IDENTIFICATION OF PROGNOSTICALLY SIGNIFICANT DNA METHYLATION SIGNATURES IN PATIENTS WITH VARIOUS BREAST CANCER TYPES

Kalinkin AI[1] ✉, Sigin VO[1], Nemtsova MV[1,2], Strelnikov VV[1,3]

[1] Research Centre of Medical Genetics, Moscow, Russia

[2] Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

[3] Pirogov Russian National Research Medical University, Moscow, Russia

Breast cancer (BC) is the most frequently diagnosed cancer and one of the major causes of female mortality. The development of prognostic models based on multiomics data is the main goal of precision oncology. Aberrant DNA methylation in BC is a diagnostic marker of carcinogenesis. Despite the existing factors of BC prognosis, introduction of methylation markers would make it possible to obtain more accurate prognostic scores. The study was aimed to assess DNA methylation signatures in various BC subtypes for clinical endpoints and patients' clinicopathological characteristics. The data on methylation of CpG dinucleotides (probes) and clinical characteristics of BC samples were obtained from The Cancer Genome Atlas Breast Cancer database. CpG dinucleotides associated with the selected endpoints were chosen by univariate Cox regression method. The LASSO method was used to search for stable probes, while further signature construction and testing of the clinical characteristics independence were performed using multivariate Cox regression. The dignostic and prognostic potential of the signatures was assessed using ROC analysis and Kaplan–Meier curves. It has been shown that the signatures of selected probes have a significant diagnostic (AUC 0.76–1) and prognostic ($p < 0.05$) potential. This approach has made it possible to identify 47 genes associated with good and poor prognosis, among these five genes have been described earlier. If the genome-wide DNA analysis results are available, the research approach applied can be used to study molecular pathogenesis of BC and other disorders.

Keywords: breast cancer, molecular subtypes, survival analysis, DNA methylation, prognostic markers

Funding: the study was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the Federal Scientific and Technical Program for the Development of Genetic Technologies in 2019–2027 (agreement № 075-15-2021-1073).

Author contribution: Kalinkin AI — study design, data acquisition, analysis and interpretation, manuscript writing; Sigin VO — manuscript writing; Nemtsova MV — study concept and design; Strelnikov VV — study concept and design, scientific editing.

✉ **Correspondence should be addressed:** Alexey I. Kalinkin
Moskvorechye, 1, Moscow, 115522; alexeika2@yandex.ru

Received: 18.10.2022 **Accepted:** 11.11.2022 **Published online:** 25.11.2022

DOI: 10.24075/brsmu.2022.056

ОПРЕДЕЛЕНИЕ ПРОГНОСТИЧЕСКИ ЗНАЧИМОЙ СИГНАТУРЫ ДНК-МЕТИЛИРОВАНИЯ У ПАЦИЕНТОК С РАЗЛИЧНЫМИ ТИПАМИ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

А. И. Калинкин[1] ✉, В. О. Сигин[1], М. В. Немцова[1,2], В. В. Стрельников[1,3]

[1] Медико-генетический научный центр имени Н. П. Бочкова, Москва, Россия

[2] Первый Московский государственный медицинский университет имени И. М. Сеченова (Сеченовский университет), Москва, Россия

[3] Российский национальный исследовательский медицинский университет имени Н. И. Пирогова, Москва, Россия

Рак молочной железы (РМЖ) — наиболее часто диагностируемое онкологическое заболевание и одна из ведущих причин смертности среди женского населения. Разработка прогностических моделей с использованием мультиомиксных данных является главной целью прецизионной онкологии. Аберрантное метилирование ДНК в РМЖ представляет собой информативный маркер канцерогенеза. Несмотря на существующие факторы прогноза РМЖ, введение маркеров метилирования позволит получать более точную прогностическую оценку. Целью работы было изучить сигнатуры метилирования ДНК в различных подтипах РМЖ для клинических конечных точек и клинико-патологических характеристик пациенток. Данные об уровнях метилирования CpG-динуклеотидов (зондов) и клинические характеристики образцов РМЖ были получены из базы данных The Cancer Genome Atlas Breast Cancer. С помощью метода одномерной регрессии Кокса были выбраны CpG-динуклеотиды, ассоциированные с выбранными конечными точками. Методом LASSO осуществляли поиск стабильных зондов, а дальнейшее построение сигнатур и независимость клинических характеристик выполняли с помощью многофакторной регрессии Кокса. Диагностический и прогностический потенциал сигнатур оценивали с помощью метода ROC-анализа и кривых Каплан–Майера. Показано, что сигнатуры отобранных зондов обладают значимым диагностическим (AUC от 0,76 до 1) и прогностическим ($p < 0,05$) потенциалом. С помощью данного подхода удалось идентифицировать 47 генов, связанных с хорошим и плохим прогнозом, из которых пять уже были описаны ранее. При наличии результатов широкогеномного анализа ДНК примененный исследовательский подход можно использовать для изучения не только молекулярного патогенеза РМЖ, но и для других заболеваний.

Ключевые слова: рак молочной железы, молекулярные подтипы, выживаемость, метилирование ДНК, прогностические маркеры

Финансирование: работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Федеральной научно-технической программы развития генетических технологий на 2019–2027 годы (соглашение № 075-15-2021-1073).

Вклад авторов: А. И. Калинкин — дизайн исследования, сбор, анализ и интерпретация данных, написание статьи; В. О. Сигин — написание статьи; М. В. Немцова — концепция и дизайн исследования; В. В. Стрельников — концепция и дизайн исследования, научное редактирование.

✉ **Для корреспонденции:** Алексей Игоревич Калинкин
ул. Москворечье, д. 1, г. Москва, 115522; alexeika2@yandex.ru

Статья получена: 18.10.2022 **Статья принята к печати:** 11.11.2022 **Опубликована онлайн:** 25.11.2022

DOI: 10.24075/vrgmu.2022.056

According to the Global Cancer Observatory (GLOBOCAN), about 2.3 million of new breast cancer (BC) cases and 684,996 deaths from BC were reported in 2020. BC, being the most common type of cancer all over the world [1], is a highly heterogeneous disease with varying molecular and clinical characteristics [2].

Today, BC subtypes are defined by immunohistochemical (IHC) staining of tumor tissue [3], particularly based on the estrogen, progesterone, HER2 receptor protein expression in the tumor and on the cancer cell proliferation rate. The development of methods for gene expression analysis involving the use of DNA microarrays played a major role in determining BC molecular subtypes. The use of the classifier based on the expression of 50 PAM50 genes makes it possible to clearly distinguish luminal A (LumA), luminal B (LumB), HER2-enriched (HER2+) molecular subtypes, as well as basal-like or triple-negative breast cancer (TNBC) [4]. TNBC that comprises 15–20% of all BC cases is characterized by aggressiveness, high metastasis rate, frequent relapses, and low survival rate compared to other BC subtypes [5]. Multigene microarray-based test systems make it possible to obtain prognostic information that is important for cancer patients, especially in cases of equivocal predictions made based on the clinical characteristics and IHC markers. Such systems include Mammaprint/Blueprint and Prosigna/PAM50, which, in addition to their predictive value, provide the possibility of division into molecular subtypes [6]. These systems can be used to define high or low risk of relapse in female BC patients, however, this option is not yet available for TNBC and HER2+ molecular subtypes due to a lack of clinical trials.

Epigenetic changes modulate genome utilization through histone modification, changes in histone variant composition, chromatin remodeling, DNA methylation, positioning of nucleosomes and non-coding RNAs (expression of specific miRNAs). For the effect to become manifest, all of the above mentioned epigenetic alterations act in concert. DNA methylation is one of the best known factors of gene expression regulation. It occurs due to covalent modification of cytosines through the methyl group attachment to the C5-positions of cytosine residues in the context of CpG dinucleotides [7]. CpG dinucleotides tend to concentrate in the GC-rich DNA regions known as CpG islands, many of which are located in promoter gene regions and long repeat regions, such as retrotransposable elements or centromere

Table 1. Clinicopathological characteristics and data on the clinical endpoint status of patients with LumAB, TNBC and HER2-enriched BC molecular subtypes taken from open source (TCGA-BRCA)

Characteristics	LumAB	TNBC	HER2-enriched
Number of samples (%)	555	134	46
Age (median), years	59	54	58
T (%)			
T1	148 (26.49)	26 (19.4)	12 (26.09)
T2	310 (55.86)	87 (64.93)	28 (60.87)
T3	81 (14.59)	16 (11.94)	3 (6.52)
T4	14 (2.52)	4 (2.99)	3 (6.52)
No information	2 (0.36)	1 (0.75)	–
N (%)			
N0	233 (41.98)	78 (58.21)	16 (34.78)
N1	197 (35.5)	41 (30.6)	17 (36.96)
N2	75 (13.51)	11 (8.21)	6 (13.04)
N3	42 (7.57)	4 (2.99)	4 (8.7)
No information	8 (1.44)	–	3 (6.52)
M (%)			
M0	431 (77.66)	110 (82.09)	37 (80.43)
M1	6 (1.08)	3 (2.24)	1 (2.17)
No information	118 (21.26)	21 (15.67)	8 (17.39)
Grade (%)			
I	98 (17.66)	16 (11.94)	4 (8.7)
II	291 (52.43)	94 (70.15)	28 (60.87)
III	154 (27.75)	19 (14.18)	12 (26.09)
IV	5 (0.9)	2 (1.49)	1 (2.17)
No information	7 (1.26)	3 (2.24)	1 (2.17)
Clinical endpoint			
Overall survival (%)			
No event	490 (88.29)	113 (84.33)	37 (80.43)
Death	65 (11.71)	21 (15.67)	9 (19.57)
Disease-free survival (%)			
No relapse	445 (80.18)	102 (76.12)	35 (76.09)
Relapse	39 (7.03)	20 (14.93)	4 (8.7)
No information	71 (12.79)	12 (8.96)	7 (15.22)
Progression-free survival (%)			
No progression	486 (87.57)	109 (81.34)	38 (82.61)
Progression	69 (12.43)	25 (18.66)	8 (17.39)

Table 2. Total number of signatures and CpG pairs obtained by the LASSO Cox regression method for each survival outcome and BC molecular subtype. For some CpG pairs it was impossible to define genes these belonged to

Clinical endpoint + molecular subtype	ID's of HM450 probes	Genes	Number of combinations obtained
OS + LumAB	cg02287630; cg20417424 cg05828605 cg00297993 cg20471297 cg08133669 cg17323488 cg08241401 cg00815177 cg08442529 cg20746134 cg01821113 cg04523731 cg11140305 cg22067527	*SLC30A7*; *EXTL2*; *ST6GALNAC5*; *C15orf41*; *DYNC1H1*; *MIA3*; *NIPAL3*; *HEY2*; *HK1*; *DIRC3*; *TMEM41A*; *SH3BP5L*; *RFX2*	32752
DFS + LumAB	cg22790777 cg23667405 cg01017355 cg09561458 cg08128789 cg27304144 cg08039281 cg13486627 cg04833210 cg27439396 cg24347894	*SLC25A39*; *BAT2*; *ZNF417*; *LRRC8B*; *HSPG2*; *PSMA6*; *RG9MTD3*; *RBM19*; *N6AMT2*; *ZNF827*	2036
PFS + LumAB	cg13792075 cg08128789 cg13447284 cg00815177 cg10466124 cg13486627 cg15481636 cg00120948 cg05564086 cg23667405 cg17960080	*LRRC8B*; *HIST3H2A*; *ABCC5*; *BAT2*; *SPAG5*; *RERE*; *NIPAL3*; *HLA-DRB5*; *RG9MTD3*	2036
OS + TNBC	cg03512997 cg07804617 cg12814969 cg14293027 cg15355719 cg17053075 cg19002462 cg26401512 cg02567719	*FAM136A*; *HNRPDL*; *ENOPH1*; *LIN54*; *DNAJB4*; *ZNF643*; *TAP1*; *RASGRP2*; *LDLRAD3*	502
DFS + TNBC	cg20154816 cg02927111 cg18701707 cg12484411 cg20222926 cg02338142 cg06667406 cg13420273 cg22512222 cg17804981 cg13745678	*FEZF1*; *PLIN5*; *KCNMB2*; *AASS*; *HDAC9*; *ZFAND1*; *TRHR*; *PKNOX1*	2036
PFS + TNBC	cg01652244 cg02927111 cg20154816 cg00355315 cg24083274 cg23390595 cg13420273 cg10170774 cg01323371	*SSU72*; *DPPA5*; *PEX5L*; *HDAC9*; *CADPS2*; *STC1*; *PKNOX1*	502
OS + HER2-enriched	cg19236995 cg01564068 cg07351262 cg23409370 cg26290926 cg22043168 cg19986472 cg01647795	*GSTM4*; *BDNF*; *SLC43A1*; *PATL2*; *DHX8*	247
DFS + HER2-enriched	cg02327465 cg11261264 cg23302638 cg27252154 cg10660854 cg02796790 cg04407660 cg23183932	*BIRC5*; *EDARADD*; *TAPBPL*; *QTRT1*; *PTPRH*; *SNRPB*	247
PFS + HER2-enriched	cg20662988 cg23409370 cg23757489 cg03880890 cg04073970 cg22284390 cg27020573 cg00297843 cg17258551	*KCNN1*; *BOD1*; *SLC35F1*; *HTR5A*; *SLC45A1*; *CCDC49*	502

repeats. Methylation of cytosine is mediated by the enzyme class known as methyltransferases (DNMT) [7]. A total of five DNMT family members have been identified in mammals: DNMT1, DNMT2, DNMT3a, DNMT3b, and DNMT3L. DNMT3a and DNMT3b are *de novo* methyltransferases that interact with non-methylated CpG dinucleotides. DNMT1 is responsible for methylation maintenance during replication in S phase. It has been shown that DNMT3L stimulates *de novo* methylation that involves DNMT3a and mediates transcriptional repression with the help of histone deacetylase 1 (HDAC1) [7]. Aberrant DNA methylation is associated with a wide range of diseases and appears to be most marked in malignant tumors [8]. Studies of recent years show that every epithelial tumor contains about 10–15 genes inactivated by the genome structural changes and hundreds of genes inactivated by DNA hypermethylation. This demonstrates the importance of this modification for tumor development. Total hypomethylation is one more feature of tumor genomes. This is a genome-wide hypomethylation that results mainly from the loss of methylation at repetitive elements and leads to genome instability and chromosomal rearrangement [8]. The increased promoter methylation in the tumor suppressor genes suppressing various mechanisms of tumor progression, that resuts in epigenetic silencing and reversible inactivation of these genes, plays an important role in BC pathogenesis [8]. Identification of the tumor-specific aberrant DNA methylation patterns can be useful for early diagnosis of cancer, differential diagnosis of malignant neoplasms, in the capacity of prognostic and predictive markers [9]. The study of specific DNA methylation patterns identified by genome-wide analysis makes an important contribution to understanding of BC pathogenesis [10]. As noted above, each cancer type is divided into subtypes. There are genomic patterns, including

Table 3. The best signature, number of probes in the signature, values of cvAUC (cross-validated area under curve; the average area under curve obtained at all stages of cross-validation), sensitivity, specificity and accuracy for each survival outcome and BC molecular subtype

Clinical endpoint + molecular subtype	Number of probes in the combination	Genes in the combination	cvAUC	Sensitivity	Specificity	Accuracy
OS + LumAB	12	*SLC30A7, EXTL2, C15orf41, MIA3, NIPAL3, HEY2, HK1, DIRC3, TMEM41A, SH3BP5L, RFX2*	0.797	0.829	0.629	0.805
DFS + LumAB	6	*SLC25A39, BAT2, ZNF417, PSMA6, RG9MTD3, ZNF827*	0.831	0.838	0.716	0.828
PFS + LumAB	9	*ABCC5, NIPAL3, HLA-DRB5, RG9MTD3, HIST3H2A, RERE, SPAG5, BAT2, cg13447284*	0.761	0.875	0.562	0.836
OS + TNBC	5	*cg03512997, LIN54, RASGRP2, LDLRAD3, ZNF643*	0.969	0.864	0.939	0.876
DFS + TNBC	5	*PKNOX1, KCNMB2, ZFAND1, HDAC9, cg13745678*	0.834	0.902	0.673	0.865
PFS + TNBC	6	*DPPA5, cg02927111, PKNOX1, SSU72, CADPS2, PEX5L*	0.844	0.952	0.674	0.900
OS + HER2-enriched	3	*GSTM4 (TSS200), GSTM4 (Body), cg26290926*	0.898	1	0.883	0.977
DFS + HER2-enriched	2	*BIRC5,cg10660854*	1	1	1	1
PFS + HER2-enriched	4	*SLC43A1, BOD1, cg00297843, KCNN1*	1	0.947	1	0.956

epigenetic ones, that are typical for these subtypes. Thus, it is necessary to perform specific genome-wide DNA methylation profiling in cancer patients, along with the conventional assessment of the promoter hypermethylation point events in certain genes [11].

Prognosis involves prediction of the possible course and outcome of cancer. Survival analysis that is based on mathematical approach to cancer prognosis makes it possible to predict the likelyhood of staying alive after a certain time. Because of their biological importance and stability, DNA methylation markers are an effective prognostic factor [12]. In one of the studies, the data of the genome-wide DNA methylation analysis of BC samples from The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) database were used to construct a model of seven CpG dinucleotides that made it possible to clearly distinguish breast tumors of all subtypes and normal tissues, and to identify six methylation sites that strongly correlated with overall survival (OS) [13]. The analysis of methylation data from open sources by LASSO regression and boosting revealed 29 and 11 CpG dinucleotides associated with OS, resperctively [14]. The study of data taken from the open source (TCGA-BRCA) also made it possible to identify three genes (*TDRD10*, *PRAC2*, and *TMEM132C*), the methylation status of which had some predictive value, however, this was true mostly for estrogen receptor-positive breast tumors [15]. A prognostic model that comprises five genes (*TGFBR2*, *EIF4EBP1*, *FOSB*, *BCL2A1*, *ADRB2*) has been developed for TNBC based on the data obtained from TCGA-BRCA. The model is equally well suited for prediction of OS and disease free survival (DFS) [16].

Research is necessary due to the lack of such signatures for HER2-enriched subtype and a rather limited number of signatures for other BC molecular subtypes. The diagnostic potential of the existing survival prediction models is also uncertain, that is why we have used a modified algorithm to search for CpG dinucleotides associated with all available clinical endpoints found in the TCGA-BRCA database.

The study was aimed to obtain various signatures based on the open data on DNA methylation in BC from The Cancer Genome Atlas Breast Cancer for prediction of various clinical endpoints (overall survival, disease-free survival, and progression-free survival) for BC molecular subtypes and test the relationship between the clinicopathological characteristics and the signatures obtained.

METHODS

The publicly available clinical parameters and the data of the genome-wide DNA methylation profiling obtained using the HumanMethylation450 (HM450) hybridization chips (Illumina Inc.; USA) within the framework of The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) project (https://portal.gdc.cancer.gov/projects/TCGA-BRCA) were acquired and processed using the TCGAbiolinks software package [17]. Inclusion criteria for patients to be used for further selection of candidate CpG pairs were as follows: appropriate BC molecular subtype, availability of accessible clinicopathological information, availability of the DNA methylation profiling data obtained using the Illumina HumanMethylation450 chips. Exclusion criteria: no data on the time values for clinical endpoints, patient's age, TNM stage and grade. Then the results obtained using the patients' FFPE (formalin fixed paraffin-embedded) blocks and cross-hybridization probes were excluded from the profiling data matrix.

Selection of CpG pairs assocciated with OS, DFS or progression-free survival (PFS) was performed by univariate Cox regression method [18]. Of all the selected CpG pairs, those subjected to multiple testing adjustment (adjusted value $p < 0.05$, Wald test was used) by the false discovery rate (FDR) method were further analyzed. The LASSO Cox regression [19] method implemented in the SurvHiDim software package [20] was used to select the most stable CpG pairs. Multivariate Cox regression [21] was used to calculate the CpG-based signatures and to test the independence between the patients' clinical parameters and these signatures. The ability to classify various outcomes was defined by logistic regression method. Stratification into the high- and low-risk categories was performed using the median. The cvROC (cross-validated receiver operative curve) method [22] was used to test the quality of the models constructed and to plot the ROC curves.

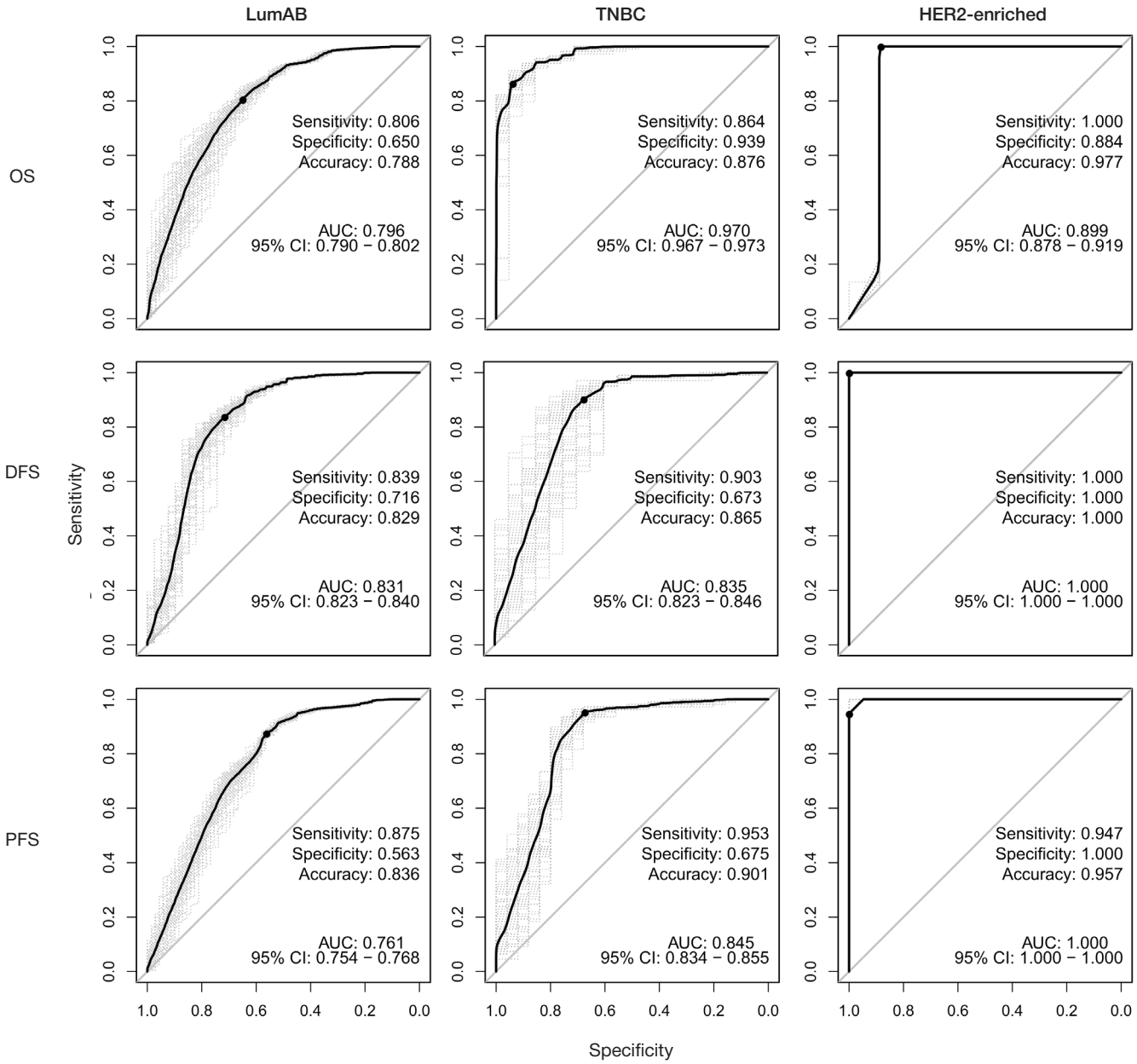

Fig. 1. cvROC curves (cross-validated receiver operative curve; ROC curve is plotted at every stage of cross-validation, then the resulting curve is constructed) for the best signatures. The vertical axis shows sensitivity (0–1), the horizontal axis shows specificity (0–1), rows show survival outcomes, columns show BC molecular subtypes

The best sensitivity and specificity values were defined by the Youden's index method. The Kaplan–Meier curves were constructed using the survminer software package [23]. The Mantel–Cox test was used to compare two survival curves. The 10-fold cross-validation method was used throughout all stages of marker selection and signature calculation. All the listed above calculations were performed using the R statistical programming language [24].

RESULTS

The studied TCGA-BRCA data set included the DNA methylation profile obtained using the HM450 chips and the clinicopathological characteristics of 735 primary BC samples. After exclusion of paraffin-embedded samples, there were a total of 555 LumA+B subtype (LumAB) samples, 134 TN subtype samples, and 46 HER2-enriched subtype samples (Table 1). Prior to selection of traits for further analysis, cross-hybridization probes were excluded from the methylation data matrix, so that the number of probes reduced from 485,577 to 456,344, respectively.

The next stage of analysis involved using univariate Cox regression to search for methylation sites that correlate with the duration of OS, DFS and PFS in various BC molecular subtypes. After the initial selection with allowance for the multiple testing adjusted p-value, the following probes were selected:

– 10,433 probes associated with OS in the LumAB subtypes, 3214 probes in the TN subtype, 6471 probes in the HER2-enriched subtype;

– 4419 probes associated with DFS in the LumAB subtypes, 168 probes in the TN subtype, 483 probes in the HER2-enriched subtype;

– 2345 probes associated with PFS in the LumAB subtypes, 43 probes in the TN subtype, 3216 probes in the HER2-enriched subtype.

LASSO Cox regression was used for each of the listed sets, allowing for selection of CpG dinucleotides that were most important for analysis. Different numbers of such CpG pairs

Table 4. Multivariate Cox regression results for the best signatures and clinicopathological characteristics. HR — hazard ratio (relative risk), P — responsible for *p*-val

Variable/ clinical endpoint + molecular subtype	OS + LumAB		DFS + LumAB		PFS + LumAB		OS + TNBC		DFS + TNBC		PFS + TNBC		OS + HER2-enriched		DFS + HER2-enriched		PFS + HER2-enriched	
	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.	HR (95% confidence interval)	P.
Risk indicator	1 (1–1)	< 0.001	1.1 (1.05–1.10)	< 0.001	1.03 (1.02–1.04)	< 0.001	1.1 (1.004–1.2)	< 0.001	1.02 (1.01–1.03)	< 0.001	1.02 (1.001–1.03)	< 0.001	1.0 (1–1)	0.03	1.02 (1.009–1.03)	0.006	1.005 (1.001–1.000)	0.01
Age (below median/over median)	1.38 (0.84–2.30)	0.201	1.2 (0.58–2.40)	0.638	2 (0.9–3.30)	0.231	0.76 (0.28–2.00)	0.575	0.51 (0.160–1.600)	0.25	0.66 (0.270–1.600)	0.35	6.2 (0.619–62.900)	0.12	1.9 (0.878–2.344)	0.512	2.06 (0.27–15)	0.47
T (T1-2/3-4)	0.87 (0.48–1.60)	0.64	1.1 (0.48-2.60)	0.78	0.77 (0.38–1.60)	0.463	2.52 (0.75-8.40)	0.134	1.06 (0.310–3.600)	0.93	1.50 (0.430–5.200)	0.52	2.6 (0.169–39.900)	0.49	0.9 (0.625–1.932)	0.404	7.00 (0.50–98)	0.14
N (N0/N1-3)	1.23 (0.74–2.00)	0.42	2.1 (0.73–6.00)	0.16	1.1 (0.56–2.20)	0.097	2.95 (0.88-9.90)	0.08	0.96 (0.280–3.300)	0.95	1.14 (0.390–3.300)	0.81	1.1 (0.174–7.500)	0.89	0.34 (0.120–1.500)	0.463	1.09 (0.07–16)	0.94
M (M0/M1)	1.03 (0.51–2.10)	0.94	1.6 (0.66–3.80)	0.30	1.9 (0.90–2.20)	0.076	0.36 (0.09–1.30)	0.125	1.56 (0.360–6.700)	0.55	0.89 (0.290–2.800)	0.83	1.2 (0.089–15.300)	0.90	0.63 (0.105–2.000	0.376	0.02 (0.00–1.05)	0.053
Grade (I-II/ III-IV)	1.20 (0.71–2.00)	0.49	2.3 (0.95–5.50)	0.06	1.54 (0.80–2.90)	0.19	1.48 (0.36–6.00)	0.58	4.19 (0.94–18.60)	0.06	3.76 (0.97–14.50)	0.055	1.10 (0.09–14.30)	0.91	2.6 0(0.78–2.80)	0.16	1.83 (0.12–27.00)	0.66

were identified during each stage of cross-validation. CpG pairs found in more than 50% of cross-validation data splits were selected (Table 2).

To select the combinations of CpG dinucleotides showing significant correlations with various survival outcomes, we assessed all possible combinations (signatures) of such CpG dinucleotides in various BC molecular subtypes. For each clinical endpoint and BC molecular subtype, cvAUC (cross-validated area under curve, the average area under curve at all stages of cross-validation), sensitivity, specificity and accuracy were defined for various combinations. The first 10 combinations showing high cvAUC values were tested for independence of the clinicopathological characteristics. The diagnostic characteristics of these combinations along with the number of probes and genes belonging to the probes are provided in Table 3.

The combination of 12 CpG dinucleotides for prediction of OS in the LumAB subtype was the largest defined combination, while the combination of two CpG dinucleotides for prediction of DFS in the HER2-expressing subtype was the smallest one. The cvROC (cross-validated receiver operative characteristics: ROC curve is plotted at every stage of cross-validation, then the average curve is constructed) curves and the Kaplan–Meier curves were plotted for each signature to show the diagnostic potential and estimate the survival function. The LumAB combinations showed lower cvAUC values (0.76–0.83), while the combinations for TN and HER2-expressing subtypes showed high cvAUC values with fewer nember of combinations (0.83–1) (Fig. 1).

Our combinations are independent of the clinical characteristics (Table 4). This makes it possible to use the risk indicators of these clinical endpoints in any group of patients.

The Kaplan–Meier curve analysis revealed a significant ($p < 0.05$) decrease in OS, DFS and PFS in the group of patients with high risk of death, relapse and disease progression compared to the group of patients with low risk. This was true for all BC molecular subtypes and all selected combinations (Fig. 2).

DISCUSSION

In this study, we considered the possibility of identifying the CpG dinucleotide differentially methylated sites to predict survival outcomes in various BC molecular subtypes using the methods of survival analysis and DNA methylation data. The approach to calculation of differential methylation based on univariate Cox regression is widely used in a variety of studies. Thus, this method was used for identification of 249,810 and 249,811 probes based on DNA methylation data

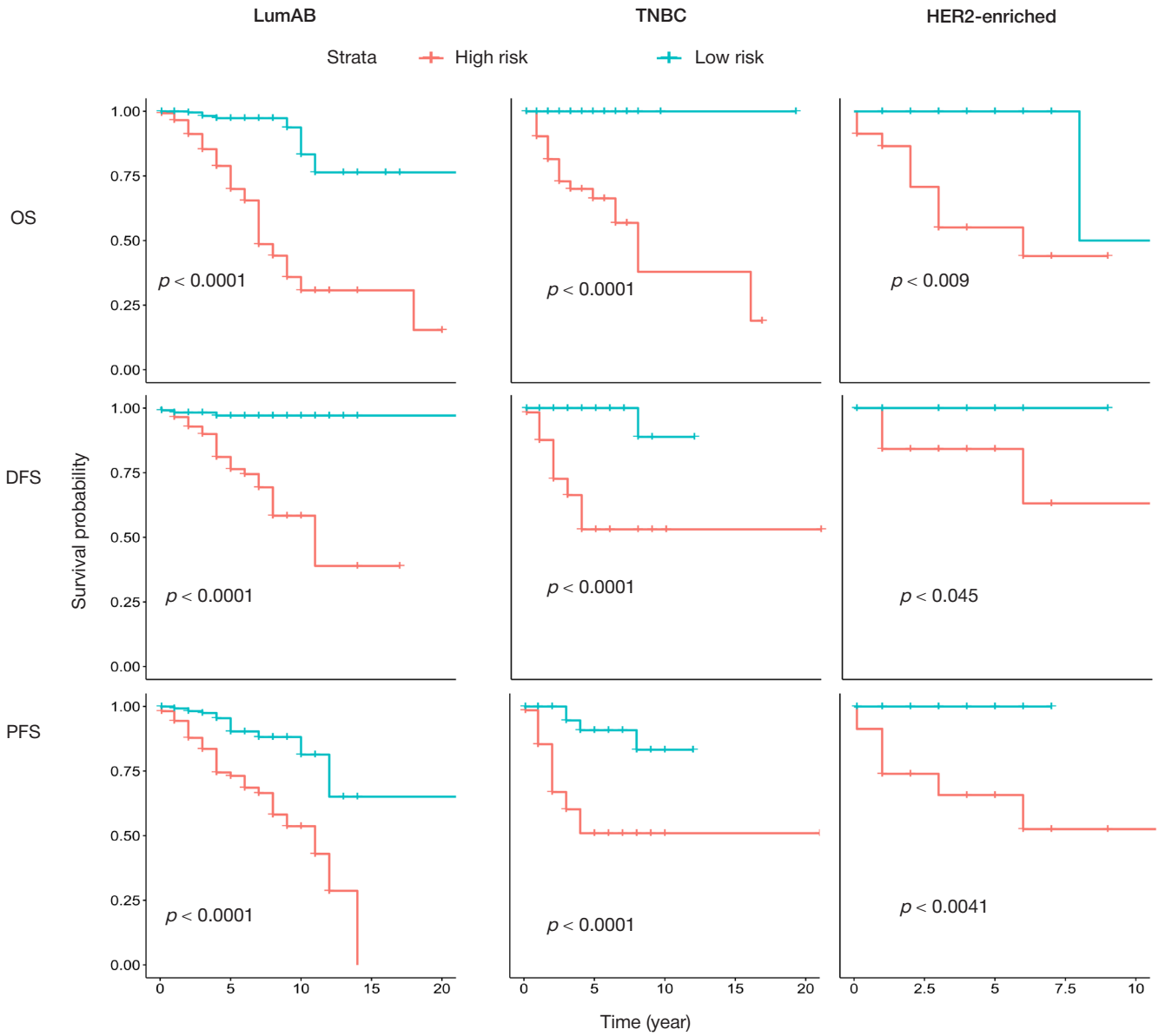

Fig. 2. Kaplan–Meier curves for the best signatures. The horizontal axis shows time (years), the vertical axis shows the likelihood of staying alive (0–1). High risk of death, progression and relapse is highlited in *red*, low risk of death, progression and relapse is highlighted in *turquoise*. Rows show survival outcomes, and columns show BC molecular subtypes

associated with ovarian cancer and BC, respectively [12], and for identification of probes based on DNA methylation data associated with cutaneous melanoma [25].

We have shown that the use of various combinations (2–12 CpG dinucleotides) makes it possible to achieve acceptable (cvAUC 0.7–0.8), high (0.8–0.9) and very high quality (0.9–1) of classification into high and low risk of death, relapse and progression. During the study we have identified 47 probes/genes (*SLC30A7, EXTL2, C15orf41, MIA3, NIPAL3, HEY2, HK1, DIRC3, TMEM41A, SH3BP5L, RFX2, SLC25A39, BAT2, ZNF417, PSMA6, RG9MTD3, ZNF827, ABCC5, HLA-DRB5, HIST3H2A, RERE, SPAG5, cg13447284, cg03512997, LIN54, RASGRP2, LDLRAD3, ZNF643, PKNOX1, KCNMB2, ZFAND1, HDAC9, cg13745678, DPPA5, cg02927111, PKNOX1, SSU72, CADPS2, PEX5L, GSTM4, cg26290926, BIRC5, cg10660854, SLC43A1, BOD1, cg00297843, KCNN1*), the methylation of which is associated with OS, DFS and PFS. We failed to define which genes seven probes (*cg13447284, cg03512997, cg13745678, cg02927111, cg26290926, cg10660854, cg00297843*) belonged to. Five of these probes (*BIRC5, PKNOX1, SPAG5, HDAC9, PSMA6*) have been previously reported in scientific literature as molecular markers of BC patients' survival based on the data of gene expression quantification [26–31].

It is noteworthy that in our study we found the same number of CpG dinucleotides (HM450 probes) for prediction of OS and DFS (based on five probes) as other researchers [16], but the probes varied according to genes they belonged to. According to our data, the signature for prediction of PFS consists of six probes; no PFS signiture was calculated during the study [16]. Other researchers suggest using invidual markers of promoter hypermethylation in seven genes (RASSF1, BRCA1, PITX2, RARB, PGR, CDH1, and PCDH10) for prediction of OS and DFS outcome in patients with ER+ BC, they also consider using the panel of three genes (GSTP1, RASSF1, and RARB) for prediction of OS based on the literature data analysis (systematic review of the reports) [26], while we have used a strategy of developing panels of six, nine and 12 methylation markers based on the marker diagnostic potential defined by statistical analysis of the experimental data set.

Among genes included in our combinations, attention is drawn to BIRC5 (encodes baculoviral IAP repeat-containing

protein 5) that is overexpressed in the majority of tumors, including BC, and is associated with poorer prognosis of overall, disease-free and progression-free survival. It has been shown that the use of taxane chemotherapy drugs may increase the expression of this gene [27]. *PKNOX1* (gene of the short arm of chromosome 21 that encodes eponymous protein and plays an important role in embryogenesis) is a tumor suppressor gene, while the increased expression of this gene is associated with poorer survival rate [28]. The increased expression of *SPAG5* (encodes protein associated with the mitotic spindle apparatus), associated with poorer prognosis of OS, DFS and PFS only in estrogen receptor-positive (ER^+) breast tumors, is also a prognostic factor [29], which is confirmed by our study. The findings of the study that involved ER^+ BC samples show that the increased expression of *HDAC9* epigenetic enzyme (encodes protein, histone deacetylase 9) in tumors is associated with the poorer prognosis of DFS [30]. Our study shows the association between the abnormal methylation of this gene and survival of patients with estrogen receptor-negative (ER^-) tumors, and more precisely with TNBC. The reduced DFS of patients with ER^+ BC was associated with the increased expression of *PSMA6* (encodes proteasome subunit alpha type-6) [31], which was also confirmed by our findings.

CONCLUSIONS

Molecular epigenetic signatures for various BC types were discovered using the survival analysis methods, combinations of various methylation sites, and estimation of diagnostic parameters. This method may be recommended to search for signatures typical for BC and other tumor diseases. In the future the discovered epigenetic signatures may be used to develop the methylation-sensitive quantitative PCR assays. After clinical trials, such assays may become a cheaper and more practical alternative to gene expression microarrays, without reducing diagnostic performance.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71 (3): 209–49. DOI: 10.3322/caac.21660
2. Bernhardt SM, Dasari P, Walsh D, Townsend AR, Price TJ, Ingman WV. Hormonal Modulation of Breast Cancer Gene Expression: Implications for Intrinsic Subtyping in Premenopausal Women. Front Oncol. 2016; 6: 241. DOI: 10.3389/fonc.2016.00241.
3. Zaha DC. Significance of immunohistochemistry in breast cancer. World J Clin Oncol. 2014; 5 (3): 382–92. DOI: 10.5306/wjco.v5.i3.382. PMID: 25114853; PMCID: PMC4127609.
4. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406 (6797): 747–52. DOI: 10.1038/35021093.
5. Echeverria GV, Ge Z, Seth S, Zhang X, Jeter-Jones S, Zhou X, et al. Resistance to neoadjuvant chemotherapy in triple-negative breast cancer mediated by a reversible drug-tolerant state. Sci Transl Med. 2019; 11 (488): eaav0936. DOI: 10.1126/scitranslmed.aav0936.
6. Blanchette P, Sivajohanathan D, Bartlett J, Eisen A, Feilotter H, Pezo R, et al. Clinical Utility of Multigene Profiling Assays in Early-Stage Invasive Breast Cancer: An Ontario Health (Cancer Care Ontario) Clinical Practice Guideline. Curr Oncol. 2022; 29 (4): 2599–615. DOI: 10.3390/curroncol29040213.
7. Edwards JR, Yarychkivska O, Boulard M, Bestor TH. DNA methylation and DNA methyltransferases. Epigenetics Chromatin. 2017; 10: 23. DOI: 10.1186/s13072-017-0130-8.
8. Vietri MT, D'Elia G, Benincasa G, Ferraro G, Caliendo G, Nicoletti GF, et al. DNA methylation and breast cancer: A way forward (Review). Int J Oncol. 2021; 59 (5): 98. DOI: 10.3892/ijo.2021.5278.
9. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 2007; 8 (4): 286–98. DOI: 10.1038/nrg2005.
10. Karami Fath M, Azargoonjahromi A, Kiani A, Jalalifar F, Osati P, Akbari Oryani M, et al. The role of epigenetic modifications in drug resistance and treatment of breast cancer. Cell Mol Biol Lett. 2022; 27 (1): 52. DOI: 10.1186/s11658-022-00344-6.
11. Lee G, Bang L, Kim SY, Kim D, Sohn KA. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. BMC Med Genomics. 2017; 10 (Suppl 1): 28. DOI: 10.1186/s12920-017-0268-z.
12. Hu WL, Zhou XH. Identification of prognostic signature in cancer based on DNA methylation interaction network. BMC Med Genomics. 2017; 10 (Suppl 4): 63. DOI: 10.1186/s12920-017-0307-9.
13. Zhang M, Wang Y, Wang Y, Jiang L, Li X, Gao H, et al. Integrative Analysis of DNA Methylation and Gene Expression to Determine Specific Diagnostic Biomarkers and Prognostic Biomarkers of Breast Cancer. Front Cell Dev Biol. 2020; 8: 529386. DOI: 10.3389/fcell.2020.529386.
14. Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, et al. DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci U S A. 2017; 114 (28): 7414–9. DOI: 10.1073/pnas.1703577114.
15. de Almeida BP, Apolónio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer. 2019; 19 (1): 219. DOI: 10.1186/s12885-019-5403-0.
16. Gao Y, Wang X, Li S, Zhang Z, Li X, Lin F. Identification of a DNA Methylation-Based Prognostic Signature for Patients with Triple-Negative Breast Cancer. Med Sci Monit. 2021; 27: e930025. DOI: 10.12659/MSM.930025.
17. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016; 44 (8): e71. DOI: 10.1093/nar/gkv1507.
18. Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. Oxid Med Cell Longev. 2021; 2021: 1302811. DOI: 10.1155/2021/1302811.
19. Utazirubanda JC, Leon T, Ngom P. Variable selection with Group LASSO approach: Application to Cox regression with frailty model. Commun Stat Simul Comput. 2021; 50 (3): 881–901. DOI: 10.1080/03610918.2019.1571605.
20. Bhattacharjee A, Pawar A. SurvHiDim: high dimensional survival data analysis. R package version 0.1.1. 2021. Available from: https://CRAN.R-project.org/package=SurvHiDim.
21. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis — an introduction to concepts and methods. Br J Cancer. 2003; 89 (3): 431–6. DOI: 10.1038/sj.bjc.6601119.
22. LeDell E, Petersen M, van der Laan M. cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals. R package version 1.1-4. 2022 Available from: http://CRAN.R-project.org/package=cvAUC.
23. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.6. Available from: https://CRAN.R-project.org/package=survminer.
24. Team RCR Foundation for Statistical Computing; Vienna, Austria: 2015.
25. Guo W, Zhu L, Zhu R, Chen Q, Wang Q, Chen JQ. A four-DNA methylation biomarker is a superior predictor of survival of patients

with cutaneous melanoma. Elife. 2019; 8: e44310. DOI: 10.7554/eLife.44310.
26. de Ruijter TC, van der Heide F, Smits KM, Aarts MJ, van Engeland M, Heijnen VCG. Prognostic DNA methylation markers for hormone receptor breast cancer: a systematic review. Breast Cancer Res. 2020; 22 (1): 13. DOI: 10.1186/s13058-020-1250-9.
27. Dai JB, Zhu B, Lin WJ, Gao HY, Dai H, Zheng L, et al. Identification of prognostic significance of BIRC5 in breast cancer using integrative bioinformatics analysis. Biosci Rep. 2020; 40 (2): BSR20193678. DOI: 10.1042/BSR20193678.
28. Jiang S, Bu X, Tang D, Yan C, Huang Y, Fang K. A tumor suppressor gene-based prognostic classifier predicts prognosis, tumor immune infiltration, and small molecule compounds in breast cancer. Front Genet. 2022; 12: 783026. DOI: 10.3389/fgene.2021.783026.
29. Mohamadalizadeh-Hanjani Z, Shahbazi S, Geranpayeh L. Investigation of the SPAG5 gene expression and amplification related to the NuMA mRNA levels in breast ductal carcinoma. World J Surg Oncol. 2020; 18 (1): 225. DOI: 10.1186/s12957-020-02001-8.
30. Linares A, Assou S, Lapierre M, Thouennon E, Duraffourd C, Fromaget C, et al. Increased expression of the HDAC9 gene is associated with antiestrogen resistance of breast cancers. Mol Oncol. 2019; 13 (7): 1534–47. DOI: 10.1002/1878-0261.
31. Li Y, Huang J, Sun J, Xiang S, Yang D, Ying X, et al. The transcription levels and prognostic values of seven proteasome alpha subunits in human cancers. Oncotarget. 2017 Jan 17; 8 (3): 4501–19. DOI: 10.18632/oncotarget.13885.

Литература

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71 (3): 209–49. DOI: 10.3322/caac.21660
2. Bernhardt SM, Dasari P, Walsh D, Townsend AR, Price TJ, Ingman WV. Hormonal Modulation of Breast Cancer Gene Expression: Implications for Intrinsic Subtyping in Premenopausal Women. Front Oncol. 2016; 6: 241. DOI: 10.3389/fonc.2016.00241.
3. Zaha DC. Significance of immunohistochemistry in breast cancer. World J Clin Oncol. 2014; 5 (3): 382–92. DOI: 10.5306/wjco.v5.i3.382. PMID: 25114853; PMCID: PMC4127609.
4. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406 (6797): 747–52. DOI: 10.1038/35021093.
5. Echeverria GV, Ge Z, Seth S, Zhang X, Jeter-Jones S, Zhou X, et al. Resistance to neoadjuvant chemotherapy in triple-negative breast cancer mediated by a reversible drug-tolerant state. Sci Transl Med. 2019; 11 (488): eaav0936. DOI: 10.1126/scitranslmed.aav0936.
6. Blanchette P, Sivajohanathan D, Bartlett J, Eisen A, Feilotter H, Pezo R, et al. Clinical Utility of Multigene Profiling Assays in Early-Stage Invasive Breast Cancer: An Ontario Health (Cancer Care Ontario) Clinical Practice Guideline. Curr Oncol. 2022; 29 (4): 2599–615. DOI: 10.3390/curroncol29040213.
7. Edwards JR, Yarychkivska O, Boulard M, Bestor TH. DNA methylation and DNA methyltransferases. Epigenetics Chromatin. 2017; 10: 23. DOI: 10.1186/s13072-017-0130-8.
8. Vietri MT, D'Elia G, Benincasa G, Ferraro G, Caliendo G, Nicoletti GF, et al. DNA methylation and breast cancer: A way forward (Review). Int J Oncol. 2021; 59 (5): 98. DOI: 10.3892/ijo.2021.5278.
9. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 2007; 8 (4): 286–98. DOI: 10.1038/nrg2005.
10. Karami Fath M, Azargoonjahromi A, Kiani A, Jalalifar F, Osati P, Akbari Oryani M, et al. The role of epigenetic modifications in drug resistance and treatment of breast cancer. Cell Mol Biol Lett. 2022; 27 (1): 52. DOI: 10.1186/s11658-022-00344-6.
11. Lee G, Bang L, Kim SY, Kim D, Sohn KA. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. BMC Med Genomics. 2017; 10 (Suppl 1): 28. DOI: 10.1186/s12920-017-0268-z.
12. Hu WL, Zhou XH. Identification of prognostic signature in cancer based on DNA methylation interaction network. BMC Med Genomics. 2017; 10 (Suppl 4): 63. DOI: 10.1186/s12920-017-0307-9.
13. Zhang M, Wang Y, Wang Y, Jiang L, Li X, Gao H, et al. Integrative Analysis of DNA Methylation and Gene Expression to Determine Specific Diagnostic Biomarkers and Prognostic Biomarkers of Breast Cancer. Front Cell Dev Biol. 2020; 8: 529386. DOI: 10.3389/fcell.2020.529386.
14. Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, et al. DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci U S A. 2017; 114 (28): 7414–9. DOI: 10.1073/pnas.1703577114.
15. de Almeida BP, Apolónio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer. 2019; 19 (1): 219. DOI: 10.1186/s12885-019-5403-0.
16. Gao Y, Wang X, Li S, Zhang Z, Li X, Lin F. Identification of a DNA Methylation-Based Prognostic Signature for Patients with Triple-Negative Breast Cancer. Med Sci Monit. 2021; 27: e930025. DOI: 10.12659/MSM.930025.
17. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016; 44 (8): e71. DOI: 10.1093/nar/gkv1507.
18. Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. Oxid Med Cell Longev. 2021; 2021: 1302811. DOI: 10.1155/2021/1302811.
19. Utazirubanda JC, Leon T, Ngom P. Variable selection with Group LASSO approach: Application to Cox regression with frailty model. Commun Stat Simul Comput. 2021; 50 (3): 881–901. DOI: 10.1080/03610918.2019.1571605.
20. Bhattacharjee A, Pawar A. SurvHiDim: high dimensional survival data analysis. R package version 0.1.1. 2021. Available from: https://CRAN.R-project.org/package=SurvHiDim.
21. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis — an introduction to concepts and methods. Br J Cancer. 2003; 89 (3): 431–6. DOI: 10.1038/sj.bjc.6601119.
22. LeDell E, Petersen M, van der Laan M. cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals. R package version 1.1-4. 2022 Available from: http://CRAN.R-project.org/package=cvAUC.
23. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.6. Available from: https://CRAN.R-project.org/package=survminer.
24. Team RCR Foundation for Statistical Computing; Vienna, Austria: 2015.
25. Guo W, Zhu L, Zhu R, Chen Q, Wang Q, Chen JQ. A four-DNA methylation biomarker is a superior predictor of survival of patients with cutaneous melanoma. Elife. 2019; 8: e44310. DOI: 10.7554/eLife.44310.
26. de Ruijter TC, van der Heide F, Smits KM, Aarts MJ, van Engeland M, Heijnen VCG. Prognostic DNA methylation markers for hormone receptor breast cancer: a systematic review. Breast Cancer Res. 2020; 22 (1): 13. DOI: 10.1186/s13058-020-1250-9.
27. Dai JB, Zhu B, Lin WJ, Gao HY, Dai H, Zheng L, et al. Identification of prognostic significance of BIRC5 in breast cancer using integrative bioinformatics analysis. Biosci Rep. 2020; 40 (2): BSR20193678. DOI: 10.1042/BSR20193678.
28. Jiang S, Bu X, Tang D, Yan C, Huang Y, Fang K. A tumor suppressor gene-based prognostic classifier predicts prognosis, tumor immune infiltration, and small molecule compounds in breast cancer. Front Genet. 2022; 12: 783026. DOI: 10.3389/

fgene.2021.783026.
29. Mohamadalizadeh-Hanjani Z, Shahbazi S, Geranpayeh L. Investigation of the SPAG5 gene expression and amplification related to the NuMA mRNA levels in breast ductal carcinoma. World J Surg Oncol. 2020; 18 (1): 225. DOI: 10.1186/s12957-020-02001-8.
30. Linares A, Assou S, Lapierre M, Thouennon E, Duraffourd C, Fromaget C, et al. Increased expression of the HDAC9 gene is associated with antiestrogen resistance of breast cancers. Mol Oncol. 2019; 13 (7): 1534–47. DOI: 10.1002/1878-0261.
31. Li Y, Huang J, Sun J, Xiang S, Yang D, Ying X, et al. The transcription levels and prognostic values of seven proteasome alpha subunits in human cancers. Oncotarget. 2017 Jan 17; 8 (3): 4501–19. DOI: 10.18632/oncotarget.13885.