www.impactjournals.com/oncotarget/ Oncotarget, 2016, Vol. 7, (No. 52), pp: 86457-86468

Research Paper

Discovery and replication of microRNAs for breast cancer risk using genome-wide profiling

Cenny Taslim[1], Daniel Y. Weng[1], Theodore M. Brasky[1], Ramona G. Dumitrescu[2], Kun Huang[1], Bhaskar V.S. Kallakury[3], Shiva Krishnan[1], Adana A. Llanos[4], Catalin Marian[1,11], Joseph McElroy[5], Sallie S. Schneider[6], Scott L. Spear[7], Melissa A. Troester[8], Jo L. Freudenheim[9], Susan Geyer[10], Peter G. Shields[1]

[1]Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA

[2]Distilled Spirits Council of the United States, Washington, DC, USA

[3]Department of Pathology, Georgetown University, Washington, DC, USA

[4]Department of Epidemiology, Rutgers University, New Brunswick, NJ, USA

[5]Center for Biostatistics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA

[6]Pioneer Valley Life Sciences Institute, Springfield, MA, USA

[7]Department of Plastic Surgery, Georgetown University Hospital, Washington, DC, USA

[8]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[9]Departement of Epidemiology and Environmental Health, School of Public Health and Health Professions, University at Buffalo, Buffalo, NY, USA

[10]Health Informatics Institute, University of South Florida, Tampa, FL, USA

[11]Victor Babes University of Medicine and Pharmacy, Timisoara, Romania

Correspondence to: *Peter G. Shields,* ***email:*** *Peter.Shields@osumc.edu*

Keywords: *microRNA, epigenetics, breast cancer risk prediction, tissue-based biomarkers, healthy women*

Received: *September 07, 2016* ***Accepted:*** *October 22, 2016* ***Published:*** *November 09, 2016*

ABSTRACT

Background: Genome-wide miRNA expression may be useful for predicting breast cancer risk and/or for the early detection of breast cancer.

Results: A 41-miRNA model distinguished breast cancer risk in the discovery study (accuracy of 83.3%), which was replicated in the independent study (accuracy = 63.4%, P=0.09). Among the 41 miRNA, 20 miRNAs were detectable in serum, and predicted breast cancer occurrence within 18 months of blood draw (accuracy 53%, *P*=0.06). These risk-related miRNAs were enriched for HER-2 and estrogen-dependent breast cancer signaling.

Materials and Methods: MiRNAs were assessed in two cross-sectional studies of women without breast cancer and a nested case-control study of breast cancer. Using breast tissues, a multivariate analysis was used to model women with high and low breast cancer risk (based upon Gail risk model) in a discovery study of women without breast cancer (*n*=90), and applied to an independent replication study (*n*=71). The model was then assessed using serum samples from the nested case-control study (*n*=410).

Conclusions: Studying breast tissues of women without breast cancer revealed miRNAs correlated with breast cancer risk, which were then found to be altered in the serum of women who later developed breast cancer. These results serve as proof-of-principle that miRNAs in women without breast cancer may be useful for predicting breast cancer risk and/or as an adjunct for breast cancer early detection. The miRNAs identified herein may be involved in breast carcinogenic pathways because they were first identified in the breast tissues of healthy women.

INTRODUCTION

Breast cancer is the most common cancer among women in the US, except for non-melanotic skin cancer, and is the second leading cause of cancer-related mortality in women [1]. Knowing which women will develop breast cancer remains elusive, and currently the most effective way of addressing breast cancer morbidity and mortality is through early detection and mammography [2, 3]. Statistical models based on breast cancer risk factors have been developed to assess life-time breast cancer risk in healthy women, guiding clinical decisions for early breast cancer detection (e.g., mammography and magnetic resonance imaging) and chemoprevention [4–10]. A widely used risk assessment method is the Breast Cancer Risk Assessment Tool, typically referred to as the "Gail model" [4], and is the only model that has been repeatedly validated in large population-based studies [5, 11–13]. The Gail model incorporates age, history of breast biopsies, family history of breast cancer, and reproductive histories. However, the predictivity of the Gail model is limited, as is its application to tailoring early detection for the general population of women. Thus, improved risk assessment and/or early detection methods are needed, because many aggressive breast cancers escape detection by mammography for some women, while at the same time mammography can lead to overdiagnosis for other women [14, 15].

One approach to improve life-time breast cancer risk prediction and/or the early detection of breast cancer is to utilize molecular signatures from normal tissue, before women develop clinical abnormalities [16–18]. For example, one study showed that epigenetic markers (DNA methylation) may improve the accuracy of the Gail model [19]. Recently, miRNAs have emerged as potential biomarkers for early detection of cancer [20–24]. miRNAs are short non-coding RNAs that are abundantly present in human cells, and negatively regulate gene and miRNA expression changes in breast cancer [25–27]. In normal cells, miRNAs affect mammary gland development and other functions [28]. In breast cancer, miRNA expression is associated with diagnosis and prognosis [29–31].

In this report, we hypothesized that the identification of miRNAs in healthy women associated with breast cancer risk can be one way of developing models for breast cancer risk assessment, and/or be used as an adjunct for enhancing early detection. To address this hypothesis, two independent cross-sectional studies of women undergoing reduction mammoplasty (RM) who had no prior history of breast cancer were used to build and evaluate a multi-miRNA model for breast cancer risk (i.e., Gail risk) prediction. We subsequently analyzed the National Institute of Environmental Health Science's Sister Study cohort (a publically available data set) [21], using a nested case-control design, to assess the utility of the multi-miRNA model using serum samples to directly assess breast cancer risk. This latter study includes women without breast cancer at the time of blood draw and were either diagnosed with breast cancer within 18 months (cases), or remained without breast cancer (controls).

RESULTS

Participants' characteristics

The characteristics of the participants are shown in Table 1. There were 90 and 71 subjects for the discovery and independent replication studies, respectively. The median age for the discovery subjects was 45 years (range: 35-76 years) and for the replication subjects it was 46 years (range: 35-66 years). The majority of subjects were Caucasians (71.1% and 78.9%, for the discovery and replication studies, respectively), and most were classified as low risk women by the Gail model; 83.3% and 67.6%, for the discovery and replication studies, respectively). Race and Gail risk, but not other characteristics, differed significantly between the two studies (*P* = 0.0005 and *P* = 0.032, respectively).

miRNAs in the RM discovery and independent replication studies

Five individual miRNAs (of the 168 miRNAs expressed above background) were differentially expressed in high vs. low risk women in the discovery study (*P* < 0.05), but none of these remained statistically significant after adjustment for multiple comparisons (Supplementary Table 1). To assess the applicability of miRNA-based model to predict breast cancer risk, we built a 41-miRNA model to distinguish between women with high vs. low risk of breast cancer (based upon the Gail risk model) in the discovery study (Supplementary Figures 1 and 2) using a projection-based multivariate classification technique for high dimensional data called sPLS-DA. The miRNAs in the model along with their weights are listed in Supplementary Table 2. Figure 1 shows the sPLS-DA components separating the two groups of women. The model had 83.3% predictive accuracy, 84% specificity, 80% sensitivity with 95.4% NPV and 50% PPV, in the discovery study. In an independent replication study, the same model achieved accuracy of 63.4% with specificity of 77.1% and sensitivity of 34.8%. The model had a NPV of 71.1% and a PPV of 42.1% for this data set (Table 2). The permutation test results show that our model predicteds high and low risk women with accuracies better than random chance (*P* = 0.09, Table 2 and Supplementary Information). In a sensitivity analysis, we additionally verified that classification using the 41-miRNA panel was not significantly influenced by race in the discovery RM study of women without breast cancer with 28% African American women (Supplementary Information).

Table 1: Characteristics of RM studies participants ≥ 35 y.o.

	Discovery study subjects (*n* = 90) [32]		Replication study subjects (*n* = 71) [34]	
Gail Characteristics	**No.**	**%**	**No.**	**%**
Race				
White	64	71.1	56	78.9
Black	25	27.8	5	7.0
Hispanic	1	1.1	8	11.3
Other	0	1.1	2	2.8
Age, years				
< 50	62	68.9	44	62.0
≥ 50	28	31.1	27	38.0
Median	45		46	
Range	35-76		35-66	
Age at menarche, years				
< 12	15	16.7	17	24.0
12-13	43	47.7	32	45.0
≥ 14	16	17.8	20	28.2
Unknown	16	17.8	2	2.8
Age at first live birth, years				
Nulliparous	22	24.4	13	18.3
< 20	9	10.0	16	22.6
20-24	12	13.3	17	23.9
25-29	16	17.8	10	14.1
≥ 30	11	12.2	13	18.3
Unknown	20	22.2	2	2.8
No. of 1st degree relatives with breast cancer				
0	57	63.3	60	84.5
1	8	8.9	9	12.7
Unknown	25	27.8	2	2.8
No. of biopsies				
0	40	44.4	65	91.6
1	6	6.7	4	5.6
≥ 2	2	2.2	2	2.8
Unknown	42	46.7	0	0
Breast Cancer risk[ǂ]				
Low	75	83.3	48	67.6
High	15	16.7	23	32.4

[ǂ] High breast cancer risk was defined as woman who has at least 10% increased risk of breast cancer relative to women at average risk of the same ethnicity and similar age, estimated by the Gail model.

Functional analysis of the miRNAs targets

To examine the biologic function of the miRNA panel, we performed Ingenuity Pathway Analysis (IPA) of the 41 miRNAs associated with increased Gail breast cancer risk. Given that a single miRNA can target many genes, the focus herein was on experimentally validated targets (based on the IPA knowledge base) of the top 10 miRNAs, ranked by their weights in the sPLS-DA model (Supplementary Table 2). Five out of the top 10 miRNAs had 94 experimentally validated targets. Figure 2A shows these five miRNAs and their respective gene targets that are known to be involved in cancer. The network that depicts the known direct interaction between these genes targets are shown in Figure 2B. This network is enriched in cell death and survival, cancer, and liver necrosis/cell death (*P* < 0.0001, Fisher's test). IPA canonical pathway analysis revealed significant enrichment in HER-2 signaling and estrogen-dependent breast cancer signaling, as well as other important cancer pathways such as PI3K/AKT signaling, PTEN signaling, and TGF-beta signaling (*P* < 0.0001, Figure 3 and Supplementary Table 3). Functional analysis indicated significant representation related to cellular growth and proliferation, cell death and survival, cell cycle and cancer (Supplementary Figure 3 and Supplementary Table 4).

For the other five miRNAs that had no experimentally validated targets, they had 98 predicted targets that are involved in breast cancer pathway (Supplementary Figure 4).

miRNAs in sister study cohort

Shifting from breast tissue to serum analysis for the prediction of actual breast cancer in the Sister Study cohort using a nested case-control design (*n* = 205 cases and 205 controls), 34 of the 41 breast miRNAs identified in the discovery study were profiled in the serum of the Sister Study cohort. There were 20 of 34 miRNAs detected above background level in more than 50 women; these 20 miRNAs were then used to build a new classification model in the discovery RM set, rather than the 41 miRNAs. This model had 81.1% accuracy, 81.3% specificity, and 80% sensitivity. The 20-miRNA model was then locked and applied to classify women in the breast cancer cohort subjects. The 20-miRNA model correctly identified 74.6% of women who remained cancer-free and 31.7% of women who were diagnosed with breast cancer (Table 2). The predictive accuracy was 53.2%. It was obtained by applying breast tissue results to serum, using a different miRNA platform, and studying women with a different body habitus (the RM subjects have a higher BMI than the general population). However, the 20-miRNA model achieved better performance than random permutation (*P* = 0.06). To be applicable in the clinical setting, we derived continuous risk scores based on the sPLS-DA model prediction. Women in the highest quartile of this score had a 53% increased risk for breast cancer (odds ratio [OR] = 1.53; *P* = 0.20), albeit based on small numbers of subjects and a statistically non-significant result.

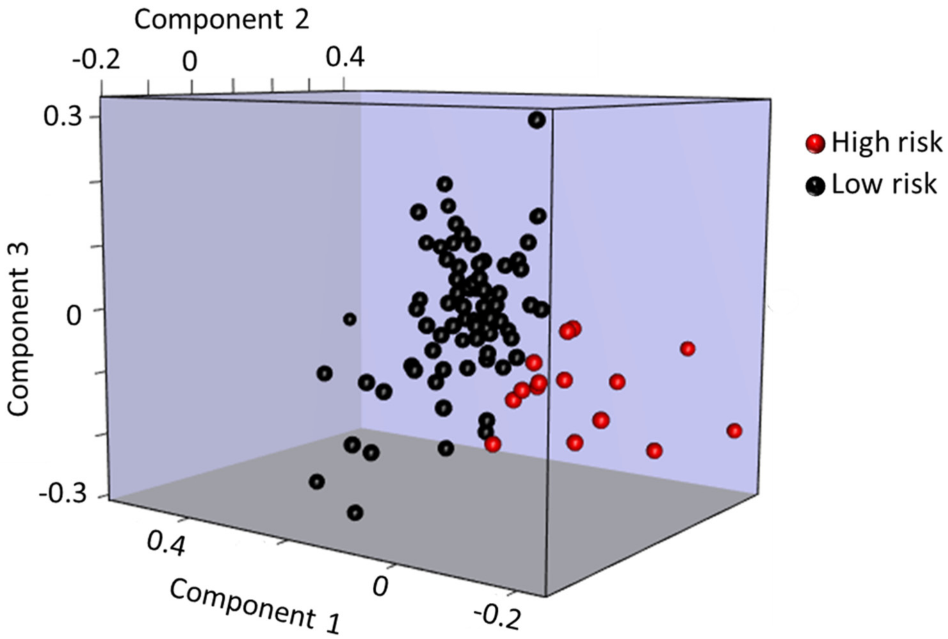

Figure 1: Graphical 3D representations of the women using sPLS-DA components. Plot of the first 3 components of the women showing a good separation between the women with high (red) and low (black) risk of developing breast cancer as calculated by Gail model.

Table 2: Classification performance of the discovery, independent replication and serum studies

Studies	Accuracy	Specificity	Sensitivity	Negative Predictive Value	Positive Predictive Value	*P*-value*
41-miRNA model performance in breast tissue						
Discovery	0.833	0.840	0.800	0.954	0.500	-
Replication	0.634	0.771	0.348	0.711	0.421	0.090
20-miRNA model performance in breast tissue and serum***						
Discovery	0.811	0.813	0.800	0.953	0.461	-
Serum (GSE44281)	0.532	0.746	0.317	-	-	0.064

**P*-value testing the performance of miRNA model based on 10,000 random permutations.
***Only 20 out of 41-miRNA were profiled and detected above background level in serum samples.

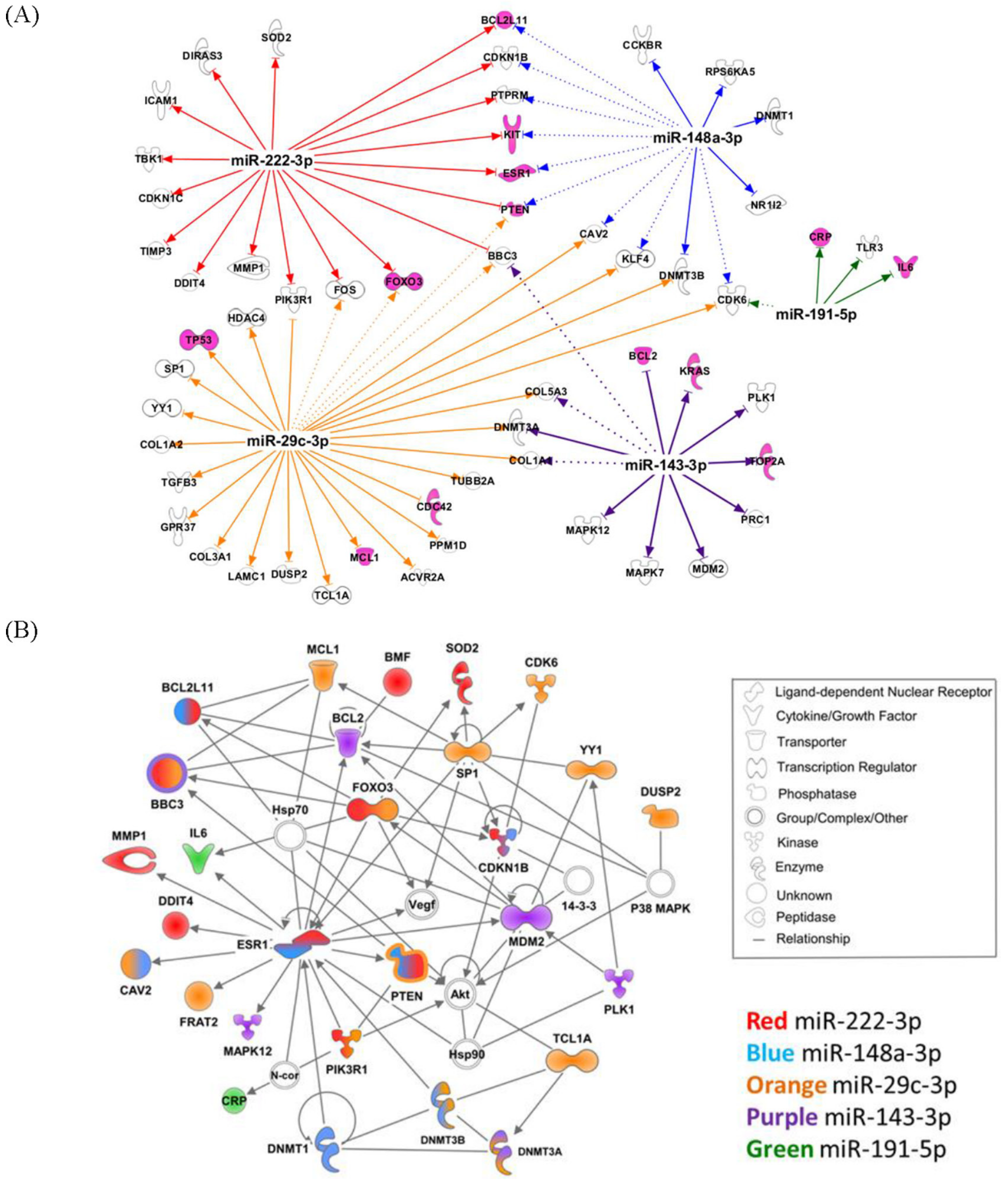

Figure 2: A. Five of the top 10 miRNAs have experimentally validated gene targets. Gene targets involved in cancer are shown. Pink molecules are important in breast cancer pathway. The connections show experimentally validated targets (solid line) and targets predicted with high confidence (dash line). **B.** The top network of the validated gene targets is enriched in cell death and survival, cancer, liver necrosis/cell death (P =10^{-50}, right-tailed Fisher's exact test). Fill colors represent molecules directly targeted by the corresponding miRNAs.

DISCUSSION

The use of molecular markers for cancer risk prediction is rapidly increasing [23, 32–35]. This study is the first to report on the accuracy of a miRNA model developed in breast tissues of women without breast cancer to predict breast cancer risk, serving as proof-of-principle that miRNAs profiled in histologically normal breast tissue may be used to predict the risk of developing breast cancer before major carcinogenic changes and/or as a way to enhance the early detection of breast cancer. We observed that this miRNA model developed using normal breast tissue may have some predictive power to differentiate women with high and low breast cancer risk and that a subset of these miRNAs detectable in the serum could identify women who were then diagnosed with breast cancer within 18 months. The results for women without breast cancer and their breast tissues indicate that miRNAs might be useful for assessing life-time breast cancer risk (as does the Gail risk model). The analysis of serum in the case-control study of breast cancer nested within the Sister Study cohort validates the results herein as a breast cancer risk predictor, but also possibly as a marker for early detection because of the short time to breast cancer diagnosis. While there was some loss in performance as the miRNA markers identified in breast tissues were applied to serum, this would be expected as serum analysis could reflect miRNA expression from many tissues, resulting in a loss of signal. The use of serum markers as surrogates for the target organ has been previously reported [22, 36–38]. Pathway analysis of the mRNA targets of the top miRNAs identified in the model suggested enrichment for HER-2 and estrogen-dependent breast cancer signaling, and other cancer-related pathways. Our study using tissues from women with no history of breast cancer provides a unique resource for risk assessment and early detection prior to abnormalities that are clinically detectable.

miR-222-3p, one of the top-ranked miRNAs detected in the model, has been shown to target the estrogen receptor 1 gene (ESR1) and was reported to be dramatically higher in ESR1 negative breast cancer cells, inhibiting ERα expression [39]. miR-222-3p can also trigger malignant transformation by altering the expression levels of genes involved in cell death and survival, such as CAV2, PTEN, FOXO3, CDK6 and promote aggressive ER-negative breast tumors by increasing proliferation and migratory activity of breast cancer cells [40, 41]. The miR-29c-3p, another top-ranked miRNA identified in the model, has been shown to up-regulate p53 and

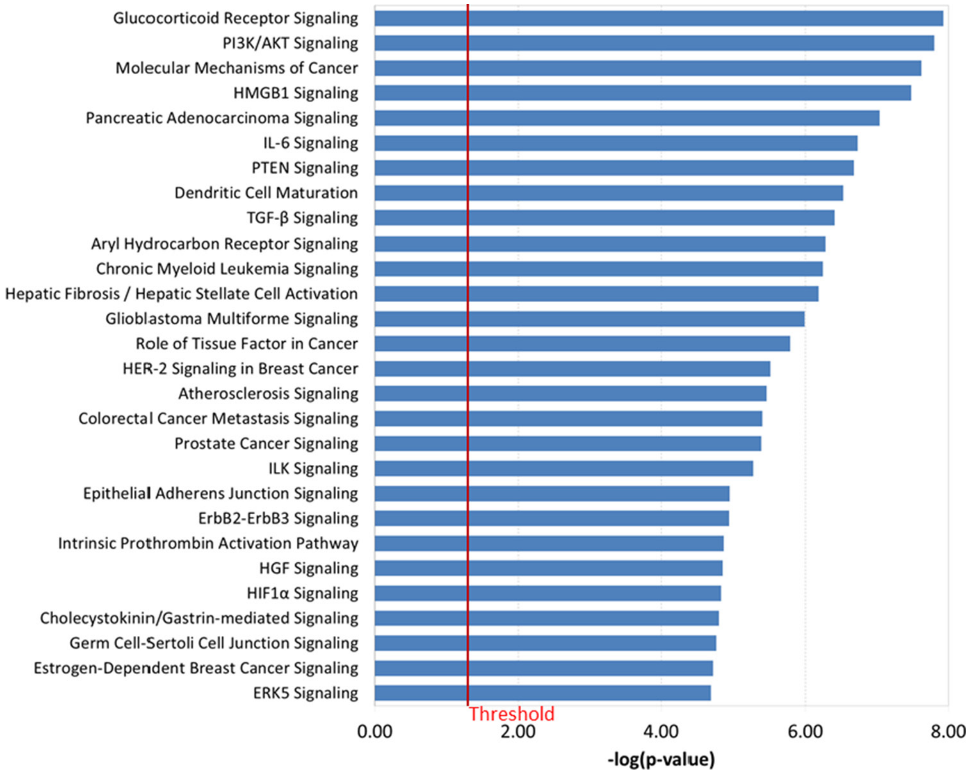

Figure 3: Canonical pathways that are significantly associated with the experimentally observed gene targets of the top 10 miRNA in the 41-miRNA panel using IPA. Fisher's exact test was used to calculate a *P* value. Values greater than the threshold implies that the association between the miRNA gene targets and the pathway is not likely due to random chance alone.

induce apoptosis in breast cancer cell lines [42]. For the other top-ranked miRNAs, five have no validated targets but have many predicted targets related to breast cancer such as AKT, AKT1, CCND1, EGFR1, ERBB2, SRC, PTEN which have been used as breast cancer biomarkers for prognosis, diagnosis, drug efficacy, and disease progression [43–45]. Thus, these miRNAs may have clinical potential as novel breast cancer biomarkers.

Three prior cohort studies have investigated miRNA expression in prospectively collected blood samples for breast cancer risk prediction, including the Sister Study Cohort used herein [21, 46, 47]. None were based on miRNA expression in the breast, none had independent replication and the miRNAs for each study do not overlap with each other. There were other substantial differences between the study reported herein and these other studies. For example, while the Hormones and Diet in the Etiology of Breast Cancer Cohort reported 20 differentially expressed miRNAs among 133 postmenopausal breast cancer cases and 133 controls, the miRNAs were assessed in leukocytes [46]. Leukocyte miRNA expression would only affect expression for that cell type, while serum miRNA levels likely reflects a contribution of multiple organs. Similarly, the Breast Cancer Family Registry, using high risk families, reported that five microRNAs were differentially expressed in blood cells, among 20 breast cancer cases and 20 controls, but none were validated in an additional 28 case/control pair from the same study [47]. Lastly, the Sister Study Cohort utilized serum drawn only a short time before diagnosis (within 18 months and a mean of 10 months), so these results reflect an assessment for early detection rather than long term risk prediction [21]. Each study also used a different laboratory assay for miRNA detection. The miRNAs identified herein using normal tissues were not observed in the Registry Study, but four overlapped with the Sister Study Cohort results (two of them were the highest expressing miRNAs in the Sister Study, i.e., miR-181a-5p and miR-222-3p) and 3 miRNAs overlapped with the Hormones and Diet Study (i.e., miR-1991-5p, miR-145-5p, and miR-199b-5p). Whether the disparate results are due to differences in study design, blood component, or assay methodology is unclear. Currently there is no consistency in the scientific literature for which miRNAs might be predictive of breast cancer risk, however no other study has used independent validation and assessed the accuracy of a miRNA-based model.

The performance of the model applied to two independent studies is very modest albeit significantly better than random chance at *P*-value < 0.10. This most likely due to the small number of women in the discovery study (and limited number of high breast cancer risk women without breast cancer) that were used to train the model. Moreover, transition from breast tissue to blood

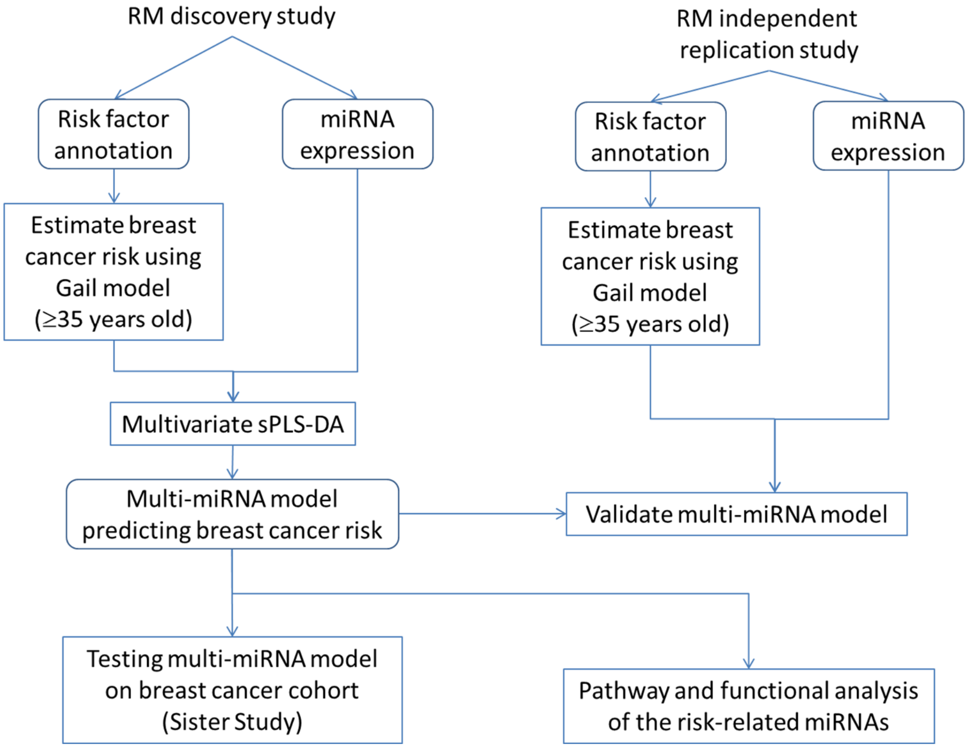
RM discovery study
Risk factor annotation
miRNA expression
Estimate breast cancer risk using Gail model (≥35 years old)
Multivariate sPLS-DA
Multi-miRNA model predicting breast cancer risk
Testing multi-miRNA model on breast cancer cohort (Sister Study)
RM independent replication study
Risk factor annotation
miRNA expression
Estimate breast cancer risk using Gail model (≥35 years old)
Validate multi-miRNA model
Pathway and functional analysis of the risk-related miRNAs

Figure 4: Workflow of the data analysis performed in this study.
Abbreviation: sPLS-DA, sparse partial least square discriminant analysis.

where some miRNAs may be expressed differently, the detection method differences, the broader range of ages and BMI, and the time from blood collection to diagnosis all could influence the results.

Findings from this study should be considered in light of some limitations. The first relates to the use of the Gail risk model, which has limited accuracy, despite the fact that it is a widely used model. Other models could have been used, such as Tyrer-Cuzick [9] and BRCAPRO [48], but these have not been validated in large population studies. Another limitation is the use of reduction mammoplasty patients for discovery and replication, where the women are mostly overweight or obese, thus limiting the generalizability of results to normal weight women. However, finding concordance with the Sister Study directly minimizes concerns about the generalizability of these findings. Separately, this study was limited to the publically available data from the Sister Study, limiting our ability to directly compare our miRNA model with Gail model and to explore additional covariates and confounding (e.g., BMI).

This study approaches the discovery of breast cancer risk miRNA models by first studying breast tissues of women without breast cancer providing comprehensive molecular measurement in target tissue. The results indicate modest concordance of miRNA expression between breast tissues and serum of women who later develop breast cancer, but given the reasons discussed above while applying breast modeling results to serum in different studies, the results indicate that the novel approach used herein has the utility to provide corroborative evidence for the ultimate development of a miRNA model for predicting breast cancer risk and/or early detection. Future prospective cohort studies with large sample size and long follow-up data are needed to confirm the promise of miRNAs in breast cancer risk prediction and early detection.

MATERIALS AND METHODS

Study population and biospecimen collection

Reduction mammoplasty (RM) discovery study

Healthy women with no prior history of breast cancer who underwent RM were studied, as described previously [49, 50]. Briefly, subjects aged 35 and older were recruited at Georgetown University Medical Center (Washington, DC), the University of Maryland (College Park, MD), the Washington Hospital Center (Washington, DC) and the Center for Plastic Surgery (Buffalo, NY) from 1997 to 2009. Recorded data by personal interview included demographics, lifestyle, reproductive history, family medical history, diet, and other exposures. Upon pathological review, subjects with gross pathology, epithelial hyperplasia, or focal microcalcifications were excluded. Tissues were dissected to remove adipose tissue, fixed in formalin, and embedded in paraffin.

RM independent replication study

Women ages ≥35 who underwent RM at Baystate Medical Center (Springfield, MA) between 2007 and 2009 were studied, as previously described [51]. Participants were interviewed by phone following surgery and data available from the questionnaires were similar to those in the discovery study. Breast tissues were collected similarly, except that the breast epithelial tissue was not dissected from fat before storage. Participants with benign biopsy results also were excluded.

Cohort study of breast cancer - Sister Study

A cohort study for breast cancer where there are publically-available serum-based miRNA profiles were then analyzed in relation to miRNA model identified from the above RM discovery study, using a nested case-control study design. The National Institute of Environmental and Health Sciences (NIEHS) Sister Study (NCBI Gene Expression Omnibus, accession number GSE44281) [21] includes 50,844 women from US or Puerto Rico who never had breast cancer but had a sister diagnosed with breast cancer. Baseline serum samples of 205 women without breast cancer who were subsequently diagnosed with breast cancer within 18 months following blood collection (mean = 10 months) were matched with 205 women who remained cancer free. The matching criteria were no prior history of cancer except non-melanoma skin cancer, race, age, date of blood draw and available blood sample. For this cohort, miRNA expression levels were determined using GeneChip miRNA 2.0 arrays (Affymetrix Inc., Santa Clara, CA, USA) as described by the original authors [21].

All participants in the three studies provided informed consent and studies were approved by the Institutional Review Boards of all participating institutions.

RNA extraction and miRNA profiling in breast tissues

Total RNA was extracted from the two RM studies using formalin-fixed paraffin embedded (FFPE) tissues according to the manufacturer's instructions (FFPE RNA purification kit, Norgen, Canada). miRNA expression was quantified using the nCounter digital detection miRNA expression assay (NanoString® Technologies, Seattle WA), with 10 % duplicates for quality control purposes. Coefficients of variation (CV) were calculated to assess assay reliability (3.76% and 4.62% in RM discovery and replication studies, respectively).

RM subjects breast cancer risk assessment

For the RM studies, each woman's Gail score was calculated [10]. We defined "high risk" women as those with a 10% or greater increased risk relative to women of the same race and age who is at average risk (e.g. > 1.1% risk if the average 5-year risk is 1%). We refer to low risk as average risk in the population for a given age and racial/ethnic demographic. Five year risks were calculated using the latest update (May 2011) of source code for the breast cancer risk calculation engine [52]. We verified the risks calculated using the source code with risks calculated using the online tools for 30 random women. All risk estimates showed complete agreement.

Data analysis

Figure 4 shows an overview of the data analysis performed in this study. Raw expressions were normalized using the top quartile mean normalization method [53], which has been shown to be more sensitive and accurate than using invariant miRNA [54] or quantile normalization [55, 56] (Supplementary Figure 5). 800 miRNAs were measured using the NanoString nCounter platform [57]. We filtered out miRNAs whose expressions were below NanoString's internal negative control in at least 50% of the samples, leaving 168 for downstream analysis. Principal Component Analysis (PCA) was used to detect the presence of potential confounding factors including technical artifacts and batch effects. In our miRNA dataset, no confounding was found (Supplementary Figure 6).

To confirm that patient characteristics in the RM studies were similar, chi-squared tests were used to compare the categorical characteristics between participants in the two studies. Two-sided Wilcoxon rank sum tests were used to compare expression of each miRNA individually between high and low risk women. *P*-values were corrected for multiple testing using the Benjamini-Hochberg False Discovery Rate (FDR) algorithm. [58] An FDR value less than 0.10 was considered statistically significant.

In order to build a multi-miRNA model associated with breast cancer risk in the discovery study, we used a projection-based multivariate sparse partial least squares discriminant analysis (sPLS-DA [59]) approach to classify women into low and high breast cancer risk groups. sPLS-DA was designed to classify high dimensional data that also performed variable selection [59]. Optimal parameters tuning was performed based on 10-fold cross validation (CV) to avoid overfitting [60].

In order to replicate the findings of the RM discovery study, the miRNA model which was developed in the discovery study was used to predict the breast cancer risks of women in the replication study. The accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the miRNA model were calculated to evaluate the discriminatory performance of the miRNA panel in classifying women with low versus high Gail risk [60]. To assess whether our model is better than random chance, we applied a permutation test by reshuffling the labels, applying the same model 10,000 times and calculating the p-value. The null distribution of the accuracies given by the permutation is the accuracy of the model when it is a random signature. Small p-value indicates that the accuracy of the model is significantly better than random chance.

To assess whether the miRNA panel can prospectively predict breast cancer risk using serum in the Sisters Study, miRNAs by Affymetrix® used for the Sister Study and NanoString for the RM studies were matched based on their sequences to identify available miRNAs for further analysis. Since only 20 out of 41 miRNAs were detected above background by the Affymetrix method in serum, we used these 20 miRNA to build a sPLS-DA model in the discovery study. This model was then applied to classify women into breast cancer and cancer-free categories in the Sister Study cohort. Finally, the same permutation test as described above was applied to assess whether the performance of our model was better than what chance alone could produce. Continuous risk scores were derived by taking the difference of the two outcomes variables predicted by sPLS-DA. Logistic regression was used to estimate breast cancer odds ratios (OR) and 95% confidence intervals (CI) for comparisons of the second through fourth quartiles of the risk score relative to the first. All analyses described in this section were performed in the R statistical environment v3.1.1. Further details are available in the Supplementary Information.

Functional and pathway analysis

Functional analysis was performed using QIAGEN's Ingenuity Pathways Analysis (IPA® QIAGEN Redwood City, www.ingenuity.com). miRNA targets were identified using TarBase [61], miRecords [62], TargetScan [63] and Ingenuity® knowledge base. Methodology and approaches are described in detail in Supplementary Information.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

GRANT SUPPORT

This work was supported by Avon Foundation [01-2010-039] and Breast Cancer Center of Excellence Award from United States Department of Defense [BC022346].

REFERENCES

1. American Cancer Society. Cancer Facts & Figures 2015. Atlanta: American Cancer Society. 2015.

2. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JDF, Feuer EJ. Effect of Screening and Adjuvant Therapy on Mortality from Breast Cancer. New England Journal of Medicine. 2005; 353: 1784–92. doi: 10.1056/NEJMoa050518.

3. Kalager M, Zelen M, Langmark F, Adami H-O. Effect of Screening Mammography on Breast-Cancer Mortality in Norway. New England Journal of Medicine. 2010; 363: 1203–10. doi: 10.1056/NEJMoa1000727.

4. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. Journal of the National Cancer Institute. 1989; 81: 1879–86. doi: 10.1093/jnci/81.24.1879.

5. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, Anderson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W, Norman SA, Simon MS, Spirtas R, et al. Projecting individualized absolute invasive breast cancer risk in African American women. Journal of the National Cancer Institute. 2007; 99: 1782–92. doi: 10.1093/jnci/djm223.

6. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DSM, Geller BM, Rosenberg R, Yankaskas BC, Kerlikowske K. Prospective breast cancer risk prediction model for women undergoing screening mammography. Journal of the National Cancer Institute. 2006; 98: 1204–14. doi: 10.1093/jnci/djj331.

7. Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail MH. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. Journal of the National Cancer Institute. 2006; 98: 1215–26. doi: 10.1093/jnci/djj332.

8. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer. 1994; 73: 643–51.

9. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine. 2004; 23: 1111–30. doi: 10.1002/sim.1668.

10. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, Gail MH. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. Journal of the National Cancer Institute. 2011; 103: 951–61. doi: 10.1093/jnci/djr154.

11. Spiegelman D, Colditz GA, Hunter D, Hertzmark E. Validation of the Gail et al. Model for Predicting Individual Breast Cancer Risk. Journal of the National Cancer Institute. 1994; 86: 600–7. doi: 10.1093/jnci/86.8.600.

12. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS. Validation Studies for Models Projecting the Risk of Invasive and Total Breast Cancer Incidence. Journal of the National Cancer Institute. 1999; 91: 1541–8. doi: 10.1093/jnci/91.18.1541.

13. Bondy ML, Lustbader ED, Halabi S, Ross E, Vogel VG. Validation of a breast cancer risk assessment model in women with a positive family history. Journal of the National Cancer Institute. 1994; 86: 620–5.

14. Nattinger AB. In the clinic. Breast cancer screening and prevention. Annals of internal medicine. 2010; 152: ITC41. doi: 10.7326/0003-4819-152-7-201004060-01004.

15. Collett K, Stefansson IM, Eide J, Braaten A, Wang H, Eide GE, Thoresen SØ, Foulkes WD, Akslen LA. A basal epithelial phenotype is more frequent in interval breast cancers compared with screen detected tumors. Cancer epidemiology, biomarkers & prevention. 2005; 14: 1108–12. doi: 10.1158/1055-9965.EPI-04-0394.

16. Pankratz VS, Hartmann LC, Degnim AC, Vierkant RA, Ghosh K, Vachon CM, Frost MH, Maloney SD, Reynolds C, Boughey JC. Assessment of the accuracy of the Gail model in women with atypical hyperplasia. Journal of clinical oncology. 2008; 26: 5374–9. doi: 10.1200/JCO.2007.14.8833.

17. Huh SJ, Oh H, Peterson MA, Almendro V, Hu R, Bowden M, Lis RT, Cotter MB, Loda M, Barry WT, Polyak K, Tamimi RM. The proliferative activity of mammary epithelial cells in normal tissue predicts breast cancer risk in premenopausal women. Cancer research. 2016; 76: 1926–34. doi: 10.1158/0008-5472.CAN-15-1927.

18. Campbell JD, Mazzilli SA, Reid ME, Dhillon SS, Platero S, Beane J, Spira AE. The Case for a Pre-Cancer Genome Atlas (PCGA). Cancer prevention research (Philadelphia, Pa). 2016; 9: 119–24. doi: 10.1158/1940-6207.CAPR-16-0024.

19. Xu Z, Bolick SCE, DeRoo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. Journal of the National Cancer Institute. 2013; 105: 694–700. doi: 10.1093/jnci/djt045.

20. Wang L-G, Gu J. Serum microRNA-29a is a promising novel marker for early detection of colorectal liver metastasis. Cancer epidemiology. 2012; 36: e61–7. doi: 10.1016/j.canep.2011.05.002.

21. Godfrey AC, Xu Z, Weinberg CR, Getts RC, Wade PA, Deroo LA, Sandler DP, Taylor JA. Serum microRNA expression as an early marker for breast cancer risk in prospectively collected samples from the Sister Study cohort. Breast cancer research. BioMed Central Ltd; 2013; 15: R42. doi: 10.1186/bcr3428.

22. Chan M, Liaw CS, Ji SM, Tan HH, Wong CY, Thike AA, Tan PH, Ho GH, Lee AS-G. Identification of circulating microRNA signatures for breast cancer detection. Clinical cancer research. 2013; 19: 4477–87. doi: 10.1158/1078-0432.CCR-12-3401.

23. Sozzi G, Boeri M, Rossi M, Verri C, Suatoni P, Bravi F, Roz L, Conte D, Grassi M, Sverzellati N, Marchiano A, Negri E, La Vecchia C, et al. Clinical Utility of a Plasma-Based miRNA Signature Classifier Within Computed Tomography Lung Cancer Screening: A Correlative MILD Trial Study.

Journal of Clinical Oncology. 2014; 32: 768–73. doi: 10.1200/JCO.2013.50.4357.

24. Massion PP. Biomarkers to the rescue in a lung nodule epidemic. Journal of clinical oncology. 2014; 32: 725–6. doi: 10.1200/JCO.2013.54.0047.

25. Iorio MV, Croce CM. MicroRNAs in cancer: small molecules with a huge impact. Journal of clinical oncology. 2009; 27: 5848–56. doi: 10.1200/JCO.2009.24.0317.

26. Bartel D. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell. 2004; 116: 281–97. doi: 10.1016/S0092-8674(04)00045-5.

27. IIorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO molecular medicine. 2012; 4: 143–59. doi: 10.1002/emmm.201100209.

28. Silveri L, Tilly G, Vilotte J-L, Le Provost F. MicroRNA involvement in mammary gland development and breast cancer. Reproduction, nutrition, development. 2006; 46: 549–56. doi: 10.1051/rnd:2006026.

29. Sakurai M, Masuda M, Miki Y, Hirakawa H, Suzuki T, Sasano H. Correlation of miRNA expression profiling in surgical pathology materials, with Ki-67, HER2, ER and PR in breast cancer patients. The International journal of biological markers. 2015; 30: e190–9. doi: 10.5301/jbm.5000141.

30. Gasparini P, Cascione L, Fassan M, Lovat F, Guler G, Balci S, Irkkan C, Morrison C, Croce CM, Shapiro CL, Huebner K. microRNA expression profiling identifies a four microRNA signature as a novel diagnostic and prognostic biomarker in triple negative breast cancers. Oncotarget. 2014; 5: 1174–84. doi: 10.18632/oncotarget.1682.

31. Parrella P, Barbano R, Pasculli B, Fontana A, Copetti M, Valori VM, Poeta ML, Perrone G, Righi D, Castelvetere M, Coco M, Balsamo T, Morritti M, et al. Evaluation of microRNA-10b prognostic significance in a prospective cohort of breast cancer patients. Molecular cancer. 2014; 13: 142. doi: 10.1186/1476-4598-13-142.

32. Sawyer S, Mitchell G, McKinley J, Chenevix-Trench G, Beesley J, Chen XQ, Bowtell D, Trainer AH, Harris M, Lindeman GJ, James PA. A role for common genomic variants in the assessment of familial breast cancer. Journal of clinical oncology. 2012; 30: 4330–6. doi: 10.1200/JCO.2012.41.7469.

33. Couch FJ, Gaudet MM, Antoniou AC, Ramus SJ, Kuchenbaecker KB, Soucy P, Beesley J, Chen X, Wang X, Kirchhoff T, McGuffog L, Barrowdale D, Lee A, et al. Common variants at the 19p13.1 and ZNF365 loci are associated with ER subtypes of breast cancer and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers. Cancer epidemiology, biomarkers & prevention. 2012; 21: 645–57. doi: 10.1158/1055-9965.EPI-11-0888.

34. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, Moran S, Heyn H, Vizoso M, Gomez A, Sanchez-Cespedes M, Assenov Y, Müller F, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. Journal of clinical oncology. 2013; 31: 4140–7. doi: 10.1200/JCO.2012.48.5516.

35. Hu Z, Chen X, Zhao Y, Tian T, Jin G, Shu Y, Chen Y, Xu L, Zen K, Zhang C, Shen H. Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. Journal of clinical oncology. 2010; 28: 1721–6. doi: 10.1200/JCO.2009.24.9342.

36. Wulfken LM, Moritz R, Ohlmann C, Holdenrieder S, Jung V, Becker F, Herrmann E, Walgenbach-Brünagel G, von Ruecker A, Müller SC, Ellinger J. MicroRNAs in renal cell carcinoma: diagnostic implications of serum miR-1233 levels. PloS one. 2011; 6: e25787. doi: 10.1371/journal.pone.0025787.

37. van Schooneveld E, Wouters MC, Van der Auwera I, Peeters DJ, Wildiers H, Van Dam PA, Vergote I, Vermeulen PB, Dirix LY, Van Laere SJ. Expression profiling of cancerous and normal breast tissues identifies microRNAs that are differentially expressed in serum from patients with (metastatic) breast cancer and healthy volunteers. Breast cancer research. 2012; 14: R34. doi: 10.1186/bcr3127.

38. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105: 10513–8. doi: 10.1073/pnas.0804549105.

39. Hu Z, Dong J, Wang L-E, Ma H, Liu J, Zhao Y, Tang J, Chen X, Dai J, Wei Q, Zhang C, Shen H. Serum microRNA profiling and breast cancer risk: the use of miR-484/191 as endogenous controls. Carcinogenesis. 2012; 33: 828–34. doi: 10.1093/carcin/bgs030.

40. Di Leva G, Gasparini P, Piovan C, Ngankeu A, Garofalo M, Taccioli C, Iorio M V, Li M, Volinia S, Alder H, Nakamura T, Nuovo G, Liu Y, et al. MicroRNA cluster 221-222 and estrogen receptor alpha interactions in breast cancer. Journal of the National Cancer Institute. 2010; 102: 706–21. doi: 10.1093/jnci/djq102.

41. Stinson S, Lackner MR, Adai AT, Yu N, Kim H-J, O'Brien C, Spoerke J, Jhunjhunwala S, Boyd Z, Januario T, Newman RJ, Yue P, Bourgon R, et al. TRPS1 targeting by miR-221/222 promotes the epithelial-to-mesenchymal transition in breast cancer. Science signaling. 2011; 4: ra41. doi: 10.1126/scisignal.2001538.

42. Park S-Y, Lee JH, Ha M, Nam J-W, Kim VN. miR-29 miRNAs activate p53 by targeting p85 alpha and CDC42. Nature structural & molecular biology. 2009; 16: 23–9. doi: 10.1038/nsmb.1533.

43. Royal Marsden NHS Foundation Trust. Evaluation Of The Role Of Nipple Aspiration, Ductal Lavage And Duct Endoscopy At The Time Of Surgery In Patients With Breast Cancer. In: ClinicalTrials.gov. Bethesda (MD): National Library of Medicine (US). 2000-[cited 2016 Nov

26]. Available from: https://clinicaltrials.gov/ct2/show/NCT00083018 NLM Identifier: NCT00083018.

44. National Cancer Institute (NCI). Phase I/II Trial of IMC-A12 in Combination With Temsirolimus in Patients With Metastatic Breast Cancer. In: ClinicalTrials.gov. Bethesda (MD): National Library of Medicine (US). 2000-[cited 2016 Nov 26]. Available from: https://clinicaltrials.gov/ct2/show/NCT00699491 NLM Identifier: NCT00699491.

45. Vanderbilt-Ingram Cancer Center; National Cancer Institute (NCI). A Phase Ib/II Study of Cisplatin, Paclitaxel, and RAD001 in Patients With Metastatic Breast Cancer. In: ClinicalTrials.gov. Bethesda (MD): National Library of Medicine (US). 200-[cited 2016 Nov 26]. Available from: https://clinicaltrials.gov/ct2/show/NCT01031446 NLM Identifier: NCT01031446.

46. Muti P, Sacconi A, Hossain A, Donzelli S, Bossel BM, Ganci F, Sieri S, Krogh V, Berrino F, Biagioni F, Strano S, Beyene J, Yarden Y, et al. Downregulation of microRNAs 145-3p and 145-5p is a Long-Term Predictor of Postmenopausal Breast Cancer Risk: the ORDET prospective study. Cancer EpidemiolBiomarkers Prev. 2014; : 2471–82. doi: 10.1158/1055-9965.EPI-14-0398.

47. Chang C-WW, Wu H-CC, Terry MB, Santella RM. microRNA Expression in Prospectively Collected Blood as a Potential Biomarker of Breast Cancer Risk in the BCFR. Anticancer Res. 2015; 35: 3969–77.

48. Berry DA, Parmigiani G, Sanchez J, Schildkraut J, Winer E. Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. Journal of the National Cancer Institute. 1997; 89: 227–38. doi: 10.1093/jnci/89.3.227.

49. Llanos AA, Brasky TM, Dumitrescu RG, Marian C, Makambi KH, Kallakury BVS, Spear SL, Perry DJ, Convit RJ, Platek ME, Adams-Campbell LL, Freudenheim JL, Shields PG. Plasma IGF-1 and IGFBP-3 may be imprecise surrogates for breast concentrations: an analysis of healthy women. Breast cancer research and treatment. 2013; 138: 571–9. doi: 10.1007/s10549-013-2452-y.

50. Dumitrescu RG, Marian C, Krishnan SS, Spear SL, Kallakury BVS, Perry DJ, Convit JR, Seillier-Moiseiwitsch F, Yang Y, Freudenheim JL, Shields PG. Familial and racial determinants of tumour suppressor genes promoter hypermethylation in breast tissues from healthy women. Journal of cellular and molecular medicine. 2010; 14: 1468–75. doi: 10.1111/j.1582-4934.2009.00924.x.

51. Pirone JR, D'Arcy M, Stewart DA, Hines WC, Johnson M, Gould MN, Yaswen P, Jerry DJ, Smith Schneider S, Troester MA. Age-associated gene expression in normal breast tissue mirrors qualitative age-at-incidence patterns for breast cancer. Cancer epidemiology, biomarkers & prevention. 2012; 21: 1735–44. doi: 10.1158/1055-9965.EPI-12-0451.

52. Breast Cancer Risk Assessment C# program. Available from http://www.cancer.gov/bcrisktool/download-source-code.aspx.

53. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, Vandesompele J. A novel and universal method for microRNA RT-qPCR data normalization. Genome biology. 2009; 10: R64. doi: 10.1186/gb-2009-10-6-r64.

54. D'haene B, Mestdagh P, Hellemans J, Vandesompele J. miRNA expression profiling: from reference genes to global mean normalization. Methods in molecular biology (Clifton, NJ). 2012; 822: 261–72. doi: 10.1007/978-1-61779-427-8_18.

55. Prokopec SD, Watson JD, Waggott DM, Smith AB, Wu AH, Okey AB, Pohjanvirta R, Boutros PC. Systematic evaluation of medium-throughput mRNA abundance platforms. RNA (New York, NY). 2013; 19: 51–62. doi: 10.1261/rna.034710.112.

56. Bolstad BM, Irizarry R., Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19: 185–93. doi: 10.1093/bioinformatics/19.2.185.

57. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nature biotechnology. 2008; 26: 317–25. doi: 10.1038/nbt1385.

58. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57: 289–300. doi: 10.2307/2346101.

59. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC bioinformatics. 2011; 12: 253. doi: 10.1186/1471-2105-12-253.

60. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. 1994. 456 p.

61. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic acids research. 2012; 40: D222–9. doi: 10.1093/nar/gkr1161.

62. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic acids research. 2009; 37: D105–10. doi: 10.1093/nar/gkn851.

63. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome research. 2009; 19: 92–105. doi: 10.1101/gr.082701.108.