A Modular Analysis of Breast Cancer Reveals a Novel Low-Grade Molecular Signature in Estrogen Receptor – Positive Tumors

Kun Yu,[1] Kumaresan Ganesan,[2] Lance D. Miller,[3] and Patrick Tan[1,2,3]

Abstract

Purpose: Previous reports using genome-wide gene expression data to classify breast tumors have typically used standard unsupervised or supervised techniques, both of which have known limitations. We hypothesized that novel clinically relevant information could be revealed in these data sets by an alternative analytic approach. Using a recently described algorithm, signature analysis (SA), we identified "modules," comprising groups of tightly coexpressed genes that are conditionally linked to particular tumors, in a series of breast tumor gene expression profiles.

Experimental Design and Results: The SA successfully identified multiple breast cancer modules specifically linked to distinct biological functions. We identified a novel module, TuM1, whose presence was not readily discernible by conventional clustering techniques. The TuM1 module is expressed in a subset of estrogen receptor (ER) – positive tumors and is significantly enriched with genes involved in apoptosis and cell death. Clinically, TuM1-expressing tumors are associated with low histopathologic grade, and this association is independent of the inherent ER status of a tumor. We confirmed the robustness and general applicability of TuM1 module by demonstrating its association with low tumor grade in multiple independent breast cancer data sets generated using different array technologies. *In vitro*, the TuM1 module is down-regulated in ER+ MCF7 cells upon treatment with tamoxifen, suggesting that TuM1 expression may be dependent on active signaling by ER. Initial data is also suggestive that TuM1 expression may be clinically associated with a patient's response to antihormonal therapy.

Conclusion: Our results suggest that modular-based approaches toward gene expression data can prove useful in identifying novel, robust, and biologically relevant signatures even from data sets that have been the subject of substantial prior analysis.

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, as individual breast tumors can exhibit tremendous variations in clinical presentation, disease aggressiveness, and treatment response (2). In recent years, several groups have reported studies using genome-wide gene expression data to classify breast cancers for the purposes of molecular taxonomy, disease prognosis, and treatment response prediction (3–7). To identify specific and informative gene expression patterns ("molecular signatures"), many of these studies have typically used standard supervised or unsupervised learning techniques—in the former, signatures are identified based on their direct association with various clinical traits (e.g., survival), whereas in the latter genes and tumors are allowed to self-associate based on their overall patterns of similarity. Despite the promising nature of these initial studies, such conventional approaches are associated with certain limitations. For example, expression signatures obtained using supervised algorithms have been criticized for their generally poor representation of biologically coherent pathogenic mechanisms and disease pathways (8), and recent studies have shown that these signatures tend to be surprisingly unstable over different training sets (9, 10). Conversely, unsupervised algorithms, such as hierarchical clustering, typically cluster genes based on their global behavior across all samples (tumors) in the data set, when in reality certain genes may only show strong regulation in a certain subset of tumors, and weak to minimal regulation in others (11, 12). Because of these challenges, it is important to develop new methods to mine the inherent richness of information present in genome-wide expression data to further identify novel, robust, and biologically relevant molecular signatures for the purposes of tumor classification and patient stratification. In particular, Barkai and colleagues (11, 12) have recently described an algorithm called signature analysis (SA), which was designed to overcome the limitations of conventional clustering approaches. The SA adopts a modular approach to gene expression data, identifying groups of tightly coregulated genes that are conditionally linked to specific samples (modules; ref. 11). Notably, SA and its variants have previously

Authors' Affiliations: [1]National Cancer Centre; [2]Agenica Research; and [3]Genome Institute of Singapore, Singapore, Republic of Singapore

Received 7/14/05; revised 12/22/05; accepted 2/28/06.

Grant support: Agenica Research (P. Tan).

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

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

Requests for reprints: Patrick Tan, National Cancer Centre, 11 Hospital Drive, 169610 Singapore, Republic of Singapore. Phone: 65-6-436-8345; Fax: 65-6-226-5694; E-mail: cmrtan@nccs.com.sg.

doi:10.1158/1078-0432.CCR-05-1530

been shown to be superior to conventional clustering algorithms for detecting gene function and defining biological relationships (11, 12).

In this study, we tested the hypothesis that novel biological information could be uncovered in these breast cancer data sets using this modular technique. We applied the SA to a set of breast cancer expression profiles and successfully defined multiple tumor modules (TuM), each associated with a distinct biological function. Most significantly, the SA identified a previously unreported module (TuM1) in a subset of estrogen receptor–positive (ER+) tumors containing genes significantly enriched in cell death and apoptosis. The TuM1 module is not discernible by conventional hierarchical clustering cluster analysis and proved to be a robust signature by repeated random sampling assays (see Results). To further characterize the biological and clinical relevance of TuM1, we show that tumors expressing the TuM1 module are associated with low histologic grade ($P < 0.001$) and that this association is independent of the inherent ER status of the tumor. The TuM1/grade association is generally applicable as it is observed across multiple independent data sets representing distinct patient populations and array technologies. We also find that *in vitro*, the TuM1 module is expressed in ER+ MCF7 cells but down-regulated upon treatment with tamoxifen, suggesting that TuM1 expression may depend on active ER signaling. Motivated by this finding, we provide clinical data suggesting that TuM1 expression in primary tumors may identify patients more likely to respond to antihormonal therapy. By identifying a novel clinically relevant molecular signature in breast cancer, our results thus show that modular approaches to gene expression data, such as SA, can successfully reveal novel biological information even from data sets that have received substantial prior analysis.

Materials and Methods

Breast tissues and clinical information. Primary human breast tumors were obtained from the National Cancer Centre of Singapore Tissue Repository, after appropriate approvals from the repository and ethics committees of the center. Profiled samples contained at least 50% tumor content. Detailed descriptions of sample collection, archiving, and histologic assessment of tumors, including techniques and parameters, have been previously reported (ref. 7; also see Supplementary Information S1).

Cell culture and tamoxifen treatment. MCF-7 breast cancer cells were obtained from American Type Culture Collection (Manassas, VA), and cells were cultured in DMEM (Life Technologies, Grand Island, NY) supplemented with 10% fetal bovine serum, 100 units/mL penicillin, 100 units/mL streptomycin, and 2 mmol/L L-glutamine. Before tamoxifen treatment, cells were washed thrice in PBS and maintained in phenol red–free DMEM with 5% dextran charcoal–stripped fetal bovine serum (HyClone Laboratories, Pittsburgh, PA) for 24 hours. Subsequently, cells were treated with 10 μmol/L tamoxifen (Sigma, St. Louis, MO) and harvested at 48 hours. Control sister cultures were treated with an equivalent volume of the vehicle (0.1% ethanol).

Sample preparation and microarray hybridization. RNA was extracted from tissues and cell lines using Trizol (Invitrogen, Carlsbad, CA) reagent and processed for Affymetrix Genechip (Affymetrix, Inc., Santa Clara, CA) hybridizations using U133A Genechips according to the instructions of the manufacturer. The expression profiling of MCF-7 cell lines was done in duplicate from two independent sets of RNA samples each comprising control untreated MCF7 cells, cells treated with 10 μmol/L tamoxifen for 48 hours, and cells treated with vehicle (0.1% ethanol) for 48 hours. The expression profiling of MCF-7 cells was done on HG-U133 plus gene chips. The hybridization signal on the chip was scanned and processed by GeneSuite software (Affymetrix).

Data processing. Raw Genechip scans were quality controlled using GeneData Refiner (Genedata, Basel, Switzerland). The expression data was preprocessed by removing genes whose expression was absent in >40% samples (i.e., "A" calls), subjecting the remaining genes (9,116 probes) to a $\log_2$ transformation, and normalization by median-centering of samples. The expression data has been deposited into the Gene Expression Omnibus database (GSE2294).

SA and iterative signature algorithm. The basic SA methodology consists of four major steps: (*a*) a predefined set of "input genes" is selected; (*b*) using these input genes, the algorithm scans the expression data set, selecting samples (i.e., tumors) where the average expression of the input genes (tumor scores) is above a threshold value ("tumor threshold"); (*c*) within the selected tumors, individual genes whose weighted (by tumor scores) average expression exceeds a "gene threshold" are then identified, resulting in (*d*) a TuM being outputted, comprising a set of genes with expression levels above a particular threshold in a specific group of tumors. A detailed description of the SA methodology is provided in ref. 11. In this report, we use an extension of SA, the iterative signature algorithm (ISA), which uses a large number of random gene sets as the initial input and subsequently refines the TuMs through multiple iterative rounds of SA (12). As the inputted genes are random, ISA does not require prior knowledge and hence constitutes an entirely unsupervised analytic approach. Specific details about the ISA workflow and parameter settings are described in Supplementary Information S2. Based on previous reports, a gene threshold of 3.0 was selected as an optimal threshold for further in-depth analysis (11).[4]

Recurrence analysis to measure module robustness. SA uses recurrence analysis to assess the robustness of a module. For a given gene set (e.g., TuM1), a collection of new derived sets are created containing both the input genes and genes randomly selected from the entire data set. SA is then done on both the input set and the derivation sets. If the input set has a meaningful coregulated pattern, then this pattern should be strongly preserved in the derivation sets, and consequently the various output modules will have a large overlap (ref. 11; see Results for illustration). On the other hand, if there is no coregulated pattern embedded in the input set, the output modules will be quite different and little overlap will be observed. The details of recurrence analysis are further described in ref. 11, which also provides a mathematical definition of the recurrence metric.

Gene ontology and pathway analysis. We used the statistical web tool GoStat to identify functional annotations or Gene Ontology groups that are highly enriched in different gene sets (13). Fisher's exact test was done to calculate the significance of the observed enrichment, combined with a Benjamini and Hochberg correction to control the false discovery rate.[5] Additional functional and pathway analysis was done using Ingenuity pathway analysis,[6] a commercial database for identifying networks and pathways of interest in genomic data that was also been used in several other published reports (14, 15). The Ingenuity pathway analysis system uses a proprietary ontology representing over 300,000 classes of biological objects and semantically encoded relationships from the public domain literature to assign biological functions to a query data set (e.g., Affymetrix probes). The significance of functional enrichment is computed by a Fisher's exact test, and represented by a range of *P* values associated with either top-level functions or related subfunctions.

[4] The SA software is available for download at: http://barkai-serv.weizmann.ac.il/GroupPage/software.htm.

[5] GoStat is available at http://gostat.wehi.edu.au/cgi-bin/goStat.pl.

[6] http://www.ingenuity.com/products/pathways_analysis.html.

Associations between TuMs and clinical data. χ^2 Tests were used to calculate the association between each TuM and the following clinical variables: patient age, lymph node status, ER status, progesterone receptor status, tumor size, histologic grade (as continuous variable), and lymphovascular invasion. The significance of each association was also confirmed by hypergeometric probability density function analysis. Linear regression was used to confirm the independence of the TuM1/grade association from ER status in multivariate analysis. For multi–data set analyses, we identified common Unigenes between the Affymetrix U133A Genechip and Stanford, Rosetta, and Ma data sets (see Results), whereas the Uppsala data sets were matched directly by probe sets. Kaplan-Meier analysis was used for survival comparisons, and Cox regression was used to confirm the prognostic significance of TuM1 in multivariate analysis.

Gene set enrichment analysis. Gene Set Enrichment Analysis (GSEA) methodology, a modification of the weighted Kolmogorov-Smirnov statistic, provides a general statistical framework to test for the enrichment of gene expression profiles (16). GSEA considers *a priori* defined gene set, such as coregulated genes, and determines whether these members are enriched at the top (or bottom) of a list of markers ranked by the degree of correlation with a specific phenotype or class distinction. Multiple hypothesis testing is adjusted by calculating false discovery rates (16). The false discovery rate is the estimated probability that the reported result is a false positive. The details of GSEA are provided in ref. (16). The default parameter settings were used in the analysis.

Results

Identification of TuMs in breast cancer by SA. The basic methodology of the SA is described in the Materials and Methods. Compared with other analytic methods, two major features of SA are worth noting. First, because SA selects individual tumors exhibiting elevated expression levels of the input gene set, it is not necessary for genes in the input set to exhibit coregulated expression across all the samples, unlike other clustering techniques. Second, the gene selection process for the output module is done independently of the original input gene set. Thus, depending on the strength of coregulation between the original input genes within the selected tumors, the genes in the outputted module may contain either all, a subset of, or very few of the original input genes. The extent of overlap between the input set genes and genes in the outputted module reflects the overall robustness of the module, which is formalized as a "recurrence score" and used later in this report (see Materials and Methods). In this report, we have used a variant of SA, the ISA, which is purely unsupervised and unbiased.

We applied the ISA to a set of 96 breast cancer gene expression profiles, resulting in a modular decomposition of the gene expression data at different gene thresholds (12). Figure 1A illustrates this concept in the form of a "module tree." At low gene thresholds, only a few TuMs are initially identified, where each TuM consists of a large number of loosely correlated tumors and genes. At higher resolutions, the expression data is decomposed into a larger number of TuMs, where each TuM now contains a smaller set of tightly correlated tumors and genes. At a gene threshold of 3.0, we defined eight TuMs in the breast cancer expression data (TuMs 1-8). To place these modules in a biological context, we used the GoStat tool to identify biological or cellular functions that were significantly overrepresented in each module. Consistent with previous reports, many of these modules could be associated with distinct biological functions, such as extracellular matrix and collagen binding activities in TuM5 (corrected P values being $P = 2.85 \times 10^{-6}$ and 8.72×10^{-6} respectively), and cell cycle/cellular proliferation in TuM7 ($P = 4.08 \times 10^{-16}$; detailed descriptions of each TuM and the GO analysis are provided in

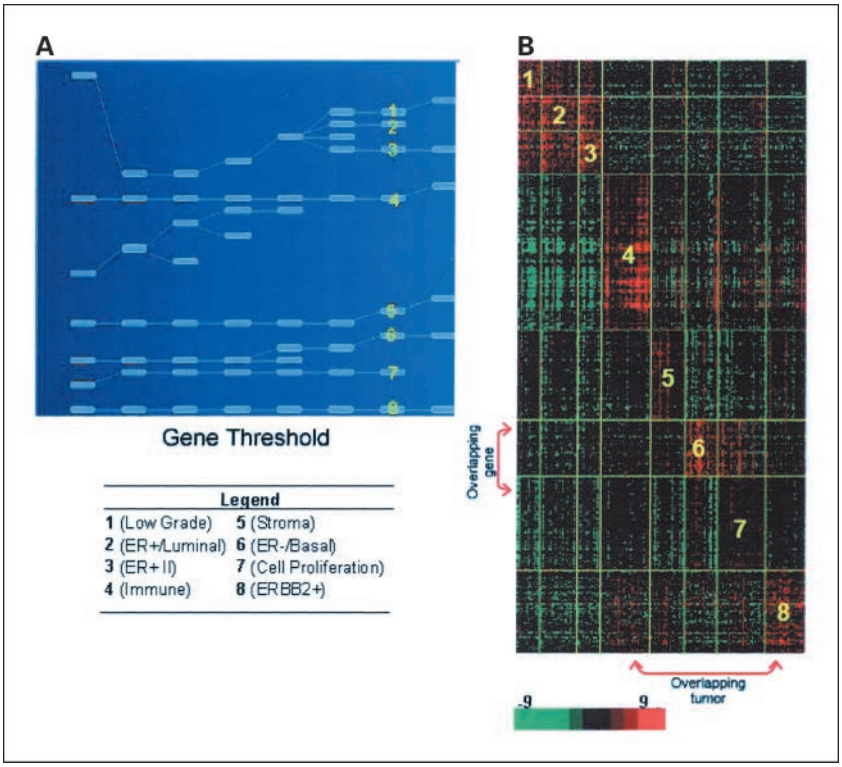

Fig. 1. TuMs of breast cancer. *A,* a module tree of the TuMs in the breast oncotranscriptome at different levels of resolution. Each node (*solid blue rectangle*) represents a TuM. Branches represent TuMs that originate from the same roots over a range of thresholds. *B,* global gene expression patterns within TuMs. Each row represents one gene and each column represents one tumor. Eight diagonal blocks (*separated by yellow grid*) represent eight modules (*under gene threshold 3.0 from A*). The legend of the eight modules is listed. The off-diagonal blocks reveal how genes in one module function in other modules. Red arrows, examples of genes and tumors that are shared between different modules. Heat-map bar, scale of gene expression value: red, high expression; green, low expression.

Supplementary Information S3). Some modules were clearly related. For example, three modules (TuM1, TuM2, and TuM3) were commonly derived from a single larger module containing genes previously reported as highly expressed in ER+ tumors, such as *ESR1*, *STC2*, and *BCL2* (3–7). Interestingly, this ER-related gene set has previously been treated in other studies as largely homogenous; however, its successful decomposition into smaller distinct units by the ISA suggests that the larger module may actually comprise multiple distinct and possibly independent biological subprograms. Although TuM2 (38 genes) and TuM3 (30 genes) exhibit substantial overlaps (~50%) in gene content (e.g., *STC2*, *BCL2*), >80% of the genes in TuM1 (33 genes) are not found in either TuM2 or TuM3. We did a survey of the literature and confirmed that the TuM1 module was previously unreported. The identification of TuM1 as a novel module thus shows the ability of the modular approach to reveal new molecular patterns in genome-wide expression data.

The TuM1 module is not obviously discernible by standard hierarchical clustering. Given that several groups have previously published extensive analyses of breast cancer expression data (3–6), the fact that hitherto the TuM1 module has remained undetected is somewhat surprising. We thus investigated if the TuM1 module could be obviously discerned using standard analytic techniques. To test this, we did a standard two-way unsupervised hierarchical clustering analysis on the gene expression data. We used a SD filter to select the top 1,500 genes exhibiting the highest variation in expression among the samples. This particular SD threshold was chosen to ensure the presence of sufficient TuM1 genes (~50%) in the filtered data. Average-linkage hierarchical clustering using a Pearson correlation metric was done on this gene set. Consistent with previous reports, the clustering analysis revealed a very large cluster of ER-related genes (~560 genes), but importantly within this group the TuM1 genes did not uniformly group with one other to form a "subcluster"—indeed, some TuM1 genes failed to localize within the ER cluster altogether (Fig. 2). Similar results were obtained when the hierarchical clustering was done on the global ISA-input gene set of 9,116 probes (Supplementary Information S4). This result indicates that it would have been highly unlikely for TuM1 to be readily discernible using conventional clustering approaches, supporting our hypothesis that novel biological information remains in these data sets despite their having received substantial prior analysis, which can be unearthed using alternative analytic methods such as SA. For the remainder of this report, we now focus on the novel TuM1 module in terms of its gene content, robustness, clinical associations, and general applicability.

TuM1 is a robust apoptotic TuM expressed in a subset of ER+ tumors. The TuM1 module was expressed in ~25% of the ER+ breast cancers in our initial data set of 96 tumors. Using a commercial database (Ingenuity), we did pathway analysis on TuM1 and found that genes related to cell death and apoptosis were significantly represented within this module ($P = 1.66 \times 10^{-5}$ to 0.034), such as *programmed cell death 4* (*PDCD4*), *mitochondrial ribosomal protein S30* (*MRPS30*), and *gap junction protein, α1, 43 kDa* (*connexin 43*; *GJA1*). Other genes in TuM1 include the xenobiotic-metabolizing enzymes *NAT1* and *FMO5*, and *PCM1*, which was recently reported to be associated with histologic grade in breast cancer (17). A fully annotated list of TuM1 genes is listed in Supplementary

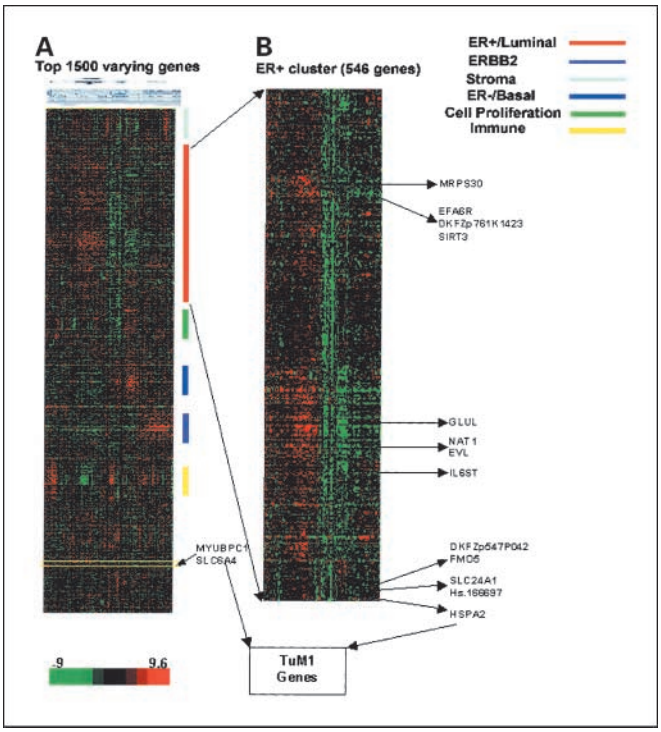

Fig. 2. *A,* unsupervised clustering of breast cancer gene expression data. The top 1,500 most highly varying genes were selected using a simple SD filter. This gene number was used to ensure that at least 50% of TuM1 genes were contained in the clustered gene set. Average linkage hierarchical clustering using a Pearson correlation metric was done using CLUSTER software and displayed by TREEVIEW. Fourteen of 16 TuM1 genes lay in the ER cluster, whereas the other two were outliers (*yellow frame*). Hierarchical clustering was also done on the entire gene set and the result is available in Supplementary Information S4. *B,* zoom-in of the ER gene cluster from hierarchical clustering. The TuM1 gene members are highlighted.

Information S5. This pathway analysis result suggests that the TuM1 module is likely to be biologically coherent and functionally significant.

To confirm that the identification of TuM1 was not dependent on the specific samples in our initial data set, we evaluated the robustness of the TuM1 module using two different techniques. First, we did recurrence analysis in an independent data set—in this method, random genes are added to TuM1 (33 members) to generate a series of TuM1-derived input gene sets, and SA is done on both TuM1 and the derived sets. The outputted modules are compared and the gene content overlap between the different output modules is determined. TuM1 is considered to be robust if the overall overlap (or "recurrence score") of the output modules is greater than a threshold (Fig. 3A) based on random input data. Specifically, we asked if the TuM1 module could be observed in an independent data set of 86 breast tumors that were not used in the original identification of TuM1. We did recurrence analysis on this independent set and found that TuM1 indeed emerged as a highly recurrent coregulated module (Fig. 3B), with the TuM1 molecular signature in this independent set also being confined to ER+ tumors at proportions similar to the original data (data not shown). Second, we further tested the robustness of TuM1 using repeated random sampling, a stringent validation technique recently proposed by Michiels et al. (9) to validate the reliability of gene signatures. We combined the original and independent test set samples and randomly

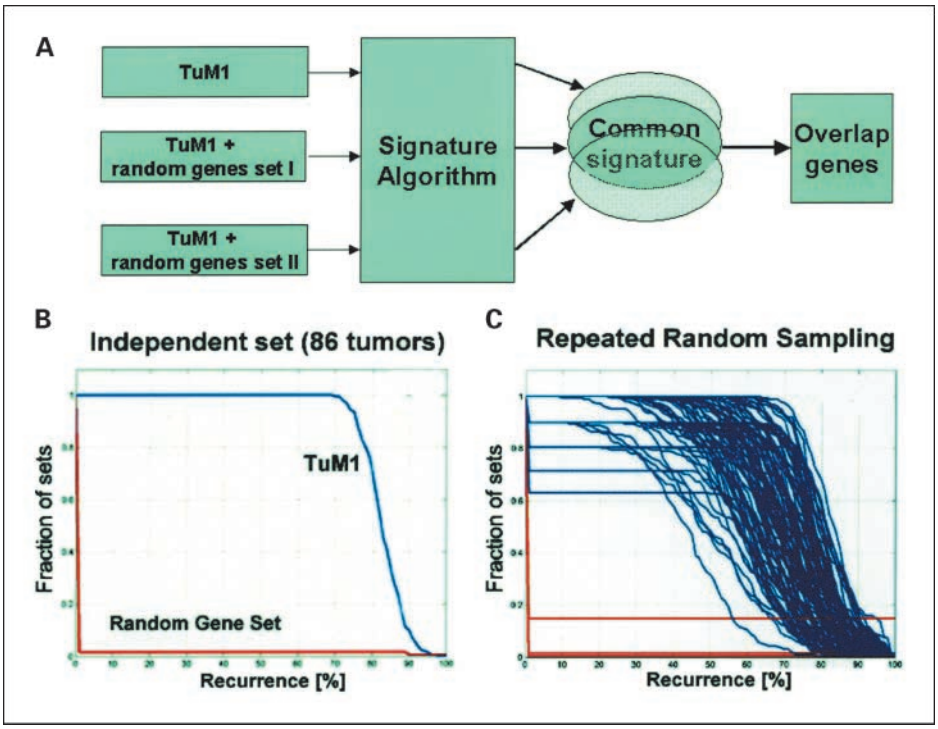

Fig. 3. Recurrence analysis. *A,* illustration of SA recurrence. The output modules of the derived input sets are compared to assess their overall overlap. A detailed explanation can be found in ref. 11. *B,* recurrence analysis of TuM1 on an independent set. TuM1 (*blue line*) shows a large number of highly overlapping outputs compared with random genes (*red line*), which yield little or no overlap in output. *X* axis (recurrence), overlap between TuM1 and 20 derivation signatures; *Y* axis, percentage of derivation signatures under a certain recurrence level. These variables are mathematically defined in ref. 11. *C,* recurrence analysis on random sampling validation (100 times). Each line represents one run of SA recurrence analysis. The recurrence of TuM1 (*blue lines*) is much higher than by chance (*red line*—there are 100 red lines). In 85% of the 100 random sampling validations, ≥70% of the derivation signatures showed >50% recurrence with TuM1.

generated one hundred sets of 96 tumors (96 being the same number as the original set). We did recurrence analysis on all 100 random sets, and found that in >85% of cases TuM1 displayed substantially higher recurrence scores compared with random data (Fig. 3C). In the remaining 15%, the failure to observe TuM1 could be attributed to the lack of TuM1-expressing tumors in the random set (Y.K., data not shown). We also independently repeated the entire ISA on a subset of these randomly generated sets and confirmed that the TuM1 module could be rederived (Y.K., data not shown). These results show that the TuM1 module is indeed highly robust within our center, and later in this report we also show that the TuM1 module is also present in breast cancer expression data sets from other groups.

TuM1 expression is associated with low histologic grade in an ER-independent fashion across multiple independent data sets. To investigate the clinical relevance of the TuM1 molecular signature, we correlated the TuM1-expressing tumors to various clinical and histopathologic variables. We observed significant positive correlations between TuM1 expression and ER positivity ($P < 0.001$), progesterone receptor positivity ($P = 0.01$), lymphovascular invasion ($P = 0.015$), small tumor size ($P = 0.015$), and low histologic grade ($P < 0.001$), but not age and lymph node status. As a comparison, TuM2 and TuM3, which are related but distinct modules to TuM1, were also significantly correlated with ER/progesterone receptor status, and TuM2 is also significantly correlated with low histologic grade (Table 1). This result suggests that despite the unsupervised nature of the ISA, many of the TuMs identified by the SA are nevertheless associated with observable and distinct clinical characteristics of breast tumors, supporting their clinical relevance (see Supplementary Information S6 for a similar analysis of the other TuMs).

The fact that TuM1 and TuM2 were significantly correlated with both ER status and histologic grade made us consider if these associations were occurring independently of one another, or if these two clinical variables (ER and grade) were mutually related. ER status has been previously shown to be strongly associated with histologic grade in breast cancer (ref. 18–26; see Discussion), and consistent with these previous

Table 1. Correlations between TuMs and clinical characteristics

	Age (≤/>55 y)	Size (≤/>3 cm)	Grade*	LN	ER	PR	LVI
TuM1	0.21	**0.0152** (≤3 †)	**<0.001**	0.15	**<0.001 (+)**	**0.0107 (+)**	**0.0152 (−)**
TuM2	0.17	0.15	**0.005**	0.06	**<0.001 (+)**	**0.0021 (+)**	0.19
TuM3	0.22	0.18	0.105	0.08	**<0.001 (+)**	**0.0015 (+)**	0.94

NOTE: Significant correlations ($P < 0.05$) are highlighted in bold. TuM1 is positively correlated with small tumor size (≤3), low grade, ER+, progesterone receptor positive, and lymphovascular invasion negative.

Abbreviations: LN, lymph node; PR, progesterone receptor; LVI, lymphovascular invasion.

*Grade is used as a continuous variable (also see Supplementary Information S7).

† The variable inside the parentheses indicates the direction of correlation with the TuMs.

reports ER is also significantly correlated with grade in our data set ($P = 0.001$). Using multivariate analysis, we tested if the correlation between TuM1 expression and low tumor grade was simply a consequence of their association with ER status or if the association between TuM1 expression and low tumor grade was independent of ER. In this analysis, we found that TuM1 expression is correlated with grade independently of ER ($P < 0.001$), but the association of TuM2 with low grade was not ($P = 0.9$; Table 2). We also repeated the univariate association studies, this time using a sample set of only ER+ tumors (unlike the previous analysis where all tumors were used). In this "ER+ only" data set, we found that TuM1 still remained significantly correlated with low tumor grade ($P < 0.001$). In contrast, TuM2 and TuM3, which both contain several ER-related genes, failed to exhibit a significant correlation with tumor grade when the ER-negative tumors were removed from the analysis ($P = 0.16$ and $P = 0.34$, respectively; Supplementary Information S7). These results indicate that the TuM1 expression signature is significantly correlated with low histologic grade in breast tumors and that this association is independent of ER status.

Having established the clinical validity of TuM1 in our in-house data set, we then tested the general applicability of the TuM1 molecular signature by applying it to an external series of patient populations. We tested a total of four independent publicly available breast cancer data sets with grade status, all using different array platforms and patient selection criteria. The first (the "Rosetta data set") consists of 117 breast tumors (71 ER+ tumors) profiled using oligonucleotide microarrays (27), the second (the "Stanford data set") consists of 122 breast tissue samples (81 ER+ tumors) profiled using cDNA microarrays (5), the third (the "Ma data set") consists of 60 ER+ tumors profiled using a separate cDNA microarray platform (28), and the fourth (the "Uppsala data set") consists of 67 ER+ tumors profiles on Affymetrix U133A arrays (29). For all these independent studies, we deliberately chose to specifically analyze only the ER+ tumors in these patient populations so as to avoid the possibility of ER status behaving as a confounding factor. We mapped the TuM1 genes identified in our study to their corresponding probes on the Rosetta, Stanford, and Ma microarrays on the basis of UniGene identifiers, and confirmed that these common subsets retain the ability to identify TuM1-overexpressing tumors in our data set (Supplementary Information S8). In all four independent data sets, the multivariate analysis showed that TuM1 is independently associated with grade (Table 3; Supplementary Information S7). This was further confirmed by univariate analysis, which showed that the TuM1 expression signature divided the ER+ tumors into two distinct subgroups, with tumors expressing high levels of the TuM1 signature being significantly associated with low histologic grade (Supplementary Information S7). These results, consistent with our own in-house series, strongly suggest that the TuM1 expression signature is likely to be a robust, specific, and generally applicable molecular signature for low histologic grade in breast cancer, as it is observed in a variety of independent data sets associated derived from a wide variety of disease stages and patient populations and profiled using different array technologies.

The TuM1 module is down-regulated by tamoxifen treatment in vitro. The observation that TuM1 is expressed in a subset of ER+ tumors raises the possibility that expression of this module may depend, at least in some part, on ER activity and signaling. To investigate the relationship between TuM1 expression and ER signaling, we decided to test the responsiveness of TuM1 to ER activity using an *in vitro* system. First, by profiling a set of breast and gastric cancer cell lines, we found that the TuM1 module was overexpressed in the ER+ breast cancer cell line MCF7 (Supplementary Information S9). Second, we treated MCF7 cells with tamoxifen, an inhibitor of ER, and using GSEA (16) further discovered that TuM1 was significantly down-regulated in tamoxifen-treated MCF7 cell lines compared with controls (false discovery rate = 0.05). As a control, none of the

Table 2. Correlation between grade and TuMs and other clinical variables in breast cancer by using linear regression multivariate analysis (SPSS)

Variable	*P*	Regression coefficient	95% Confidence interval for regression coefficient	
			Lower bound	Upper bound
TUM1	**<0.001**	0.783	0.404	1.162
TUM2	0.898	−0.025	−0.418	0.367
TUM3	0.586	−0.111	−0.516	0.294
TuM4	0.353	−0.125	−0.391	0.141
TuM5	0.426	0.120	−0.179	0.420
TuM6	0.405	0.127	−0.174	0.427
TuM7	0.192	−0.184	−0.462	0.094
TuM8	0.337	−0.137	−0.420	0.146
Age	0.197	0.006	−0.003	0.016
Size	0.317	0.003	−0.003	0.009
Node	0.106	0.183	−0.040	0.406
ER	0.091	−0.255	−0.551	0.041
PR	**0.020**	0.315	0.052	0.579

NOTE: Besides TuM1, only progesterone receptor is marginally correlated with grade. The positive regression coefficient means that the variable is associated with low grade.

Table 3. Correlation between TuM1 and grade within ER+ tumors in four public data sets

Data set	*P*	Regression coefficient	95% Confidence interval for regression coefficient	
			Lower bound	Upper bound
Rosetta	**0.014**	0.414	0.085	0.744
Stanford	**<0.001**	0.499	0.230	0.768
Ma	**0.015**	0.395	0.082	0.707
Uppsala	**0.017**	0.330	0.062	0.599

NOTE: The full list of multivariate analysis result is provided in Supplementary Information S7.

other TuMs were affected by tamoxifen treatment, with the exception of TuM2, which was marginally correlated with tamoxifen treatment (false discovery rate = 0.19). The details of this analysis are given in Supplementary Information S10. This result suggests that at least *in vitro*, TuM1 expression may be dependent on active ER signaling and may thus represent a molecular signature of ER activity.

A possible association between TuM1 expression and treatment response or clinical outcome. Tamoxifen is a standard antihormonal therapy used to treat ER+ breast cancer patients. Our finding that expression of the TuM1 module is dependent on active ER signaling made us investigate if the presence of this module in primary tumors might function as a molecular biomarker for active ER activity, and identify tumors that are likely to respond to tamoxifen or other antihormonal treatments. Supporting this possibility, certain genes in TuM1 have also been independently shown to be associated with therapeutic response in breast cancer (see Discussion). As clinical response information was not available in our in-house data, we tested three independent data sets where such data was available. First, we tested the Stanford series, which consists of patients who received adjuvant endocrine therapy if their tumors were ER+ (5). Using Kaplan-Meier survival analysis, patients with TuM1-expressing ER+ tumors exhibited better survival outcomes compared with patients with ER+ tumors where TuM1 was not expressed ($P = 0.0001$ for overall survival; $P = 0.0036$ for relapse-free survival; Fig. 4A and Supplementary Information S11). In a multivariate analysis of TuM1, grade, age, lymph node, and tumor size, TuM1 behaved as an independent predictor of survival outcome, whereas grade did not, demonstrating that TuM1 is more directly prognostic of patient survival than grade status alone (Supplementary Information S12). Second, we tested the Ma data set, which comprises a set of preselected tamoxifen-responsive and resistant ER+ tumors (28). Once again, TuM1-ovexpressing patients exhibited significantly better outcome than low TuM1

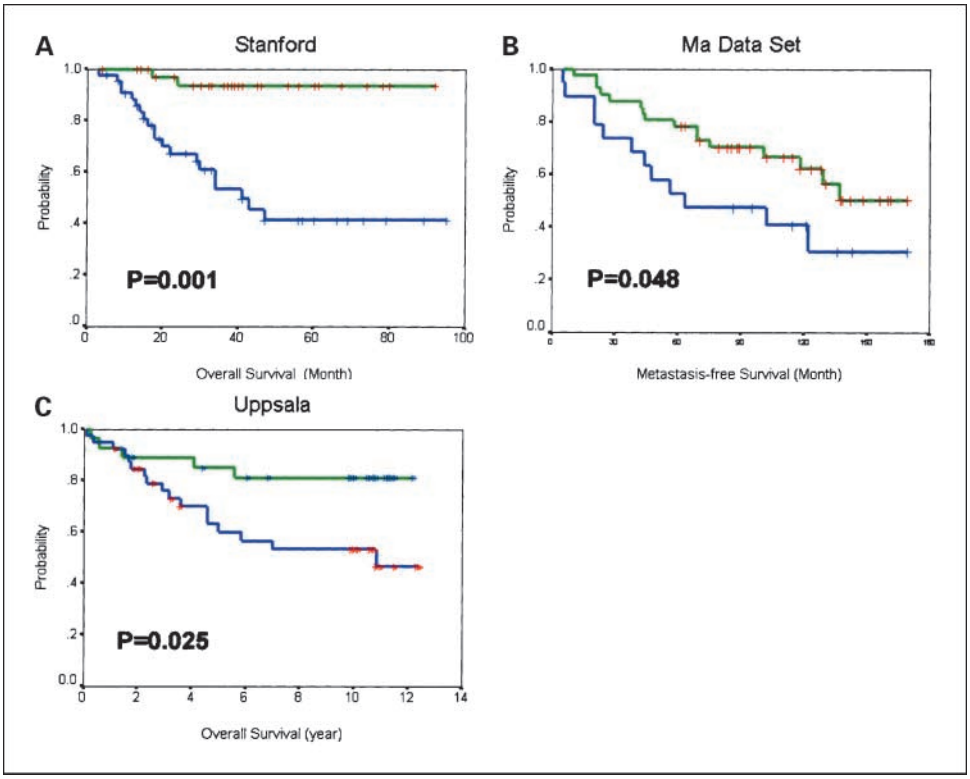

Fig. 4. Analysis of TuM1-disease outcome associations in three independent patient groups that received antihormonal treatment by using Kaplan-Meier analysis: *A,* Stanford data set: Overall survival for 81 ER+ patients who received adjuvant endocrine treatment (5). *B,* Ma data set: Metastasis-free survival for 60 ER+ patients receiving tamoxifen monotherapy (28). *C,* Uppsala data set: Overall survival for 67 ER+ patients receiving tamoxifen monotherapy (29).

patients (P = 0.048; Fig. 4B). By multivariate Cox regression analysis, TuM1 was the sole independent prognosis factor (P = 0.03; Supplementary Information S12); as grade, tumor size, node, and age are controlled in the Ma patient cohort (28). This observation was also tested using GSEA, which confirmed that TuM1 expression was significantly associated with tamoxifen response (P = 0.024; Supplementary Information S13). Third, the prognostic ability of TuM1 was tested on the Uppsala set, an independent patient cohort of 67 ER+ patients who received tamoxifen as monotherapy (29). Once again, patients with TuM1-expressing tumors experienced significantly improved overall survival outcomes compared with low TuM1-expressing patients (P = 0.025; Fig. 4C). By multivariate Cox regression analysis, TuM1 remained significantly associated with survival (P = 0.024), whereas grade, tumor size, and lymph node status did not (Supplementary Information S12). Taken collectively, these preliminary results raise the possibility that TuM1 expression in primary tumors might also be associated with the response of a tumor to clinical treatment, in particular antihormonal therapy.

Discussion

Gene expression profiling has been applied extensively in cancer research. However, recent reports have revealed significant limitations in many of the clustering algorithms and supervised classification techniques commonly used in such studies (8, 9, 11, 30). Given the complex genetic and molecular heterogeneity of cancer, it is perhaps not surprising that the identification of robust molecular signatures capable of directly reflecting disease behavior and clinical outcome remains a challenging task, and it has been proposed this will first require the comprehensive identification of gene signatures representing specific biological mechanisms and pathways (8). To achieve this aim, a number of powerful "modular" tools, such as SA and others (30, 31), have been developed, which are capable of identifying sets of genes associated with specific functions that are conditionally coregulated in tumors. In this report, we applied SA to characterize a set of breast tumor expression profiles, and identified a novel cell death and apoptosis–related gene expression signature (TuM1) that was not readily discernible using conventional clustering approaches. Notably, these analyses were all done on breast cancer data sets that had previously been extensively analyzed by multiple groups (3–7)—the successful identification of TuM1 as a novel module thus highlights the richness of information that likely remains embedded in such genome-wide data and awaiting discovery. We further found that the TuM1 module was highly robust across multiple independent data sets and was significantly enriched in genes associated with cell death and apoptosis, supportive of its biological coherence. Taken collectively, these results show that module-based approaches can successfully identify novel, robust, and biologically meaningful gene signatures in breast cancer.

Many of the genes in TuM1 have intriguing functions relevant to tumor biology, cell death, and treatment response. A few such examples are discussed here. For example, *PDCD4* has been shown to inhibit the growth of tumor cells (32), whereas *GJA1* has been reported to suppress cell proliferation and tumorigenicity of human glioblastoma cells (33) and to enhance apoptosis in response to chemotherapeutic agents (34). In addition, *MRPS30* has been reported as a proapoptotic gene that encodes protein programmed cell death 9 (35), whereas *leucine-rich repeats and immunoglobulin-like domains 1* (*LRIG1*) is a negative regulator of the ErbB family of receptor tyrosine kinases and has been suggested to suppress ErbB receptor function (36). Besides apoptosis-related genes, TuM1 also contains *β-TrCP1* (also known as *Fbwla* or *FWD1*), a component of the SKP1-cullin-F-box ubiquitin protein ligase complex, which can activate the nuclear factor-κB pathway and repress cell proliferation (37). Intriguingly, some genes in TuM1 have also been linked to clinical treatment response as well: Inactivation of *PDCD4* in human cancers has also been reported to cause decreased sensitivity to both geldanamycin and tamoxifen in breast cancer *in vitro* (38), whereas *NAT1*, another TuM1 gene, has been reported as an independent prognostic factor of breast cancer relapse and potential predictor of tamoxifen response (39).

Clinically, a major feature of the TuM1 module is its association with low histologic grade in an ER-independent manner. It is well known that histologic grade strongly correlates with ER status in breast cancer (18–26), with ER-negative tumors being predominantly high grade (grade 3). Indeed, consistent with these previous reports, there is a clear bias between ER and grade in all the data sets analyzed in this report (Supplementary Information S14). Because of the strong association between ER status and grade, previous reports attempting to identify "grade signatures" using supervised learning methods, in which genes exhibiting the strongest expression differences between high-grade and low-grade breast tumors are selected, have tended to define low-grade signatures containing multiple ER-related genes, such as *GATA3* (6), which could represent possible confounders. In contrast, the TuM1/low-grade association is independent of ER status, as confirmed by multivariate analysis. As for genes up-regulated in high-grade breast tumors (high-grade signatures), the majority seem to be related to cellular proliferation (6). Of interest, we have previously identified a gene signature for the Nottingham Prognostic Index in ER+ tumors, where tumor grade is a major component of the Nottingham Prognostic Index. This previous result also suggests that cell proliferation gene signatures are correlated with grade in an ER-independent manner as well (40).

Functionally, we have also shown in this report that the TuM1 module is expressed in the ER+ MCF7 cell line and is the only breast cancer TuM that is significantly responsive to tamoxifen treatment. This result suggests that expression of the TuM1 module may depend on continuous ER signaling and that TuM1 might represent a potential molecular signature of ER activity. The use of TuM1 as an *in vivo* biomarker of ER signaling is further supported by our observation that TuM1 is associated with clinical outcome in multiple independent patient cohorts receiving adjuvant hormonal treatment (the Stanford, Ma, and Uppsala cohorts; Fig. 4; Supplementary Information S12). This intriguing but preliminary finding definitely deserves further study and validation on a larger cohort of patients, supported by careful experiment design and data analysis. Interestingly, in the two independent patient cohorts where patients did not receive adjuvant treatment, patients with TuM1-expressing tumors also exhibited a trend toward improved clinical outcome; however, these differences were not statistically significant (P = 0.48 for Rosetta data set and P = 0.07 for Veridex data set; Supplementary

Information S15). This is consistent with the hypothesis that the TuM1 module may have a better ability to predict a patient's response to treatment than the intrinsic aggressive of the disease (i.e., the TuM1 signature is a predictive, rather than prognostic, signature).

In conclusion, our result shows the feasibility and utility of applying modular analytic approaches, such as SA, on cancer expression data. Besides breast cancer, our results suggest that, with the increasing availability of larger and comprehensive expression data sets, sophisticated analytic tools, such as SA, may be useful in refining our global understanding of the gene expression pathways in various malignancies.

Acknowledgments

We thank Tan Puay Hoon, Hong Ga Sze, Wee Siew Bok, and Wong Chow Yin for their clinical assistance; and Pablo Tamayo and Jill P. Mesirov of the Broad Institute for assistance with GSEA analysis.

References

1. Chia KS, Seow A, Lee HP, Shanmugaratnam K. Cancer incidence in Singapore, 1993-1997. Singapore: Singapore Cancer Registry; 2000.
2. Tavassoli FA, Schnitt SJ. Pathology of the breast. New York: Elsevier; 1992.
3. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumors. Nature 2000;406:747–52.
4. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869–74.
5. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 2003;100:8418–23.
6. 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 U S A 2003;100:10393–8.
7. Yu K, Lee CH, Tan PH, Tan P. Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations. Clin Cancer Res 2004;10:5508–17.
8. Chang HY, Nuyten DS, Sneddon JB, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci U S A 2005;102:3738–43.
9. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays. Lancet 2005;365:1684–5.
10. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171–8.
11. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. Nat Genet 2002;31:370–7.
12. Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data. Bioinformatics 2004;20:1993–2003.
13. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004;20:1464–5.
14. Bredel M, Bredel C, Juric D, et al. Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. Cancer Res 2005;65:8679–89.
15. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
16. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
17. Armes JE, Hammet F, De Silva M, et al. Candidate tumor-suppressor genes on chromosome arm 8p in early-onset and high-grade breast cancers. Oncogene 2004;23:5697–702.
18. Nishimura R, Misumi A, Kimura M, Tokunaga T, Akagi M. Relationship between the content of estrogen and progesterone receptors and the pathological characteristics in human breast cancer. Jpn J Surg 1982;12:191–7.
19. Fisher ER, Osborne CK, McGuire WL, et al. Correlation of primary breast cancer histopathology and estrogen receptor content. Breast Cancer Res Treat 1981;1:37–41.
20. Maynard PV, Davies CJ, Blamey RW, Elston CW, Johnson J, Griffiths K. Relationship between oestrogen-receptor content and histological grade in human primary breast tumours. Br J Cancer 1978;38:745–8.
21. Blanco G, Alavaikko M, Ojala A, et al. Estrogen and progesterone receptors in breast cancer: relationships to tumour histopathology and survival of patients. Anticancer Res 1984;4:383–9.
22. Chua DY, Pang MW, Rauff A, Aw SE, Chan SH. Correlation of steroid receptors with histologic differentiation in mammary carcinoma. A Singapore experience. Cancer 1985;56:2228–34.
23. Barbi GP, Marroni P, Bruzzi P, Nicolo G, Paganuzzi M, Ferrara GB. Correlation between steroid hormone receptors and prognostic factors in human breast cancer. Oncology 1987;44:265–9.
24. Bhatavdekar JM, Trivedi SN, Shah NG, et al. Correlation of steroid receptors with histopathologic characteristics in breast carcinoma. Neoplasma 1988;35:413–23.
25. Reiner A, Reiner G, Spona J, Schemper M, Holzner JH. Histopathologic characterization of human breast cancer in correlation with estrogen receptor status. A comparison of immunocytochemical and biochemical analysis. Cancer 1988;61:1149–54.
26. Komaki K, Mori T, Morimoto T, Sasa M, Monden Y, Ii K. Correlation between estrogen receptor status and histological malignancy in human breast cancer. J Surg Oncol 1991;46:185–9.
27. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
28. Ma XJ, Wang Z, Ryan PD, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004;5:607–16.
29. Miller LD, Smeds J, George J, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 2005;102:13550–5.
30. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet 2004;36:1090–8.
31. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. Nat Genet 2005;37:S38–45.
32. Lankat-Buttgereit B, Goke R. Programmed cell death protein 4 (pdcd4): a novel target for antineoplastic therapy? Biol Cell 2003;95:515–9.
33. Huang RP, Fan Y, Hossain MZ, Peng A, Zeng ZL, Boynton AL. Reversion of the neoplastic phenotype of human glioblastoma cells by connexin 43 (cx43). Cancer Res 1998;58:5089–96.
34. Huang RP, Hossain MZ, Huang R, Gano J, Fan Y, Boynton AL. Connexin 43 (cx43) enhances chemotherapy-induced apoptosis in human glioblastoma cells. Int J Cancer 2001;92:130–8.
35. Carim L, Sumoy L, Nadal M, Estivill X, Escarceller M. Cloning, expression, and mapping of PDCD9, the human homolog of *Gallus gallus* pro-apoptotic protein p52. Cytogenet Cell Genet 1999;87:85–8.
36. Laederich MB, Funes-Duran M, Yen L, et al. The leucine-rich repeat protein LRIG1 is a negative regulator of ErbB family receptor tyrosine kinases. J Biol Chem 2004;279:47050–6.
37. Nakayama K, Hatakeyama S, Maruyama S, et al. Impaired degradation of inhibitory subunit of NF-κB (IκB) and β-catenin as a result of targeted disruption of the β-TrCP1 gene. Proc Natl Acad Sci U S A 2003;100:8752–7.
38. Jansen AP, Camalier CE, Stark C, Colburn NH. Characterization of programmed cell death 4 in multiple human cancers reveals a novel enhancer of drug sensitivity. Mol Cancer Ther 2004;3:103–10.
39. Bieche I, Girault I, Urbain E, Tozlu S, Lidereau R. Relationship between intratumoral expression of genes coding for xenobiotic-metabolizing enzymes and benefit from adjuvant tamoxifen in estrogen receptor α-positive postmenopausal breast carcinoma. Breast Cancer Res 2004;6:R252–63.
40. Yu K, Lee CH, Tan PH, et al. A molecular signature of the Nottingham prognostic index in breast cancer. Cancer Res 2004;64:2962–8.