

RESEARCH

Open Access



Novel methylation biomarkers in liquid biopsy and classifying biosignatures for the clinical management of breast cancer

Maria Panagopoulou^{2*}, Maria A. Papadaki³, Makrina Karaglani^{1,2}, Theodosios Theodosiou^{1,10}, Kleita Michaelidou³, Stavroula Baritaki⁴, Ioannis Tsamardinos^{5,6,7}, Stylianos Kakolyris⁸, Sofia Agelaki^{3,9} and Ekaterini Chatzaki^{1,2,10*}

Abstract

Background Breast Cancer (BrCa) remains a devastating disease presenting emerging needs for effective management. Recently, epigenetic biomarkers are assessed in liquid biopsy for diagnostic and prognostic applications. This study applies a 3-step data-driven biomarker discovery pipeline to identify robust methylation biomarkers and generate high-performance biosignatures specific for clinically significant BrCa end-points, followed by laboratory validation in patient cell-free DNA (cfDNA).

Methods Publicly available genome-wide methylomes from 520 BrCa and 185 non-diseased breast tissues (discovery dataset) were analyzed via Automated Machine Learning (AutoML, JADBio) to identify BrCa-specifically methylated promoters. Bioinformatic search revealed any BrCa biological relevance. Next, the methylation of identified promoters was experimentally validated in plasma cfDNA from 195 BrCa patients and 135 healthy individuals by Methylation Specific qPCR (qMSP) (validation cohort). Finally, autoML analyzed experimental and clinical data to develop optimized classifying biosignatures for diagnosis, prognosis, and prediction.

Results AutoML identified 3 BrCa-specific methylated promoters in *CLDN15*, *MRGPRD* and *ZNF430*. Pathway analysis revealed implications with biological processes such as signaling and transcription. Laboratory validation using clinical cfDNA samples confirmed elevated methylation levels in BrCa patients for all 3 promoters, which were correlated with poor prognostic and predictive parameters. Classification analysis by autoML of experimental methylation measurements and patients' clinical data built 5 specific models: a diagnostic biosignature distinguishing BrCa from health (AUC 0.79, CI: 0.75–0.84), a classification biosignature differentiating BrCa disease status (adjuvant, neoadjuvant, and metastatic group) (AUC 0.68, CI: 0.62–0.72), a prognostic biosignature predicting relapse (AUC 0.79, CI: 0.74–0.83), a biosignature predicting treatment response in metastatic patients (AUC 0.86, CI: 0.67–1.00), and a biosignature differentiating distinct molecular subtypes (AUC of 0.71, CI: 0.64–0.77), underscoring their possible clinical utility.

Conclusion Our data-driven approach successfully identified 3 BrCa-specifically methylated promoters in genes not previously implicated in BrCa. Their role in pathology needs further attention as they could also represent novel

*Correspondence:
Maria Panagopoulou
mpanagop@med.duth.gr
Ekaterini Chatzaki
achatzak@med.duth.gr

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

targets. Moreover, the laboratory validation in clinical BrCa cfDNA samples led to the development of 5 biosignatures, some demonstrating strong predictive performance. The low number of features and the minimally invasive nature of liquid biopsy highlight the potential for clinical implementation of great value.

Keywords Methylation, Promoter, Liquid biopsy, Breast cancer, Prognosis, Diagnosis, Biosignature, Model, Machine learning

Background

One of the most commonly diagnosed cancers in women worldwide is Breast Cancer (BrCa), with more than 2 million new cases and 660,000 new deaths each year [1]. The main risk factors are genetics, hormones, lifestyle, and age [2]. As BrCa is a heterogeneous disease, hormonal receptor status, Human Epidermal growth factor Receptor 2 (HER2), tumor size, stage, and in some cases, genetic tests are used to guide clinical decisions, prognosis, and treatment choice. However, resistance to therapy and relapse are the main reasons for BrCa mortality [3]. Emerging personalized choices can address these unmet clinical needs by tailoring treatment to an individual's genetic, molecular, and tumor profile. The precision approach enables the selection of targeted therapies that bypass resistance mechanisms, reduce toxicity, and improve efficacy, and clinicians can adapt treatment in real time, preventing relapse and ensuring sustained response. Personalized medicine promises to fill gaps left by traditional, one-size-fits-all protocols. This shift enhances outcomes and quality of life for diverse patient populations.

DNA methylation of gene promoters is an epigenetic mechanism that affects gene expression [4]. In cancer, methylation of specific genes such as tumor suppressors leads to transcriptional silencing promoting tumorigenesis [5, 6]. DNA methylation is a common event in BrCa and it is well established that it occurs early in the carcinogenic process until the last stages [7–10]. Thus, it is considered a suitable biomarker for multiple clinically important end-points related to prognosis and treatment [8, 11]. Gene promoter methylation can be assessed in liquid biopsy biomaterial, such as cell-free DNA (cfDNA), and multiple efforts have adopted this approach to develop suitable solutions [12]. cfDNA is a biomaterial consisting of extracellular DNA fragments of different sizes, mainly originating from apoptosis, necrosis, and active release [13]. It has been proved that it is an emerging minimally invasive biomaterial in oncology as it is enriched with tumorous DNA that enters the bloodstream during tumor lifespan [12]. Thus, it could be exploited for cancer early diagnosis, prognosis, and monitoring through single or repetitive blood tests.

It has become increasingly apparent that a single biomarker is impossible to offer great sensitivity and specificity to meet the needs for personalized patient management. Recently, high-throughput genome-wide

methylation methodologies such as next-generation sequencing or microarrays have revolutionized the acquisition of methylomics readings providing a large source of information regarding biological events and functions. Machine Learning (ML) has opened a new opportunity in interpreting and exploiting these vast omics datasets. ML involves a variety of algorithms and performs intelligent prediction and feature selection [14]. Machine learning offers significant advantages over traditional statistics in analyzing omics data for biomarker discovery. Unlike traditional methods, which often rely on predefined models and assumptions, machine learning can handle high-dimensional, noisy, and non-linear data typical of omics. It autonomously detects complex patterns and interactions among thousands of variables, uncovering subtle biomarkers that might be missed otherwise. Machine learning also enables feature selection, classification, and prediction with greater accuracy and scalability. By learning directly from data, it continuously improves with new inputs, making it highly adaptable for personalized medicine and real-time clinical decision-making based on evolving biological insights. Recently, Automated Machine Learning (AutoML), democratizes its use among non-experts as it automatizes the whole process task by task (e.g. data preprocessing, algorithm selection, feature selection, model training, hyperparameter tuning, model evaluation, analysis visualization, model output) [15]. Our team has previously employed a commercially available AutoML tool, ad-hoc for life science research (jadbio.com), that builds predictive models to develop efficient novel biomarkers with promising results in various clinical settings [8, 11, 16, 17]. Here, we applied a three-step data-driven pipeline (Fig. 1) analyzing by the same JADBIO AutoML tool publicly available breast cancer and normal breast methylomes to identify cancer-specific biomarkers, which we then validated clinically in cfDNA from BrCa patients from two different oncology clinics. Further machine learning in our experimental data builds classification biosignatures with diagnostic and prognostic value.

Step 1—In silico analysis: At first, TCGA and GEO databases were searched to retrieve methylome datasets from BrCa tumor tissues and healthy breast adjacent tissues. AutoML was employed to build classification models of differentially methylated genes. Selected genes were also investigated for their biological relevance to BrCa, including gene ontology, pathway analysis,

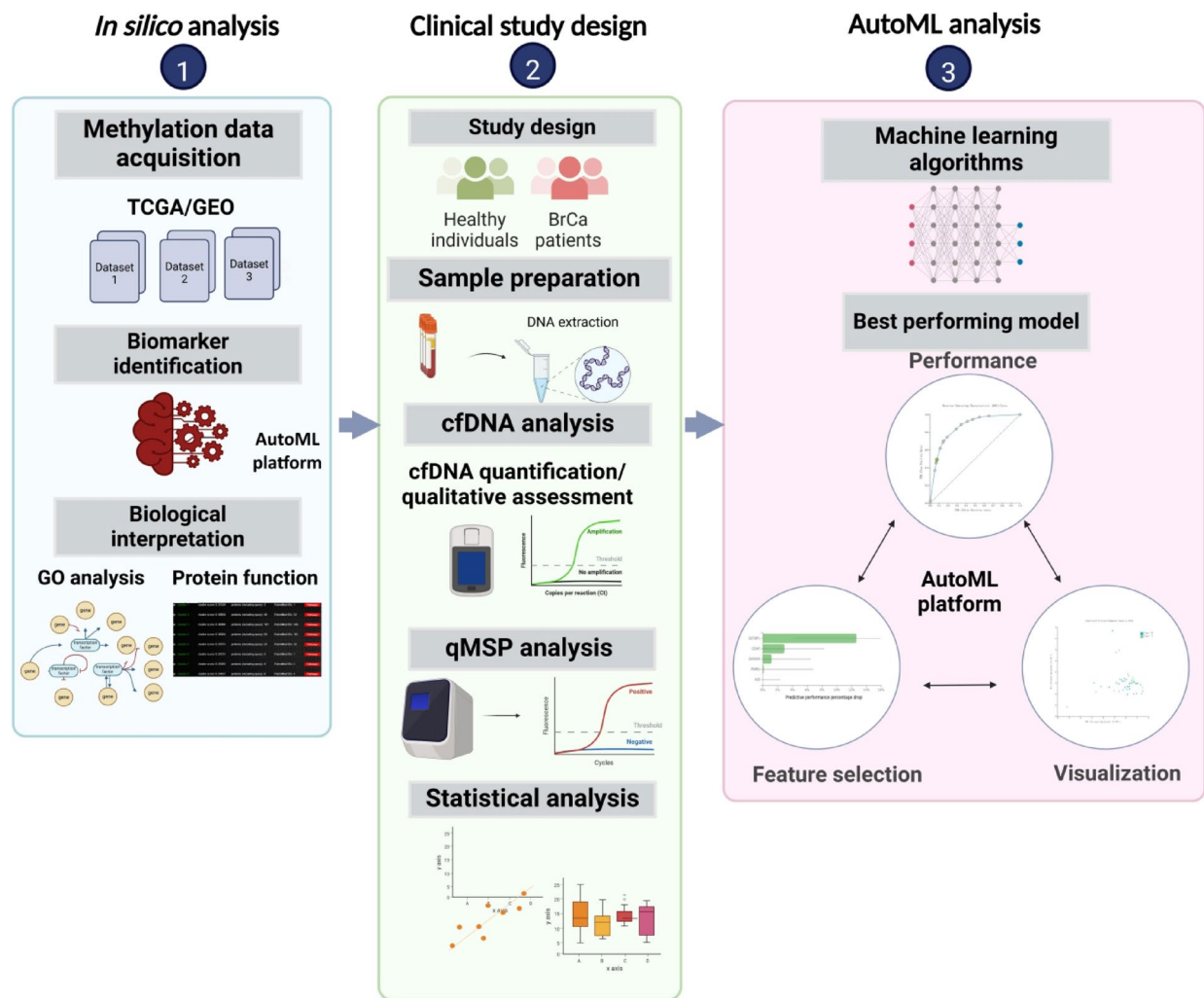


Fig. 1 Workflow of the study. Our pipeline uses a three-step approach to identify, develop and validate highly performing methylation biosignatures of clinical value in BrCa assessed in liquid biopsies

protein–protein interaction and functional relationships.

Step 2 – Clinical study: The quantity of cfDNA and the methylation status of the selected genes were measured in cfDNA of 195 BrCa patients compared to 135 healthy individuals, followed by standard statistical analysis. **Step 3—AutoML Analysis:** Finally, AutoML optimized further the biosignatures against clinically important end-points, based on our experimental patient methylation measurements combined with clinical and demographic data. Abbreviations: TCGA: The Cancer Genome Atlas, GEO Gene Expression Omnibus, BrCa: Breast Cancer, cfDNA: cell-free DNA, AutoML Automated Machine Learning. The figure was designed using Biorender software (<https://app.biorender.com/gallery>).

Methods

In-silico analysis

Data sources

Raw DNA methylome data from BrCa and non-diseased breast tissues along with corresponding clinical data, were obtained from TCGA (www.cancer.gov/tcga, accessed on 5 February 2020) and GEO (www.ncbi.nlm.nih.gov/geo, accessed on 10 February 2020) [18] databases. Notably, the term non-diseased breast tissue refers to the following types of samples: In the TCGA-BRCA project, these tissues are designated as “Solid Tissue Normal,” meaning non-tumor breast tissue resected from the same patient at the time of surgery, typically adjacent to the tumor. In the two GEO studies, non-diseased breast tissue biopsy specimens were obtained from cancer-free women, as specified in their descriptions. The BrCa tissues and non-diseased breast tissue groups exhibited the same median age, and all normal breast tissues from the

TCGA project were matched for age and ethnicity, as they were derived from the same individual TCGA case inclusion criteria were: 1. Platform: Infinium Human Methylation 450 K bead-chip 2. Primary site: breast; 3. Project: TCGA-BRCA; 4. Gender: female; 5. Age at diagnosis: 26–80 years; 6. Race: white, black or African American and Asian. A total of 730 cases were downloaded. The GEO database was searched using ‘Breast cancer’ and ‘Metastatic Breast cancer’ as keywords and ‘Methylation profiling by array’ as study type: 84 and 10 studies were found, respectively. Those using the Infinium Human Methylation 450 K bead-chip array and providing adequate raw IDAT files along with clinical annotations, including age, disease status, and information on metastasis, were selected for further analysis. In addition, all GEO cohorts included samples deriving from female patients, and those with diagnosis between the ages of 26 and 80 years were selected for further analysis. Subsequently, the following five studies were selected, namely GSE72245 [19], GSE72251 [19], GSE88883 [20], GSE108576 [21], and GSE74214 [20].

Data preprocessing and DNA methylation analysis

To ensure that all samples underwent identical quality control and normalization, thereby harmonizing signal intensities and batch effect across studies, downloaded raw DNA methylation data (IDAT files) and sample annotation files were subjected to the Bioconductor R package RnBeads v2.0 [22] for DNA methylation analysis as described in our previous work [8]. Samples with adequate clinical and IDAT files were further processed using the RnBeads pipeline. During quality control, samples with polymorphic probes or probes outside of the CpG context, as well as probes on sex chromosomes, and probes/samples with the highest fraction of unreliable measurements were automatically excluded by RnBeads. After filtering and normalization, 705 high-quality samples including 520 breast cancer tissues and 185 non-diseased breast tissues were used for downstream analysis.

In our workflow, promoters (1500 bp upstream to 500 bp downstream of the transcription start site (TSS)) were chosen as the genomic region of interest and analysis was made to distinguish between BrCa tissues and non-diseased breast tissues. Normalized methylation beta-values for each gene promoter were generated, representing the average methylated probe intensity divided by the overall intensity (the sum of methylated and unmethylated probe intensities), plus an offset of 100 [23]. Methylation values are expressed as decimal values between 0.0 (no methylation) and 1.0 (full methylation). Also, Differentially Methylated Promoters (DMPs) between the two groups were automatically ranked by RnBeads, and the 300 top-ranking promoters were chosen for further functional analysis. We apply this cut-off,

as the top 300 DMPs presented strong statistically significant scores based on FDR (between $2.49\text{E-}148$ and $8.07\text{E-}66$) and correspond to a sufficiently large and manageable number of genes for a meaningful functional enrichment analysis. Gene ontology (GO) analysis of the 300 DMPs was carried out using the DAVID tool [24]. The GO annotation covers three domains: molecular function, the activities of a gene product at the molecular level; cellular component, the location of a gene in parts of a cell or its extracellular environment; biological process, chemical reactions that a gene is participating or other events that are involved and are pertinent to the functioning of an organism. The Protein–Protein Interaction (PPI) network of the 300 DMPs was constructed using the STRING database (version 12.0) [25]. Finally, the REACTOME database was used for pathway analysis of the 300 DMPs [26].

Identification of BrCa-specific methylation biomarkers

To identify methylation biomarkers, we performed feature selection on the whole 28,591 normalized methylation beta-values of the promoters via the AutoML platform JADBio (version 1.2.8.) [27] that identifies a minimum set of biomarkers bearing the maximal classifying ability. JADBio applies to low or high-sample data, and high or low-scale omics data, and produces accurate predictive models estimating the out-of-sample model's performance after bootstrap correction and cross-validation. JADBio preprocesses data, including mean and mode imputation, constant removal, and standardization. There is no need for selecting appropriate algorithms, nor tuning their hyperparameter values. The available classification algorithms are random forest classification, support vector machine (SVM) and ridge logistic regression and classification decision trees. JADBio is also suitable for small sample size datasets through a stratified, K-fold, repeated cross-validation BBC-CV algorithm protocol [28]. The predictive performance of the models was primarily evaluated using the Area Under the Curve (AUC), with values >0.5 considered better than random guessing. To provide a more comprehensive assessment of model performance, we also report additional metrics including precision, accuracy, balanced accuracy, specificity, sensitivity (recall), and positive and negative predictive values.

Biological interpretation

Auto ML-identified biomarkers were further studied by bioinformatic tools to reveal their biological role and relevance to BrCa carcinogenesis. Gene Ontology (GO) and pathway data for each respective gene were retrieved via the GeneCards Suite [29]. Then, we employed the multiple UniReD tool [30] to identify any functional relationships of respective proteins. Multiple UniReD is a mining

tool of published biomedical literature that associates the proteins of interest (query list) to a list of reference proteins that are known and verified to be involved in the disease under investigation (reference list). An IA score for each protein of interest is produced that signifies its relatedness to the proteins in the reference list. The higher the score the higher the functional association of the proteins of interest to the reference list proteins. Finally, the STRING database was used to reveal protein interactions between emerging biomarkers and their neighboring genes.

Clinical validation study

Clinical samples

Blood samples were collected from two different Oncology clinics: The Department of Medical Oncology of the University General Hospital of Alexandroupoli (UGHA) (Onco₁ group) and the Department of Medical Oncology of the University Hospital of Heraklion (UHH) (Onco₂ group). Patients who visited the clinics between 2011–2020 and 2013–2016, respectively, were included in the study. Follow-up data until 2023 were also available. Blood samples were collected following diagnosis from a total of 195 BrCa patients, all Caucasian, and allocated in distinct groups according to patient status, as follows: 1) 131 patients having undergone surgery for primary breast cancer a month before sampling and not yet initiated adjuvant therapy (adjuvant group), (b) 21 patients upon diagnosis of breast cancer, having no previous surgery, before the initiation of neo-adjuvant therapy (neo-adjuvant group), (c) 43 patients upon diagnosis for metastatic disease before the initiation of first-line chemotherapy (metastatic group). Treatment protocols included chemotherapy (taxanes and/or anthracyclines) and/or targeted therapy (such as hormonal or other monoclonal antibodies). The Response Evaluation Criteria in Solid Tumors (RESIST) criteria version 1.1 [31] was used to evaluate the treatment response of patients with metastatic disease at the first clinical check after first-line treatment and categorized into four groups: Complete Response (CR), Partial Response (PR), Stable Disease (SD), Progression Disease (PD).

The clinicopathological features for all patient groups for both clinics are presented in Table 1. Inclusion criteria were: 1) Age at diagnosis: 20–80, 2) Drug-naïve patients, 3) Primary cancer site: Breast, 4) Not previously diagnosed with another type of cancer or had a second primary cancer, 5) female sex. Information on PAM50 molecular subtypes was not available for the studied patients. However, data on estrogen receptor (ER), progesterone receptor (PR), and Human Epidermal Growth Factor Receptor 2 (HER2) status were available and used to categorize tumors according to molecular subtypes. Approvals were obtained by the Scientific Board of

UGHA and UHH, following an assessment by the Ethics Committee (decision 14/895/28.11.11 and 9286/15-01-2013) and were conducted according to the ethical principles of the 1964 Declaration of Helsinki and its later amendments. In parallel, blood samples were also collected from 132 healthy control women recruited from the blood donation unit of the UGHA (Table 1). Inclusion criteria for controls were: (1) age-matched with breast cancer patients, (2) female sex, (3) Caucasian, and (4) no history of cancer or other major disease. Regarding ethnicity, all participants were Europeans of Greek origin. All participants signed a voluntary informed consent.

Pre-analytical procedures and qualitative assessment of cfDNA

Plasma was isolated within 2 h from blood sampling in EDTA-coated tubes following each clinic's different protocol. In the Onco₁ group blood samples were centrifuged at 2000×g for 10 min and an additional high-speed centrifugation step at 14,000×g for 10 min was performed to remove cellular debris and contaminants. In the Onco₂ group blood samples were centrifuged at 2500×rpm for 15 min at 4 °C. The supernatant was then collected and subjected to a second centrifugation at 2000×g for 15 min at 4 °C.

Plasma samples were stored at −80 °C until further identical analysis for both Onco₁ and Onco₂ groups. CfDNA was quantified directly in unpurified plasma using a Qubit fluorometer 3.0 (Invitrogen Ltd., Life Technologies, UK) as previously described [7]. Then, cfDNA was extracted automatically from 1200 µL of plasma using the MagCore Plasma DNA Extraction kit in the MagCore system (RBRCA Bioscience, New Taipei City, Taiwan) according to the manufacturer's instructions, and its quantity was estimated by Qubit. Extracted cfDNA samples were stored at −20 °C until further processing. The extracted DNA quality was assessed by a Taqman probe-based qPCR assay using the nuclear Glyceraldehyd-3-phosphat-dehydrogenase (GAPDH) reference gene, as previously described [16]. Primers and probe of the GAPDH assay are presented in Supplementary Table 1. Samples with a quantification cycle (Ct) > 35 were excluded from further analysis, this applied to only 3–4 samples, that not included in the total samples used in the study.

Bisulfite conversion was performed by EZ DNA Methylation-Gold™ Kit (ZYMO Research Co., Orange, CA) as described by the manufacturer. Specifically, 40 µl of extracted cfDNA was subjected to conversion, and cfDNA was then eluted to a 20 µl volume and stored at −80°C. A fixed volume of 40 µl of extracted cfDNA for each sample was used for bisulfite conversion to maximize sensitivity, particularly given the relatively low abundance of cfDNA. After conversion, a methylation-independent

Table 1 Demographic and clinicopathological characteristics of BrCa patients and Healthy women volunteers

	Healthy (n = 132)	Total BrCa patients (n = 195)	Onco ₁ group (n = 137)	Onco ₂ group (n = 58)
<i>Age</i>				
Mean (± SD)	57.32(± 14.66)	58.17(± 13.28)	57.69(± 13.30)	60.16(± 13.17)
Median (range)	57(26–80)	58.50(30–80)	58(30–80)	61(33–80)
<i>Patient status</i>				
Adjuvant	–	131	95	36
Metastatic		43	21	22
Neoadjuvant		21	21	–
<i>Menopause</i>				
Yes		113	78	35
No		77	54	23
Not available		5	5	–
<i>Type</i>				
Ductal	–	152	103	49
Others		25	16	9
Not available		18	18	–
<i>Stage</i>				
I	–	27	20	7
II		57	40	17
III		59	47	12
IV		43	21	22
Not available		8	8	–
<i>Grade</i>				
1	–	4	4	
2		91	58	33
3		88	67	21
Not available		12	8	4
<i>Sites of metastasis</i>				
Lung	–	22	12	10
lymph nodes		12	4	8
Bone		20	10	10
Liver		13	2	11
Skin		2	0	2
Locally		7	1	6
Brain		1	0	1
Other		5	2	3
Visceral		31	13	18
Non-Visceral		30	12	18
Not available		2	1	1
<i>Hormone Receptor subtypes</i>				
ER/PR+	–	90	57	33
HER2+		65	53	12
TNBC		21	20	1
Not available		19	7	12
<i>ER status</i>				
Positive	–	131	94	37
Negative		43	36	7
Not available		21	7	14
<i>PR status</i>				
Positive	–	113	81	32
Negative		61	49	12
Not available		21	7	14
<i>Her2 status</i>				
Positive	–	65	53	12

Table 1 (continued)

	Healthy (n = 132)	Total BrCa patients (n = 195)	Onco ₁ group (n = 137)	Onco ₂ group (n = 58)
Negative		111	77	34
Not available		19	7	12

ER Estrogen receptor; PR Progesterone receptor; Her2 Human epidermal growth factor 2 receptor; TNB Triple negative breast cancer

Taqman probe-based dual-qPCR assay with non-CpG including primers for the β -actin (ACTB) and collagen type II alpha 1 chain (COL2A1) genes was used to account for variability in input DNA, to verify DNA quality after conversion, and to normalize results. Primers and probes were designed using the Oligo Primer Analysis Software v. 7 [32] and are presented in Supplementary Table 1. Each qPCR was carried out in 20 μ L of total reaction volume containing 10.9 μ L H₂O, 4 μ L 5X Platinum II buffer, 1.2 μ L Mg, 0.5 μ L dNTPs mix, 0.4 μ L of each ACTB and COL2A1 primer mix (10 μ M), 0.2 μ L of a 10 μ M CY5-labeled ACTB-probe, and 0.2 μ L of a 10 μ M HEX-labeled COL2A1-probe. For each reaction, 2 μ L of cfdNA was added. All qPCR reactions were performed using the Rotor-Gene 6000 Series (Qiagen, Darmstadt, Germany). The results were calculated using Rotor-Gene Software 1.7 (Qiagen). The analysis was performed using the RQ_{sample} (Relative Quantification) = $2^{-\Delta\Delta CT}$ method for each gene [33]. The mean RQ_{sample} value of ACTB and COL2A1 genes for each sample was used to normalize methylation results.

Methylation analysis in cfdNA

Methylation levels of identified gene promoters were analyzed using quantitative SYBR Green-based methylation-specific PCR (qMSP) assays. Primers specific for the methylated sequence of each gene promoter were newly designed using the MethPrimer software [34]. Primer sequences are provided in Supplementary Table 1. Each qPCR was carried out in 20 μ L of total reaction volume containing 4 μ L 5X Platinum II buffer, 0.8 μ L to 0.4 μ L Mg, 0.5 μ L dNTPs mix, 0.6 μ L to 0.4 μ L of each primer mix (10 μ M) for the methylated sequence of the *CLDN15*, *MKGPRD* and *ZNF430* promoters, 0.6 μ L 1:10,000 diluted SYBR® Green I dye (Molecular Probes, Inc., Invitrogen Ltd) and RNase-free water till the total volume of 19 μ L. For each reaction, 1 μ L of cfdNA was added. Extensive optimization was performed to develop robust qMSP assays. The specificity and cross-reactivity of primers were evaluated using unconverted gDNA and converted methylated and non-methylated DNA standards. The analytical specificity of qMSP assays was assessed using mixes of converted methylated and non-methylated DNA standards (100%, 50%, 10%, 1%, 0%). The analytical sensitivity of assays was evaluated using serial dilutions of converted methylated DNA controls in H₂O. The reproducibility (calculated as coefficients of variation, CVs), efficiency, and linearity were also evaluated

to complete the validation file of the established assays. The results were calculated using the Rotor-Gene 6000 Series Software 1.7 (Qiagen). Relative methylation levels were analyzed using the RQ_{sample} (Relative Quantification) = $2^{-\Delta\Delta CT}$ method [33]. Specifically, $\Delta\Delta CT$ values were generated for each target after normalization by the mean RQ_{sample} value of ACTB and COL2A1 and using 1% methylated control as a calibrator. An amplification signal > 40 cycles was considered negative ($RQ = 0$). For statistical comparisons, RQ values were analyzed as continuous variables. For categorical variables analyses, a positivity cut-off was applied, and samples with $Ct \leq 40$ were considered positive and methylated (methylated value = 1), while at $Ct > 40$ they were considered unmethylated (methylated value = 0).

Statistical analysis

The Kolmogorov–Smirnov test was used to check for normality in the distribution. A Kruskal–Wallis test was applied to compare continuous variables between subgroups and Mann–Whitney U test was also applied to compare continuous binary variables. In the case of categorical variables, the chi-square test was applied. The Spearman(r) correlation was used to compare two continuous variables. Survival curves were calculated using the Kaplan–Meier, and comparisons were performed using the log-rank test. We used overall survival (OS), progression-free survival (PFS), and disease-free survival (DFI) as end-points. Cox proportional hazards regression was applied to investigate the relationship between OS or DFI and independent variables like age or cancer stage. All statistical tests employed in our analysis were two-sided. Statistical significance was placed at a p -value < 5×10^{-2} . No correction for multiple testing was applied, as the analyses were based on a limited number of predefined, hypothesis-driven comparisons. Continuous variables are expressed as median (minimum–maximum) or mean $\pm$ standard deviation. Categorical variables are shown as absolute frequencies. Statistical analysis was conducted with the IBM SPSS 19.0 statistical software (IBM Corp. 2010. IBM SPSS Statistics for Windows, Version 19.0., Armonk, NY, USA).

Biosignature development

To develop optimal BrCa-specific biosignatures of clinical importance, we created datasets with patients’ methylation measurements and clinical, and demographic data. Methylation values were expressed as continuous

variables using the RQ values and as categorical variables (methylation positives and negatives) as described above. Then, we further analyzed them by the AutoML technology JADBio [27]. For the analysis, we used an extensive model tuning effort, and we chose the AUC metric for performance optimization and report additional metrics, including precision, accuracy, balanced accuracy, specificity, sensitivity (recall), and positive and negative predictive values, to provide a more comprehensive assessment of model performance. To ensure that biosignatures are robust and well-performing on unseen data, we divided the data into the training (70% of samples) and the validation (30%) sets.

Results

Differential promoter methylation in silico analysis between BrCa and healthy tissues

Differential Rnbeads analysis of raw genome-wide methylomes of 28,591 gene promoters from 520 breast cancer tissues and 185 non-diseased breast tissues revealed Differentially Methylated Promoters (DMPs), and the 300 top-ranking were further analyzed as per gene ontology and pathway analysis. The complete list of the 300 DMPs is presented in Supplementary Table 2. All these DMPs were hypermethylated in BrCa in relation to normal tissues. Gene ontology analysis by DAVID tool is provided in Supplementary Fig. 1. Molecular function analysis showed enrichment mostly in DNA-binding transcription factor activity and in RNA polymerase II cis-regulatory region DNA binding. For biological process enrichment analysis, DMPs were found to mainly regulate RNA polymerase II transcription and neuroblast proliferation. Finally, the cellular component analysis showed mainly chromatin and nucleus enrichment. The Protein–Protein Interaction (PPI) network was constructed using the STRING database (version 12.0) and showed statistically significant interactions between the respective 300 DMPs as demonstrated in Supplementary Fig. 2. REACTOME database showed enrichment of DMPs in developmental biology, transcription, neuronal system, and signal transduction pathways (Supplementary Fig. 3).

In silico identification of BrCa-specific methylation biomarkers

The promoter methylomes from the breast cancer tissues and the non-diseased breast tissues were also analyzed using the AutoML platform JADBio to build BrCa-specific methylation models and identify novel biomarkers. The whole normalized methylation beta-values dataset that corresponded to 28,581 promoters was uploaded to the platform rather than the top DMPs in order to perform an unbiased feature selection. Analysis was conducted automatically using various machine-learning

algorithms. The workflow also included the training of the model, the optimization of the hyperparameters, as well as the post-analysis of the output model. The original dataset (520 primary and metastatic BrCa and 185 normal tissues) was automatically and randomly split into a training dataset of 364 BrCa and 130 normal tissues and a validation dataset of 156 BrCa and 55 normal tissues. Analysis of the training dataset delivered a biosignature comprising the minimal set of features with the highest classification ability between BrCa and health. Specifically, a biosignature was built via Ridge Logistic Regression, containing three feature protein-coding gene promoters, namely *CLDN15*, *MRGPRD*, and *ZNF430* (Fig. 2A). JADBio trained 3017 different model types. Each one was employed many times during cross-validation (a repeated tenfold CV without dropping, max. repeats=20), leading to fitting 30,170 model instances. In discriminating BrCa from healthy tissues, this signature reached an AUC of 0.994 (Confidence Interval (CI): 0.985–0.999). The overall model performance, along with detailed metrics including precision, accuracy, balanced accuracy, specificity, sensitivity (recall), and positive and negative predictive values, are available at the following links: (<https://app.jadbio.com/share/ebda6680-6d06-4580-a46d-c55cab2179dc>). Upon validation, the model showed an AUC of 0.99, verifying the stability and accuracy of the model's performance. The model's performance and validation results are depicted in Fig. 2B and D.

The biological role of the respective genes of DMPs included in the model, i.e. *Claudin 15* (*CLDN15*), *MAS Related GPR Family Member D* (*MRGPRD*), *Zinc Finger Protein 430* (*ZNF430*), was further studied. According to Gene Card, they participate in transcription, cell junction, and adhesion and are plasma membrane and nucleus components. A detailed description of the biological role of each gene is presented in Table 2. To examine the potential role of their protein products in BrCa carcinogenesis, we used the literature mining tool multiple UniReD to assess functional associations between proteins according to published data, as previously applied [11]. A reference list of proteins known to be involved in breast cancer pathophysiology was used as a reference list in UniReD. This list included 10 proteins (Supplementary Table 3) selected through manual curation of literature reporting functional or genetic involvement in BrCa. Each entry includes UniProt ID, gene/protein name, and citation to the original studies. This curated list served as the validated “reference protein list” against which UniReD identified known functional associations with the novel proteins. *CLDN15* had a moderate association with BrCa, reaching a score of 4 out of 10, while *MRGPRD* reached a score of 2.5 out of 10, indicating a less known association with BrCa

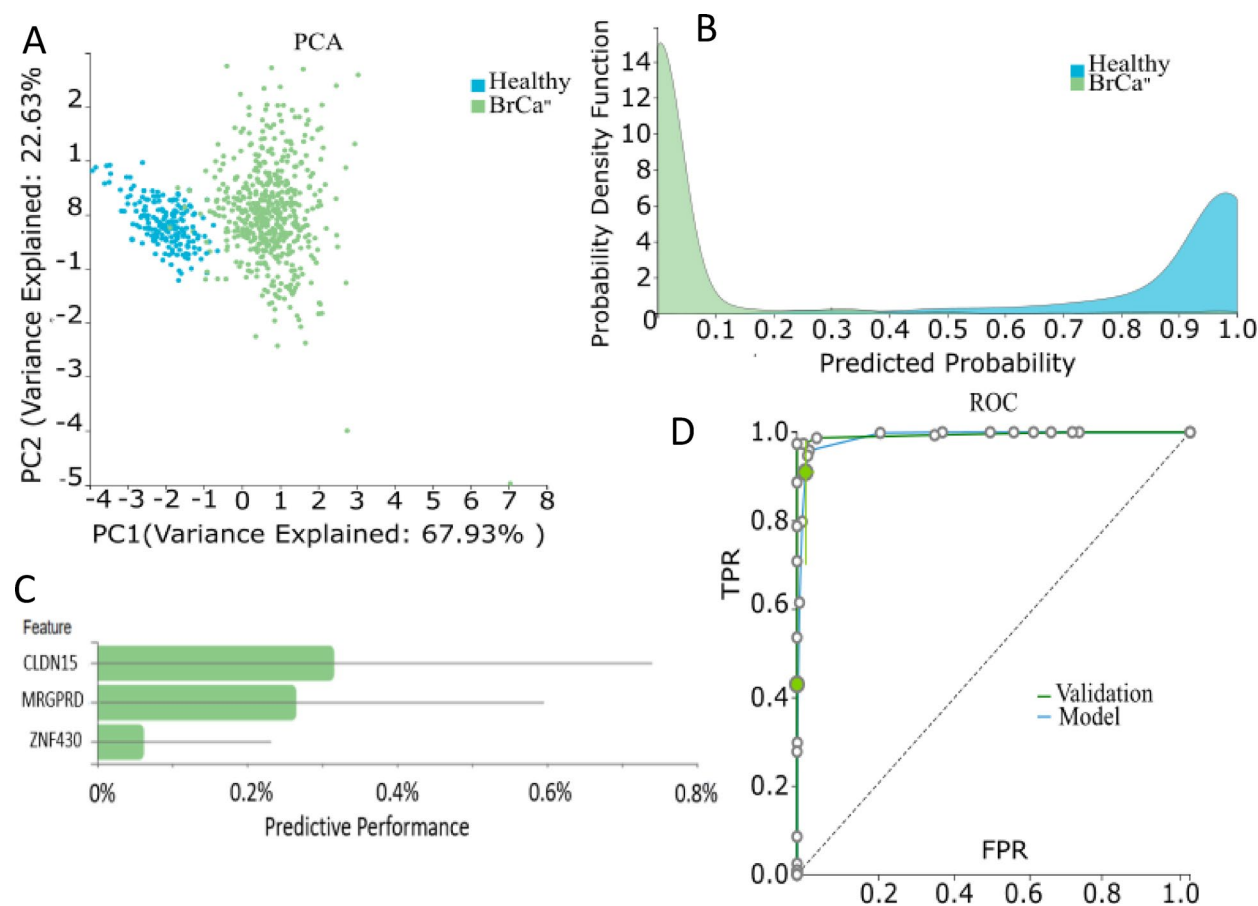


Fig. 2 BrCa-specific promoter methylation biosignature built on *in-silico* genome-wide methylation data. **A** PCA plot shows the separation between BrCa patients (green) and healthy individuals (blue), **B** Probability density plot depicting distinct distributions among BrCa patients and healthy individuals, **C** Feature Importance plot of the features of the model. Feature importance is defined as the percentage drop in predictive performance when the feature is removed from the model. **D** ROC curves of training (blue line) and external validation (green line) models. Abbreviations: ROC: Receiver Operating Characteristic; PCA: Principal Component Analysis; TPR: True Positive Rate; FPR: False Positive Rate

Table 2 Genes with DMPs selected by AutoML in the classifying model discriminating BrCa from healthy breast tissues: bioinformatic search of their biological characteristics and functions of their respective proteins

Gene	Description	Gene type	Pathway	GO-Molecular function	GO-Cellular component	GO-Biological process
CLDN15	Claudin 15	Protein coding	Cell junction organization, Blood–Brain Barrier, Immune Cell Transmigration	Protein binding and structural molecule activity	Plasma membrane, Nucleus	Cell adhesion, monoatomic ion transport, bicellular tight junction
MRGPRD	MAS Related GPR Family Member D	Protein coding	Signaling	Functions as a specific membrane receptor	Cell membrane	G protein-coupled receptor signaling pathway
ZNF430	Zinc Finger Protein 430	Protein coding	Transcription	DNA-binding transcription factor activity, RNA polymerase II-specific DNA binding	Nucleus	Regulation of transcription by RNA polymerase II and DNA-templated transcription

GO Gene ontology; Gene Ontology was carried out using the GeneCards database and covers three domains: molecular function, the elemental activities of a gene product at the molecular level; cellular component, the parts of a cell or its extracellular environment; biological process, chemical reactions, or other events that are involved and are pertinent to the functioning of integrated living units

according to published literature. Multiple UniReD did not include records for ZNF430. STRING database was used to reveal protein interactions between products of these and neighboring genes (Supplementary Fig. 4).

CLDN15 was found to interact with proteins of the Claudin family, all participating in tight junction integrity. Transcription factor ZNF430 was shown to interact with the still uncharacterized C19orf44 and MRGPRD with

the MRGPRES, also bearing G protein-coupled receptor activity and the NPFF neuropeptide. The above results showed that the emerged methylation biomarkers are located in genes quite novel and barely studied in BrCa.

Clinical validation of novel methylation biomarkers

Quantification of cfDNA in BrCa patients and healthy individuals

The concentration of cfDNA was measured directly in plasma samples using the Qubit fluorometer. Levels of cfDNA in cancer patients were significantly higher in

relation to the healthy volunteer control group ($U = 3135$, $Z = -4.92$, $p < 0.001$, Fig. 3A). In specific, the adjuvant group, the metastatic group, and the neo-adjuvant group all presented higher cfDNA levels in relation to the healthy group ($U = 2207$, $Z = -4.47$, $p < 0.001$, $U = 438.5$, $Z = -3.20$, $p = 0.001$ and $U = 489.5$, $Z = -2.76$, $p = 0.006$, respectively) (Fig. 3B). Extracted cfDNA levels in the adjuvant group above or equal to $0.25 \text{ ng}/\mu\text{l}$ correlated with shorter OS ($p = 0.003$) and DFI ($p = 0.005$) (Fig. 3C, 3). Cox proportional hazards regression analysis showed that the covariates of age and stage were not significantly

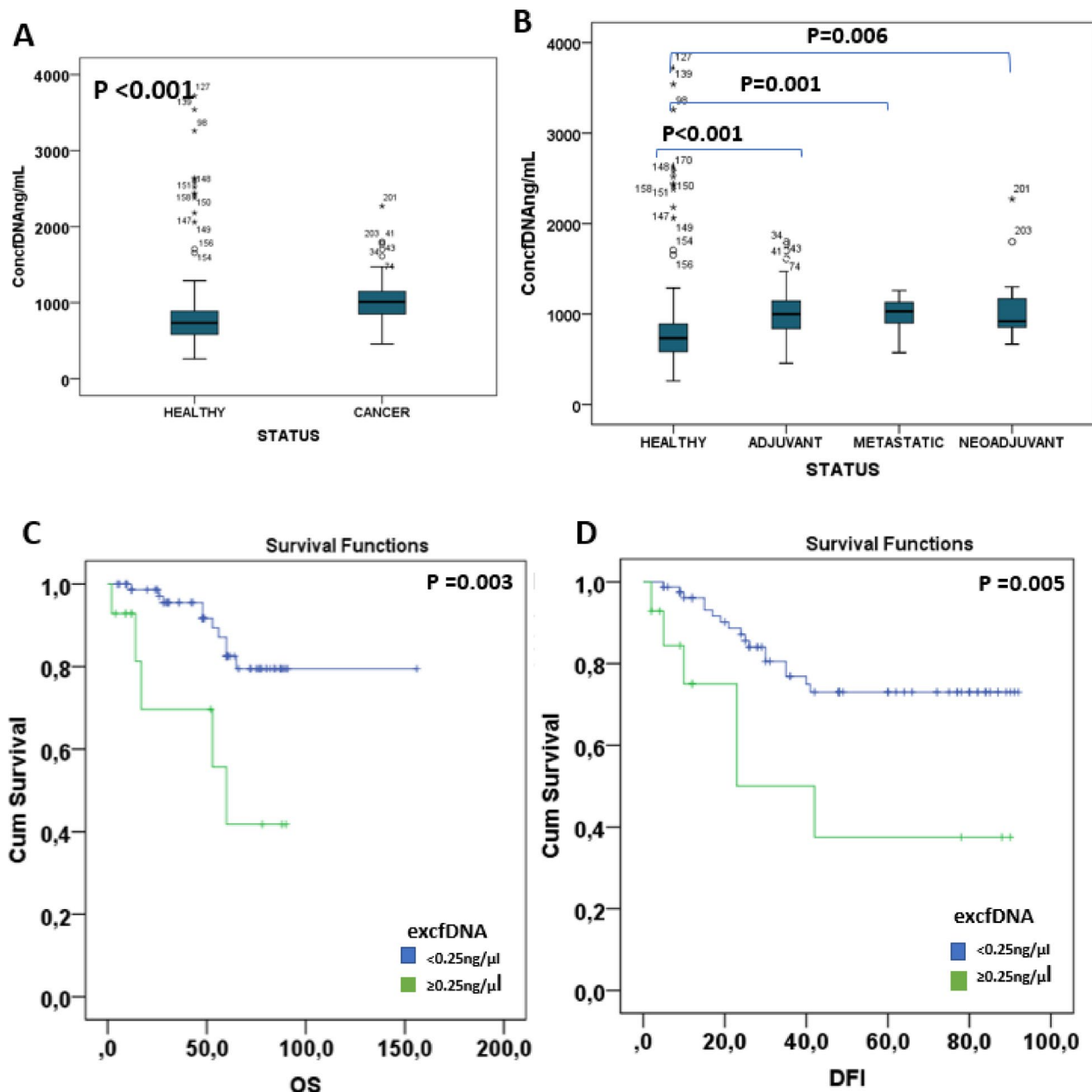


Fig. 3 Mean cfDNA concentration as quantified directly in the plasma of healthy individuals and BrCa patients (A); and in BrCa subgroups (B). Kaplan-Meiers depicts OS (C) and DFI (D) in relation to the median value of extracted cfDNA concentration ($0.25 \text{ ng}/\mu\text{l}$) in the adjuvant group of patients. Abbreviations: OS: Overall Survival, DFI: Disease-Free Interval

associated with the OS ($p=0.858$), ($B_{\text{stage}}=0.18$, Hazard ratio (HR)=1.20, CI: 0.17–8.60), ($B_{\text{age}}=-0.11$, HR=0.90, CI: 0.27–2.99) or DFI ($p=0.055$), ($B_{\text{stage}}=0.72$, HR=2.05, CI: 0.98–4.29), ($B_{\text{age}}=-0.37$, HR=0.69, CI: 0.47–1.01) (Supplementary Fig. 5). No other correlation of cfDNA quantity with patient clinicopathological characteristics like stage or with demographic data like age has emerged (Supplementary Fig. 5).

Quantification of DMPs methylation biomarkers in BrCa patients and healthy individuals

At first, the analytical sensitivity and efficiency of all qMSP assays developed were evaluated. All assays could detect down to 0.01 ng of methylated DNA. Efficiency was 100% ($R^2=0.99$, slope = -3.32) for *MRGPRD*, 110% ($R^2=0.98$, slope = -3.10) for *CLDN15* and 101% for *ZNF430* ($R^2=0.97$, slope = -3.30) and 1% of methylated alleles among 99% of unmethylated alleles could be detected (Supplementary Fig. 6).

Promoter methylation of *CLDN15*, *MRGPRD*, and *ZNF430* was investigated in the converted cfDNA of all 195 BrCa patients and 132 healthy individuals. Methylation levels were higher in cancer in relation to the healthy group for all studied gene promoters (*CLDN15* $U=10,614$, $Z=-3.02$, $p=0.002$, *MRGPRD* $U=9460$, $Z=-4.14$, $p<0.001$, *ZNF430* $U=11,968$, $Z=-2.12$, $p=0.034$) (Fig. 4A, C, E). In specific, adjuvant and metastatic groups of patients had higher methylation compared to the healthy group for *MRGPRD* (adjuvant $U=6429$, $Z=-3.70$, $p<0.001$, metastatic $U=1840$, $Z=-3.47$, $p<0.001$, Fig. 4D) and *ZNF430* (adjuvant $U=7128$, $Z=-4.03$, $p<0.001$, metastatic $U=2371$, $Z=-3.00$, $p<0.001$, Fig. 4F) promoters. Additionally, the metastatic and neoadjuvant groups showed higher methylation levels in relation to the healthy group for *CLDN15* ($U=1826$, $Z=-3.85$, $p<0.001$; $U=855$, $Z=-3.14$, $p=0.003$, respectively; Fig. 4B). Between BrCa patients, higher methylation levels of *CLDN15* and *MRGPRD* promoters were noticed in stage III BrCa in relation to stage II ($U=946$, $Z=-2.86$, $p=0.004$) and stage I ($U=375$, $Z=-2.49$, $p=0.013$), respectively (Fig. 5A, 5). Median relative methylation values, along with their minimum and maximum values, for different patient statuses and stages are presented in Table 3. Statistically significant correlations also emerged when methylation was considered binary as methylation positives and negatives. Specifically, all genes were more often methylated in the metastatic group as compared to the healthy group (*CLDN15*: Chi-square=8.28, $p=0.006$, *MRGPRD*: Chi-square=5.40, $p=0.013$, *ZNF430*: Chi-square=7.68, $p=0.015$). Furthermore, positive methylation of at least two genes was more often present in the metastatic and neo-adjuvant groups of patients (Chi-square=12.54, $p<0.001$, Chi-square=5.81, $p=0.017$, respectively) in

relation to the healthy group of patients. Additionally, positive methylation of all three genes was observed more frequently in the metastatic group of patients compared to the healthy group (Chi-square=16.12, $p<0.001$).

Analysis of 131 adjuvant women showed that those who relapsed more often presented higher levels of *CLDN15* promoter methylation ($U=1451$, $Z=-2.09$, $p=0.037$) (Fig. 5C). Also, liver metastasis was associated with higher methylation levels of *ZNF430* promoter ($U=358$, $Z=-2.45$, $p=0.014$) (Fig. 5D). Similarly, in the group of metastatic patients, higher levels of *CLDN15* methylation were associated with the recurrence of the disease ($U=107$, $Z=-2.06$, $p=0.044$), especially with liver metastasis ($U=120$, $Z=-2.02$, $p=0.048$) (Fig. 6A, 6). Also, survival analysis showed that frequent methylation of *MRGPRD* promoter in the metastatic group of patients was associated with shorter PFS (Chi-square=4.79, $p=0.029$) (Fig. 6C), and positive *ZNF430* methylation was associated with absence of response in the 1st-line treatment at the first check (Fig. 6D) (Chi-square=5.86, $p=0.043$). Additionally, *MRGPRD* methylation in all BrCa patients was correlated with age, showing higher levels in patients >58y (median value of age) as compared to younger patients ($U=2882$, $Z=-2.81$, $p=0.05$) (Supplementary Fig. 5). No other correlation was shown between gene methylation and grade, histological cancer subtype, tumor size, nodal, or vascular infiltration. Furthermore, no statistically significant correlations emerged in the 21 patients of the neo-adjuvant group, obviously limited by the small group size. These results demonstrate the relevance of the identified methylation biomarkers to important clinical parameters in BrCa.

Methylation-based BrCa biosignature development

The experimental methylation data combined with clinical and demographic patient data were analyzed by AutoML to build optimized BrCa biosignatures of clinical value for implementation in liquid biopsies. In the first classification analysis, the task was the discrimination between BrCa and health. In this AutoML analysis, JADBio trained 3481 different machine learning pipelines (also called configurations), corresponding to different model types. Each one was employed many times during cross-validation (a repeated tenfold CV without dropping, max. repeats=20), leading to fitting 104,430 model instances. Using data from 132 healthy and 195 BrCa patients, autoML produced a best-performing biosignature via the Classification Random Forests algorithm, consisting of six features including gene methylation and cfDNA concentration measurements (Fig. 7). AUC was 0.79 (CI: 0.75–0.84) and average precision was 0.81 (CI: 0.77, 0.85). The overall model performance and detailed metrics are available in the following link:

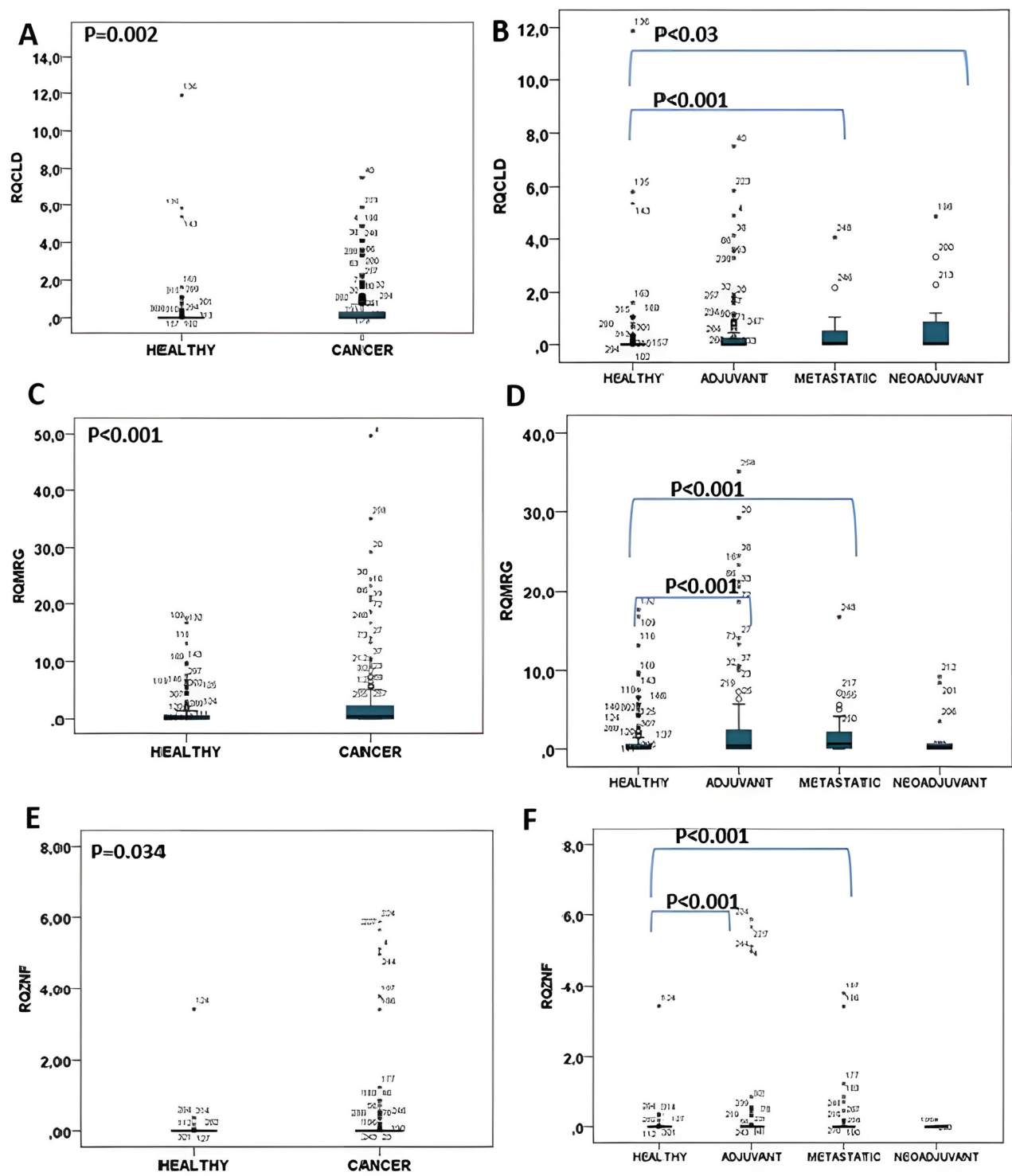


Fig. 4 Methylation levels of *CLDN15*, *MRGPRD*, *ZNF430* gene promoters as detected by qMSP in cfDNA: comparisons between the healthy and BrCa groups (A, B, C) and the BrCa subgroups (adjuvant, metastatic, neoadjuvant) (D, E, F). RQCLD: Relative Quantification of methylation for *CLDN15*, RQZNF: Relative Quantification of methylation for *ZNF430*, RQMRGPRD: Relative Quantification of methylation for *MRGPRD*

(<https://app.jadbio.com/share/e1c78f2e-538f-4754-a031-2e42679f2149>). AutoML at JADBio is specifically designed to prevent overfitting and to perform automatic internal validation,

ensuring that no samples are overlooked during the validation process [27, 28]. Still, in order to further confirm the robustness of our results, we applied an additional conventional validation strategy by randomly splitting

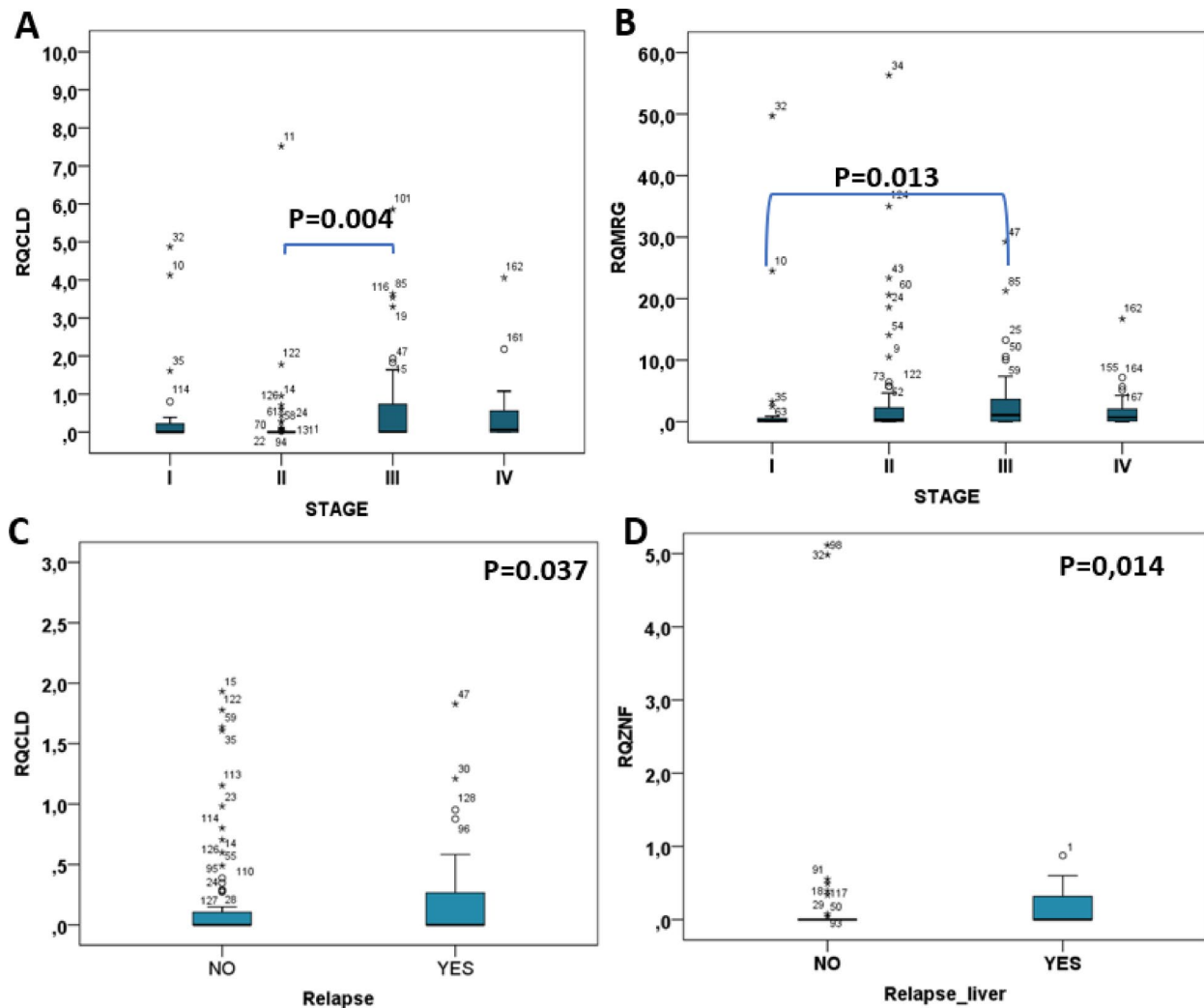


Fig. 5 Promoter methylation levels as detected by qMSP in cfDNA: among different BrCa stages of all BrCa patients for *CLDN15* (A), *MRGPRD* (B), and in relation to relapse and liver relapse in the adjuvant group of patients for *CLDN15* (C) and *ZNF430* (D) genes

the dataset into a training set (141 BrCa and 89 healthy participants; 70%) and a validation set (55 BrCa and 43 healthy participants; 30%). Using this approach on the same classification task, the training dataset generated a biosignature of seven features through Classification Random Forest, including the same features identified previously, along with the methylation of *CLDN15*, achieving a comparable AUC of 0.79 (CI: 0.74–0.84) and an average precision of 0.82 (CI: 0.77–0.85). (<https://app.jadbio.com/share/304ec46f-a4b1-4879-b023-904b735685be>). Validation showed an AUC of 0.76, confirming stable performance of the model (Supplementary Fig. 7).

The next task was a classification analysis distinguishing between healthy individuals and patients across different clinical statuses: adjuvant, metastatic, and neoadjuvant. JADBio trained 3481 different model types. Each one was employed many times during cross-validation (a repeated tenfold CV without dropping, max.

repeats = 20), leading to fitting 174,050 model instances. The analysis of 132 healthy individuals, 131 adjuvant, 43 metastatic and 21 neoadjuvant patients generated a best-performing biosignature using the Classification Random Forest algorithm, including gene methylation and cfDNA concentration measurements, with an AUC of 0.68 (CI: 0.62–0.72) and an average precision of 0.74 (CI: 0.69–0.77) (Fig. 8). The overall model performance and detailed metrics are available in the following link: (<http://app.jadbio.com/share/ddba0c04-c253-4b78-b3c2-31f1acee9dfe>). To further validate the model, we randomly split the same dataset into training (70%) (comprising 93 healthy, 95 adjuvant, 31 metastatic, and 11 neoadjuvant patients) and a validation set (30%) (comprising 39 healthy, 36 adjuvant, 12 metastatic, and 10 neoadjuvant patients). The training dataset produced a biosignature via Classification Random Forest using the same features, with an AUC of 0.71 (CI: 0.66–0.75) and an average

Table 3 Median relative methylation values, along with their minimum and maximum values, for different patient statuses and stages

Status	Healthy	Adjuvant	Metastatic	Neoadjuvant
RQ _{CLDN15}	0.001	0.020	0.058	0.043
Median(min–max)	(0.000–20.252)	(0.000–23.344)	(0.000–4.055)	(0.000–4.839)
RQ _{MGRPRD}	0.054	0.445(0.000–56.297)	0.648	0.170
Median(min–max)	(0.000–17.569)	(0.000–16.679)	(0.000–16.679)	(0.000–9.221)
RQ _{ZNF430}	0.000(0.000–3.410)	0.001(0.000–15.779)	0.648	0.000(0.000–22.238)
Median(min–max)	(0.000–3.410)	(0.000–15.779)	(0.000–16.679)	(0.000–22.238)
BrCa patient stages	Stage I	Stage II	Stage III	Stage IV
RQ _{CLDN15}	0.001(0.000–4.873)	0.001(0.000–23.344)	0.077(0.000–5.856)	0.065(0.000–4.055)
Median(min–max)	(0.000–4.873)	(0.000–23.344)	(0.000–5.856)	(0.000–4.055)
RQ _{MGRPRD}	0.158	0.298	1.094	0.712
Median(min–max)	(0.000–49.694)	(0.000–56.297)	(0.000–29.242)	(0.000–16.679)
RQ _{ZNF430}	0.000(0.000–5.115)	0.000(0.000–5.876)	0.000(0.000–15.779)	0.001(0.000–3.768)
Median(min–max)	(0.000–5.115)	(0.000–5.876)	(0.000–15.779)	(0.000–3.768)

RQ_{CLDN15} Relative Quantification of methylation for *CLDN15*; RQ_{ZNF430} Relative Quantification of methylation for *ZNF430*; RQ_{MGRPRD} Relative Quantification of methylation for *MGRPRD*

precision of 0.77 (CI: 0.74–0.81) (<https://app.jadbio.com/share/d3de7aa3-66e1-453f-888f-de1e2dc70071>). Validation confirmed consistent performance, yielding an AUC of 0.72 (Supplementary Fig. 8).

Next classification analysis was performed on all BrCa patients to develop a predictive model for relapse. JADBIO trained 3,481 machine learning pipelines. Each one was employed many times during cross-validation (a repeated tenfold CV without dropping, max. repeats = 20), resulting in the fitting of 278,480 model instances. The analysis of 174 BrCa patients yielded a best-performing biosignature using the Classification Random Forests algorithm. This biosignature included as features the methylation of all genes, disease stage, vessel infiltration, HER2 status, and hormonal therapy as a maintenance treatment. It achieved an AUC of 0.78 (95% CI: 0.74–0.83) and an average precision of 0.82 (95% CI: 0.78–0.85). (<https://app.jadbio.com/share/e82acb61-9bc2-47f2-8e0c-e872401705b0>) (Fig. 9). When the dataset was split (training: 121 BrCa patients, validation: 53 patients), the training set yielded a four-feature biosignature with the same features except HER2 status. This model achieved an AUC of 0.79 (95% CI: 0.74–0.86) and an average precision of 0.89 (95% CI: 0.79–0.88) (<https://app.jadbio.com/share/78479666-120c-407c-b30e-a3ba06a68edf>). In validation, the model reached an AUC of 0.71 (Supplementary Fig. 9).

In the metastatic patient cohort, a classification task was set to identify patients unlikely to respond to 1st line treatment and who will develop PD. JADBIO trained

3481 machine learning pipelines, each one was employed many times during cross-validation (a repeated tenfold CV without dropping, max. repeats = 20) resulting in the fitting of 111,392 model instances. From the analysis of 43 metastatic patients, a best-performing model was produced using the Support Vector Machine (SVM) algorithm, with 5 features: methylation of *CLDN15* and *ZNF430*, presence of bone metastasis at diagnosis, presence of liver metastasis at diagnosis, and type of first-line treatment. It achieved an AUC of 0.86 (95% CI: 0.67–1.00) and an average precision of 0.89 (95% CI: 0.79–1.00). (Fig. 10) (<https://app.jadbio.com/share/43b961fc-5229-4914-9627-fea760ede895>). When the dataset was split into a training set (24 patients) and a validation set (19 patients), a 4-feature biosignature comprising the methylation of *CLDN15*, *MGRPRD*, *ZNF430*, and ER status was built, with an AUC of 0.97 (95% CI: 0.87–1.00) in the training cohort (<https://app.jadbio.com/share/46d025a2-29fe-459e-a769-2a3696d329f5>). However, its performance did not generalize well, as in the validation set AUC was 0.44 (Supplementary Fig. 10).

Finally, in the adjuvant group of patients, a classification task was set to differentiate between distinct molecular subtypes (ER/PR+, HER2+, TNBC). The analysis of 131 patients of the adjuvant group resulted in a best-performing model via a Classification Random Forest algorithm that contained the following five features: methylation of *MGRPRD* and *ZNF430*, the quantity of extracted cfDNA, overall survival, and disease-free interval. AUC was 0.71 (95% CI: 0.64–0.77) and average precision of 0.79 (95% CI: 0.75–0.84). (<https://app.jadbio.com/share/44aca8ee-411d-4204-8ac9-f12b4a3f7136>) (Supplementary Fig. 11).

Discussion

Our study's ambition is to provide new liquid biopsy solutions in the diagnosis and management of BrCa. So far, there are few available FDA-approved blood-based tests. For example, the CellSearch® System is used in the prognosis of metastatic BrCa by enumerating circulating tumor cells (CTCs) [35]. While the Guardant360 CDx is appropriate for driving treatment choice in advanced ESR1 or PIK3CA-mutated BrCa [36], this is an option for a limited percentage of patients. Thus, there is an urgent need to discover new, reliable, minimally invasive tests in BrCa.

To this end, we introduce an AutoML-based pipeline that can identify BrCa-specific methylation biomarkers through feature selection and, upon laboratory validation in liquid biopsy, delivers diagnostic and prognostic biosignatures with clinical value. In total, we developed 5 sufficiently performing methylation-based biosignatures: 1) A diagnostic biosignature of BrCa with AUC of 0.79 discriminating BrCa from health, 2) A biosignature

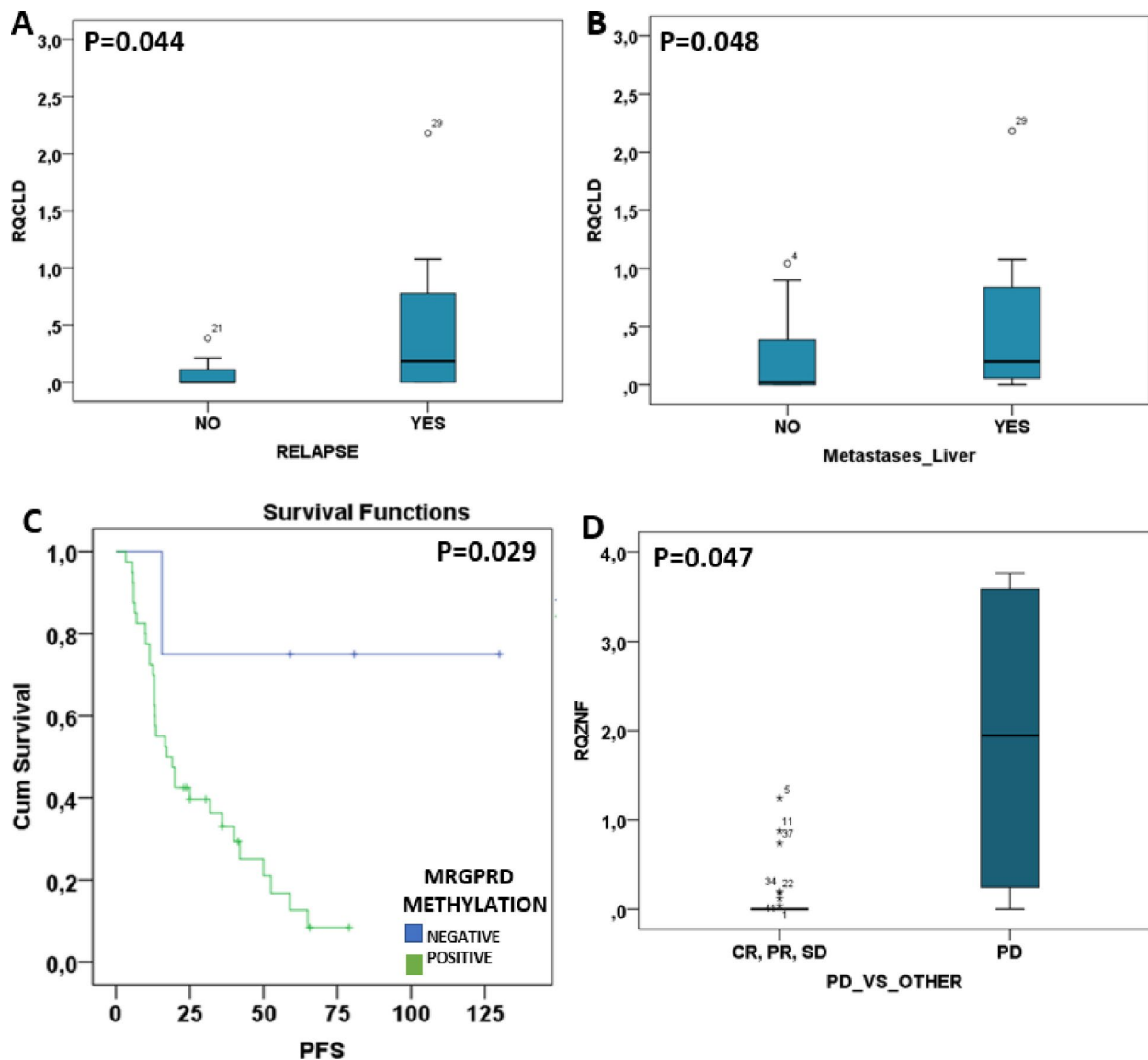


Fig. 6 Promoter methylation levels assessed by qMSP in cfDNA of the group of BrCa metastatic patients for *CLDN15* in relation to new relapse (**A**) or metastasis in the liver (**B**), and *ZNF430* in relation to response in the 1st line treatment (**D**). Kaplan–Meier depicts PFS (**C**) in relation to *MRGPRD* methylation. Abbreviations: CR: Complete Response, PR: Progressive Disease, SD: Stable Disease

classifying BrCa subgroups with an AUC of 0.68, 3) A prognostic biosignature for predicting relapse, with an AUC of 0.79, 4) A predictive biosignature for treatment response in the metastatic group with an AUC of 0.86 and 5) A biosignature classifying BrCa distinct molecular subtypes with an AUC of 0.71.

In the first place, in silico bioinformatics analysis of 28,591 gene promoters from 520 BrCa tissues and 185 non-diseased breast tissues led to the 300 top-ranking DMPs between BrCa and normal tissues. There was a great level of heterogeneity between the BrCa tissue samples, due to different disease stages, molecular subtypes, and race; however, this was necessary to increase the predictive power of the features identified at this discovery

stage and the robustness of the output. We chose to analyze gene promoters to identify candidate biomarkers as promoters are key regulators of gene expression, and their epigenetic silencing by DNA methylation is a well-established hallmark of cancer initiation and progression. In contrast, analyzing the methylation CpG-by-CpG in the whole gene or gene body is more often associated with alternative splicing and genomic instability, and therefore less suitable for biomarker development. All of these genes were hypermethylated in tumors in relation to healthy tissue. Indeed, it is well-established that methylation is a key event in cancer [37], leading to gene silencing and is strongly correlated with other cancer-associated events such as mutations, inflammation, and

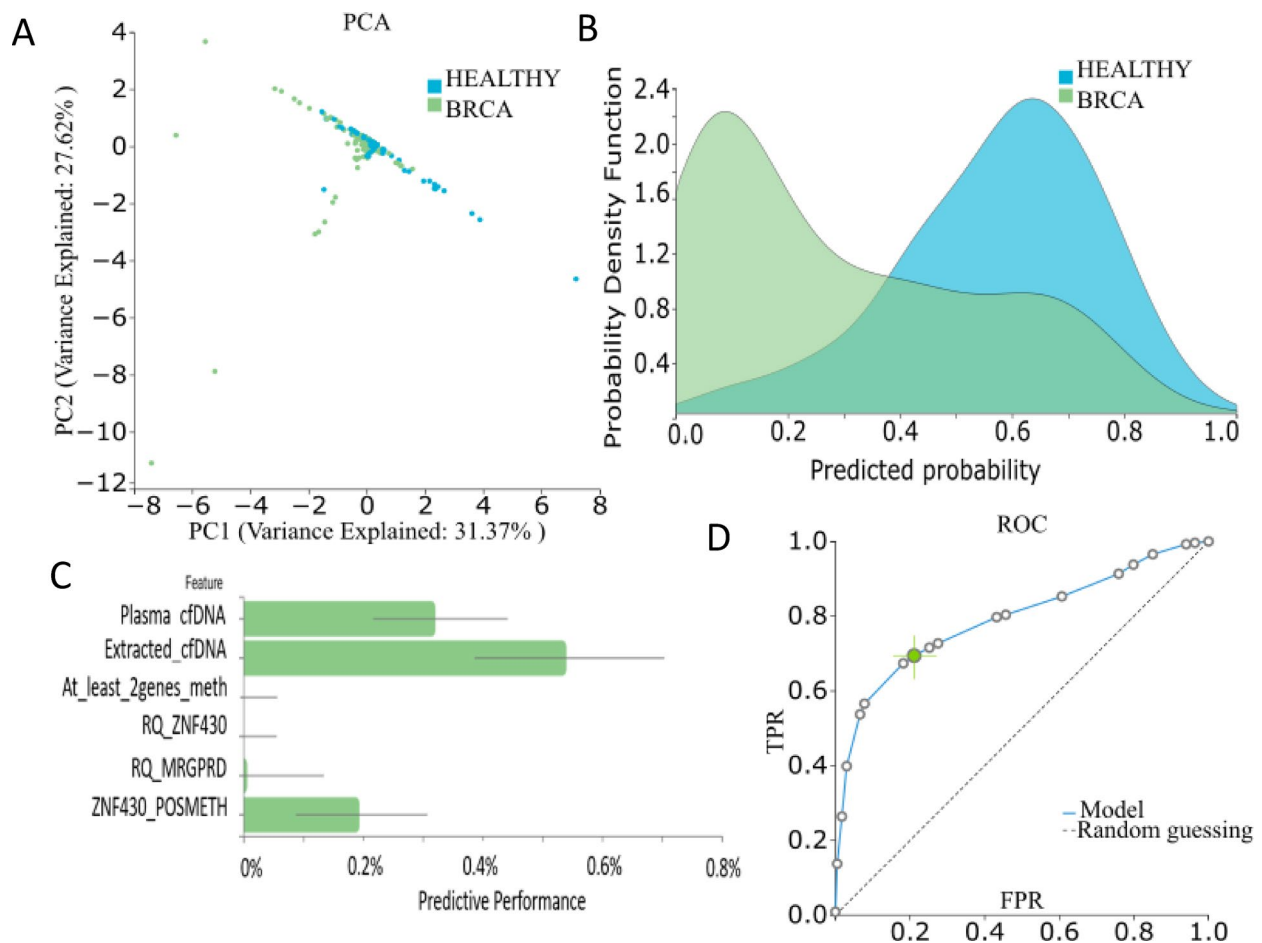


Fig. 7 BrCa diagnostic biosignature. **A** PCA plot shows the separation performance between BrCa patients (green) and healthy individuals (blue), **B** Probability density plot depicts distinct distributions among BrCa patients and healthy individuals, **C** Feature Importance plot of the features included in the model. **D** ROC curve of the model. Abbreviations: ROC: Receiver Operating Characteristic; PCA: Principal Component Analysis; TPR: True Positive Rate; FPR: False Positive Rate

hypoxia [38]. Gene ontology analysis of those 300 DMPs and respective genes showed mostly a regulatory role of gene expression e.g. DNA-binding transcription factor activity and RNA polymerase II cis-regulatory region DNA binding, indicating that they could be tumor suppressors. Alterations in these genes, like abnormal methylation, can disrupt cellular functions and contribute to carcinogenesis [39]. In addition, pathway analysis revealed enrichment of these genes in cancer-affected pathways like transcription and signal transduction [40].

Next, AutoML analysis by JADBio of 28,591 gene promoters delivered a biosignature comprising three DMPs (*CLDN15*, *MRGPRD* and *ZNF430*) showing the highest classification performance in the training dataset (AUC:0.994, CI: 0.985–0.999) and validation (AUC of 0.995). The emerged biomarkers belong to protein-coding genes. *CLDN15* participates in cell junction organization with other Claudin family proteins. It has been shown that tight junction disruption is a hallmark of Epithelial-to-Mesenchymal Transition (EMT), a critical

step in cancer progression and metastasis [41]. According to the literature, *CLDN15* was a positive marker in malignant mesothelioma [42] and lung adenocarcinoma [43]. On the other hand, downregulation of its expression was noticed in BrCa [44], colon cancer [45], and ovarian cancer, where low expression was associated with short OS [46]. *MRGPRD* and *ZNF430* are transcription factors and have a regulatory role in gene expression, which, upon distraction, could lead to cancer progression. Both of them were found to interact with almost unknown proteins: *MRGPRD* with *MRGPRED*, which has no previous records in relation to cancer or other pathology, and the *NPFF*, which was recently correlated with the progression of Ovarian Cancer [47], and *ZNF430* with the also understudied *C19orf44*. In a relevant study, *MRGPRD* promoted tumorigenesis in lung cancer and was highly expressed [48]. In contrast, a recent study showed that *ZNF430* was expressed at significantly lower levels in another hormone-dependent adenocarcinoma of the endometrium and was associated with OS [49]. Previous

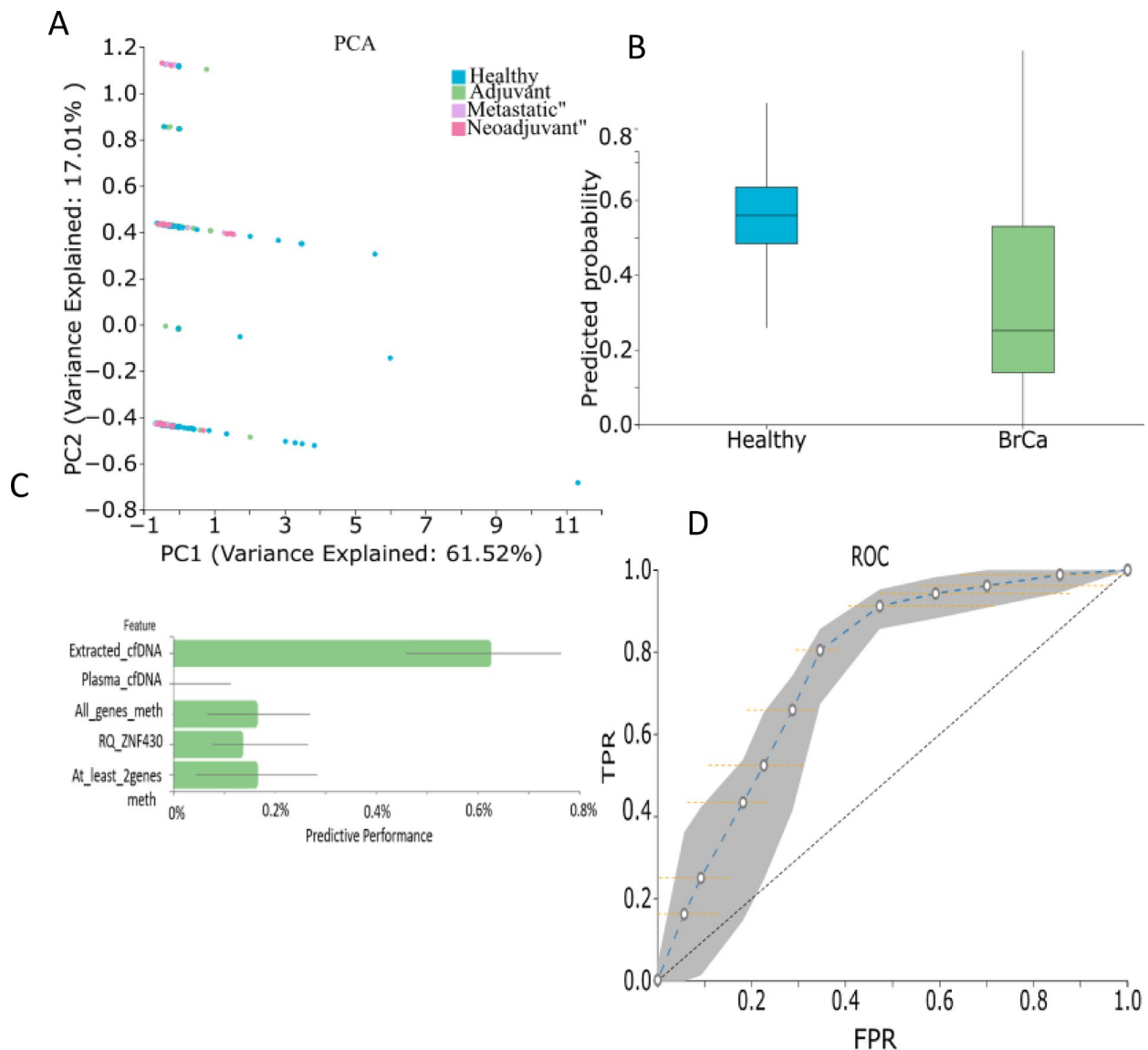


Fig. 8 BrCa classifying biosignature between healthy, adjuvant, metastatic, and neoadjuvant groups. **A** PCA plot, **B** Box-plot predicted probability between healthy and all BrCa groups **C** Feature Importance plot of the model. **D** ROC curve. Abbreviations: ROC: Receiver Operating Characteristic; PCA: Principal Component Analysis; TPR: True Positive Rate; FPR: False Positive Rate

studies have shown that many members of the Claudin family are methylated in promoter regions in BrCa and other cancers, leading to the downregulation of their expression [50]. Unfortunately, there are no available data for the *CLDN15*. On the other hand, methylation of *MRGPRD* in the blood of pregnant women was associated with exercise [51]. The above findings show that methylation is a mechanism that regulates the studied genes and could stand as a possible biomarker or therapeutic target upon investigation. To our knowledge, no other information is available for these genes in cancer, as also indicated by the multiUnired tool. For a glimpse into the expression of these proteins in breast and breast cancer, we analysed expression of the three identified genes

in the human atlas expression data and found detectable expression in breast, especially for *CLDN15* and *ZNF430*, in a non-tissue specific manner. In addition, we analysed strand-specific RNA sequencing data from 22 primary invasive breast cancer carcinoma expressing estrogen receptors and their paired adjacent mammary healthy tissues (GSE103001) and found no statistically significant differences (non-presented observation). We believe that larger studies and most importantly functional in vitro and in vivo experiments would be very useful for unraveling their role and contribution to biological processes in this setting. Our study is the first to demonstrate aberrant methylation of *CLDN15*, *MRGPRD*, and *ZNF430* promoters in BrCa. Beyond the potential significance

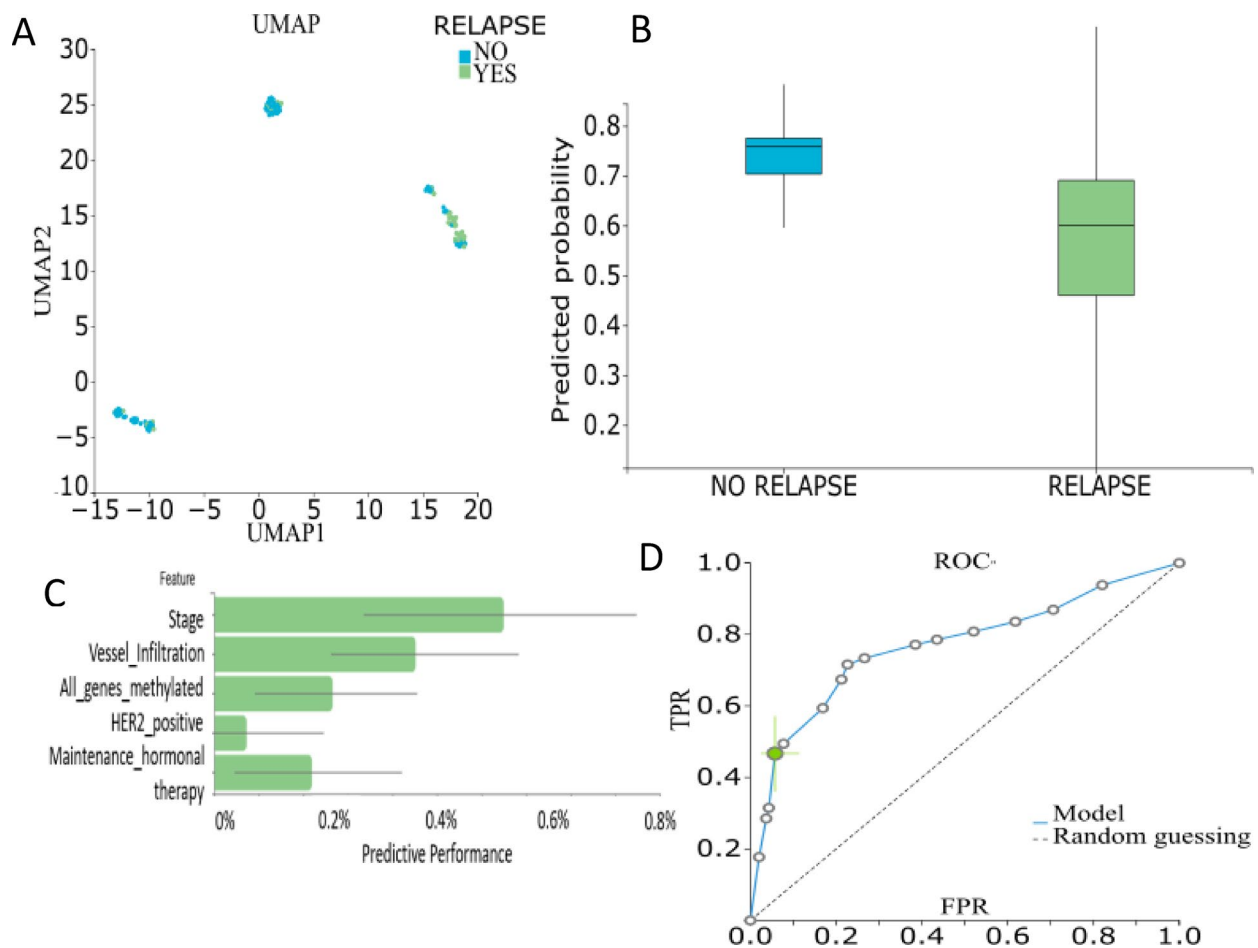


Fig. 9 BrCa prognostic methylation biosignature predicting relapse. **A** UMAP plot, **B** Box-plot predicted probability between those patients who relapsed and those who did not relapse **C** Feature Importance plot of the model. **D** ROC curve. Abbreviations: ROC: Receiver Operating Characteristic; TPR: True Positive Rate; FPR: False Positive Rate; UMAP: Uniform Manifold Approximation and Projection

of these methylation sites as biomarkers, the respective proteins could present new targets for therapeutic interventions, and this is worthy of further study. Moreover, we need to recognize how powerful this data-driven approach has been in unfolding novel genes of interest.

We pursued these results through an experimental pilot clinical validation in BrCa patient liquid biopsies. We first measured the quantity of cfDNA, finding it higher in BrCa patients than in healthy individuals, confirming previous results of our group [7] and others [12, 52]. This is a common finding among other types of cancers [53]. Measurement of plasma cfDNA concentration was effective in detecting lung cancer (AUC:0.94) and high-risk individuals [54]. Frattini et al. suggested cfDNA concentration as a diagnostic and monitoring tool in colon cancer [55]. Similar findings were also described in gastric cancer [56]. In our study, cfDNA levels were correlated with shorter OS and DFI in the adjuvant BrCa group, supporting some prognostic value. Whereas, we have previously shown that higher cfDNA levels were associated with shorter PFS in metastatic BrCa patients

[7]. As with other cancer types, it has been shown that increased plasma cfDNA was associated with poorer 5-year survival in lung cancer [57]. Similarly, the cfDNA fraction was a strong predictor of OS, PFS, and treatment response in advanced prostate cancer [58].

Interestingly, when the promoter methylation of the ML-identified genes was studied in the BrCa cfDNA significant differences emerged between BrCa patients and healthy individuals. Specifically, methylation levels were elevated in BrCa patients across all studied gene promoters, consistent with our in silico results as well as the established role of aberrant methylation in cancer [59]. Specifically, adjuvant and metastatic groups of patients had higher *MKGPRD* and *ZNF430* promoter methylation as compared to the healthy group. Also, the metastatic and neoadjuvant groups presented higher *CLDN15* methylation in relation to the healthy group. These results corroborate previous studies showing that promoter methylation of tumor suppressor and cancer-related genes is elevated in cancer in relation to health [5, 60–62] and could lead to gene silencing and cancer

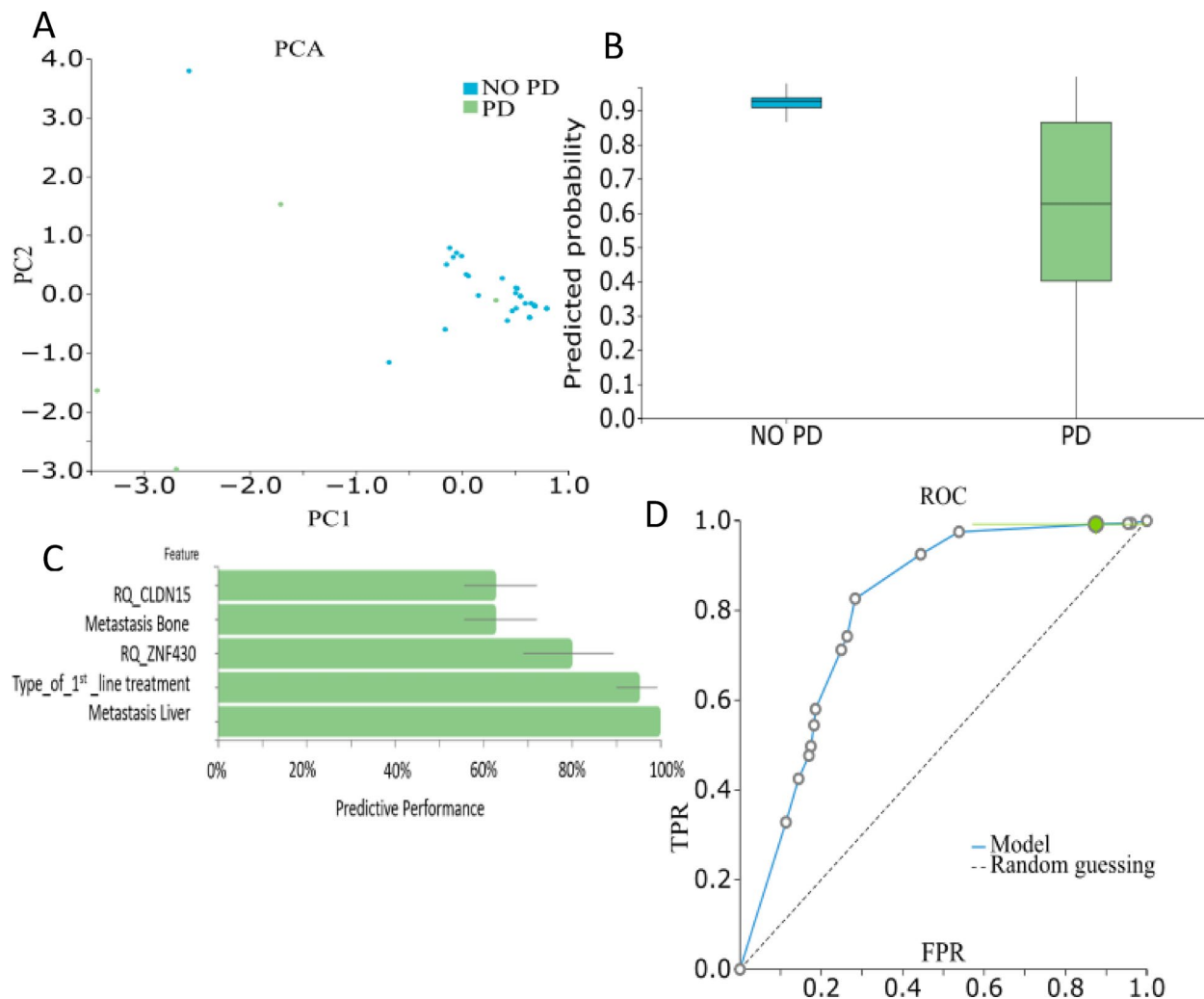


Fig. 10 BrCa prognostic biosignature of progressive disease in metastatic patients. **A** PCA plot, **B** Box-plot predicted probability between those patients who had PD and those who had not **C** Feature Importance plot of the model. **D** ROC curve. Abbreviations: ROC: Receiver Operating Characteristic; PCA: Principal Component Analysis; TPR: True Positive Rate; FPR: False Positive Rate; PD: Progression Disease

progression [63]. A quick search in GEO and Human Protein Atlas revealed expression of mRNA of *CLDN15* and *ZNF430* in both benign and cancer breast tissue, less so that of *MGRPRD*, and it would be interesting to study differences in mRNA and protein levels between groups and their correlation to gene promoter methylation. Of great clinical implication is that the gene methylation profile of tumor tissue can be traced in liquid biopsy and can be a valuable tool for cancer diagnosis [64, 65] and monitoring [66].

To further support this, our experimental analysis revealed significant correlations between promoter methylation levels and clinical parameters of BrCa. Specifically, elevated methylation of *CLDN15* and *MGRPRD* was observed in stage III samples as compared to earlier stages, showing a possible connection of methylation with cancer progression. Similarly, stratified analysis within each group showed that increased *CLDN15*

methylation levels were associated with frequent recurrence both in the adjuvant and the metastatic group. We also found that all genes were more often methylated in the metastatic group as compared to the healthy group. Our findings align with previous research indicating that higher DNA methylation of cancer-related genes is associated with cancer progression and aggressiveness [67, 68]. The observed increase in *CLDN15* methylation in advanced BrCa stage agrees with previous findings for other claudin members, showing that loss of mRNA and protein expression is associated with advanced-stage colorectal cancer and other adverse clinical characteristics [69]. Similarly, the frequent methylation of the *MGRPRD* promoter in older patients highlights the potential role of age-related epigenetic changes, as previously reported [70]. In the metastatic group, frequent methylation of *MGRPRD* was associated with shorter PFS, and higher levels of *ZNF430* were associated with an absence

of response in the 1st-line treatment at the first check, showing their role in aggressive cancer phenotype and highlighting the potential of using methylation markers to predict treatment response. Functional studies investigating the downstream effects of promoter methylation of these genes could provide further insights into the pathogenic role in BrCa.

As a final step, we performed multi-parametric analyses by machine learning to maximize the exploitation of our data. A total of four highly-performing biosignatures were produced. The first biosignature classified between BrCa and health, achieving a high AUC of 0.789. The second discriminated subgroups of the study, achieving a lower AUC of 0.678, possibly due to the low number of patients in the metastatic and adjuvant groups and the nature of the analysis (discrimination of cancer patients' groups that share common methylation patterns). Both signatures contained the methylation of studied genes and the quantity of cfDNA as features, and no other parameters like demographic characteristics, showing that methylation and cfDNA bear high diagnostic power as previously reported [71, 72]. Even though the used autoML platform, JADBio has been shown to shield against typical methodological pitfalls in data analysis that lead to overfitting and overestimating performance, our original datasets were also analyzed after splitting the dataset into training and validation datasets. In our first two classification analyses between BrCa and health and between subgroups, training, and validation analyses achieved similar AUCs with those of the original datasets, showing stable performance of the model and proving no overfitting.

A prognostic biosignature was also built to predict cancer recurrence and included the methylation of all studied genes, stage, vessel infiltration, HER2 status, and hormonal therapy as maintenance treatment. The biosignature showed high performance, which was also confirmed by validation. Cancer stage, HER2, Vascular infiltration, and hormonal therapy are all known and well-established major prognostic factors [73–78]. Thus, it was expected to be included in the biosignature, the performance of which, however, is strengthened by the cfDNA methylation features. Our analysis was made in all BrCa patients without considering the molecular subtype, like Luminal A or HER2+ enriched, due to the restricted number of samples, presenting a study limitation.

The last biosignature predicts PD after 1st line treatment and includes as features the methylation of *CLDN15* and *ZNF430*, along with bone and liver metastasis at diagnosis. Although performance was high at training, it dropped at validation, possibly due to the low number of metastatic patients and the very few cases having PD. Previous studies have shown that the first

metastatic sites strongly influence future prognosis [79, 80]. Furthermore, it has been shown that gene-specific methylation biomarkers are associated with treatment response and are reliable for treatment monitoring [81]. In a relevant study, researchers found 9 significantly differentially methylation regions that are associated with neoadjuvant treatment response in the triple-negative BrCa [82]. Upon validation in a larger cohort of patients in sequential time points during treatment, this approach could enhance decision-making for treatment options in metastatic BrCa.

Other researchers have developed in silico machine-learning models in BrCa employing data-driven approaches. Wang et al. [83] used TCGA and GEO BrCa methylation data to build models with 99% accuracy in predicting BrCa invasiveness. Similarly, Zhu et al. [84] used a data-driven approach to build methylation-based models for the prognosis of BrCa with an AUC of 0.740. Gomes et al. [85] used TCGA data and deep learning to construct predictive models of BrCa with an accuracy of almost 99%. However, none of these studies validated their results in tissues or in the liquid biopsy of cancer patients. We have previously analyzed with autoML our experimental BrCa methylation data to build diagnostic and prognostic biosignatures of high performance [8], however, the genes studied were selected based on existing biological evidence. To the best of our knowledge, the present study is the first to employ machine learning models to identify biomarkers through a data-driven pipeline, as well as validate them in clinical liquid biopsy samples and develop optimized models with clinical significance.

Our study has certain limitations. The small sample size of the neoadjuvant and metastatic groups restricted the robustness of model validation in these subgroups. Additionally, some of our analyses were performed on the combined set of BrCa samples across stages I to IV. While this approach may introduce heterogeneity, it was chosen to enhance the diagnostic power of our analysis. Also, some of the developed biosignatures, like the BrCa-specific diagnostic biosignature, achieved sufficient AUC, precision, and CIs, but exhibited a performance imbalance in metrics such as True Positives and True Negatives. These findings highlight the need for validation in larger BrCa cohorts to refine these metrics and confirm the biosignatures' performance before translation into clinical practice.

The choice to collect samples a month after surgery and before initiation of adjuvant therapy in the 'adjuvant' group was based on previous findings showing that after surgery and during treatment, clearance of cfDNA can create a bias [86]. To date, it is still not clear which is the best time point to collect cfDNA for these studies and the ideal study design should include multiple

timepoints [12]. Also, clinical samples were collected from two different oncology clinics and were processed with slightly different centrifugation protocols. While such variations can influence biomarker sensitivity detection, cfDNA yield, and integrity, we did not observe systematic differences [87, 88]. Therefore, we are not capable of estimating the exact influence on biosignature performance. Nevertheless, we acknowledge centrifugation parameters as a potential source of pre-analytical variability, which could have introduced some heterogeneity between samples and affected biosignature performance. On the other hand, it strengthens the validity of the results as validation included two different medical settings and pre-analytical protocols. Future studies will evaluate the developed biosignatures in a larger patient cohort to confirm their clinical utility in independent settings. Moreover, advanced machine learning will allow the integration of multiple biological entity datasets like genomics, epigenomics, transcriptomics, proteomics, and others to provide a more comprehensive and holistic understanding of BrCa pathogenesis, accurately identify distinct molecular fingerprints, with applications in early detection, personalized pharmacotherapy, and novel treatment targets.

Conclusion

Collectively, our data further support the value of cfDNA as a minimally invasive tool for BrCa management. For the first time, the methylation of *CLDN15*, *MRGPRD*, and *ZNF430* emerged as potential biomarkers in BrCa. Our data-driven, machine-learning pipeline produced five biosignatures: three diagnostic, one prognostic, and one predictive, validated in the laboratory, and combined with clinical data, demonstrated great performance. Additionally, the models' low number of features, the minimally invasive approach of liquid biopsy, and the use of relatively simple qPCR technology present significant advantages for potential clinical implementation. We believe that the adopted AutoML-based pipeline can be applied to the majority of cancer types and provide mature, clinically relevant solutions in the era of personalized medicine.

Abbreviations

AutoML	Automated machine learning
ML	Machine learning
DMPs	Differentially methylated promoters
BrCa	Breast cancer
cfDNA	Cell-free DNA
HER2	Human Epidermal growth factor Receptor 2
GO	Gene ontology
PPI	Protein–protein interaction
SVM	Support vector machine
AUC	Area under the curve
RESIST	The response evaluation criteria in solid tumors
CR	Complete response
PR	Partial response
SD	Stable disease

PD	Progression disease
GAPDH	Glyceraldehyd-3-phosphat-dehydrogenase
ACTB	β-Actin
COL2A1	Collagen type II Alpha 1 chain
RQ	Relative quantification
qMSP	Methylation-specific PCR
OS	Overall survival
PFS	Progression free survival
DFI	Disease free survival
CI	Confidence interval
ROC	Receiver operating characteristic
PCA	Principal component analysis
TPR	True positive rate
FPR	False positive rate
CLDN15	Claudin 15
MRGPRD	MAS related GPR family member D
ZNF430	Zinc finger protein 430
UMAP	Uniform manifold approximation and projection
EMT	Epithelial-to-mesenchymal transition

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13058-025-02170-y>.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

Not applicable

Author contributions

Conceptualization: M.P. and E.C.; methodology, M.P., M.P.D., M.K., software, M.P., I.T., T.T., results analysis, M.P., T.T., E.C., K.M., S.K., writing—original draft preparation, M.P.; writing—review and editing, M.P., I.T., E.C., S.K. and S.A.; visualization, M.P., M.K.; supervision, E.C.; project administration, M.P., E.C.; funding acquisition, E.C. All authors have read and agreed to the published version of the manuscript.

Funding

This publication is financed by the Project 'Strengthening and optimizing the operation of MODY services and academic and research units of the Hellenic Mediterranean University; funded by the Public Investment Program of the Greek Ministry of Education and Religious Affairs.

Availability of data and materials

The Github link ([<https://github.com/MariaPanPan/BrCa-study-data.git>] (<https://github.com/MariaPanPan/BrCa-study-data.git>)) contains the following: 1. Datasets that were retrieved from public resources (TCGA-BRCA project and GEO studies) 2. Differentially methylated Promoters as they emerged from the RnBeads analysis 3. BrCa working sheets 4. SPSS Results GEO studies analyzed: GSE74214: [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74214>] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74214>) GSE72245: [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72245>] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72245>) GSE108576: [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108576>] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108576>) GSE72251: [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72251>] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72251>) GSE88883: [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE88883>] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE88883>).

Declarations

Ethics approval and consent to participate

For the clinical sample analysis, approvals were obtained from the Scientific Board of UGHA and UHH, following an assessment by the Ethics Committee (decisions 14/895/28.11.11 and 9286/15-01-2013). The study was conducted in accordance with the ethical principles outlined in the 1964 Declaration of Helsinki and its subsequent amendments. All participants signed a voluntary informed consent.

Consent for publication

Not applicable.

Competing interests

T. Theodosiou and E. Chatzaki are co-founders of ABCureD PC, while M. Panagopoulou and M. Karagiani were briefly employed under short-term contracts.

Author details

¹Laboratory of Pharmacology, Medical School, Democritus University of Thrace, 68100 Alexandroupolis, Greece

²Institute of Agri-Food and Life Sciences, Hellenic Mediterranean University Research Centre, 71410 Heraklion, Crete, Greece

³Laboratory of Translational Oncology, Medical School, University of Crete, 70013 Heraklion, Crete, Greece

⁴Laboratory of Experimental Oncology, Division of Surgery, Medical School, University of Crete, 71003 Heraklion, Greece

⁵JADBio Gnosis DA S.A., Science and Technology Park of Crete, 70013 Heraklion, Greece

⁶Department of Computer Science, University of Crete, 70013 Heraklion, Greece

⁷Institute of Applied and Computational Mathematics, Foundation for Research and Technology–Hellas, 70013 Heraklion, Greece

⁸Department of Medical Oncology, University General Hospital of Alexandroupoli, 68100 Alexandroupoli, Greece

⁹Department of Medical Oncology, University General Hospital of Heraklion, 71500 Heraklion, Greece

¹⁰ABCureD PC, 68100 Alexandroupoli, Greece

Received: 19 May 2025 / Accepted: 3 November 2025

Published online: 05 January 2026

References

1. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: an overview. *Int J Cancer*. 2021. <https://doi.org/10.1002/ijc.33588>.
2. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk factors and prevention of breast cancer. *Int J Biol Sci*. 2017;13(11):1387–97.
3. Luque-Bolivar A, Pérez-Mora E, Villegas VE, Rondón-Lagos M. Resistance and overcoming resistance in breast cancer. *Breast cancer* (Dove Medical Press). 2020;12:211–29.
4. Kulis M, Esteller M. DNA methylation and cancer. *Adv Genet*. 2010;70:27–56.
5. Panagopoulou M, Drosouni A, Fanidis D, Karagiani M, Balgkouranidou I, Xenidis N, et al. ENPP2 promoter methylation correlates with decreased gene expression in breast cancer: implementation as a liquid biopsy biomarker. *Int J Mol Sci*. 2022. <https://doi.org/10.3390/ijms23073717>.
6. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. *Trends Genet*. 2021;37(11):1012–27.
7. Panagopoulou M, Karagiani M, Balgkouranidou I, Bizioti E, Koukaki T, Karamitrousis E, et al. Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. *Oncogene*. 2019;38(18):3387–401.
8. Panagopoulou M, Karagiani M, Manolopoulos VG, Iliopoulos I, Tsamardinos I, Chatzaki E. Deciphering the methylation landscape in breast cancer: diagnostic and prognostic biosignatures through automated machine learning. *Cancers (Basel)*. 2021. <https://doi.org/10.3390/cancers13071677>.
9. Zhang W, Wang H, Qi Y, Li S, Geng C. Epigenetic study of early breast cancer (EBC) based on DNA methylation and gene integration analysis. *Sci Rep*. 2022;12(1):1989.
10. Panagopoulou M, Lambropoulou M, Balgkouranidou I, Nena E, Karagiani M, Nicolaidou C, et al. Gene promoter methylation and protein expression of BRMS1 in uterine cervix in relation to high-risk human papilloma virus infection and cancer. *Tumor Biol*. 2017;39(4):1010428317697557.
11. Karagiani M, Panagopoulou M, Baltasava I, Apalaki P, Theodosiou T, Iliopoulos I, et al. Tissue-specific methylation biosignatures for monitoring diseases: an in silico approach. *Int J Mol Sci*. 2022;23(6):2959.
12. Panagopoulou M, Esteller M, Chatzaki E. Circulating cell-free DNA in breast cancer: searching for hidden information towards precision medicine. *Cancers*. 2021. <https://doi.org/10.3390/cancers13040728>.
13. Panagopoulou M, Karagiani M, Balgkouranidou I, Pantazi C, Kolios G, Kakolyris S, et al. Circulating cell-free DNA release in vitro: kinetics, size profiling, and cancer-related gene methylation. *J Cell Physiol*. 2019;234(8):14079–89.
14. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*. 2022;23(1):40–55.
15. Salehin I, Islam MS, Saha P, Noman SM, Tunj A, Hasan MM, et al. AutoML: a systematic review on automated machine learning with neural architecture search. *J Inf Intell*. 2024;2(1):52–81.
16. Panagopoulou M, Karagiani M, Tzitzikou K, Kessari N, Arvanitidis K, Amarantidis K, et al. Mitochondrial fraction of circulating cell-free DNA as an indicator of human pathology. *Int J Mol Sci*. 2024;25(8):4199.
17. Karagiani M, Agorastos A, Panagopoulou M, Parlapani E, Athanasios P, Bitsios P, et al. A novel blood-based epigenetic biosignature in first-episode schizophrenia patients through automated machine learning. *Transl Psychiatry*. 2024;14(1):257.
18. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207–10.
19. Jeschke J, Bizet M, Desmedt C, Calonne E, Dedeurwaerder S, Garaud S, et al. DNA methylation-based immune response signature improves patient diagnosis in multiple cancers. *J Clin Invest*. 2017;127(8):3090–102.
20. Johnson KC, Houseman EA, King JE, Christensen BC. Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. *Breast Cancer Res*. 2017;19(1):81.
21. Orozco JIJ, Knijnenburg TA, Manughian-Peter AO, Salomon MP, Barkhoudarian G, Jallas JR, et al. Epigenetic profiling for the molecular classification of metastatic brain tumors. *Nat Commun*. 2018;9(1):4627.
22. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol*. 2019;20(1):55.
23. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11:587.
24. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216–21.
25. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605–12.
26. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res*. 2023;52(D1):D672–8.
27. Tsamardinos I, Charonyktakis P, Papoutsoglou G, Borboudakis G, Lakiotaki K, Zenklusen JC, et al. Just add data: automated predictive modeling for knowledge discovery and feature selection. *NPJ Precis Oncol*. 2022;6(1):38.
28. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach Learn*. 2018;107(12):1895–922.
29. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics*. 2016;54:1.30.1–1.3.
30. Baltasava I, Theodosiou T, Papanikolaou N, Pavlopoulos GA, Amoutzias GD, Panagopoulou M, et al. Prediction and ranking of biomarkers using multiple UniReD. *Int J Mol Sci*. 2022. <https://doi.org/10.3390/ijms231911112>.
31. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–47.
32. Rychlik W. OLIGO 7 primer analysis software. *Methods Mol Biol* (Clifton, NJ). 2007;402:35–60.
33. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2⁻(Delta Delta C(T)) method. *Methods* (San Diego, Calif). 2001;25(4):402–8.
34. Li L-C, Dahiya R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics*. 2002;18(11):1427–31.
35. Riethdorf S, O'Flaherty L, Hille C, Pantel K. Clinical applications of the cell search platform in cancer patients. *Adv Drug Deliv Rev*. 2018;125:102–21.
36. Abbasi HQ, Maryum A, Khan AM, Shahnoor S, Oduoye MO, Wechuli PN. Advancing precision oncology in breast cancer: the FDA approval of elacestrant and Guardant360 CDx: a correspondence. *Int J Surg*. 2023;109(7):2157–8.
37. Esteller M. Epigenetics in cancer. *N Engl J Med*. 2008;358(11):1148–59.
38. Baylin SB, Jones PA. A decade of exploring the cancer epigenome — biological and translational implications. *Nat Rev Cancer*. 2011;11(10):726–34.
39. Shiah JV, Johnson DE, Grandis JR. Transcription factors and cancer: approaches to targeting. *Cancer J*. 2023;29(1):38–46.

40. Fu D, Hu Z, Xu X, Dai X, Liu Z. Key signal transduction pathways and crosstalk in cancer: biological and therapeutic opportunities. *Transl Oncol*. 2022;26:101510.
41. Kyuno D, Takasawa A, Kikuchi S, Takemasa I, Osanai M, Kojima T. Role of tight junctions in the epithelial-to-mesenchymal transition of cancer cells. *Biochimica et Biophysica Acta (BBA)*. 2021;1863(3):183503.
42. Watanabe M, Higashi T, Ozeki K, Higashi AY, Sugimoto K, Mine H, et al. CLDN15 is a novel diagnostic marker for malignant pleural mesothelioma. *Sci Rep*. 2021;11(1):12554.
43. Chaouche-Mazouni S, Scherpereel A, Zaamoun R, Mihalache A, Amir Z-C, Lebailli N, et al. Claudin 3, 4, and 15 expression in solid tumors of lung adenocarcinoma versus malignant pleural mesothelioma. *Ann Diagn Pathol*. 2015;19(4):193–7.
44. Yang G, Jian L, Chen Q. Comprehensive analysis of expression and prognostic value of the claudin family in human breast cancer. *Aging*. 2021;13(6):8777–96.
45. Alghamdi RA, Al-Zahrani MH. Identification of key claudin genes associated with survival prognosis and diagnosis in colon cancer through integrated bioinformatic analysis. *Front Genet*. 2023. <https://doi.org/10.3389/fgene.2023.1221815>.
46. Gao P, Peng T, Cao C, Lin S, Wu P, Huang X, et al. Association of CLDN6 and CLDN10 with immune microenvironment in ovarian cancer: a study of the claudin family. *Front Genet*. 2021;12:595436.
47. Wu Z, Jia Q, Liu B, Fang L, Leung PCK, Cheng J-C. NPFF stimulates human ovarian cancer cell invasion by upregulating MMP-9 via ERK1/2 signaling. *Exp Cell Res*. 2023;430(1):113693.
48. Nishimura S, Uno M, Kaneta Y, Fukuchi K, Nishigohri H, Hasegawa J, et al. MRGD, a MAS-related G-protein coupled receptor, promotes tumorigenesis and is highly expressed in lung cancer. *PLoS ONE*. 2012;7(6):e38618.
49. Mamoor S. Differential expression of zinc finger protein 430 in human endometrial cancer. 2021.
50. Hana C, Thaw Dar NN, Galo Venegas M, Vulfovich M. Claudins in cancer: a current and future therapeutic target. *Int J Mol Sci*. 2024;25(9):4634.
51. Yan J, Wang C, Wei Y, Yang H. Exercise intervention during pregnancy induces DNA methylation alterations in maternal blood and cord blood. *Chin Med J*. 2023;136(13):1624–6.
52. Khurram I, Khan MU, Ibrahim S, Saleem A, Khan Z, Mubeen M, et al. Efficacy of cell-free DNA as a diagnostic biomarker in breast cancer patients. *Sci Rep*. 2023;13(1):15347.
53. Su Y, Wang L, Jiang C, Yue Z, Fan H, Hong H, et al. Increased plasma concentration of cell-free DNA precedes disease recurrence in children with high-risk neuroblastoma. *BMC Cancer*. 2020;20(1):102.
54. Sozzi G, Conte D, Leon M, Ciricione R, Roz L, Ratcliffe C, et al. Quantification of free circulating DNA as a diagnostic marker in lung cancer. *J Clin Oncol Off J Am Soc Clin Oncol*. 2003;21(21):3902–8.
55. Frattini M, Gallino G, Signoroni S, Balestra D, Battaglia L, Sozzi G, et al. Quantitative analysis of plasma DNA in colorectal cancer patients: a novel prognostic tool. *Ann N Y Acad Sci*. 2006;1075:185–90.
56. Sai S, Ichikawa D, Tomita H, Ikoma D, Tani N, Ikoma H, et al. Quantification of plasma cell-free DNA in patients with gastric cancer. *Anticancer Res*. 2007;27(4c):2747–51.
57. Sozzi G, Roz L, Conte D, Mariani L, Andriani F, Lo Vullo S, et al. Plasma DNA quantification in lung cancer computed tomography screening: five-year results of a prospective study. *Am J Respir Crit Care Med*. 2009;179(1):69–74.
58. Fonseca NM, Maurice-Dror C, Herberts C, Tu W, Fan W, Murtha AJ, et al. Prediction of plasma ctDNA fraction and prognostic implications of liquid biopsy in advanced prostate cancer. *Nat Commun*. 2024;15(1):1828.
59. Esteller M. Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol*. 2005;45:629–56.
60. Panagopoulou M, Panou T, Gkoutakos A, Tarapatzi G, Karaglanis M, Tsamardinos I, et al. BRCA1 & BRCA2 methylation as a prognostic and predictive biomarker in cancer: implementation in liquid biopsy in the era of precision medicine. *Clin Epigenetics*. 2024;16(1):178.
61. Nikolaidis C, Nena E, Panagopoulou M, Balgouranidou I, Karaglanis M, Chatzaki E, et al. PAX1 methylation as an auxiliary biomarker for cervical cancer screening: a meta-analysis. *Cancer Epidemiol*. 2015;39(5):682–6.
62. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*. 2002;21(35):5427–40.
63. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer*. 2011;11(10):726–34.
64. Hulbert A, Jusue-Torres I, Stark A, Chen C, Rodgers K, Lee B, et al. Early detection of lung cancer using DNA Promoter hypermethylation in plasma and sputum. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2017;23(8):1998–2005.
65. Ibrahim J, Peeters M, Van Camp G, de Op Beeck K. Methylation biomarkers for early cancer detection and diagnosis: Current and future perspectives. *Eur J Cancer*. 2023;178:91–113.
66. Kristiansen S, Nielsen D, Sölétormos G. Detection and monitoring of hypermethylated RASSF1A in serum from patients with metastatic breast cancer. *Clin Epigenetics*. 2016;8(1):35.
67. de Unamuno BB, Murria Estal R, Pérez Simó G, Simarro Farinos J, Pujol Marco C, Navarro Mira M, et al. Aberrant DNA methylation is associated with aggressive clinicopathological features and poor survival in cutaneous melanoma. *Br J Dermatol*. 2018;179(2):394–404.
68. Okamoto Y, Sawaki A, Ito S, Nishida T, Takahashi T, Toyota M, et al. Aberrant DNA methylation associated with aggressiveness of gastrointestinal stromal tumour. *Gut*. 2012;61(3):392–401.
69. Cox KE, Liu S, Hoffman RM, Batra SK, Dhawan P, Bouvet M. The expression of the claudin family of proteins in colorectal cancer. *Biomolecules*. 2024;14(3):272.
70. Johnson AA, Akman K, Calimport SR, Wuttke D, Stolzing A, de Magalhães JP. The role of DNA methylation in aging, rejuvenation, and age-related disease. *Rejuven Res*. 2012;15(5):483–94.
71. Zhang X, Zhao D, Yin Y, Yang T, You Z, Li D, et al. Circulating cell-free DNA-based methylation patterns for breast cancer diagnosis. *NPJ Breast Cancer*. 2021;7(1):106.
72. Panagopoulou M, Cheretaki A, Karaglanis M, Balgouranidou I, Bizioti E, Amarantidis K, et al. Methylation status of corticotropin-releasing factor (CRF) receptor genes in colorectal cancer. *J Clin Med*. 2021. <https://doi.org/10.3390/jcm10122680>.
73. Nors I, Iversen LH, Erichsen R, Gotschalck KA, Andersen CL. Incidence of recurrence and time to recurrence in stage I to III colorectal cancer: a nationwide Danish cohort study. *JAMA Oncol*. 2024;10(1):54–62.
74. Pedersen RN, Esen BÖ, Mellemkjær L, Christiansen P, Ejlersen B, Lash TL, et al. The incidence of breast cancer recurrence 10–32 years after primary diagnosis. *JNCI J Natl Cancer Inst*. 2021;114(3):391–9.
75. Bradley R, Braybrooke J, Gray R, Hills R, Liu Z, Peto R, et al. Taylor C (EBCTCG) EBCTCG. Trastuzumab for early-stage, HER2-positive breast cancer: a meta-analysis of 13 864 women in seven randomised trials. *Lancet Oncol*. 2021;22(8):1139–50.
76. Macchiarini P, Fontanini G, Hardin MJ, Chuanchieh H, Bigini D, Vignati S, et al. Blood vessel invasion by tumor cells predicts recurrence in completely resected T1 N0 M0 non-small-cell lung cancer. *J Thorac Cardiovasc Surg*. 1993;106(1):80–9.
77. Shimada Y, Ishii G, Hishida T, Yoshida J, Nishimura M, Nagai K. Extratumoral vascular invasion is a significant prognostic indicator and a predicting factor of distant metastasis in non-small cell lung cancer. *J Thorac Oncol*. 2010;5(7):970–5.
78. Dufresne A, Pivot X, Tournigand C, Facchini T, Alweeg T, Chaigneau L, et al. Maintenance hormonal treatment improves progression free survival after a first line chemotherapy in patients with metastatic breast cancer. *Int J Med Sci*. 2008;5(2):100–5.
79. Yamamura J, Kamigaki S, Fujita J, Osato H, Manabe H, Tanaka Y, et al. New insights into patterns of first metastatic sites influencing survival of patients with hormone receptor-positive, HER2-negative breast cancer: a multicenter study of 271 patients. *BMC Cancer*. 2021;21(1):476.
80. Newton PK, Mason J, Venkatappa N, Jochelson MS, Hurt B, Nieva J, et al. Spatiotemporal progression of metastatic breast cancer: a Markov chain model highlighting the role of early metastatic sites. *NPJ Breast Cancer*. 2015;1(1):15018.
81. Lu C-Y, Hsiao C-Y, Peng P-J, Huang S-C, Chuang M-R, Su H-J, et al. DNA methylation biomarkers as prediction tools for therapeutic response and prognosis in intermediate-stage hepatocellular carcinoma. *Cancers (Basel)*. 2023;15(18):4465.
82. Meyer B, Clifton S, Locke W, Luu P-L, Du Q, Lam D, et al. Identification of DNA methylation biomarkers with potential to predict response to neoadjuvant chemotherapy in triple-negative breast cancer. *Clin Epigenetics*. 2021;13(1):226.
83. Wang C, Zhao N, Yuan L, Liu X. Computational detection of breast cancer invasiveness with DNA methylation biomarkers. *Cells*. 2020. <https://doi.org/10.3390/cells9020326>.

84. Zhu C, Zhang S, Liu D, Wang Q, Yang N, Zheng Z, et al. A novel gene prognostic signature based on differential DNA methylation in breast cancer. *Front Genet.* 2021;12:742578.
85. Gomes R, Paul N, He N, Huber AF, Jansen RJ. Application of feature selection and deep learning for cancer prediction using DNA methylation markers. *Genes.* 2022;13(9):1557.
86. Garcia-Murillas I, Cutts RJ, Walsh-Crestani G, Phillips E, Hrebien S, Dunne K, et al. Longitudinal monitoring of circulating tumor DNA to detect relapse early and predict outcome in early breast cancer. *Breast Cancer Res Treat.* 2025;209(3):493–502.
87. Sorber L, Zwaenepoel K, Jacobs J, De Winne K, Goethals S, Reclusa P, et al. Circulating cell-free DNA and RNA analysis as liquid biopsy: optimal centrifugation protocol. *Cancers (Basel).* 2019. <https://doi.org/10.3390/cancer11040458>.
88. Shin K-H, Lee SM, Park K, Choi H, Kim I-s, Yoon SH, et al. Effects of different centrifugation protocols on the detection of EGFR mutations in plasma cell-free DNA. *Am J Clin Pathol.* 2022;158(2):206–11.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.