35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer

C. Cava, I. Zoppis, G. Mauri, M. Ripamonti, F. Gallivanone, C. Salvatore,
M. C. Gilardi and I. Castiglioni

Abstract— **Specific genome copy number alterations, such as deletions and amplifications are an important factor in tumor development and progression, and are also associated with changes in gene expression. By combining analyses of gene expression and genome copy number we identified genes as candidate biomarkers of BC which were validated as prognostic factors of the disease progression. These results suggest that the proposed combined approach may become a valuable method for BC prognosis.**

I. Introduction

Breast Cancer (BC) is one of the most common cancers worldwide, with more than 1,300,000 cases and 450,000 deaths each year [1].

Histologic grading quantifies the aggressive behavior of a tumor classifying breast tumors into grade 1 (G1; well-differentiated, slow-growing), grade 2 (G2; moderately differentiated), and grade 3 (G3; poorly differentiated, highly proliferative) malignancies [2].

The effects of chromosome copy number changes on gene expression levels have remained largely unknown although several studies have explored gene expression changes occurring in CNA regions [3]. Many studies in this context applied a combination of cDNA and Array comparative genomic hybridization (CGH) (to detect CNA) . However, SNP array (in this study) offers higher resolution than CGH microarray increasing the ability to detect small CNA [4].

CGH microarrays has been used in BC samples to identify recurrent genome copy number alterations (CNA) between tumors classified according to molecular subtypes, such as histological type or receptor expression [5]. The identification of CNA in genes that are responsible for gene expression regulation is crucial in order to define key genetic events leading to malignant transformation and progression of disease. By combining gene expression data and copy number data these regulators can be revealed. In a limited number of studies [6-8] this approach was adopted for BC prognosis. Callagy et al. measured both genome copy number and gene expression profiles in 101 primary BC samples and found that high-level amplification and/or overexpression of genes at 8p11, 11q13, 17q12, and/or 20q13 were strongly associated with worse prognosis [7]. Other authors used CGH arrays, and by matching gene expression array data showed a significant correlation between DNA copy number alterations and mRNA levels [8].

In our study, we used a combination approach of gene expression and copy number alteration to identify possible genes able to differentiate the progression of BC disease and we identified 49 genes (that had not been detected previously).The prognostic power of this combination approach was validated in case-control studies by the use of a machine learning algorithm compared to existing methods.

Encouraging results suggest that the proposed approach may become a valuable method for BC prognosis.

C. Cava, M. Ripamonti, F. Gallivanone, C. Salvatore, 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: 0039-02-26432715; fax: 0039-02-26415202, 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).

C. Salvatore is with the Department of Physics, University of Milano–Bicocca, Milano, Italy. (e-mail: christian.salvatore@unimib.it).

II. Materials and Methods

A. Gene Expression Analysis

We used two public BC microarray data sets (CEL files) from the Gene Expression Omnibus (GEO) database (GSE11121, GSE2990) and the dataset used by Foekens et al. in [9] (on a collaboration agreement), containing 200, 125 and 180 samples, respectively, for a total of 505 BC microarray data sets. The datasets came from the same Affymetrix GeneChip Human Genome U133A platform. All the samples came from lymph-node-negative patients who were not subjected to any adjuvant systemic treatment, and included both patients without distant metastases and patients with distant metastases.

-Normalization

Gene expression values were computed from microarray data using a Robust Multi-array Average (RMA) method [10].

-Data merging

With the purpose to combine the gene expression data coming from the three different datasets it was necessary to detect and remove the batch effects (experimental variations of datasets generated by different laboratories). An Empirical Bayes method, Combining Batches of Gene Expression Microarray Data (ComBat) was used, and the systematic

difference of differently normalized data generated by the three different laboratories were adjusted [11].

-Identification of up/down regulated genes

We selected two groups of patients from the gene expression dataset (505 samples): 394 patients without distant metastases (class A) and 111 patients with distant metastases (class B). Our aim was in fact to select significant genes based on differential expression between these two classes of samples.

To discover associations between gene expression and the presence/absence of metastasis, a Significance Analysis of Microarray (SAM) was used [12]. SAM identifies statistically significant genes by carrying out gene specific t-tests and computes a statistic measure for each gene, which represents the strength of the relationship between the gene expression and a response variable (e.g. false discovery rate, FDR). More specifically, as a first step, a SAM analysis was used to obtain DNA probes discriminating between the two classes of interest. The genes were considered up/down-regulated if their mean expression in class B were significantly higher/lower (FDR, q value <0.01) than in class A. 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 (www.affymetrix.com/analysis/index.affx).

B. Copy Number Analysis

We used one public BC SNP array data sets (raw and normalized CEL files) from the GEO database (GSE7545) containing 51 samples. All the samples include both patients without distant metastases and patients with distant metastases.

-Normalization

Affymetrix 500K Mapping Array intensity signal CEL files were processed by dChip 2005 (Build date Nov 30, 2005) using the PM/MM difference model and an invariant set normalization.

-Identification of copy number gain/losses

We used Copy Number Analyser for GeneChip (CNAG) [13] to identify the chromosomal regions with gains and losses of DNA.

C. Combination of gene expression and genome copy number alteration (Venn analysis)

Two different combined analysis were performed: (a) up regulated genes versus copy number gains, and (b) down regulated genes versus copy number losses.

Figure 1 gives a simple representation of the proposed combined methodology.

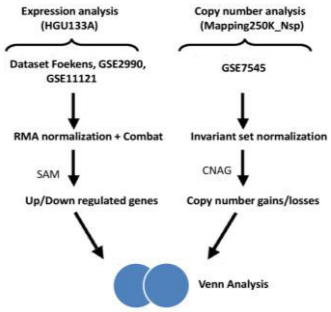

Figure 1. Representation of the proposed combined methodology

D. Validation

To evaluate the performance of the proposed approach based on the combination of gene expression and CNA, we used a machine learning algorithm, trained on the identified up-regulated genes and tested on the ability to differentiate progression of BC samples with respect to the up regulation of these genes. We used the mean expression value of each gene (up-regulated) from the probes.

For this purpose we used published microarray expression BC datasets (CEL files) from the Gene Expression Omnibus (GEO) database (GSE7390 and GSE6532).

Desmedt et al. [14] (GSE7390) validated 76 gene prognostic signature to predict distant metastases and to compare the outcome with clinical risk assessment. They observed a strong time dependence of this signature for time to distant metastasis (TDM) and overall survival. curves.

Loi et al. [15] (GSE6532) reported *PIK3CA* mutation–associated gene signature (*PIK3CA*-GS) and found a relationship with clinical outcome. *PIK3CA*-GS could identify better clinical outcomes in ER+/HER2- disease.

All the samples came from lymph-node-negative patients who were not subjected to any adjuvant systemic treatment.

From the first dataset (GSE7390) 3 classes were obtained as follows: 30 grade I patients (class GI), 83 grade II patients (class GII) and 83 grade III patients (class GIII), considering histological grade has prognostic factor [e.g. 16].

The second dataset (GSE6532) was selected as follows: 112 tumors of lymph-node-negative patients, 34 patients with grade I, 49 with grade II and 29 with grade III.

-Normalization

Expression values from Affymetrix GeneChip Human Genome U133A platform were calculated using the Affymetrix GeneChip analysis software MAS 5.0 (for GSE7390) and the standard quantile normalization method in RMA (for GSE6532).

-Machine learning

To evaluate the performances of proposed approach we designed a Rapid Miner (RM) *workflow* (WF) [17]. 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., [18].

The RM workflow designed for our evaluation implements standard Support Vector Machine (SVM) algorithms to forecast the patient grade. The main issues of this workflow are characterized by the following processes:

a) SVM Parameter Optimization. We iteratively changed SVM parameters to optimize its performance. This was performed by a cross validation process, which in turn trained and tested the SVM algorithm. we optimized the inference accuracy over a space of given SVM feasible learning parameters .The following values are used. kernel.γ: from 0 to 5, step 30; kernel.C: from 0 to 5, step 30; kernel.type ∈ {ANOVA, NEURAL, RADIAL}.

b) Cross Validation. The SVM was validated by a two-step process based on a k-fold cross-validation 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 set.

Figure 2 shows the RM-WF designed for our evaluation.

The performance of the classification was obtained in terms of Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Values (NPV), Accuracy for the following case-control study: GI (control group) vs GII (case group), GI (control group) vs GIII(case group) and GII (control group) vs GIII (case group).

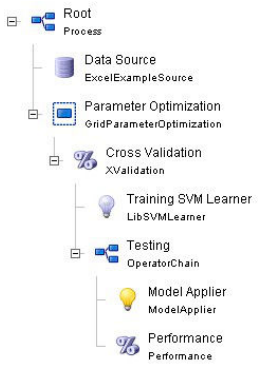

Figure 2. **Data Source Operator** reads data from files (reporting patient representation instances). **Parameter Optimization Operator** assigns a set of defined values to the learning parameters and performs the inner operators for all possible combinations of them. **Cross Validation Operator** encapsulates a n-fold cross validation process: the input data set S is split up into subsets {S1,S2,...,Sn}. The inner operators are applied n times using at each iteration i the set Si as test and S\Si as training set. **SVM Operator** implements a Support Vector Machine algorithm to deliver an inference model**. Model Applier Operator** applies the model delivered by the SVM operator. Finally, **Performance Operator** collects the performance evaluation of the classification task and outputs performance measures.

III. RESULTS

A. Gene Expression Analysis

RMA provided approximately 22.000 probes. Among these genes, 253 were identified with up (193) or down (60) regulation in expression, by the comparison of BC samples of patients without distant metastases (class A) and BC samples of patients with distant metastases (class B).

B. Copy Number Analysis

Copy number gains were frequently observed within regions 1q, 8q, 17q and 20, copy number losses were frequently observed within regions 13q, 1p, and 3. Our findings were consistent with published cytogenetic studies [e.g. 19].

C. Combination of gene expression and genome copy number alteration (Venn analysis)

A low number (<10) of down-regulated genes and copy number losses was found.

Forty-nine up-regulated genes and copy number gains were found. The list of these genes is shown in table I. All genes have functional annotations that could be directly linked with cancer.

By using biological pathway-based analysis Reactome we determined whether 49 up amplified genes are enriched for a particular pathway.

Most part of these genes are implicated in development and progression of BC, e.g.

- Cell cycle (p-value= 3.6e-21)
- Mitotic M-M/G1 phase (p-value= 3.8e-12)
- Phosphorylation of Gorasp1, Golga2 and RAB1A by CDK1-CCNB (p-value= 3.6e-06)
- CDK1 phosphrylates Mastl (p-value=5.0e-05)

TABLE I. GENES AND THEIR POSITION

Gene	Position	Gene	Position
NEK2	1q32.3	BUB1	2q13
CCNE2	8q22.1	CCNA2	4q27
ASPM	1q31.3	CCNB1	5q13.2
VAPB	20q13.32	CCNB2	15q22.2
TTC13	1q42.2	CENPE	4q24
RAE1	20q13.31	HMMR	5q34
TAF5L	1q42.13	LHX2	9q33.3
DDX27	20q13.13	MSH6	2p16.3
BIRC5	17q25.3	CBX5	12q13.13
RBL1	20q11.23	ENC1	5q13.2
SKP2	5p13.2	GTSE1	22q13.31
TPX2	20q11.21	NEIL3	4q34.3
KIF14	1q32.1	ORC6	16q11.2
RRM2	2p25.1	ZWINT	10q21.1
BRD2	6p21.32	SMC4	3q25.33
CMC2	16q23.2	CDK1	10q21.2
RSRC1	3q25.32	KIF23	15q23
SPC25	2q31.1	TTK	6q14.1
CDC20	1p34.2	EZH2	7q36.1
CDKN3	14q22.2	KIFC1	6p21.32
E2F3	6p22.3	MELK	9p13.2
ECT2	3q26.31	PTTG1	5q33.3
FXR1	3q26.33	RFC4	3q27.3

SENP5	3q29	SOX4	6p22.3
RAD1	5p13.2		

D. Validation

Results of Sensitivity, Specificity, PPV, NPV, and Accuracy of the SVM classification are shown in Table II (GSE7390) and Table III (GSE6532), respectively, for the considered case-control study: GI vs GII, GI vs GIII and GII vs GIII. The prognostic power of this combination approach (Comb) was compared to SAM methods. The principal component analysis (PCA) was used for reducing the dimension of SAM genes.

TABLE II. CLASSIFICATION PERFORMANCE (GSE7390)

	GI vs GII		GI vs GIII		GII vs GIII	
	SAM	Comb	SAM	Comb	SAM	Comb
Sensitivity	**100.00%**	91.57%	**89.16%**	85.54%	71.08%	85.54%
Specificity	6.67%	50.00%	56.67%	83.33%	79.52%	79.52%
PPV	74.77%	83.52%	85.06%	93.42%	77.63%	80.68%
NPV	**100.00%**	68.18%	65.38%	67.57%	73.33%	84.62%
Accuracy	75.23%	80.56%	80.49%	84.95%	75.32%	81.33%

TABLE III. CLASSIFICATION PERFORMANCE (GSE6532)

	GI vs GII		GI vs GIII		GII vs GIII	
	SAM	Comb	SAM	Comb	SAM	Comb
Sensitivity	**91.84%**	83.67%	86.21%	93.10%	51.72%	72.41%
Specificity	44.12%	50.00%	88.24%	88.24%	89.80%	89.80%
PPV	70.31%	70.69%	86.21%	87.10%	75.00%	80.77%
NPV	**78.95%**	68.00%	88.24%	93.75%	75.86%	84.62%
Accuracy	**72.22%**	69.89%	87.30%	90.48%	75.64%	83.33%

First we notice that GI vs GIII has a better behaviour providing the ability of the proposed combined approach in differentiating low (GI) and high grade (GIII) BC tumors (GSE7390 accuracy: 84.95% PPV: 93.42%, GSE6532 accuracy: 90.48% NPV: 93.75%). On the contrary, GI vs GII reports the worst behavior (GSE7390 accuracy: 80.56%, specifity: 50% GSE6532 accuracy: 69.89% specifity: 50%), anyway accuracy values are always greater than or equal 70%. We notice an overall judgment that the indexes are clearly better for our methodology (Comb). GI vs GII with Comb (GSE7390: 50%) balances better the specifity of SAM (6.67%). GI vs GII (GSE6532) reports the worst behavior with Comb but the accuracy values are similar (Comb: 69.89%, SAM: 72.22%).This suggests that the proposed approach may become a valuable method for BC prognosis.

IV. CONCLUSIONS

In this study, we performed a genome-wide analysis of genome copy number and gene expression changes in BC to identify genes whose expression were deregulated due to altered copy number. We obtained 49 genes as potential molecular BC markers with biological roles in the development and progression of BC. This suggests evidence of a candidate oncogene role in tumorigenesis. The effects of these copy number changes on gene expression levels have been largely unknown and we have demonstrated that a classification analysis, concerning the disease progression as characterized by progression markers (e.g. grade), shows high performance when using a combination of CNA-based information and gene expression.

V. REFERENCES

[1] Koboldt DC et al. Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61{70, October 2012.

[2] Anna V. et al Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer Cancer research, Vol. 66, No. 21. (1 November 2006), pp. 10292-10301.

[3] E. Hyman, et al. Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer Cancer Research 2002 Vol 62 pages 6240--6245

[4] Wang et al Copy Number Variation Detection via High-Density SNP Genotyping Cold Spring Harb Protoc, Vol. 2008, No. 6. (1 June 2008)

[5] Albertson DG et al. Chromosome aberrations in solid tumors. Nat Genet 2003;34:369–76.

[6] Callagy, G.,et al . (2005). Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. J. Pathol. 205, 388–396.

[7] Chin SF et al. High-resolution array- CGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol 2007;8:R215.

[8] Andre F, et al. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. Clin Cancer Res 2009;15:441–51.

[9] 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.

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

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

[12] 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.

[13] Nannya, Y., et al. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide poly-morphism genotyping arrays. Cancer Research, 65(14):6071–6079.

[14] Desmedt et al. Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series *Clinical cancer research* Vol. 13 2007 pages 3207-3214

[15] Loi S et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* 2007 Apr1;25(10):1239-46.

[16] Rakha EA,et al. Histologic grading is an independent prognostic factor in invasive lobular carcinoma of the breast. Breast Cancer Res Treat 2008, 111:121-127.

[17] 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.

[18] Italo Zoppis,et al. Mutual information optimization for mass spectra data alignment. IEEE/ACM Trans. Comput. Biol. Bioinf., 9(3):934–939, 2012.

[19] Pollack JR et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. PNAS 99(20) pages 12963–12968 October 1, 2002