Journal of Breast Cancer

J Breast Cancer 2018 December; 21(4): 363-370 https://doi.org/10.4048/jbc.2018.21.e56

ORIGINAL ARTICLE

Identification and Validation of Circulating MicroRNA Signatures for Breast Cancer Early Detection Based on Large Scale Tissue-Derived Data

Xiaokang Yu, Jinsheng Liang, Jiarui Xu[1], Xingsong Li, Shan Xing[2], Huilan Li[2], Wanli Liu[2], Dongdong Liu[1], Jianhua Xu[1], Lizhen Huang, Hongli Du

School of Biology and Biological Engineering, South China University of Technology, Guangzhou; [1]Department of Laboratory Science, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou; [2]Department of Laboratory Medicine, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China

Purpose: Breast cancer is the most commonly occurring cancer among women worldwide, and therefore, improved approaches for its early detection are urgently needed. As microRNAs (miRNAs) are increasingly recognized as critical regulators in tumorigenesis and possess excellent stability in plasma, this study focused on using miRNAs to develop a method for identifying noninvasive biomarkers. **Methods:** To discover critical candidates, differential expression analysis was performed on tissue-originated miRNA profiles of 409 early breast cancer patients and 87 healthy controls from The Cancer Genome Atlas database. We selected candidates from the differentially expressed miRNAs and then evaluated every possible molecular signature formed by the candidates. The best signature was validated in independent serum samples from 113 early breast cancer patients and 47 healthy controls using reverse transcription quantitative real-time polymerase chain reaction. **Results:** The miRNA candidates in our method were revealed to be associated with breast cancer according to previous studies and showed potential as useful biomarkers. When validated in independent serum samples, the area under curve of the final miRNA signature (miR-21-3p, miR-21-5p, and miR-99a-5p) was 0.895. Diagnostic sensitivity and specificity were 97.9% and 73.5%, respectively. **Conclusion:** The present study established a novel and effective method to identify biomarkers for early breast cancer. And the method, is also suitable for other cancer types. Furthermore, a combination of three miRNAs was identified as a prospective biomarker for breast cancer early detection.

Key Words: *Breast neoplasms, Data mining, Early detection of cancer, MicroRNAs, Tumor biomarkers*

INTRODUCTION

Breast cancer is the most prevalent type of cancer among women around the world and has the highest fatality rate [1]. The most reliable detection methods are mammography and core needle biopsy. However, these methods are not sensitive or comfortable enough for women to select as routine examinations. For liquid biopsy, existing markers such as carcinoembryonic antigen or carbohydrate antigen 153 are not recommended for screening or diagnosis of breast cancer because of their low sensitivity in early detection [2]. Therefore, a convenient, effective method for early detection is urgently needed.

MicroRNAs (miRNAs) are noncoding RNAs approximately 22 nucleotides in length. Recent evidence [3,4] demonstrates that miRNAs could be utilized as biomarkers for different cancers. They widely regulate life processes including cell proliferation, differentiation, apoptosis, and metabolism through a complicated network of the miRNAs and their target genes. Mitchell et al. [5] identified miR-16, let-7a, and other miRNAs from plasma RNA isolated from healthy volunteers and found that these miRNAs remain intact and are safe from endogenous RNase. Other studies [6,7] identified specific expression patterns of serum miRNAs for lung cancer, colorectal cancer, and diabetes, and found that the levels of miRNAs in unfrozen serum remained stable over a 4 hours period at room

Correspondence to: Hongli Du
https://orcid.org/0000-0002-6419-0160
School of Biology and Biological Engineering, South China University of Technology, 382 Zhonghuan Road East, Higher Education Mega Centre, Panyu District, Guangzhou 510006, China
Tel: +86-02039380667, Fax: +86-02039380667, E-mail: hldu@scut.edu.cn

This work was supported by the National Key R&D Program of China (2018YFC0910201), the Science and Technology Planning Project of Guangzhou (201704020176, 201508020040, and 201510010044), the Fundamental Research Funds for the Central Universities (2015ZZ125), and Science and Technology Planning Project of Guangdong Province, China (2014A020212274).

Received: August 6, 2018 **Accepted:** November 9, 2018

temperature and are minimally affected by twice freezing and re-thawing. These result suggested there are stable and detectable miRNAs in serum that can serve as nonintrusive biomarkers in tumor diagnosis.

Wang et al. [8] suggested tumor cells could communicate with each other by exporting specific miRNAs. miRNAs with consistent expression status in tumor tissue and serum samples are likely to be message molecules released from tumor tissue to the circulatory system, which can be used as indicators for tumor detection. Previous studies [9] suggested there are miRNAs whose expression deregulation status was consistent between tumor tissue samples and serum samples. Thus, effective miRNA signatures identified from tissue profiles may also be prospective serum miRNA signatures.

Large scale tumor data from cancer databases have been used widely in research to obtain reliable evidence. The Cancer Genome Atlas (TCGA) is one of these preeminent databases, which contains information on over 1,000 breast cancer cases, including clinicopathological information and transcriptomic data. The present research is based on tissue-originated public miRNA expression profiles from TCGA, aiming to establish a novel method for identifying effective miRNA signatures which can detect early breast cancer in patients.

METHODS

Data and serum samples

The present study consists of two stages: signature discovery stage and signature validation stage. Analysis design and strategies are summarized in Figure 1.

In the discovery stage, the sequencing data normalized as reads per million (RPM) of mature miRNAs was downloaded from the breast cancer project of TCGA (containing information from 99 stage I patients, 310 stage II patients, and 87 healthy controls) through Broad GDAC Firehose data portal (http://gdac.broadinstitute.org/). These data were used in method establishment and identifying prospective tissue-based signatures.

In the validation stage, 113 breast cancer patients in early stages and 47 healthy controls from Sun Yat-sen University Cancer Center were recruited as the validation cohort. Serum of the subjects was collected based on the following criteria: (1) collected at diagnosis before receiving any surgery or treatment; (2) breast cancer serum samples were collected from patients diagnosed as having early breast cancer, including invasive breast cancer at stage I, stage IIA, or stage IIB; (3) control serum samples were collected from healthy volunteers without any history of cancer or inflammatory conditions currently.

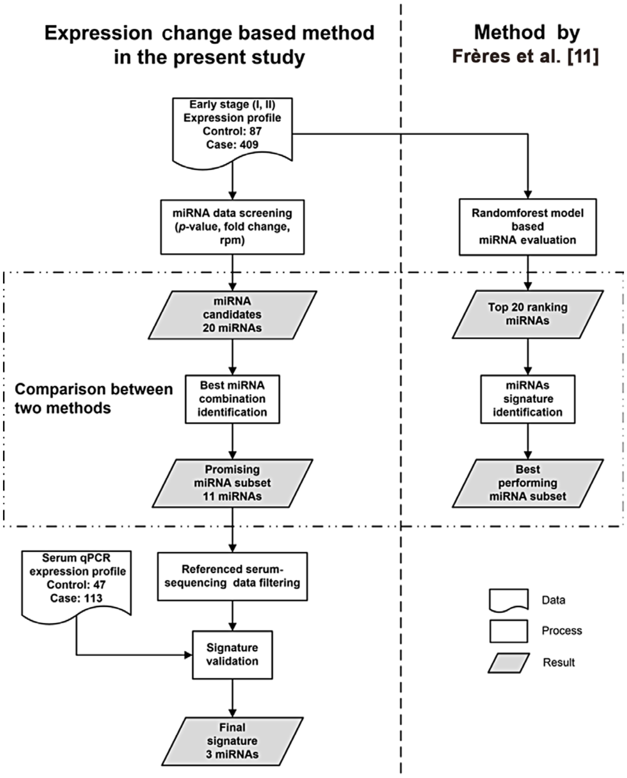

Expression change based method in the present study
Method by Frères et al. [11]
Early stage (I, II) Expression profile Control: 87 Case: 409
miRNA data screening (p-value, fold change, rpm)
Randomforest model based miRNA evaluation
miRNA candidates 20 miRNAs
Top 20 ranking miRNAs
Comparison between two methods
Best miRNA combination identification
miRNAs signature identification
Promising miRNA subset 11 miRNAs
Best performing miRNA subset
Serum qPCR expression profile Control: 47 Case: 113
Referenced serum-sequencing data filtering
Signature validation
Final signature 3 miRNAs
Data
Process
Result

Figure 1. Flow chart of the analysis design in the present study. The expression change-based method pipeline was described on the left and the random forest algorithm-based method on the right. Tissue profiles were used in discovery stage while independent serum profiles were used in validation stage. Intermediate results of the expression change-based method were compared with those of the random forest algorithm-based method [11] for evaluation purpose.
miRNA = microRNA; qPCR = quantitative real-time polymerase chain reaction.

Serum samples were prepared by retaining the supernatant after double centrifugation of blood samples at 4°C (10 minutes at 3,000 rpm and 10 minutes at 13,400 rpm) and stored at –80°C immediately until use. In the present study, the tumor stage was classified according to the revised American Joint Committee on Cancer tumor-node-metastasis (TNM) classification. Histopathological information was obtained by reviewing medical records. This study followed the principles outlined in the Helsinki Declaration and was reviewed and approved by the Institutional Review Board and Ethics Committee of Sun Yat-sen University Cancer Center (GZR2017-186). Written informed consent was obtained from all of the enrolled participants.

The characteristics of patients enrolled in the present study were summarized in Table 1.

Establishment of expression change based method for signature identification

To concisely identify biomarkers of early stage breast can-

Table 1. Clinical characteristics of breast cancer patients and healthy controls

Characteristic	Discovery stage		Validation stage	
	Cancer tissue (n=409) No. (%)	Healthy control (n=87) No. (%)	Patient serum (n=113) No. (%)	Healthy control (n=47) No. (%)
Age (yr)*	58 (26–90)	56 (27–85)	49 (29–73)	51 (41–64)
Clinical TNM stage				
I	99 (24.2)	-	32 (28.3)	-
II	310 (75.8)	-	81 (71.7)	-
ER status				
Negative	89 (21.8)	-	26 (23.0)	-
Positive	298 (72.8)	-	83 (73.5)	-
Unknown	22 (5.4)	-	4 (3.5)	-
PR status				
Negative	126 (30.8)	-	37 (32.7)	-
Positive	261 (63.9)	-	71 (62.8)	-
Unknown	22 (5.4)	-	5 (4.4)	-
HER2 status				
Negative	2 (0.5)	-	41 (36.3)	-
Positive	188 (46.0)	-	67 (59.3)	-
Unknown	219 (53.5)	-	5 (4.4)	-
Ki-67 status				
Negative (<14%)	NA	-	13 (11.5)	-
Positive (≥14%)	NA	-	95 (84.1)	-
Unknown	NA	-	5 (4.4)	-
Histologic subtype				
IDC	274 (67.0)	-	98 (86.7)	-
ILC	89 (21.8)	-	3 (2.7)	-
Other	46 (11.2)	-	12 (10.6)	-

ER=estrogen receptor; PR=progesterone receptor; HER2=human epidermal growth factor receptor 2; NA=not assessed; IDC=invasive ductal carcinoma; ILC=invasive lobular carcinoma.
*Median (range).

cer, only data of early stage patients were selected from the complete dataset. Missing expression values were replaced by the stage average value, and miRNAs with data missing rate >10% were eliminated.

In the discovery stage, Student t-test was performed between the expression profiles of early stage patients and healthy controls. *p*-values were adjusted using the Benjamini-Hochberg procedure. Adjusted *p*-values <0.05 were considered statistically significant. miRNAs with high expression levels and high expression fold change were preferred to insure detectability and reliability. Consequently, only those miRNAs with control average rpm >100 and absolute fold change >3.5 were further evaluated. Statistical analysis was performed with R software environment (version 3.3.1, 2016) using bayesreg.R script developed by Cyber-T workspace [10].

To identify the most promising prospective signature, all possible combinations of eligible miRNAs in the screening process were considered as potential signatures, and the effectiveness of each signature was measured by Youden Index (specificity+sensitivity−1). The number of half miRNAs in signature was chosen as diagnostic rule. The threshold of normal expression of each miRNA was defined as a certain expression value in healthy controls, which maximized the Youden Index when classifying samples using this miRNA alone (called balanced value in the text).

Comparison with random forest algorithm based method

For evaluation purposes, the expression change (EC)-based method in the present study was compared with a well-designed method, which makes use of the random forest algorithm (RF) [11].

The best results from the identification process were compared with the best signatures identified by the RF-based method using the same data from TCGA. A subset of miRNAs of the same number as the candidates in the EC-based method was selected from the importance matrix of the RF-based method. Secondly, all possible combinations defined from this miRNA subset were considered as potential classifiers, the specificity and sensitivity of which were obtained using the diagnostic rule mentioned in the reference.

The comparison was performed using the following aspects: miRNA candidates, performance of combinations consist of different number miRNA candidates, and characteristics of the top 25 combinations.

Validation of the final signature

Further validation of the best performing signature was conducted with serum data of an independent cohort obtained through reverse transcription quantitative real-time polymerase chain reaction (RT-qPCR). The essential Minimum Information for Publication of RT-qPCR Experiments guidelines were followed during specimen preparation.

In the circulating miRNAs extraction step, according to the manufacturer's instructions of the miRNeasy Serum/Plasma Kit (Qiagen, Duesseldorf, Germany), 3.5 μL *Caenorhabditis elegans* miR-39 miRNA mimic (1.6×10^7 copies/μL; Qiagen) was added to each 200 μL serum sample as a normalization control before miRNAs were extracted. Reverse transcription was performed using mir-X miRNA First-Strand Synthesis Kit (Takara, Kusatsu, Japan). RT-qPCR was performed with PowerUp^{TM} SYBR^{TM} Green Master Mix (Thermo Fisher, Waltham, USA) according to the manufacturer's instructions. Most of primers used in qPCR reaction were designed and synthesized by Tiangen Biotech, Co., Ltd. (Beijing, China) according to miRNA sequences. Detailed sequences of the commercial primers are classified due to privacy policy, but the sequence of one published primer for miR-10b-5p we used was ACACTCCAGCT-

GGGTACCCTGTAGAA [12]. RT-qPCR was performed on Roche LightCycler® 480 System (Roche, Basel, Switzerland).

Data normalization was conducted using the $2^{-\Delta\Delta Cq}$ method $\{\Delta\Delta Cq = [(Cq_{\text{cancer}} - Cq_{\text{reference gene}})\ \text{-mean}(Cq_{\text{control}} - Cq_{\text{reference gene}})]\}$ for each sample to obtain a relative expression value. Diagnostic rule was the same as described in the EC-based method. Receiver operating characteristic (ROC) curve was obtained to evaluate the performance of the final signature in the validation cohort.

RESULTS

Comparison between different normal expression range defining methods

In addition to the balanced value that maximized the Youden Index of each miRNA, the 80th percentile (or 20th, depending on whether the expression level of controls was lower than that of patients) and the 90th percentile (or 10th) of expression values in healthy controls were also taken into consideration when defining the normal expression threshold. A suitable threshold should lead to the majority of the healthy controls having expression in the normal range. To compare these threshold defining methods, signature analysis was conducted using all three methods. Considering the best performance (highest specificity and acceptable sensitivity), we chose the balanced value as the normal range threshold in this study. The comparison is shown in Figure 2.

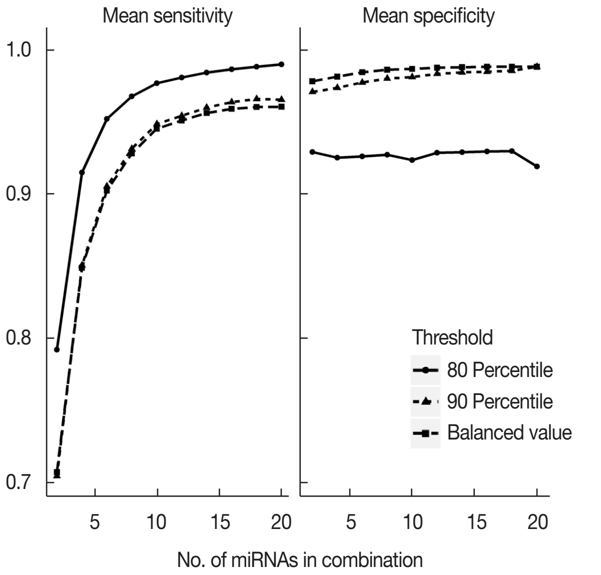


Figure 2. Comparison between three threshold defining methods. Signatures from the 20 microRNA (miRNA) candidates in the expression change-based method were grouped by the number of miRNAs, and the mean sensitivity and specificity were calculated respectively.

Tissue-based miRNA candidate selection and comparison with random forest algorithm based method

After tissue profiling data preprocessing, 402 miRNAs were eligible for screening. Most of these were eliminated in the screening process, and 20 miRNAs were ultimately left in the candidate list of the EC-based method. When comparing these candidates to the 20 most important miRNAs from the RF-based method, we found eight common miRNAs. Details are shown in Supplementary Figure 1 (available online).

Next, we compared the performance of the signatures from the 20 eligible candidates in the EC-based method and signatures from the 20 most important miRNAs in the RF-based method. Average performance of signatures composed of equal numbers of miRNAs were compared, and results are shown in Supplementary Figure 2 (available online). Additionally, the top 25 combinations in both methods were compared. The average of miRNA number per combination, specificity, and sensitivity were 6.920, 0.989, and 0.990 for the EC-based method while the values were 6.640, 0.985, and 0.960 for the RF-based method. Specifically, the best combination obtained from the RF-based method consisted of seven miRNAs, of which the specificity and sensitivity were 0.989 and 0.980, respectively. However, the best combination identified using the EC-based method consisted of nine miRNAs, of which the specificity and sensitivity were 0.989 and 0.993 (Supplementary Tables 1 and 2, available online).

Identification of the best combination

Among all the signatures from 20 eligible candidates in the EC-based method, two tied for the best performance. Both were composed of nine miRNAs, seven of which were in common. In other words, there were 11 miRNAs enrolled in the two best combinations: miR-183-5p, miR-182-5p, miR-141-3p, miR-21-5p, miR-21-3p, miR-10b-5p, miR-99a-5p, miR-378a-5p, miR-144-5p, miR-451a, and miR-486-5p. Result of Student t-test and breast cancer related references of these miRNAs were listed in Table 2 [13-22].

Validation in serum samples

To further narrow the number of miRNAs in the validation stage, the RPM data of a previous study [23], which had performed small RNA-sequencing on serum samples from breast cancer patients, were taken into consideration. Four miRNAs in the best signature (miR-21-5p, miR-21-3p, miR-99a-5p, miR-10b-5p) with great fold changes and high RPM values in serum sequencing data and similar deregulation status as in the tissue samples of TCGA dataset (Table 3) were selected into validation stage. However, although we used two primers designed by Tiangen Biotech Company and one published

Table 2. The expression status of 11 miRNAs of the best combinations in the present study and in other researches

miRNA	Data form TCGA		Data from other article			
	Fold change	*p*-value	Fold change	*p*-value	Application	Reference
miR-183-5p	+8.26	<0.01	+3.22	<0.01	Prognosis	[13,14]
miR-182-5p	+5.61	<0.01	+7.75	<0.01	Prognosis	[14,15]
miR-141-3p	+5.58	<0.01	NA	NA	NA	NA
miR-21-5p	+4.69	<0.01	+3.2	<0.01	Diagnosis (no stage information, AUC=0.607)	[16]
miR-21-3p	+4.35	<0.01	+	<0.05		[17]
miR-10b-5p	–3.65	<0.01	–	<0.05	Diagnosis (stage I, II, III, AUC=0.950)	[17,18]
miR-99a-5p	–3.80	<0.01	NA	NA	Prognosis	[19]
miR-378a-5p	–5.29	<0.01	NA	NA	NA	NA
miR-144-5p	–8.77	<0.01	–2.50	<0.01	Prognosis	[20,21]
miR-451a	–8.91	<0.01	–	<0.05	NA	[17]
miR-486-5p	–22.18	<0.01	–	< 0.05	Metastasis detection	[21,22]

The plus sign or minus sign before the fold change values indicated the deregulation status of miRNAs. Some values are absent because no specific value was listed in the corresponding reference. "+" represents upregulated while "–" represents downregulated. In the parentheses was stage information of patient cohorts in which the miRNA was applied as diagnostic marker.
miRNA=microRNA; TCGA=The Cancer Genome Atlas; NA=not assessed; AUC=area under curve.

Table 3. The tissue-based and serum-based expression status of the 11 best miRNAs obtained through EC-based method

miRNA	Tissue sequencing data (control [n=87]; early stage cancer [n=409])				Serum sequencing data (control [n=8]; breast cancer [n=8])		
	MeanC	MeanE	FC	*p*-value	MeanC	MeanE	FC
miR-183-5p	2,123.70	17,552.28	+8.26	<0.01	13,750.50	18,643.21	+1.36
miR-182-5p	8,575.70	48,118.67	+5.61	<0.01	7,319.85	9,721.51	+1.33
miR-141-3p	205.07	1,144.28	+5.58	<0.01	4,859.21	5,765.71	+1.19
miR-21-5p	53,402.99	250,410.52	+4.69	<0.01	249,722.74	457,962.70	+1.83
miR-21-3p	622.85	2,710.56	+4.35	<0.01	1,234.56	3,389.78	+2.75
miR-99a-5p	3,014.67	794.02	–3.80	<0.01	7,917.29	3,491.94	–2.27
miR-378a-5p	287.29	54.26	–5.29	<0.01	49.53	36.63	–1.35
miR-10b-5p	266,581.80	73,028.62	–3.65	<0.01	3,520.15	1,773.58	–1.98
miR-144-5p	519.16	59.18	–8.77	<0.01	21.03	34.62	+1.65
miR-451a	2,800.20	314.28	–8.91	<0.01	1,802.82	3,164.29	+1.76
miR-486-5p	1,774.80	80.02	–22.18	<0.01	693.11	996.39	+1.44

Average expression values of miRNAs in control group (MeanC) and patient group (MeanE) were both listed. The plus sign or minus sign before the fold change values indicated the deregulation status of miRNAs. "+" represents upregulated while "–" represents downregulated.
miRNA=microRNA; EC=expression change; FC=fold change.

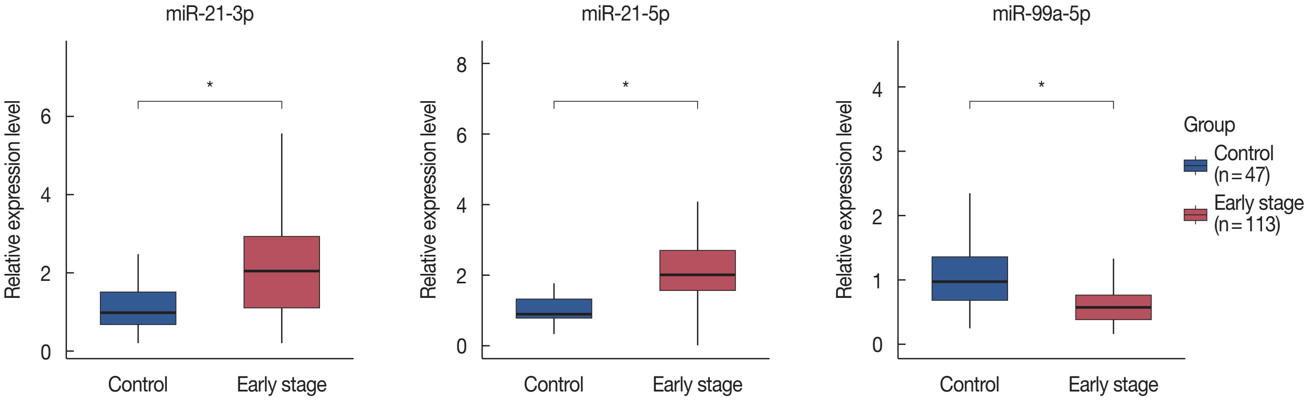


Figure 3. Expression levels of the three final microRNAs (miRNAs) in serum samples. The relative expression level of miRNAs was normalized to $2^{-\Delta\Delta Cq}$ value and two-sided Student t-test was used to compare miRNA expression level.
**p*-value <0.01.

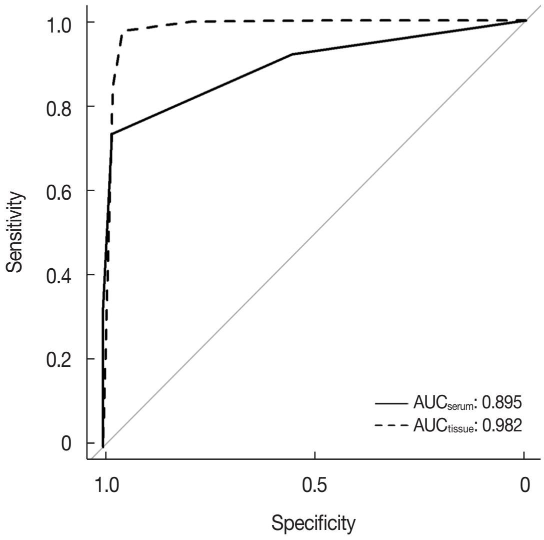


Figure 4. Receiver operating characteristic curve of the final signature based on tissue data and independent serum data. The number of normal expressed microRNAs in signature was used as diagnostic index in this analysis.
AUC = area under curve.

primer for miR-10b-5p, we were unable to obtain specific and stable amplification results of this miRNA. Therefore, miR-10b-5p was excluded from the final signature temporarily.

According to the qPCR results, the Cq values of control miRNA among all the samples were stable (22.59 ± 0.77 cycles) and were in a normal range confirmed by the provider. The relative expression levels of the three selected miRNAs remained significantly different between patients and healthy controls. miR-21-5p and miR-21-3p were upregulated while miR-99a-5p was downregulated, which was consistent with the tissue data (Figure 3). Similar expression status of these miRNAs was also found in previous studies [19,24]. An area under curve (AUC) of 0.895, diagnostic specificity of 73.5%, and sensitivity of 97.9% were obtained when the 3-miR signature was tested in serum samples, whereas an AUC of 0.982, specificity of 97.6%, and sensitivity of 95.4% were obtained when tested in tissue samples. The high sensitivity and considerable specificity suggested the 3-miR signature was well validated in serum data. ROC is shown in Figure 4.

DISCUSSION

Instead of collecting samples and obtaining data from patients of all stages, the present study made use of a large set of early stage breast cancer miRNA expression profiles from the TCGA database. Since TCGA is one of the largest public molecular data sources which also contains detailed clinical information, such as the TNM stage of patients, this could lead to more reliable and valuable results, particularly in cancer early detection. Moreover, with normalized data from a larger cohort, as well as more detailed and uniform clinical information, we can conduct further studies more easily, such as identifying signatures for cancer subtypes, conducting survival correlation analysis, and so on.

Although serum miRNA profiles were used to identify diagnostic signatures in many other studies [3,11,25,26], the present study was designed to identify signatures based on tissue profiles and validate them with serum profiles. Using serum profiles to determine cancer can be challenging; since blood circulates the entire body, factors such as exercise and diet can influence circulating miRNAs. In contrast, tissue miRNA profiles described what exactly happened in the tumor and were minimally affected by irrelevant factors. Though the relationship between expression patterns of tissue miRNA and circulating miRNA is still unclear [12], previously studies [25,26] did find miRNAs whose expression deregulation statuses were consistent between tissue samples and serum samples. These findings suggested the expression patterns of circulating miRNAs in cancer patients were somehow influenced by the expression change of miRNAs in tumor tissues. Thus, effective miRNA signatures selected from tissue profiles could also be useful serum miRNA signatures.

Another concern is the expression abundance and the expression fold change of miRNAs in signatures. miRNAs with low expression or little fold change were selected as diagnostic signatures in previous studies [11,25,26] but are impractical in application. In contrast, great fold change and high expression of miRNAs were preferred in the present study because these miRNAs were more likely to be the key factors in oncogenesis and more detectable in the serum samples as biomarkers.

After miR-21-3p, miR-21-5p, and miR-99a-5p were chosen as the final signature, we further investigated the expression levels of these miRNAs in every clinical stage using tissue-derived data from TCGA. Compared to healthy controls, the fold-change values of stage I group to stage III group for miR-21-5p were +4.67, +4.69, +4.67, and +4.01, +4.45, +4.08 for miR-21-3p, and −3.41, −3.94, −3.51 for miR-99a-5p. Adjusted *p*-values for these results were all less than 0.01. Results of the stage IV group were invalid since the sample size was too small. According to previous studies, increased miR-21, in response to transforming growth factor β1 signaling, is associated with tumor invasion and chemoresistance *in vitro* [27]. Specifically, overexpression of miR-21-3p could strongly augment L1 cell adhesion molecular expression in renal, endometrial, and ovarian carcinoma-derived cell lines, which promotes cell motility, invasion, chemoresistance and metastasis formation [28]. Upregulated miR-99a could suppress the proliferation, migration and invasion of the MDA-MB-231 breast cancer cells *in vitro* and inhibited the growth of xeno-transplant tumor *in vivo* [29]. Although few serum-derived data of these three miRNAs was found, the expression levels of serum circulating miR-21 in different histological tumor grades from a recent study [30] matched our results. These findings suggest the three final miRNAs play an important role in the development of breast cancer and could serve as useful biomarkers for early detection of breast cancer.

It is worth mentioning that majority of the tissue samples used in the present study were donated by Hispanic patients, while all serum samples were from Chinese patients. In the future, more uniform study cohorts should be used when possible. A previous study [25] identified a panel of nine miRNAs (miR-15a, miR-18a, miR-107, miR-133a, miR-139-5p, miR-143, miR-145, miR-365, and miR-425) that can distinguish early stage patients from healthy controls, achieving an AUC of 0.665, sensitivity of 83.3%, and specificity of 41.2% in a validation cohort. Another study [18] showed that a combination of serum miRNAs (miR-145, miR-155, and miR-382) can achieve an AUC of 0.988, sensitivity of 97.6%, and specificity of 100% in cancer detection but this was not validated in an independent cohort. A recent study [26] found a combination of five miRNAs (miR-1246, miR-1307-3p, miR-4634, miR-6861-5p, and miR-6875-5p) from early stage serum samples using microarray profiling, which achieved an AUC of 0.971, sensitivity of 97.3%, and specificity of 82.9%. However, four miRNAs in the combination were unable to be validated by qPCR. All of these study had certain limitations, such as the number of samples in the validation cohort was small or the sensitivity, specificity, and accuracy were not satisfying or not well validated.

In conclusion, the present study established a novel and effective method to identify miRNA signatures for breast cancer early detection based on large scale data from the TCGA database, which could be applied to other cancer types in the future. Furthermore, a prospective biomarker combination of three miRNA (miR-21-5p, miR-21-3p, and miR-99a-5p) for early detection of breast cancer was identified using this method and verified using Chinese clinical serum samples.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65:87-108.
2. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Oncol Pract 2007;3:336-9.
3. Lin XJ, Chong Y, Guo ZW, Xie C, Yang XJ, Zhang Q, et al. A serum microRNA classifier for early detection of hepatocellular carcinoma: a multicentre, retrospective, longitudinal biomarker identification study with a nested case-control study. Lancet Oncol 2015;16:804-15.
4. Du M, Shi D, Yuan L, Li P, Chu H, Qin C, et al. Circulating miR-497 and miR-663b in plasma are potential novel biomarkers for bladder cancer. Sci Rep 2015;5:10437.
5. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A 2008;105:10513-8.
6. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res 2008;18:997-1006.
7. Gilad S, Meiri E, Yogev Y, Benjamin S, Lebanony D, Yerushalmi N, et al. Serum microRNAs are promising novel biomarkers. PLoS One 2008;3:e3148.
8. Wang K, Zhang S, Weber J, Baxter D, Galas DJ. Export of microRNAs and microRNA-protective protein by mammalian cells. Nucleic Acids Res 2010;38:7248-59.
9. Pigati L, Yaddanapudi SC, Iyengar R, Kim DJ, Hearn SA, Danforth D, et al. Selective release of microRNA species from normal and malignant mammary epithelial cells. PLoS One 2010;5:e13515.
10. Kayala MA, Baldi P. Cyber-T web server: differential analysis of high-throughput data. Nucleic Acids Res 2012;40:W553-9.
11. Frères P, Wenric S, Boukerroucha M, Fasquelle C, Thiry J, Bovy N, et al. Circulating microRNA-based screening tool for breast cancer. Oncotarget 2016;7:5416-28.
12. Allaya N, Khabir A, Sallemi-Boudawara T, Sellami N, Daoud J, Ghorbel A, et al. Over-expression of miR-10b in NPC patients: correlation with LMP1 and Twist1. Tumour Biol 2015;36:3807-14.
13. Chang YY, Kuo WH, Hung JH, Lee CY, Lee YH, Chang YC, et al. Deregulated microRNAs in triple-negative breast cancer revealed by deep sequencing. Mol Cancer 2015;14:36.
14. Song C, Zhang L, Wang J, Huang Z, Li X, Wu M, et al. High expression of microRNA-183/182/96 cluster as a prognostic biomarker for breast cancer. Sci Rep 2016;6:24502.
15. Calvano Filho CM, Calvano-Mendes DC, Carvalho KC, Maciel GA, Ricci MD, Torres AP, et al. Triple-negative and luminal A breast tumors: differential expression of miR-18a-5p, miR-17-5p, and miR-20a-5p. Tumour Biol 2014;35:7733-41.
16. Matamala N, Vargas MT, González-Cámpora R, Miñambres R, Arias JI, Menéndez P, et al. Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection. Clin Chem 2015;61:1098-106.
17. Ouyang M, Li Y, Ye S, Ma J, Lu L, Lv W, et al. MicroRNA profiling implies new markers of chemoresistance of triple-negative breast cancer. PLoS One 2014;9:e96228.
18. Mar-Aguilar F, Mendoza-Ramírez JA, Malagón-Santiago I, Espino-Silva PK, Santuario-Facio SK, Ruiz-Flores P, et al. Serum circulating microRNA profiling for identification of potential breast cancer biomarkers. Dis Markers 2013;34:163-9.

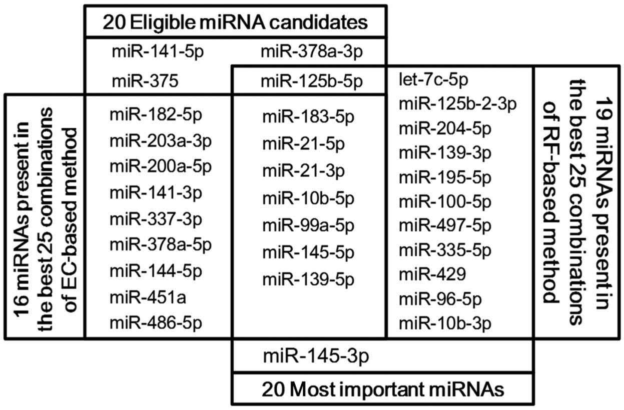
19. Li J, Song ZJ, Wang YY, Yin Y, Liu Y, Nan X. Low levels of serum miR-99a is a predictor of poor prognosis in breast cancer. Genet Mol Res 2016; 15(3):gmr8338.
20. Pan Y, Zhang J, Fu H, Shen L. miR-144 functions as a tumor suppressor in breast cancer through inhibiting ZEB1/2-mediated epithelial mesenchymal transition process. Onco Targets Ther 2016;9:6247-55.
21. Madhavan D, Peng C, Wallwiener M, Zucknick M, Nees J, Schott S, et al. Circulating miRNAs with prognostic value in metastatic breast cancer and for early detection of metastasis. Carcinogenesis 2016;37:461-70.
22. Zhang G, Liu Z, Cui G, Wang X, Yang Z. MicroRNA-486-5p targeting PIM-1 suppresses cell proliferation in breast cancer cells. Tumour Biol 2014;35:11137-45.
23. Zhu J, Zheng Z, Wang J, Sun J, Wang P, Cheng X, et al. Different miRNA expression profiles between human breast cancer tumors and serum. Front Genet 2014;5:149.
24. Si H, Sun X, Chen Y, Cao Y, Chen S, Wang H, et al. Circulating microRNA-92a and microRNA-21 as novel minimally invasive biomarkers for primary breast cancer. J Cancer Res Clin Oncol 2013;139:223-9.
25. Kodahl AR, Lyng MB, Binder H, Cold S, Gravgaard K, Knoop AS, et al. Novel circulating microRNA signature as a potential non-invasive multi-marker test in ER-positive early-stage breast cancer: a case control study. Mol Oncol 2014;8:874-83.
26. Shimomura A, Shiino S, Kawauchi J, Takizawa S, Sakamoto H, Matsuzaki J, et al. Novel combination of serum microRNA for detecting breast cancer in the early stage. Cancer Sci 2016;107:326-34.
27. Dai X, Fang M, Li S, Yan Y, Zhong Y, Du B. miR-21 is involved in transforming growth factor beta1-induced chemoresistance and invasion by targeting PTEN in breast cancer. Oncol Lett 2017;14:6929-36.
28. Doberstein K, Bretz NP, Schirmer U, Fiegl H, Blaheta R, Breunig C, et al. miR-21-3p is a positive regulator of L1CAM in several human carcinomas. Cancer Lett 2014;354:455-66.
29. Xia M, Li H, Wang JJ, Zeng HJ, Wang SH. MiR-99a suppress proliferation, migration and invasion through regulating insulin-like growth factor 1 receptor in breast cancer. Eur Rev Med Pharmacol Sci 2016;20:1755-63.
30. Fan T, Mao Y, Sun Q, Liu F, Lin JS, Liu Y, et al. Branched rolling circle amplification method for measuring serum circulating microRNA levels for early breast cancer detection. Cancer Sci 2018;109:2897-906.

Supplementary Table 1. Top 25 combinations obtained from the expression change-based method

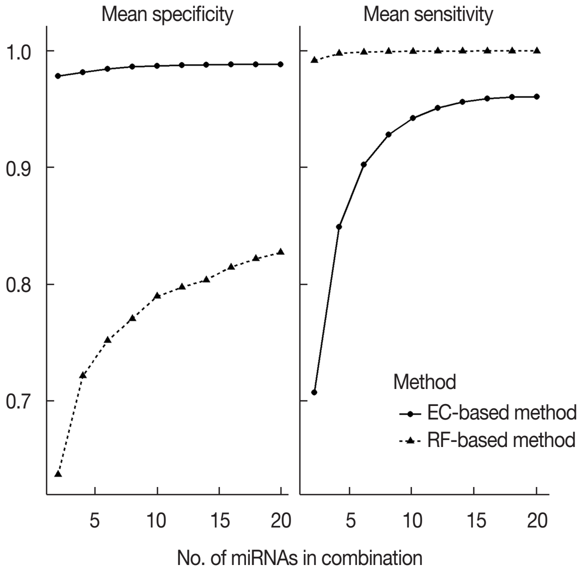
Combination	Specificity	Sensitivity
miR-183-5p,miR-141-3p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-451a,miR-486-5p	0.988506	0.992665
miR-182-5p,miR-141-3p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-144-5p,miR-451a	0.988506	0.992665
miR-183-5p,miR-141-3p,miR-21-5p,miR-10b-5p,miR-451a	0.988506	0.99022
miR-141-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-451a	0.988506	0.99022
miR-141-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-486-5p	0.988506	0.99022
miR-21-5p,miR-200a-5p,miR-10b-5p,miR-99a-5p,miR-451a	0.988506	0.99022
miR-21-5p,miR-200a-5p,miR-10b-5p,miR-99a-5p,miR-486-5p	0.988506	0.99022
miR-183-5p,miR-182-5p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-378a-5p	0.988506	0.99022
miR-183-5p,miR-141-3p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-183-5p,miR-141-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-451a	0.988506	0.99022
miR-183-5p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-139-5p	0.988506	0.99022
miR-182-5p,miR-141-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-144-5p	0.988506	0.99022
miR-182-5p,miR-141-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-451a	0.988506	0.99022
miR-182-5p,miR-141-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-182-5p,miR-203a-3p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-139-5p	0.988506	0.99022
miR-182-5p,miR-21-5p,miR-200a-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-451a	0.988506	0.99022
miR-141-3p,miR-203a-3p,miR-21-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-141-3p,miR-21-5p,miR-200a-5p,miR-21-3p,miR-10b-5p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-141-3p,miR-21-5p,miR-200a-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-141-3p,miR-21-5p,miR-200a-5p,miR-10b-5p,miR-99a-5p,miR-145-5p,miR-486-5p	0.988506	0.99022
miR-141-3p,miR-21-5p,miR-200a-5p,miR-10b-5p,miR-99a-5p,miR-139-5p,miR-451a	0.988506	0.99022
miR-203a-3p,miR-21-5p,miR-200a-5p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-21-5p,miR-200a-5p,miR-21-3p,miR-10b-5p,miR-337-3p,miR-378a-5p,miR-486-5p	0.988506	0.99022
miR-183-5p,miR-182-5p,miR-141-3p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-144-5p	0.988506	0.99022
miR-183-5p,miR-182-5p,miR-141-3p,miR-21-5p,miR-21-3p,miR-10b-5p,miR-99a-5p,miR-378a-5p,miR-451a	0.988506	0.99022
Average	0.988506	0.9904156

Supplementary Table 2. Top 25 combinations obtained from the random forest algorithm-based method

Combination	Specificity	Sensitivity
miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-195-5p,miR-429,miR-10b-3p	0.988506	0.98044
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-195-5p,miR-497-5p,miR-21-3p	0.988506	0.978215
miR-21-5p,miR-139-3p,miR-10b-5p,miR-183-5p,miR-195-5p	0.988506	0.977995
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-195-5p,miR-10b-3p	0.988506	0.97555
miR-21-5p,miR-139-3p,miR-10b-5p,miR-195-5p,miR-100-5p,miR-497-5p,miR-10b-3p	0.988506	0.97066
miR-21-5p,miR-139-3p,miR-10b-5p,miR-195-5p,miR-10b-3p	0.988506	0.968215
miR-139-5p,miR-21-5p,miR-10b-5p,miR-183-5p,miR-195-5p,miR-497-5p,miR-96-5p,miR-10b-3p	0.988506	0.968215
miR-139-5p,miR-21-5p,miR-10b-5p,miR-195-5p,miR-100-5p	0.977011	0.977995
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-183-5p,miR-195-5p,miR-497-5p	0.988506	0.963325
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-125b-5p,miR-195-5p	0.988506	0.958655
miR-139-5p,miR-21-5p,miR-204-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-195-5p	0.977011	0.968215
miR-139-5p,miR-21-5p,miR-10b-5p,miR-195-5p,miR-497-5p,miR-429	0.988506	0.95599
miR-21-5p,miR-139-3p,miR-10b-5p,miR-125b-5p,miR-195-5p,miR-497-5p,miR-21-3p,miR-10b-3p	0.988506	0.95599
miR-139-5p,miR-21-5p,miR-10b-5p,miR-195-5p,miR-497-5p,miR-21-3p	0.988506	0.95599
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-195-5p,miR-497-5p	0.977011	0.96577
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-195-5p,miR-497-5p,miR-335-5p,miR-10b-3p	0.988506	0.953545
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-125b-5p,miR-195-5p,miR-335-5p	0.977011	0.963325
miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-145-5p,miR-195-5p,miR-100-5p	0.977011	0.963325
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-195-5p,miR-125b-2-3p	0.977011	0.96088
miR-21-5p,miR-139-3p,miR-10b-5p,let-7c-5p,miR-183-5p,miR-195-5p	0.988506	0.948655
miR-139-5p,miR-21-5p,miR-10b-5p,miR-183-5p,miR-195-5p,miR-497-5p	0.988506	0.948655
miR-21-5p,miR-139-3p,miR-10b-5p,miR-195-5p,miR-497-5p,miR-10b-3p	0.988506	0.948655
miR-21-5p,miR-139-3p,miR-10b-5p,miR-99a-5p,miR-195-5p,miR-497-5p	0.988506	0.94132
miR-139-5p,miR-21-5p,miR-139-3p,miR-10b-5p,miR-125b-5p,miR-195-5p,miR-497-5p,miR-335-5p	0.977011	0.948655
miR-21-5p,miR-139-3p,miR-10b-5p,let-7c-5p,miR-195-5p,miR-497-5p	0.988506	0.933985
Average	0.9852874	0.9612888



Supplementary Figure 1. Intersection of the top 20 microRNA (miRNA) candidates and the miRNAs presented in the top 25 combinations of expression change (EC)-based method and random forest algorithm (RF)-based method. The early stage breast cancer tissue data from The Cancer Genome Atlas were applied to both methods to obtain miRNA candidates and miRNA combinations.



Supplementary Figure 2. Comparison between expression change (EC)-based method and random forest algorithm (RF)-based method using tissue data from The Cancer Genome Atlas. Sensitivity and specificity of every combination consisting of ones in the top 20 microRNA (miRNA) candidates of each method were calculated and were grouped by the number of miRNAs in combinations. As the number of miRNAs increased in combinations, both mean sensitivity and mean specificity of EC-based method can reach a high level, which presented a more balanced performance than RF-based method.