</td>
<td>Human hepatocellular carcinoma associated protein (JCL-1)</td>
</tr>
<tr>
<td>9</td>
<td>2471 AF086028</td>
<td>EST</td>
</tr>
<tr>
<td>10</td>
<td>690 AL161952</td>
<td>Glutamine synthetase</td>
</tr>
<tr>
<td>11</td>
<td>6711 BC001663</td>
<td>Hypothetical protein MGC2771</td>
</tr>
<tr>
<td>12</td>
<td>7001 AF051882</td>
<td>Homo sapiens carbonic anhydrase precursor (CA XII)</td>
</tr>
<tr>
<td>13</td>
<td>1176 BC000187</td>
<td>p6 = cytochrome c oxidase subunit Vlc homolog/COSVlc/prostatic carcinoma upregulated gene</td>
</tr>
<tr>
<td>14</td>
<td>1472 X55543</td>
<td>Homo sapiens hXBP-1 transcription factor (TREB)</td>
</tr>
<tr>
<td>15</td>
<td>7583 AF000231</td>
<td>Homo sapiens rab11a GTPase</td>
</tr>
<tr>
<td>16</td>
<td>7116 AK026956</td>
<td>Homo sapiens HSPC055</td>
</tr>
<tr>
<td>17</td>
<td>7715 AB011159</td>
<td>NCK-associated protein 1 (KIAA0587 protein)</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>7435 AF132948</td>
<td>Homo sapiens CGI-14 protein</td>
</tr>
<tr>
<td>19</td>
<td>2307 X51754</td>
<td>Human mRNA for Ig lambda-chain</td>
</tr>
<tr>
<td>20</td>
<td>2828 X51755</td>
<td>λ-Immunoglobulin constant region</td>
</tr>
</tbody>
</table>

*Gene group in Figure 2.

Figure 3. PCA of gene expression in breast cancers. The variation in expression levels of 21 genes is reduced to a three-dimensional space. The first, second and third factor scores are plotted in the x, y and z axis, respectively. Each sphere represents a single case of breast cancer; the color indicates the group affiliation: green, group 1 of Figure 2; red, group 2; blue, group 3.
prognosis using biopsy samples without surgical operation, contributing to more appropriate treatment decisions.

MATERIALS AND METHODS

Samples

At surgery, tumor specimens were obtained from 301 female Japanese breast cancer patients, during a period from November 1996 to September 2000 (Table 2). Normal breast tissues were also obtained from 10 individuals. Specimens were snap frozen in liquid nitrogen and kept at –80°C until use. Of these 301 patients, 110 who underwent surgery due to invasive breast cancer during the period from November 1996 to March 1998 [median follow-up, 48 months (range 39–55 months)] were recruited in the study where the relationship between patients prognosis and the gene expression profile assessed by ATAC–PCR was examined. No patients were lost to follow-up during this follow-up period. Information on clinical parameters was obtained shortly after the surgery, and noted during prospective follow-up. These processes followed the ethical guidelines set by the Ministry of Education, Science and Culture.

Selection of genes and ATAC–PCR assay

Total RNA was purified from clinical materials utilizing Trizol reagent (Gibco BRL). A 3’ end cDNA library was constructed using a mixture of RNA from 12 malignant samples, as described by Matoba et al. (12). 2379 unique sequences (1928 known genes, 161 EST-matched genes and 290 unknown genes, searched against GenBank release no. 123.0) were obtained from 7572 sequencing reactions. Utilizing software developed in our laboratory, we designed PCR primers for the ATAC–PCR reaction for a total of 2412 genes including additional 33 genes such as Hypoxia-Inducible Factor 1α (HIF-1α) gene. The ATAC–PCR reactions were performed as described with the following modification (12). The protocol and the primary data may be obtained at http://love2.aist-nara.ac.jp/laboratory/index_frame.html. Briefly, seven adapters were used, two of which were assigned to the mixture of the 10 malignant tissues used to create the library. Each reaction mixture contained 10 portions of the MB-1 control, two
portions of the MB-2 control, and three portions of each sample, where one portion is equivalent to 6 ng total RNA. Amplified products were separated by an ABI 3700 DNA analyzer. We then calculated the relative expression levels as compared to the control. The obtained data matrix was normalized by the median and converted to a logarithmic scale. The detailed protocols are available in the supplementary document. The selected gene expression patterns were confirmed utilizing real-time RT–PCR with a LightCycler (Roche Molecular Biochemicals) (data not shown).

Six genes exhibiting expression patterns similar to the median, including glyceraldehyde-3-phosphate dehydrogenase (GAPDH), were selected from the original pool of 2412 genes. For the validation studies, the relative expression levels of these six genes were measured by ATAC–PCR; the average of the six genes was used to normalize the data, instead of the median.

**Enzyme immunoassay for ER**

We measured ER protein levels in breast cancers using an enzyme immunoassay kit (Abbott Research Laboratories) according to the manufacturer’s instructions. The detection limit for ER was 5 fmol/mg total protein.

**Immunohistochemistry for ErbB2 protein**

Paraffin sections of tumor tissues were subjected to hematoxylin and eosin staining and immunohistochemical staining of erbB2 protein by the avidin–biotin–peroxidase method using a rabbit polyclonal anti-erbB2 antibody (Nichirei Co. Ltd). The immunostaining results of erbB2 were considered positive when >10% of the cancer cells were stained regardless of staining intensity. Only those cancer cells with membrane staining for erbB2 were counted.

**Statistical analysis**

Genes lacking more than 30 values were excluded from the following analysis. We calculated the variance among both the malignant and normal samples for each gene. 407 genes with a tumor to normal variance ratio exceeding 1.18 were selected for cluster analysis. This ratio (\(tgr\)) was described in the following equation:

\[
tgr = \frac{\sum_{i=1}^{n} (\chi_{i} - \overline{\chi})^2}{\sum_{i=1}^{m} (\chi_{i} - \overline{\chi})^2 + \sum_{i=m+1}^{n} (\chi_{i} - \overline{\chi})^2}
\]

\(\chi_{i}\), gene expression level; \(1 \leq i \leq m\), ER-positive samples; \(m + 1 \leq i \leq n\), ER-negative samples; \(\overline{\chi}_{tgr}\), average of expression levels of ER-positive samples; \(\overline{\chi}_{ERN}\), average of expression levels of ER-negative samples; \(\overline{\chi}\), average of all samples.

Additional statistical analyses were done using either SYSTAT 8.0 (SPSS Inc.) or StatView–J5.0 (SAS Institute Inc.). The details of data processing and statistical analysis are described in the supplementary document for data analysis. All supplementary materials described in the text are available at http://love2.aist-nara.ac.jp/laboratory/index_frame.html.

**ACKNOWLEDGEMENTS**

We would like to thank Mr Riu Yamashita for his aid in the analysis of the data, Dr Yoshimasa Sakakibara and Dr Ichiro Takemasa for their helpful suggestions and Ms Satoko Maki, Ms Keiko Miyaoka and Mrs Yoko Ishida for their valuable technical assistance.

**REFERENCES**