A new gene expression signature related to breast cancer Estrogen Receptor status

E. Christodoulou, M. Ioannou, M. Kafousi, E. Sanidas, G. Papagiannakis,
V. Danilatou, G. Tsiliki, T. Margaritis, H. Kondylakis, D. Manakanatas, L. Koumakis,
A. Kanterakis, S. Vassilaros, M. Tsiknakis, A. Analyti,
G. Potamias, D. Tsiftsis, E. Stathopoulos, D. Kafetzopoulos

Abstract— **The aim of this study is to identify a gene expression signature which is characteristic of ER status in breast cancer patients. To our knowledge, this is the first microarray study in Greece involving clinical samples. We identified 97 genes that are characteristic for ER status and can well distinguish the ER+ from the ER- samples. We shrank our list to a 11-gene list correlating to the same patient ER status. We found a significant overlap of these genes with published ER status characteristic signatures like the ones of West et.al. [1] and of Van't Veer et. al. [2]. This fact is very important given the minimal overlap of such genes reported by others [3]. In order to obtain a molecular insight into how the expression of estrogen receptor activates cancer cells, we found associations with biological pathways. Interestingly, the vast majority of these genes are highly related to breast cancer.**

I. INTRODUCTION

Expression profiling is a relatively recent technology that has gained the respect of the scientific community and it is now increasingly used. Its effectiveness however may overcome other caveats such as complexity and cost. In the present work we are examining the levels of Estrogen Receptor (ER) and related genes using expression profiling of breast tissue samples.

It is well known that ER status is a strong marker that distinguishes different pathologic subtypes of breast cancer with prognostic implications [2], [1]. ER is a protein found mainly inside the cells of the female reproductive tissue and in some cancer cells. The hormone "estrogen" binds to the receptors inside the cells and may cause the cells to grow [4]. Many microarray studies exist on breast cancer but few of them conclude in a definitive of ER status gene signature [5]. In this study, we have identified a set of genes whose expression profile highly correlates with ER expression and have associated them with possible biological meaning. The motivation of our retrieving genes which are co-expressed with Estrogen Receptor is much deeper than just detecting if the Estrogen Receptor is up or down regulated; this could be done in a straightforward way by simply measuring the expression of the ESR1 gene. We are rather motivated by the fact that the pathways in which these genes participate may reveal invaluable knowledge on how breast cancer is regulated. To our knowledge, this is the first clinical microarray study in Greece aiming at the definition of an ER-related gene signature. This fact issues the use of a new microarray platform and also the interest in examining whether the Greek patient population could be categorized into ER positive and ER negative patients based on known genes or whether there are some specific to the population genes that regulate the Estrogen Receptor status. This attempt is part of the *Prognochip* project. The *Prognochip* project aims to develop an infrastructure that would allow the integration of clinical and genomic information towards the identification and validation of *signature* gene expression profiles of breast tumors correlating with other epidemiological or clinical parameters [6].

II. MATERIALS AND METHODS

A. Samples

26 dissected breast tumors (17 ER+ and 9 ER-) were stored using the RNAlater protocol and were collected from patients treated at University Hospital of Heraklion after institutional review board approval. The classification of the samples as ER+ or ER- was based on immunohistochemical analysis by the pathology laboratory. We assessed the size of our sample using the R package *samr* and found that for a mean difference of $log_2 2$ (between the two groups) the False Discovery Rate (FDR) is 0.05. The False Negative Rate (FNR) is very low.

B. Array fabrication

The human library for our two-color array was obtained from the Qiagen and contains $34,580$ approximately 70mer probes representing $24,650$ genes and $37,123$ gene transcripts. There are $\sim 50\%$ gene transcripts more than genes because 7027 genes have more than one transcript and this is due to alternative splicing. All transcripts of a gene are represented by a probe which corresponds to their common sequence. Further information can be obtained by the human

This work was partially funded by EPAN (AP14394/4-11-03), AKMON (AP.6260EFA1250/17-5-2004) co-funded from the European Regional Development Fund by 70% and from the Hellenic Ministry of Development by 30% through the Operational Programmae Competitiveness, PEP (KR8/16/6/2006) and Crete Innovation Pole 11-PPK-06 projects

The authors are with 1) FORTH, Institute of Molecular Biology and Biotechnology and Institute of Computer Science, P.O. Box 1385, 711 10, Heraklion, Greece, 2)The University of Crete, Medical School, P.O. 2208, Voutes, 71003 Heraklion, Crete, 3)Prolipsis Diagnostic Breast Center, 3 Sevastias Street, 115 28 Athens, Greece
echristo@imbb.forth.gr, mioannou@imbb.forth.gr

V3.0.1 datasheet under [7]. The lyophilized oligo set was resuspended in 3x SSC, 5% DMSO, 0.01% maltoside at a of concentration 10μM by using the liquid handling robot Biomek 2000. The oligos were printed in duplicate spots onto aminosilane glass slides activated with PDITC, using the Packard Array Spotter 24 printer.

C. RNA extraction

Breast cancer tissues were homogenized by using the homogenizer Dia Max of HEIDOLPH and total RNA was extracted with Qiazol Lysis reagent (Qiagen) and further purified on Rneasy columns (Qiagen) according to the manufacturer's instructions.

D. Probe preparation

To obtain enough amplified RNA for a microarray experiment, a round of RNA amplification was performed on all samples. To serve as reference in microarray hybridizations, a human universal reference RNA from Stratagene was amplified identically. Reverse transcription was performed in the presence of 10mM each of dATP, dCTP, dGTP, dTTP(Invitrogen), 0.1M dithiothreitol, 5x First strand buffer, and 200U Superscript III (Invitrogen) using an oligo(dT) T7 primer. The second strand synthesis of cDNA was catalyzed by 20U of E.coli DNA polymerase I (New England Biolabs) while 2U of RNease H (Invitrogen) produces the primers for the cDNA synthesis. The cDNA molecules were transcribed by T7 RNA polymerase (Epicentre) in 20μl reactions at 42^oC for 6 hours. The amplified RNA was purified on RNeasy columns (Qiagen) and quantified in Nanodrop.

E. Probe Labeling and Hybridization

Three mixes of external RNA controls, produced by in vitro transcription, were spiked into the RNA samples. NHS ester of Alexa 647 was added to the tissue RNA reaction and Alexa 555 dye (Molecular probes) was added to the reference RNA reaction. Both reactions were incubated at 50^oC for 3 hours. The unincorporated dye was removed using Microspin G-50 columns (GE Healthcare). Labeling efficiency and quantity of labeled RNA was determined with the spectrophotometer Nanodrop. The ratio of unlabeled to labeled nucleotides was typically between 20 to 30 bases. Slides were prehybridized in 5xSSC, 0.1% SDS, 1%BSA for 90 min and washed with 5x SSC, 0.1% SDS. When two-color arrays are used, breast cancer samples can be compared by hybridizing each sample with a common reference RNA and this is what we followed in the present work. The labeled probes Cy3 and Cy5 were combined and diluted in 85.5μl hybridization buffer (5x SSC, 0.1% SDS, 50% formamide). 15μg fragmented salmon sperm DNA were added to combined samples that were subsequently denatured at 80^oC for 5 min. Hybridizations were carried out using Tecan HyB Station 4800 at 48^oC for 16h, followed by washing in : 2x SSC, 0.1% SDS at 42^oC twice, 0.1X SSC, 0.1% SDS at room temperature twice, and 0.1X SSC at room temperature three successive runs.

F. Scanning and image processing

Arrays were scanned using a GSI Lumonics ScanArray5000. Data were collected in Cy3 and Cy5 channels and stored as TIFF images. Fluorescent intensities of Cy5 and Cy3 channels on each slide were subjected to spot filtering and normalization. In a microarray study, "normalization" identifies and removes systematic sources of variation in the measured intensities due to separate reverse transcription and labeling, different scanning parameters etc. In our study the normalization was performed by using the print-tip loess normalization method which is one of the most commonly utilized normalization techniques.

G. Statistical analysis

We applied the Significance Analysis of Microarrays (SAM) statistic [8] to detect the probes that are differentially expressed in ER+ with respect to ER- samples. This method implements a modified t-statistic in order to reveal the significant probes. More specifically, for every probe i, it calculates a metric d_i.

$$d_i = \frac{r_i}{s_i + s_0} \tag{1}$$

where r_i is the mean difference of each probe's i expressions between the two groups, s_i is the standard deviation of probe's i expression across all samples and s_0 is a small number used to avoid the possibility that the fraction goes to infinity. The larger this metric, the more significant the probe is. The interested reader can refer to the *SAM* manual [9] for further information. The use of permutations in the *SAM* method has been criticized from time to time [10]. However, *SAM* has been extensively used in the past [11] and it is a well recognized method in detecting differentially expressed genes. According to the Science Citation Index, this method is mostly stated among the statistical microarray analysis methods. We support the use of permutation statistics in this paper because it is proven to determine if the expression of any gene is significantly related to a certain characteristic, which here is ER expression. The null hypothesis in this case is that the structure of the data (categorization into ER+ and ER- samples) does not affect the expression of a probe i. For every probe i the statistic (1) is calculated. Then 100 permutations of the samples are realized and another statistic, $\hat{d}_i$, is calculated. For these probes that the calculated statistic is much different from the observed statistic the null hypothesis is rejected; these probes are differentially expressed between the ER+ and ER- samples.

H. Data

In this study we have decided to restrict the training set to 20 out of 26 samples because we selected only the strongly expressed ones ($> 70\%$) ER+ aiming at a clearest signature. The value 70% indicates the percentage in cells of a tissue section where over-expression of the estrogen receptor is observed. For the computational analysis described hereafter we used the *R* [12] language. Before proceeding to the main analysis, we reduced the size of the data in order to minimize computing power and increase efficiency. Thus, we applied a 2-fold filter to the probes, keeping only the probes that were more than 2-fold up/down-regulated in the examined tissues with respect to the reference sample. This procedure resulted in $3,653$ probes, with which we did our manipulations, from the initial set of $34,580$ probes.

SAM analysis requires that its input data are linearly normalized across probes and across samples [9]. We performed the print tip loess normalization, mentioned in the *Scanning and image processing* subsection, across the probes expressed in each tumor; Between array normalization was not required as the means of the probe expressions and the standard deviations between the samples (arrays) were almost equal. It should be assured that the $3,653$ gene expressions in each sample have median 0 according to [9]. The data were thus scaled to median=0 and were ensured to follow the normal distribution depicted in Figure 1.

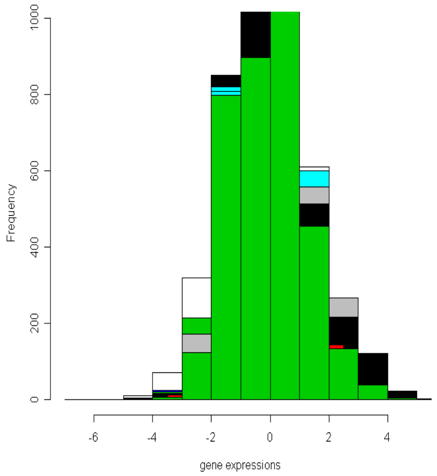

Fig. 1. Superimposed histograms of the distributions of the 3653 probe expressions in each sample. The distributions are normal and that allows us to apply our computational techniques

To exhaustively search the research literature for genes and biological processes we used Biolab Experiment Assistant (BEA) [13], [14] and DAVID tools [15].

III. RESULTS

In order to retrieve the final set of statistically significant genes, the value of the *delta* parameter should be determined. After performing the SAM analysis for our two-class unpaired data, the *delta* parameter was set according to the allowed False Discovery Rate (FDR). The FDR was selected in a way that it was as small as possible (ideally 0) but corresponding to a descent number of probes. It has been declared that an FDR of up to $10\% - 20\%$ is allowed [16]. A *delta* of approximately 0.52 was chosen which corresponds to an FDR of 9%. In this case we retrieved 112 statistically differentially expressed significant probes, 51 of which are ER+ indicators and 61 of which are ER- indicators.

We were interested in comparing the above results with other methods, to check the consistency of the methods. One of the secondary methods that we applied is the *Rank Products* method [17]. By taking a cutoff at the first 60 more significant probes according to their correlation with each category (ER+, ER-), we tried to detect the ones which overlap with our signature. We detected 30 common probes in the ER+ group and 31 common probes in the ER- group. Another method that we applied was the *Pearson Correlation Coefficient* with ER status. This method was adapted from Van't Veer et. al. [2]. More specifically, we calculated the correlation of each probe's expression across experiments with ER status, which is defined as a vector of 0 and 1 values (0 corresponds to ER- and 1 to ER+ tissues). We randomly put a correlation coefficient cutoff at $+0.5$ and at -0.5. The probes whose correlation coefficient was > 0.5 are highly expressed in ER+ samples and the ones whose correlation coefficient was $< (-0.5)$ are highly expressed in ER- samples. In the first case we found 55 probes and in the second 149 probes. 42 of our ER+ correlated probes and 55 of our ER- correlated probes were found in these lists. The above findings are very encouraging; they indicate that the differences in gene expressions reflect real biological significance rather than just statistical significance.

A. Clustering

The heatmap of the probe expression levels of the retrieved significant probes is given in Figure 2, where the detected probes are linked to their corresponding genes. Hierarchical clustering using *euclidean* distance and *complete* linkage is applied to both genes and tumors. The clustering procedure has almost the same results when other distances and other types of linkage are used. The genes are put in clusters according to their expression levels which in turn are indicative of the sample category. The real status of the samples is given in a color scale: blue for ER+, red for ER-. As it can be noticed there exist two cases where ER- samples 88a and 95a are put in the wrong cluster; they are clustered with ER+ samples although they are declared as ER-. By recurring to the raw data, these two samples have indeed values that are closer to the ones observed in ER+ than to the ones observed in ER- samples. As far as the clustering of genes is concerned, three main clusters can be observed; two with high expression in ER- samples and low in ER+, and one cluster with the opposite pattern of expression. In one of the two highly expressed genes in ER- sample clusters, the high gene expressions are much higher than in the other.

B. Classification

In this section we classified our tumors into one of the ER+ and ER- categories. The method we used is very similar to the one reported by Van't Veer et al. [2]. We keep in mind of course that the number of variables that Veer et.a l used (genes and samples) might have led to overoptimistic results regarding the retrieved error [19], [20]. However, many supportive of van't Veer et.al studies exist [21], [22]. Considering that this paper describes one of the first works with satisfying results in microarray analysis, we decided to follow its proposed methodology for both classification and

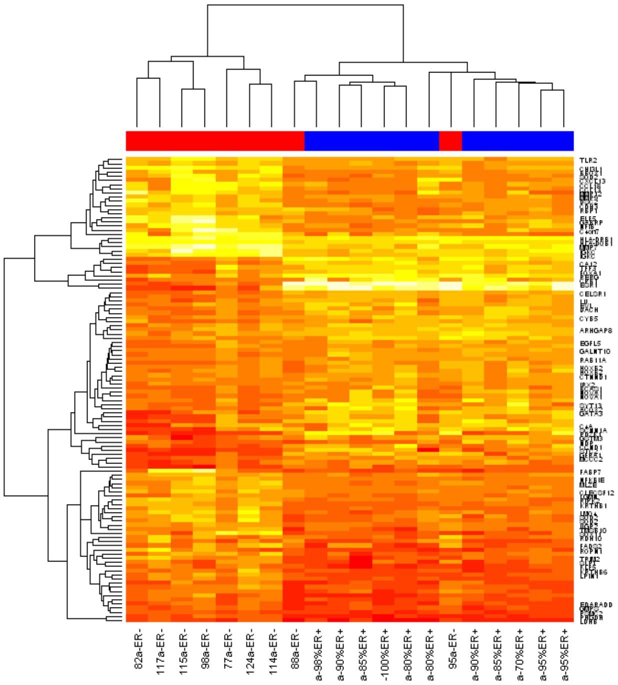

Fig. 2. Expression levels of the significant probes across the 20 samples. Yellow indicates over-expression and red under-expression. The columns correspond to the breast cancer samples. The real status of the samples is given in a color scale: blue for ER+, red for ER-.

extraction of the optimal signature. We took one sample out and measured its correlation coefficient (Pearson correlation) with the average good and poor prognosis expression levels of the remaining samples, using only the expression levels of the 112 significant probes. We repeated this procedure 20 times, so as to cover all the samples. This method's results regarding the two also mentioned before cases (88a and 95a) were discrepant when correlated to immunohistological data produced by the Pathologists. We turned pensive because of the clustering and classification results regarding these two samples and thus sent for a review the respective slides. Indeed, while the infiltrating component of the carcinoma was negative in these cases, in the representative tissue sections an ER+ in situ component (routinely not reported by Pathologists) was present (in a percentage of about 1%), which was detected by our molecular method, most evidently.

C. *A new signature*

Taking the above results into account, we decided to exclude from our analysis the two samples for which the errors were reported. and extract a new signature using the remaining 18 samples. Out of these samples, 11 are ER+ and 7 ER-. The method used was the Significance Analysis of Microarrays (SAM) and the data were confirmed to be correctly normalized also in this case. Using this set of samples, we extracted 97 significant probes, 61 of which were significantly more highly expressed in the ER+ samples and 36 of which were significantly more highly expressed in the ER- samples. A full list of the significant probes and of corresponding gene symbols (in the annotated cases) is given in Supplementary Information 1. Tables I and II represent significant probes and genes along with information on biological pathways and published involvement in breast cancer. These tables were populated with the use of *R* [12] functions and the DAVID tool [15]. The clustering of the 18 tumors based only on the 97 significant probes is given in Figure 3. In this case too, *euclidean distance* and *complete linkage* were used. As it can be noticed, this time all 18 samples were put in the correct cluster. In order to better validate our signature, we also applied a classification method similar to the one described above. All 18 samples were assigned to the correct category. We were also very interested in examining if these 97 significant probes can assign to the correct cluster all of the initial samples which we analyzed using microarrays (all ER+ and not only the ones that are $>70\%$ ER+ - 26 samples). In this case the two problematic samples, 88a and 95a, were labeled as ER+. By applying the same clustering method as above we retrieved the heatmap of Figure 4. As it can be noticed, there exist two samples, 77a and 114a which, although they are ER-, they are put in the ER+ cluster. Nevertheless, two ER+ clusters exist; containing the more highly and less expressed (%) ER+ samples respectively. The reported ER- samples were assigned to this second one cluster which justifies in a way this result. Of 97 probes which correspond to different genes, 19 overlap with the signature found by Veer et.al. [2] and 33 with the signature by West et.al. [1]. This observation is very encouraging; it verifies that there exist some genes that can indeed be indicative of ER status.

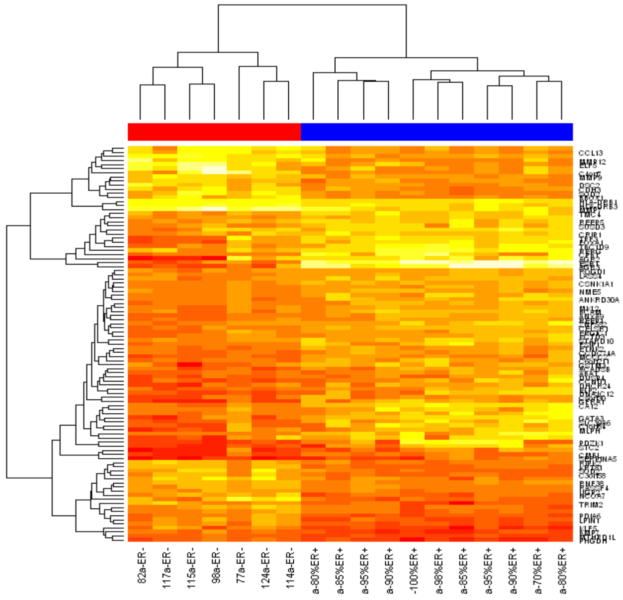

Fig. 3. Expression levels of the significant probes across the 18 samples. Yellow indicates over-expression and red under-expression. The columns correspond to the breast cancer samples. The real status of the samples is given in a color scale: blue for ER+, red for ER-.

D. *Optimal signature*

An intriguing issue is minimizing the false positive rate (fpr) using as less probes as possible.

TABLE I

ER- indicator genes

Probe id	Gene name	Pathway	BC association
H200000442	MMP12		
H200000574	MMP7	Signal transduction (wnt path)	Associated with poor prognosis, wound healing and metastasis)
H200013790	MMP9		Tumor invasion and angiogenesis in BC and other cancers
H200015680	DSC2	Cell communication	
H300022173	KRTHB1	Cell communication	
H200000695	CDH3	Signaling molecules	Tumor aggressiveness in BC LOH events of Chr16p in BC
H300022893	HLA-DRB1	Signaling molecules, immune system	
H300021886	HLA-DRB3	Signaling molecules, immune system	
H200001995	CCL13	Signaling molecules, immune system	
H200005719	GABRP	Signaling molecules: Neuroactive ligand-receptor interaction	Down regulated in BC, Index of tumor progression, prognostic marker
H300004703	CCL18	Signaling molecules: Cytokine-cytokine receptor interaction	
H200007916	BMP7	Signaling molecules, signal transduction Expressed in BC, may associated with bone metastasis	Expressed in various breast cancer cell lines
H200007119	KLF5		Suppresses tumor cell growth in breast cancer
H200006022	CHI3L1		High serum levels of YKL-40 associated with poor prognosis
H200004198	SOX11		Play a role in tumorigenesis
H200009652	RASSF4		Potential tumor suppressor. May promote apoptosis and cell cycle arrest
H300021244	UGP2	Carbohydrate metabolism,	Forms UDP-glycose which in mammary lactating gland is converted to Udp-galactose and lactose
H200004673	MTHFD1L	Carbohydrate metabolism, metabolism of cofactors and vitamins	
H300006924	PHGDH	Aminoacid metabolism	
H300002199	BBOX1	Aminoacid metabolism (lysine degradation)	

TABLE II

ER+ indicator genes

Probe ID	Gene name	Pathway	BC association
H200000435	ESR1		Stimulation of growth of breast cancer. Response to endocrine therapy
H300000645	NME5		Confers protection from cell death by Bax and alters the cellular levels of several antioxidant enzymes including Gpx5.
H200007883	GATA3		Highly associated with ER & PgR, tumor grade. Involved in growth control & maintenance of differential state in epithelial cells
H300003818	FOXA1		Decreased expression in BC. Mediated ER in BC cells
H200003045	RERG		Loss may contribute to tumorigenesis in breast. Decreased in BC with poor prognosis
H200010467	AGR2		Associated with ER+ BC. Interacts with metastasis genes Potential therapeutic target and molecular marker in prostate cancer
H200014049	STC2		Expression induced by estrogen, altered in BC
H200006989	CCND1	Cell growth and death	Regulated positively by Rb. Mutations, amplification and overexpression of this gene, are observed frequently in a variety of tumors and may contribute to tumorigenesis.Better outcome
H200006652	BCL2	Upregulated in response of human prolactin treatment in BC cancer cell lines.	Expressed in BC, inverse correlation with cytological grade
H300002542	NAT1/NAT2	Caffeine metabolism, drug metabolism	Polymorphisms associated with BC risk
H200006150	DHCR24	Protects cells from oxidative stress by reducing caspase 3	Activity during apoptosis induced by oxidative stress
H200000512	GSTM3	Aminoacid metabolism	
H200017772	ABAT	aminoacid and carbohydrate metabolism	
H200007735	MCCC2	Aminoacid metabolism	
H300004674	ETNK2	Lipid metabolism	
H200006864	ACADSB	Lipid metabolism	
H200001041	CA12	Energy metabolism	
H300015296	CSNK1A1	Signal transduction	Association with BC metastasis
H200000600	DUSP4	Signal transduction	
H200014021	BCAM		Up-regulated following malignant transformation in some cell types. Play a role in epithelial cell cancer
H200006636	SLC39A6		Better outcome in BC. Upregulated by estrogen in BC cell lines.
H200016503	DNAJC12		(DeBessa SA 2006) Correlation with ER in BC
H300003702	PDZK1		May play a role in the cellular mechanisms associated with multidrug resistance through its interaction with ABCC2 and PDZK1IP1.
H200019227	ANKRD30A		NY-BR-1 is a differentiation antigen present in BC Possible antigenic target for antibody treatment
H200020432	CMBL	Xenobiotics biodegradation	
H200006282	SERPINA5	Immune system	Positive prognostic factor (suppression of tumor invasion)

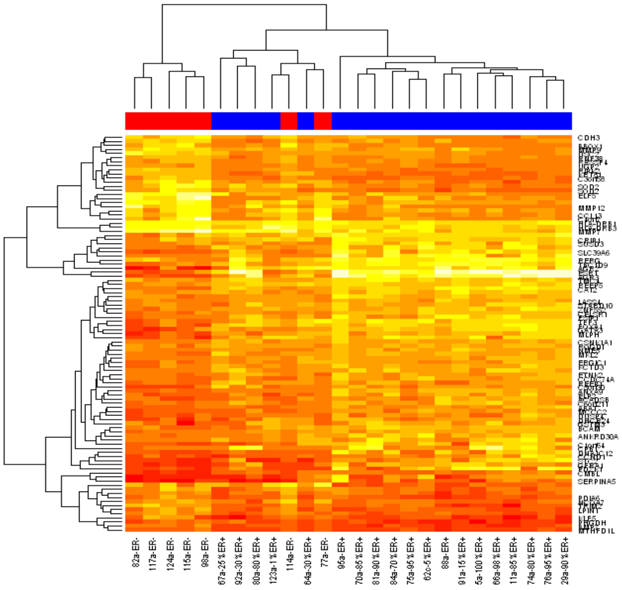

Fig. 4. Expression levels of the significant genes across all the 26 analyzed samples. Yellow indicates over-expression and red under-expression. The columns correspond to the breast cancer samples. The real status of the samples is given in a color scale: blue for ER+, red for ER-.

Thus, they try to find a sub-signature of the original one which has the same, or even better, result. We followed the following procedure described in [2]:

- Rank the retrieved genes in decreasing order according to the correlation of their score with the ER status of the patients
- Take one by one the genes, starting from the top of the ordered list, and classify the tumors based on their expression profiles
- Define the cutoff point, meaning the minimum number of genes having the minimum error

In our case, we ranked the probes and started classifying the tumors using a signature starting from 2 genes and ending at 97 genes, adding one gene at each step. The number of misclassifications was high for 2 genes and continued to decrease until the use of 11 genes (Figure 5) where it finally reached its optimal number of 0 misclassifications in both the ER+ and the ER- tumors. This zero error remained stable from there on.

IV. CONCLUSIONS AND FUTURE WORK

A. Conclusions

In this paper we defined a gene signature indicative of the ER status of breast cancer patients from Greece. The very interesting finding of this work is that two samples that were initially defined to be ER-, were found to be ER+ using either classification or clustering methods. By resorting to the pathologists, they investigated into these cases and re-annotated these two samples as ER+, a fact that strengthens the computational method used.

As it can be noticed from Tables I and II, the vast majority of the retrieved genes in our signature are highly related to breast cancer. Moreover, there exists a great overlap with pre-defined such prognostic signatures. Two of the genes that are retrieved as highly significant in our list, which also appear in our 11-gene signature and overlap with both the Veer et.al. [2] and the West et.al. [1] datasets, are *FOXA1* and *GATA3*. Supporting bibliography exists that these genes are co-expressed with ER [18], [23]. *FOXA1* is a transcription factor with favorable prognostic significance which seems to play an essential role in the expression of approximately 50% of ERα target genes and it has already been suggested as a possible therapeutic target for breast cancer [24]. It is also declared to be involved in a growth inhibitory role [25]. The zinc-finger *GATA3* transcription factor plays an important role in breast cancer as it is involved in the growth control and differentiation of breast epithelial cells. Moreover, *STC2*, *CCND1*, *BCL2* and *TFF3* (see ER positive genes supplementary information) are four other ER+ indicative genes that are induced by estrogen, involved in apoptosis and response to injury and xenobiotics respectively [26], [27], [28], [29]. *TFF3* is also reported to be a marker of disseminated breast cancer cells [30]. From the list of the ER- related genes, great interest is given on *KRTHB1*, which is the only ER- gene that exists in the 11 gene signature and is involved in cell communication, as noticed in Table I. It is one of the genes found to mark and mediate breast cancer metastasis to the lungs in mice [31]. Another set of genes that exist in our signature and are related to breast cancer are the matrix metalloproteinases (MMP-family), *MMP7*, *MMP9* and *MMP12*. *MMP9* is highly expressed in luminal A breast carcinomas [32], whereas *MMP7* and *MMP12* have been correlated with poor prognosis breast cancer [33], [34]. *CDH3*, *BMP7* and *KLF5* are other markers with potential biological significance in breast cancer [15].

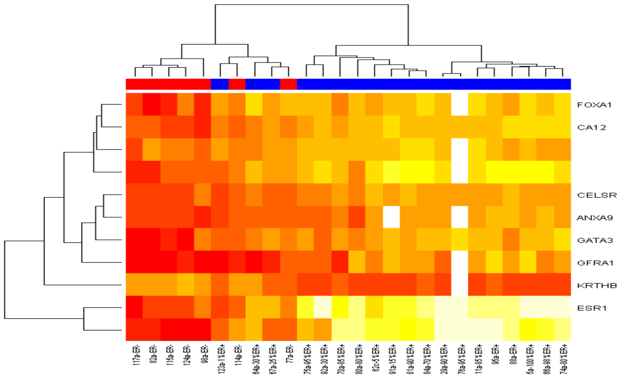

Fig. 5. Heatmap of the 11 most significant genes across the samples. Yellow color indicates over-expression and red under-expression. There exists 1 gene that is up-regulated in ER- tumors and 10 genes that are up-regulated in ER+ tumors. The clustering of tumors and genes is made using euclidean distance and complete linkage.

B. Future Work

One of the future aims is to increase the sample size. With the use of the *samr* package of *R* [12], it was found that depending on the number of genes truly changed at 2-fold rate, the sample size should be increased to 36 or 54, in order to get FDR$<$ 0.05, which means smaller than the one in the current study.

The associations of the gene signatures with biological pathways and other important processes in breast cancer tumorigenesis might give new insight in disease pathogenesis and reveal new molecular targets in the treatment of breast cancer. We intend to study the involvement of whole biological pathways in order to establish pathway signatures instead of simple gene lists. In that way we may will be able to improve our knowledge in that field and use the vast amount of information produced by microarrays in the most efficient way.

REFERENCES

[1] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, Jr., J. R. Marks and J. R. Nevins, ”Predicting the clinical status of human breast cancer by using gene expression profiles”, *PNAS*, vol. 98, no. 20, pp 11462-11467 (2001)

[2] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, ”Gene expression profiling predicts clinical outcome of breast cancer”, *Nature*, vol. 415, pp 530-536 (2002)

[3] L. Ein-Dor, I. Kela, G. Getz, D. Givol and E. Domany, ”Outcome signature genes in breast cancer: is there a unique set”, *Bioinformatics*, vol. 21, pp 171-178 (2005)

[4] URL: http://www.cancer.gov/

[5] C. R. Acharya, D. S. Hsu, C. K. Anders, A. Anguiano, K. H. Salter, K. S. Walters, R. C. Redman, S. A. Tuchman, C. A. Moylan, S. Mukherjee, W. T. Barry, H. K. Dressman, G. S. Ginsburg, K. P. Marcom, K. S. Garman, G. H. Lyman, J. R. Nevins and A. Potti, ”Gene Expression Signatures, Clinicopathological Features, and Individualized Therapy in Breast Cancer”, *Jama*, vol. 299, No. 13, pp 1574-1587 (2008)

[6] G. Potamias, A. Analyti, D. Kafetzopoulos, M. Kafousi, T. Margaritis, D. Plexousakis, P. Poirazi, M. Reczko, I.G. Tollis, M. E. Sanidas, E. Stathopoulos, M. Tsiknakis, S. Vassilaros, ”Breast Cancer and Biomedical Informatics: The PrognoChip Project”, *IMACS 2005: Computer Science and Artificial Intelligence - Bioinfoamtics session*, Paris, France, July 11-15 (2005)

[7] URL: http://www.biotech.kth.se/molbio/microarray

[8] V. Tusher, R. Tibshirani, and G. Chu., ”Significance analysis of microarrays applied to transcriptional responses to ionizing radiation.”, *Proc. Natl. Acad. Sci. USA.*, vol. 98, pp 5116-5121 (2001)

[9] G. Chu, B. Narasimhan, R. Tibshirani and V. Tusher, ”SAM Significance Analysis of Microarrays Users guide and technical document”

[10] T. Hsing, S. Attoor, E. Dougherty, ”Relation Between Permutation-Test P Values and Classifier Error Estimates”, *Machine Learning*, vol. 52, pp. 11-30 (2003)

[11] T. Sorlie, C. M. Peroua, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning and A.-L. Borresen-Dale, ”Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”, *PNAS*, vol. 98, no. 19, pp 10869-10874 (2001)

[12] URL: http://www.r-project.org/

[13] URL: http://www.biovista.com/bea

[14] D. Mastellos, C. Andronis, A. Persidis and J. D. Lambris, ”Novel biological networks modulated by complement”, *Clin Immunol.*, vol 115, issue 3, pp 225-35 (2005)

[15] URL: http://david.abcc.ncifcrf.gov/

[16] S. G. Baker and B. S. Kramer, ”Using microarrays to study the microenvironment in tumor biology: The crucial role of statistics.”, *Seminars in Cancer Biology* (2008), doi:10.1016/j.semcancer.2008.03.001

[17] R. Breitling, P. Armengaud, A. Amtmann. and P. Herzyk, ”Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments”, *FEBS Letters*, vol. 573, pp 83-92 (2004)

[18] M. Brown and J. S. Carroll, ”Estrogen receptor target gene: an evolving concept.”, *Mol Endocrinol*, vol. 20 , pp 1707-14 (2006)

[19] S. K. Gruvberger, M. Ringnr, P. Edn, . Borg, M. Fern, C. Peterson and P. S. Meltzer, ”Expression profiling to predict outcome in breast cancer: the influence of sample selection”, *Breast Cancer Res.*, vol. 5, pp. 23-26 (2003)

[20] Braga-Neto, U. ”Fads and fallacies in the name of small-sample microarray classification - A highlight of misunderstanding and erroneous usage in the applications of genomic signal processing”, *IEEE Signal Processing Magazine*, vol. 24, pp. 91-97 (2007)

[21] S. R. Morris and L. A. Carey, ”Molecular profiling in breast cancer”, *Rev. Endocr. Metab. Disord.*, vol. 8, pp 185-198 (2007)

[22] H. Moon, H. Ahn, R. L. Kodell, C.-J. Lin, S. Baek and J. J. Chen, ”Classification methods for the development of genomic signatures from high-dimensional data”, *Genome Biology*, vol. 7, issue 12, R121 (2006)

[23] J. Schneider, M. Ruschhaupt, A. Buneß, M. Asslaber, P. Regitnig, K. Zatloukal, W. Schippinger, F. Ploner, A. Poustka and H. Sultmann, ”Identification and meta-analysis of a small gene expression signature for the diagnosis of estrogen receptor status in invasive ductal breast cancer”, *Int. J. Cancer*, vol. 119, pp 2974-2979 (2006)

[24] H. Nakshatri and S. Badve, ”FOXA1 as a therapeutic target for breast cancer”, *Expert Opin Ther Targets*, vol. 11, pp 507-14 (2007)

[25] H. P. Koeffler, C. W. Miller, B. Y. Karlan, I. Wolf, S. Bose and E. A. Williamson, ”FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer.”, *Int J Cancer*, vol. 120 , pp 1013-22 (2007)

[26] N.M. Malara, A. Leottab, A. Sidotia, S. Liob, R. DAngeloa, B. Caparellob, F. Munaoc, F. Pinoa and A. Amato, ”Ageing, hormonal behaviour and cyclin D1 in ductal breast carcinomas”, *The Breast*, vol. 15, Issue 1, pp 81-89 (2006)

[27] T. Bouras,M. C. Southey, A. C. Chang, R. R. Reddel, D. Willhite, R. Glynne, M. A. Henderson, J. E. Armes and D. J. Venter, ”Stanniocalcin 2 Is an Estrogen-responsive Gene Coexpressed with the Estrogen Receptor in Human Breast Cancer”, vol. *Cancer Research*, vol. 62, pp 1289-1295 (2002)

[28] G. M. Callagy, M.J. Webber, P.D Pharoah and C. Caldas, ”Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer”, *BMC Cancer*, vol. 8, issue 153 (2008)

[29] S. Tozlu, I. Girault, S. Vacher, J. Vendrell, C. Andrieu, F. Spyratos, P. Cohen, R. Lidereau and I. Bieche, ”Identification of novel genes that co-cluster with estrogen receptor alpha in breast tumor biopsy specimens, using a largescale real-time reverse transcription-PCR approach”, *Endocrine-Related Cancer*, vol. 13, pp 1109-1120 (2006)

[30] M. Lacroix, ”Significance, detection and markers of disseminated breast cancer cells”, *Endocrine-Related Cancer*, vol. 13, pp 1033-1067 (2006)

[31] A. J. Minn, G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, A. Viale, A. B. Olshen, W. L. Gerald and J. Massague, ”Genes that mediate breast cancer metastasis to lung”, *Nature*, vol. 436, pp 518-524 (2005)

[32] P. Neven, R. Paridaens, H. Wildiers, A. Smeets, W. Hendrickx, M. Drijkoningen, J. Decock, ”Matrix metalloproteinase expression patterns in luminal A type breast carcinomas.”, *Dis Markers*, vol. 23, pp 189-96 (2007)

[33] S. Mochizuki, M. Shimoda, T. Shiomi, Y. Fujii and Y. Okada, ”ADAM28 is activated by MMP-7 (matrilysin-1) and cleaves insulin-like growth factor binding protein-3”, *Biiochemical and Biophysical Research Associations*, vol. 315, Issue 1, pp 79-84 (2004)

[34] P. M. McGowan and M. J. Duffy, ”Matrix metalloproteinase expression and outcome in patients with breast cancer: analysis of a published database”, *Annals of Oncology*, doi:10.1093/annonc/mdn180 (2008)