Research Article

Cancer Epidemiology, Biomarkers & Prevention

# Breast Cancer Prognostic Biomarker Using Attractor Metagenes and the *FGD3–SUSD3* Metagene

Tai-Hsien Ou Yang[1,2], Wei-Yi Cheng[1,2], Tian Zheng[3], Matthew A. Maurer[4], and Dimitris Anastassiou[1,2]

## Abstract

**Background:** The winning model of the Sage Bionetworks/DREAM Breast Cancer Prognosis Challenge made use of several molecular features, called attractor metagenes, as well as another metagene defined by the average expression level of the two genes *FGD3* and *SUSD3*. This is a follow-up study toward developing a breast cancer prognostic test derived from and improving upon that model.

**Methods:** We designed a feature selector facility calculating the prognostic scores of combinations of features, including those that we had used earlier, as well as those used in existing breast cancer biomarker assays, identifying the optimal selection of features for the test.

**Results:** The resulting test, called BCAM (Breast Cancer Attractor Metagenes), is universally applicable to all clinical subtypes and stages of breast cancer and does not make any use of breast cancer molecular subtype or hormonal status information, none of which provided additional prognostic value. BCAM is composed of several molecular features: the breast cancer–specific *FGD3–SUSD3* metagene, four attractor metagenes present in multiple cancer types (CIN, MES, LYM, and END), three additional individual genes (*CD68*, *DNAJB9*, and *CXCL12*), tumor size, and the number of positive lymph nodes.

**Conclusions:** Our analysis leads to the unexpected and remarkable suggestion that ER, PR, and HER2 status, or molecular subtype classification, do not provide additional prognostic value when the values of the *FGD3–SUSD3* and attractor metagenes are taken into consideration.

**Impact:** Our results suggest that BCAM's prognostic predictions show potential to outperform those resulting from existing breast cancer biomarker assays. *Cancer Epidemiol Biomarkers Prev; 23(12); 2850–6. ©2014 AACR.*

## Introduction

Several prognostic models for breast cancer using molecular features have been used in biomarker products (1–3), which have also proven to be of value to medical decision making, such as predicting whether an early-stage patient will benefit from adjuvant chemotherapy. A recent crowd-sourced research study, the Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge (BCC; ref. 4) used the METABRIC dataset (5) containing molecular and clinical features from 1,981 patients with breast cancer. The winning model (6, 7) as well as all five top-scoring models made use of several molecular features, called attractor metagenes (8), as well as the *FGD3–SUSD3* metagene defined by the average of the expression levels of the two genes, *FGD3* and *SUSD3*, which are located directly adjacent to each other at Chr9q22.31.

To make a prognostic tool usable in a clinical setting derived from our newly identified metagenes, we optimized a new model based on the disease-specific survival information included in the METABRIC dataset, providing an estimate of the breast cancer–specific 10-year survival rate for each patient. Here, we present the derivation of the model to which we refer as the BCAM (Breast Cancer Attractor Metagenes) biomarker. We derived the model using the uniformly renormalized (4) 1,981-sample METABRIC dataset, in which the two genes whose high expression is most associated with good prognosis are *FGD3* and *SUSD3*; in fact the most prognostic molecular feature that we had found (6) was the *FGD3–SUSD3* metagene. At the other extreme, the genes whose high

[1]Department of Systems Biology, Columbia University, New York, New York. [2]Department of Electrical Engineering, Columbia University, New York, New York. [3]Department of Statistics, Columbia University, New York, New York. [4]Division of Hematology/Oncology of the Department of Medicine, Columbia University, New York, New York.

**Note:** Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Current Address for W.-Y. Cheng: Icahn School of Medicine, Mount Sinai Hospital, New York, New York.

**Corresponding Authors:** Matthew A. Maurer, Division of Hematology/Oncology, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, 1130 St. Nicholas Avenue, Room 217C, New York, NY 10032. Phone: 212-851-4761; Fax: 212-851-4572; E-mail: mm2058@columbia.edu; and Dimitris Anastassiou, Department of Systems Biology and Department of Electrical Engineering, Columbia University, 1312 S.W. Mudd Building, Mail Code 4712, 500 West 120th Str., New York, NY 10027. Phone: 212-854-3113; Fax: 212-932-9421; E-mail: da8@columbia.edu.

**doi:** 10.1158/1055-9965.EPI-14-0399

expression is most associated with poor prognosis were members of the mitotic chromosomal instability ("CIN") attractor metagene, which we previously identified (8) as a "pan-cancer" molecular signature using unsupervised analysis of other datasets from different cancer types.

Because many models submitted in the BCC were similar to those used in existing biomarker products, we hypothesized that the features that we used in our models can improve the accuracy of existing prognostic assays. We therefore compared our features with those used in the 21-gene Oncotype DX (1), 70-gene MammaPrint (2), and 50-gene PAM50 (3) assays, using several breast cancer datasets. Given the lack of the actual implemented biomarker scores for the patients in these datasets, we used the available microarray values to produce an estimated performance. On the basis of these comparisons, our results suggest that the combination of features used in the BCAM model may compare favorably with those used in these products, and therefore we conclude that it is promising and deserving of formal evaluation within the context of several ongoing adjuvant clinical trials.

## Materials and Methods

### Datasets, preprocessing, end points of survival analysis

Because most breast cancer datasets do not include the number of positive lymph nodes, we relaxed the requirements for acceptable validation datasets to allow for those that merely provide a binary (negative/positive) lymph node status. Still, we found only four datasets in addition to METABRIC (5), available through Sage Synapse under accession number syn1710250, with the requirements that they include probes for genes *FGD3* and *SUSD3*, tumor size, lymph node status, and disease-specific survival or recurrence data, from which we could extract at least one statistically significant ($P < 0.05$) comparison between the BCAM formula and those used in other genomic assays. We refer to these four datasets as Loi (9), Buffa (10), Wang (11), and Miller (12), and they are available from the Gene Expression Omnibus under accession numbers GSE6532, GSE22219, GSE19615, and GSE3494, respectively. Only the Buffa dataset provides the number of positive lymph nodes; in the other datasets, we used the BCAM formula setting the number of positive lymph nodes for lymph node–positive patients to 1. The tumor size and the lymph node number were logarithmically transformed.

The datasets generated from Affymetrix U133A/B, and Plus2.0 arrays were renormalized using Robust Multi-array Average (RMA), as implemented in the *affy* package in Bioconductor (www.bioconductor.org) in the R software. If there was more than one platform provided for each patient, the measurements were combined and renormalized using RMA. The METABRIC dataset was renormalized by Sage Synapse (4). Because the BCAM formula is the linear combination of heterogeneous covariates, we corrected the distribution of genomic assays in each dataset by multiplying the size and the lymph node number with the ratio of the standard deviations of the genomic assays in each dataset to the standard deviation of the genomic assays in the METABRIC dataset.

For survival analysis, because each dataset uses different end point for censoring, we used the end point defined closest to disease-specific survival available in the METABRIC dataset and in the Miller dataset. We used time to recurrence in the Loi and Wang datasets and distant-relapse–free survival in the Buffa dataset.

### Comparison of predictive models

The concordance index (13) is used to assess the accuracy of the rankings of patients' risk. It is defined as the relative frequency of accurate pairwise predictions of survival ranking over all pairs of patients for which such a determination can be achieved. To compare the performances of the predictive models, we estimated the distribution of the concordance index as the overall C-index (13) for each model on each subset of samples. Because the overall C estimator is proven to be asymptotically normal (13), the null distribution of the C-index can be approximated by a normal distribution with mean 0.5 and the sampling variance of C-index when the sample size is sufficiently large. Standardized by the mean under the null hypothesis and estimated variance from data, the C-index follows a Student $t$ distribution approximately. The difference between two estimated C-indices, after standardization, also follows a $t$ distribution approximately under the null hypothesis that the two C-indices are equal. Therefore, the comparison between two overall C-indices can be carried out by a Student $t$ test and the $P$ value is evaluated accordingly. The overall C-index estimation and $t$ test were performed by the *survcomp* package (14) in the R software.

### Feature selector facility

The prognostic score displayed for each combination of selected features was designed to be resistant to overfitting. It is evaluated as the asymptotic average of the concordance indices resulting from random 2-fold cross-validation experiments in the METABRIC dataset. Each experiment uses the selected features as covariates to train a Cox proportional hazards model on half of the dataset based on random splitting, and evaluates the corresponding concordance index of the fitted model on the other half. Each experiment is also repeated by reversing the training/validation roles of the same subsets.

### Estimation of survival rate

The final BCAM score between 0 and 100 is generated as the corresponding percentile value from the Cox model formula against the 1,981-sample METABRIC dataset. The breast cancer–specific 10-year survival rate associated with the BCAM score is found by calculating the Kaplan–Meier hazard ratio at 10 years for the METABRIC subpopulation inside a sliding window

containing 20% of the samples (10% in each side) with the closest BCAM scores. If there are not enough patients on one side of the window, we reduce the window size so that it remains symmetric.

### Other breast cancer prognostic formulas

We compared BCAM with four biomarkers used in other genomic assays: The 21-gene Oncotype DX signature (1), the 70-gene MammaPrint signature representing a good prognosis gene expression profile (2), the 50-gene ROR-S signature whose different expression profiles constitute centroids for four intrinsic PAM50 subtypes (3); and the ROR-C signature combining the PAM50 subtypes with original tumor size (3). The definition of each of the four groups in the 21-gene signature and the formula for combining them were obtained in (1) without applying the cutoff thresholds, as the expression levels of the groups for the microarray values and RT-PCR values were not compatible. The score of the 70-gene assay was derived as described in the original articles (2, 15). The centroids of intrinsic subtypes were obtained from the Bioconductor package *genefu*. The formula of combining the individual scores for the four subtypes and tumor size were obtained from the original article (3).

## Results

### Validation of the *FGD3–SUSD3* metagene

We first confirmed that the breast cancer–specific *FGD3–SUSD3* metagene, which was the most prognostic molecular feature in METABRIC, remains highly prognostic in all other datasets. Figure 1A shows the Kaplan–Meier survival curves of the *FGD3–SUSD3* metagene, demonstrating statistical significance in all five datasets. The gene most associated with the *FGD3–SUSD3* metagene in METABRIC (also among the most associated ones in all the other datasets) is the estrogen receptor *ESR1*, which is less prognostic than *FGD3–SUSD3* in all five datasets (Fig. 1B).

### Feature selection

We then defined the features, which, when combined, would optimize prognostic performance in the METABRIC datasets. The *FGD3–SUSD3* metagene was an obvious such candidate feature. We also included the metagenes that we had used in the top-ranked models of the BCC, namely CIN (mitotic chromosomal instability), MES (mesenchymal transition), and LYM (lymphocyte infiltration) and two conditioned versions: MES*, restricted to early-stage tumors defined as lymph node negative with tumor size less than 30 mm, and LYM*, restricted to samples with more than three positive lymph nodes [During our participation in the BCC we had found (6) that MES was prognostic only in early-stage cancers and that LYM, although protective overall, was associated with poor prognosis in the presence of multiple positive lymph nodes]. Following our additional results, analyzing other datasets from 12 different cancer types (16) in collaboration with The Cancer Genome Atlas (TCGA) Pan-Cancer project (17), we also considered the next top-ranked attractor metagene, END, a newly discovered multi-cancer molecular signature of endothelial markers. We also included all the molecular features whose combination is used in existing breast cancer prognostic assays: Oncotype DX (proliferation, invasion, estrogen, HER2 groups, *CD68*, *GSTM1*, and *BAG1* genes); PAM50 defined molecular subtypes (basal, luminal A, luminal B, and HER2 features); the single 70-gene Mammaprint feature; and the three genes *ESR1*, *PGR*, *ERBB2* used (18) in the TargetPrint assay. Finally, we included the number of positive lymph nodes and tumor size, as we had identified them in the BCC as two of the most prognostic clinical features.

We designed a feature selection Web-based facility (www.ee.columbia.edu/~anastas/featureselector), which evaluates a prognostic score after selecting a specified number among the above features. The score was designed so that it will cease increasing when overfitting has occurred. We include logarithmic versions for the number of lymph nodes and the tumor size, because we found that the score becomes consistently higher if we include these versions rather than the direct values. The purpose of the overall facility is to provide an estimate of the performance of each of the existing assays by selecting the corresponding features, as well as to provide insight on the relative contribution of individual features when combined with other ones, leading to the selection of an optimal biomarker. Instructive results, noted in the facility, are the identified best selection of a given number $N$ of features.

For $N = 1$, the most prognostic feature among those listed in the facility is the "Luminal A" feature of PAM50, which measures the degree of correspondence with a good prognosis subtype. However, the luminal A feature is eliminated from the best choice of features when $N = 2$, in which case the optimal choice is the *FGD3–SUSD3* metagene combined with the number of positive lymph nodes. At $N = 3$, the CIN metagene is also selected, followed in increasing order by tumor size, MES*, LYM, LYM*, *CD68*, and *END*, each of which increases the score, at which point ($N = 9$) it reaches the value of 0.741. Following this selection of nine features, no additional feature increases the score. To further increase performance, we then implemented a heuristic optimization algorithm by including randomly chosen single genes in combination with some or all of the selected features, retaining genes with both high cross-validation scores and known roles in cancer literature. We thus identified two additional genes, *DNAJB9* and *CXCL12*, for a total number of 11 features increasing the score to 0.747. *DNAJB9* has the property that, if included among the potential features, is selected as early as $N = 4$ (www.ee.columbia.edu/~anastas/featureselector2). The other gene, *CXCL12*, is selected at $N = 7$. Both of these genes are known to play important roles in cancer (19, 20).

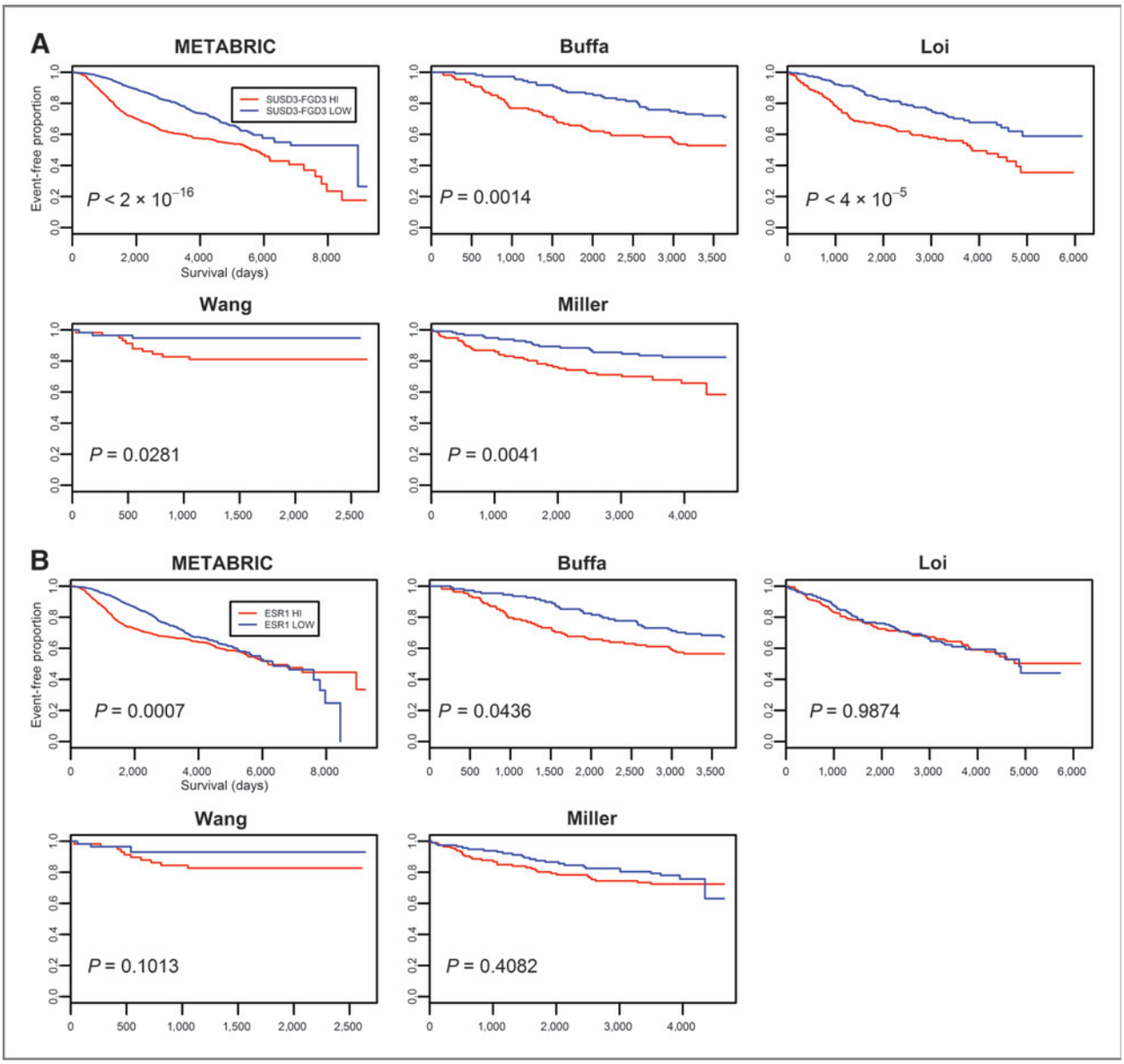


Figure 1. Kaplan–Meier survival curves on the basis of the *FGD3–SUSD3* metagene (A), the *ESR1* gene (B), in five datasets. *P* values were derived using the log-rank test after dividing each dataset into two equal-sized subgroups.

## BCAM biomarker

We thus defined our model based on the Cox model formula defined by the full METABRIC dataset using the 11 features, *FGD3–SUSD3*, CIN, MES*, LYM, END, LYM*, *CD68*, *DNAJB9*, *CXCL12*, number of positive lymph nodes, and tumor size (Supplementary Methods and Materials).

The final BCAM score between 0 and 100 is generated from the Cox model formula as the percentile value against the 1,981-sample METABRIC dataset. Figure 2 shows the estimated breast cancer–specific 10-year survival rate as a function of the BCAM score.

## Validation in other datasets

We compared the prognostic performance of the BCAM formula with formulas of other genomic assays: Oncotype DX, MammaPrint, ROR-S (using PAM50 subtype information alone), and ROR-C (using PAM50 subtype information and tumor size). We used other breast cancer datasets appropriate for evaluating prognostic values to which we refer as Loi (9), Buffa (10), Wang (11), and Miller (12). For each dataset, we also consider the two subsets, (i) lymph node-negative (LNN) patients, and (ii) estrogen receptor-positive (ERP) patients (regardless of PR and HER2 status). Additional intersection of these sets would not lead to results of statistical significance.

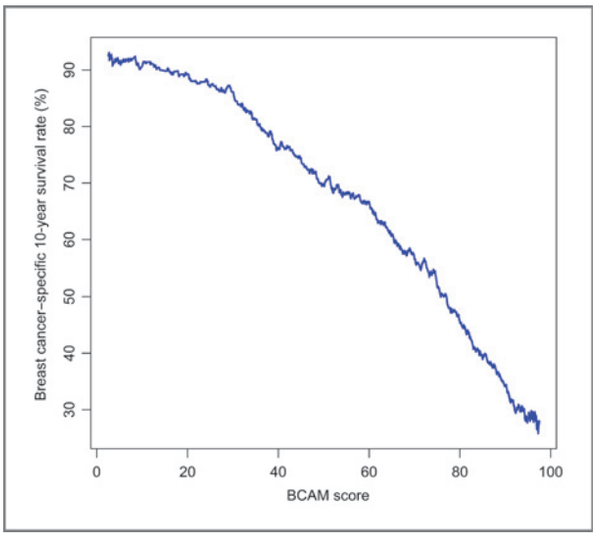


Figure 2. Breast cancer–specific 10-year survival rate as a function of the BCAM score normalized as the percentile value against the 1,981-sample METABRIC dataset.

BCAM outperformed the other genomic assays in all cases in which comparisons had statistical significance (Table 1). In most of these comparisons (except when comparing BCAM with ROR-C in the LNN subsets) BCAM makes use of clinical information not used in the other assays. These results demonstrate the advantage of integrating clinical stage with molecular feature information into one product with enhanced prognostic power.

## Discussion

The results of our analysis of the METABRIC dataset lead to the unexpected and remarkable suggestion that breast cancer subtype classification, as well as estrogen/progesterone receptor and HER2 status do not provide any additional prognostic information as long as the expression levels of the *FGD3–SUSD3* and the attractor metagenes are known and taken into consideration. This suggestion is strengthened by the fact that the uniformly renormalized 1,981-sample METABRIC dataset is uniquely rich and useful for reaching results of statistical significance in survival analysis.

In support of the above suggestion, using the publicly available Web-based feature selector facility, for all feature combinations that we tried:

1. Selecting the Oncotype DX Estrogen group, or any of genes *ESR1* and *PGR*, in addition to any selected feature combination that includes metagenes *FGD3–SUSD3* and CIN, does not increase, and in most cases decreases the score.
2. Replacing the selection of the Oncotype DX Estrogen group or any of genes *ESR1* and *PGR* (including any multiple selection of these features) with *FGD3–SUSD3*, in any selected feature combination, increases the score.

The expression values of Her2, PR, and ER and subtype designation are correlated with the metagenes of BCAM and therefore are indirectly taken into account. They obviously provide vital information and improved understanding of breast cancer biology, which has led to effective treatments. However, our results suggest that an optimum breast cancer biomarker product does not need to include them.

Many early versions of microarray platforms, notably the popular Affymetrix U133A, do not contain probes for *FGD3* and *SUSD3*, which may provide some explanation as to why these genes were not found earlier as highly prognostic in breast cancer. The two genes are genomically adjacent to each other and are correlated with *ESR1* and *PGR*. The simultaneous silencing of *FGD3* and *SUSD3* is strongly associated (6) with poor prognosis. Furthermore, a recent study (21) identified *SUSD3* as the single most predictive gene (more than *ESR1*) of response to aromatase inhibitor therapy. On the basis of the above facts, we believe that existing breast cancer prognostic assays should be reevaluated, and that the biologic mechanism responsible for *FGD3–SUSD3* silencing, as well as the corresponding phenotypic associations, should be priority research topics.

The alternative offered by the BCAM biomarker is one universal prognostic assay applicable to all breast cancer subtypes and stages, integrating tumor biology across stages. Indeed, as evidenced by the feature selector facility, the LYM and MES metagenes would not be prognostic in the absence of stage information, and the conditioned LYM* and MES* features add significantly to the overall prognostic power. BCAM is also independent of tumor grade, as the CIN metagene is a proxy for, and more prognostic than, grade, or the expression of the *Ki67* gene.

We also observed that inclusion of gene *CD68*, used in the Oncotype DX assay, improves the prognostic performance of our model, of which we were not aware during our BCC participation. The expression of gene *CD68*, a marker of tumor-associated macrophages, is associated with worse prognosis, although it is positively correlated with the protective LYM lymphocyte infiltration signature, and their combination improves prognostic ability.

The CIN, MES, and LYM and END features were previously identified and precisely defined by multicancer analysis (8, 16) as representing attributes of cancer in general. Furthermore, they were all found without any use of the METABRIC dataset, and their prognostic power was independently confirmed in METABRIC. This raises the exciting possibility that they will also prove to be useful serving as "building blocks" of biomarkers in other types of cancer as well.

There are a few limitations to this study, one being the lack of large independent datasets that have outcomes as well as the necessary data for applying the BCAM

**Table 1.** List of scores, measured by the corresponding concordance index, after applying the formula of each prognostic assay on cancer datasets and their LNN and ERP subsets

| | Number of samples | Number of events | BCAM | ROR-S | ROR-C | 21-gene | 70-gene |
|---|---|---|---|---|---|---|---|
| Buffa | | | | | | | |
| Full | 216 | 82 | 0.757 | **0.673** | **0.681** | **0.684** | **0.628** |
| LNN | 125 | 43 | 0.720 | 0.663 | **0.658** | 0.681 | **0.628** |
| ERP | 134 | 49 | 0.725 | 0.710 | **0.677** | 0.727 | **0.647** |
| Loi | | | | | | | |
| Full | 393 | 139 | 0.716 | **0.604** | **0.668** | **0.635** | **0.605** |
| LNN | 250 | 85 | 0.695 | **0.610** | 0.670 | **0.6346** | **0.604** |
| ERP | 348 | 117 | 0.714 | **0.605** | **0.677** | **0.640** | **0.606** |
| Wang | | | | | | | |
| Full | 115 | 14 | 0.782 | **0.640** | **0.686** | **0.642** | **0.594** |
| LNN | 64 | 5 | 0.839 | **0.638** | **0.648** | **0.665** | 0.674 |
| ERP | 66 | 3 | 0.660 | **0.367** | 0.545 | **0.435** | **0.372** |
| Miller | | | | | | | |
| Full | 236 | 55 | 0.764 | **0.639** | 0.726 | **0.643** | **0.636** |
| LNN | 158 | 22 | 0.690 | 0.604 | 0.702 | 0.608 | 0.604 |
| ERP | 201 | 49 | 0.755 | **0.646** | 0.727 | **0.645** | **0.650** |
| METABRIC | | | | | | | |
| Full | 1981 | 623 | (0.755) | 0.654 | 0.670 | 0.671 | 0.634 |
| LNN | 1037 | 333 | (0.718) | 0.637 | 0.648 | 0.650 | 0.641 |
| ERP | 1526 | 447 | (0.749) | 0.641 | 0.668 | 0.657 | 0.612 |

NOTE: Highlighted in blue are the values achieving highest score in each case. Highlighted in yellow and bold are the scores for which the corresponding *P* value of comparison with the BCAM score is less than 0.05. The last set of rows contains the scores from the METABRIC dataset and the listed BCAM scores result from applying the formula on the entire dataset. These cannot be compared with other scores because METABRIC was used for BCAM training. ROR-S uses the gene expression–based PAM50 assay; ROR-C uses the gene expression plus tumor size–based PAM50 assay; 21-gene uses the Oncotype DX 21-gene assay; and 70-gene uses the Mammaprint 70-gene assay.

model, such as the expression levels of genes *FGD3* and *SUSD3*. The four available independent datasets that we used have relatively small sample sizes and are difficult to pool. Furthermore, the estimated performance of existing biomarker assays on platforms different from those used in the actual product implementation is only suggestive and cannot be used as a definitive proxy of the actual performance. However, as a general justifying principle, relative gene expression has been shown to be reliable across platforms and measurement techniques. For example, a high level of interplatform concordance in terms of genes identified as differentially expressed was demonstrated (22) by comparing three RT-PCR platforms and seven microarrays, including Illumina, Affymetrix, and Agilent.

There are several published examples of across-platform comparisons of breast cancer biomarker products: the predictions from five gene expression–based models, including those used in Oncotype DX, MammaPrint, and intrinsic subtypes have been compared with each other (23) using Agilent microarray data in all cases. Furthermore, the ability of six genomic signatures, including Oncotype DX, MammaPrint, and PAM50-ROR to predict relapse in breast cancer patients with $ER^+$ tumors treated with adjuvant tamoxifen, was evaluated (24) using only four Affymetrix microarray datasets. As the authors acknowledge in that work, one important caveat to their analyses that must be recognized and always kept in mind when interpreting across platform genomic studies, such as the one presented here, is that "although we strove to implement each predictor as published, signatures developed on platforms other than the Affymetrix U133A (used in their work) were suboptimally implemented." Such imperfect comparisons, however, are still valuable. Many other publications, including some recently published ones (25, 26), also use microarray values to compare breast cancer prognostic signatures.

Accordingly, the aim of our study is to demonstrate that our results are promising and to suggest that they should

be rigorously evaluated in the context of larger-scale clinical studies either under way or being planned.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Authors' Contributions

**Conception and design:** T.-H. Ou Yang, W.-Y. Cheng, M.A. Maurer, D. Anastassiou
**Development of methodology:** T.-H. Ou Yang, W.-Y. Cheng, D. Anastassiou
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** T.-H. Ou Yang
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** T.-H. Ou Yang, W.-Y. Cheng, T. Zheng, M.A. Maurer, D. Anastassiou
**Writing, review, and/or revision of the manuscript:** T.-H. Ou Yang, W.-Y. Cheng, T. Zheng, M.A. Maurer, D. Anastassiou
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** T.-H. Ou Yang
**Study supervision:** D. Anastassiou

## Acknowledgments

We thank Sage Bionetworks for providing the uniformly renormalized METABRIC dataset which was accessed through Synapse (synapse.sagebase.org). The original METABRIC dataset (5) was generated by the Molecular Taxonomy of Breast Cancer International Consortium.

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.

Received April 18, 2014; revised August 6, 2014; accepted September 5, 2014; published OnlineFirst September 23, 2014.

## References

1. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
2. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
3. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.
4. Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Sci Transl Med 2013;5:181re1.
5. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–52.
6. Cheng WY, Ou Yang TH, Anastassiou D. Development of a prognostic model for breast cancer survival in an open challenge environment. Sci Transl Med 2013;5:181ra50.
7. McCarthy N. Prognostic models: rising to the challenge. Nat Rev Cancer 2013;13:378.
8. Cheng WY, Ou Yang TH, Anastassiou D. Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput Biol 2013;9:e1002920.
9. Loi S, Haibe-Kains B, Majjaj S, Lallemand F, Durbecq V, Larsimont D, et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. Proc Natl Acad Sci U S A 2010;107:10208–13.
10. Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, et al. microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res 2011;71:5635–45.
11. Li Y, Zou L, Li Q, Haibe-Kains B, Tian R, Li Y, et al. Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. Nat Med 2010;16:214–8.
12. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, 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.
13. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004;23:2109–23.
14. Schroder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics 2011;27:3206–8.
15. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
16. Cheng WY, Ou Yang TH, Shen H, Laird PW, Anastassiou DThe Cancer Genome Atlas Research Network. Multi-cancer molecular signatures and their interrelationships. arXiv:1306.2584v2; 2013.
17. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
18. Roepman P, Horlings HM, Krijgsman O, Kok M, Bueno-de-Mesquita JM, Bender R, et al. Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. Clin Cancer Res 2009;15:7003–11.
19. Sterrenberg JN, Blatch GL, Edkins AL. Human DNAJ in cancer and stem cells. Cancer Lett 2011;312:129–42.
20. Boimel PJ, Smirnova T, Zhou ZN, Wyckoff J, Park H, Coniglio SJ, et al. Contribution of CXCL12 secretion to invasion of breast cancer cells. Breast Cancer Res 2012;14:R23.
21. Moy I, Todorovic V, Dubash AD, Coon JS, Parker JB, Buranapramest M, et al. Estrogen-dependent sushi domain containing 3 regulates cytoskeleton organization and migration in breast cancer cells. Oncogene 2014 Jan 13. [Epub ahead of print].
22. MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006;24:1151–61.
23. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, et al. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006;355:560–9.
24. Prat A, Parker JS, Fan C, Cheang MC, Miller LD, Bergh J, et al. Concordance among gene expression-based predictors for ER-positive breast cancer treated with adjuvant tamoxifen. Ann Oncol 2012;23:2866–73.
25. Jonsdottir K, Assmus J, Slewa A, Gudlaugsson E, Skaland I, Baak JP, et al. Prognostic value of gene signatures and proliferation in lymph-node-negative breast cancer. PLoS ONE 2014;9:e90642.
26. Zhao X, Rodland EA, Sorlie T, Vollan HK, Russnes HG, Kristensen VN, et al. Systematic assessment of prognostic gene signatures for breast cancer shows distinct influence of time and ER status. BMC Cancer 2014;14:211.