PLOS | ONE

RESEARCH ARTICLE

DNA Methylation Patterns Can Estimate Nonequivalent Outcomes of Breast Cancer with the Same Receptor Subtypes

Min Zhang[1], Shaojun Zhang[1], Yanhua Wen[1], Yihan Wang[1], Yanjun Wei[1], Hongbo Liu[1], Dongwei Zhang[2], Jianzhong Su[1], Fang Wang[1]*, Yan Zhang[1]*

1 College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China, **2** Department of General Surgery, The Second Affiliated Hospital of Harbin Medical University, Harbin 150086, China

* wangfang@ems.hrbmu.edu.cn (FW); tyozhang@ems.hrbmu.edu.cn (YZ)

OPEN ACCESS

Citation: Zhang M, Zhang S, Wen Y, Wang Y, Wei Y, Liu H, et al. (2015) DNA Methylation Patterns Can Estimate Nonequivalent Outcomes of Breast Cancer with the Same Receptor Subtypes. PLoS ONE 10 (11): e0142279. doi:10.1371/journal.pone.0142279

Editor: William B. Coleman, University of North Carolina School of Medicine, UNITED STATES

Received: July 22, 2015

Accepted: October 20, 2015

Published: November 9, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors thank the support of National Natural Science Foundation of China [grant number 61402139, 31401075, 31371334, 61403112, 61203262], the Natural Science Foundation of Heilongjiang Province [grant number ZD2015003] and the Natural Scientific Research Fund of Heilongjiang Provincial [grant number QC2011C061 to SZ].

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Breast cancer has various molecular subtypes and displays high heterogeneity. Aberrant DNA methylation is involved in tumor origin, development and progression. Moreover, distinct DNA methylation patterns are associated with specific breast cancer subtypes. We explored DNA methylation patterns in association with gene expression to assess their impact on the prognosis of breast cancer based on Infinium 450K arrays (training set) from The Cancer Genome Atlas (TCGA). The DNA methylation patterns of 12 featured genes that had a high correlation with gene expression were identified through univariate and multivariable Cox proportional hazards models and used to define the methylation risk score (MRS). An improved ability to distinguish the power of the DNA methylation pattern from the 12 featured genes ($p = 0.00103$) was observed compared with the average methylation levels ($p = 0.956$) or gene expression ($p = 0.909$). Furthermore, MRS provided a good prognostic value for breast cancers even when the patients had the same receptor status. We found that ER-, PR- or Her2- samples with high-MRS had the worst 5-year survival rate and overall survival time. An independent test set including 28 patients with death as an outcome was used to test the validity of the MRS of the 12 featured genes; this analysis obtained a prognostic value equivalent to the training set. The predict power was validated through two independent datasets from the GEO database. The DNA methylation pattern is a powerful predictor of breast cancer survival, and can predict outcomes of the same breast cancer molecular subtypes.

Introduction

Breast cancer is the second largest cause of morbidity worldwide, the first cause of tumors in women [1], and the leading cause of cancer death in women. Moreover, the incidence rates of breast cancer are continuing increase [2]. Breast cancer has multiple molecular subtypes that are classified using tumor biomarkers, such as hormone receptors (HR) (i.e., estrogen receptor (ER) and progesterone receptor (PR)) and human epidermal growth factor receptor 2 (Her2)

[3]. Four clinical subtypes of breast cancer can be separated according to HR expression and the epithelial cell of origin (luminal or basal): Luminal A (HR+/Her2-), Luminal B (HR+/Her2+), Her2-enriched (HR-/Her2+), and triple-negative (HR-/Her2-) [4–7]. Breast cancer is a highly heterogeneous disease between and within tumors as well as among cancer-bearing individuals, which is a challenge for the diagnosis, treatment and prognosis [8].

DNA methylation is an epigenetic modification that plays important roles in gene expression regulation, cellular differentiation, development and even tumorigenesis. DNA methylation often occurs at the C-5 position of cytosine, especially cystonsines located in C-phosphate-G (CpG) sites. DNA hypermethylation in gene promoter or CpG islands can result in tumor suppressor silencing, leading to tumorigenesis. Therefore, a large number of differentially methylated regions in cancer have been identified to explore the epigenetic regulation mechanisms underlying oncogenesis [9]. Recently, DNA methylation biomarkers for the diagnosis, molecular typing and prognosis of breast cancer were identified. For example, hypermethylation of RASSF1A can be used to detect breast cancer during the early stages using a CpG island that is hypermethylated in 60–70% of breast cancers [10, 11] or a promoter that is hypermethylated in 70% of breast cancer individuals [12]. Methylated RASSF1A is strongly associated with metastasis, tumor size, and an increased risk of death [13]. BRCA is a well known tumor suppressor for both breast and ovarian cancer whose mutations are more likely to be higher grade, poorly differentiated, highly proliferative, ER negative, PR negative and harbor p53 mutations [14, 15]. However, Xu et al. found that although methylation of the BRCA1 promoter was more prevalent in cancers with tumors size greater than 2 cm, hypermethylation of BRCA1 from breast cancers with BRCA1 mutations had no overall correlation with ER, PR or grade [16]. Aberrant DNA methylation may be correlated with more advanced tumor stages at the time of diagnosis, but it is independent of BRCA1 mutation. Furthermore, DNA methylation has several advantages over sequence mutations as a cancer biomarker [17]. First, the aberrant methylation of specific CpG islands or gene promoters is more frequent than mutations. Second, aberrant methylation patterns can be detected even when they are embedded in an excess amount of normal DNA molecules. Third, techniques for the detection of methylation patterns are relatively simple [18].

Breast cancers can have different treatment responses and overall outcomes even when they are at the same stage of the disease or have the same subtype. Therefore, a good prognostic biomarker of breast cancer can not only contribute to the accurate classification of the subtype but also guide clinical treatment and improve breast cancer outcomes. Signatures predicting the clinical outcomes of breast cancer based on gene expression profiling have been identified and will provide benefits for adjuvant therapy [19]. However, the identification of prognostic predictors in breast cancer that regulate gene expression may have more benefits than unstable gene expression. For example, aberrant methylation of the TSC [20], SFRP1 [21], and RASSF1A [22] genes was associated with an unfavorable prognosis of breast cancer and could be regarded as independent predictors. Although several methylation biomarkers have been identified to predict breast cancer survival, they are usually limited to average methylation levels of several genes based on experiential knowledge. However, there is a weak correlation between the average DNA methylation levels of gene promoter and gene expressions in genome wide [23]. This finding prompted us to hypothesize that methylated CpGs in promoters might be not have equivalent regulatory effects on gene expression. Here, we used canonical correlation analysis to obtain the methylation patterns of CpGs with the strongest correlation with gene expression, and identified predictors of breast cancer prognosis based on high throughput DNA methylation data. The methylated features showed a good distinction of breast cancer outcomes even in samples with the same receptor status.

Materials and Methods

Data downloading and processing

DNA methylation data sets of breast cancers based on Human Infinium 450K arrays were obtained from TCGA (http://cancergenome.nih.gov/). Gene expression data sets of breast cancers based on AgilentG4502A_07 were also downloaded from TCGA. Breast cancer samples that had both DNA methylation and gene expression information were retained as the training set. Other samples with only DNA methylation and patient outcome information were regarded as the test set. Samples or CpG sites missing data in the training set were filtered in the following analysis. Additionally, two DNA methylation datasets based on Human Infinium 450K arrays (GSE37754) and Human Infinium 27K arrays (GSE20712) were downloaded from GEO database (http://www.ncbi.nlm.nih.gov/geo/) as independent test sets to assess the predictive power of the DNA methylation biomarkers.

Canonical correlation analysis

Due to the multiple CpGs located in the gene promoters and the variability of the DNA methylation levels of multiple CpGs located in the same gene, canonical correlation analysis was used to estimate the correlation between gene expression and DNA methylation levels from multiple CpGs. Let $Y^i = \{y_1^i, y_2^i, \ldots, y_n^i\}$ denote the expression levels of the i*th* gene among *n* samples, and $X^i = \{X_1^i, X_2^i, \ldots, X_p^i\}$ denote the DNA methylation levels of *p* CpGs from the i*th* gene, where $X_j^i = \{x_{1j}^i, x_{2j}^i, \ldots, x_{nj}^i\}^t$ is the methylation levels of the j*th* CpG among n samples. For each gene, we can describe the methylation and expression data using the following matrix.

$$D = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{bmatrix}_{n\times(p+1)}$$

Assuming the mean and covariance of X and Y were

$$E(X) = \bar{X} \text{ and } Cov(X) = \Sigma_{XX}$$

$$E(Y) = \bar{Y} \text{ and } Cov(Y) = \Sigma_{YY}$$

The corresponding covariance matrix was Σ.

$$\sum = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

U and V were the linear combination of X_j and Y respectively.

$$U = \alpha_1 \cdot X_1 + \alpha_2 \cdot X_2 + \cdots \alpha_p \cdot X_p = \alpha' \cdot X$$
$$V = \beta \cdot Y$$

Thus

$$E(U) = \alpha' \cdot \bar{X}$$

$$E(V) = \beta \cdot \bar{Y}$$

$$\sum_{UU} = \alpha' \sum_{XX} \alpha$$

$$\sum_{UV} = \sum_{VU} = \alpha' \sum_{XY} \beta$$

$$\sum_{VV} = \beta \sum_{YY} \beta$$

The objective function that obtained the maximum correlation between U and V was solved based on the Lagrange multiplier method.

$$\begin{cases} \max\left(r(U,V) = \dfrac{\alpha' \sum_{XY} \beta}{\sqrt{\alpha' \sum_{XX} \alpha} \cdot \sqrt{\beta' \sum_{YY} \beta}} \right) \\ s.t. \sum_{UU} = \sum_{VV} = 1 \end{cases}$$

The estimated value $\hat{\alpha}$ and $\hat{\beta}$ constituted the canonical variables U and V that obtained the maximum correlation. Therefore, the methylation pattern of multiple CpGs located in the gene promoter was represented by U which was estimated through the following equation.

$$U = \hat{\alpha}_1 \cdot X_1 + \hat{\alpha}_2 \cdot X_2 + \cdots \hat{\alpha}_p \cdot X_p$$

The expression pattern was represented by V which was estimated through the following equation.

$$V = \hat{\beta} \cdot Y$$

where $\hat{\beta}$ denoted the regulatory direction of methylation on expression. A $\hat{\beta}$ less than 0 indicated negative regulation; conversely, a $\hat{\beta}$ greater than 0 indicated positive regulation. Additionally, $r(U,V)$ was the correlation degree between the methylation pattern and expression pattern.

For each gene, the methylation pattern score (MPS) was defined as

$$MPS_k = \hat{\alpha}_1 \cdot x_{k1} + \hat{\alpha}_2 \cdot x_{k2} + \cdots \hat{\alpha}_p \cdot x_{kp}$$

where, x_{kj} was the DNA methylation levels of the j*th* CpG in the k*th* sample.

Survival analysis

The log-rank test was used to identify a subset of genes whose MPS showed significant differences between the high and low groups and to obtain a *p* value. Univariate and multivariable analyses were performed using Cox proportional hazards models incorporating MPS and known prognostic clinical factors, including age at diagnosis ($\leq$55 vs $\geq$56 years), tumor pathological stage (I & II vs III & IV), and tumor size (1–2 vs 3–4) as categorical variables. Univariate Cox regression analysis was performed to assess the survival prognosis capabilities of the selected gene set using the overall survival time as a dependent variable. To create an optimal feature for genes based on methylation patterns to assess breast cancer outcomes, the methylation risk score (MRS) was defined through featured genes identified based on multivariable Cox proportional hazards models with the MPS as a continuous variable.

$$MRS_k = c_1 \cdot gene_{1k} + \cdots + c_m \cdot gene_{mk}$$

where k was the k*th* sample. m was the number of feature genes filtered by the multivariable Cox proportional hazards models and c_j (j = 1, 2, . . ., m) was the coefficient estimated by the multivariable Cox proportional hazards models. The 5-year overall survival for each MRS scoring group (high vs low) was calculated using the Kaplan–Meier method, and the statistical significance was assessed using the log-rank test. The significance level of all statistical tests was 0.05.

Multiple test correction

To control the false discover rate (FDR) of the featured genes, we adopted two methods to correct the p value of the statistical test. For the identification of the gene subset based on MPS through the log-rank test, sample labels were permutated 100 times and the log-rank test was re-performed. Empirical p values were calculated according to the order of the observed p value among the 100 permutations. If the empirical p value was less than 0.01, the gene was retained. Thus, all p values from the 100 permutations were larger than the observed p value. To obtain the significant candidate gene set with the univariate Cox proportional hazards models, we again permuted the sample labels 100 times. In each permutation, we obtained the p-value of the univariate Cox proportional hazards models. According to the FDR equation, the cutoff of p (p_0) was determined through FDR = 0.01.

$$FDR = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} \#(p < p_0)_{permutation_i}}{\#(p < p_0)_{observed}}$$

where the numerator was the expected number of genes whose p value from the univariate Cox proportional hazards models was random less than cut off p_0 in random and the denominator was the number of genes whose p value was less than p_0 in the real situation.

Results

Identification of featured biomarkers based on the effect of DNA methylation regulatory patterns of CpGs on expression

The gene expression and DNA methylation data from 209 breast cancer patients were obtained after processing the missing data from TCGA, which included 15,801 genes and 281,066 CpGs located in gene promoters. The breast cancer details are shown in S1 Table. The methylation patterns of CpGs in gene promoters that showed the maximum correlation with the gene expression patterns were obtained through canonical correlation analysis, which measured the maximum regulatory effect of DNA methylation on gene expression. Pearson's correlation analysis was performed between the average methylation of CpGs and gene expression. However, the canonical correlation degree was significantly higher than the Pearson's correlation degree (Wilcoxon p value $< 2.2 \times 10^{-16}$, S1A Fig). Moreover, we found that DNA methylation had a negative regulatory effect on the expression of some genes and a positive regulatory effect on others (S1B Fig).

The MPS was calculated through the canonical variable from the canonical correlation analysis to estimate the methylation patterns of CpGs. High- and low-MPS groups were classified based on the median MPS among the samples. The log-rank test was used to identify genes that showed significant difference in outcomes between the high- and low-MPS groups. The p value of the log-rank test of these genes was less than 0.05 and the lowest of 100 permutations. A gene subset including 151 genes was identified. Furthermore, 38 genes were retained through

Table 1. Multivariate Cox proportional hazard model of risk gene set.

Gene	Coef	HR	*p* value	95% CI
PFN1	-54.0	3.60×10^{-24}	0.002025	[4.67×10^{-39}, 2.78×10^{-9}]
ZFAND5	-45.9	1.23×10^{-20}	0.002271	[2.01×10^{-33}, 7.48×10^{-8}]
TMEM184B	-42.5	3.35×10^{-19}	0.008188	[6.76×10^{-33}, 1.66×10^{-5}]
PPP1R12C	-41.2	1.24×10^{-18}	0.000633	[6.64×10^{-29}, 2.31×10^{-8}]
NRIP2	-40.6	2.45×10^{-18}	0.000455	[3.50×10^{-28}, 1.71×10^{-8}]
DPAGT1	-34.2	1.36×10^{-15}	0.001741	[6.71×10^{-25}, 2.75×10^{-6}]
HPX	-29.2	2.06×10^{-13}	0.002975	[8.77×10^{-22}, 4.84×10^{-5}]
HERPUD2	27.9	1.32×10^{12}	0.009686	[8.65×10^{2}, 2.01×10^{21}]
ZNF592	30.5	1.83×10^{13}	0.00128	[1.55×10^{5}, 2.17×10^{21}]
ZNF536	51.4	2.03×10^{22}	0.000563	[4.27×10^{9}, 9.60×10^{34}]
SLC25A21	51.6	2.52×10^{22}	0.008259	[6.01×10^{5}, 1.06×10^{39}]
NARS	53.3	1.44×10^{23}	0.004414	[1.64×10^{7}, 1.26×10^{39}]

Abbreviations: Coef is the Cox proportional hazard model regression coefficient. HR: hazard ratio; CI: confidence interval; p value: cox regression model p value.

doi:10.1371/journal.pone.0142279.t001

univariate Cox proportional hazards models with $p < 0.05$ and FDR = 0.01 among 100 permutations (S2 Table).

The methylation prognostic biomarkers of breast cancer were identified through multivariable Cox proportional hazards models based on 38 genes with $p < 0.01$. The methylation prognostic biomarkers consisted of 7 protective genes with negative Cox proportional hazard model regression coefficient, indicating that the survival time increased as the MPS increased, and 5 risk genes with positive coefficients, indicating that the MPS increased as the survival time decreased (Table 1).

Estimating breast cancer outcomes according to the methylation risk score

We assessed the MRS of each sample based on the 12 methylation biomarkers described as follows to predict the breast cancer outcomes.

$$MRS = 51.36 \cdot ZNF536 + 53.32 \cdot NARS + 30.54 \cdot ZNF592 + 51.58 \cdot SLC25A21 + 27.91 \cdot HERPUD2$$
$$-53.98 \cdot PFN1 - 45.85 \cdot ZFAND5 - 40.55 \cdot NRIP2 - 29.21 \cdot HPX - 34.23 \cdot DPAGT1$$
$$-42.54 \cdot TMEM184B - 41.23 \cdot PPP1E12C$$

The Kaplan–Meier method was used to estimate the 5-year overall survival rate between the high- and low-MRS groups classified through the median MRS among the samples. As shown in Fig 1A, the high-MRS group had a significantly shorter 5-year overall survival rate than the low-MRS group (68.5% vs 94.1%, log-rank $p = 0.00103$). The average methylation and expression levels of 12 featured genes were adopted to perform survival analysis. No significant difference were detected between the high and low groups classified through the median methylation or expression levels (Fig 1B and 1C). Additionally, we used the 7 protective and 5 risk genes for the survival analysis. The overall survival time and 5-year survival rates were significant different between the high and low groups; the 7 protective genes especially allowed more distinct estimations (S2 Fig).

MRS had a significant association with hazard ratio (HR) of death in both the univariate and multivariable Cox proportional hazards models compared with the prognostic clinical

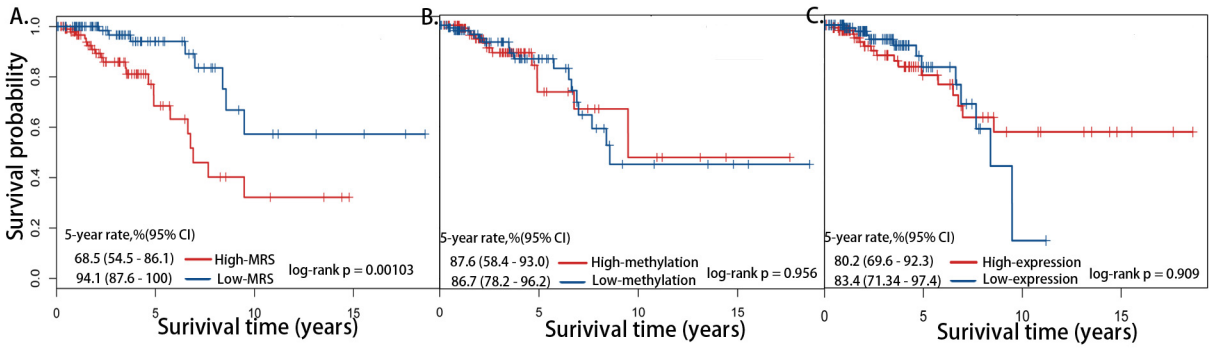

Fig 1. Kaplan-Meier survival analysis of overall survival of 209 breast patients based on feature genes. (A) MRS. (B) Average DNA methylation levels. (C) Average gene expression levels.

doi:10.1371/journal.pone.0142279.g001

factors, including age at diagnosis (≤55 vs ≥56 years), tumor pathological stage (I & II vs III & IV), and TNM (1–2 vs 3–4) (S3 Table). This result suggested that the regulatory effect of the DNA methylation pattern on expression might be better able to predict the outcome of breast cancer than depending on methylation or expression alone.

Differential outcomes of receptor states with differential MRS

High heterogeneity and differential outcomes of breast cancer were found among patients from the same subtype or receptor status. However, no significant differences in the overall survival time and 5-year survival rate were found between ER+ and ER-, PR+ and PR-, Her2 + and Her2- patients (Fig 2A–2C). When we combined these factors with the MRS, found a significant difference between the positive and negative receptor status and the breast cancer outcome. For example, ER- & high-MRS samples resulted in the lowest 5-year survival rat, whereas ER+ & low-MRS samples had the highest 5-year survival rate (Fig 2D). Similarly, PR- & high-MRS and Her2- & high-MRS samples exhibited the lowest 5-year survival rates, whereas PR+ & low-MRS and Her2+ & low-MRS samples reached the highest 5-year survival rates (Fig 2E and 2F). Significantly differential outcomes were observed between the high- and low-MRS groups, although the patients had the same receptor state (Fig 2G–2L). In conclusion, differential outcomes were observed when the MRS and receptor status were combined, even when the patients had the same receptor state. We found that ER-, PR- or Her2- patients with high-MRS had the worst outcomes, which was consistent with the known conclusion.

Application of the methylation risk feature on breast cancers that resulted in death state and independent test datasets

To validate the prognostic value of DNA methylation biomarkers on other breast samples, the MRS of 12 featured genes were applied to 28 breast samples from patients with a death outcome from the TCGA dataset that only had DNA methylation information and were not included in the previous dataset. We found a significantly differential survival time between the low- and high-MRS groups (Fig 3A). Moreover, the overall survival time of the high-MRS group was less than 5 years. MRS showed a good ability to distinguish outcomes between ER + and ER-, PR+ and PR-, and Her2+ and Her2- samples (Fig 3B–3D). We observed similar results with the training set, with ER-, PR- or Her2- samples with high-MRS showing the worst prognosis; indeed, less than one year of survival was estimated for the Her2- & high-MRS group. The effect of the prognostic outcome in the independent samples suggested that a DNA

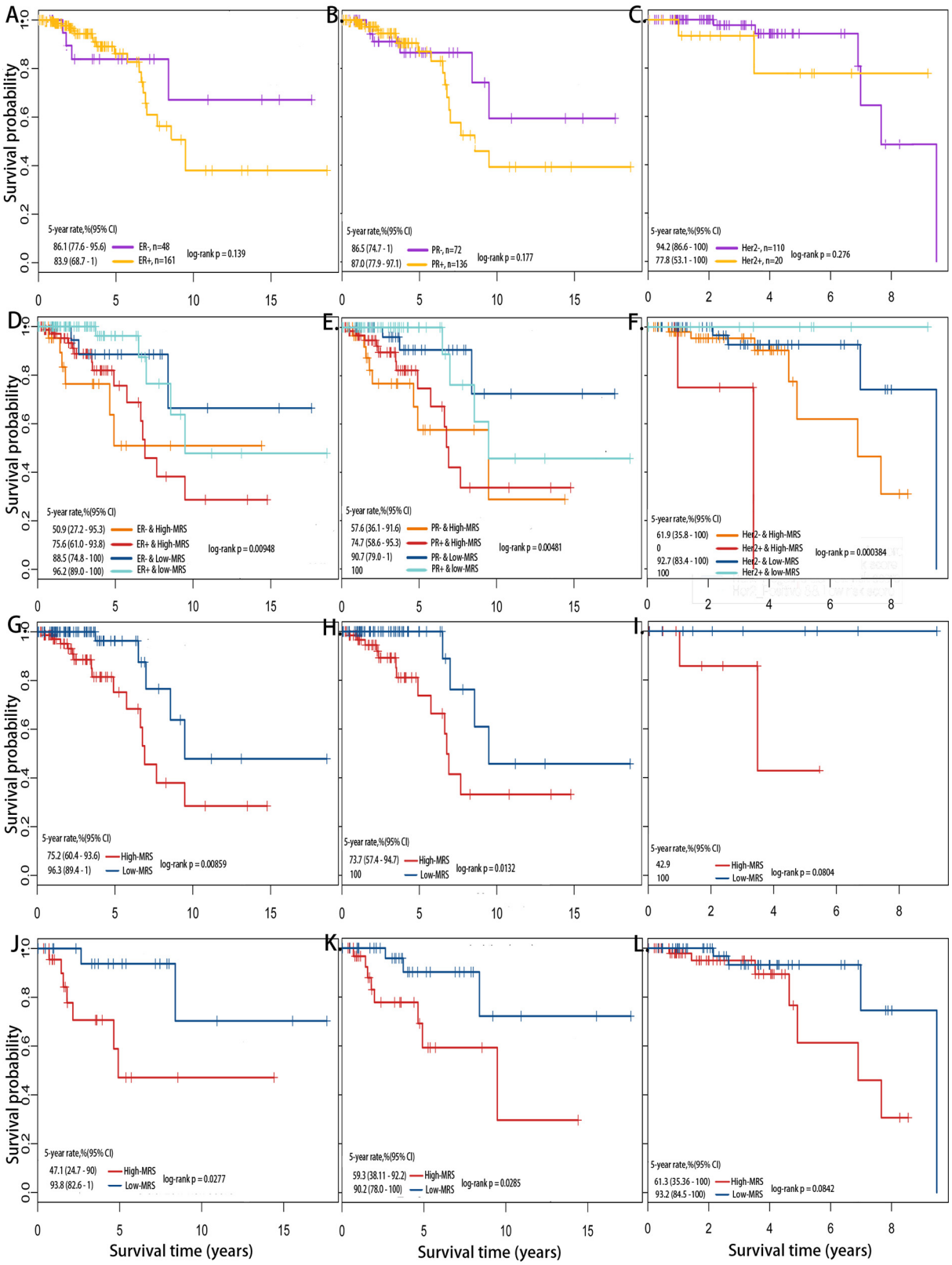
A
Survival probability
5-year rate,%(95% CI)
86.1 (77.6 - 95.6) ER-, n=48
83.9 (68.7 - 1) ER+, n=161
log-rank p = 0.139
B
5-year rate,%(95% CI)
86.5 (74.7 - 1) PR-, n=72
87.0 (77.9 - 97.1) PR+, n=136
log-rank p = 0.177
C
5-year rate,%(95% CI)
94.2 (86.6 - 100) Her2-, n=110
77.8 (53.1 - 100) Her2+, n=20
log-rank p = 0.276
D
5-year rate,%(95% CI)
50.9 (27.2 - 95.3) ER- & High-MRS
75.6 (61.0 - 93.8) ER+ & High-MRS
88.5 (74.8 - 100) ER- & Low-MRS
96.2 (89.0 - 100) ER+ & low-MRS
log-rank p = 0.00948
E
5-year rate,%(95% CI)
57.6 (36.1 - 91.6) PR- & High-MRS
74.7 (58.6 - 95.3) PR+ & High-MRS
90.7 (79.0 - 1) PR- & Low-MRS
100 PR+ & low-MRS
log-rank p = 0.00481
F
5-year rate,%(95% CI)
61.9 (35.8 - 100) Her2- & High-MRS
0 Her2+ & High-MRS
92.7 (83.4 - 100) Her2- & Low-MRS
100 Her2+ & low-MRS
log-rank p = 0.000384
G
5-year rate,%(95% CI)
75.2 (60.4 - 93.6) High-MRS
96.3 (89.4 - 1) Low-MRS
log-rank p = 0.00859
H
5-year rate,%(95% CI)
73.7 (57.4 - 94.7) High-MRS
100 Low-MRS
log-rank p = 0.0132
I
5-year rate,%(95% CI)
42.9 High-MRS
100 Low-MRS
log-rank p = 0.0804
J
5-year rate,%(95% CI)
47.1 (24.7 - 90) High-MRS
93.8 (82.6 - 1) Low-MRS
log-rank p = 0.0277
K
5-year rate,%(95% CI)
59.3 (38.11 - 92.2) High-MRS
90.2 (78.0 - 100) Low-MRS
log-rank p = 0.0285
L
5-year rate,%(95% CI)
61.3 (35.36 - 100) High-MRS
93.2 (84.5 -100) Low-MRS
log-rank p = 0.0842
Survival time (years)
Survival time (years)
Survival time (years)

Fig 2. Kaplan-Meier survival analysis of overall survival for receptor status. (A) Survival comparison between ER+ and ER- patients. (B) Survival comparison between PR+ and PR- patients. (C) Survival comparison between Her2+ and Her2- patients. (D) Survival comparison through combination of ER states and MRS. (E) Survival comparison through combination of PR states and MRS. (F) Survival comparison through combination of Her2 states and MRS. (G) Survival comparison between high-MRS and low-MRS from ER+. (H) Survival comparison between high-MRS and low-MRS from PR+. (I) Survival comparison between high-MRS and low-MRS groups from Her2+. (J) Survival comparison between high-MRS and low-MRS groups from ER-. (K) Survival comparison between high-MRS and low-MRS groups from PR-. (L) Survival comparison between high-MRS and low-MRS groups from Her2-.

doi:10.1371/journal.pone.0142279.g002

methylation biomarker identified through the regulation of gene expression by DNA methylation patterns in CpGs, is believable and has good predictive value for breast cancer outcomes.

Additionally, we evaluated the performance of 12 featured genes using two independent breast cancer datasets (GSE37754 and GSE20712). NARS and SLC25A21 were not included in the 27K DNA methylation dataset (GSE20712), and an additional 10 featured genes were used. All 12 featured genes were used in the 450K dataset (GSE37754). Using two independent cohorts, we found differential outcomes between the high- and low-MRS groups (Fig 4); p values of the log-rank test were especially significant in GSE20712 (Fig 4B). MRS showed a preferable distinguishing power between the high- and low-MRS groups. This finding suggested that DNA methylation biomarkers might be robust factors for the prediction of breast cancer outcomes.

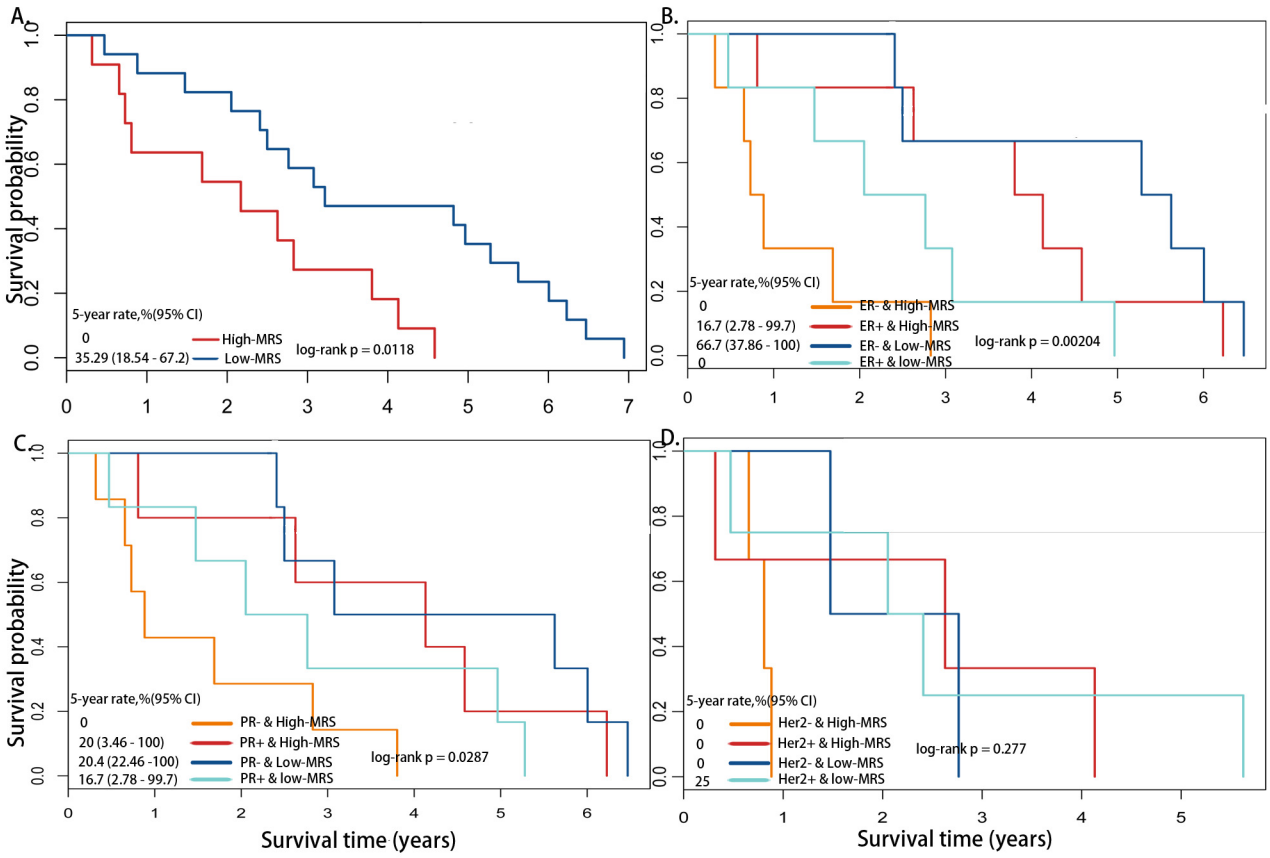

Fig 3. Kaplan-Meier survival analysis of overall survival on 28 patients with death outcome. (A) Survival comparison between High-MRS and Low-MRS groups. (B) Survival comparison among ER+/- patients. (C) Survival comparison among PR+/- patients. (D) Survival comparison among Her2+/- patients.

doi:10.1371/journal.pone.0142279.g003

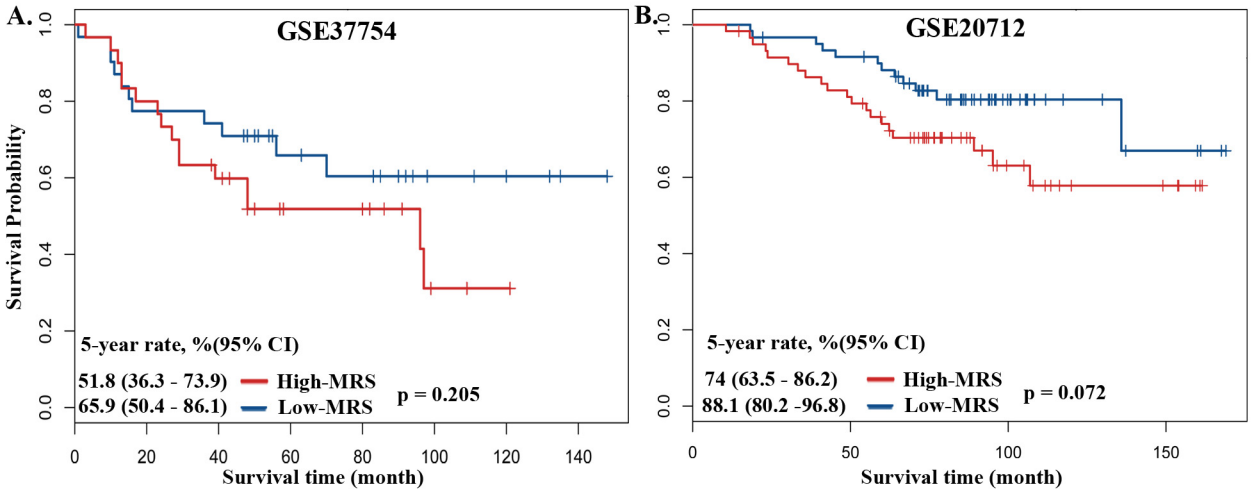

Fig 4. The Kaplan-Meier survival analysis of overall survival on four independent dataset from GEO database. (A) GSE37754 from 450K arrays. (B) GSE20712 from 27K arrays.

doi:10.1371/journal.pone.0142279.g004

Discussion

Breast cancer is a heterogeneous tumor. Molecular classification has been successfully used to design individualized therapies, leading to significant improvements in disease-specific survival [24]. Recently, breast cancer was classified into three major subtypes based on luminal, Her2 + and basal-like based gene expression profiling [25, 26]. Moreover, DNA methylation showed distinct patterns among subtypes of breast cancer (especially between luminal B and basal-like) [27]. Each of the breast cancer subtypes has different risk factors for incidence, response to treatment, risk of disease progression and outcomes. For example, triple negative breast cancer which usually includes basal-like tumors that lack HR and Her2, has a worse outcomes than the other subtypes because no specific molecular targets have been identified [28]. Recently, the identification of differential methylated regions of triple-negative breast cancer based on whole-genome methylation sequencing has provided diagnostic and prognostic value for personalized management [29].

We identified DNA methylation patterns of CpGs located within gene promoters that had a maximal regulatory effect on gene expression based on canonical correlation analysis to define the methylation pattern score of adjacent CpGs. DNA methylation featured genes associated with breast cancer outcomes that were obtained according to the DNA methylation patterns, which contributed to the construction of the methylation risk score. The methylation risk score of the featured genes showed the best ability to estimate the survival time between the high and low-risk groups compared to average DNA methylation or gene expression. Moreover, we found significant differential outcomes between the high- and low-MRS groups even though the breast samples had the same HR or Her2 status (especially ER-, PR- or Her2-), with the high-MRS group having the worst outcomes. A similar conclusion was supported using 28 breast samples with death as an outcome. We evaluated the estimation ability of the DNA methylation pattern of featured genes in other breast cancers based on Human Infinium 450K and 27K arrays from the GEO database. Due to the absence of some CpGs and genes in the 27K arrays compared with the 450K arrays, we used shared featured genes and CpGs between the 450K and 27K datasets to measure the MRS. Only 10 featured genes were found (excluding NARS and SLC25A21) in the 27K datasets; these genes were used to predict breast cancer

outcomes. We also found differential outcomes between the high- and low-MRS groups in two independent cohorts. However, the p values of the log-rank test from the 450K test sets were not significant. We speculate that the high heterogeneity in cancer may lead to differential outcomes. Notably, the breast samples from GSE37754 (450K arrays) were in the early stages of cancer with tumor pathological stages I or II, which might have caused the non-significant p-value. Although differences in the platform between the 27K and 450K arrays led to the absence of some featured genes and CpGs, the p value of the log-rank test was significant and the breast cancer outcomes were clearly distinguished between the high- and low-MRS groups. This finding implied that the survival model based on the methylation patterns of featured genes showed a good survival prognostic capability in breast cancer.

The DNA methylation featured genes we identified were ZNF536, ZNF592, NARS, SLC25A21, HERPUD2, NRIP2, PPP1R12C, DPAGT1, PFN1, ZFAND5, HPX and TMEM184B. ZNF536, ZNF592 and ZFAND5 were zinc finger proteins that had important roles in transcript regulation, embryonic development and cell differentiation. Several studies reported that SLC25A21 [30], PFN1 [31], HPX [32]and TMEM184B [33] were associated with a risk of breast cancer. Additionally, other genes were associated with the risk of other diseases, such as NARS, which causes nonsyndromic hearing loss, Leigh syndrome [34] and Alpers syndrome [35], and DPAGT1, which is involved in the pathogenesis of oral cancer [36]. The methylation featured genes did not include the BRCA gene. This finding was consistent with the conclusion of Xu et al, who reported that the methylation of BRCA had no overall correlation with ER, PR or grade. These results suggest that DNA methylation is an independent predictor of breast cancer prognosis and is independent of BRCA1 mutation. We attempted to compare the survival time between the BRCA mutation and non-mutation samples through MRS, but the analysis was limited by the small sample number with BRCA mutations. The good prognostic power of DNA methylation biomarkers can help guide clinical treatment and predict the outcome of breast cancer.

Supporting Information

S1 Fig. Feature of correlation between DNA methylation pattern and gene expression based on canonical correlation analysis. (A) Comparison of canonical and pearson's correlation coefficient. (A) Regulatory effect of DNA methylation pattern on gene expression based on canonical correlation analysis.
(TIF)

S2 Fig. Kaplan-Meier survival analysis of overall survival of 209 breast patients. (A) Protected genes. (B) Risk genes.
(TIF)

S1 Table. Information of Breast invasive carcinome patients from TCGA.
(DOC)

S2 Table. Results of univariate Cox proportional hazard model about significant gene set.
(DOC)

S3 Table. Cox proportional hazards analyses using different predictors.
(DOC)

Acknowledgments

We thank the support of National Natural Science Foundation of China [grant number 61402139, 31401075, 31371334, 61403112, 61203262], the Natural Science Foundation of

Heilongjiang Province [grant number ZD2015003] and the Natural Scientific Research Fund of Heilongjiang Provincial [grant number QC2011C061 to SZ].

Author Contributions

Conceived and designed the experiments: FW YZ. Performed the experiments: MZ SZ. Analyzed the data: MZ Y. Wen Y. Wei. Contributed reagents/materials/analysis tools: Y. Wang HL DZ JS. Wrote the paper: SZ HL FW YZ.

References

1. DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. 2011 Nov-Dec; 61(6):409–18. doi: 10.3322/caac.20134 PMID: 21969133
2. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin. 2010 Sep-Oct; 60(5):277–300. doi: 10.3322/caac.20073 PMID: 20610543
3. Kohler BA, Sherman RL, Howlader N, Jemal A, Ryerson AB, Henry KA, et al. Annual Report to the Nation on the Status of Cancer, 1975–2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State. J Natl Cancer Inst. 2015 Jun; 107(6):djv048. doi: 10.1093/jnci/djv048 PMID: 25825511
4. Anderson WF, Katki HA, Rosenberg PS. Incidence of breast cancer in the United States: current and future trends. J Natl Cancer Inst. 2011 Sep 21; 103(18):1397–402. doi: 10.1093/jnci/djr257 PMID: 21753181
5. Anderson WF, Luo S, Chatterjee N, Rosenberg PS, Matsuno RK, Goodman MT, et al. Human epidermal growth factor receptor-2 and estrogen receptor expression, a demonstration project using the residual tissue repository of the Surveillance, Epidemiology, and End Results (SEER) program. Breast Cancer Res Treat. 2009 Jan; 113(1):189–96. doi: 10.1007/s10549-008-9918-3 PMID: 18256926
6. Kurian AW, Fish K, Shema SJ, Clarke CA. Lifetime risks of specific breast cancer subtypes among women in four racial/ethnic groups. Breast Cancer Res. 2010; 12(6):R99. doi: 10.1186/bcr2780 PMID: 21092082
7. Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? J Natl Cancer Inst. 2014 Aug; 106(8).
8. Polyak K. Heterogeneity in breast cancer. J Clin Invest. 2011 Oct; 121(10):3786–8. doi: 10.1172/JCI60534 PMID: 21965334
9. Daniel J. Weisenberger GL. Contributions of DNA methylation aberrancies in shaping the cancer epigenome. Translational Cancer Research. 2015; 4(3):219–34.
10. Campan M, Weisenberger DJ, Laird PW. DNA methylation profiles of female steroid hormone-driven human malignancies. Curr Top Microbiol Immunol. 2006; 310:141–78. PMID: 16909910
11. Sui M, Huang Y, Park BH, Davidson NE, Fan W. Estrogen receptor alpha mediates breast cancer cell resistance to paclitaxel through inhibition of apoptotic cell death. Cancer Res. 2007 Jun 1; 67(11):5337–44. PMID: 17545614
12. Lewis CM, Cler LR, Bu DW, Zochbauer-Muller S, Milchgrub S, Naftalis EZ, et al. Promoter hypermethylation in benign breast epithelium in relation to predicted breast cancer risk. Clin Cancer Res. 2005 Jan 1; 11(1):166–72. PMID: 15671542
13. Fiegl H, Millinger S, Mueller-Holzner E, Marth C, Ensinger C, Berger A, et al. Circulating tumor-specific DNA: a marker for monitoring efficacy of adjuvant therapy in cancer patients. Cancer Res. 2005 Feb 15; 65(4):1141–5. PMID: 15734995
14. Mirza S, Sharma G, Prasad CP, Parshad R, Srivastava A, Gupta SD, et al. Promoter hypermethylation of TMS1, BRCA1, ERalpha and PRB in serum and tumor DNA of invasive ductal breast carcinoma patients. Life Sci. 2007 Jul 4; 81(4):280–7. PMID: 17599361
15. Vincent-Salomon A, Ganem-Elbaz C, Manie E, Raynal V, Sastre-Garau X, Stoppa-Lyonnet D, et al. X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors. Cancer Res. 2007 Jun 1; 67(11):5134–40. PMID: 17545591
16. Honrado E, Osorio A, Milne RL, Paz MF, Melchor L, Cascon A, et al. Immunohistochemical classification of non-BRCA1/2 tumors identifies different groups that demonstrate the heterogeneity of BRCAX families. Mod Pathol. 2007 Dec; 20(12):1298–306. PMID: 17885670
17. Dworkin AM, Huang TH, Toland AE. Epigenetic alterations in the breast: Implications for breast cancer detection, prognosis and treatment. Semin Cancer Biol. 2009 Jun; 19(3):165–71. doi: 10.1016/j.semcancer.2009.02.007 PMID: 19429480

18. Wajed SA, Laird PW, DeMeester TR. DNA methylation: an alternative pathway to cancer. Ann Surg. 2001 Jul; 234(1):10–20. PMID: 11420478

19. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31; 415(6871):530–6. PMID: 11823860

20. Jiang WG, Sampson J, Martin TA, Lee-Jones L, Watkins G, Douglas-Jones A, et al. Tuberin and hamartin are aberrantly expressed and linked to clinical outcome in human breast cancer: the role of promoter methylation of TSC genes. Eur J Cancer. 2005 Jul; 41(11):1628–36. PMID: 15951164

21. Veeck J, Niederacher D, An H, Klopocki E, Wiesmann F, Betz B, et al. Aberrant methylation of the Wnt antagonist SFRP1 in breast cancer is associated with unfavourable prognosis. Oncogene. 2006 Jun 8; 25(24):3479–88. PMID: 16449975

22. Muller HM, Widschwendter A, Fiegl H, Ivarsson L, Goebel G, Perkmann E, et al. DNA methylation in serum of breast cancer patients: an independent prognostic marker. Cancer Res. 2003 Nov 15; 63 (22):7641–5. PMID: 14633683

23. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015 Jun 1.

24. Perez EA. Breast cancer management: opportunities and barriers to an individualized approach. Oncologist. 2011; 16 Suppl 1:20–2. doi: 10.1634/theoncologist.2011-S1-20 PMID: 21278437

25. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001 Sep 11; 98(19):10869–74. PMID: 11553815

26. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug 17; 406(6797):747–52. PMID: 10963602

27. Stefansson OA, Moran S, Gomez A, Sayols S, Arribas-Jorba C, Sandoval J, et al. A DNA methylation-based definition of biologically distinct breast cancer subtypes. Mol Oncol. 2015 Mar; 9(3):555–68. doi: 10.1016/j.molonc.2014.10.012 PMID: 25468711

28. Silver DP, Richardson AL, Eklund AC, Wang ZC, Szallasi Z, Li Q, et al. Efficacy of neoadjuvant Cisplatin in triple-negative breast cancer. J Clin Oncol. 2010 Mar 1; 28(7):1145–53. doi: 10.1200/JCO.2009.22.4725 PMID: 20100965

29. Clare Stirzaker EZSJC. Genome-wide DNA methylation profiling in triple negative breast cancer reveals epigenetic signatures with important clinical value. Molecular & Cellular Oncology. 2015.

30. Rudolph A, Hein R, Lindstrom S, Beckmann L, Behrens S, Liu J, et al. Genetic modifiers of menopausal hormone replacement therapy and breast cancer risk: a genome-wide interaction study. Endocr Relat Cancer. 2013 Dec; 20(6):875–87. doi: 10.1530/ERC-13-0349 PMID: 24080446

31. Gau DM, Lesnock JL, Hood BL, Bhargava R, Sun M, Darcy K, et al. BRCA1 deficiency in ovarian cancer is associated with alteration in expression of several key regulators of cell motility—A proteomics study. Cell Cycle. 2015 Jun 18; 14(12):1884–92. doi: 10.1080/15384101.2015.1036203 PMID: 25927284

32. Cine N, Baykal AT, Sunnetci D, Canturk Z, Serhatli M, Savli H. Identification of ApoA1, HPX and POTEE genes by omic analysis in breast cancer. Oncol Rep. 2014 Sep; 32(3):1078–86. doi: 10.3892/or.2014.3277 PMID: 24969553

33. Lindstrom S, Thompson DJ, Paterson AD, Li J, Gierach GL, Scott C, et al. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. Nat Commun. 2014; 5:5303. doi: 10.1038/ncomms6303 PMID: 25342443

34. Simon M, Richard EM, Wang X, Shahzad M, Huang VH, Qaiser TA, et al. Mutations of human NARS2, encoding the mitochondrial asparaginyl-tRNA synthetase, cause nonsyndromic deafness and Leigh syndrome. PLoS Genet. 2015 Mar; 11(3):e1005097. doi: 10.1371/journal.pgen.1005097 PMID: 25807530

35. Sofou K, Kollberg G, Holmstrom M, Davila M, Darin N, Gustafsson CM, et al. Whole exome sequencing reveals mutations in NARS2 and PARS2, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. Mol Genet Genomic Med. 2015 Jan; 3 (1):59–68. doi: 10.1002/mgg3.115 PMID: 25629079

36. Varelas X, Bouchie MP, Kukuruzinska MA. Protein N-glycosylation in oral cancer: dysregulated cellular networks among DPAGT1, E-cadherin adhesion and canonical Wnt signaling. Glycobiology. 2014 Jul; 24(7):579–91. doi: 10.1093/glycob/cwu031 PMID: 24742667