Candidate Biomarkers for Response to Tamoxifen in Breast Cancer metastatic patients

Claudia Cava, Gloria Bertoli, Italo Zoppis, Giancarlo Mauri, Maria Carla Gilardi and Isabella Castiglioni

Abstract— **Tamoxifen is currently used for the treatment of breast cancer. Response to tamoxifen in metastatic conditions is a primary issue in cancer development. We used a cohort of breast cancer patients, treated or not with tamoxifen, and combined these data with the gene signature of metastatic samples in order to investigate the genetic mechanism of metastasis development, in search of a possible therapeutic effect of tamoxifen in metastatic conditions,. The analysis revealed a group of 21 genes common both to the set of up regulated genes in metastatic BC patients and to the set of down regulated genes in tamoxifen treated patients. These genes could be used as biomarkers for tamoxifen-sensitivity in order to optimize BC treatment.**

I. Introduction

BREAST cancer (BC) is an heterogeneous disease with regards to its clinical behavior and response to therapies [1].

Multiple factors are responsible for BC and a number of treatment options are also available for the disease [2]. Hormone treatment is one of the key strategies in the management of metastatic breast cancer. Antiestrogen tamoxifen is the most widely used hormonal therapy for its treatment, though resistance to tamoxifen is still a concern in recurrent disease (affecting 25-35% of patients) [3]. Furthermore, therapy resistant metastatic tumor cells are a major cause of death. The underlying biological mechanisms of resistance to tamoxifen are not completely understood [3]. The observed clinical efficacy of tamoxifen has been attributed to both growth interruptionand induction of apoptosis within the breast cancer cells [3]. For instance, this compound is a nonsteroidal agent with potent antiestrogenic properties, which competes with estrogen for estrogen receptor (ER) binding sites in breast and in other tissues. The complex between ER and tamoxifen interacts with DNA, causing the cells to remain in the G_0 and G_1 phases of the cell cycle, thus leading to their death.

It's interesting uncover a gene profile that may accurately predict whether a cancer will be tamoxifen sensitive.

While conventionally used in ER-positive tumors, which comprise approximately 75% of breast cancers, Tamoxifen, has also been used to successfully treat some ER-negative breast tumors in recent years [4].

In previous studies, gene signatures have been developed that can distinguish subgroups of patients with different prognoses or response to different treatment [e.g.3].

Metastasis is also a major issue in studying treatment efficacy for many cancers. However, a characterization of genes associated with tamoxifen efficacy in metastatic conditions is lacking.

In this study, we used a cohort of breast cancer patients, treated or not with tamoxifen, and combined these data with the gene signature of metastatic samples.

We focused on a group of genes as potential tamoxifen-targets in metastatic BC, looking for common gene among metastatic and tamoxifen-treated BC samples.

The challenge is to identify the driving pathway of the metastatic BC tumor that could be target of the drug treatment, thus addressing BC patients to the best treatment.

II. Materials and Methods

A. Metastatic vs Non-Metastatic BC Patients

Three different publicly available microarray datasets from BC patients were combined together to create a final data set of 505 BC microarrays: GSE11121, GSE2990 and the dataset used by Foekens et al. [5] containing 200, 125 and 180 samples, respectively. The datasets came from the same Affymetrix GeneChip Human Genome U133A. All the samples came from BC patients who were not subjected to any adjuvant systemic treatment.

From the final gene expression dataset (505 patients), we selected two groups of patients: 394 patients without distant metastases (class A) and 111 patients with distant metastases (class B).

- *Identification of up/down regulated genes*

Class A and class B gene expression data were processed using Robust Multi-array Average (RMA) method in order to generate probe-level intensities. To correct the batch effects (experimental variations of datasets generated by different laboratories) Combining Batches of Gene Expression Microarray Data (ComBat) was used [6,7].

Significance Analysis of Microarray (SAM) [8] was applied. SAM analysis was used to obtain DNA probes discriminating between the two groups of interest. The genes were considered up/down regulated from SAM if their mean expression in class B were significantly higher/lower (FDR

C. Cava, G. Bertoli, M.C. Gilardi and I. Castiglioni are with the Institute of Molecular Bioimaging and Physiology of the National Research Council (IBFM-CNR), Milan, Italy (corresponding author to provide phone: 0039-02-21717511; fax: 0039-02-21717502, e-mail: isabella.castiglioni@ibfm.cnr.it).

G. Mauri and I. Zoppis are with the Department of Informatics, Systems and Communications, University of Milano–Bicocca, Milano, Italy. (e-mail: mauri@disco.unimib.it, zoppis@disco.unimib.it).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

q-value < 0.01) than in class A.

B. Tamoxifen-treated vs untreated BC Patients

One publicly available microarray datasets from BC patients (GSE6532) was considered. From GSE6532 two groups were selected: 123 BC patients without treatment (class C) and 125 BC patients treated with tamoxifen (class D). The dataset came from the Affymetrix GeneChip Human Genome U133B.

- *Identification of up/down regulated genes*

Class C and class D gene expression data were processed using RMA method in order to generate probe-level intensities. SAM was used to obtain DNA probes discriminating between the two groups of interest. The genes were considered up/down regulated if their mean expression in class D were significantly higher/lower (FDR q-value < 0.01) than in class C.

C. Filtered gene signatures in Tamoxifen-treated patients with metastasis

Different combined analysis were considered: a) up regulated genes in metastatic patients vs down regulated genes in tamoxifen-treated patients; and b) down regulated genes in metastatic patients vs up regulated genes in tamoxifen-treated patients.

D. Assessment

To evaluate the performance of the proposed approach we used machine learning algorithms trained on the identified gene signatures and tested on the ability to differentiate the different classes.

For this purpose we used classes of homogeneous samples (similar number of samples) of BC microarray expression from the FK and GSE11121 datasets.

Specifically, from the FK dataset 36 metastatic patients and 36 no-metastatic patients were selected; from GSE11121, 46 metastatic patients and 46 non-metastatic patients were selected. To check the robustness of the derived signature in a blinfold analysis with Naive Bayes (NB) classification we used two new datasets never utilized in the previous phases. We selected from GSE20685 dataset 35 metastatic patients and 35 no-metastatic patients and from GSE7390 40 metastatic patients and 40 no-metastatic patients.

- *Machine Learning*

We designed a Rapid Miner (RM) workflow [9]. RM is a software environment for rapid prototyping of machine learning processes. It is currently used for classification, clustering, and also data integration tasks e.g., [10].

Two types of classification were applied: i) NB ii) SVM machine.

i) NB has been successfully used in Bioinformatics for classification and prediction. It is a simple probabilistic classifier based on applying Bayes theorem with naive independence assumptions[11-13].

ii) SVM allows modeling different parameters in order to perform the best solution for each datatype. We optimized the accuracy over a space of given SVM feasible learning parameters. The following parameters were used. kernel.γ: from 0 to 5, step 30; kernel.C: from 0 to 5, step 30; kernel.type $\in$ {ANOVA, NEURAL, RADIAL}. The optimization allowed us to change iteratively the SVM parameters to optimize the performance of the learning scheme.

-*Validation*

The classification algorithms were validated by a k-fold cross validation (k=10), a two step process: in the first step a classifier is built describing a predetermined set of data classes. In the second step, the model (a trained SVM) is used for testing new classification examples; the generalization performance of the classifier is estimated using a new test data set. The input data set S is split into subsets $\{S_1,S_2,...,S_k\}$. The first inner operator (SVM) realizes the learning step described above. SVM is applied k times using at each iteration i the set S_i as the test set and $S\text{-}S_i$ as the training set.

The performance of classification was obtained in terms of AUC and Receiver Operating Characteristic (ROC) curve. AUC is a measure of test accuracy of the classifier.

ROC curve is a two dimensional measure of classification performance, widely used in biomedical research to assess the performance of diagnostic tests [14].

The classification methods (NB and SVM) were applied to gene features in patients with metastasis (SAM) and in filtered gene signatures in tamoxifen-treated patients in metastatic conditions.

E. Identification of gene signature

In a second step, the genes, as found up or down regulated in expression, were identified by submitting IDs probes from the HGU133Array to Affymetrix through the Netaffxtool [14]. By using biological pathway-based analysis Reactome we determined whether gene signatures were enriched for a particular pathway.

III. Results

A. Metastatic vs Non-Metastatic BC Patients

SAM provided 253 genes (512 probes). Among these genes, 193 were identified with up and 60 with down regulation in expression, through comparison of BC samples of patients without distant metastases (class A) and with distant metastases (class B).

B. Treated vs Untreated BC Patients

SAM provided 4687 genes (8773 probes). All these genes were identified with down regulation in expression, by comparing BC samples of patients treated with tamoxifen (class D) and BC samples of untreated patients (class C).

C. *Filtered gene signatures in Tamoxifen-treated patients with metastasis*

Since no up in tamoxifen-treated patients were found, only up regulated genes in metastatic patients vs down regulated genes in tamoxifen-treated patients were studied. From this combination we obtained 21 genes (42 probes).

D. *Assessment*

Table I shows AUC performance of the two machine learning algorithms (NB and SVM).

We can observe that AUC values in both NB and SVM have a better performance when filtered gene signatures were used (an AUC close to 1 indicates very reliable diagnostic test).

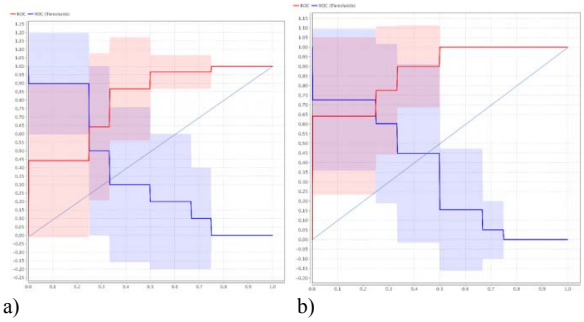
Figure 1 and 2 show ROC Curve (red line) at various threshold settings (blue line). In particular, Figure 1 shows

TABLE I
AUC PERFORMANCE

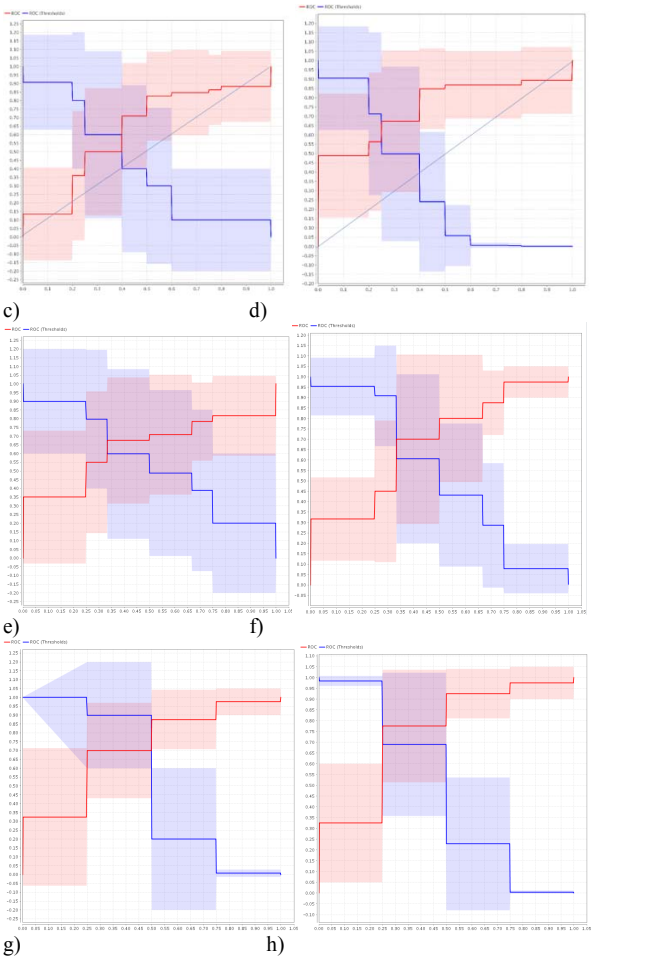
Dataset	Gene signatures (AUC)		Filtered Gene signatures (AUC)	
	NB	SVM	NB	SVM
FK dataset	0.80	0.91	0.87	0.92
GSE11121	0.62	0.73	0.75	0.78
GSE20685	0.63		0.68	
GSE7390	0.72		0.75	

NB ROC curve for each datatype: gene signatures (SAM) and filtered gene signatures. Figure 2 shows SVM ROC curve for each datatype: gene signatures (SAM) and filtered gene signatures.

In both NB and SVM (FK and GSE11121), the ROC curves show an area close to 1 in filtered gene signatures, confirming the good performance (if the ROC curve rises rapidly towards the upper right-hand corner of the graph, or if the value of area under the curve is large, the test can be said to perform well. An area close to 1.0 indicates that the test is reliable, while an area close to 0.5 shows that the test is unreliable.). The blinfold analysis (GSE20685 and GSE7390) confirmed that machine learning performed better for gene filtering.

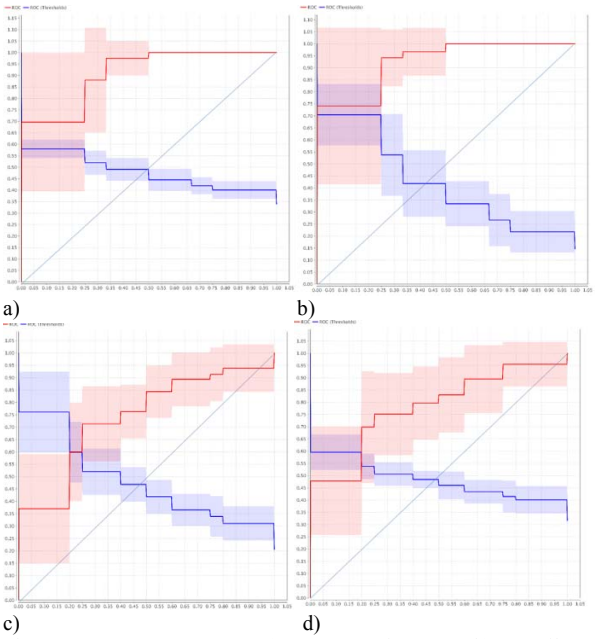


a) b)



c) d) e) f) g) h)

Fig. 1. NB ROC Curve: a) FK Gene signatures b) FK Filtered Gene Signatures c) GSE11121 Gene signatures d) GSE11121 Filtered Gene Signatures e) GSE20685 Gene signatures f) GSE20685 Filtered Gene Signatures g) GSE7390 Gene signatures h) GSE7390Filtered Gene Signatures



a) b) c) d)

Fig. 2. SVM ROC Curve: a) FK Gene signatures b) FK Filtered Gene Signatures c) GSE11121 Gene signatures d) GSE11121 Filtered Gene Signature

E. *Identification of gene signatures*

Table II shows the gene and relative position found by Reactome analysis.

Most parts of the obtained gene signatures are implicated in development and progression of BC, e.g.

1. Mitotic G1 phases (p-value= 7.7e-05)
2. Cell Cycle (p-value= 1.1e-04)
3. Phosphorylation of Gorasp1, Golga2 and RAB1A by CDK1-CCNB (p-value= 8.9e-05).

TABLE II
FILTERED GENE SIGNATURES

Gene	Position	Function*
ASPM	1q31.3	1
CBX5	12q13.13	1
CCNB2	15q22.2	1., 2. and 3.
CDK1	10q21.1	1., 2. and 3.
CENPN	16q23.2	1. and 2.
ECT2	3q26.31	2
EWSR1	22q12.2	Transcription
FANCI	15q26.1	DNA repair
FXR1	3q26.33	Metabolism
GALT	9p13	Metabolism
GMPS	3q25.31	DNA and RNA synthesis
GTSE1	22q13.31	Microtubule Assembly
H2AFY	5q31.1	2
HMGB2	4q34.1	2
HTATIP2	11p15.1	2
LOC128322	1p13.2	-
MAGOHB	12p13.2	mRNA
NUP62	19q13.33	Splicing
RAD1	5p13.2	1. and 2.
SMC1A	Xp11.22	2.
ZDHHC13	11p15.1	1. and 2.

1. Mitotic G1 phases 2 Cell Cycle 3. Phosphorylation
* This classification is by Reactome and Gene Card function definition

SAM analysis identified a high number of genes either in tamoxifen-treated vs un-treated samples and in metastatic vs non metastatic samples. The identification of just down-regulated genes in tamoxifen-treated patients necessarily excluded the down regulated genes in metastatic patients, as we wanted to identify only the genes that are both involved in metastatic process (up) and affected by tamoxifen treatment (down). Thus it is not unexpected to find only few genes that are both down in tamoxifen-treated samples and up in metastatic samples.

By analyzing the function of the 21 found genes by Reactome, we found that only few pathways are underlined by the program: cell cycle, mitosis and a phosphorylation of Gorasp1 and RAB1A, two proteins involved in the Golgi apparatus structure maintenance. Given that cell cycle includes also mitosis and as phosphorylation of Gorasp1 and Rab1, it represents a critical event for Golgi structure control during mitosis, thus, appearing to be the pathway majorly affected by tamoxifen treatment in metastatic tumors.

IV. CONCLUSION

In this study we performed an analysis of a cohort of BC patients, with and without metastasis, and combined these data with the gene signature of patients treated or not with tamoxifen. We obtained 21 genes as potential molecular BC markers with biological roles in the sensitivity of the tumor to the tamoxifen treatment.

Our findings could be used for further studies with the purpose to identify the driving pathway of the metastatic tumor, which could represent the target of drug treatment.

REFERENCES

[1] Kornelia Polyak Heterogeneity in breast cancer Published in Volume 121, Issue 10 J Clin Invest. 2011; 121(10):3786–3788

[2] Yang G, Nowsheen S, Aziz K, Georgakilas AG Toxicity and adverse effects of Tamoxifen and other anti-estrogen drugs Pharmacol Ther. 2013 May 24. pii: S0163-7258(13)00120-4.

[3] Loi et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* 2008, **9**:239

[4] H.S. Rugo, The breast cancer continuum in hormone-receptor-positive breast cancer in postmenopausal women: evolving management options focusing on aromatase inhibitors. Ann Oncol, 19 (2008), pp. 16–27.

[5] Yixin Wang,et al. Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet, 365(9460):671{679}, February 2005.

[6] Vitoantonio Bevilacqua, et al. Comparison of data-merging methods with svm attribute selection and classification in breast cancer gene expression. In ICIC (3), volume 6840 of Lecture Notes in Computer Science, pages 498{507}. Springer, 2011.

[7] W. Evan Johnson, et al. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1):118{127}, January 2007.

[8] B. Smith, V. G. Tusher, et al. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 98(9):5116{5121}, April 2001

[8] Ingo Mierswa, et al. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, KDD '06: Proc. Of the 12th ACM SIGKDD int. conf. on Know. disc. and data mining, pages 935–940, 2006.

[9] Cava, C., Zoppis I., Gariboldi M., Castiglioni I., Mauri G. and Antoniotti M. (2013) Copy–Number Alterations for Tumor Progression Inference. Artificial Intelligence in Medicine. vol. 7885 of the Lecture Notes in Computer Science series Pages 104-109

[10] 1. Cao J, Panetta R, Yue S, Steyaert A, Young-Bellido M, Ahmad S: A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics 2003, 19(2):234-240.

[11] Ferrari LD, Aitken S: Mining housekeeping genes with a Naive Bayes classifier. BMC Genomics 2006, 7(277).

[12] Sandberg R, Winberg G, Bränden C-I, Kaske A, Ernberg I, Cöster J: Capturing Whole- Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier. Genome Research 2001, 11(8):1404-1409.

[13] Witten, I. H., & Frank, E. (1999). Data mining: Practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann Publishers. pp. 89–97, pp. 125–127, pp. 159–161.

[14] Netaffxtool : www.affymetrix.com/analysis/index.affx