NAR Genomics and Bioinformatics, 2024, **6**, lqae022
https://doi.org/10.1093/nargab/lqae022
Advance access publication date: 24 February 2024
Standard Article

An evolutionary learning-based method for identifying a circulating miRNA signature for breast cancer diagnosis prediction

Srinivasulu Yerukala Sathipati[1,*,†], Ming-Ju Tsai[2,3,†], Nikhila Aimalla[4], Luke Moat[1], Sanjay K. Shukla[1], Patrick Allaire[1], Scott Hebbring[1], Afshin Beheshti[5,6], Rohit Sharma[7] and Shinn-Ying Ho[8,9,10]

[1]Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA
[2]Hinda and Arthur Marcus Institute for Aging Research at Hebrew Senior Life, Boston, MA 02131, USA
[3]Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02131, USA
[4]Department of Internal Medicine-Pediatrics, Marshfield Clinic Health System, Marshfield, WI 54449, USA
[5]Blue Marble Space Institute of Science, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA94035, USA
[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[7]Department of Surgical Oncology, Marshfield Clinic Health System, Marshfield, WI 54449, USA
[8]Institute of Bioinformatics and Systems biology, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[9]College of Health Sciences, Kaohsiung Medical University, Kaohsiung 807378, Taiwan
[10]Biomedical Engineering, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan

[*]To whom correspondence should be addressed. Tel: +1 715 2216470; Fax: 715 389 4950; Email: sathipathi.srinivasulu@marshfieldclinic.org
[†]The first and second authors should be regarded as Joint First Authors.

Abstract

Breast cancer (BC) is one of the most commonly diagnosed cancers worldwide. As key regulatory molecules in several biological processes, microRNAs (miRNAs) are potential biomarkers for cancer. Understanding the miRNA markers that can detect BC may improve survival rates and develop new targeted therapeutic strategies. To identify a circulating miRNA signature for diagnostic prediction in patients with BC, we developed an evolutionary learning-based method called BSig. BSig established a compact set of miRNAs as potential markers from 1280 patients with BC and 2686 healthy controls retrieved from the serum miRNA expression profiles for the diagnostic prediction. BSig demonstrated outstanding prediction performance, with an independent test accuracy and area under the receiver operating characteristic curve were 99.90% and 0.99, respectively. We identified 12 miRNAs, including hsa-miR-3185, hsa-miR-3648, hsa-miR-4530, hsa-miR-4763-5p, hsa-miR-5100, hsa-miR-5698, hsa-miR-6124, hsa-miR-6768-5p, hsa-miR-6800-5p, hsa-miR-6807-5p, hsa-miR-642a-3p, and hsa-miR-6836-3p, which significantly contributed towards diagnostic prediction in BC. Moreover, through bioinformatics analysis, this study identified 65 miRNA-target genes specific to BC cell lines. A comprehensive gene-set enrichment analysis was also performed to understand the underlying mechanisms of these target genes. BSig, a tool capable of BC detection and facilitating therapeutic selection, is publicly available at https://github.com/mingjutsai/BSig.

Introduction

Breast cancer (BC) is one of the most commonly diagnosed cancer worldwide with 2.26 million cases according to World Health Organization (1). The Surveillance, Epidemiology and End Result program (SEER 2020) estimates ∼281, 550 new cases and 43, 600 cancer deaths for US women in 2021 (2). Amongst females with BC, the highest death rate occurs between ages 65–74 (1). BC is a multifactorial disease with risk factors that include age; sex; family history; reproductive factors such as early menarche, late menopause and low parity; exposure to sex hormones; and genetic, lifestyle and environmental factors (3–7). The stage of BC is strongly predictive of a patient's survival. For Stages I-III the survival rates are between 99% and 72% but this drops to 22% for Stage IV (2). The current standard treatment modalities of BC include a combination of surgery, drug therapy and radiation therapy. In recent years, significant improvements have been made in the diagnosis (8,9), treatment and prognosis of BC (10,11). However, understanding of recurrence and metastasis remains elusive for curing the tumors. Most tumors are curable if detected early before it has advanced and metastasized (12). Therefore, early diagnosis improves chance of survival by providing appropriate treatment options at the earlier stages to effective management of BC.

The ideal biomarker for any disease should be reliably indicative of the disease prior to symptomatic presentation, minimally- or non-invasive in its testing, and inexpensive to deploy (13,14). Researchers have traditionally relied on protein-based biomarkers, but developing new biomarkers based on proteins faces several challenges. These include the complexity of protein compositions in biological samples, low abundance of proteins of interest in blood samples, and the disease's heterogeneity (15). In addition, the development of

Received: November 21, 2023. **Revised:** January 11, 2024. **Editorial Decision:** February 5, 2024. **Accepted:** February 13, 2024

diagnostic specific protein-based biomarkers is also an expensive and a time-consuming task.

Recently, nucleic acid marker-based targeted therapies, particularly microRNA (miRNA) based therapies that are in development, have shown potential for prognosticating and early diagnosis of various diseases (16). MiRNAs are a novel class of endogenous, non-coding, single-stranded RNAs (17) approximately 22 nucleotides in length. They play a crucial role in post-transcriptional regulation of gene expression by blocking translation of messenger RNA (mRNA) targets (18). Several studies have demonstrated that these small regulatory RNA molecules can be released into extracellular fluids, making them suitable biomarkers for a range of diseases, including various human cancers, such as BC (19,20). These circulating miRNAs are increasingly recognized as potential minimally invasive biomarkers associated with relapse-free survival, overall survival, and response to therapy in various cancers (21–25). The stability of circulating miRNAs in the bloodstream, combined with their altered expression reflecting the physiological conditions (26), enhance their reliability as cancer biomarkers. For instance, circulating miRNAs such as miR-195, miR-214, miR-127, miR-15a and miR-18a have been identified as diagnostic and prognostic biomarkers for BC (22). The recent literature indicates that circulating miRNAs, particularly miR-21-5p and miR-155-5p, have been identified as predictive biomarkers in the neoadjuvant treatment of BC patients (27). A combination of serum miRNAs, including hsa-miR-145, hsa-miR-382, and hsa-miR-21, along with glutamic acid levels and circulating HER2 concentrations, shows potential as non-invasive diagnostic tool for the early prediction of BC in Egyptian patients (28). Additionally, increased expression of miR-145 has been associated with improved recurrence-free survival in a study of 124 BC patients (29). Mishra *et al.* (30) identified five aberrantly expressed circulating miRNAs (miR-195-5p, miR-495, miR-34a-5p, miR-106a-5p, and miR-454-3p) in BC patients compared to healthy individuals. Additionally, Matsuzaki *et al.* aimed to non-invasively detect early-stage cancers and predict tumor tissue-of-origin (31). The study utilized serum miRNA profiles and machine learning techniques to develop a classifier, demonstrating optimal diagnostic performance using 100 miRNAs across various cancers. This research underscores the potential of combining serum miRNomics with machine learning to establish a blood-based cancer classification system, with implications for clinical applications of circulating miRNA diagnostics.

Advances in machine learning methods continue to improve in handling larger biomedical datasets and are well-adapted to addressing numerous biological problems (32). However, one of the major challenges of biomedical data analysis includes how higher dimensionality impedes identification of potential biomarkers from many candidate features (33). New methods for handling large datasets will expedite the application of miRNA signatures to BC detection. To cope with the dimensionality issues in miRNA expression data, we previously developed multiple machine learning-based cancer prediction models that contributed to survival predictions in glioblastoma, lung adenocarcinoma, bladder urothelial carcinoma, and stomach and esophageal carcinomas (34–37), as well as early-stage detection in BC and hepatocellular carcinoma (38,39). This study proposes a machine learning method using an optimal feature selection algorithm and support vector machines (SVM) to identify a circulating miRNA signature that could aid in predicting the diagnosis of BC.

Materials and methods

Data availability

A total of 1280 serum samples of BC patients and 2686 serum samples were obtained from healthy controls with non-coding RNA profiling by array. The data are in the GEO under accession number GSE73002 (40). This dataset contains BC serum samples collected prior to therapy. The criteria used to define the healthy control population are consistent with the specifications provided in the study (40). For more comprehensive information regarding both the healthy and BC cohorts, please refer to this study. In our study, we chose to utilize samples from the GSE73002 dataset as part of a strategic effort to improve clarity and precision. This approach was adopted to eliminate any potential sources of ambiguity. The dataset was divided into distinct training and test subsets, maintaining a balanced 70:30 ratio. The training subset included 1793 samples, comprising both BC and healthy cases. In contrast, the test subset comprised 768 samples, encompassing both BC and healthy instances. Additionally, to validate the robustness of our approach, we employed the remaining 1464 healthy samples as an independent cohort.

Detailed steps involved in the study are explained in the following sub sections.

MiRNA signature selection

We developed BSig to identify the miRNA signature associated with BC and to distinguish between BC and healthy groups. During the machine learning process, we used BC individuals as a standard for positive and healthy individuals as a standard for negative. BSig was developed using a SVM classifier and an optimal feature selection algorithm, inheritable bi-objective combinatorial genetic algorithm (IBCGA), which effectively solves bi-objective combinatorial problems. These objectives include selecting a small set of informative features and optimizing the fitness function in terms of accuracy. IBCGA has been successfully applied in several cancer biomarker discoveries (34,36–39,41). SVM has several applications which were successfully implemented and applying in various biomedical fields (42,43). SVM uses non-linear transformation to map data from an input space to a higher-dimensional space to establish an accurate prediction model. The flowchart of BSig method is shown in Supplementary Figure S3.

BSig used an intelligent evolutionary algorithm (IGA) (44) to solve the large-scale parameter optimization problem. IGA has been successfully applied in our previous work in reconstructing gene regulatory networks (45) and also been used to successfully predict the regulatory role of CRP (46). In the optimization process, IBCGA was used to identify a miRNA signature while maximizing the mean accuracy as its fitness function. All the candidate features were encoded into binary variables, including the parameters C, and γ of the SVM. We used 354 miRNA expression profiles ($n = 354$) from patients with BC as candidate features. The detailed steps involved in IBCGA, including initialization, evaluation, selection, crossover, mutation, termination, and inheritance, can be found in the following lines. Additionally, both the miRNA signature and prediction model files have been made available

on the GitHub page. After identifying the miRNA signature, permutation-based analysis was used to prioritize the miRNAs in the signature. The parameter setting of BSig was $I_{start} = 10$ and $I_{end} = 30$, meaning that the search for the n value ranged from 10 to 30. The fitness function was to maximize prediction accuracy of 10-CV. The main steps of feature selection algorithm for identifying a signature of m miRNAs are as follows.

- Step 1: Randomly generate a population of N_{pop} individuals. In this work, $N_{pop} = 50$, $G_{max} = 60$, $r_{start} = 10$, $r_{end} = 30$, $r = r_{start}$.
- Step 2: Evaluate the fitness value of all individuals using the fitness function mean of accuracies.
- Step 3: Use a tournament selection method that selects the winner from two randomly selected individuals to generate a mating pool.
- Step 4: Select two parents from the mating pool to perform an orthogonal array crossover operation.
- Step 5: Apply a conventional mutation operator to the randomly selected individuals in the new population. To prevent the highest fitness value from deteriorating, mutation is not applied to the best individuals.
- Step 6: If the stopping condition of G_{max} generation is satisfied, the best individual is the solution Sr. Otherwise, go to Step 2.
- Step 7: If $r < r_{end}$, randomly change one bit in the binary genes for each individual from 0 to 1; increase the number r by one, and go to Step 2. Otherwise, output the solution Sj with j miRNAs as a signature where Sj is the most accurate solution among the Sr solutions and stop the algorithm.

Feature appearance score

We selected a robust miRNA signature (RMS) from a nondeterministic feature set sections using the feature appearance score (FAS). A feature set with a more significant appearance score suggests that the higher FAS contributed significantly towards BC diagnosis prediction. The robust signature among $S = 30$ solutions was selected using the following procedure.

- Step 1: Perform S independent runs of IBCGA to obtain RMS. There are K_t miRNAs in the t-th signatures, $t = 1, \ldots, K$.
- Step 2: The appearance score of a signature is calculated as follows:

Calculate the feature frequency score $f(k)$ for each miRNA that ever appears in the K signatures.

Calculate the score Mt, $t = 1, \ldots, K$ where R_{ti} is the ith miRNA in the tth signature:

$$M_t = \sum_{i=1}^{K_t} f(R_{ti}) / K_t \tag{1}$$

- Step 3: Output the tth signature with the largest appearance score M_t as the RMS.

Using various machine learning methods to build breast cancer prediction model

This study utilized the stepwise feature forward selection method, a feature selection algorithm, to select informative features for predicting the diagnosis of BC. Subsequently, we used six machine learning methods to build prediction models with the same training dataset, and their performance was evaluated by the same test dataset. We utilized the scikit-learn Python package (47) to employ three methods: Neural Network classification used Multi-layer Perceptron algorithm, Random Forest classification, and Extree Tree classification. Additionally, three boosting algorithms were utilized, including XGBoost (48), LightGBM (49) and CatBoost (50). The Optuna framework (51) with 10-fold cross-validation optimized all parameters for the six methods.

Performance measures

This work used the following equations to measure the performance evaluation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

Where TP is true positive; TN is true negative; FP is false positive and FN is false negative.

Identification of miRNA-targets in breast lineage

To identify the target genes from the selected miRNAs, we use the miRTarBase (9.0 beta) database (52) to extract the experimentally verified microRNA-target interactions by various techniques, including CLASH, CLIP-seq, HITS-CLIP, Immunofluorescence, Luciferase reporter assay, qRT-PCR, western blotting, and PAR-CLIP. In addition, we utilized the Chronos Gene Effect Scores from Cancer Cell Line Encyclopedia (53) to assess the functional significance of the identified genes in BC cell lines. Chronos assigns a gene effect score based on its effect on cell growth and survival in CRISPR-based knockout screens.

miRNA-target enrichment analysis

Gene-set libraries are utilized to organize and categorize accumulated knowledge about the functions of groups of genes. In this study, we employed Enrichr (54), a web-based application that features the latest gene-set libraries, to perform gene-set enrichment analysis. We evaluated the performance of Enrichr in ranking terms from gene-set libraries by combining the P-value obtained through Fisher's exact test with the z-score of deviation from the expected rank, as computed by the following formula:

$$c = \log(p) \cdot z \tag{5}$$

This study used six Gene-set libraries to perform gene-set enrichment analysis, including Wikipathways (55), Elsevier collection (56), Gene Ontology (57), ChEA (58), Achilles (59) and Cancer Cell Line Encyclopedia (60).

Results

Identification of a robust miRNA signature to accurately predict breast cancer

We utilized miRNA expression profile data from 1280 patients with BC and 2686 healthy controls retrieved from the

Gene Expression Omnibus database (GSE73002) (40) to develop a machine learning-based diagnosis prediction method named BSig (**B**reast cancer **Sig**nature). This method identifies a miRNA signature and distinguishes individuals with BC from healthy individuals. The BSig method selects a robust set of miRNAs as a biomarker signature that can accurately predict the BC diagnosis and combines this identification with the prioritization of those signature miRNAs which contribute most to diagnostic prediction. The overview of this study is shown in Figure 1. BSig selected 12 of these miRNAs as a robust signature and achieved 10-fold cross-validation (10-CV) accuracy, sensitivity, specificity, area under the curve of receiver characteristic operator (AUROC), and area under the precision-recall curve (AUPRC) of 100%, 1.0, 1.0, 1.0, and 1.0, respectively. Additionally, BSig achieved a test accuracy, sensitivity, specificity, AUROC, and AUPRC of 99.36%, 0.99, 0.99, 1.0, and 1.0, respectively, in distinguishing BC and healthy individuals.

Next, we validated the BSig on an independent test dataset (GSE211692) (31), which encompasses serum samples from 598 BC patients and 5643 healthy controls. For additional clinical and expression details, please refer to the original study (31). Notably, the BSig demonstrated remarkable performance on this independent test cohort, achieving a sensitivity of 0.89 (537 out of 598), specificity of 0.74 (4224 out of 5643), and an AUC value of 0.94, effectively distinguishing between BC and healthy controls. The evaluation of BSig's predictive performance on this test cohort was assessed using an ROC curve, as illustrated in Supplementary Figure S1.

Performance comparison of breast cancer prediction

We tested the efficacy of the BSig method by comparing with six other machine learning methods, including Random Forest, Extra Trees, Neural Network, XG Boost (48), Light GBM (61), and CatBoost (50). The prediction comparison results are shown in Table 1&2. To make the comparison feasible, we employed feature selection algorithm, stepwise feature forward selection method and selected 16 features to predict the BC diagnosis. The Random Forest method achieved a training accuracy, sensitivity, specificity, AUROC, AUPRC, and test accuracies of 95.02%, 0.90, 0.97, 0.98, 0.97, and 93.31%, respectively. Meanwhile, XGBoost achieved a training accuracy, sensitivity, specificity, AUROC, AUPRC, and test accuracies of 96.24%, 0.92, 0.98, 0.99, 0.98, and 95.96%, respectively. Light GBM achieved a training accuracy, sensitivity, specificity, AUROC, AUPRC, and test accuracies of 96.24%, 0.92, 0.98, 0.96, 0.99, and 95.71%, respectively. Extra tree achieved a training accuracy, sensitivity, specificity, AUROC, AUPRC, and test accuracies of 96.53%, 0.93, 0.98, 0.99, 0.98, and 94.45%, respectively. Catboost achieved a training accuracy, sensitivity, specificity, AUROC, AUPRC, and test accuracies of 96.53%, 0.93, 0.98, 0.99, 0.98, and 95.96%, respectively. Neural network achieved a training accuracy, sensitivity, specificity, AUROC, AUPRC, and test accuracies of 95.96%, 0.92, 0.97, 0.99, 0.97, and 96.21%, respectively. The comparison of prediction results for the training cohort is presented in Table 1, while the corresponding results for the test cohort are displayed in Table 2. The prediction comparison results showed that BSig prediction performance is better than these standard machine learning methods. The performance evaluation of prediction methods using AUROCs and AUPRCs are shown in Figure 2 and Supplementary Figure S2, respectively.

To assess the predictive capabilities of BSig, a comparative analysis was conducted with the previous study (40). In order to mitigate concerns related to overtraining and potential biases, a strategic data partitioning was employed, segregating data sets for distinct scenarios: BC vs. benign BC and BC vs. prostate cancer. We included 54 benign samples and 93 prostate cancer samples in our analysis. Notably, the SVM parameters for these prediction models differ, reflecting separate training for the BC vs. benign and BC vs. prostate cancer classifications. As a result, the signatures derived from these models differed from the established BC signature consisting of 12 miRNAs. The implementation of BSig yielded consistent outcomes. Specifically, after conducting 50 independent runs, BSig achieved a mean 10-CV accuracy of $100 \pm 0\%$ and a test accuracy of $91.23 \pm 5\%$ (Supplementary Table S1).

Expanding the scope, BSig's predictive capabilities extended to differentiating between BC and prostate cancer samples. Across the 50 independent runs, BSig consistently exhibited a mean 10-CV accuracy and test accuracy of $100 \pm 0\%$ and $99.57 \pm 1.17\%$, respectively (Supplementary Table S2). This substantiates the BSig's robustness in accurately distinguishing between these two sample categories.

Robust feature set selection and feature prioritization

We conducted 30 independent runs of BSig to select a robust feature set. The average number of features obtained across the 30 runs was 15 ± 4. We then calculated the frequency appearance score (FAS) for each independent run. The FAS reflects the frequency of each feature in the optimization process, with higher FAS indicating a robust feature set in the optimization modeling process and vice versa. The results of the FAS for the independent runs are presented in Supplementary Table S3. The highest FAS (13.83) was obtained in run 22 with 12 features, and the lowest FAS (5.84) was obtained in run 29 with 25 features. We used the feature set with the highest FAS to build the BSig prediction model, which included hsa-miR-3185, hsa-miR-3648, hsa-miR-4530, hsa-miR-4763-5p, hsa-miR-642a-3p, hsa-miR-5100, hsa-miR-5698, hsa-miR-6124, hsa-miR-6768-5p, hsa-miR-6800-5p, hsa-miR-6807-5, and hsa-miR-6836-3p.

Permutation-based miRNA signature prioritization

The miRNAs in the identified miRNA signature were ranked based on the accuracy difference obtained from the permutation feature importance analysis. The miRNAs with the greatest AUROC difference indicate a higher contribution to the model's predictive power. According to the permutation feature analysis, the top five ranked miRNAs are hsa-miR-5100, hsa-miR-6836-3p, hsa-miR-3185, hsa-miR-4530, and hsa-miR-6800-5p (Figure 3).

We measured the miRNA signature expression difference across BC and healthy using the non-parametric t-test. Among the 12 miRNAs of signature, 11 miRNAs, including hsa-miR-3185, hsa-miR-3648, hsa-miR-4530, hsa-miR-4763-5p, hsa-miR-5100, hsa-miR-5698, hsa-miR-6124, hsa-miR-6768-5p, hsa-miR-6800-5p, hsa-miR-6807-5p, and hsa-miR-6836-3p were significantly ($P \leq 0.005$) expressed between BC and healthy samples. However, the remaining miRNA, hsa-miR-

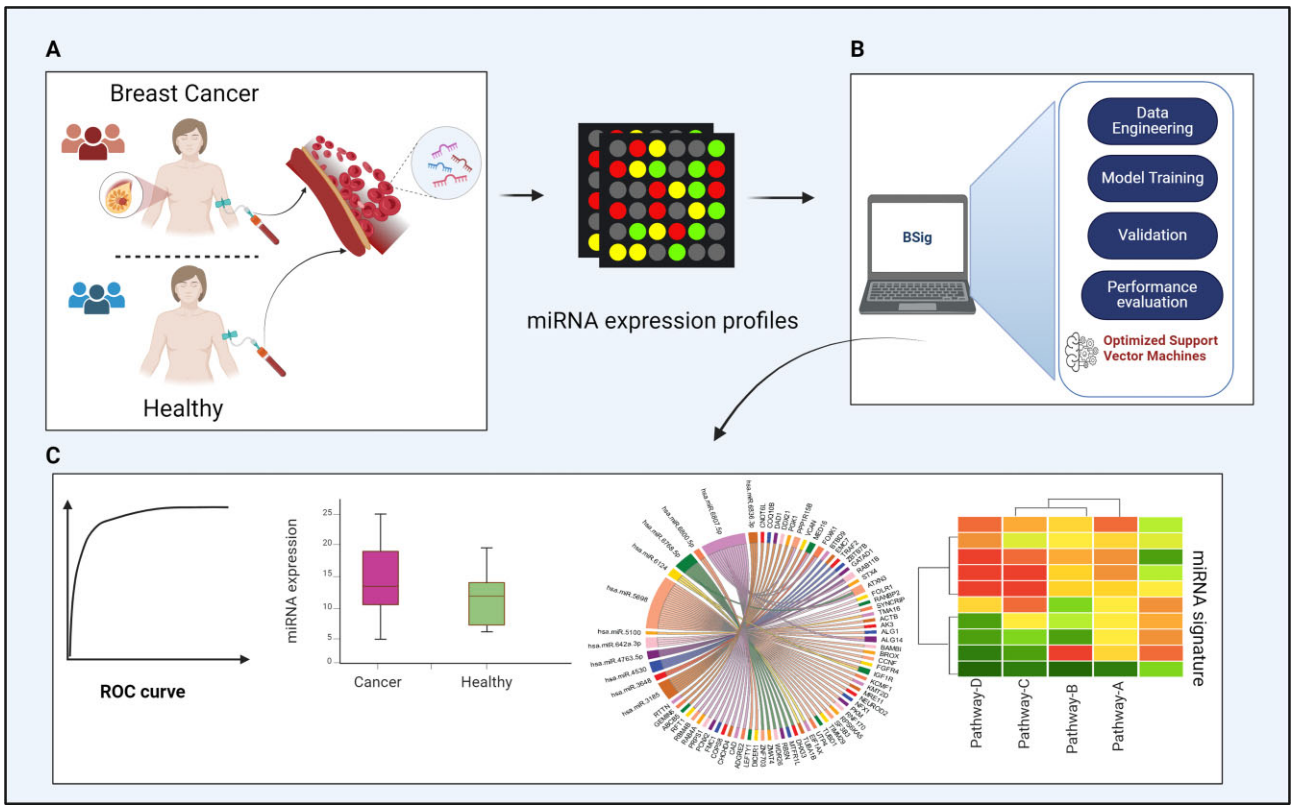

Figure 1. The system overview of the study. (**A**) Extraction of miRNA expression profiles data of patients with breast cancer and healthy individuals from Gene Expression Omnibus database, (**B**) overview of the BSig method for miRNA signature identification, and (**C**) analysis of the identified miRNA signature.

Table 1. The comparison of training prediction performance

Method	Accuracy (%)	AUROC	AUPRC	Sensitivity	Specificity
BSig	100	100	100	1.00	1.00
Neural Network	95.97	99.10	97.20	0.92	0.97
Random Forest	95.02	98.84	97.80	0.90	0.97
XGBoost	96.25	99.36	98.80	0.92	0.98
Light GBM	96.25	99.36	98.80	0.92	0.98
Extra tree	96.53	99.24	98.60	0.93	0.98
Catboost	96.53	99.35	98.80	0.93	0.98

Table 2. The comparison of test prediction performance

Method	Accuracy (%)	AUROC (%)	AUPRC	Sensitivity	Specificity
BSig	99.37	100	100	0.99	0.99
Neural Network	96.21	99.10	98.70	0.93	0.97
Random Forest	93.31	98.70	97.50	0.89	0.95
XGBoost	95.96	99.40	98.80	0.92	0.97
Light GBM	95.71	99.40	98.80	0.92	0.97
Extra tree	94.45	99.00	98.20	0.89	0.96
Catboost	95.96	99.40	98.70	0.92	0.97

642a-3p showed no significance difference between the two groups. Box-plot analysis of the miRNA signature between BC and healthy samples is conducted in Figure 4.

Identification of miRNA-target genes

This study used the miRTarBase database to identify 1968 target genes from 5355 miRNA-target interactions involving 12 selected miRNAs (Supplementary Table S4). To identify BC cell line-specific target genes, the Chronos Gene Effect Scores from Cancer Cell Line Encyclopedia were utilized to assess the functional significance of the identified genes in BC cell lines. Using Chronos Gene Effect Score in BC cell lines, 65 target genes were identified from 189 miRNA-target interactions (see Supplementary Table S5 for details). The

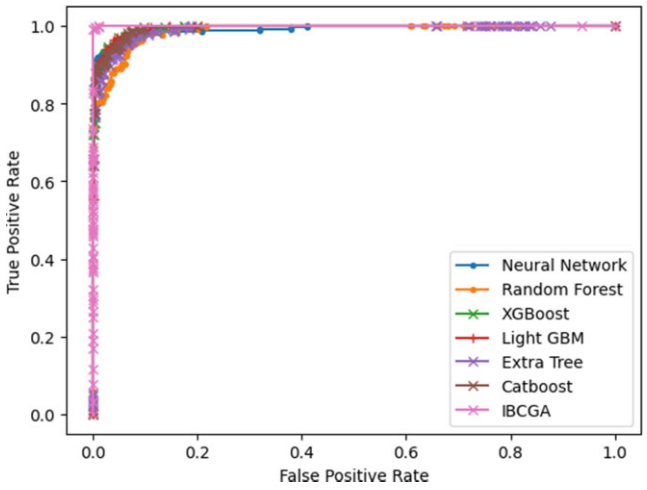

Figure 2. The comparison of BSig prediction performance. Evaluation of the prediction performance using AUROC between BSig and standard machine learning methods.

miRNA signature and its BC-specific target genes are shown in Figure 5A, B.

Additionally, we identified miRNA signature targeted BC-associated known genes using miRDB (62), miRWalk (63), and miRDIP (64). The miRNA signature targeted 41 BC-associated known genes, including, BRAP, BCCIP, BARD1, PTEN, TP53INP1, CFLAR, RAD50, MAP3K19, FGFR2, and CDH11, to name a few. The highest number of genes were targeted by the miRNAs, hsa-miR-6124 and hsa-miR-6768-5p, which targeted more than ten genes. The miRNA signature and BC-associated target genes are shown in Figure 5C, D.

Biological relevance of the miRNA-targeted genes

To validate the biological pathways and mechanisms of the identified miRNA-target genes, we performed pathway enrichment analyses, including Wikipathways, Elsevier collection, and GO biological process. The top three highly enriched pathways in each pathway analysis are listed as follows. In the Wikipathways, the targeted genes were enriched in glycolysis in senescence (WP5049), aerobic glycolysis (WP4629), and insulin signaling (WP481). Elsevier enrichment pathways also included glycolysis in addition to MTHFR activity regulation, and SIRT2 signaling in aging. GO biological analysis showed glucose catabolic process to pyruvate (GO: 0061718), canonical glycolysis (GO: 0061621), and glycolytic process through glucose-6-phosphate (GO: 61620). All the pathway enrichments are shown in Figure 6A–C. The enriched Wiki pathways, Elsevier pathway collection, and biological processes are listed in Supplementary Tables S6–S8, with their corresponding *P*-values.

We also conducted three gene-set enrichment analysis to validate the mechanism of the miRNA-target genes in BC cell lines, including ChEA (58), Project Achilles (59) and Cancer Cell Line Encyclopedia (60). The ChEA database contained comprehensive target genes of transcription factors from CHIP-chip, ChIP-seq, and other transcription factor binding site profiling studies. The results showed that MYC and MYCN were enriched from several studies, including the human breast adenocarcinoma cell line: MYC 28411283 ChIP-Seq MDA231-LM2-4175 Human BC (*P*-value: 2.78e-7, odds ratio: 4.24). Project Achilles is a systematic effort aimed at identifying and cataloging gene essentiality across hundreds of genomically characterized cancer cell lines (59). To infer gene fitness effects from CRISPR knockout screens computationally, researchers developed CERES (65), a test in which a more negative CERES score indicates that the gene is essential for cell viability in the certain cell line. In our study, the Achilles Fitness Decrease results showed that the knockout gene-set decreased cell viability in certain cell lines, with two breast-relevant cell lines appearing in the top three results, including HCC2218-breast (*P*-value: 3.38e-4, odds ratio: 9.15) and HCC70-breast (*P*-value: 6.29e-4, odds ratio: 7.94). The Cancer Cell Line Encyclopedia (60) results indicated the upregulated genes across certain cancer cell lines, with only three cell lines meeting the enriched criteria (P-value < 0.05), and two of them being BC cell line, including BT483-breast (*P*-value: 0.028, odds ratio: 4.77), and CAL51-breast (*P*-value: 0.049, odds ratio: 5.88). The miRNA signature targeted gene set enrichment in ChEA cell lines, Achilles fitness, and cancer cell line encyclopedia are depicted in Figure 6D–F, and corresponding statistics are listed in Supplementary Tables S9-S11, respectively.

In addition, to comprehensively investigate the circulating miRNAs that have been previously identified in serum, we conducted an in-depth review of the existing literature. Numerous studies have investigated the roles of serum miRNAs in BC (66–68). This review encompassed an assessment of circulating miRNAs specifically associated with BC, considering both their identification and their corresponding expression levels in BC samples. The detailed compilation of these BC-associated circulating miRNAs, along with their corresponding expression profiles, is presented in Supplementary Table S12.

Discussion

Recently, circulating tumor miRNA in blood become a promising biomarker to detect cancer. The stability of circulating miRNAs has been demonstrated to have utility for the minimally-invasive detection of various cancers (69–71). The identification of cancer-specific non-invasive miRNA signature leads to more accurate assessments to improve therapeutic strategies and early-stage detection of cancer.

In this study, we utilized miRNA expression profile data from serum samples of patients with BC and healthy individuals to develop a diagnosis prediction method BSig. BSig is an advanced evolutionary learning method designed to select a compact set of features (miRNAs) from a vast pool of features (miRNA expression profiles) specific to BC. BSig employs an integration of feature selection and machine learning techniques to select a distinctive miRNA signature capable of accurately predicting BC diagnosis. Through the utilization of BSig, we have successfully identified a miRNA signature composed of 12 BC-specific miRNAs, collectively demonstrating a robust ability to precisely predict BC diagnosis. This platform merges miRNA expression profile data with BC diagnosis information, generating outcomes that indicate whether input samples belong to the category of BC or not, employing an optimization approach. Although, there has been a growing interest in applying machine learning techniques to predict cancer diagnosis and prognosis, there is still a need for more accurate and customizable models to improve clinical decision-making. One key aspect we emphasize is the generalizability of the BSig method for cancer diagnosis predictions. Through

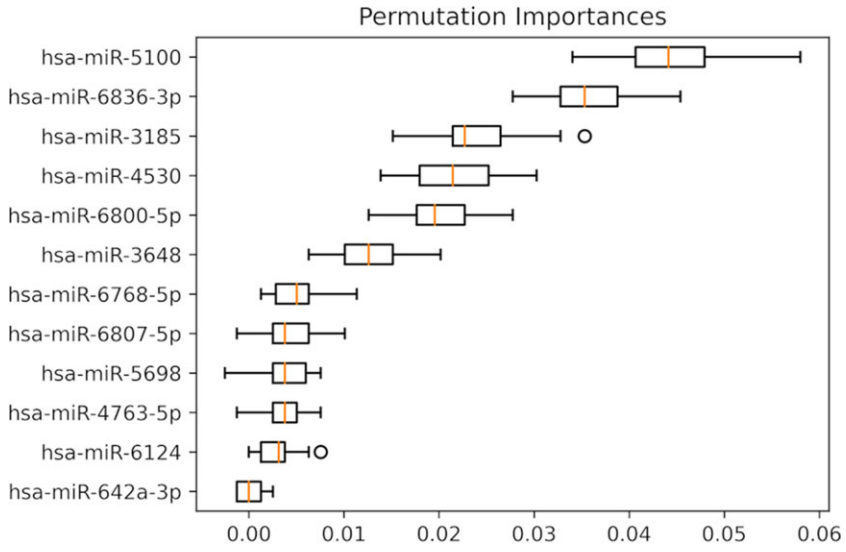

Figure 3. Permutation-based feature prioritization of the miRNA signature.

rigorous testing and validation, we have demonstrated its robustness and effectiveness in different datasets, diverse patient populations, and cancer types, indicating its potential to provide valuable insights across a wide range of clinical scenarios (34,36–39,41). By leveraging the power of evolutionary learning and optimization techniques, our approach leads to improved and personalized cancer diagnostics, ultimately contributing to more efficient and accurate clinical decision-making processes.

The identified miRNA signature accurately predicted BC diagnosis in both training and validation datasets. The BSig method prediction performance was better than the standard machine learning methods. To evaluate the efficacy of the individual miRNAs of the signature, we performed the permutation test and prioritize the miRNAs according to their prediction capability. Among these 12 miRNAs, 11 miRNAs, including hsa-miR-3185, hsa-miR-3648, hsa-miR-4530, hsa-miR-4763-5p, hsa-miR-5100, hsa-miR-5698, hsa-miR-6124, hsa-miR-6768-5p, hsa-miR-6800-5p, hsa-miR-6807-5p, and hsa-miR-6836-3p were significantly ($P \leq 0.005$) expressed between BC and healthy samples.

We identified 1968 target genes of the miRNA signature using miRTarBase (52). In addition, the BC-specific gene targets were identified using Chronos Gene Effect Scores from Cancer Cell Line Encyclopedia (60). The analysis showed that miRNA signature targeted 65 BC-specific genes, in which seven genes, including ATXN3, PPP1R15B, MED16, FOXK1, RAB11B, ALG14, and IGF1R were targeted by more than one miRNA of the signature. The pathway analysis of Wikipathways, Elsevier enrichment, and GO biological processes revealed that targeted genes were commonly enriched in glycolysis pathway. Many cancer cells preferentially consume glucose and utilize glycolysis to meet the demand for their survival, and particularly aerobic glycolysis is considered a hallmark of cancer (72). Emerging roles of aerobic glycolysis, known as the Warburg effect, in BC was well demonstrated in the studies (73,74). Altered metabolism is often observed in BC progression and metabolism indeed varied across subtypes of BC (75). Triple-negative BC cells are characterized by high glucose intake and low mitochondrial respiration (76) and HER-positive BC tumor characteristics were manifested in these studies by higher glutamine metabolic activity than other BC subtypes (77). MiRNAs mediate glycolytic pathways associated with cancer progression either directly or indirectly via targeting oncogenes. Furthermore, numerous studies have demonstrated the miRNA regulated glycolysis pathways in cancers (78,79). In BC patients, miR-16-1-3p regulates aerobic glycolysis which is critical for modulating BC cell proliferation *in vitro* and *in vivo* (80). Expression of miR-210-3p regulate aerobic glycolysis through modulating the glycolytic genes of HIF-1α and p53 in triple negative BC (81). Additionally, there are some miRNAs such as miR-27b, miR-31, miR-155, miR-340, miR-30a-5p, and miR-342-3p that facilitate glycolysis by targeting PDHX, DNMT3, PIK3R1, MCU, LDHA, and MCT1, respectively, in BC (82–84).

Gene-set enrichment analysis on the ChEA (58) database revealed that the enrichment of MYC and MYCN from the human breast adenocarcinoma cell lines. MYC amplification or overexpression has been linked to more aggressive tumor behavior such as a higher rate of tumor recurrence, tumor invasion, and metastasis. Moreover, high levels of MYC expression have been associated with a worse prognosis in BC patients, and MYC expression may also contribute to the development of resistance to some BC therapies. As a result, MYC is considered a potential therapeutic target for BC treatment (85).

We also explored the published roles of the 12 miRNAs in other cancers besides BC. Two distinct roles of hsa-miR-3185 were observed in hepatocellular carcinoma (86) and gastric cancers (87). In hepatocellular carcinoma, higher expression of miR-3185 correlated with improved survival whereas in gastric cancer, higher expression levels are associated with the poorer survival (87). Correlative analysis of miRNA expression revealed that the expression of hsa-miR-3648 positively correlated with risk of recurrence of estrogen receptor positive BC and lymph node-negative mammary carcinomas (88). The expression of hsa-miR-3648 was found to be associated with bladder, prostate, lung, hepatocellular carcinoma, and esophageal carcinoma. In these cancers, hsa-miR-3648 showed either positive correlation with cancer progression or regulate oncogenic function via binding to TFC21, APC2, VE-cadherin, Z0-1, SOCS, DLL4, PANX2, and NKAIN1

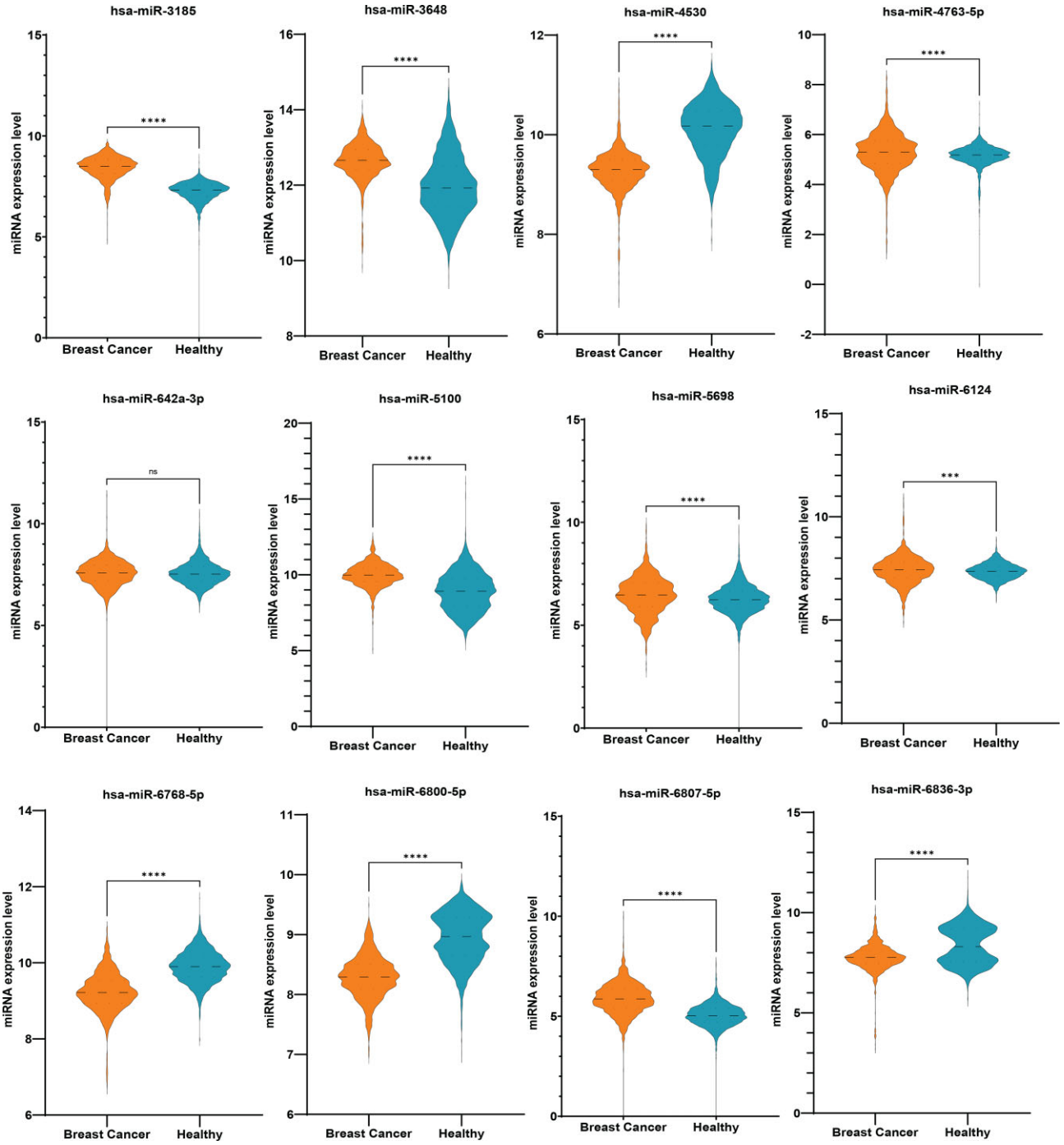

Figure 4. Comparison of miRNA expression difference across breast cancer and healthy individuals (* indicates *P*-value < 0.05).

(89–92). Wang *et al.*, identified that hsa-miR-4530 targets RUNX2 and associated with breast chemosensitivity (93). Expression of hsa-miR-4530 negatively correlated with tumor suppressive function resulting in tumor progression by targeting VASH1 and RTEL1 in different cancer types, including pancreatic, glioma and liver cancers (94,95). Hsa-miR-642a-3p encompassing tissue biomarkers and functional/target identification studies in cancer cell lines. Evidences suggested that expression of hsa-miR-642-3p targets FOXO4 and regulate oncogenic function in gastric, gallbladder, and cervical cancers (96,97). Hsa-miR-5100 possesses diverse role in various cancers, including BC. A BC study on xenografts reported that release of hsa-miR-5100 via exosome from PGRN-/-TAM (macrophages) inhibited invasion, migration and EMT of BC cells (98). The oncogenic function of hsa-miR-5100 was identified in lung cancer, oral squamous, esophageal, and melanoma (99–101), whereas it's tumor suppressor activity identified in pancreatic, childhood leukemia, gastric, prostate, multiple myeloma, gastrointestinal stromal, and BC (102–104). Additionally, hsa-miR-5100 has also been used in integrated panel as a prognostic biomarker for cancer detection in colon, oral squamous, pancreatic, and prostate cancers (105,106). The expression levels of hsa-miR-5698 was significantly associated with overall survival after the ini-

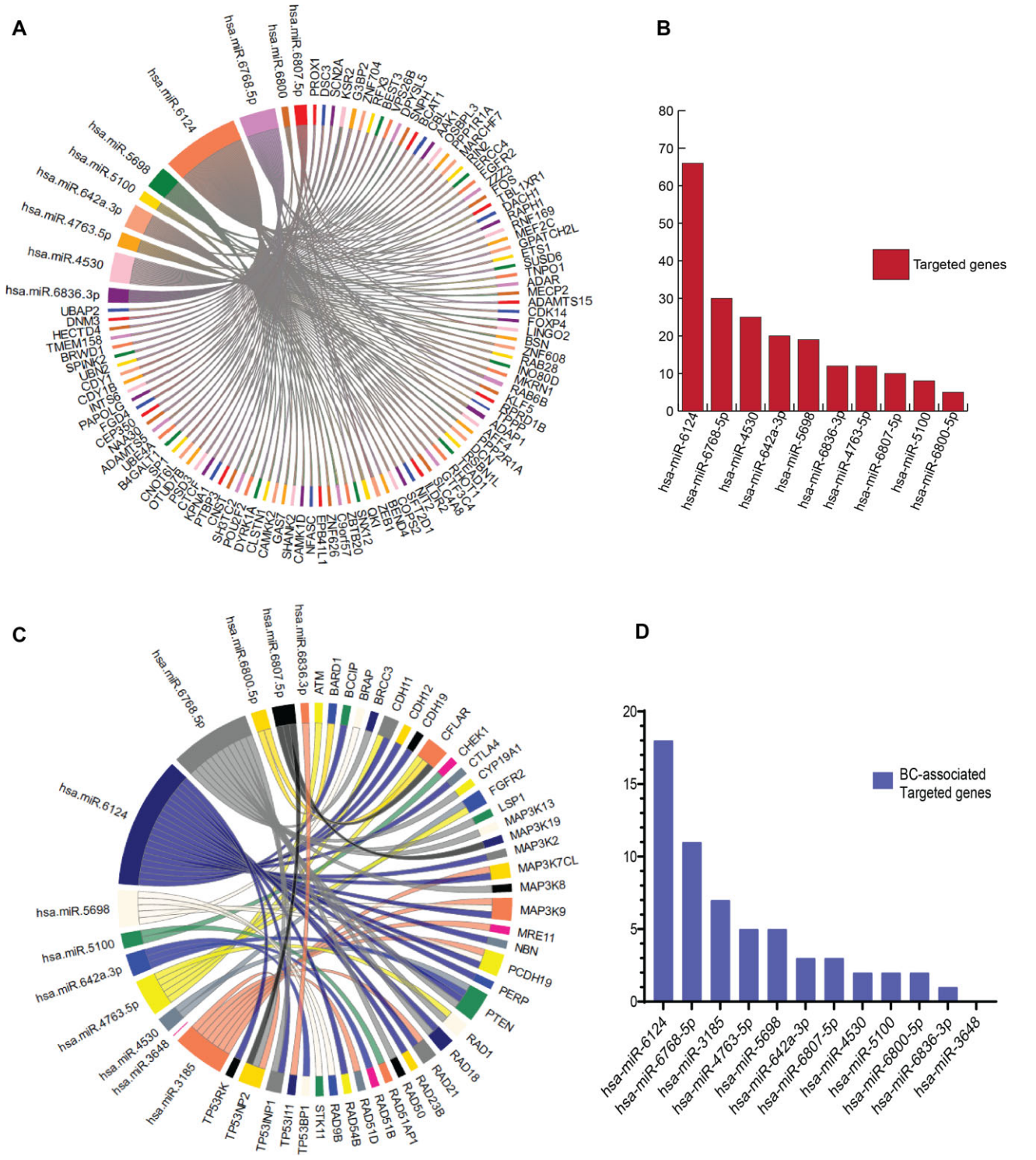

Figure 5. The miRNA and its breast cancer-specific target gene interaction. (**A**) 12 miRNAs and their 65 target genes in which seven genes, ATXN3, PPP1R15B, MED16, FOXK1, RAB11B, ALG14, and IGF1R were targeted by more than one miRNA and (**B**) the number of genes targeted by the individual miRNAs. (**C**) 12 miRNAs and their 41 breast cancer-associated known genes, and (**D**) the number of breast cancer known genes targeted by the individual miRNAs.

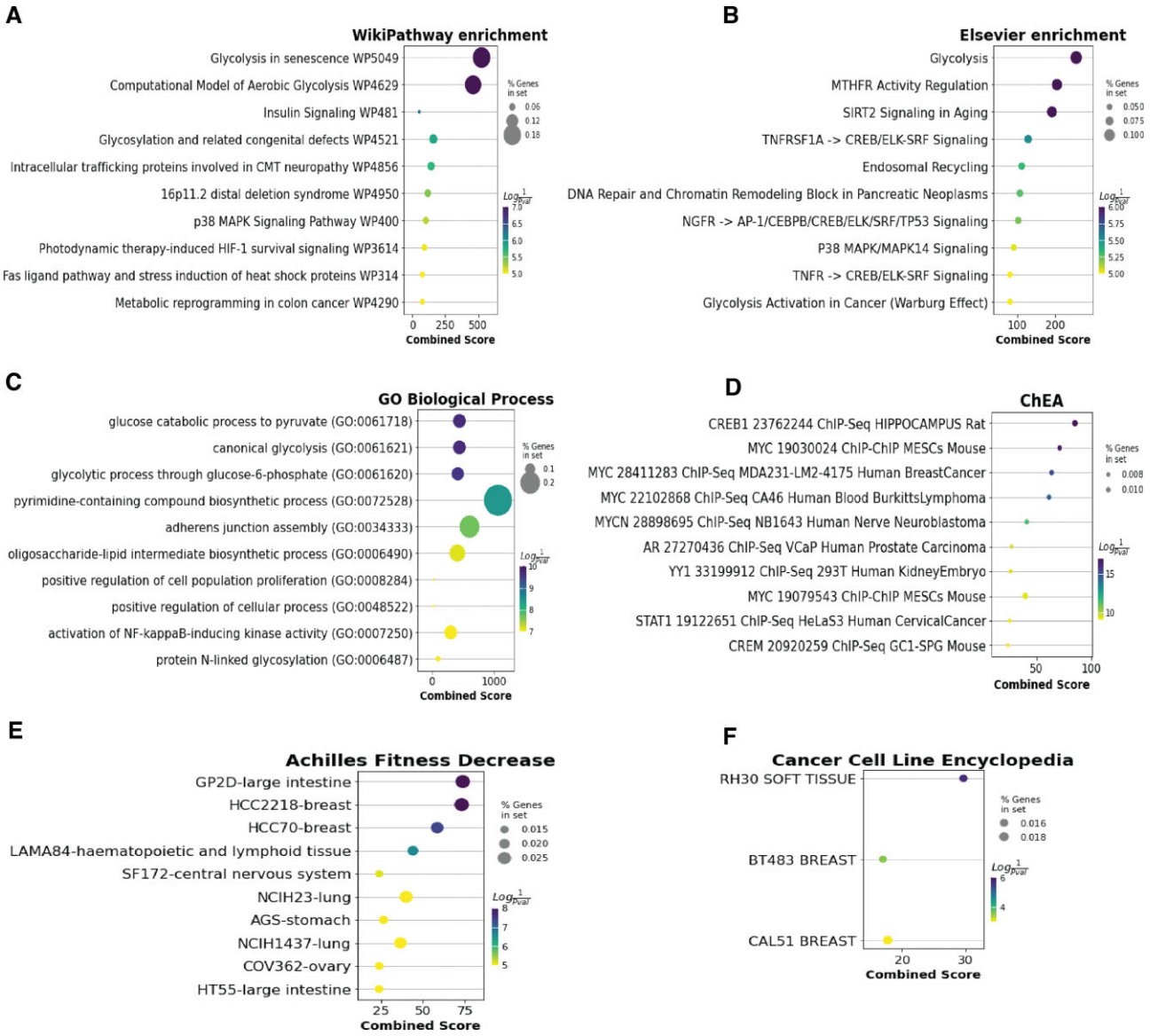

Figure 6. Gene enrichment analysis. The miRNA signature targeted gene set enrichment analysis in (**A**) Wiki pathways, (**B**) Elsevier pathways, (**C**) biological processes, (**D**) ChEA, (**E**) Achilles fitness decrease and (**F**) cancer cell line encyclopedia.

tiation of eribulin treatment in metastatic BC (107). Expression levels of hsa-miR-6124 and hsa-miR6807-5p were detected in urine, and their higher expression levels are associated with bladder cancer (108) and hepatocellular carcinoma (109). Bioinformatics analysis on pan cancer revealed that hsa-miR-6124 promotes tumor progression by binding 3′ untranslated regions (3′-UTRs) of ALYREF and eIF4A3 in hepatocellular carcinoma (109). Studies on hsa-miR-6836-3p indicated that its expression level increases upon Apigenin treatment. Apigenin significantly inhibited Huh7 cell proliferation, cell cycle, colony formation, and cell invasion in a concentration-dependent manner in hepatocellular carcinoma (110).

Taken together, our findings show that miRNA signature consisting of 12 miRNAs could be potentially useful to BC cancer screening and help in developing the minimally-invasive methods in BC. The fact that several of the miRNAs identified in this study has shown to have a role in other cancers suggests a shared signaling pathways across several cancers. The enrichment of glycolysis pathways in our study further validates a role of it cancers in general.

Utilizing miRNAs for diagnostic predictions presents certain limitations, including variability in miRNA regulation and functions, as well as technological and methodological challenges in detecting and quantifying miRNAs. It is important to consider the influence of various physiological and pathological conditions on miRNA profiles and the consequent implications for their reliability and specificity in diagnostics. This comprehensive examination highlights the potential of miRNA-based diagnostics while emphasizing critical aspects that require attention for their effective implementation in clinical practice. Though the prediction ability of the BSig is accurate, the prediction models are currently limited to the GEO datasets: GSE73002 and GSE211692. This is due to the lack of standardization strategies for isolation, quantification and normalization of miRNA expression profiles. A larger sample sets including prospective clinical data are required for validating the models to applicable for clinical

routine. However, the miRNA signature identified here is robust with higher specificity and parameter tuning of the BSig method could improve the prediction performances on larger datasets. Additionally, BSig can be customized for diagnosis prediction in other cancer types, and would have potential in clinical applications.

In conclusion, the identified miRNA signature could serve as a valuable adjunct to the existing diagnostic methodologies, particularly in scenarios where imaging and biopsy results may be inconclusive or require further validation. In this context, the miRNA signature could enhance the diagnostic accuracy and help refine the diagnostic process. While our focus primarily centers on diagnostic support, we acknowledge that the signature's accuracy and sensitivity could potentially position it as a candidate for future studies in BC screening, particularly in asymptomatic individuals who might be at risk.

Data availability

The datasets utilized in this study are available at the following GEO accessions: GSE73002 and GSE211692. The BSig implementation and model files can be found in GitHub (https://github.com/mingjutsai/BSig) and Zenodo (https://doi.org/10.5281/zenodo.10625880).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors acknowledge David Puthoff, PhD, from the MCRI for manuscript editing assistance.

Author contributions: S.Y.S. designed the system and supervised and carried out the detail study. S.Y.S., M.T., N.A., M.L., S.H., P.A., S.K.S., A.B., R.S. and S.Y.H. participated in data analysis, manuscript preparation and discussed the results. All authors have read and approved the final manuscript.

Funding

This work was supported in part by institutional funding from Marshfield Clinic Research Institute (MCRI) to SYS and MCRI Weber Endowment to SKS, Marshfield, WI. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest statement

None declared.

References

1. Ferlay,J., Colombet,M., Soerjomataram,I., Parkin,D.M., Piñeros,M., Znaor,A. and Bray,F. (2021) Cancer statistics for the year 2020: an overview. *Int. J. Cancer*, **149**, 778–789.
2. Siegel,R.L., Miller,K.D., Fuchs,H.E. and Jemal,A. (2021) Cancer statistics, 2021. *CA Cancer J. Clin.*, **71**, 7–33.
3. Benz,C.C. (2008) Impact of aging on the biology of breast cancer. *Crit. Rev. Oncol. Hematol.*, **66**, 65–74.
4. Brewer,H.R., Jones,M.E., Schoemaker,M.J., Ashworth,A. and Swerdlow,A.J. (2017) Family history and risk of breast cancer: an analysis accounting for family structure. *Breast Cancer Res. Treat.*, **165**, 193–200.
5. Rosato,V., Bosetti,C., Negri,E., Talamini,R., Dal Maso,L., Malvezzi,M., Falcini,F., Montella,M. and La Vecchia,C. (2014) Reproductive and hormonal factors, family history, and breast cancer according to the hormonal receptor status. *Eur. J. Cancer Prev.*, **23**, 412–417.
6. Key,T.J., Appleby,P.N., Reeves,G.K., Travis,R.C., Alberg,A.J., Barricarte,A., Berrino,F., Krogh,V., Sieri,S., Brinton,L.A., *et al.* (2013) Sex hormones and risk of breast cancer in premenopausal women: A collaborative reanalysis of individual participant data from seven prospective studies. *Lancet Oncol.*, **14**, 1009–1019.
7. Shiovitz,S. and Korde,L.A. (2015) Genetics of breast cancer: A topic in evolution. *Ann. Oncol.*, **26**, 1291–1299.
8. Guo,R., Lu,G., Qin,B. and Fei,B. (2018) Ultrasound imaging technologies for breast cancer detection and management: A review. *Ultrasound Med. Biol.*, **44**, 37–70.
9. Byra,M., Galperin,M., Ojeda-Fournier,H., Olson,L., O'Boyle,M., Comstock,C. and Andre,M. (2019) Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Med. Phys.*, **46**, 746–755.
10. Olopade,O.I., Grushko,T.A., Nanda,R. and Huo,D. (2008) Advances in breast cancer: pathways to personalized medicine. *Clin. Cancer Res.*, **14**, 7988–7999.
11. Alamdari,S.G., Amini,M., Jalilzadeh,N., Baradaran,B., Mohammadzadeh,R., Mokhtarzadeh,A. and Oroojalian,F. (2022) Recent advances in nanoparticle-based photothermal therapy for breast cancer. *J. Controlled Release*, **349**, 269–303.
12. Roy,P.S. and Saikia,B.J. (2016) Cancer and cure: a critical analysis. *Indian J. Cancer*, **53**, 441–442.
13. Weber,J.A., Baxter,D.H., Zhang,S., Huang,D.Y., Huang,K.H., Lee,M.J., Galas,D.J. and Wang,K. (2010) The microRNA spectrum in 12 body fluids. *Clin. Chem.*, **56**, 1733–1741.
14. Bennett,M.R. and Devarajan,P. (2011) In: Edelstein,C.L. (ed.) *Biomar kers of Kidney Disease*. Academic Press, San Diego, pp. 1–24.
15. Hanash,S.M. (2011) Why have protein biomarkers not reached the clinic? *Genome Medicine*, **3**, 66.
16. Condrat,C.E., Thompson,D.C., Barbu,M.G., Bugnar,O.L., Boboc,A., Cretoiu,D., Suciu,N., Cretoiu,S.M. and Voinea,S.C. (2020) miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis. *Cells*, **9**, 276.
17. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
18. Jackson,R.J. and Standart,N. (2007) How do microRNAs regulate gene expression? *Sci. STKE*, **2007**, re1.
19. Peng,Y. and Croce,C.M. (2016) The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*, **1**, 15004.
20. Baer,C., Claus,R. and Plass,C. (2013) Genome-wide epigenetic regulation of miRNAs in cancer. *Cancer Res.*, **73**, 473–477.
21. Wang,W.-T. and Chen,Y.-Q. (2014) Circulating miRNAs in cancer: from detection to therapy. *J. Hematol. Oncol.*, **7**, 86.
22. Hamam,R., Hamam,D., Alsaleh,K.A., Kassem,M., Zaher,W., Alfayez,M., Aldahmash,A. and Alajez,N.M. (2017) Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death. Dis.*, **8**, e3045.
23. Escuin,D., López-Vilaró,L., Mora,J., Bell,O., Moral,A., Pérez,I., Arqueros,C., García-Valdecasas,B., Ramón,Y.C.T., Lerma,E., *et al.* (2021) Circulating microRNAs in early breast cancer patients and its association with lymph node metastases. *Front. Oncol.*, **11**, 627811.
24. Martino,E., D'Onofrio,N., Anastasio,C., Abate,M., Zappavigna,S., Caraglia,M. and Balestrieri,M.L. (2023) MicroRNA-nanoparticles against cancer: opportunities and challenges for personalized medicine. *Mol. Ther. Nucleic Acids*, **32**, 371–384.
25. Becker,V., Yuan,X., Boewe,A.S., Ampofo,E., Ebert,E., Hohneck,J., Bohle,R.M., Meese,E., Zhao,Y., Menger,M.D., *et al.* (2023) Hypoxia-induced downregulation of microRNA-186-5p in endothelial cells promotes non-small cell lung cancer

angiogenesis by upregulating protein kinase C alpha. *Mol. Ther. Nucleic Acids*, **31**, 421–436.
26. Gilad,S., Meiri,E., Yogev,Y., Benjamin,S., Lebanony,D., Yerushalmi,N., Benjamin,H., Kushnir,M., Cholakh,H., Melamed,N., *et al.* (2008) Serum microRNAs are promising novel biomarkers. *PLoS One*, **3**, e3148.
27. Tiberio,P., Gaudio,M., Belloni,S., Pindilli,S., Benvenuti,C., Jacobs,F., Saltalamacchia,G., Zambelli,A., Santoro,A. and De Sanctis,R. (2023) Unlocking the potential of circulating miRNAs in the breast cancer neoadjuvant setting: a systematic review and meta-analysis. *Cancers*, **15**, 3424.
28. El-Toukhy,S.E., El-Daly,S.M., Kamel,M.M. and Nabih,H.K. (2023) The diagnostic significance of circulating miRNAs and metabolite profiling in early prediction of breast cancer in Egyptian women. *J. Cancer Res. Clin. Oncol.*, **149**, 5437–5451.
29. Davey,M.G., McGuire,A., Casey,M.C., Waldron,R.M., Paganga,M., Holian,E., Newell,J., Heneghan,H.M., McDermott,A.M., Keane,M.M., *et al.* (2023) Evaluating the role of circulating microRNAs in predicting long-term survival outcomes in breast cancer: a prospective, multicenter clinical trial. *J. Am. Coll. Surg.*, **236**, 317–327.
30. Mishra,S., Srivastava,A.K., Suman,S., Kumar,V. and Shukla,Y. (2015) Circulating miRNAs revealed as surrogate molecular signatures for the early detection of breast cancer. *Cancer Lett.*, **369**, 67–75.
31. Matsuzaki,J., Kato,K., Oono,K., Tsuchiya,N., Sudo,K., Shimomura,A., Tamura,K., Shiino,S., Kinoshita,T., Daiko,H., *et al.* (2023) Prediction of tissue-of-origin of early stage cancers using serum miRNomes. *JNCI Cancer Spectrum*, **7**, pkac080.
32. Xu,C. and Jackson,S.A. (2019) Machine learning and complex biological data. *Genome Biol.*, **20**, 76.
33. Wang,S., Celebi,M.E., Zhang,Y.-D., Yu,X., Lu,S., Yao,X., Zhou,Q., Miguel,M.-G., Tian,Y., Gorriz,J.M., *et al.* (2021) Advances in data preprocessing for biomedical data fusion: an overview of the methods, challenges, and prospects. *Information Fusion*, **76**, 376–421.
34. Yerukala Sathipati,S. and Ho,S.Y. (2017) Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. *Sci. Rep.*, **7**, 7507.
35. Yerukala Sathipati,S., Huang,H.-L. and Ho,S.-Y. (2016) Estimating survival time of patients with glioblastoma multiforme and characterization of the identified microRNA signatures. *Bmc Genomics [Electronic Resource]*, **17**, 1022.
36. Yerukala Sathipati,S., Tsai,M.J., Shukla,S.K., Ho,S.Y., Liu,Y. and Beheshti,A. (2022) MicroRNA signature for estimating the survival time in patients with bladder urothelial carcinoma. *Sci. Rep.*, **12**, 4141.
37. Yerukala Sathipati,S., Tsai,M.-J., Carter,T., Allaire,P., Shukla,S.K., Beheshti,A. and Ho,S.-Y. (2022) Survival estimation in patients with stomach and esophageal carcinoma using miRNA expression profiles. *Comput. Struct. Biotechnol. J.*, **20**, 4490–4500.
38. Yerukala Sathipati,S. and Ho,S.-Y. (2018) Identifying a miRNA signature for predicting the stage of breast cancer. *Sci. Rep.*, **8**, 16138.
39. Yerukala Sathipati,S. and Ho,S.-Y. (2020) Novel miRNA signature for predicting the stage of hepatocellular carcinoma. *Sci. Rep.*, **10**, 14452.
40. Shimomura,A., Shiino,S., Kawauchi,J., Takizawa,S., Sakamoto,H., Matsuzaki,J., Ono,M., Takeshita,F., Niida,S., Shimizu,C., *et al.* (2016) Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci.*, **107**, 326–334.
41. Yerukala Sathipati,S., Tsai,M.-J., Shukla,S.K. and Ho,S.-Y. (2023) Artificial intelligence-driven pan-cancer analysis reveals miRNA signatures for cancer stage prediction. *Hum. Genet. Genomics Adv.*, **4**, 100190.
42. Noble,W.S. (2006) What is a support vector machine?*Nat. Biotechnol.*, **24**, 1565–1567.
43. Yerukala Sathipati,S., Tsai,M.J., Carter,T., Shukla,S.K. and Ho,S.Y. (2022) SPIKES: identification of physicochemical properties of spike proteins across diverse host species of SARS-CoV-2. *STAR Protoc*, **3**, 101460.
44. Ho,S.-Y., Shu,L.-S. and Chen,J.-H. (2004) Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Trans. Evol. Comput.*, **8**, 522–541.
45. Tsai,M.J., Wang,J.R., Ho,S.J., Shu,L.S., Huang,W.L. and Ho,S.Y. (2020) GREMA: modelling of emulated gene regulatory networks with confidence levels based on evolutionary intelligence to cope with the underdetermined problem. *Bioinformatics*, **36**, 3833–3840.
46. Tsai,M.J., Wang,J.R., Yang,C.D., Kao,K.C., Huang,W.L., Huang,H.Y., Tseng,C.P., Huang,H.D. and Ho,S.Y. (2018) PredCRP: predicting and analysing the regulatory roles of CRP from its binding sites in *Escherichia coli*. *Sci. Rep.*, **8**, 951.
47. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
48. Chen,T. and Guestrin,C. (2016) In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, San Francisco, California, USA, pp. 785–794.
49. Ke,G., Meng,Q., Finley,T., Wang,T., Chen,W., Ma,W., Ye,Q. and Liu,T.-Y. (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inform. Process. Syst.*, **30**, 3149–3157.
50. Dorogush,A.V., Ershov,V. and Gulin,A. (2018) CatBoost: gradient boosting with categorical features support. arXiv doi: https://arxiv.org/abs/1810.11363, 24 October 2018, preprint: not peer reviewed.
51. Akiba,T., Sano,S., Yanase,T., Ohta,T. and Koyama,M. (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Vol. **19**, pp. 2623–2631.
52. Huang,H.-Y., Lin,Y.-C.-D., Li,J., Huang,K.-Y., Shrestha,S., Hong,H.-C., Tang,Y., Chen,Y.-G., Jin,C.-N. and Yu,Y. (2020) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.
53. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V. and Sonkin,D. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
54. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M. and Lachmann,A. (2016) Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
55. Kutmon,M., Riutta,A., Nunes,N., Hanspers,K., Willighagen,E.L., Bohler,A., Mélius,J., Waagmeester,A., Sinha,S.R., Miller,R., *et al.* (2016) WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.*, **44**, D488–D494.
56. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A., *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
57. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
58. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma'ayan,A. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.
59. Tsherniak,A., Vazquez,F., Montgomery,P.G., Weir,B.A., Kryukov,G., Cowley,G.S., Gill,S., Harrington,W.F., Pantel,S.,

Krill-Burger,J.M., *et al.* (2017) Defining a cancer dependency map. *Cell*, **170**, 564–576.
60. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V., Sonkin,D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
61. Ke,G., Meng,Q., Finley,T., Wang,T., Chen,W., Ma,W., Ye,Q. and Liu,T.-Y. (2017), LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS*, **30**, 3149–3157.
62. Chen,Y. and Wang,X. (2020) miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.*, **48**, D127–D131.
63. Sticht,C., De La Torre,C., Parveen,A. and Gretz,N. (2018) miRWalk: An online resource for prediction of microRNA binding sites. *PLoS One*, **13**, e0206239.
64. Tokar,T., Pastrello,C., Rossos,A.E.M., Abovsky,M., Hauschild,A.-C., Tsay,M., Lu,R. and Jurisica,I. (2018) mirDIP 4.1—integrative database of human microRNA target predictions. *Nucleic Acids Res.*, **46**, D360–D370.
65. Meyers,R.M., Bryan,J.G., McFarland,J.M., Weir,B.A., Sizemore,A.E., Xu,H., Dharia,N.V., Montgomery,P.G., Cowley,G.S., Pantel,S., *et al.* (2017) Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.*, **49**, 1779–1784.
66. Zhang,Z., Zhang,H., Yu,J., Xu,L., Pang,X., Xiang,Q., Liu,Q. and Cui,Y. (2022) miRNAs as therapeutic predictors and prognostic biomarkers of neoadjuvant chemotherapy in breast cancer: A systematic review and meta-analysis. *Breast Cancer Res. Treat.*, **194**, 483–505.
67. Tarighati,E., Keivan,H. and Mahani,H. (2023) A review of prognostic and predictive biomarkers in breast cancer. *Clin. Exp. Med.*, **23**, 1–16.
68. Hamam,R., Hamam,D., Alsaleh,K.A., Kassem,M., Zaher,W., Alfayez,M., Aldahmash,A. and Alajez,N.M. (2017) Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death. Dis.*, **8**, e3045.
69. Brase,J.C., Wuttig,D., Kuner,R. and Sültmann,H. (2010) Serum microRNAs as non-invasive biomarkers for cancer. *Mol. Cancer*, **9**, 306.
70. Lu,Z., He,Q., Liang,J., Li,W., Su,Q., Chen,Z., Wan,Q., Zhou,X., Cao,L., Sun,J., *et al.* (2019) miR-31-5p is a potential circulating biomarker and therapeutic target for oral cancer. *Mol. Ther. Nucleic Acids*, **16**, 471–480.
71. Li,X., Li,Y., Zhao,L., Zhang,D., Yao,X., Zhang,H., Wang,Y.-c., Wang,X.-y., Xia,H., Yan,J., *et al.* (2014) Circulating muscle-specific miRNAs in Duchenne muscular dystrophy patients. *Mol. Ther. Nucleic Acids*, **3**, e177.
72. Warburg,O. (1956) On the origin of cancer cells. *Science*, **123**, 309–314.
73. Wu,Z., Wu,J., Zhao,Q., Fu,S. and Jin,J. (2020) Emerging roles of aerobic glycolysis in breast cancer. *Clin. Transl. Oncol.*, **22**, 631–646.
74. Li,W., Xu,M., Li,Y., Huang,Z., Zhou,J., Zhao,Q., Le,K., Dong,F., Wan,C. and Yi,P. (2020) Comprehensive analysis of the association between tumor glycolysis and immune/inflammation function in breast cancer. *J. Transl. Med.*, **18**, 92.
75. Brglez,V., Pucer,A., Pungerčar,J., Lambeau,G. and Petan,T. (2014) Secreted phospholipases A2 are differentially expressed and epigenetically silenced in human breast cancer cells. *Biochem. Biophys. Res. Commun.*, **445**, 230–235.
76. Pelicano,H., Zhang,W., Liu,J., Hammoudi,N., Dai,J., Xu,R.-H., Pusztai,L. and Huang,P. (2014) Mitochondrial dysfunction in some triple-negative breast cancer cell lines: role of mTOR pathway and therapeutic potential. *Breast Cancer Res.*, **16**, 434.
77. Kim,S., Lee,Y. and Koo,J.S. (2015) Differential expression of lipid metabolism-related proteins in different breast cancer subtypes. *PLoS One*, **10**, e0119473.
78. Ahmad,A., Aboukameel,A., Kong,D., Wang,Z., Sethi,S., Chen,W., Sarkar,F.H. and Raz,A. (2011) Phosphoglucose isomerase/autocrine motility factor mediates epithelial-mesenchymal transition regulated by miR-200 in breast cancer cells EMT regulation by PGI/AMF is mediated by miR-200. *Cancer Res.*, **71**, 3400–3409.
79. Fabani,M.M. and Gait,M.J. (2008) miR-122 targeting with LNA/2′-O-methyl oligonucleotide mixmers, peptide nucleic acids (PNA), and PNA–peptide conjugates. *RNA*, **14**, 336–346.
80. Ye,T., Liang,Y., Zhang,D. and Zhang,X. (2020) MicroRNA-16-1-3p represses breast tumor growth and metastasis by inhibiting PGK1-mediated Warburg effect. *Front. Cell Dev. Biol.*, **8**, 615154.
81. Du,Y., Wei,N., Ma,R., Jiang,S. and Song,D. (2020) A miR-210-3p regulon that controls the Warburg effect by modulating HIF-1α and p53 activity in triple-negative breast cancer. *Cell Death. Dis.*, **11**, 731.
82. Eastlack,S.C., Dong,S., Ivan,C. and Alahari,S.K. (2018) Suppression of PDHX by microRNA-27b deregulates cell metabolism and promotes growth in breast cancer. *Mol. Cancer*, **17**, 1–16.
83. Zhao,Y., He,J., Yang,L., Luo,Q. and Liu,Z. (2018) Histone deacetylase-3 modification of MicroRNA-31 promotes cell proliferation and aerobic glycolysis in breast cancer and is predictive of poor prognosis. *J. Breast Cancer*, **21**, 112–123.
84. Romero-Cordoba,S.L., Rodriguez-Cuevas,S., Bautista-Pina,V., Maffuz-Aziz,A., D'Ippolito,E., Cosentino,G., Baroni,S., Iorio,M.V. and Hidalgo-Miranda,A. (2018) Loss of function of miR-342-3p results in MCT1 over-expression and contributes to oncogenic metabolic reprogramming in triple negative breast cancer. *Sci. Rep.*, **8**, 12252.
85. Xu,J., Chen,Y. and Olopade,O.I. (2010) MYC and breast cancer. *Genes Cancer*, **1**, 629–640.
86. Pascut,D., Pratama,M.Y., Gilardi,F., Giuffrè,M., Crocè,L.S. and Tiribelli,C. (2020) Weighted miRNA co-expression networks analysis identifies circulating miRNA predicting overall survival in hepatocellular carcinoma patients. *Sci. Rep.*, **10**, 18967.
87. Yu,Z., Rong,Z., Sheng,J., Luo,Z., Zhang,J., Li,T., Zhu,Z., Fu,Z., Qiu,Z. and Huang,C. (2021) Aberrant non-coding RNA expressed in gastric cancer and its diagnostic value. *Front. Oncol.*, **11**, 606764.
88. Emmadi,R., Canestrari,E., Arbieva,Z.H., Mu,W., Dai,Y., Frasor,J. and Wiley,E. (2015) Correlative analysis of miRNA expression and oncotype Dx recurrence score in estrogen receptor positive breast carcinomas. *PLoS One*, **10**, e0145346.
89. Hu,Y., Dingerdissen,H., Gupta,S., Kahsay,R., Shanker,V., Wan,Q., Yan,C. and Mazumder,R. (2018) Identification of key differentially expressed microRNAs in cancer patients through pan-cancer analysis. *Comput. Biol. Med.*, **103**, 183–197.
90. Sun,W., Li,S., Yu,Y., Jin,H., Xie,Q., Hua,X., Wang,S., Tian,Z., Zhang,H., Jiang,G., *et al.* (2019) MicroRNA-3648 is upregulated to suppress TCF21, resulting in promotion of invasion and metastasis of human bladder cancer. *Mol. Ther. Nucleic Acids*, **16**, 519–530.
91. Liu,X., Jiang,T., Li,X., Zhao,C., Li,J., Zhou,F., Zhang,L., Zhao,S., Jia,Y., Shi,J., *et al.* (2020) Exosomes transmit T790M mutation-induced resistance in EGFR-mutant NSCLC by activating PI3K/AKT signalling pathway. *J. Cell. Mol. Med.*, **24**, 1529–1540.
92. Yokota,Y., Noda,T., Okumura,Y., Kobayashi,S., Iwagami,Y., Yamada,D., Tomimaru,Y., Akita,H., Gotoh,K., Takeda,Y., *et al.* (2021) Serum exosomal miR-638 is a prognostic marker of HCC via downregulation of VE-cadherin and ZO-1 of endothelial cells. *Cancer Sci.*, **112**, 1275–1288.
93. Wang,X.X., Ye,F.G., Zhang,J., Li,J.J., Chen,Q.X., Lin,P.Y. and Song,C.G. (2018) Serum miR-4530 sensitizes breast cancer to neoadjuvant chemotherapy by suppressing RUNX2. *Cancer Manag Res*, **10**, 4393–4400.

94. Kojima,M., Sudo,H., Kawauchi,J., Takizawa,S., Kondou,S., Nobumasa,H. and Ochiai,A. (2015) MicroRNA markers for the diagnosis of pancreatic and biliary-tract cancers. *PLoS One*, **10**, e0118220.
95. Pratama,M.Y., Visintin,A., Crocè,L.S., Tiribelli,C. and Pascut,D. (2020) Circulatory miRNA as a biomarker for therapy response and disease-free survival in hepatocellular carcinoma. *Cancers (Basel)*, **12**, 2810.
96. Cao,J., Shao,H., Hu,J., Jin,R., Feng,A., Zhang,B., Li,S., Chen,T., Jeungpanich,S., Topatana,W., *et al.* (2022) Identification of invasion-metastasis associated MiRNAs in gallbladder cancer by bioinformatics and experimental validation. *J. Transl. Med.*, **20**, 188.
97. Zhang,L., Zhang,Y., Zhou,J., Wang,Y., Wang,H., Huang,M., Yu,Q. and Qi,S. (2022) LncRNA NR2F1-AS1 inhibits the malignant properties of cervical cancer cells via targeting miR-642a-3p/NR2F1 Axis. *Rev. Invest. Clin.*, **74**, 181–192.
98. Yue,S., Ye,X., Zhou,T., Gan,D., Qian,H., Fang,W., Yao,M., Zhang,D., Shi,H. and Chen,T. (2021) PGRN(-/-) TAMs-derived exosomes inhibit breast cancer cell invasion and migration and its mechanism exploration. *Life Sci.*, **264**, 118687.
99. Shi,J., Bao,X., Liu,Z., Zhang,Z., Chen,W. and Xu,Q. (2019) Serum miR-626 and miR-5100 are Promising Prognosis Predictors for Oral Squamous Cell Carcinoma. *Theranostics*, **9**, 920–931.
100. Ma,Z., Zhu,T., Wang,H., Wang,B., Fu,L. and Yu,G. (2022) Investigation of serum markers of esophageal squamous cell carcinoma based on machine learning methods. *J. Biochem.*, **172**, 29–36.
101. Suwei,D., Yanbin,X., Jianqiang,W., Xiang,M., Zhuohui,P., Jianping,K., Yunqing,W. and Zhen,L. (2022) Metformin inhibits melanoma cell metastasis by suppressing the miR-5100/SPINK5/STAT3 axis. *Cell. Mol. Biol. Lett.*, **27**, 48.
102. Zhang,H., Wang,J., Wang,Y., Li,J., Zhao,L., Zhang,T. and Liao,X. (2022) Long non-coding LEF1-AS1 sponge miR-5100 regulates apoptosis and autophagy in gastric cancer cells via the miR-5100/DEK/AMPK-mTOR Axis. *Int. J. Mol. Sci.*, **23**, 4787.
103. Mello-Grand,M., Bruno,A., Sacchetto,L., Cristoni,S., Gregnanin,I., Dematteis,A., Zitella,A., Gontero,P., Peraldo-Neia,C., Ricotta,R., *et al.* (2021) Two novel ceramide-like molecules and miR-5100 levels as biomarkers improve prediction of prostate cancer in gray-zone PSA. *Front. Oncol.*, **11**, 769158.
104. Saltarella,I., Lamanuzzi,A., Desantis,V., Di Marzo,L., Melaccio,A., Curci,P., Annese,T., Nico,B., Solimando,A.G., Bartoli,G., *et al.* (2022) Myeloma cells regulate miRNA transfer from fibroblast-derived exosomes by expression of lncRNAs. *J. Pathol.*, **256**, 402–413.
105. Jacob,H., Stanisavljevic,L., Storli,K.E., Hestetun,K.E., Dahl,O. and Myklebust,M.P. (2018) A four-microRNA classifier as a novel prognostic marker for tumor recurrence in stage II colon cancer. *Sci. Rep.*, **8**, 6157.
106. Liu,H.P., Lai,H.M. and Guo,Z. (2021) Prostate cancer early diagnosis: circulating microRNA pairs potentially beyond single microRNAs upon 1231 serum samples. *Brief. Bioinform.*, **22**, bbaa111.
107. Satomi-Tsushita,N., Shimomura,A., Matsuzaki,J., Yamamoto,Y., Kawauchi,J., Takizawa,S., Aoki,Y., Sakamoto,H., Kato,K., Shimizu,C., *et al.* (2019) Serum microRNA-based prediction of responsiveness to eribulin in metastatic breast cancer. *PLoS One*, **14**, e0222024.
108. Piao,X.M., Jeong,P., Kim,Y.H., Byun,Y.J., Xu,Y., Kang,H.W., Ha,Y.S., Kim,W.T., Lee,J.Y., Woo,S.H., *et al.* (2019) Urinary cell-free microRNA biomarker could discriminate bladder cancer from benign hematuria. *Int. J. Cancer*, **144**, 380–388.
109. Xue,C., Zhao,Y., Li,G. and Li,L. (2021) Multi-omic analyses of the m(5)C regulator ALYREF reveal its essential roles in hepatocellular carcinoma. *Front. Oncol.*, **11**, 633415.
110. Wang,S.M., Yang,P.W., Feng,X.J., Zhu,Y.W., Qiu,F.J., Hu,X.D. and Zhang,S.H. (2021) Apigenin inhibits the growth of hepatocellular carcinoma cells by affecting the expression of microRNA transcriptome. *Front. Oncol.*, **11**, 657665.

Received: November 21, 2023. **Revised:** January 11, 2024. **Editorial Decision:** February 5, 2024. **Accepted:** February 13, 2024