RESEARCH NOTE

REVISED Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning [version 3; referees: 2 approved]

Eliseos J. Mucaki[1], Katherina Baranova[1], Huy Q. Pham[2], Iman Rezaeian[2], Dimo Angelov[3], Alioune Ngom[2], Luis Rueda[2], Peter K. Rogan [1,3,4]

[1]Deparment of Biochemistry , University of Western Ontario, London, Canada
[2]School of Computer Science, University of Windsor, Windsor, Canada
[3]Department of Computer Science, University of Western Ontario, London, Canada
[4]CytoGnomix Inc, London, Canada

v3 **First published:** 31 Aug 2016, **5**:2124 (doi: 10.12688/f1000research.9417.1)
Second version: 27 Jan 2017, **5**:2124 (doi: 10.12688/f1000research.9417.2)
Latest published: 12 May 2017, **5**:2124 (doi: 10.12688/f1000research.9417.3)

Abstract

Genomic aberrations and gene expression-defined subtypes in the large METABRIC patient cohort have been used to stratify and predict survival. The present study used normalized gene expression signatures of paclitaxel drug response to predict outcome for different survival times in METABRIC patients receiving hormone (HT) and, in some cases, chemotherapy (CT) agents. This machine learning method, which distinguishes sensitivity vs. resistance in breast cancer cell lines and validates predictions in patients; was also used to derive gene signatures of other HT (tamoxifen) and CT agents (methotrexate, epirubicin, doxorubicin, and 5-fluorouracil) used in METABRIC. Paclitaxel gene signatures exhibited the best performance, however the other agents also predicted survival with acceptable accuracies. A support vector machine (SVM) model of paclitaxel response containing genes *ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCL2, BCL2L1, BMF, CYP2C8, CYP3A4, MAP2, MAP4, MAPT, NR1I2, SLCO1B3, TUBB1, TUBB4A,* and *TUBB4B* was 78.6% accurate in predicting survival of 84 patients treated with both HT and CT (median survival ≥ 4.4 yr). Accuracy was lower (73.4%) in 304 untreated patients. The performance of other machine learning approaches was also evaluated at different survival thresholds. Minimum redundancy maximum relevance feature selection of a paclitaxel-based SVM classifier based on expression of genes *BCL2L1, BBC3, FGF2, FN1,* and *TWIST1* was 81.1% accurate in 53 CT patients. In addition, a random forest (RF) classifier using a gene signature (*ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCL2, BCL2L1, BMF, CYP2C8, CYP3A4, MAP2, MAP4, MAPT, NR1I2,SLCO1B3, TUBB1, TUBB4A,* and *TUBB4B*) predicted >3-year survival with 85.5% accuracy in 420 HT patients. A similar RF gene signature showed 82.7% accuracy in 504 patients treated with CT and/or HT. These results suggest that

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 3 published 12 May 2017	✓ report	
REVISED version 2 published 27 Jan 2017	? report	✓ report
version 1 published 31 Aug 2016	? report	? report

1 **Elana Judith Fertig** , Johns Hopkins University USA

2 **Chun-Wei Tung** , Kaohsiung Medical University Taiwan

Discuss this article

Comments (0)

tumor gene expression signatures refined by machine learning techniques can be useful for predicting survival after drug therapies.

This article is included in the Machine learning: life sciences collection.

Corresponding author: Peter K. Rogan (progan@uwo.ca)

Competing interests: PKR cofounded CytoGnomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

How to cite this article: Mucaki EJ, Baranova K, Pham HQ *et al.* **Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning [version 3; referees: 2 approved]** *F1000Research* 2017, **5**:2124 (doi: 10.12688/f1000research.9417.3)

Grant information: AN and LR are funded by NSERC grants RGPIN-2016-05017 and RGPIN-2014-05084 and by the Windsor Essex County Cancer Centre Foundation under a Seeds4Hope grant. PKR has been supported by NSERC [Discovery Grant RGPIN-2015-06290], Canadian Foundation for Innovation, Canada Research Chairs and CytoGnomix Inc.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 31 Aug 2016, **5**:2124 (doi: 10.12688/f1000research.9417.1)

REVISED **Amendments from Version 2**

We have addressed the reviewers' comments regarding overfitting by 1) deriving and validation biochemically inspired machine learning models using the METABRIC Validation patient dataset independently of the Discovery data and 2) assessing the accuracy of the Discovery dataset-based models with patient data derived from an independent source (reference 5). In addition, we have stratified the patients by breast cancer subtype and evaluated each subtype with the combined Discovery+Validation dataset-based models using all of the feature selection methods (Supplementary File 1).

See referee reports

Introduction

Current pharmacogenetic analysis of chemotherapy makes qualitative decisions about drug efficacy in patients (determination of good, intermediate or poor metabolizer phenotypes) based on variants present in genes involved in the transport, biotransformation, or disposition of a drug. We have applied a supervised machine learning (ML) approach to derive accurate gene signatures, based on the biochemically-guided response to chemotherapies with breast cancer cell lines[1], which show variable responses to growth inhibition by paclitaxel and gemcitabine therapies[2,3]. We analyzed stable[4] and linked unstable genes in pathways that determine their disposition. This involved investigating the correspondence between 50% growth inhibitory concentrations (GI_{50}) of paclitaxel and gemcitabine and gene copy number, mutation, and expression first in breast cancer cell lines and then in patients[1]. Genes encoding direct targets of these drugs, metabolizing enzymes, transporters, and those previously associated with chemo-resistance to paclitaxel (n=31 genes) were then pruned by multiple factor analysis (MFA), which indicated that expression levels of genes *ABCC10*, *BCL2*, *BCL2L1*, *BIRC5*, *BMF*, *FGF2*, *FN1*, *MAP4*, *MAPT*, *NKFB2*, *SLCO1B3*, *TLR6*, *TMEM243*, *TWIST1*, and *CSAG2* could predict sensitivity in breast cancer cell lines with 84% accuracy. The cell line-based paclitaxel-gene signature predicted sensitivity in 84% of patients with no or minimal residual disease (n=56; data from 5). The present study derives related gene signatures with ML approaches that predict outcome of hormone- and chemotherapies in the large METABRIC breast cancer cohort[6].

Methods

SVM (Support Vector Machine) learning: Previously, paclitaxel-related response genes were identified from peer-reviewed literature, and their expression and copy number in breast cancer cell lines were analyzed by multiple factor analysis of GI_{50} values of these lines[2] (Figure 1). Given the expression levels of each gene, a SVM is evaluated on patients by classifying those with shorter survival time as resistant and longer survival as sensitive to hormone and/or chemotherapy using paclitaxel, tamoxifen, methotrexate, 5-fluorouracil, epirubicin, and doxorubicin. The SVM was trained using the function *fitcsvm* in MATLAB R2014a[7] and tested with either leave-one-out or 9 fold cross-validation (indicated in Table 1). The Gaussian kernel was used for this study, unlike Dorman *et al.*[1] which used the linear kernel. The SVM requires selection of two different parameters, C (misclassification cost) and sigma (which controls the flexibility and smoothness of Gaussians)[8]; these parameters determine how strictly the SVM learns the training set, and hence if not selected properly, can lead to overfitting. A grid search evaluates a wide range of combinations of these values by parallelization. A Gaussian kernel selects the C and sigma combination that lead to the lowest cross-validation misclassification rate. A backwards feature selection (greedy) algorithm was designed and implemented in MATLAB in which one gene of the set is left out in a reduced gene set and the classification is then assessed; genes that maintain or lower the misclassification rate are kept in the signature. The procedure is repeated until the subset with the lowest misclassification rate is selected as the optimal subset of genes. These SVMs were then assessed for their ability to predict patient outcomes based on available metadata (see Figure 1 and reference 1). Interactive prediction using normalized expression values as input is available at http://chemotherapy.cytognomix.com.

RF (Random Forest) learning: RF was trained using the WEKA 3.7[9] data mining tool. This classifier uses multiple random trees for classification, which are combined via a voting scheme to make a decision on the given input gene set. A grid search was used to optimize the maximum number of randomly selected genes for each tree in RF, where k (maximum number of selected genes for each tree) was set from 1 to 19. Figure 2 depicts the therapy outcome prediction process of a given patient using a RF consisting of a series of decision trees derived from different subsets of paclitaxel-related genes.

Augmented Gene Selection: The most relevant genes (features) for therapy outcome prediction were found using the Minimum Redundancy and Maximum Relevance (mRMR) approach[10]. mRMR is a wrapper approach that incrementally selects genes by maximizing the average mutual information between gene expression features and classes, while minimizing their redundancies:

$$mRMR = \max_{s}\left[\frac{1}{|s|}\sum_{f_i \in S} I(f_i, C) - \frac{1}{|s|^2}\sum_{f_i, f_j \in S} I(f_i, f_j)\right]$$

where f_i corresponds to a feature in gene set S, $I(f_i, C)$ is the mutual information between f_i and class C, and $I(f_i, f_j)$ is the mutual information between features f_i and f_j.

For this experiment, we used a 26-gene signature (genes *ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCL2, BCL2L1, BMF, CYP2C8, CYP3A4, MAP2, MAP4, MAPT, NR1I2, SLCO1B3, TUBB1, TUBB4A, TUBB4B, FGF2, FN1, GBP1, NFKB2, OPRK1, TLR6,* and *TWIST1*) as the base feature set. These genes were selected (in Dorman *et al.*[1]) based either on their known involvement in paclitaxel metabolism, or evidence that their expression levels and/or copy numbers correlate with paclitaxel GI_{50} values. mRMR and SVM were combined to obtain a subset of genes that can accurately predict patient survival outcomes; here, we considered 3, 4 and 5 years as survival thresholds for breast cancer patients.

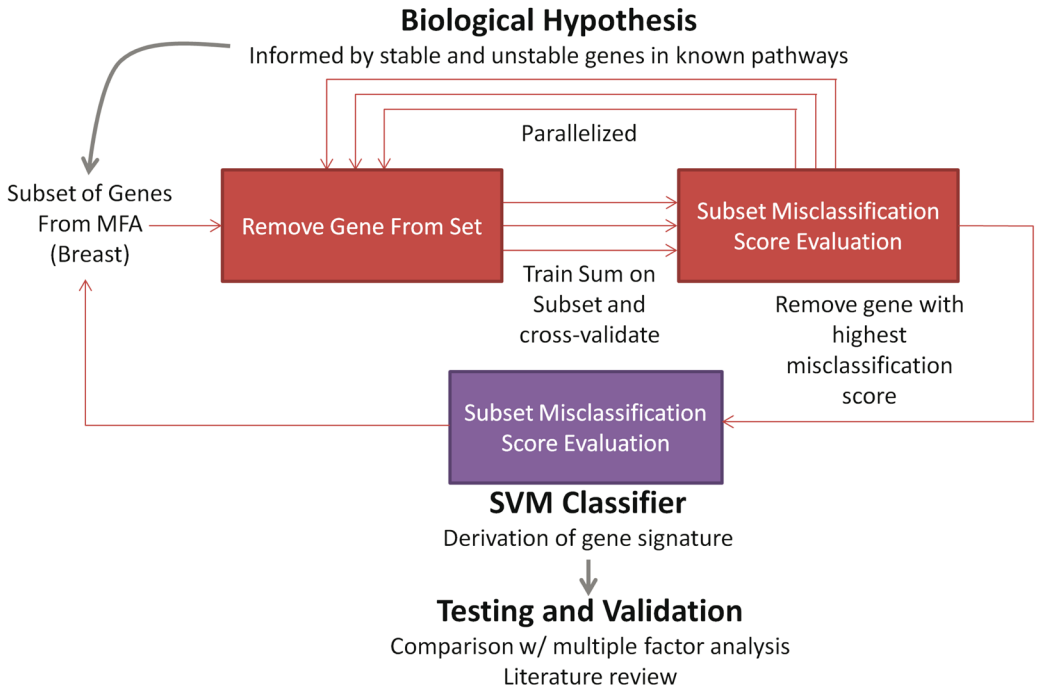

Figure 1. Biochemically-inspired SVM gene signature derivation workflow. The initial set of genes is carefully selected through the understanding of the drug and the pathways associated with it. A multiple factor analysis of the GI_{50} values of a training set of breast cancer cell lines and the corresponding expression levels of each gene in the initial set reduces the list of genes.

Performance was evaluated with several metrics. WEKA determined accuracy (ACC), the weighted average of precision and F-measure, the Matthews Correlation Coefficient (MCC) and the area under ROC curve (AUC).

Results and discussion

Dataset 1. Predicted treatment response for each individual METABRIC patient[11]

http://dx.doi.org/10.5256/f1000research.9417.d149864

The predicted and expected response to treatment for each individual METABRIC patient for each analyses listed in Table 1, Table 2 and Table 3 are indexed. Patients sensitive to treatment are labeled with '0' while resistant patients are labeled '1'.

The performances of several ML techniques have been compared such that they distinguish paclitaxel sensitivity and resistance in METABRIC patients using its tumour gene expression datasets. We used mRMR to generate gene signatures and determine which genes are important for treatment response in METABRIC patients. The paclitaxel models are more accurate for prediction of outcomes in patients receiving HT and/or CT compared to other patient groups.

SVMs and RF were trained using expression of genes associated with paclitaxel response, mechanism of action and stable genes in the biological pathways of these targets (Figure 3). Pair-wise comparisons of these genes with those from MammaPrint and Oncotype Dx (other genomic classifiers for breast cancer) find that these signatures are nearly independent of each other, with only a single gene overlap. The distinct differences of these signatures are due to their methodology of derivation, based on different principles and for different purposes (i.e. drug response for a specific reagent). SVM models for drugs used to treat these patients were derived by backwards feature selection on patient subsets stratified by treatment or outcome (Table 1). The highest SVM accuracy was found for the paclitaxel signature in patients treated with HT and/or adjuvant chemotherapy (78.6%). Since some CT patients were also treated with tamoxifen, methotraxate, epirubicin, doxorubicin and 5-fluorouracil, we also evaluated the performance of models developed for these drugs using the same algorithm. These gene signatures also had acceptable performance (accuracies between 71–76%; AUCs between 0.686 – 0.766). Leave-one-out validation (CT and HT, no treatment, and deceased patients) exhibited higher model performance than 9-fold crossvalidation (CT and/or HT, including patients treated with radiation).

The RF classifier was used to predict paclitaxel therapy outcome for patients that underwent CT and/or HT (Table 2). The best performance achieved with RF showed an 85.5% overall accuracy using a 3-year survival threshold for distinguishing therapeutic resistance vs. sensitivity for those patients that underwent HT.

The best overall accuracy and AUC (sensitivity and specificity) for CT/HT patients using mRMR feature selection for SVM predicting outcome of paclitaxel therapy was obtained for CT patients with 4-year survival (Table 3). Outcomes for HT patients with

Table 1. SVM gene expression signature performance on METABRIC patients.

Patient treatment	# of patients	Agent: *final gene signature (C and sigma)*	Accuracy (%)	Precision	F-Measure	MCC[1]	AUC[2]
Both CT and HT[3]	84	Paclitaxel: *ABCC1, ABCC10, BAD, BIRC5, FN1, GBP1, MAPT, SLCO1B3, TMEM243, TUBB3, TUBB4B (C=10000, σ=10)*	78.6	0.787	0.782	0.559	0.814
		Tamoxifen: *ABCC2, ALB, CCNA2, E2F7, FLAD1, FMO1, NCOA2, NR1I2, PIAS4, SULT1E1 (C=100000, σ=100)*	76.2	0.761	0.760	0.510	0.701
		Methotrexate: *ABCC2, ABCG2, CDK2, DHFRL1 (C=10, σ=1)*	71.4	0.712	0.711	0.410	0.766
		Epirubicin*: ABCB1, CDA, CYP1B1, ERBB3, ERCC1, MTHFR, PON1, SEMA4D, TFDP2 (C=1000, σ=10)*	72.6	0.725	0.723	0.434	0.686
		Doxorubicin: *ABCC2, ABCD3, CBR1, FTH1, GPX1, NCF4, RAC2, TXNRD1 (C=100000, σ=100)*	75.0	0.749	0.750	0.488	0.701
		5-Fluorouracil: *ABCB1, ABCC3, MTHFR, TP53 (C=10000, σ=100)*	71.4	0.714	0.714	0.417	0.718
CT and/or HT[3,4,5,6]	735	Paclitaxel: *BAD, BCAP29, BCL2, BMF, CNGA3, CYP2C8, CYP3A4, FGF2, FN1, NFKB2, NR1I2, OPRK1, SLCO1B3, TLR6, TUBB1, TUBB3, TUBB4A, TUBB4B, TWIST1 (C=10000, σ=100)*	66.1	0.652	0.643	0.287	0.660
Deceased only[2,6,7] (CT and/or HT)	327	Paclitaxel: *ABCB11, BAD, BBC3, BCL2, BCL2L1, BIRC5, CYP2C8, FGF2, FN1, GBP1, MAPT, NFKB2, OPRK1, SLCO1B3, TMEM243 (C=100, σ=10)*	75.3	0.752	0.752	0.505	0.763
No treatment[3]	304	Paclitaxel: *ABCB1, ABCB11, BBC3, BCL2L1, BMF, CYP3A4, FGF2, GBP1, MAP4, MAPT, NR1I2, OPRK1, SLCO1B3, TUBB4A, TUBB4B, TWIST2 (C=100, σ=10)*	73.4	0.734	0.733	0.467	0.769

Initial gene sets preceding feature selection: Paclitaxel - *ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCAP29, BCL2, BCL2L1, BIRC5, BMF, CNGA3, CYP2C8, CYP3A4, FGF2, FN1, GBP1, MAP2, MAP4, MAPT, NFKB2, NR1I2, OPRK1, SLCO1B3, TLR6, TUBB1, TWIST1.* Tamoxifen - *ABCB1, ABCC2, ALB, C10ORF11, CCNA2, CYP3A4, E2F7, F5, FLAD1, FMO1, IGF1, IGFBP3, IRS2, NCOA2, NR1H4, NR1I2, PIAS4, PPARA, PROC, RXRA, SMARCD3, SULT1B1, SULT1E1, SULT2A1.* Methotrexate - *ABCB1, ABCC2, ABCG2, CDK18, CDK2, CDK6, CDK8, CENPA, DHFRL1.* Epirubicin - *ABCB1, CDA, CYP1B1, ERBB3, ERCC1, GSTP1, MTHFR, NOS3, ODC1, PON1, RAD50, SEMA4D, TFDP2.* Doxorubicin - *ABCB1, ABCC2, ABCD3, AKR1B1, AKR1C1, CBR1, CYBA, FTH1, FTL, GPX1, MT2A, NCF4, RAC2, SLC22A16, TXNRD1.* 5-Fluorouracil - *ABCB1, ABCC3, CFLAR, IL6, MTHFR, TP53, UCK2.* [1]MCC: Matthews Correlation Coefficient. [2]AUC: Area under receiver operating curve. [3] Surviving patients; [4] Analysis included patients in the METABRIC 'discovery' dataset only; [5] SVMs tested with 9 fold cross-validation, all others tested with leave-one-out cross-validation; [6] Includes all patients treated with HT,CT, combination CT/HT, either with or without combination radiotherapy; [7] Median time after treatment until death (> 4.4 years) was used to distinguish favorable outcome, ie. sensitivity to therapy.

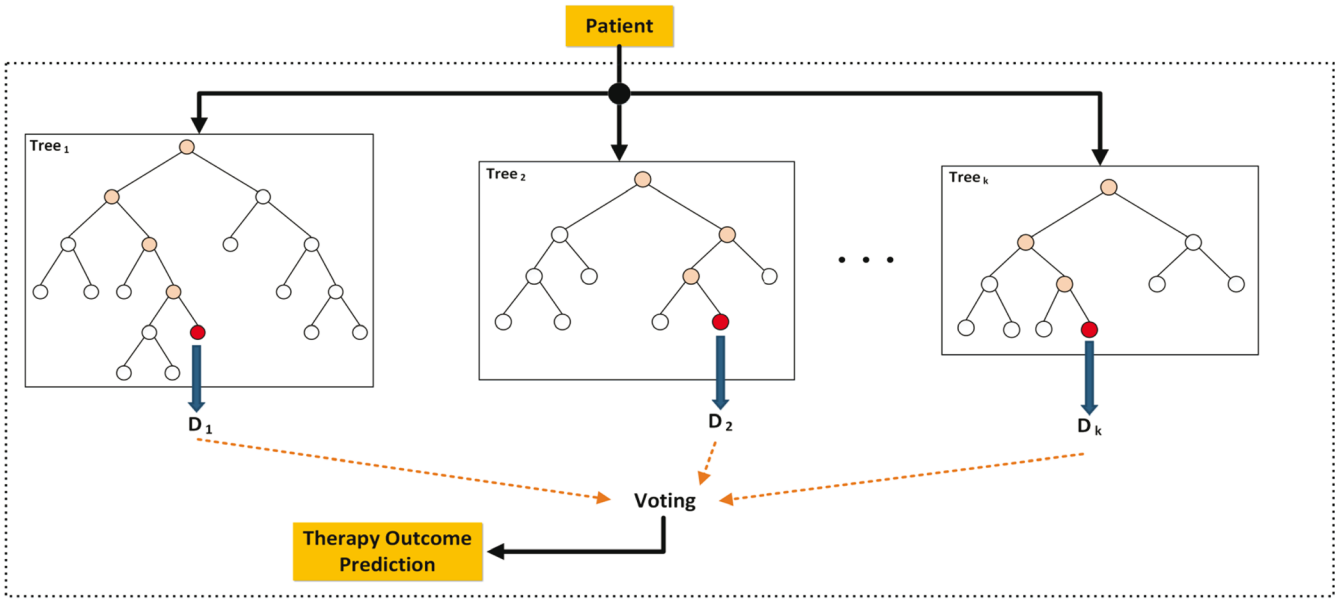

Figure 2. RF decision tree diagram depicts the therapy outcome prediction process of a given patient, using a RF consisting of *k* decision trees. Several DTs are built using different subsets of paclitaxel-related genes. The process starts from the root of each tree and if the expression of the gene corresponding to that node is greater than a specific value, the process continues through the right branch, otherwise it continues through the left branch until it reaches a leaf node; that leaf represents the prediction of the tree for that specific input. The decisions of all trees are considered and the one with the largest number of votes is selected as the patient outcome.

Table 2. Results of applying RF to predict outcome of paclitaxel therapy.

Type of treatment	Survival years (as threshold)	# Patients	*K* (number of genes to be used in random selection)	Accuracy (True Positive - TP) (%)	Precision	F-Measure	MCC[1]	AUC[2]
Chemotherapy (CT)	3	53	7	56.6	0.510	0.524	-0.095	0.441
	4		7	69.8	0.698	0.698	0.396	0.700
	5		19	66.0	0.645	0.636	0.230	0.653
Hormone therapy (HT)	3	420	19	85.5	0.731	0.788	0.000	0.606
	4		9	78.6	0.715	0.706	0.069	0.559
	5		9	71.0	0.634	0.627	0.059	0.632
CT and/or HT	3	504	9	82.7	0.685	0.749	0.000	0.506
	4		19	73.6	0.647	0.648	0.039	0.527
	5		7	65.3	0.602	0.593	0.086	0.588

[1]MCC: Matthews Correlation Coefficient. [2]AUC: Area under receiver operating curve; both Discovery and Validation patient datasets analyzed. RF predictions done using a gene panel consisting of 19 genes (*ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCL2, BCL2L1, BMF, CYP2C8, CYP3A4, MAP2, MAP4, MAPT, NR1I2, SLCO1B3, TUBB1, TUBB4A, TUBB4B*).

Table 3. Results of mRMR feature selection for an SVM for predicting outcome of paclitaxel therapy.

Data	CT[1]			HT			CT+HT		
Survival years (as threshold)	3	4	5	3	4	5	3	4	5
# patients[2]	53			420			504		
Accuracy (TP) (%)	81.1	81.1	84.9	85.7	79.5	72.9	83.1	74.8	67.9
Precision	0.809	0.813	0.852	0.878	0.765	0.692	0.795	0.703	0.662
F-Measure	0.809	0.811	0.845	0.794	0.726	0.663	0.772	0.672	0.666
MCC	0.582	0.625	0.675	0.119	0.17	0.173	0.161	0.137	0.238
AUC	0.783	0.812	0.82	0.508	0.533	0.548	0.53	0.531	0.61
SVM Par. (gamma)	0.0	0.5	1.0	1.0	0.75	1.5	0.75	0.5	1.0
SVM Par. (cost)	64	128	8	2	64	2	16	2	2
Selected genes	*MAP4, GBP1, FN1, MAPT, BBC3, FGF2, NFKB2, TUBB4B*	*TWIST1, FN1, BBC3, FGF2, BCL2L1*	*ABCB11, BCL2, GBP1, SLCO1B3, ABCB1, BAD, TUBB4A, MAPT, NFKB2, TUBB4B*	*ABCB11, BCL2, MAP4, TUBB1, GBP1, SLCO1B3, ABCB1, BAD, TWIST1, FN1, TUBB4A, MAPT, OPRK1, BBC3, FGF2, NFKB2, ABCC1, NR1I2*	*BAD, GBP1, MAPT, BBC3*	*ABCB11, MAP4, SLCO1B3, BAD, FN1, OPRK1, BBC3, NFKB2, NR1I2, TUBB4B*	*ABCB11, SLCO1B3, BAD, TUBB4A, MAPT, BBC3, FGF2, NFKB2, ABCC1, NR1I2*	*ABCB11, BMF, BCL2, MAP4, TUBB1, GBP1, SLCO1B3, ABCB1, BAD, TWIST1, FN1, MAPT, OPRK1, BBC3, FGF2, NFKB2, ABCC1, NR1I2, TUBB4B*	*MAP4, GBP1, SLCO1B3, BAD, MAPT, OPRK1, BBC3, NFKB2, ABCC1, NR1I2, TUBB4B*

[1]For patients treated with CT with ≥4 Yr survival and CT+ HT for ≥ 5 Yr, the cost for the mRMR model was set to 64. Of those treated with CT for ≥ 4 Yr, genes were selected using a greedy, stepwise forward search, while in other cases, greedy stepwise backward search was used. Also, gamma = 0 in all cases. [2]Predicted responses for individual METABRIC patients are provided in Dataset 1.

3-year survival were predicted with 85.7% accuracy; however, the specificity was lower in this group. SVM combined with mRMR further improved accuracy of feature selection and prediction of response to hormone and/or chemotherapy based on survival time than either SVM or RF alone. Predicted treatment responses for individual METABRIC patients using the described ML techniques are indicated in Dataset 1.

Tumor co-variate information was provided by METABRIC, which included Estrogen receptors (ER), Progesterone Receptor (PR), HER2, Lymph Node (LN) and PAM50 subtypes. To assess model co-variate accuracy, predictions described in Table 1–Table 3 were broken down by subtype (available in Supplementary file 1). Subtypes with <20 individuals for a particular treatment combination were not analyzed. The deviation in classification accuracy between subtypes was mostly consistent with the average. One exception involved the RF and mRMR analyses, which was 8.3 to 23.0% below the average for (ER)-negative, (HER2)-positive and basal subtypes in patients treated with HT. However, this deviation was not observed for CT-treated patients with the (ER)-negative subtype, which was consistent with the fact that CT response was derived from the paclitaxel gene set. (ER)-negative patients primarily received CT[6]. Further, the accuracy of the SVM models tested with CT and HT-treated patients was significantly higher for (HER2)-positive patients (26 correct, 3 misclassified; 90% accurate) compared to (HER2)-negative patients (40 correct, 15 misclassified; 73% accurate). *MAPT* expression (present in reduced 'CT and HT' paclitaxel

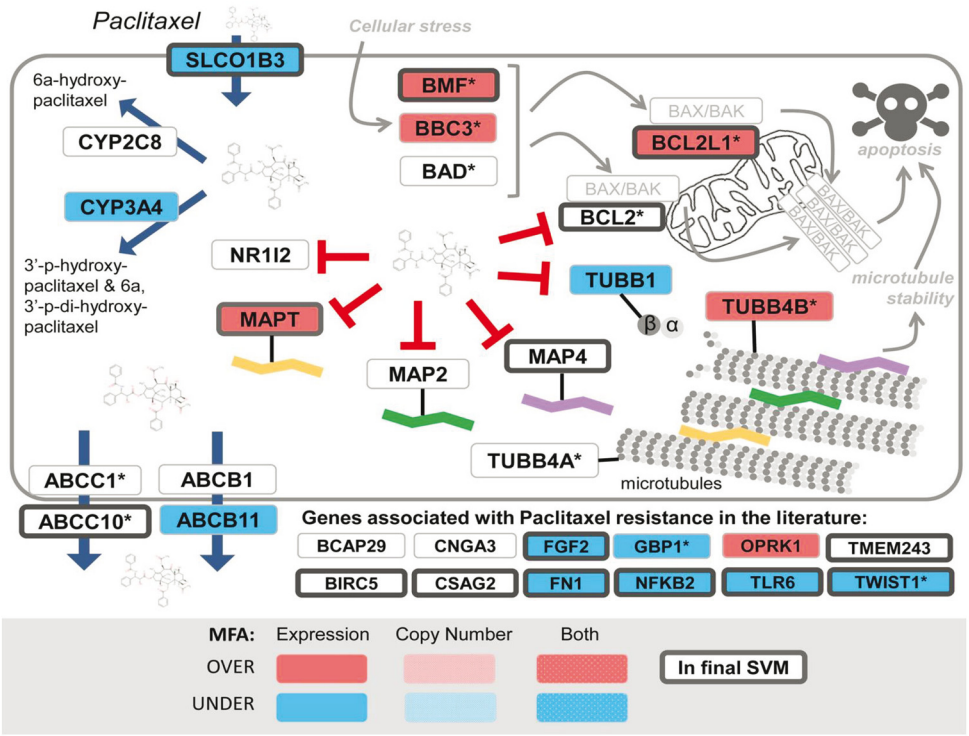

Figure 3. Schematic elements of gene expression changes associated with response to paclitaxel. Red boxes indicate genes with a positive correlation between gene expression or copy number, and resistance using multiple factor analysis. Blue demonstrates a negative correlation. Genes outlined in dark grey are those in a previously published paclitaxel SVM model (reproduced from reference 1 with permission).

Table 4. Results of applying RF to predict outcome of the paclitaxel signature for the METABRIC *Discovery* patient set.

Type of treatment	Survival years (as threshold)	# Patients	K (number of genes to be used in random selection)	Accuracy (True Positive - TP) (%)	Precision	F-Measure	MCC	AUC
Chemotherapy (CT)	3	22	7	61.1	0.617	0.612	0.224	0.444
	4		7	66.7	0.643	0.646	0.189	0.715
	5		19	66.7	0.722	0.687	0.189	0.571
Hormone therapy (HT)	3	185	19	77.0	0.780	0.775	0.018	0.524
	4		9	79.1	0.733	0.710	0.084	0.527
	5		9	68.9	0.533	0.601	-0.133	0.594
CT and/or HT	3	221	9	80.2	0.677	0.734	-0.07	0.389
	4		19	54.8	0.554	0.551	-0.143	0.395
	5		7	60.5	0.567	0.579	0.016	0.479

Paclitaxel gene panel consisted of 19 genes (*ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCL2, BCL2L1, BMF, CYP2C8, CYP3A4, MAP2, MAP4, MAPT, NR1I2, SLCO1B3, TUBB1, TUBB4A, TUBB4B*).

Table 5. Results of mRMR feature selection for an SVM for predicting outcome of the paclitaxel signature for the METABRIC *Discovery* patient set.

Treatment	CT[1]			HT			CT+HT		
Survival years (as threshold)	3	4	5	3	4	5	3	4	5
# patients	22			185			221		
Accuracy (TP) (%)	57.14	57.14	85.7	81.8	70.9	63.6	71.2	69.7	71.2
Precision	0.595	0.686	0.735	0.726	0.670	0.532	0.647	0.629	0.693
F-Measure	0.571	0.623	0.791	0.769	0.686	0.562	0.668	0.628	0.666
MCC	0.167	-0.258	0.000	-0.080	0.032	-0.075	0.035	0.071	0.245
AUC	0.583	0.333	0.500	0.479	0.514	0.477	0.513	0.521	0.586
SVM Par. (gamma)	0.0	0.5	1.0	1.0	0.75	1.5	0.75	0.5	1.0
SVM Par. (cost)	64	128	8	2	64	2	16	2	2
Selected genes	*TWIST1 BMF CYP2C8 CYP3A4 BCL2L1 BBC3 BAD MAP2 MAPT NFKB2 FN1*	*BCL2 BMF CYP2C8 CYP3A4 BAD ABCC10 NFKB2*	*MAP2 BCL2 BCL2L1 BBC3 MAPT GBP1 NFKB2*	*TWIST1 BCL2 BMF CYP2C8 CYP3A4 BCL2L1 BBC3 TLR6 BAD ABCB11 ABCC1 ABCC10 MAP4 MAPT NR1I2 GBP1 NFKB2 OPRK1 FN1*	*TWIST1 CYP2C8 CYP3A4 BCL2L1 BBC3 TLR6 ABCB11 ABCC1 ABCC10 MAP2 MAPT NR1I2 GBP1 NFKB2 FN1*	*TWIST1 BMF CYP2C8 CYP3A4 BCL2L1 BBC3 ABCB11 ABCC1 ABCC10 MAP2 MAP4 MAPT NR1I2 GBP1 NFKB2 OPRK1*	*BMF CYP2C8 BCL2L1 BBC3 BAD ABCC1 ABCC10 MAP4 NR1I2 GBP1 NFKB2 OPRK1 FN1*	*TWIST1 BMF CYP2C8 CYP3A4 BCL2L1 BBC3 TLR6 ABCB11 ABCC1 ABCC10 MAP2 MAP4 MAPT NR1I2 GBP1 NFKB2 OPRK1 FN1*	*TWIST1 BMF CYP3A4 BCL2L1 BBC3 TLR6 BAD ABCB11 ABCC1 MAP2 MAP4 MAPT NR1I2 GBP1 NFKB2 OPRK1 FN1*

[1]For patients treated with CT with ≥4 Yr survival and CT+ HT for ≥ 5 Yr, the cost for the mRMR model was set to 64. Of those treated with CT for ≥ 4 Yr, genes were selected using a greedy, stepwise forward search, while in other cases, greedy stepwise backward search was used. Also, gamma = 0 in all cases.

Table 6. Comparison between our mRMR+SVM method and K-TSP method on *Discovery* patient set of the METABRIC data.

Data	CT			HT			CT+HT		
Survival years	3	4	5	3	4	5	3	4	5
# patients	22			185			221		
mRMR+SVM Accuracy (%)	57.14	57.14	85.7	81.8	70.9	63.6	71.21	69.70	71.21
K-TSP[12] Accuracy (%)	57.14	28.57	28.57	80.91	68.18	69.19	71.21	54.55	53.03

The performances of several ML techniques have been compared such that they distinguish paclitaxel sensitivity and resistance in METABRIC patients using its tumour gene expression datasets. We used mRMR to generate gene signatures and determine which genes are important for treatment response in METABRIC patients. The paclitaxel models are more accurate for prediction of outcomes in patients receiving HT and/or CT compared to other patient groups.

model; Table 1) has been shown to segregate well with PAM50 luminal and basal subtypes[1]. When analyzing METABRIC patients, however, the accuracy of these two subtypes are nearly identical to the average (78.6%, where basal and luminal classification accuracy is 76.7% [n=30] and 76.2% [n=21], respectively).

We assessed the separate *Discovery* and *Validation* datasets, respectively, as training and test sets and repeated the previous experiments. In this scenario, the performance of the model was poor (slightly better than random). This occurred because the gene expression distributions of many of the paclitaxel-related genes in our signature were not reproducible between these two sets (based on Wilcoxon rank sum test, Kruskal-Wallis test and t-tests; Supplementary file 2). Cross-study validation allows for the comparison of classification accuracy between the generated gene signatures. The observed heterogeneity in gene expression highlights one of the many challenges of cross-validation of gene signatures between these data from the same study exhibit drastic differences (for example, *BCL2L1*; Supplementary file 2). Furthermore, these gene expression differences also affect the performance of these methods when these datasets were combined (compare Table 2 and Table 4 for RF; Table 3 and Table 5 for mRMR). We considered the possibility that the Discovery model might be subject to overfitting. We therefore performed cross-study validation of the Discovery set-signature with an independently-derived dataset (319 invasive breast cancer patients treated with paclitaxel and anthracycline chemotherapy[5]). The mRMR+SVM CT-models performed well (4-year threshold model had an overall accuracy of 68.7%; 3-year threshold model exhibited lower overall accuracy [52%], but was significantly better at predicting patients in remission [74.2%]).

To evaluate the paclitaxel models without relying on the *Validation* dataset, the *Discovery* set was split into two distinct parts, consisting of 70% of the patient samples randomly selected for training, and a different set of 30% of samples for testing. This procedure was repeated 100 times using different combinations of training and test samples, and the median performance of these runs is reported (Table 4 and Table 5). We also compared the performance of our mRMR+SVM model with the *K-TSP* model[12] (Table 6). In most cases, our method outperformed K-TSP, based on its accuracy in classifying new patients. Starting with the same set of *Discovery* genes, we also trained a separate model using the *Validation* data, and tested this data by 70/30% cross-validation (accuracy for RF: 56–67% [CT], 67–83% [HT], 56–81% [CT-HT]; accuracy for mRMR: 33–56% [CT], 70–84% [HT], 64–82% [CT-HT]). In addition, we evaluated the performance of the model derived from the Discovery set on a different set of patients treated with paclitaxel[5]. These results suggest that the aforementioned issue with *Discovery* training and *Validation* testing was primarily due to a batch effect, rather than to overfitting.

While not a replication study *sensu stricto*, the initial paclitaxel gene set used for feature selection was the same as in our previous study[1]. Predictions for the METABRIC patient cohort, which was independent of the previous validation set[5] used in Dorman *et al.*[1], of the either same (SVM) or different ML methods (RF and SVM with mRMR) exhibited comparable or better accuracies than our previous gene signature[1].

These techniques are powerful tools which can be used to identify genes that may be involved in drug resistance, as well as predict patient survival after treatment. Future efforts to expand these models to other drugs may assist in suggesting preferred treatments in specific patients, with the potential impact of improving efficacy and reducing duration of therapy.

Conclusion

In this study we used METABRIC dataset to predict outcome for different survival times in patients receiving hormone (HT) and, in some cases, chemotherapy (CT) agents. We used published literature and various machine learning methods in order to identify optimal subsets of genes from a biologically-relevant initial gene set that can accurately predict therapeutic response of patients who have received chemotherapy, hormone therapy or a combination of both treatments. The SVM methodology has been previously shown to outperform randomized gene sets[1]. The predictions made by our method are based on the level of an individual drug. Genomic information has been shown to correlate with tumor therapy response in previous studies[5,13–17]. From these studies, analytical methods have been used to develop gene signatures for chemotherapy resistance prediction[5], subtypes (PAM50), and metastatic risk stratification (Oncotype DX™, MammaPrint®). We also examined the method exhibiting the best performance in the Sage Bionetworks / DREAM Breast Cancer Prognosis Challenge[18], which was also phenotype-based, however it produces outcome signatures based on molecular processes rather than the cancer drugs themselves. While interesting and informative, the results cannot be directly compared. Our approach may be useful for selecting specific therapies in patients that would be expected to produce a favorable response.

Data availability

Patient data: The METABRIC datasets are accessible from the European Genome-Phenome Archive (EGA) using the accession number EGAS00000000083 (https://www.ebi.ac.uk/ega/studies/EGAS00000000083). Normalized patient expression data for the *Discovery* (EGAD00010000210) and *Validation* sets (EGAD00010000211) were retrieved with permission from EGA. Corresponding clinical data was obtained from the literature[6]. While not individually curated, HT patients were treated with tamoxifen and/or aromatase inhibitors, while CT patients were most commonly treated with cyclophosphamide-methotrexate-fluorouracil (CMF), epirubicin-CMF, or doxorubicin-cyclophosphamide.

F1000Research: Dataset 1. Predicted treatment response for each individual METABRIC patient, 10.5256/f1000research.9417.d149864[11]

Author contributions

PKR, AN and LR designed the methodology and oversaw the project. SVM feature selection with MATLAB was automated by DA. EJM and KB selected the initial gene signatures, and

performed processing of the METABRIC data using SVM methods. EJM and IR performed the preprocessing of the METABRIC dataset using RF; EJM, IR and HQ designed feature selection and classification modules using WEKA. PKR and EJM wrote the revised manuscript.

Competing interests

PKR cofounded CytoGnomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Grant information

AN and LR are funded by NSERC grants RGPIN-2016-05017 and RGPIN-2014-05084 and by the Windsor Essex County Cancer Centre Foundation under a Seeds4Hope grant. PKR has been supported by NSERC [Discovery Grant RGPIN-2015-06290], Canadian Foundation for Innovation, Canada Research Chairs and CytoGnomix Inc.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: Accuracy of SVM, RF and mRMR by Patient Co-variates.
Accuracy of all models described in Table 1–Table 3 were further broken down by provided patient subtype information (ER, HER, PR, PAM50, and LN).

Click here to access the data.

Supplementary File 2: Variation of Gene Expression Distribution between Discovery and Validation Datasets.
Whisker plots showing the distribution of expression in the *Discovery* and *Validation* METABRIC datasets for 26 genes used in the paclitaxel gene signature.

Click here to access the data.

References

1. Dorman SN, Baranova K, Knoll JH, *et al.*: **Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning.** *Mol Oncol.* 2016; **10**(1): 85–100.
PubMed Abstract | Publisher Full Text
2. Daemen A, Griffith OL, Heiser LM, *et al.*: **Modeling precision treatment of breast cancer.** *Genome Biol.* 2013; **14**(10): R110.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Shoemaker RH: **The NCI60 human tumour cell line anticancer drug screen.** *Nat Rev Cancer.* 2006; **6**(10): 813–823.
PubMed Abstract | Publisher Full Text
4. Park NI, Rogan PK, Tarnowski HE, *et al.*: **Structural and genic characterization of stable genomic regions in breast cancer: Relevance to chemotherapy.** *Mol Oncol.* 2012; **6**(3): 347–59.
PubMed Abstract | Publisher Full Text
5. Hatzis C, Pusztai L, Valero V, *et al.*: **A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer.** *JAMA.* 2011; **305**(18): 1873–1881.
PubMed Abstract | Publisher Full Text
6. Curtis C, Shah SP, Chin SF, *et al.*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature.* 2012; **486**(7403): 346–352.
PubMed Abstract | Publisher Full Text | Free Full Text
7. **MATLAB and Statistics Toolbox Release 2014a.** The MathWorks Inc., Natick, Massachusetts, United States.
8. Ben-Hur A, Weston J: **A user's guide to support vector machines.** *Methods Mol Biol.* 2010; **609**: 223–39.
PubMed Abstract | Publisher Full Text
9. Hall M, Frank E, Holmes G, *et al.*: **The WEKA data mining software: an update.** *ACM SIGKDD Explorations Newsletter.* 2009; **11**(1): 10–18.
Publisher Full Text
10. Ding C, Peng H: **Minimum redundancy feature selection from microarray gene expression data.** *J Bioinform Comput Biol.* 2005; **3**(2): 185–205.
PubMed Abstract | Publisher Full Text
11. Rezaeian I, Mucaki EJ, Baranova K, *et al.*: **Dataset 1 in: Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Machine Learning.** *F1000Research.* 2016.
Data Source
12. Marchionni L, Afsari B, Geman D, *et al.*: **A simple and reproducible breast cancer prognostic test.** *BMC Genomics.* 2013; **14**: 336.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Van't Veer LJ, Dai H, van de Vijver MJ, *et al.*: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature.* 2002; **415**(6871): 530–536.
PubMed Abstract | Publisher Full Text
14. Duan Z, Duan Y, Lamendola DE, *et al.*: **Overexpression of MAGE/GAGE genes in paclitaxel/doxorubicin-resistant human cancer cell lines.** *Clin Cancer Res.* 2003; **9**(7): 2778–2785.
PubMed Abstract
15. 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**(6): 607–616.
PubMed Abstract | Publisher Full Text
16. Glinsky GV, Berezovska O, Glinskii AB: **Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer.** *J Clin Invest.* 2005; **115**(6): 1503–1521.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Rajput S, Volk-Draper LD, Ran S: **TLR4 is a novel determinant of the response to paclitaxel in breast cancer.** *Mol Cancer Ther.* 2013; **12**(8): 1676–1687.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Cheng WY, Ou Yang TH, Anastassiou D: **Biomolecular events in cancer revealed by attractor metagenes.** *PLoS Comput Biol.* 2013; **9**(2): e1002920.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 3

Referee Report 31 May 2017

doi:10.5256/f1000research.12412.r22653

✔ **Elana Judith Fertig**
Division of Oncology Biostatistics and Bioinformatics, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

The authors have addressed all previous comments.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Referee Report 21 February 2017

doi:10.5256/f1000research.11525.r19726

✔ **Chun-Wei Tung**
School of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan

The authors have addressed all the concerns raised from the previous review. A minor comment for the batch effects is given in the follows. As batch effects are expected for heterogeneous datasets, the direct application of prediction model built on the discovery dataset to the validation dataset would be incorrect and usually result in poor performance. To show the usefulness of the gene signatures while minimizing the batch effects, the authors might consider to run cross-validation on the validation dataset alone using the gene signatures obtained from discovery dataset.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 21 Apr 2017

Peter Rogan, University of Western Ontario, Canada

Thank you for your suggestion. As recommended, we have repeated the 70/30% cross-validation analysis performed in the manuscript (Tables 4 and 5) with the same genes obtained from the *Discovery* dataset (Tables 4 and 5), but using the *Validation* dataset alone for training and testing. We found that this analysis had a similar performance level as the analysis reported in the main manuscript (Tables 4 and 5). There are exceptions. The mRMR+SVM gene signature developed using "CT-only" patients at a 5-year threshold was much less accurate using the *Validation* data. However, the "CT-only" subset of the *Validation* dataset is small (N=31), and thus variability is not unexpected. Overall, this analysis suggests that the cross-validation issue was indeed mostly due to batch effects.

The following sentence was written in the main text which describes this result:

"Starting with the same set of *Discovery* genes, we also trained a separate model using the *Validation* data, and tested this data by 70/30% cross-validation (accuracy for RF: 56-67% [CT], 67-83% [HT], 56-81% [CT-HT]; accuracy for mRMR: 33-56% [CT], 70-84% [HT], 64-82% [CT-HT])."

Competing Interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Referee Report 03 February 2017

doi:10.5256/f1000research.11525.r19727

Elana Judith Fertig
Division of Oncology Biostatistics and Bioinformatics, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

The authors were very responsive to the previous round of reviews, including more robust cross-validation and cross-study validation and comparison with other classifiers. Particular concerns remain that the author's conclusions that it is inappropriate to perform cross-study validation due to batch effects are incorrect, particularly since this challenging task is essential to assess overfitting and for clinical translation of classifiers. In addition, the conclusion was insufficiently revised to place their classifier in the context of the broader literature in this field.

Methods

1. Abbreviations SVM and RF must be spelled out as Support Vector Machine and Random Forest on first use. This was not addressed in the revised methods section.

Results

1. The authors did perform a robust cross-study validation, as requested in the previous review. We agree this is challenging, due in part to batch effects as reported in this manuscript. However, such cross-study validation is essential to assess the accuracy of classifiers. It is also essential to have translation of genomic signatures into the clinic, where even different assays may be used. To address these concerns the authors must do the following: (a) Remove the sentence "This heterogeneity indicates that it is inappropriate to test our gene expression signatures derived by

one of these datasets using the other dataset." (b) Discuss the importance of cross-study validation, challenges in this application, and potential of overfitting of suggested by these results.

2. The author's response that specific therapies were not provided in METABRIC is incorrect. According to Curtis *et al.*, (2012) "Nearly all oestrogen receptor (ER)-positive and/or lymph node (LN)-negative patients did not receive chemotherapy, whereas ER-negative and LN-positive patients did. Additionally, none of the $HER2^+$ patients received trastuzumab. As such, the treatments were homogeneous with respect to clinically relevant groupings." Therefore, the previous criticism #12 remains. Covariates such as ER/HER2/LN or PAM50 subtypes must be included in a table describing the sample cohorts remains. In addition, accuracy must be computed separately for these co-variates or included in the machine learning model.

Conclusion

1. The discussion is insufficient. It still lacks sufficient context of existing genomics classifiers in the literature. The discrepancy between their algorithm and clinical assays is confusing in revised sentence "Unlike Mammaprint and Oncotype Dx tests, this model focuses on predicting survival prediction based on gene expression in the tumor, presumably before or during drug therapy." As written, it appears to disregard the long history of predicting clinical outcome from gene expression involved in developing these classifiers from gene expression data (e.g., van't Veer *et al.,* 2002) into clinical assays based upon expression of smaller numbers of genes.

2. Based on the previous review, the authors include context with other predictions of the METABRIC data in the response to the reviewers. This must also be included in the Conclusion to assess the relevance of their findings in the literature.

References

1. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, METABRIC Group, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.*Nature*. 2012; **486** (7403): 346-52 PubMed Abstract | Publisher Full Text
2. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: Gene expression profiling predicts clinical outcome of breast cancer.*Nature*. 2002; **415** (6871): 530-6 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Apr 2017

Peter Rogan, University of Western Ontario, Canada

Methods

Comment 1. Abbreviations SVM and RF must be spelled out as Support Vector Machine and

Random Forest on first use. This was not addressed in the revised methods section.

Response: These abbreviations are now spelled out upon their first use in the main text (Methods section).

Results

Comment 2. The authors did perform a robust cross-study validation, as requested in the previous review. We agree this is challenging, due in part to batch effects as reported in this manuscript. However, such cross-study validation is essential to assess the accuracy of classifiers. It is also essential to have translation of genomic signatures into the clinic, where even different assays may be used. To address these concerns the authors must do the following: (a) Remove the sentence "This heterogeneity indicates that it is inappropriate to test our gene expression signatures derived by one of these datasets using the other dataset." (b) Discuss the importance of cross-study validation, challenges in this application, and potential of overfitting of suggested by these results.

Response: In regards to this point:
(a) This sentence has been removed, as requested.
(b) To address concerns regarding potential overfitting of our models, we cross-validate the acquired models to a non-METABRIC data set (from an independent study). In the Sage Bionetworks / DREAM Breast Cancer Prognosis Challenge, cross-study validation was performed using the "OsloVal" data set, which consists of gene expression and copy number data from 184 breast cancer patients (Margolin *et al.*, 2013). However, this dataset is not publically available and requires Ethics Board / IRB Review which we did not believe to be worth the effort. Instead, we performed cross-study validation on the gene expression of 310 breast cancer patients made publically available by Hatzis *et al.* (2011).

Analysis of this dataset was successful for the mRMR + SVM models developed using chemotherapy-treated patient ("CT" models), where the threshold for resistance was set to 3-years and 4-years. The "CT 3-year" model performed well predicting responsive patients (74.2% accuracy), while the "CT 4-year" model performed better predicting non-responsive patients (75.1% accuracy). The "CT 4-year" model outperformed the "CT 5-year" model for both sensitive and resistant patient data sets.

Random Forest and mRMR+SVM models which used hormone-treated patients ("HT" and "CT+HT") were much less accurate compared to the "CT-only" models, and predict patients a large percentage of patients from the Hatzis data as sensitive.

In the main manuscript, we have replaced the removed sentence from (a) and have written the following:
"Cross-study validation allows for the comparison of classification accuracy between the generated gene signatures. The observed heterogeneity in gene expression highlights one of the many challenges of cross-validation of gene signatures between these data from the same study exhibit drastic differences (for example, *BCL2L1*; Supplementary file 2). Furthermore, these gene expression differences also affect the performance of these methods when these datasets were combined (compare Table 2 and Table 4 for RF; Table 3 and Table 5 for mRMR). We considered the possibility that the Discovery model might be subject to overfitting. We therefore performed cross-study validation of the Discovery set-signature with an independently-derived dataset (319 invasive breast cancer patients treated with paclitaxel and anthracycline chemotherapy[5]). The

mRMR+SVM CT-models performed well (4-year threshold model had an overall accuracy of 68.7%; 3-year threshold model exhibited lower overall accuracy [52%], but was significantly better at predicting patients in remission [74.2%])."

References

Margolin AA, *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Sci Transl Med. 2013 Apr 17;5(181):181re1. doi: 10.1126/scitranslmed.3006112.
Hatzis, C., *et al.* 2011. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA. 305, 1873–1881.

Comment 3. The author's response that specific therapies were not provided in METABRIC is incorrect. According to Curtis et al., (2012) "Nearly all oestrogen receptor (ER)-positive and/or lymph node (LN)-negative patients did not receive chemotherapy, whereas ER-negative and LN-positive patients did. Additionally, none of the HER2$^+$ patients received trastuzumab. As such, the treatments were homogeneous with respect to clinically relevant groupings." Therefore, the previous criticism #12 remains. Covariates such as ER/HER2/LN or PAM50 subtypes must be included in a table describing the sample cohorts remains. In addition, accuracy must be computed separately for these co-variates or included in the machine learning model.

Response: Thank you for the clarification regarding patient treatment. As a response, we have added an additional supplementary table which breaks down the accuracy of our models by subtype (ER, HER2, PR, LN and PAM50; Dataset 2). In the main text, we note that accuracy of most models are consistent between subtypes (+/- 10% deviation in accuracy). Subtypes with less than twenty individuals were ignored due to its small sample size. The following deviations in accuracy were noted:

1. Random Forest and mRMR models are shown to be consistently more accurate in predicting ER+, HER2- when treated with hormone therapy (both "HT" and "CT and/or HT" categories), when compared to ER- and HER2+ patients. The PAM50 basal subtype is consistently low in accuracy when testing patients treated with hormone therapy. This is most likely partially influenced by the RF and mRMR models for 'HT' to more often predict patients as sensitive, combined with the fact that ER+ and HER2- patients were more likely to response to therapy. It is important to note that the accuracy of predictions by RF and mRMR with patients treated only with chemotherapy was fairly consistent across all available subtypes (+/- 10% accuracy).
2. SVM paclitaxel models performed significantly better with HER2+ patients (26 correct, 3 misclassified; 90% accurate) in HER2- patients (40 correct, 15 misclassified; 73% accurate) when tested on patients treated with both hormone and chemotherapy. In Dorman et al (2016), it was stated that *MAPT* expression (which is present in the paclitaxel model) segregated with PAM50 luminal and basal subtypes. For this model, the accuracies of these subtypes are nearly identical to the accuracy of the entire subset.

Text describing these results can be found in the third paragraph of the results.

Conclusion

Comment 4. The discussion is insufficient. It still lacks sufficient context of existing genomics classifiers in the literature. The discrepancy between their algorithm and clinical assays is confusing in revised sentence:

"Unlike Mammaprint and Oncotype Dx tests, this model focuses on predicting survival prediction based on gene expression in the tumor, presumably before or during drug therapy."
As written, it appears to disregard the long history of predicting clinical outcome from gene expression involved in developing these classifiers from gene expression data (e.g., van't Veer et al., 2002) into clinical assays based upon expression of smaller numbers of genes.

Response: We have removed the indicated sentence, which we agree was insufficient to the comment from the previous iteration of this article: "Must be discussed in the context of existing genomics classifiers for breast cancer (e.g., OncotypeDx and/or Mammaprint)".

We in no way meant to ignore the long history of predicting clinical outcome from gene expression (as well as other genomic factors). A discussion on this topic was not included in earlier submissions as it initially had an imposed word length limit (upon first submission). We did, however, reference other articles which do discuss this topic. In Dorman *et al.* (2016), which described some of the methodology for initial gene selection that this study was based on, these contributions are well-referenced, including the history of the prediction of clinical outcome from genomic status:
"Previous studies have derived associations between the genomic status of one or more genes and tumor response to certain therapies (Duan *et al.*, 2003; Glinsky *et al.*, 2005; Hatzis *et al.*, 2011; Ma *et al.*, 2004; Rajput *et al.*, 2013; van't Veer *et al.*, 2002).

Correlations between single gene expression and tumor resistance (Duan *et al.*, 2003, 1999) do not take into account multiple mechanisms of resistance or assess interactions between multiple genes. ABC transporter overexpression has long been shown to confer resistance, but enzymatic or functional inhibition has not substantially improve patient response to chemotherapy (Samuels *et al.*, 1997).

Multi-gene analytical approaches have previously been successful in deriving prognostic gene signatures for metastatic risk stratification (Oncotype DX™, MammaPrint®), subtypes (PAM50), and efforts have been made to predict chemotherapy resistance (Hess *et al.*, 2006; Hatzis *et al.*, 2011). "

In response to Dr. Fertig's comments, we have added a short discussion with citations of previously published approaches (including MammaPrint and Oncotype DX):
"Genomic information has been shown to correlate with tumor therapy response in previous studies[5,12-16]. From these studies, analytical methods have been used to develop gene signatures for chemotherapy resistance prediction[5], subtypes (PAM50), and metastatic risk stratification (Oncotype DX™, MammaPrint®)."

Comment 5. Based on the previous review, the authors include context with other predictions of the METABRIC data in the response to the reviewers. This must also be included in the Conclusion to assess the relevance of their findings in the literature.

Response: We have added the indicated text from the previous 'response to the reviewers' (modified) to the Conclusions:

"We also examined the method exhibiting the best performance in the Sage Bionetworks / DREAM Breast Cancer Prognosis Challenge[17], which was also phenotype-based, however it produces outcome signatures based on molecular processes, rather than the cancer drugs themselves.

While interesting and informative, the results cannot be directly compared."

Please note that the majority of entries in the DREAM project were not fully curated and only exist as source code. Analyzing these files to determine what methodology was attempted by these groups is beyond the scope of our study. A description of the second place of the METABRIC phase of the DREAM challenge is provided in the link below. This link describes how the METABRIC data is trained using a bipartite graphing as input for linear models, boost models, and RankSVM. While they state that RankSVM was the least successful between the three methods, it does not appear that this particular study has been published to the literature. As a result, we cannot fully review their results, and thus cannot be compared to our methodology in the main manuscript.

https://sagesynapse.wordpress.com/2012/11/01/breast-cancer-challenge-team-pitttransmed-places-s

Competing Interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Version 1

Referee Report 03 October 2016

doi:10.5256/f1000research.10141.r16345

Chun-Wei Tung
School of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan

This study proposed prediction methods using SVM and RF classifiers with mRMR selected feature sets from cell line data and demonstrate its prediction ability for outcomes from METABRIC patient cohort. The classifiers with good prediction performance show the usefulness of combining domain knowledge with feature selection techniques. However, some details essential for reproducibility and interpretation are missing.
Required information is listed in the following.

1. What are the values of parameters for SVM and RF classifiers and the methods for parameter selection (by default or other selection methods)?

2. The development and evaluation of models for patient data are not clear. Whether the models were trained using partial data from METABRIC or only leave-one-out cross-validation was applied? If cross-validation is the case, then what is the model offered at the online server because there will be more than one models created, and whether the cross-validation is involved in the feature selection process that often leads to an overestimation of the performance. For the case of training on partial data, both training and test performance are essential information for evaluating the robustness of models.

3. Since some of the datasets are highly imbalanced, the numbers of positives and negatives, as well as sensitivity and specificity are more important than accuracy for interpreting the results as a high

accuracy with a low AUC could be the result of all positive/negative predictions on an imbalanced dataset. Listing all the information along with the accuracy and AUC will help the interpretation of prediction performances.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Jan 2017

Peter Rogan, University of Western Ontario, Canada

Comment 1:What are the values of parameters for SVM and RF classifiers and the methods for parameter selection (by default or other selection methods)?

Response: The parameter values for these classifiers have been added to the Tables 1-5.

In regards to parameter selection, the first paragraph of the methods now describes C and Sigma selection as a grid search to find the values with the lowest cross-validation misclassification rate. Similarly for RF, a grid search was used to optimize the maximum number of randomly selected genes for each tree (second paragraph of Methods section).

Comment 2: The development and evaluation of models for patient data are not clear. Whether the models were trained using partial data from METABRIC or only leave-one-out cross-validation was applied? If cross-validation is the case, then what is the model offered at the online server because there will be more than one models created, and whether the cross-validation is involved in the feature selection process that often leads to an overestimation of the performance. For the case of training on partial data, both training and test performance are essential information for evaluating the robustness of models.

Response: We obtained new results for both RF and mRMR+SVM models when we use discovery set as training set and validation set as test set, the performance of the model was poor. After more investigation we found that there happened to be a large variation between gene expression of 26 targeted genes between discovery and validation set (please see Supplementary Dataset 2). Hence, building any classifier using discovery and validation set as training and test set in their current forms will result of poor performance, since the training and test sets are vastly different.

However, we did carry out another experiment on discovery set solely and used 70% of data for training and remaining 30% for test the performance of the model. The results have been added to the manuscript (Tables 4 and 5).

Comment 3: Since some of the datasets are highly imbalanced, the numbers of positives and negatives, as well as sensitivity and specificity are more important than accuracy for interpreting the results as a high accuracy with a low AUC could be the result of all positive/negative predictions on an imbalanced dataset. Listing all the information along with the accuracy and AUC will help the interpretation of prediction performances.

Response: As previously mentioned, we have added more performance measures including MCC

and AUC. They have been added Tables 1-5 of the manuscript.

Competing Interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Referee Report 30 September 2016

doi:10.5256/f1000research.10141.r16733

Elana Judith Fertig
Division of Oncology Biostatistics and Bioinformatics, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

This study develops SVM and RF algorithms built upon previously learned gene signatures of therapeutic response to breast cancer. The algorithms are applied and compared to predict patient survival under different treatment conditions in METABRIC data. The analyses and comparisons are robust and this study provides a useful assessment of biologically-driven classifiers. The three major areas that require improvement before the article is indexed are as follows, and described in further detail below.

1. The methods require further clarification to distinguish differences between this study and the previous study as well as the parameters of the machine learning algorithms.
2. Accuracy in the results must better distinguish results on independent test and training sets.
3. Classifiers must be put in the context of other existing genomics classifiers used in breast cancer and/or previously published in Mammaprint data.

Title and Abstract

Acceptable

Article content

Methods

1. Abbreviations SVM and RF must be spelled out as Support Vector Machine and Random Forrest on first use in Methods.
2. Writing in *SVM learning* subsection of *Methods* requires clarification to distinguish which of these methods were developed in the previous *Molecular Oncology* publication and which were developed as part of this publications.
3. Details about the SVM learning algorithm are included in the caption to Figure 1, but must also be included and completely described in text for the corresponding section of the methods.
4. No equations are provided to describe the role of the parameters C and sigma. It is also unclear whether this greedy search is implemented by the Matlab function *fitcsvm* or uses custom code developed by the authors.

Results

1. Need to specify whether reported accuracies are computed with leave-one-out cross validation or 9-fold cross validation (described in Methods).

2. Ideally, given the size of METABRIC data they would be calculated on independent training (first 1000 patient samples) and training (last 1000 patient samples) datasets.
3. AUC must be computed separately for discovery and validation sets (Table 2).
4. It is unclear whether the previous validation set described in the sentence “Predictions for the METABRIC patient cohort, which was independent of the previous validation set” refers to a validation set used in this publication or the previous publication.
5. Covariates such as ER/PR or PAM50 subtypes must be included in a table describing the sample cohorts. Accuracy must be computed separately for these co-variates or they must also be included as co-variates in the machine-learning model.
6. Ideally accuracy would be compared to existing breast cancer classifiers (e.g., using code from Marchionni *et al., BMC Genomics*, 2013) and/or survival curves reported in the literature.

Conclusions

1. Must be discussed in the context of existing genomics classifiers for breast cancer (e.g., OncotypeDx and/or Mammaprint).
2. Results must be put in context with other predictions on METABRIC data, e.g., outcomes from the DREAM contest.

Data

Acceptable

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Jan 2017

Peter Rogan, University of Western Ontario, Canada

Comment 1: The methods require further clarification to distinguish differences between this study and the previous study as well as the parameters of the machine learning algorithms.

Response: The first paragraph of the Methods describes Support Vector Machine learning, which has been greatly expanded upon. Differences in SVM methodology between the two studies are indicated there (i.e. a Gaussian kernel was used instead of a linear kernel). All other feature selection methods described in the manuscript (Random Forest, mRMR) were not used in Dorman *et al.*, 2016.

The parameters for machine learning algorithms have been incorporated in the manuscript, and can be found in the footnote section of each data table.

Comment 2: Accuracy in the results must better distinguish results on independent test and training sets.

Response: The Validation dataset showed a distinct overall expression profile from the Discovery set, possibly due to batch effects, which are well known. We added another experiment to the manuscript by splitting the Discovery set into Training and Test sets. The model was trained using

70% of the data and then tested using the remaining 30% of data as test set. We repeated this procedure 100 times and took the median as the final performance result. The results are presented in Tables 4 and 5 of the manuscript.

Comment 3: Classifiers must be put in the context of other existing genomics classifiers used in breast cancer and/or previously published in Mammaprint data.

Response: We have added two sentences in the second paragraph of the "Results and Discussion" section which describes the comparison of our gene signature to those from MammaPrint and Oncotype Dx. Pair-wise comparison of these three signatures show that they are nearly independent of one another.

Methods

Comment 4: Abbreviations SVM and RF must be spelled out as Support Vector Machine and Random Forest on first use in Methods.

Response: We thank the reviewer for this suggestion. It has been addressed in the Methods section of the manuscript.

Comment 5: Writing in SVM learning subsection of Methods requires clarification to distinguish which of these methods were developed in the previous Molecular Oncology publication and which were developed as part of this publications.

Response: This is now clarified within the first paragraph of the Methods section in the manuscript. The SVM classifier was adopted from previous Molecular Oncology publication, while the feature selection method has been developed as part of this publication.

Comment 6: Details about the SVM learning algorithm are included in the caption to Figure 1, but must also be included and completely described in text for the corresponding section of the methods.

Response: Thanks for the reviewer's suggestion. This description of the SVM learning algorithm has been moved from the Figure 1 legend and integrated into the first paragraph of the methods section.

Comment 7: No equations are provided to describe the role of the parameters C and sigma. It is also unclear whether this greedy search is implemented by the Matlab function fitcsvm or uses custom code developed by the authors.

Response: A brief description of the role of each parameter has been added to the first paragraph of the methods section of the manuscript. Readers are also now directed to a reference (Ben-Hur and Weston, 2010) if more detail is desired.

The greedy search, also called sequential backward feature selection, was implemented as a script by our lab in MATLAB. It is not a MATLAB function. This is clarified by changing a few words in the first paragraph of the methods section: "A backwards feature selection (greedy) algorithm was designed and implemented in MATLAB in which…"

Moreover, as described above, the SVM classifier was adopted from previous Molecular Oncology publication (Dorman *et al.* 2016), while the feature selection method has been developed as part of this publication.

Results

Comment 8: Need to specify whether reported accuracies are computed with leave-one-out cross validation or 9-fold cross validation (described in Methods).

Response: All SVM models described in the manuscript used leave-one-out cross validation except one, and this is clearly indicated in Table 1, and is now commented on in the methods. A 9-fold cross-validation was used to build a model using 735 patients who were treated with Chemotherapy and/or Hormone therapy, as leave-one-out cross validation of this many patients took an unreasonably long time to complete (it exceeded 3 weeks on a dedicated I7 Intel processor).

Comment 9: Ideally, given the size of METABRIC data they would be calculated on independent training (first 1000 patient samples) and test (last 1000 patient samples) datasets.

Response: We obtained new results for both RF and mRMR+SVM models using Discovery patient set for training and Validation set for testing, however the performance of the model was poor. After further investigation, we found that there were large differences between gene expression levels of the 26 model signature genes in the Discovery versus Validation sets (we used Wilcoxon rank sum test, Kruskal-Wallis test and t-test to evaluate the results – shown in the plotted distributions of gene expression in Supplemental Dataset 2) regardless of patient status (alive or dead). Hence, building any classifier using discovery and validation set as training and test set in their current forms will result of poor performance due to this source of heterogeneity.

To address this issue, we did carry out another experiment based on data from the Discovery patient dataset alone; using 70% of data for training and remaining 30% for testing, the performance of the model was significantly better. We speculate that the discrepancy between the expression distributions in the Discovery and Validation sets were the result of batch effects. The results have been added to the manuscript (Tables 4,5).

Comment 10: AUC must be computed separately for discovery and validation sets (Table 2).

Response: We have included additional performance measures to Tables 1-5, including Area Under Curve (AUC).

Comment 11: It is unclear whether the previous validation set described in the sentence "Predictions for the METABRIC patient cohort, which was independent of the previous validation set" refers to a validation set used in this publication or the previous publication.

Response: This sentence is referring to breast cancer patient data from Hatzis *et al.* (2013), which was used as a validation set in Dorman *et al.* (2016), not this publication. We have modified this sentence to clarify the issue.

Comment 12: Covariates such as ER/PR or PAM50 subtypes must be included in a table describing the sample cohorts. Accuracy must be computed separately for these co-variates or

they must also be included as co-variates in the machine-learning model.

Response: Even with the subtype as covariant, it is not possible to perform the analysis the reviewer requested. Certain therapies are definitely more effective in particular subtypes (eg. etoposide, docetaxel, and cisplatin are preferentially active in basal or claudin-low cell lines, as observed clinically; Heiser *et al.,* 2012). The public METABRIC dataset (or the corresponding publication) does not provide the specific therapies used to treat individual patients. Had they done so, it would have made sense to look at these covariates.
Reference: Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F, *et al.* (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci US A* **109**:2724-2729.

Comment 13: Ideally accuracy would be compared to existing breast cancer classifiers (e.g., using code from Marchionni et al., BMC Genomics, 2013) and/or survival curves reported in the literature.

Response: The proposed method has been compared against the K-TSP (Marchionni *et al.*, BMC Genomics, 2013) as per reviewer's suggestion and the results are presented in Table 6 of the manuscript.

Conclusions
Comment 14: Must be discussed in the context of existing genomics classifiers for breast cancer (e.g., OncotypeDx and/or Mammaprint).

Response: We have added text to both the second paragraph of the "Results and Discussion" paragraph and to the conclusion of the paper.

Comment 15: Results must be put in context with other predictions on METABRIC data, e.g., outcomes from the DREAM contest.

Response: An important distinction to note in regards to our methodology is that the predictions are based on the genes known to be associated with the response to specific drugs used to treat breast cancer. In the DREAM contest, the method with the highest METABRIC score (as described in Cheng *et al.*, 2013) was phenotype-based, finding signatures for molecular processes that are disregulated in METABRIC, rather than responses to the cancer therapies themselves. While this is an interesting prediction method, the results cannot compared to our approach. The gene signatures that we have derived contain components of many different pathways.
Reference: Cheng WY, Ou Yang TH, Anastassiou D. Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput Biol. 2013;9(2):e1002920.

Competing Interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.