doi:10.4149/neo_2015_099

A 39-gene signature is associated with early occurrence of distant metastasis in primary lymph-node negative breast cancers

R. XIA[1,‡,*], S. CHEN[2,‡], W. ZHANG[3], Y. CHEN[3], R. ZHU[3], A. DENG[3]

[1]Department of Transfusion, Huashan Hospital, Fudan University, Shanghai 200040, China; [2]Department of Dermatology, Changzheng Hospital, Second Military Medical University, Shanghai 200003, China; [3]Department of Laboratory Diagnosis, Changhai Hospital, Second Military Medical University, Shanghai 200433, P. R. China

**Correspondence: rongxiahs@163.com*
‡Contributed equally to this work.

Received March 13, 2015 / Accepted May 25, 2015

Risk factors of the development of distant metastasis in primary node-negative breast cancer patients are heterogeneous. Identification of patients at high risk of early distant metastasis is of important clinical significance. In the current study, using the already published datasets, we develop a gene signature that can robustly predict early distant metastasis for patients with primary node-negative breast cancer. We identified a 39-gene signature, which were associated with distant metastasis and shorter distant metastasis free survival (DMFS) in node-negative breast cancers. Using the survival prediction analysis method in BRB-Array tools, this signature can stratify patients into early- and late- distant metastasis subgroups with different DMFS in VDX training dataset (AUC=0.734, $P < 0.01$). And we further validated the reliability of the prognostic value of this 39-gene signature in another two independent breast cancer cohorts (NKI dataset, AUC=0.642, $P<0.0167$; TRANSBIG dataset, AUC=0.711, $P<0.0167$). Furthermore, the early distant metastasis subgroups defined by the 39-gene signature exhibited a significant association with ER negative status and more aggressive molecular subtypes in all three datasets, and with poor differentiation status in two datasets. In summary, we developed a novel distant metastasis-related gene signature for predicting early occurrence of distant metastasis in node-negative breast cancers, what might be useful in making treatment decisions for these early metastasis patients.

Key words: breast cancer, distant metastasis, gene signature, microarray

Despite considerable progress in regional and systemic treatments, up to 30% of lymph-node negative breast cancer patients are at a high risk of relapse and distant metastasis [1]. Currently, metastasis at distant sites is still a key cause of mortality from breast cancer. Gene-expression profiles have been widely used to establish molecular signatures to improve prognostic accuracy, treatment choice, and disease outcomes prediction in multiple cancers including breast cancer [2 3]. In recent decades, a panel of gene signatures has been described for predicting distant metastatic risk or survival in primary breast cancer [4 5].

Primary node-negative breast cancer is highly heterogeneous in prognosis, even among the patients with distant metastasis, the duration of developing first distant metastasis after diagnosis of primary tumors may also vary greatly, which may span 20 years [6]. Therefore, it is of importance to more accurately predict the risk of early occurrence of distant metastasis for node-negative breast cancers.

In this study, using the previously published gene expression microarray datasets, we developed a novel gene signature which can be used to identify node-negative breast cancer patients at high risk of early occurrence of distant metastasis.

Patients and methods

Datasets. Four previously published microarray datasets were used in our study. GSE46141 dataset was used to identify distant metastasis-specific genes, it is derived from fine-needle aspiration biopsies of breast cancer metastases from different anatomical sites including 23 distant metastases (18 liver, 4 bone, 1 ascite) and 67 local metastases (breast, regional lymph

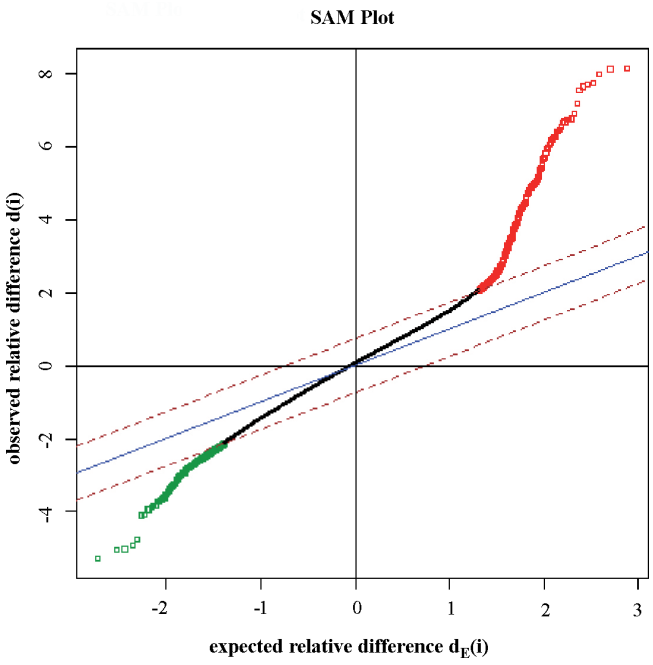

Figure 1. Significance Analysis of Microarrays (SAM) analysis of differently expressed genes between local metastatic and distant metastatic breast cancers in GSE46141 dataset. 297 genes were found to be differentially expressed in metastatic breast cancers compared with metastatic breast cancers including 212 up-regulated and 85 down-regulated genes.

node or skin) [7]. The expression, annotation and clinical data of GSE46141 were downloaded from Gene Expression Omnibus (GEO) database.

Training dataset VDX was used to develop the predictive gene signature for early distant metastasis, derived from two different consecutive series [8 9]. Datasets NKI and TRANSBIG were used to validate the predictive value of developed gene signature. Dataset NKI contains the gene expression data by van't Veer et al. [10] and van de Vijver et al. [11]. TRANSBIG dataset contains the gene expression data as published from TRANSBIG in 2006 [12]. The expression, annotation and clinical data of training and validation datasets were extracted from VDX, NKI and TRANSBIG R/Bioconductor package. In this study, only node-positive breast cancer patients with positive distant metastasis were included (VDX, n=118; NKI, n=62; TRANSBIG, n=62).

Comparison of differently expressed genes between distant and local metastatic breast cancers in GSE46141 dataset. Significance Analysis of Microarrays (SAM) was used to compare the expression profiles of 23 distant and 67 local metastatic tumors in GSE46141 dataset by BRB-ArrayTools as described previously [13]. One thousand permutations were used to estimate the FDR and to select differentially expressed genes. Target proportion of false discoveries was set at 0.05.

Genes correlated with distant metastasis free survival (DMFS) in training VDX dataset. The differently expressed metastasis-related genes identified by SAM analysis were further used in univariate Cox proportional hazards model to test which of these genes significantly influenced patients' DMFS in VDX dataset. This procedure was also performed by BRB-ArrayTools. The parameter of Significance threshold of univariate tests was set at 0.001, and number of permutations at 1000.

Development and validation of gene signature for early distant metastasis risk prediction. Survival risk prediction of the gene expression data by the metastatic gene signature was performed based on principal components with BRB-ArrayTools software. Survival risk prediction was based on 10-fold cross-validation. The patients were dichotomized into groups at high- or low- risk at early distant metastasis using the 50th percentile of the prognostic indexes (above and below the median expression across all samples in each training and validation datasets).

Gene ontology analysis of gene signature. The gene list of early distant metastasis predictive gene signature was put into Protein Analysis THrough Evolutionary Relationships (PANTHER), a software system for inferring the functions of genes based on their evolutionary relationships [14]. The enrichment of these genes in GO molecular function, biological process and PATHER class was analyzed.

Statistics. Distributions of DMFS were assessed using the Kaplan-Meier curve method and log-rank statistics. Correlation between predictive risk subgroups with clinical characteristics was analyzed using the X^2 test. Time-dependent receiver operating characteristic (ROC) curves of DMFS were measured by R package survivalROC implemented in BRB-Arraytools [15]. All the other statistic analyses were performed by Medcalc Software. $P < 0.05$ was considered statistically significant.

Results

Using the Significance Analysis of Microarray (SAM) method, 297 genes were found to be differentially expressed between distant and local metastatic breast cancers from GSE46141 dataset (figure 1). A large proportion of these genes are known to regulate metastasis in cancer, with 212 up-regulated in distant metastatic samples and 85 down-regulated (data not shown).

Using the BRB-ArrayTools, we found 39 from 297 genes were significantly associated with patients' DMFS in VDX dataset including 17 positively associated and 22 negatively associated (permutated $P < 0.001$) (table 1).

Using the survival prediction analysis model in the BRB-arrayTools, the prognostic indexes were calculated based on this 39-gene signature. Patients in VDX training dataset were partitioned into early- and late distant metastasis subgroups.

As seen in figure 2, early metastasis group had a significantly shorter DMFS than late metastasis group ($P<0.0001$).

The predictive accuracy of the 39-gene signature for early-distant metastasis is 0.734 ($P < 0.01$) as measured by the AUC. Furthermore, we also found that early-distant metastasis was significantly associated with ER negative status, that is 64.3% of predicted early-distant metastasis cases were ER negative, whereas only two of late metastasis cases were ER negative ($P<0.0001$) (table 2).

The reliability of the 39-gene signature in predicting the risk of early distant metastasis was further validated in other two independent for lymph node negative breast cancer datasets (NKI and TRABSBIG). As seen in figure 2, in both validation datasets, the patients predicted as early metastasis subgroup exhibited shorter DMFS than the late counterparts, and demonstrated a moderate discriminatory power (NKI dataset, AUC=0.642, $P<0.01$; TRANSBIG dataset, AUC=0.711, $P<0.01$). Same as training VDX dataset, high-risk of early metastasis was positively correlated with ER negative status in both of the validation datasets. Furthermore, we found that high risk of early distant metastasis was also positively associated with poor differentiation status

Table 1. Identification of a 39-gene signature associated with early occurrence of distant metastasis in lymph-node negative breast cancers from VDX dataset

	Parametric p-value	FDR	Permutation p-value	Hazard Ratio	SD of log intensities	Gene
1	< 1e-07	< 1e-07	< 1e-07	1.439	1.65	WWTR1
2	0.0000392	0.00136	< 1e-07	1.991	0.57	PRDX4
3	0.000027	0.00102	< 1e-07	1.944	0.692	KIF2C
4	0.0000783	0.00191	< 1e-07	1.87	0.635	PTDSS1
5	0.000679	0.00801	< 1e-07	1.75	0.603	ATP1B3
6	0.0007607	0.00853	0.002	1.714	0.682	KIF20A
7	0.0005502	0.00737	< 1e-07	1.696	0.754	HJURP
8	0.0002651	0.00472	0.001	1.57	0.904	CCNA2
9	0.0005867	0.00761	0.001	1.556	0.836	DLGAP5
10	0.0000514	0.00164	< 1e-07	1.468	1.167	CDCA3
11	< 1e-07	0.0000208	< 1e-07	1.458	1.595	CDCA8
12	0.0002937	0.00488	< 1e-07	1.429	0.916	HMGB3
13	0.0000157	0.000814	< 1e-07	1.39	1.367	LTBP1
14	0.0000218	0.000905	< 1e-07	1.37	1.465	KIFC1
15	0.000693	0.00801	< 1e-07	1.257	1.469	CRYAB
16	0.0000118	0.000814	< 1e-07	1.217	2.427	CDH3
17	0.0005015	0.00694	< 1e-07	1.164	2.162	NFIB
18	0.000695	0.00801	0.001	0.844	1.937	DNAJC12
19	0.0000013	0.00018	< 1e-07	0.813	2.37	ESR1
20	0.0000654	0.0017	< 1e-07	0.813	1.949	SLC44A4
21	0.0002106	0.00416	< 1e-07	0.79	1.373	ARNT2
22	0.0004165	0.00617	0.001	0.785	1.484	GPRC5C
23	0.0000139	0.000814	< 1e-07	0.782	1.782	ABAT
24	0.0000612	0.00169	< 1e-07	0.782	1.824	MAPT
25	0.0002277	0.0043	0.001	0.776	1.376	CYP21A2
26	0.0003673	0.00565	0.001	0.715	1.06	MST1
27	0.0001712	0.0039	< 1e-07	0.694	1.081	QDPR
28	0.0007929	0.00866	0.001	0.69	1.002	FBP1
29	0.0009766	0.0104	< 1e-07	0.682	0.827	ABCG1
30	0.0000153	0.000814	< 1e-07	0.674	1.133	BCL2
31	0.0001786	0.0039	< 1e-07	0.665	0.86	SORL1
32	0.0006765	0.00801	0.001	0.66	0.788	FLT3
33	0.0000565	0.00167	< 1e-07	0.644	0.912	IQGAP2
34	0.0000184	0.000848	< 1e-07	0.615	0.813	PBLD
35	0.0004377	0.00626	0.002	0.609	0.739	SLC29A3
36	0.0002727	0.00472	< 1e-07	0.582	0.723	P4HTM
37	0.0003323	0.0053	< 1e-07	0.569	0.586	ROGDI
38	0.0001993	0.00414	< 1e-07	0.483	0.574	NBR1
39	0.0000038	0.000394	< 1e-07	0.441	0.613	MYO15B

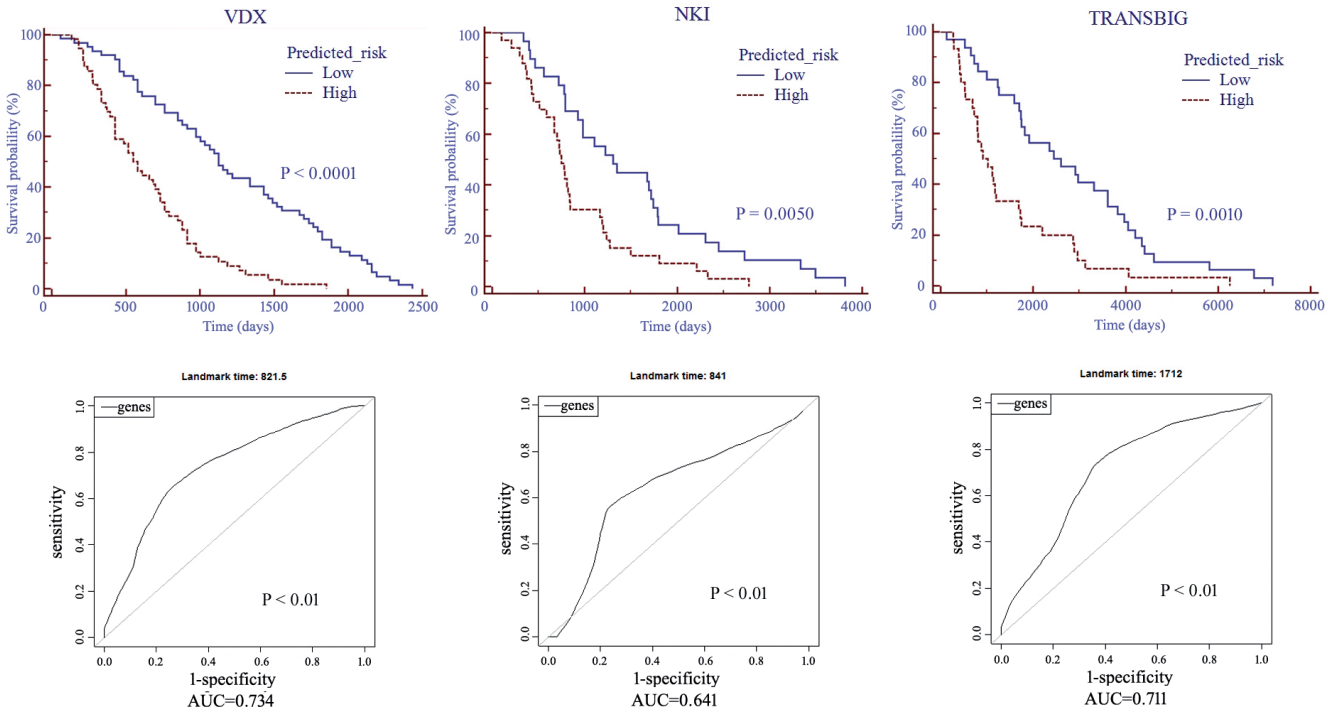

Figure 2. Distant metastasis-free survival (DMFS) prediction and ROC curves using a 39-gene signature in training (VDX) and validation (NKI and TRANSBIG) datasets. In all of the datasets, patients at high risk of early distant metastasis demonstrated a significantly shorter DMFS than those at low risk, and ROC curves demonstrated a moderate discriminatory power of this gene signature.

(P=0.0263 in NKI dataset; P<0.0001 in TRANSBIG dataset) (table 2). Interestingly, we found that this 39 gene signature was significantly correlated with more aggressive molecular subtypes (basal-like and HER2 breast cancers) in all the three datasets (P<0.001) (table 2).

Finally, we performed a Gene ontology analysis on the 39 genes comprising of gene signature. As shown in figure 3, our results indicated that this signature showed enrichment in many molecular function, biological process and PATHER protein class related to tumor progression and metastasis.

Discussion

For clinical heterogeneity, predicting the metastasis of breast cancer patients using gene expression signature has been widely investigated. Currently, three series of gene signatures have been developed to estimate the risk of distant metastasis in breast cancers. Tutt et al. [16] developed an RT-PCR based 14-gene signature significantly associated with risk of distant metastasis in node-negative, estrogen receptor-positive breast cancer. Recently, Drukker et al. [17] demonstrated a significant difference in DMFS between high- and low- risk subgroups classified by the MammaPrint 70-gene signature in both node-negative and -positive breast cancer patients. Mittempergher et al. [6] identified a 241-gene signature for late metastasis in breast cancer.

In this study, we first identified a panel of metastasis-related genes that were associated with DMFS in the training test. Based on these genes, we have identified a 39-gene expression signature which can classify node-negative, metastasis-positive breast cancer patients into high- and low- risk groups with different DMFS. The predictive value of the 39-gene signature risk evaluation system was further validated in two independent cohorts of datasets. Our 39-gene signature is a novel DMFS-related gene classifier; there is no gene overlap with the 14-gene panel, 70-gene panel, only one gene ESR1 overlap with the 241-gene panel. However, this 39-gene signature correlated closely with aggressive molecular subtypes. In particular, nearly all the basal-like breast cancers were predicted as high-risk of early distant metastasis by this signature.

Additionally, different from the previous studies, all patients in our analysis were distant metastasis-positive; cases with censored DFMS data were excluded. Therefore, we emphasized more its utility in predicting subgroup at high risk of early occurrence from the other patients which would develop distant metastasis relative lately. That also means these patients at high-risk of early metastasis might need more intensive treatments.

For their relationship with rapid onset of distant metastasis, it is reasonable to speculate that the 39-gene signature might be of functional significance in the metastasis development of breast cancer. Ontology and protein

Table 2. Association between high- and low- risk of early distant metastasis determined by a 39-gene signature and clinical parameters

	Low risk	High risk	P value
	Dataset VDX		
Number of cases	62	56	
Age (years)	54.8033±12.8813	50.1304±10.2775	0.0457
Differentiation			
Well	0	0	
Moderate	7	3	0.4899
Poor	32	29	
Unknown	23	24	
ER status			
Negative	2	36	<0.0001
Positive	60	20	
Molecular subtype			
Luminal A	16	0	
Luminal B	27	7	
HER2	6	14	P < 0.0001
Basal-like	1	20	
Normal breast-like	7	3	
	Dataset NKI		
	29	33	
Age (years)	42.2069±5.8394	42.3636±7.2233	0.9261
Tumor size			
<2cm	11	12	0.6958
≥2cm	22	17	
Differentiation			
Well	4	0	
Moderate	9	8	0.0263
Poor	16	25	
ER status			
Negative	0	21	<0.0001
Positive	29	12	
Molecular subtype			
Luminal A	8	1	
Luminal B	14	1	
HER2	2	8	<0.0001
Basal-like	1	14	
Normal breast-like	2	3	
	Dataset TRANSBIG		
Number of cases	32	30	
Age (years)	47.6875±7.5752	45.7000±8.8167	0.3440
Tumor size			
<2cm	14	14	0.9803
≥2cm	18	16	
Differentiation			
Well	6	1	
Moderate	21	7	<0.0001
Poor	5	22	
ER status			
Negative	4	23	<0.0001
Positive	28	7	
Molecular subtype			
Luminal A	5	0	
Luminal B	12	3	
HER2	3	8	<0.0001
Basal-like	1	18	
Normal breast-like	9	1	

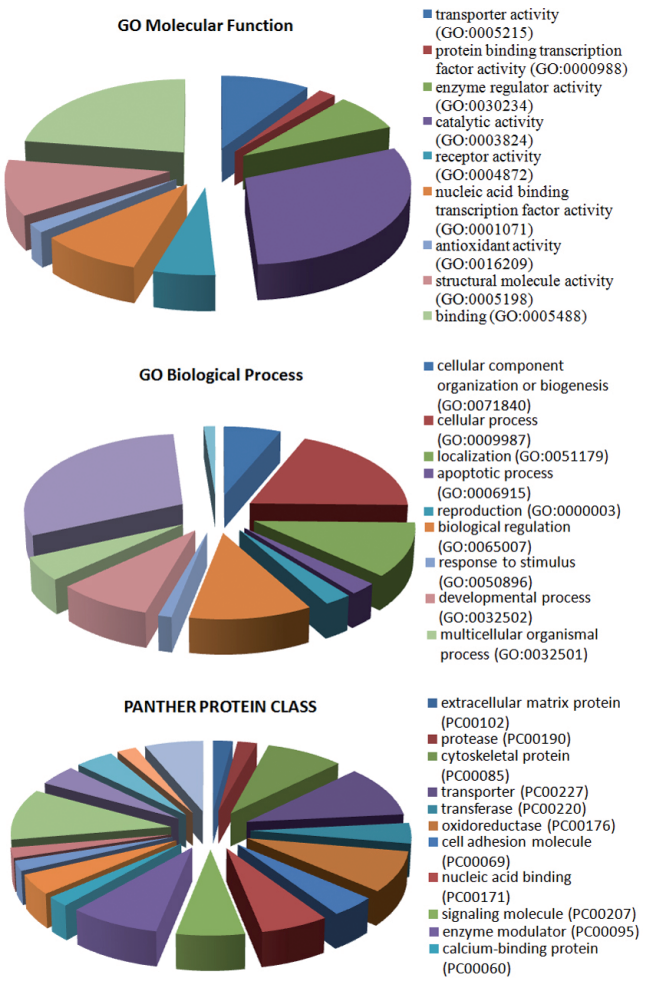

Figure 3. Gene ontology and protein class analysis of 39-gene signature by PANTHER software. Many of these genes have been found to be involved in metastasis-related processes.

class analyses revealed that many of these genes have been found to be involved in metastasis-related processes such as adhesion and apoptosis. And this gene signature also covered 9 confirmed genes which play critical roles in the migration, invasion of cancer cells including 6 positive regulators (WWTR1, PRDX4, CCNA2, DLGAP5, CDCA3 and CDH3) and 3 negative regulators (MAPT, MST1, and IQGAP2). In particular, CCNA2 gene has been confirmed to be a significant predictive power for DMFS in breast cancer [18]. Therefore, it is of importance to further explore the biological and therapeutic significance of this 39-gene signature during cancer metastasis.

In conclusion, in this study, we proposed a novel 39-gene signature for early occurrence of distant metastasis in node-positive and metastatic breast cancer patients, and highlighted its potential clinical implication for treatment decisions.

Acknowledgements: This study was supported by grants from National Science Foundation of China (81273282, 81202353, 81371786, 81302579), Changhai Hospital (CH125530300), Grant of Nanjing District (12MA056).

References

[1] CARDOSO F, CASTIGLIONE M. Locally recurrent or metastatic breast cancer: ESMO clinical recommendations for diagnosis, treatment and follow-up. Ann Oncol. 2009; 20 Suppl 4: 15–18. http://dx.doi.org/10.1093/annonc/mdp115

[2] CADOO KA, FORNIER MN, MORRIS PG. Biological subtypes of breast cancer: current concepts and implications for recurrence patterns. Q J Nucl Med Mol Imaging. 2013; 57: 312–321.

[3] STADLER ZK, COME SE. Review of gene-expression profiling and its clinical use in breast cancer. Crit Rev Oncol Hematol. 2009; 69: 1–11. http://dx.doi.org/10.1016/j.critrevonc.2008.05.004

[4] KELLY CM, WARNER E, TSOI DT, VERMA S, PRITCHARD KI. Review of the clinical studies using the 21-gene assay. Oncologist. 2010; 15: 447–456. http://dx.doi.org/10.1634/theoncologist.2009-0277

[5] SLODKOWSKA EA, ROSS JS. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn. 2009; 9: 417–422. http://dx.doi.org/10.1586/erm.09.32

[6] MITTEMPERGHER L, SAGHATCHIAN M, WOLF DM, MICHIELS S, CANISIUS S, et al. A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. Mol Oncol. 2013; 7: 987–999. http://dx.doi.org/10.1016/j.molonc.2013.07.006

[7] KIMBUNG S, KOVACS A, BENDAHL PO, MALMSTROM P, FERNO M, et al. Claudin-2 is an independent negative prognostic factor in breast cancer and specifically predicts early liver recurrences. Mol Oncol. 2014; 8: 119–128. http://dx.doi.org/10.1016/j.molonc.2013.10.002

[8] WANG Y, KLIJN JG, ZHANG Y, SIEUWERTS AM, LOOK MP, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365: 671–679. http://dx.doi.org/10.1016/S0140-6736-(05)70933-8

[9] MINN AJ, GUPTA GP, PADUA D, BOS P, NGUYEN DX, et al. Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci U S A. 2007; 104: 6740–5. http://dx.doi.org/10.1073/pnas.0701138104

[10] VAN ,T VEER LJ, DAI H, VAN DE VIJVER MJ, HE YD, HART AA, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415: 530–536. http://dx.doi.org/10.1038/415530a

[11] VAN DE VIJVER MJ, HE YD, VAN'T VEER LJ, DAI H, HART AA, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002; 347: 1999–2009. http://dx.doi.org/10.1056/NEJMoa021967

[12] DESMEDT C, PIETTE F, LOI S, WANG Y, LALLEMAND F, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res. 2007; 13: 3207–3214. http://dx.doi.org/10.1158/1078-0432.CCR-06-2765

[13] SIMON R, LAM A, LI M, NGAN M, MENENZES S, et al. Analysis of gene expression data using BRB-ArrayTools. Cancer informatics. 2007; 3: 11.

[14] MI H, MURUGANUJAN A, THOMAS PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013; 41: D377–386. http://dx.doi.org/10.1093/nar/gks1118

[15] HEAGERTY PJ, ZHENG Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005; 61: 92–105. http://dx.doi.org/10.1111/j.0006-341X.2005.030814.x

[16] TUTT A, WANG A, ROWLAND C, GILLETT C, LAU K, et al. Risk estimation of distant metastasis in node-negative, estrogen receptor-positive breast cancer patients using an RT-PCR based prognostic expression signature. BMC Cancer. 2008; 8: 339. http://dx.doi.org/10.1186/1471-2407-8-339

[17] DRUKKER CA, VAN TINTEREN H, SCHMIDT MK, RUTGERS EJ, BERNARDS R, et al. Long-term impact of the 70-gene signature on breast cancer outcome. Breast Cancer Res Treat. 2014; 143: 587–592. http://dx.doi.org/10.1007/s10549-013-2831-4

[18] GAO T, HAN Y, YU L, AO S, LI Z, et al. CCNA2 is a prognostic biomarker for ER+ breast cancer and tamoxifen resistance. PLoS One. 2014; 9: e91771. http://dx.doi.org/10.1371/journal.pone.0091771