Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Network and Pathway Analysis of Cancer Susceptibility (A)

Pathway-based Biomarkers for Breast Cancer in Proteomics

Fan Zhang[1,2], Youping Deng[3], Mu Wang[4,5], Li Cui[6] and Renee Drabier[1]

[1]Department of Academic and Institutional Resources and Technology, University of North Texas Health Science Center, Fort Worth, TX, USA. [2]Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA. [3]Department of Internal Medicine, Rush University Medical Center, Chicago, IL, USA. [4]Department of Biochemistry and Molecular Biology, IU School of Medicine, Indianapolis, IN, USA. [5]Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN, USA. [6]Department of Neurosciences, School of Medicine, University of San Diego, La Jolla, CA, USA.

ABSTRACT: Genes do not function alone but through complex biological pathways. Pathway-based biomarkers may be a reliable diagnostic tool for early detection of breast cancer due to the fact that breast cancer is not a single homogeneous disease. We applied Integrated Pathway Analysis Database (IPAD) and Gene Set Enrichment Analysis (GSEA) approaches to the study of pathway-based biomarker discovery problem in breast cancer proteomics. Our strategy for identifying and analyzing pathway-based biomarkers are threefold. Firstly, we performed pathway analysis with IPAD to build the gene set database. Secondly, we ran GSEA to identify 16 pathway-based biomarkers. Lastly, we built a Support Vector Machine model with three-way data split and fivefold cross-validation to validate the biomarkers. The approach–unraveling the intricate pathways, networks, and functional contexts in which genes or proteins function–is essential to the understanding molecular mechanisms of pathway-based biomarkers in breast cancer.

KEYWORDS: pathway analysis, breast cancer, proteomics, machine learning

SUPPLEMENT: Network and Pathway Analysis of Cancer Susceptibility (A)

CITATION: Zhang et al. Pathway-based Biomarkers for Breast Cancer in Proteomics. *Cancer Informatics* 2014:13(S5) 101–108 doi: 10.4137/CIN.S14069.

RECEIVED: July 1, 2014. **RESUBMITTED:** August 29, 2014. **ACCEPTED FOR PUBLICATION:** September 3, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Methodology

FUNDING: Funding support was provided by the Bioinformatics Program at the University of North Texas Health Science Center. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Fan.Zhang@unthsc.edu

Introduction

Breast cancer is the most common cancer among American women, except for skin cancers. About 12% women in the US will develop invasive breast cancer during their lifetime.[1] According to the American Cancer Society, in 2014 in the US, approximately 232,670 new cases of invasive breast cancer will be diagnosed in women, about 62,570 new cases of carcinoma in situ (CIS) will be diagnosed (CIS is noninvasive and is the earliest form of breast cancer), and about 40,000 women will die from breast cancer. Early detection of malignant breast cancer tumor is critical to the prevention of cancer deaths and successful cancer treatment.

Researchers have shown that functional genomics studies using DNA Microarrays is effective in detecting the difference between healthy breast tissues and breast cancer tissue by measuring thousands of differentially expressed genes simultaneously.[2–4] However, early cancer detection and treatment are still challenging due to multiple reasons. In addition to the difficulty of obtaining tissue samples for microarray analysis, another reason is that breast cancer consists of multiple disease status, each arising from a distinct molecular mechanism and having a distinct clinical progression path,[5] which makes the disease difficult to detect in early stages. Therefore, there is a need for us to develop more reliable diagnostic tools for early detection of breast cancer. Pathway-based biomarkers can be one of the tools because compelling evidences have been provided that these genes function in a pathway regulating human cancers.[6] Another example supporting for focusing on pathways rather than individual genes is the studies of the TP53 tumor-suppressor gene,[6] which is a transcription factor that normally inhibits cell growth and stimulates cell death when induced by cellular stress.[7–10]

Genes do not function alone but through complex biological pathways. As more information is revealed through large-scale "omics" techniques, it is becoming increasingly apparent. Unraveling these intricate pathways and analyzing their functions is essential to understanding biological mechanisms. Extensive function, pathway, and network analysis allowed for the discovery of highly significant pathways from a set of disease versus healthy samples. Knowledge of activation of these processes will lead to novel assays identifying their proteomic signatures in the plasma of patients at high risk for cancer disease.

Gene set enrichment analysis (GSEA) has been widely used for finding significantly affected pathways from various experimental platforms. For example, Wang et al developed an R package named SeqGSEA to derive biological insight by integrating differential expression and splicing from RNA-Seq data with functional gene set analysis. This approach built count data model with negative binomial distributions to first score differential expression and splicing in each gene, respectively, followed by two strategies to combine the two scores for integrated GSEA. The R package SeqGSEA is particularly useful for efficiently translating RNA-Seq data to biological discoveries.[11,12] Holden et al combined Single-nucleotide polymorphism (SNP) association analysis with the pathway-driven GSEA, to facilitate handling of genome-wide gene expression data. The method they developed, GSEA-SNP, may facilitate the identification of disease-associated SNPs and pathways, as well as the understanding of the underlying biological mechanisms.[13]

Therefore, we presented an approach that combines Integrated Pathway Analysis Database (IPAD) and GSEA together for the identification of pathway-based biomarkers from breast cancer proteomics. Moreover, a Support Vector Machine (SVM) model with three-way data split and five-fold cross-validation was used to validate the prediction performance for the pathway-based biomarkers identified. We believe that the approach can help us understand molecular mechanisms of pathway-based biomarkers in breast cancer.

Methods

Human plasma samples. Ammonium carbonate, ammonium bicarbonate, urea, formic acid, lysozyme, 2-iodoethanol, and triethylphosphine were all purchased from Sigma-Aldrich (St. Louis, MO, USA). Acetonitrile and Mass Spectrometry (MS)-grade water were purchased from Honey Burdick & Jackson (Morristown, NJ, USA). Trypsin was purchased from Worthington Biochemical Corporation (Lakewood, NJ, USA). Seppro tip IgY-12 and reagent kit were purchased from GenWay Biotech (San Diego, CA, USA).

The Hoosier Oncology Group (HOG; Indianapolis, IN, USA) collected the two batches of plasma samples: Study A and Study B. Each study contains 40 plasma samples from women diagnosed with breast cancer and 40 plasma samples from healthy volunteer women as control. All participants gave their written, informed consent for collection and use of the samples. This study was approved by the Indiana University institutional review board and conducted in accordance with the principles of the Declaration of Helsinki. We collected all samples in the two studies with the same standard operating procedure and stored them in a central repository in Indianapolis, IN, USA. The two batches were processed at different times in the same laboratory. We analyzed each sample in a single batch by mass spectrometry. The demography and clinical distribution of breast cancer stages for the two studies are comparable.

Liquid Chromatography Tandem Mass Spectrometry (LC/MS/MS) plasma proteomics analysis. Thermo-Fisher Scientific linear ion-trap mass spectrometer (Linear Trap Quadropole) coupled with a Surveyor HPLC system was used to analyze tryptic peptides and to identify proteins. We first eluted peptides with a gradient from 5% to 45% acetonitrile, which was developed over 120 minutes at a flow rate of 50 µL/minute. Then, we collected data in the "triple-play" mode (primary MS scan, zoom scan, and MS/MS scan).[14] Lastly, we generated the raw peak list data with XCalibur (version 2.0), which were further analyzed by a label-free identification and quantitative algorithm described by Higgs et al.[15]

The MS database searching was performed against the International Protein Index (version 3.6).[16] We used the same algorithm in[15] to carry out protein quantification. First, we aligned all extracted ion chromatograms along retention time. We matched each aligned peak by parent ion, charge state, daughter ions (MS/MS data), and retention time within a one-minute window. Then, we measured, normalized, and compared the area under the curve (AUC) for each individually aligned peak for their relative abundance using methods described in Wang et al and Higgs et al.[14,15]

Linear mixed model. All peak intensities were transformed to a log scale before quantile normalization,[17] and the protein intensity is fit by a separate analysis of variance statistical model for each protein as y_{ijk} using the following:

$$y_{ijk} = \mu + T_j + S_k + I_i + \varepsilon_{ijk} \tag{1}$$

where $I_i \in N(0,\sigma_1^2), S_k \in N(0,\sigma_2^2), \varepsilon_{ijk} \in N(0,\sigma^2)$ Here, μ is the average intensity value, T_j is the fixed group effect that is caused by the experimental conditions or treatments, S_k is the random sample effect that is from either individual biological samples or sample preparations, I_i is the random replicate effect that is from replicate injections of the same sample, and ε_{ijk} is the within-group errors. S_k and I_i are assumed independent of each other and independent of the within-group errors ε_{ijk}.

Pathway-based biomarker discovery. We use the GSEA R package (GSEA-P-R–1.0)[18] to determine all significantly affected pathway-based biomarkers. GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states.[18] The primary result of the GSEA is the enrichment score (ES), which reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes.[18] GSEA calculates the ES by walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not. The magnitude of the increment depends on the correlation of the gene with the phenotype. The ES is the maximum deviation from zero encountered in walking down the list. A positive ES indicates gene set enrichment at the top of the ranked list; a negative ES indicates gene set enrichment at the bottom of the ranked list.

The basic steps we run the analysis in GSEA are as follows: 1) building the gene set database (.gmt file) by performing pathway analysis with the IPAD (http://bioinfo.hsc.unt.edu/ipad/);[19] 2) creating the expression dataset file (.gct file) and phenotype labels file (.cls file); 3) setting the analysis parameters and run the analysis; and 4) analyzing the results and identifying the significantly affected pathway-based biomarker.

Support vector machine. An SVM is a discriminative classifier formally defined by a separating hyperplane $<w, x> + b = 0$. The separation is considered to be optimal if the set of patterns is separated by the hyperplane without error and the distance between the closest pattern to the hyperplane is maximal. Without loss of generality, it is appropriate to consider a canonical hyperplane,[20] where the parameters w, b are constrained by $\min_{i} \text{in}|<w, x_i> +b| = 1$.

For the use of the SVM as an appropriate tool for validating the pathway-based breast cancer biomarkers, a three-way data split is applied for training, validation, and testing. Briefly, we used Study A as both the "training set" for learning to fit the parameters of the classifier, and the "validation set" to tune the parameters of the classifier, and used Study B as the "testing set" only to assess the performance of the fully trained classifier. We randomly partitioned Study A into subsamples. For each subsample, a cross-section of the data was flagged for use as the validation set and a new model was created by training on the remaining data, which are the training set and not in the subsample. We then repeated the cross-validation process, with each of the subsamples used exactly once as the validation data. Lastly, we averaged the results from the folds to produce a single estimation.

Performance measurements.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TP}+\text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{TP+\text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}=\text{TN}=\text{FP}+\text{FN}}$$

where TP = true positive rate, TN = true negative rate, FP = false positive rate, and FN = false negative rate.

We used the following five measurements for our evaluation: (1) Sensitivity (also called recall), the proportion of actual positive pairs which are correctly identified; (2) Specificity, the proportion of negative pairs that are correctly identified; (3) Precision, the probability of correct positive prediction; (4) Accuracy, the proportion of correctly predicted pairs; and (5) AUC.

Results

The plasma proteome sets from Study A and B contain 1423 and 1389 proteins. A total of 615 proteins are in common between the two datasets. We used the IPAD to perform pathway analysis and obtained 1261 pathways in Study A and 1277 pathways in Study B.

We downloaded the R source code (GSEA-P-R-1.0),[18] and built the gene set databases (.gmt files) for Study A based on the result of pathway analysis. The geneset database (.gmt) file is a tab-separated text file containing one gene set per line. The first column is the pathway names. The second column is a brief description of the pathway. The remaining columns contain the names of the genes in the pathway.

To analyze experimental data, we needed to create two text files: a sample phenotype file (.cls) and a gene expression file (.gct). A sample phenotype (.cls) file is a text file containing three lines. The first line contains three numbers separated by spaces. The first number is the number of samples. The second and third numbers are the constants 2 and 1, respectively. The second line begins with # and is followed by a space-separated list of "long" phenotype names. The third line consists of a space-separated list of "short" phenotype labels for each of the samples in the gene expression file, in the same order in which they occur there. A gene expression (.gct) file is a tab-separated text file. The first line is the constant #1.2. The second line contains two numbers: the number of genes and the number of samples. The third line contains column headers for the table on the following lines. The fourth line and below each contain expression values for one gene: 1) the first column is the gene name, 2) the second column is the gene description, 3) the third and subsequent columns are the expression values for each sample. In our case, the gene expression file is actually a protein intensity file. In order to keep consistent with pathway analysis, we used gene IDs instead of the original protein IDs in the first column.

GSEA reports ES and false discovery rate (FDR) for the GSEA. The ES reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes. GSEA calculates the ES by walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not.[18] The FDR is the estimated probability that a gene set with a given Normalized Enrichment Score (NES) represents a false positive finding.[18] For example, an FDR of 25% indicates that the result is likely to be valid three out of four times. We chose ES greater than 0.5 and FDR less than 0.25 as thresholds and identified 16 pathway-based biomarkers (Table 1). In general, given the lack of coherence in most expression datasets and the relatively small number of gene sets being analyzed, an FDR cutoff of 25% and ES cutoff of 0.5% are appropriate. The average values for ER and FDR for the 16 pathway-based biomarkers are 0.6428 and 0.1646, respectively.

We built a SVM model with fivefold cross-validation and three-way data split to validate the 16 pathway-based biomarkers. We first randomly partitioned Study A into subsamples with fivefold cross-validation. For each subsample, a cross-section of the data is flagged for use as the validation set, and a new model is created by training on the remaining data, which are the training set and not in the subsample. We then used Study B as the testing set to assess the performance of the fully trained classifier. The testing set is totally independent of the training set. No data from the testing set were utilized in 1) identification of 16 pathway-based biomarkers or 2) development of the SVM model.

We trained the SVM with radius basis function kernels function and fivefold cross-validation. Sensitivity, specificity, precision, accuracy, receiver operating characteristic (ROC) curve, and AUC were calculated to help evaluate the predictive performance of the 16 pathway-based biomarkers.

We obtained high performances: for the training set (mean AUC = 0.9075, mean precision = 80.76%, mean accuracy = 80.70%, mean sensitivity = 80.63%, mean specificity = 80.78%) and for the testing set (mean AUC = 0.8350, mean precision = 73.29%, mean accuracy = 76.56%, mean sensitivity = 82.03%, mean specificity = 71.09%) (Table 2). The result shows that pathway-based biomarkers we identified using GSEA can be used as predictors for improving the prediction accuracy for both the training set and the testing set. For example, the G Protein-Coupled Receptor (GPCR) downstream signaling pathway achieves the highest performance with AUC = 0.9481, precision = 82.98%, accuracy = 88.75%, sensitivity = 97.50%, and specificity = 80.00% (Table 3 and Figure 1).

Discussion

In this study, we combined IPAD and GSEA together to identify 16 pathway-based biomarkers from breast cancer proteomics and then we used SVM with three-way data split and fivefold cross-validation to validate the prediction performance of the 16 pathway-based biomarkers.

GSEA appears to have greater power than single-gene analysis in detecting small but biologically important changes in a set of genes. It can help to find significantly affected pathways within a gene expression data.

The IPAD contains about 22,498 genes, 25,469 proteins, 1,956 pathways, 6,704 diseases, 5,615 drugs, and 52 organs integrated from databases including the BioCarta,[17] KEGG,[21] NCI-Nature curated,[22] Reactome,[23] CTD,[24] PharmGKB,[25] DrugBank,[26] and HOMER.[27] It can provide reliable pathway–gene relationship for the gene set database in GSEA.

The 16 pathway-based biomarkers we identified are signaling, complement pathway, binding receptors, and metabolism (Table 1), which are consistent with previous findings.[28]

Table 1. 16 pathway-based biomarkers identified.

PATHWAY ID	PATHWAY NAME	AE	N	ES	FDR
h_classicPathway	Classical complement pathway	14	14	0.70549	0.07245
500792	GPCR ligand binding	16	409	0.60928	0.08991
hsa05322	Systemic lupus erythematosus	17	141	0.68254	0.10214
166663	Initial triggering of complement	13	63	0.64665	0.10883
h_lectinPathway	Lectin-induced complement pathway	10	12	0.76798	0.1105
h_compPathway	Complement pathway	16	19	0.66422	0.12402
200149	PDGFR-beta signaling pathway	12	129	0.70522	0.16918
hsa04610	Complement and coagulation cascades	42	71	0.50477	0.17074
373076	Class A/1 (Rhodopsin-like receptors)	10	304	0.7634	0.1732
375276	Peptide ligand–binding receptors	10	185	0.7634	0.1732
392499	Metabolism of proteins	18	378	0.56802	0.21487
381150	Diabetes pathways	15	137	0.57968	0.21726
388396	GPCR downstream signaling	27	866	0.48634	0.22031
hsa05133	Pertussis	18	84	0.5792	0.22171
hsa05150	*Staphylococcus aureus* infection	16	69	0.58669	0.22746
418594	G alpha (i) signaling events	12	199	0.67202	0.23845

Notes: AE is the number of genes presented in the pathway. *N* is the total number of genes in the pathway.

Table 2. Validation with SVM for the 16 pathway-based biomarkers.

	TRAINING SET					TESTING SET				
PATHWAY ID	PRECISION (%)	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	ROC (%)	PRECISION (%)	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	ROC (%)
h_classicPathway	80.56	77.50	72.50	82.50	87.88	71.43	72.50	75.00	70.00	78.63
500792	79.49	78.75	77.50	80.00	90.63	74.47	78.75	87.50	70.00	89.00
hsa05322	73.17	73.75	75.00	72.50	88.88	69.39	73.75	85.00	62.50	82.19
166663	77.27	80.00	85.00	75.00	88.94	69.39	73.75	85.00	62.50	77.81
h_lectinPathway	73.91	77.50	85.00	70.00	89.56	64.29	65.00	67.50	62.50	75.63
h_compPathway	80.00	80.00	80.00	80.00	90.44	70.45	72.50	77.50	67.50	79.19
200149	72.22	70.00	65.00	75.00	78.69	57.14	53.75	30.00	77.50	65.19
hsa04610	86.67	91.25	97.50	85.00	96.94	79.59	86.25	97.50	75.00	88.81
375276	80.56	77.50	72.50	82.50	89.69	75.56	78.75	85.00	72.50	86.75
373076	80.56	77.50	72.50	82.50	89.69	75.56	78.75	85.00	72.50	86.75
392499	82.05	81.25	80.00	82.50	93.25	78.26	82.50	90.00	75.00	91.88
381150	81.40	83.75	87.50	80.00	93.44	76.47	83.75	97.50	70.00	88.38
388396	92.50	92.50	92.50	92.50	98.19	82.98	88.75	97.50	80.00	94.81
hsa05133	81.58	80.00	77.50	82.50	90.94	73.81	75.00	77.50	72.50	81.44
hsa05150	84.09	87.50	92.50	82.50	93.81	73.91	77.50	85.00	70.00	81.88
418594	86.11	82.50	77.50	87.50	91.13	80.00	83.75	90.00	77.50	87.75
Mean	80.76	80.70	80.63	80.78	90.75	73.29	76.56	82.03	71.09	83.50

We further evaluated the prediction performance of our pathway-based biomarkers by comparing our results with prediction performances in previously published findings. For example, Aaroe et al identified a set of 738 differentially expressed probes that achieved an estimated prediction accuracy of 79.5% with a sensitivity of 80.6% and a specificity of 78.3%[28] and Sharma et al identified a panel of 37 genes that permitted early detection with a classification accuracy of 82%.[29] These prediction results were based on the training set, not the independent testing set. For the testing set, our pathway-based biomarkers show a similar prediction performance to theirs. However, for the training set, prediction performance of our pathway-based biomarkers is higher than theirs (Table 2). In addition, we also did the feature selection based on all proteins. With a *P* value cutoff <0.001, we identified 72 protein biomarkers in Study A. After SVM learning, we obtained the prediction performances: for the training set (AUC = 0.9769, precision = 88.10%, accuracy = 90.00%, sensitivity = 92.50%, specificity = 87.50%) and for the testing set (AUC = 0.9188, precision = 81.25%, accuracy = 87.50%, sensitivity = 97.50%, specificity = 77.50%). The prediction performances based on proteins are lower than the highest performance of our pathway-based approach. When we chose the top 17 proteins in Study A as biomarkers, we obtained lower prediction performance (AUC = 0.8138 for the testing set) than the mean performance based on the pathway-based biomarkers (AUC = 0.8350 for the testing set).

Table 3. Prediction result for the GPCR downstream signaling.

	TRAINING SET		TESTING SET	
PREDICTED	CANCER	NORMAL	CANCER	NORMAL
Cancer	37	3	39	8
Normal	3	37	1	32
Precision		92.50%		82.98%
Accuracy		92.50%		88.75%
Sensitivity		92.50%		97.50%
Specificity		92.50%		80.00%

An interesting observation in our study is that some of genes in the pathway-based biomarker are not differentially expressed between cancer and normal, for example, in the GPCR downstream signaling pathway, ADRBK1 (*P* value = 0.68), AGT (*P* value = 0.94) and OR7D4 (*P* value = 0.69) with high a *P* value. After we removed all proteins with *P* value ≥ 0.001 in the pathway, the prediction performances dropped dramatically (Table 4). It suggests that the genes with a high *P* value can still be valuable in a pathway, compared with conventional methods, which usually limit genes to those with change below a *P* value threshold such as 0.001. In conventional methods, a maximal *P* value threshold had to be enforced for selection of differentially expressed genes/proteins to control false positives. This is because while gene expression profiling using microarray is a powerful technology with potential to enhance the molecular understanding of tumors, the sources of variability due to patient heterogeneity, tumor heterogeneity, replicate variability, and technical variability makes it difficult to set a

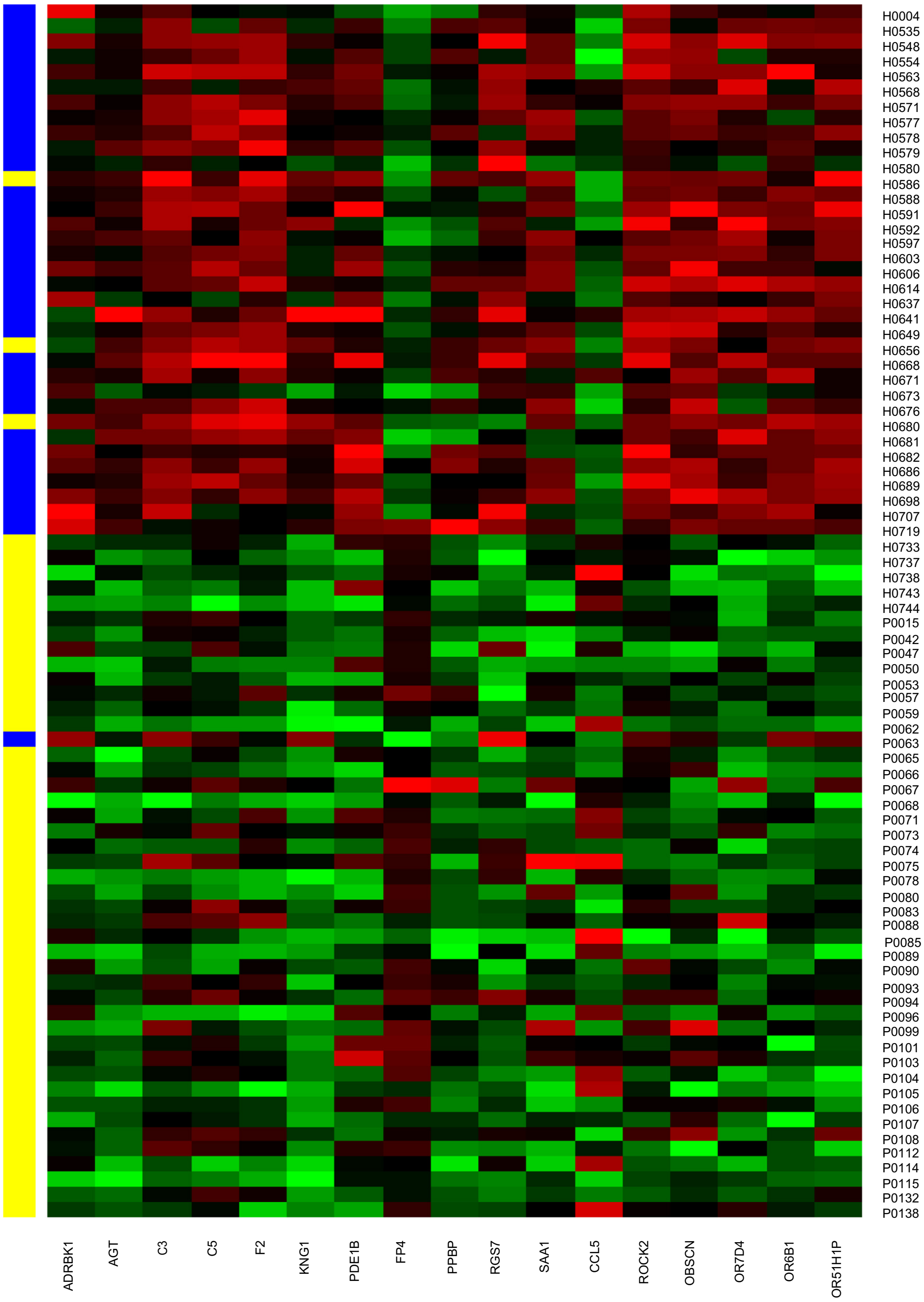

Figure 1. Seventeen genes pathway-based biomarkers predicting the healthy and breast cancer samples in testing set. X axis shows the 17 genes in the GPCR downstream signaling pathway. Y-axis shows the 40 breast cancer and 40 healthy samples.

Notes: H, healthy; P, cancer; blue, predicted as healthy; yellow, predicted as cancer.

P value threshold. Prior genomic studies have shown that simple *P* value thresholds were too stringent at high expression values and not stringent enough at low expression values.[30] For example, at low expression values, replicate variability is much higher than commonly used thresholds. Our SVM method, on the other hand, was able to pick up gene expression change patterns at *P* value levels above those usually used in conventional method, primarily because nonlinear cooperative relationships among biomarker expression patterns on the same network were trained and learned by SVM.

Conclusion

We developed an integrated computational approach that addressed a challenging pathway network biomarker development problem in the early detection of breast cancer from proteomics. The approach combined GSEA with IPAD and SVM. Briefly, first, we performed pathway analysis using IPAD and built a gene set for GSEA; then we ran GSEA to identify the 16 pathway-based biomarkers; lastly, we validated the prediction accuracy using an SVM with three-way data split and fivefold cross-validation.

Table 4. Prediction performance after removing all "insignificant" proteins in pathways.

	TRAINING SET					TESTING SET				
PATHWAY ID	PRECISION (%)	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	ROC (%)	PRECISION (%)	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	ROC (%)
h_classicPathway	71.25	67.50	75.00	85.44	65.85	66.25	67.50	65.00	75.00	71.25
500792	70.00	77.50	62.50	76.75	57.14	56.25	50.00	62.50	63.00	70.00
hsa05322	73.75	75.00	72.50	86.31	69.39	73.75	85.00	62.50	75.25	73.75
166663	75.00	67.50	82.50	84.06	60.00	55.00	30.00	80.00	74.06	75.00
h_lectinPathway	70.00	65.00	75.00	85.00	62.86	61.25	55.00	67.50	74.25	70.00
h_compPathway	71.25	67.50	75.00	85.44	65.85	66.25	67.50	65.00	75.00	71.25
200149	70.00	60.00	80.00	71.41	57.89	53.75	27.50	80.00	47.28	70.00
hsa04610	77.50	75.00	80.00	88.50	71.79	71.25	70.00	72.50	78.81	77.50
375276	70.00	77.50	62.50	76.75	57.14	56.25	50.00	62.50	63.00	70.00
373076	70.00	77.50	62.50	76.75	57.14	56.25	50.00	62.50	63.00	70.00
392499	73.75	70.00	77.50	71.16	66.67	58.75	35.00	82.50	60.06	73.75
381150	71.25	75.00	67.50	83.38	56.00	53.75	35.00	72.50	71.88	71.25
388396	70.00	77.50	62.50	76.75	57.14	56.25	50.00	62.50	63.00	70.00
hsa05133	75.00	70.00	80.00	85.00	62.07	58.75	45.00	72.50	74.00	75.00
hsa05150	73.75	65.00	82.50	84.00	64.52	61.25	50.00	72.50	74.06	73.75
418594	70.00	77.50	62.50	76.75	57.14	56.25	50.00	62.50	63.00	70.00
MEAN	72.03	71.56	72.50	80.84	61.79	60.08	51.09	69.06	68.42	72.03

The approach achieved high prediction performances: for the training set (mean AUC = 0.9075, mean precision = 80.76%, mean accuracy = 80.70%, mean sensitivity = 80.63%, mean specificity = 80.78%) and for the testing set (mean AUC = 0.8350, mean precision = 73.29%, mean accuracy = 76.56%, mean sensitivity = 82.03%, mean specificity = 71.09%) (Table 2). Our results show that the pathway-based biomarker identification method can be used as a predictor to improve the prediction accuracy for both the training set and the testing set. We believe this computational approach can be helpful for biomarker discovery in early stages of breast cancer and can also provide general guidance for pathway network discovery applications in other diseases. In the future, we will follow-up with biological experiments to validate these biomarkers with our collaborators.

Acknowledgments

We thank HOG for collecting breast cancer plasma samples.

Author Contributions

Conceived the initial work and designed the method: FZ, RD. Developed the data mining and validation method and performed all the computational analyses: FZ. Designed the experiment: MW. Performed literature search: LC. Validated markers for breast cancer: YD. Made critical revisions and approved final version: FZ, YD, MW, LC, RD. All authors reviewed and approved of the final manuscript.

REFERENCES

1. What are the key statistics about breast cancer? http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics.
2. Hu X, Zhang Y, Zhang A, et al. Comparative serum proteome analysis of human lymph node negative/positive invasive ductal carcinoma of the breast and benign breast disease controls via label-free semiquantitative shotgun technology. *OMICS*. 2009;13(4):291–300.
3. Zeidan BA, Cutress RI, Murray N, et al. Proteomic analysis of archival breast cancer serum. *Cancer Genomics Proteomics*. 2009;6(3):141–7.
4. Lebrecht A, Boehm D, Schmidt M, Koelbl H, Schwirz RL, Grus FH. Diagnosis of breast cancer by tear proteomic pattern. *Cancer Genomics Proteomics*. 2009;6(3):177–82.
5. Polyak K. Breast cancer: origins and evolution. *J Clin Invest*. 2007;117(11):3155–63.
6. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med*. 2004;10(8):789–99.
7. Oren M. Decision making by p53: life, death and cancer. *Cell Death Differ*. 2003;10(4):431–42.
8. Prives C, Hall PA. The p53 pathway. *J Pathol*. 1999;187(1):112–26.
9. Ichimura K, Bolin MB, Goike HM, Schmidt EE, Moshref A, Collins VP. Deregulation of the p14ARF/MDM2/p53 pathway is a prerequisite for human astrocytic gliomas with G1-S transition control gene abnormalities. *Cancer Res*. 2000;60(2):417–24.
10. Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. *Nature*. 2000;408(6810):307–10.
11. Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. *Bioinformatics*. 2014;30(12):1777–9.
12. Wang X, Cairns MJ. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinformatics*. 2013;14(suppl 5):S16.
13. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*. 2008;24(23):2784–5.
14. Wang M, You J, Bemis KG, Tegeler TJ, Brown DP. Label-free mass spectrometry-based protein quantification technologies in proteomic analysis. *Brief Funct Genomic Proteomic*. 2008;7(5):329–39.
15. Higgs RE, Knierman MD, Gelfanova V, Butler JP, Hale JE. Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res*. 2005;4(4):1442–50.

16. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*. 2004;4(7):1985–8.
17. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185–93.
18. 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(43):15545–50.
19. Zhang F, Drabier R. IPAD: the integrated pathway analysis database for systematic enrichment analysis. *BMC Bioinformatics*. 2012;13(Suppl 15):S7.
20. Vapnik VN. *Statistical Learning Theory*. New York: Springer; 1998.
21. Victor KG, Rady JM, Cross JV, Templeton DJ. Proteomic profile of reversible protein oxidation using prop, purification of reversibly oxidized proteins. *PLoS One*. 2012;7(2):e32527.
22. Schaefer CF, Anthony K, Krupa S, et al. PID: the pathway interaction database. *Nucleic Acids Res*. 2009;37(Database issue):D674–9.
23. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39(Database issue):D691–7.
24. Davis AP, King BL, Mockus S, et al. The comparative toxicogenomics database: update 2011. *Nucleic Acids Res*. 2011;39(Database issue):D1067–72.
25. McDonagh EM, Whirl-Carrillo M, Garten Y, Altman RB, Klein TE. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark Med*. 2011;5(6):795–806.
26. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. 2011;39(Database issue):D1035–41.
27. Zhang F, Chen JY. HOMER: a human organ-specific molecular electronic repository. *BMC Bioinformatics*. 2011;12(suppl 10):S4.
28. Aaroe J, Lindahl T, Dumeaux V, et al. Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Res*. 2010;12(1):R7.
29. Sharma P, Sahni NS, Tibshirani R, et al. Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. *Breast Cancer Res*. 2005;7(5):R634–44.
30. Mariani TJ, Budhraja V, Mecham BH, Gu CC, Watson MA, Sadovsky Y. A variable fold change threshold determines significance for expression microarrays. *FASEB J*. 2003;17(2):321–3.