BioDecoder: A miRNA Bio-interpretable Neural Network Model for Noninvasive Diagnosis of Breast Cancer

Lei Liu[1, #], Weili Lin[1, #], Suqi Cao[2], Liu Yang[2], Sheng Gao[1], Na Jiao[2], Lixin Zhu[3], Ruixin Zhu[1, *], Dingfeng Wu[2, *]

1. Department of Gastroenterology, The Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200072, P. R. China.

2. National Clinical Research Center for Child Health, the Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310058, Zhejiang, P. R. China.

3. Department of Colorectal Surgery, Guangdong Institute of Gastroenterology, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, P. R. China.

[#]These authors contributed equally: Lei Liu, Suqi Cao, and Liu Yang.

[*]Corresponding authors:

Ruixin Zhu (rxzhu@tongji.edu.cn)

Department of Gastroenterology, The Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200072, P. R. China.

Tel: 86-21-6598-1041

Dingfeng Wu (dfw_bioinfo@126.com)

National Clinical Research Center for Child Health, the Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310058, Zhejiang, P.R. China.

Tel: 86-571-8173-2391

Email

Lei Liu: leiliu@tongji.edu.cn

Weili Lin: linweili@tongji.edu.cn

Suqi Cao: suqi_cao@outlook.com

Liu Yang: 13676735203@163.com

Sheng Gao: gaos@tongji.edu.cn

Na Jiao: jiaona@zju.edu.cn

Lixin Zhu: zhulx6@mail.sysu.edu.cn

Ruixin Zhu: rxzhu@tongji.edu.cn

Dingfeng Wu: dfw_bioinfo@126.com

Abstract

Early diagnosis of breast cancer remains a major clinical challenge. Liquid biopsy has become a powerful tool for cancer diagnosis by the aid of various the state-of-the-art detection technologies and artificial intelligence (AI) methods. Although the prediction performance is superior, the clinical application of existing AI models is greatly limited due to their poor interpretability. Here, we designed a miRNA-Gene-Module-Pathway-Disease biological decoding path, and developed BioDecoder thereof, a miRNA bio-interpretable neural network model for breast cancer early diagnosis. We demonstrated that BioDecoder could achieve early non-invasive diagnosis of breast cancer with a remarkable performance (AUC = 0.989) and showed strong generalizability in an external cohort (AUC = 0.890). Meanwhile, the biologically interpretable results of BioDecoder revealed that significant changes in metabolic pathway and oxidative phosphorylation were the main action pathways of circulating miRNA in breast cancer. Our study indicates that BioDecoder offers the promise of non-invasive early diagnosis of breast cancer and can be generalized to other cancers and corresponding biomarkers.

Keywords: breast cancer, circulating miRNA, biological interpretability, noninvasive diagnosis, liquid biopsy

Introduction

Breast cancer is one of the most common malignancies worldwide. According to the data released by GLOBOCAN 2020, female breast cancer has surpassed lung cancer as the most commonly diagnosed cancer with about 2.3 million (11.7%) new cases in 2020, being the leading cause of cancer mortality among women [1]. Published research has revealed that the 5-year average survival rate of in situ female breast cancer reaches 99.0%, while those of regional- and distant-stage breast cancer are only 86.0% and 29.0%, respectively [2]. Lokong et al. [3] also reported that delayed diagnosis was an important reason for the higher breast cancer mortality in low-income countries. Therefore, early screening and diagnosis are essential to improve the overall survival rate of breast cancer. Tissue biopsy is the gold standard for clinical diagnosis of breast cancer; however, as an invasive test it is not suitable for early detection [4]. Currently, mammogram screening has been commonly used for early diagnosis of breast cancer but with risks of overdiagnosis and radiation exposure [5-7]. Hence, it is imperative to develop an accurate and non-invasive alternative tool for the early detection of breast cancer.

Liquid biopsy has become an important means of clinical early screening of cancer[8, 9]. It can detect and analyze circulating tumor DNA (ctDNA), RNA (i.e., mRNA, miRNA), circulating tumor cells (CTC), and exosomes in plasma, urine, and other body fluids, providing information that is difficult to capture in medical imaging[10, 11]. Compared with tissue biopsy, liquid biopsy is non-invasive and easier to monitor tumor oncogenesis, metastasis and treatment response in real time [12, 13]. Although the

diagnosis of breast cancer is challenging due to heterogeneity [14], circulating carcinoma proteins, circulating tumor cells, ctDNA, circulating miRNA, and other biomarkers have been applied in liquid biopsy research of breast cancer and achieved good predictive performance [15, 16]. Among them, the circulating miRNA plays an important role in tumor pathogenesis as oncogenes or tumor suppressors [12, 17], making it a promising biomarker for breast cancer diagnosis. In previous research, using machine learning algorithms, a panel of five miRNAs (miR-1246, miR-1307-3p, miR-4634, miR-6861-5p and miR-6875-5p) was demonstrated to detect breast cancer with 89.7% accuracy [18], and another set of seven miRNAs including has-miR-126-5p and has-miR-144-3p showed predictive power for triple-negative breast cancer with an area under the receiver operating characteristic curve (AUC) of 0.814 [19].

Artificial intelligence (AI), including traditional machine learning algorithms and deep learning architectures, has greatly altered the research paradigm in medical science, and has brought new breakthroughs in precise diagnosis, treatment and prognosis of cancer [20]. Relying on the development of AI, liquid biopsy has become a powerful tool for cancer diagnosis [21]. The inherent black-box nature of most AI models, however, hinders their interpretability and widespread clinical application [22]. To help alleviate this problem, eXplainable AI (XAI) [23] has been introduced. Research have revealed that feature importance, model perturbation, feature association, and prior knowledge, etc., can be utilized to improve the interpretability of AI models [24, 25]. By integrating prior biological knowledge, bio-interpretable models (white-box solution) can be constructed to capture potential causality and uncover the underlying

biological process of diseases with better model credibility and generalizability, thereby promoting the research of disease mechanisms and the identification of therapeutic targets. For example, a recent study by Elmarakeby et al. [26] have demonstrated the capacity of biological XAI model for revealing novel molecularly altered candidates and predicting the staging of prostate cancer patients. Consequently, development of a breast cancer early diagnostic biological XAI model promises great benefits for further popularizing the clinical application of breast cancer liquid biopsy.

This study was undertaken to design a miRNA biological decoding path (miBDP) and develop BioDecoder, a miRNA bio-interpretable neural network model, for breast cancer early screening and diagnosis. Integrating prior biological knowledge and AI technology, BioDecoder dramatically ameliorated its biological interpretation ability under the premise of ensuring prediction performance. The findings drawing from BioDecoder provide new insights into the pathogenesis and treatment of breast cancer.

Results

A set of 4113 serum samples, including 2833 control samples (i.e., 2686 non-cancer samples, 93 prostate disease samples and 54 benign breast disease samples) and 1280 breast cancer samples, and corresponding profiles of 2540 circulating miRNAs were obtained as the discovery cohort (Table S1) [18]. We developed a miRNA bio-interpretable neural network model (BioDecoder) to diagnose breast cancer, whose performance was compared with traditional black-box models (i.e., random forest [RF] and fully connected neural network [FCN]). The potential mechanism of miRNA in

87 breast cancer was then explained through BioDecoder. Finally, the predictive
88 performance of BioDecoder was validated on an external cohort (11 control samples
89 and 122 breast cancer samples, Table S2) by transfer learning (Figure 1A).
90
91 **Differential circulating miRNAs as biomarkers for breast cancer diagnosis**
92 Circulating miRNAs in serum have potential for breast cancer diagnosis [27, 28], which
93 was confirmed in our discovery cohort (Figure 1B). Seven hundred and ten miRNAs
94 with significant differences between breast cancer and control samples were screened
95 out ($|\log_2 FC| > 1$ and FDR < 0.05; FC: fold change), including 704 up-regulated
96 miRNAs and 6 down-regulated miRNAs (Table S3). Among them, has-miR-1246 and
97 has-miR-1307-3p, which were significantly overexpressed in breast cancer patients,
98 have been proven to be potent combined markers for early detection of breast cancer
99 in published studies, with a sensitivity of 97.3%, a specificity of 82.9%, and an
100 accuracy of 89.7% [18, 29].
101 These 710 differential miRNAs were mapped to 11,418 target genes in the
102 miRTarBase database (Table S4). As shown in Figure 1C, the results of Kyoto
103 encyclopedia of genes and genomes (KEGG) pathway enrichment analysis revealed
104 that proteoglycans in cancer, hippo signaling pathway and signaling pathway
105 regulating pluripotency of stem cells, etc. were regulated by differential miRNAs and
106 might participate in the onset and progression of cancer. In particular, these target
107 genes were also significantly enriched in breast cancer pathway (FDR < 0.001; Table
108 S5), which was consistent with the results of gene set enrichment analysis (GSEA)

(enrichment score = 0.758, FDR < 0.001; Figure 1D, Table S6).

BioDecoder enabled precise diagnosis of breast cancer

By leveraging the prior biological knowledge, a miRNA-Gene-Module-Pathway-Disease biological decoding path (miBDP) was extracted from databases to characterize the biological process of miRNAs in the body (Figure 2A). Based upon miBDP, we constructed the miRNA bio-interpretable neural network model (BioDecoder) for breast cancer diagnosis. The 710 differential miRNAs were fed into BioDecoder as input, and then 11,418 targeted genes, 116 modules and 70 pathways from miRTarBase and KEGG were used as hidden layers for information extraction (Table S4), followed by a disease layer that outputs the probability of breast cancer (Figure 2A and Figure S1). For comparison, similar neural network architecture was used in FCN. However, different from FCN, each neuron in BioDecoder represented a specific biological entity, and the links between adjacent layers were partially connected according to the real biological relationship, rather than fully connected (Figure S1). Moreover, in view of the class imbalance issue in the discovery cohort, the synthetic minority oversampling technique (SMOTE) was performed to balance the sample size of the two classes (i.e., control and cancer), thereby improving model stability (Figure 2B, C).

After 100 epochs training, the validation losses were minimized (Figure 2D) and the area under the receiver operating characteristic curves (AUC) was stable at the highest scores (Figure 2E). The results confirmed that RF, FCN and BioDecoder all had

131 excellent prediction performance (AUC > 0.97; Table 1, Figure 2F and G, and Figure
132 S2). Nevertheless, the validation AUC of RF was significantly higher than its test
133 AUC, suggesting an overfitting problem. BioDecoder showed a comparable
134 performance to FCN although it had more restrictions on model architecture (Table 1).
135 BioDecoder with oversampling achieved the best performance for predicting risk of
136 breast cancer on the test set (AUC = 0.989, balanced accuracy = 0.960, precision =
137 0.949, recall = 0.943) and was used for subsequent analysis.
138
139 **BioDecoder revealed the underlying pathological mechanisms of miRNA in**
140 **breast cancer**
141 BioDecoder is a neural network architecture with bio-entity connections between
142 adjacent layers (i.e., miRNA, gene, module and pathway), which can reflect the
143 specific changes of these bio-entities in breast cancer. Ranking the pathways in
144 BioDecoder by weights, it was found that several pathways, such as metabolic
145 pathway, ribosome, oxidative phosphorylation, and DNA replication, were
146 significantly different between the control and cancer samples ($P < 0.001$), and were
147 of prime importance to breast cancer early diagnosis (Figure 3A, B).
148 Specifically, hsa-miR-3659 and hsa-miR-190a-3p had high weights in the metabolic
149 pathway (including 97 modules, 406 genes, and 403 miRNAs, Figure S3A) that has
150 an important influence on breast cancer occurrence [30]. The ribosome pathway is
151 involved in the proliferation and metastasis of breast cancer cells [31-35], in which hsa-
152 miR-17-3p and hsa-miR-3622b-3p were the key factors (Figure S3B). Oxidative

phosphorylation contained 10 energy metabolism modules (e.g., F-type ATPase and V-type ATPase), 84 genes, and 81 miRNAs (Figure 3C). A subset of miRNAs targeting these modules also obtained good diagnostic capabilities for breast cancer. For instance, a set of 38 miRNAs (such as hsa-miR-3146 and hsa-miR-330-3p) in the F-type ATPase module achieved excellent diagnostic performance (AUC = 0.953), while the only miRNA (hsa-miR-3664-3p) in the V-type ATPase module yielded an AUC up to 0.868 (Figure 3C). Interestingly, we found that some target genes of miRNAs could affect the prognosis of breast cancer (Figure 3D, E). Low expression of ATP5F1B ($P < 0.001$) and ATP6AP137 ($P < 0.001$) significantly improved the breast cancer prognosis, and the 5-year survival increased from 70% and 77% to 85%, respectively.

Extended application and validation of BioDecoder

The superior biological interpretability of BioDecoder opens up encouraging prospects in its clinical practice. As presented in Figure 4A, besides significantly distinguishing non-cancer and breast cancer samples (P = 2.666e-224), BioDecoder could also accurately identify other diseases, such as prostate disease (P = 2.302e-9) and benign breast disease (P = 1.539e-12). Meanwhile, although patients with benign breast disease were highly likely to develop cancer at miRNA level, the probability was still significantly lower than that of breast cancer patients (P = 0.040, Figure 4A). It indicated that BioDecoder has the potential for early screening of breast cancer.

BioDecoder contained the biological decoding path of miRNA, and the sample

distribution of each level in miBDP is presented in Figure 4B. At the miRNA level, the model could roughly distinguish between control and breast cancer samples; nevertheless, disease samples were chaotic at principal component analysis (PCA) space. As the decoding proceeded, the distinctions between different categories increased gradually. At the module and pathway levels, there were significant differences among breast cancer, prostate disease and non-cancer samples, while benign breast disease samples were close to breast cancer samples (Figure 4B).

To evaluate the robustness and generalization ability of BioDecoder, an external validation was performed by transfer learning. The external validation cohort included miRNA expression profiles of breast tissue from 122 breast tumor patients and 11 healthy individuals [36]. The first four layers of BioDecoder were frozen, and only the weights of the output layer were updated in transfer learning (Figure 4C). Taking into consideration different sampling proportions (10%–70%) of the external training set, BioDecoder exhibited better generalizability than FCN—BioDecoder achieved an excellent diagnostic performance with only a few external training samples, yielding AUC up to 0.890 (Figure 4D, E).

Discussion

Breast cancer is the most common malignant cancer in women and its early diagnosis can effectively reduce mortality [37]. The accuracy of breast cancer early screening has always been the coalface of research [2, 38]. In this study, we designed a miRNA bio-interpretable neural network model, BioDecoder, for noninvasive diagnosis of breast

197 cancer. Based upon miRNA expression profile of serum sample, BioDecoder achieved
198 superior predictive performance for breast cancer (area under the receiver operating
199 characteristic curve [AUC] = 0.989). In addition, BioDecoder showed strong
200 robustness and clinical generalizability (AUC = 0.890) through transfer learning, even
201 for breast tissue samples.

202 Liquid biopsy has been commonly used in cancer diagnosis due to its high
203 sensitivity and specificity, especially with the aid of artificial intelligence (AI) [39, 40].
204 Such black-box models, however, could hardly integrate into daily clinical practice
205 owing to their poor bio-interpretability [22]. To alleviate this issue, here we developed
206 BioDecoder based on the architecture of miRNA biological decoding path (miBDP,
207 miRNA-Gene-Module-Pathway-Disease) [24, 41]. The bio-interpretable miBDP
208 architecture not merely considerably reduces the number of parameters and enhances
209 modeling efficiency (Figure S1) [26], but also guarantees BioDecoder great benefits in
210 digging into the pathogenesis of breast cancer and discovering potential therapeutic
211 targets (Figure 2, 3). The construction of Biodecoder is a decoding process of miRNA
212 expression information according to miBDP, in which different diseases and their
213 stages can be effectively distinguished (Figure 4B). Besides the excellent predictive
214 power, BioDecoder exhibited outstanding performance in transfer learning (Figure
215 4D), implying that biologically interpretable architecture has an edge in terms of
216 model generalizability and clinical application.

217 Clinically, circulating miRNAs has been proved to be related to the pathogenesis
218 of breast cancer [28, 42], and can be used as biomarkers for breast cancer diagnosis, such

as has-miR-1246 and has-miR-1307-3p [18, 29]. In our results, 710 differential miRNAs were significantly enriched in breast cancer related pathways. Through the biological interpretation of miBDP, metabolism and oxidative phosphorylation were found to be the key pathways for miRNA to regulate the development of breast cancer, which indicates that metabolism may be reprogrammed in breast cancer [27, 43, 44]. Moreover, under the regulation of miRNAs, some genes, such as ATP5F1B and ATP6AP1, could affect the prognosis of breast cancer. Studies have shown that the increased expression of genes in oxidative phosphorylation pathway plays an major role in the immunotherapeutic drug resistance of breast cancer, which could be reversed by the knockdown or inhibition of ATP synthase [45]. Our findings suggests that BioDecoder's interpretability can offer new thoughts for refining clinical diagnosis and precise treatment of breast cancer.

Although BioDecoder uncovered the key pathways for the onset and progression of breast cancer, the mechanism of miRNA targeting these pathways still needs experimental verification. In essence, the interpretability of BioDecoder comes from prior biological knowledge, and therefore detailed biological knowledge (e.g., genetic information, clinical characteristics, and various molecular experimental data) can improve model performance in capturing the real causality. Also, the transferability and predictive performance of similar architecture applied to other biomarkers and diseases need to be further evaluated.

Conclusions

Our study proposed a bio-interpretable neural network architecture, namely BioDecoder, which can accurately diagnose breast cancer and reveal the potential mechanism of miRNA in breast cancer. Based on reliable prior knowledge, this bio-interpretable architecture has great potential to be applied to other types of biomarkers and diseases.

Methods

Discovery cohort

The data used in this work can be acquired from the ArrayExpress database (https://www.ebi.ac.uk/biostudies/arrayexpress). The discovery cohort (E-GEOD-73002, Table S1) consists of 1280 serum samples from breast cancer patients and 2833 serum samples from control samples (i.e., 2686 non-cancer samples, 93 prostate disease samples and 54 benign breast disease samples) [18]. Samples from breast cancer patients with the following characteristics were excluded: (1) given drugs before serum collection and (2) with concurrent or previously diagnosed advanced cancer in other organs. Serum samples of control samples with no history of cancer or hospitalization within the past 3 months were included for analysis. The miRNA expression profiles of all samples were obtained by microarray analysis and verified by quantitative Reverse Transcription-Polymerase Chain Reaction (RT-PCR).

External validation cohort

The external cohort includes 133 Spanish breast tissue samples (i.e., 122 breast cancer

samples and 11 control samples, Table S2), which was reported by Matamala et. al. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58606) [36]. The miRNA expression profiles of all tissue samples were obtained by microarray analysis and verified by quantitative RT-PCR.

Experimental setup

A stratified random sampling was performed to divide the discovery cohort into two subsets: 60% for training set (768 breast cancer samples and 1700 control samples) and 40% for test set (512 breast cancer samples and 1133 control samples). Then the training set was oversampled to balance the number of positive and negative samples using the synthetic minority oversampling technique (SMOTE) algorithm of the *imblearn* package (version 0.9.1) [46].

Construction of artificial intelligence (AI) models

Random Forest (RF) model

The RF model was constructed by the *scikit-learn* (version 0.21.3) package. In the training set, we performed feature selection through recursive feature elimination using cross-validation. Subsequently, a 5-fold cross-validation and grid search were used for model training and hyperparameter tuning. Area under the receiver operating characteristic curve (AUC) was used as the primary evaluation measure for model selection. Finally, the RF model was constructed using the optimal features and hyperparameters (max_features = 0.1, n_estimators = 101, max_depth = None,

max_samples = None, criterion = gini, and class_weight = balanced).

BioDecoder and Fully Connected Neural Network (FCN) models

The neural network models were constructed by the *pytorch* (version 1.13) package [47]. The architecture of neural network consisted of one input layer (miRNA), three hidden layers (gene, module and pathway), and one output layer (disease). The input and hidden layers included linear function, rectified linear unit (ReLU) function, batch normalization (BatchNorm1d) function and dropout function, while the output layer contained only linear and BatchNorm1d functions, followed by the softmax function for classification (Figure S1). To make each layer biologically interpretable, we fixed the number of neurons according to the corresponding miRNA, gene, module and pathway, and links between adjacent layers were partially connected through a mask matrix, which was a boolean matrix representing real biological connections between layers, thereby providing biological meaning for the neurons between each layer. Notably, FCN had the same configuration as BioDecoder, except that the layers of FCN were fully connected and were not biologically meaningful (Figure S1).

The model was trained by Adam optimizer [48] (learning rate = 0.01, batch size = 64, and minimal epoch = 100) with batch gradient descent, and used cross entropy as the loss function. To prevent overfitting, the model was early stopped when the validation loss was minimized. The model was then applied to the test set to assess model performance. Evaluation metrics such as balanced accuracy, precision, recall and AUC were reported.

Assessment of transfer learning robustness

Transfer learning [49, 50] was used to validate the predictive performance of BioDecoder on an external cohort. The first four layers of the BioDecoder were frozen, while the weights of pathway-disease were retrained by external training set (Figure S1). By the stratified random sampling, the external cohort was divided into two unequal parts—that is, the external training set and external test set. The training set was used to tune the transfer learning model (at the sampling proportion from 10% to 70%), and the test set was used to evaluate the model performance.

Statistics Analysis

All statistical analysis was performed using *R* software (version 4.2.1) or *Python* software (version 3.9.6). Statistical significance was assessed using the Wilcoxon signed-rank test, unless otherwise specified. The differentially expressed miRNAs between breast cancer samples and control samples were established using a linear regression model in the R package *limma* [51]. The resulting *P* values were corrected using the Benjamini-Hochberg (BH) method. The biomarkers that were differentially expressed miRNAs were screened by false discovery rate (FDR) < 0.05 and fold change (FC) > 2, or FDR < 0.05 and FC < 0.5. The corresponding target genes of differential miRNAs were obtained from miRTarBase database [52], and module and pathway information were extracted from the Kyoto encyclopedia of genes and genomes (KEGG). Pathway enrichment analysis was performed using the R packages

329 *clusterProfiler* [53] and *GESA* [54]. The PCA method from the python package *scikit-learn*
330 was applied for principal component analysis (PCA). Gene expression data for breast
331 cancer survival analysis were collected from the Human Protein Atlas
332 (https://www.proteinatlas.org/). The network graph was visualized by *Cytoscape*
333 (https://cytoscape.org/, version 3.9.0).

334

335 **Abbreviations**

336 AI, Artificial Intelligence;
337 AUC, Area Under the ROC Curve
338 BH, Benjamini-Hochberg;
339 CTC, Circulating Tumor Cells;
340 ctDNA, circulating tumor DNA;
341 ctRNA, circulating tumor RNA;
342 GSEA, Gene Set Enrichment Analysis
343 FC, Fold Change;
344 FCN, Fully Connected Neural Network;
345 FDR, False Discovery Rate;
346 KEGG, Kyoto Encyclopedia of Genes and Genomes;
347 miBDP, miRNA Biological Decoding Path;
348 OS, Oversampling;
349 RF, random forest;
350 ROC, Receiver Operating Characteristic;
351 RT-PCR, Reverse Transcription-Polymerase Chain Reaction;
352 SMOTE, Synthetic Minority Oversampling Technique;
353 XAI, eXplainable Artificial Intelligence;

354

355 **Key Points**

356 ● Artificial intelligence technology combines prior biological knowledge greatly

improves the model interpretability while ensuring the prediction performance.

- BioDecoder achieved accurate early diagnosis of breast cancer and showed strong robustness and clinical expandability.
- The pathways, such as metabolic, ribosome, oxidative phosphorylation and DNA replication, played key roles in the pathogenesis of breast cancer.

Acknowledgements

We are grateful for all the subjects who participated in this study. BioDecoder framework and its biological prior knowledge schematic were created with *BioRender* (https://biorender.com/), and network graph were draw with *cytoscape*.

Authors' contributions

Dingfeng Wu and Ruixin Zhu conceived and designed the project. Each author has contributed significantly to the submitted work. Lei Liu, Suqi Cao and Liu Yang was responsible for the data analysis and drafted the manuscript. Lixin Zhu, Sheng Gao, Weili Lin, Na Jiao, RuixinZhu and Dingfeng Wu revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (32200529 to Dingfeng Wu, 82170542 to Ruixin Zhu, 92251307 to Ruixin Zhu, 82000536 to Na Jiao), and the National Key Research and Development Program of China (2021YFF0703700/2021YFF0703702 to Ruixin Zhu).

Ethics approval and consent to participate

N/A

Consent for publication

Obtained.

Competing interests

The authors declared no potential conflicts of interest in terms of the research, authorship, and/or publication of this article.

Availability of data and materials

No new sequencing data was used in this paper. All the software packages used in this study are open source and publicly available and the code used in this study is available on GitHub at https://github.com/ddhmed/BioDecoder.

Reference

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209-249. doi:10.3322/caac.21660
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. Jan 2022;72(1):7-33. doi:10.3322/caac.21708
3. Lukong KE, Ogunbolude Y, Kamdem JP. Breast cancer in Africa: prevalence, treatment options, herbal medicines, and socioeconomic determinants. *Breast Cancer Res Treat*. Nov 2017;166(2):351-365. doi:10.1007/s10549-017-4408-0
4. Lowry KP, Bissell MCS, Miglioretti DL, et al. Breast Biopsy Recommendations and Breast Cancers Diagnosed during the COVID-19 Pandemic. *Radiology*. May 2022;303(2):287-294. doi:10.1148/radiol.2021211808
5. Ryser MD, Lange J, Inoue LYT, et al. Estimation of Breast Cancer Overdiagnosis in a U.S. Breast Screening Cohort. *Ann Intern Med*. Apr 2022;175(4):471-478. doi:10.7326/m21-3577
6. Cardoso F, Kyriakides S, Ohno S, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Ann Oncol*. Aug 1 2019;30(8):1194-1220. doi:10.1093/annonc/mdz173
7. Jatoi I, Pinsky PF. Breast Cancer Screening Trials: Endpoints and Overdiagnosis. *J Natl Cancer Inst*. Sep 4 2021;113(9):1131-1135. doi:10.1093/jnci/djaa140
8. Liu NN, Jiao N, Tan JC, et al. Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. *Nat Microbiol*. Feb 2022;7(2):238-250. doi:10.1038/s41564-021-01030-7
9. Alix-Panabières C, Pantel K. Liquid Biopsy: From Discovery to Clinical Application. *Cancer Discov*.

Apr 2021;11(4):858-873. doi:10.1158/2159-8290.Cd-20-1311

10. Krebs MG, Malapelle U, André F, et al. Practical Considerations for the Use of Circulating Tumor DNA in the Treatment of Patients With Cancer: A Narrative Review. *JAMA Oncology*. 2022;8(12):1830-1839. doi:10.1001/jamaoncol.2022.4457

11. Yu W, Hurley J, Roberts D, et al. Exosome-based liquid biopsies in cancer: opportunities and challenges. *Ann Oncol*. Apr 2021;32(4):466-477. doi:10.1016/j.annonc.2021.01.074

12. Raza A, Khan AQ, Inchakalody VP, et al. Dynamic liquid biopsy components as predictive and prognostic biomarkers in colorectal cancer. *J Exp Clin Cancer Res*. Mar 15 2022;41(1):99. doi:10.1186/s13046-022-02318-0

13. Nikanjam M, Kato S, Kurzrock R. Liquid biopsy: current technology and clinical applications. *J Hematol Oncol*. Sep 12 2022;15(1):131. doi:10.1186/s13045-022-01351-y

14. Giordano SB, Gradishar W. Breast cancer: updates and advances in 2016. *Curr Opin Obstet Gynecol*. Feb 2017;29(1):12-17. doi:10.1097/GCO.0000000000000343

15. Zhou E, Li Y, Wu F, et al. Circulating extracellular vesicles are effective biomarkers for predicting response to cancer therapy. *EBioMedicine*. May 2021;67:103365. doi:10.1016/j.ebiom.2021.103365

16. Luo H, Wei W, Ye Z, Zheng J, Xu RH. Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. *Trends Mol Med*. May 2021;27(5):482-500. doi:10.1016/j.molmed.2020.12.011

17. Zhang S, Zhou Y, Wang Y, et al. The mechanistic, diagnostic and therapeutic novel nucleic acids for hepatocellular carcinoma emerging in past score years. *Brief Bioinform*. Mar 22 2021;22(2):1860-1883. doi:10.1093/bib/bbaa023

18. Shimomura A, Shiino S, Kawauchi J, et al. Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci*. Mar 2016;107(3):326-34. doi:10.1111/cas.12880

19. Kahraman M, Röske A, Laufer T, et al. MicroRNA in diagnosis and therapy monitoring of early-stage triple-negative breast cancer. *Sci Rep*. Aug 2 2018;8(1):11584. doi:10.1038/s41598-018-29917-2

20. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. Jan 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

21. Cucchiara F, Petrini I, Romei C, et al. Combining liquid biopsy and radiomics for personalized treatment of lung cancer patients. State of the art and new perspectives. *Pharmacol Res*. Jul 2021;169:105643. doi:10.1016/j.phrs.2021.105643

22. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. *Brief Bioinform*. May 20 2021;22(3)doi:10.1093/bib/bbaa177

23. Gunning D, Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*. 2019;40(2):44-58. doi:https://doi.org/10.1609/aimag.v40i2.2850

24. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet*. Oct 3 2022;doi:10.1038/s41576-022-00532-2

25. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel)*. Dec 25 2020;23(1)doi:10.3390/e23010018

26. Elmarakeby HA, Hwang J, Arafeh R, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*. 2021;598(7880):348-352. doi:10.1038/s41586-021-03922-4

27. Kandettu A, Radhakrishnan R, Chakrabarty S, Sriharikrishnaa S, Kabekkodu SP. The emerging role of miRNA clusters in breast cancer progression. *Biochim Biophys Acta Rev Cancer*. Dec 2020;1874(2):188413. doi:10.1016/j.bbcan.2020.188413

28. Soheilifar MH, Masoudi-Khoram N, Madadi S, et al. Angioregulatory microRNAs in breast cancer:

463 Molecular mechanistic basis and implications for therapeutic strategies. *J Adv Res*. Mar 2022;37:235-
464 253. doi:10.1016/j.jare.2021.06.019
465 29. Zhai LY, Li MX, Pan WL, et al. In Situ Detection of Plasma Exosomal MicroRNA-1246 for Breast
466 Cancer Diagnostics by a Au Nanoflare Probe. *ACS Appl Mater Interfaces*. Nov 21 2018;10(46):39478-
467 39486. doi:10.1021/acsami.8b12725
468 30. Wang T, Fahrmann JF, Lee H, et al. JAK/STAT3-Regulated Fatty Acid β-Oxidation Is Critical for
469 Breast Cancer Stem Cell Self-Renewal and Chemoresistance. *Cell Metab*. Jan 9 2018;27(1):136-150.e5.
470 doi:10.1016/j.cmet.2017.11.001
471 31. Ebright RY, Lee S, Wittner BS, et al. Deregulation of ribosomal protein expression and translation
472 promotes breast cancer metastasis. *Science*. Mar 27 2020;367(6485):1468-1473.
473 doi:10.1126/science.aay0939
474 32. Jin J, Qiu S, Wang P, et al. Cardamonin inhibits breast cancer growth by repressing HIF-1α-
475 dependent metabolic reprogramming. *J Exp Clin Cancer Res*. Aug 27 2019;38(1):377.
476 doi:10.1186/s13046-019-1351-4
477 33. Chu W, Zhang X, Qi L, et al. The EZH2-PHACTR2-AS1-Ribosome Axis induces Genomic Instability
478 and Promotes Growth and Metastasis in Breast Cancer. *Cancer Res*. Jul 1 2020;80(13):2737-2750.
479 doi:10.1158/0008-5472.Can-19-3326
480 34. Li X, Wang M, Li S, et al. HIF-1-induced mitochondrial ribosome protein L52: a mechanism for
481 breast cancer cellular adaptation and metastatic initiation in response to hypoxia. *Theranostics*.
482 2021;11(15):7337-7359. doi:10.7150/thno.57804
483 35. Chu W, Zhang X, Qi L, et al. The EZH2–PHACTR2–AS1–Ribosome Axis induces Genomic Instability
484 and Promotes Growth and Metastasis in Breast Cancer. *Cancer Research*. 2020;80(13):2737-2750.
485 doi:10.1158/0008-5472.can-19-3326
486 36. Matamala N, Vargas MT, González-Cámpora R, et al. Tumor MicroRNA Expression Profiling
487 Identifies Circulating MicroRNAs for Early Breast Cancer Detection. *Clinical Chemistry*.
488 2015;61(8):1098-1106. doi:10.1373/clinchem.2015.238691
489 37. Mann RM, Hooley R, Barr RG, Moy L. Novel Approaches to Screening for Breast Cancer. *Radiology*.
490 Nov 2020;297(2):266-285. doi:10.1148/radiol.2020200172
491 38. Alba-Bernal A, Lavado-Valenzuela R, Domínguez-Recio ME, et al. Challenges and achievements of
492 liquid biopsy technologies employed in early breast cancer. *EBioMedicine*. Dec 2020;62:103100.
493 doi:10.1016/j.ebiom.2020.103100
494 39. Li J, Guan X, Fan Z, et al. Non-Invasive Biomarkers for Early Detection of Breast Cancer. *Cancers*
495 *(Basel)*. Sep 27 2020;12(10)doi:10.3390/cancers12102767
496 40. Hamam R, Hamam D, Alsaleh KA, et al. Circulating microRNAs in breast cancer: novel diagnostic
497 and prognostic biomarkers. *Cell Death Dis*. Sep 7 2017;8(9):e3045. doi:10.1038/cddis.2017.440
498 41. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning.
499 2017:arXiv:1702.08608. Accessed February 01, 2017.
500 https://ui.adsabs.harvard.edu/abs/2017arXiv170208608D
501 42. Nassar FJ, Nasr R, Talhouk R. MicroRNAs as biomarkers for early breast cancer diagnosis,
502 prognosis and therapy prediction. *Pharmacol Ther*. Apr 2017;172:34-49.
503 doi:10.1016/j.pharmthera.2016.11.012
504 43. Viale A, Pettazzoni P, Lyssiotis CA, et al. Oncogene ablation-resistant pancreatic cancer cells
505 depend on mitochondrial function. *Nature*. Oct 30 2014;514(7524):628-32. doi:10.1038/nature13611
506 44. Bacci M, Giannoni E, Fearns A, et al. miR-155 Drives Metabolic Reprogramming of ER+ Breast

Cancer Cells Following Long-Term Estrogen Deprivation and Predicts Clinical Response to Aromatase Inhibitors. *Cancer Res*. Mar 15 2016;76(6):1615-26. doi:10.1158/0008-5472.Can-15-2038

45. Gale M, Li Y, Cao J, et al. Acquired Resistance to HER2-Targeted Therapies Creates Vulnerability to ATP Synthase Inhibition. *Cancer Res*. Feb 1 2020;80(3):524-535. doi:10.1158/0008-5472.CAN-18-3985

46. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority over-Sampling Technique. *J Artif Int Res*. 2002;16(1):321–357.

47. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019;

48. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*. 2014;

49. Novakovsky G, Saraswat M, Fornes O, Mostafavi S, Wasserman WW. Biologically relevant transfer learning improves transcription factor binding prediction. *Genome Biology*. 2021;22(1)doi:10.1186/s13059-021-02499-5

50. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun*. Oct 12 2020;11(1):5131. doi:10.1038/s41467-020-18918-3

51. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, eds. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer New York; 2005:397-420.

52. Huang H-Y, Lin Y-C-D, Li J, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic acids research*. 2020;48(D1):D148-D154. doi:10.1093/nar/gkz896

53. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287. doi:10.1089/omi.2011.0118

54. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. Oct 25 2005;102(43):15545-50. doi:10.1073/pnas.0506580102

Table

Table 1. The predictive performance of random forest, fully connected neural network, and BioDecoder.

	Training	Validation		Test			
Model	Loss	Loss	AUC	AUC	Accuracy*	Precision	Recall
Raw data							
RF	NA	NA	0.991	0.977	0.980	0.967	0.969
FCN	0.128	0.352	0.990	0.989	0.964	0.960	0.945
BioDecoder	0.889	0.359	0.988	0.986	0.955	0.948	0.934
Oversampled data							
RF	NA	NA	0.995	0.977	0.979	0.963	0.971
FCN	0.125	0.348	0.989	0.988	0.958	0.951	0.938
BioDecoder	0.246	0.353	0.988	0.989	0.960	0.949	0.943

* Balanced accuracy; RF: random forest; FCN: fully connected neural network; NA: not

538 applicable.

539

540 **Figure Legends**

541 **Figure 1. Differential changes of circulating miRNAs in breast cancer.** (A) The
542 workflow of this study for breast cancer diagnosis based on circulating miRNAs. (B)
543 Principal component analysis (PCA) of miRNA profiles showed different distribution
544 between breast cancer and control samples in the discovery cohort. (C) The top 25
545 KEGG pathways enriched by the target genes of differential miRNAs. (D) Gene set
546 enrichment analysis results of differential miRNA target genes in breast cancer
547 pathways.

548 **Figure 2. The architecture and performance of BioDecoder.** (A) BioDecoder
549 framework. This figure was created with BioRender.com (https://biorender.com/). (B)
550 The distribution of control samples and breast cancer samples in the discovery cohort.
551 (C) The distribution of control samples and breast cancer samples in the discovery
552 cohort after oversampling. (D) The validation loss calculated by cross_entropy
553 function during model training. (E) The AUC scores obtained during model training.
554 (F) AUC of BioDecoder-OS on the test set. (G) Confusion matrix of BioDecoder-OS
555 on the test set. SMOTE: synthetic minority oversampling technique; FCN: fully
556 connected neural network; OS: oversampling; AUC: area under the receiver operating
557 characteristic curve.

558 **Figure 3. Biological interpretation of BioDecoder.** (A) The pathway importance
559 ranked by weights. (B) Boxplot of differential expression between control and breast

cancer samples for the four important pathways (metabolic pathway, ribosome, oxidative phosphorylation, and DNA expression). (C) The biological network of oxidative phosphorylation pathway. Logistic regression was performed using miRNAs in each module and the receiver operating characteristic curves (ROC) were showed. (D). Survival curves of breast cancer patients based on ATP5F1B expression. Survival curves of breast cancer patients based on ATP6AP1 expression. *** , $P <$ 0.001.

Figure 4. Application and validation of BioDecoder. (A) The probability of breast cancer predicted by BioDecoder in non-cancer, prostate disease, benign breast disease and breast cancer samples. The differences between groups were shown. (B) The distribution of test set samples at different miRNA biological decoding path levels of BioDecoder. (C) The flow chart of transfer learning for applying BioDecoder on the external cohort. (D) The transfer learning performance of BioDecoder and fully connected neural network on external cohort with different sampling proportions of the training set. (E) The receiver operating characteristic curve of BioDecoder's transfer learning performance on the full external cohort. OS: oversampling.

Supplementary material

Figure S1. The neural network architecture of fully connected neural network, BioDecoder and transfer learning.

Figure S2. The receiver operating characteristic curve and confusion matrix of random forest, fully connected neural network and Biodecoder in raw data and oversampling data.

Figure S3. The biological network of metabolic pathway and ribosome pathway.

584 Table S1. Discovery cohort (E-GEOD-73002).

585 Table S2. The external validation cohort (GSE58606).

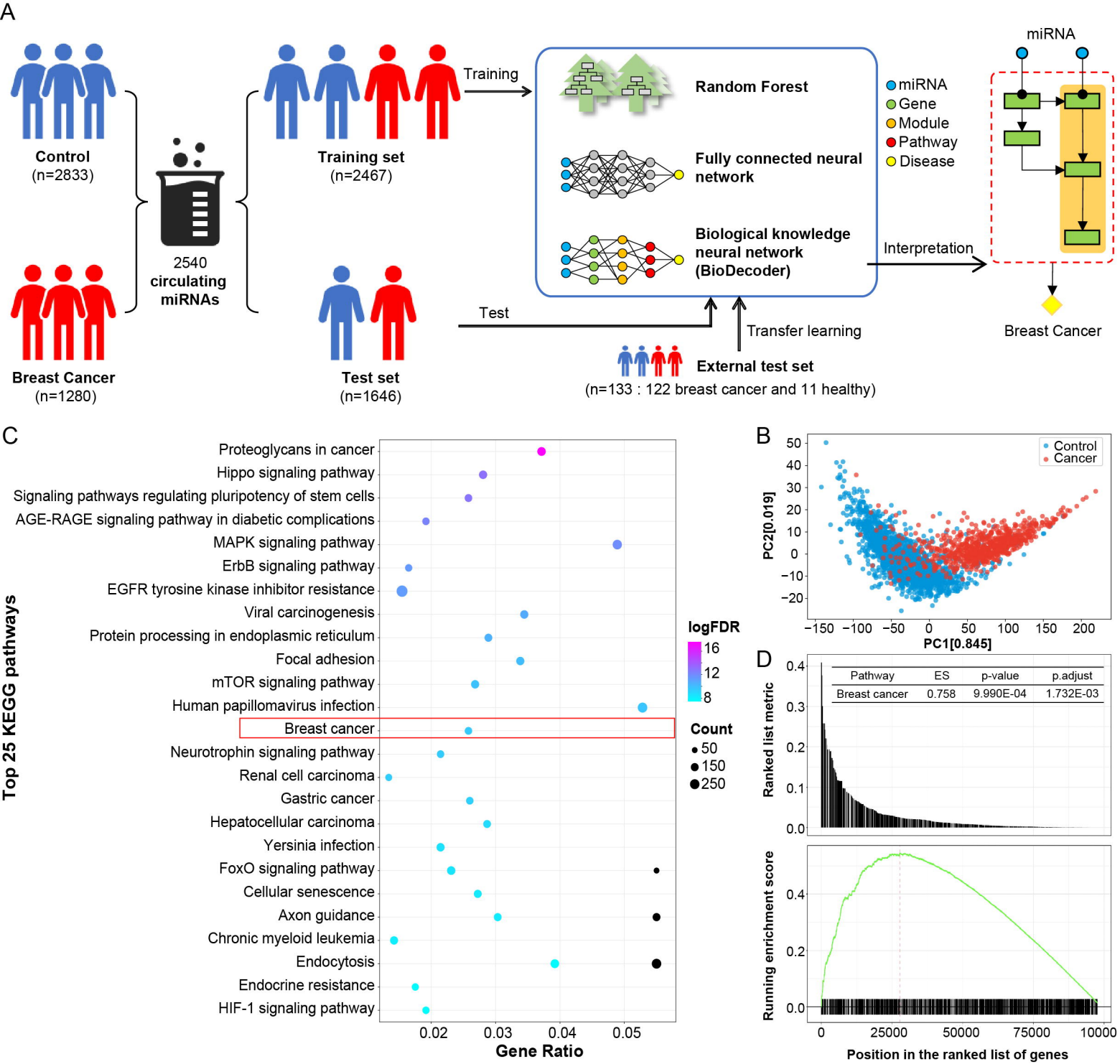
586 Table S3. The 710 miRNAs with significant differences between breast cancer and
587 control samples.

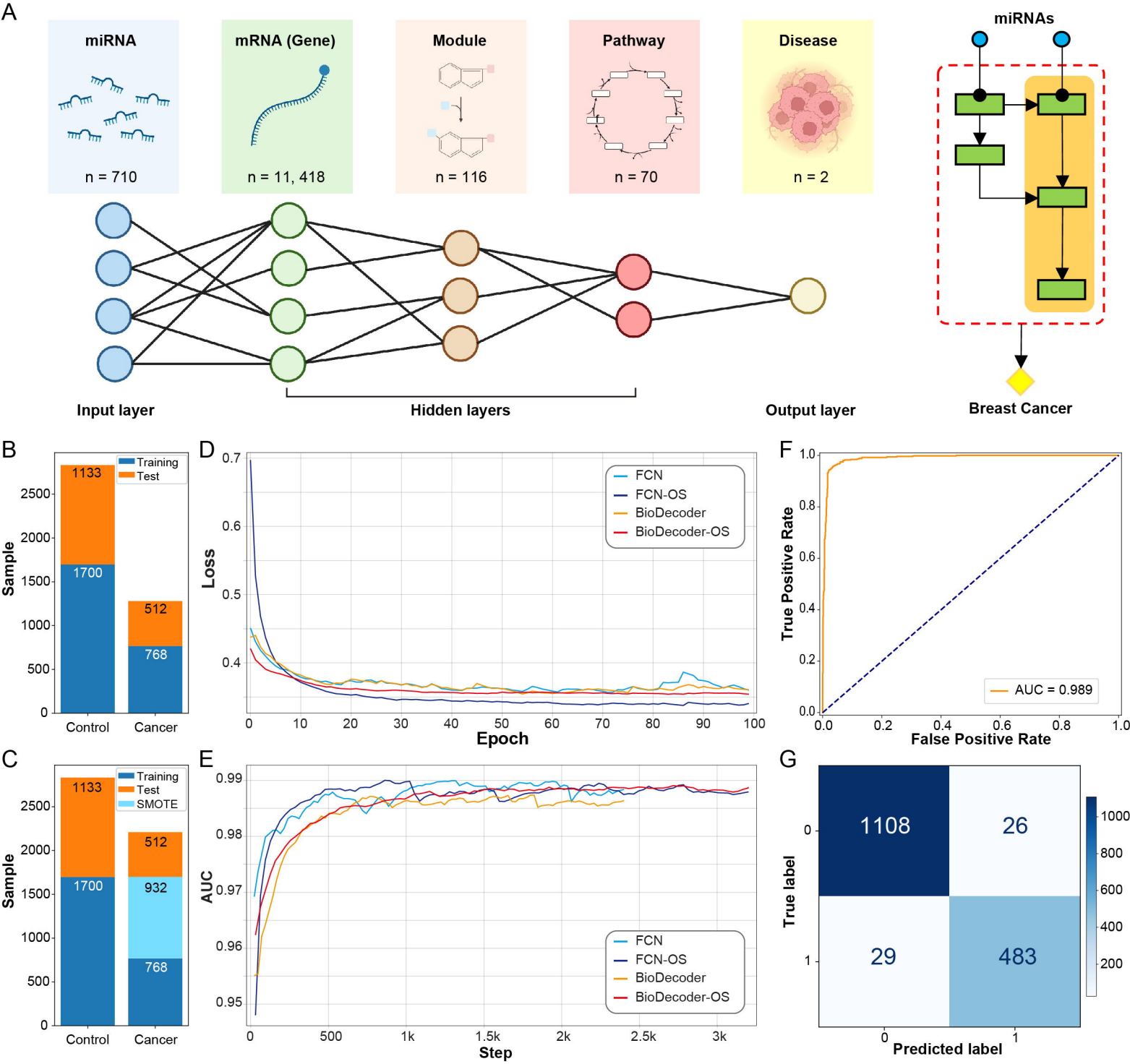
588 Table S4. The correspondence of biological entries in miBDP.

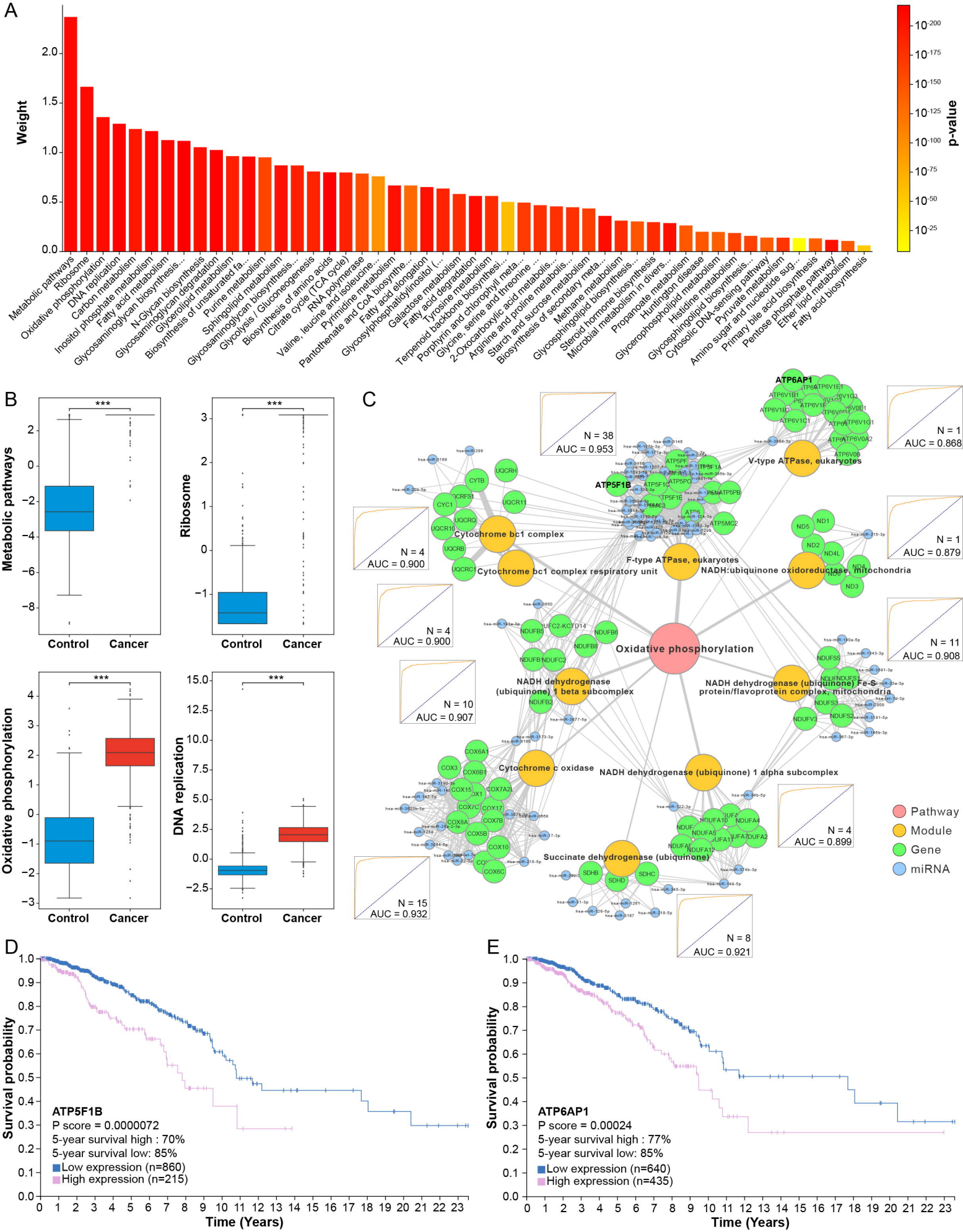
589 Table S5. Pathway enrichment of miRNA targeted genes by clusterProfiler.

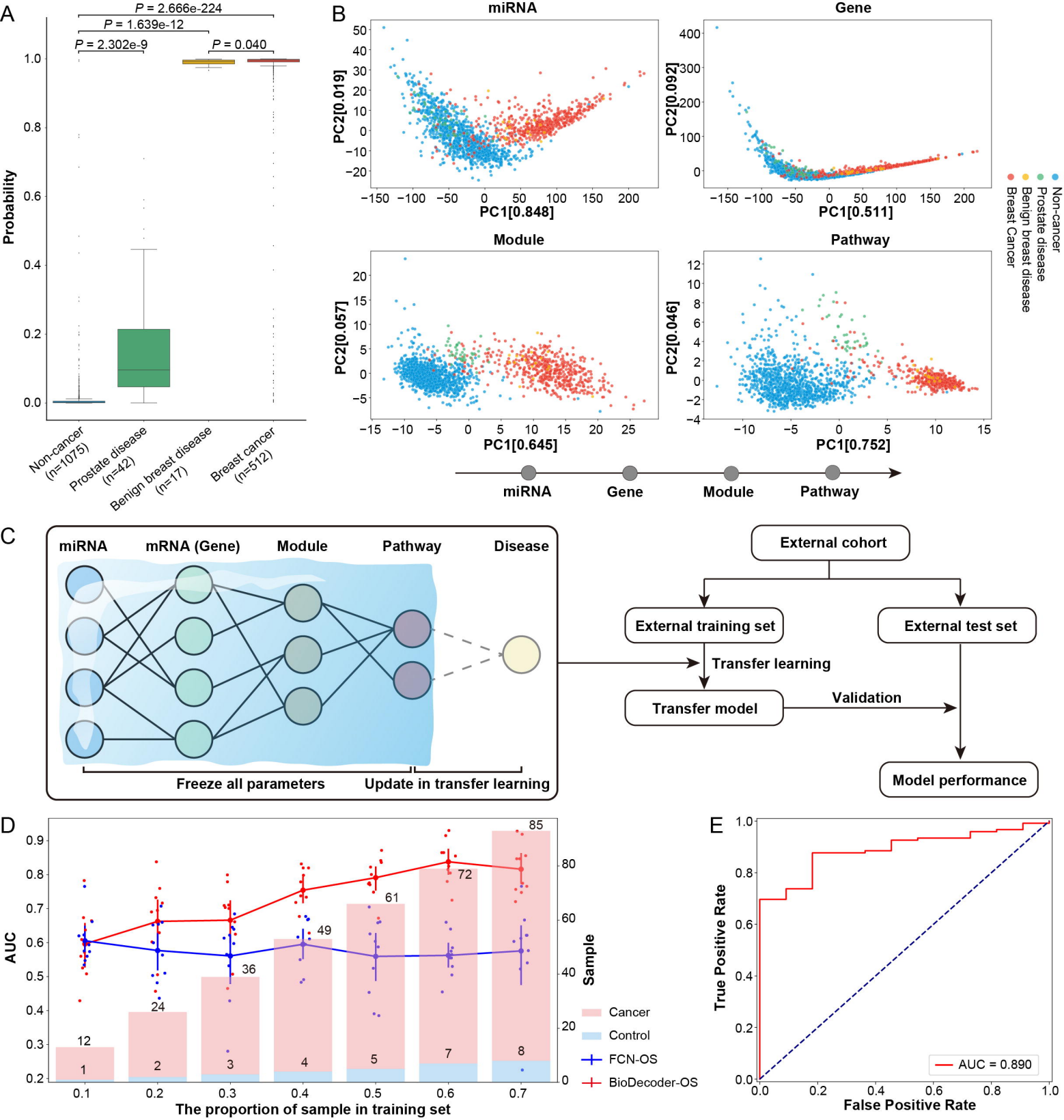
590 Table S6. Pathway enrichment of miRNA targeted genes by GSEA.

591


A
Control
(n=2833)
Breast Cancer
(n=1280)
2540
circulating
miRNAs
Training set
(n=2467)
Test set
(n=1646)
Training
Test
Random Forest
Fully connected neural network
Biological knowledge neural network (BioDecoder)
miRNA
Gene
Module
Pathway
Disease
Interpretation
Transfer learning
External test set
(n=133 : 122 breast cancer and 11 healthy)
Breast Cancer
C
Top 25 KEGG pathways
Proteoglycans in cancer
Hippo signaling pathway
Signaling pathways regulating pluripotency of stem cells
AGE-RAGE signaling pathway in diabetic complications
MAPK signaling pathway
ErbB signaling pathway
EGFR tyrosine kinase inhibitor resistance
Viral carcinogenesis
Protein processing in endoplasmic reticulum
Focal adhesion
mTOR signaling pathway
Human papillomavirus infection
Breast cancer
Neurotrophin signaling pathway
Renal cell carcinoma
Gastric cancer
Hepatocellular carcinoma
Yersinia infection
FoxO signaling pathway
Cellular senescence
Axon guidance
Chronic myeloid leukemia
Endocytosis
Endocrine resistance
HIF-1 signaling pathway
0.02
0.03
0.04
0.05
Gene Ratio
logFDR
16
12
8
Count
50
150
250
B
Control
Cancer
PC2[0.019]
PC1[0.845]
D
Pathway ES p-value p.adjust
Breast cancer 0.758 9.990E-04 1.732E-03
Ranked list metric
Running enrichment score
Position in the ranked list of genes


A
miRNA
n = 710
mRNA (Gene)
n = 11, 418
Module
n = 116
Pathway
n = 70
Disease
n = 2
miRNAs
Input layer
Hidden layers
Output layer
Breast Cancer
B
Training
Test
1133
1700
512
768
Sample
Control
Cancer
C
Training
Test
SMOTE
1133
1700
512
932
768
Sample
Control
Cancer
D
FCN
FCN-OS
BioDecoder
BioDecoder-OS
Loss
Epoch
E
FCN
FCN-OS
BioDecoder
BioDecoder-OS
AUC
Step
F
True Positive Rate
False Positive Rate
AUC = 0.989
G
1108
26
29
483
True label
Predicted label


A
Weight
p-value
Metabolic pathways
Ribosome
Oxidative phosphorylation
DNA replication
Carbon metabolism
Inositol phosphate metabolism
Fatty acid metabolism
Glycosaminoglycan biosynthesis...
N-Glycan biosynthesis
Glycosaminoglycan degradation
Glycerolipid metabolism
Biosynthesis of unsaturated fa...
Purine metabolism
Sphingolipid metabolism
Glycosaminoglycan biosynthesis...
Glycolysis / Gluconeogenesis
Biosynthesis of amino acids
Citrate cycle (TCA cycle)
RNA polymerase
Valine, leucine and isoleucine...
Pyrimidine metabolism
Pantothenate and CoA biosynthe...
Fatty acid elongation
Glycosylphosphatidylinositol (...
Galactose metabolism
Fatty acid degradation
Tyrosine metabolism
Terpenoid backbone biosynthesi...
Porphyrin and chlorophyll meta...
Glycine, serine and threonine ...
2-Oxocarboxylic acid metabolis...
Arginine and proline metabolis...
Starch and sucrose metabolism
Biosynthesis of secondary meta...
Methane metabolism
Glycosphingolipid biosynthesis...
Steroid hormone biosynthesis
Microbial metabolism in divers...
Propanoate metabolism
Huntington disease
Histidine metabolism
Glycerophospholipid metabolism
Glycosphingolipid biosynthesis...
Cytosolic DNA-sensing pathway
Pyruvate metabolism
Amino sugar and nucleotide sug...
Primary bile acid biosynthesis
Pentose phosphate pathway
Ether lipid metabolism
Fatty acid biosynthesis
B
Metabolic pathways
Ribosome
Oxidative phosphorylation
DNA replication
Control
Cancer
***
C
ATP6AP1
V-type ATPase, eukaryotes
ATP5F1B
Cytochrome bc1 complex
Cytochrome bc1 complex respiratory unit
F-type ATPase, eukaryotes
NADH:ubiquinone oxidoreductase, mitochondria
Oxidative phosphorylation
NADH dehydrogenase (ubiquinone) 1 beta subcomplex
NADH dehydrogenase (ubiquinone) Fe-S protein/flavoprotein complex, mitochondria
Cytochrome c oxidase
NADH dehydrogenase (ubiquinone) 1 alpha subcomplex
Succinate dehydrogenase (ubiquinone)
N = 38
AUC = 0.953
N = 1
AUC = 0.868
N = 4
AUC = 0.900
N = 1
AUC = 0.879
N = 4
AUC = 0.900
N = 11
AUC = 0.908
N = 10
AUC = 0.907
N = 4
AUC = 0.899
N = 15
AUC = 0.932
N = 8
AUC = 0.921
Pathway
Module
Gene
miRNA
D
Survival probability
Time (Years)
ATP5F1B
P score = 0.0000072
5-year survival high : 70%
5-year survival low: 85%
Low expression (n=860)
High expression (n=215)
E
Survival probability
Time (Years)
ATP6AP1
P score = 0.00024
5-year survival high : 77%
5-year survival low: 85%
Low expression (n=640)
High expression (n=435)


A
P = 2.666e-224
P = 1.639e-12
P = 2.302e-9
P = 0.040
Probability
Non-cancer (n=1075)
Prostate disease (n=42)
Benign breast disease (n=17)
Breast cancer (n=512)
B
miRNA
Gene
Module
Pathway
PC2[0.019]
PC1[0.848]
PC2[0.092]
PC1[0.511]
PC2[0.057]
PC1[0.645]
PC2[0.046]
PC1[0.752]
Non-cancer
Prostate disease
Benign breast disease
Breast Cancer
miRNA
Gene
Module
Pathway
C
miRNA
mRNA (Gene)
Module
Pathway
Disease
Freeze all parameters
Update in transfer learning
External cohort
External training set
External test set
Transfer learning
Transfer model
Validation
Model performance
D
AUC
Sample
Cancer
Control
FCN-OS
BioDecoder-OS
The proportion of sample in training set
E
True Positive Rate
False Positive Rate
AUC = 0.890