Breast cancer identification via modeling of peripherally circulating miRNAs

Xiaomeng Cui [1,2] , Zhangming Li [3] , Yilei Zhao [4] , Anqi Song [5] , Yunbo Shi [Corresp., 1,2] , Xin Hai [4] , Wenliang Zhu [6]

[1] The higher educational key laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province, Harbin, China

[2] School of Measurement-Control Tech & Communications Engineering, Harbin University of Science and Technology, Harbin, China

[3] Department of Pharmacy, Guangdong Hospital of Integrated Chinese and Western Medicine, Foshan, China

[4] Department of Pharmacy, First Affiliated Hospital of Harbin Medical University, Harbin, China

[5] Department of Student Affairs, Second Affiliated Hospital of Harbin Medical University, Harbin, China

[6] Department of Pharmacy, Second Affiliated Hospital of Harbin Medical University, Harbin, China

Corresponding Author: Yunbo Shi
Email address: shiyunbo@hrbust.edu.cn

Prolonged life expectancy in humans has been accompanied by an increase in the prevalence of cancers. Breast cancer (BC) is the leading cause of cancer-related deaths. It accounts for one fourth of all diagnosed cancers and affects one in eight females worldwide. Given the high BC prevalence, there is a practical need for demographic screening of the disease. In the present study, we re-analyzed a large microRNA (miRNA) expression dataset (GSE73002), with the goal of optimizing miRNA biomarker selection using neural network cascade (NNC) modeling. Our results identified numerous candidate miRNA biomarkers that are technically suitable for BC detection. We combined three miRNAs (miR-1246, miR-6756-5p, and miR-8073) into a single panel to generate a NNC model, which successfully detected BC with 97.1% accuracy in an independent validation cohort comprising 429 BC patients and 895 healthy controls. In contrast, at least seven miRNAs were merged in a multiple linear regression model to obtain equivalent diagnostic performance (96.4% accuracy in the independent validation set). Our findings suggested that suitable modeling can effectively reduce the number of miRNAs required in a biomarker panel without compromising prediction accuracy, thereby, increasing the technical possibility of early detection of BC.

1   **Breast cancer identification via modeling of peripherally circulating miRNAs**

2   Xiaomeng Cui[1,2], Zhangming Li[3], Yilei Zhao[4], Anqi Song[5], Yunbo Shi[1,2,*], Xin Hai[4,*], Wenliang Zhu[6]

3   [1]The higher educational key laboratory for Measuring & Control Technology and Instrumentations of

4   Heilongjiang Province, Harbin, 150080, China

5   [2]School of Measurement-Control Tech & Communications Engineering, Harbin University of Science and

6   Technology, Harbin, 150080, China

7   [3]Department of Pharmacy, Guangdong Hospital of Integrated Chinese and Western Medicine, Foshan,

8   528200, China.

9   [4]Department of Pharmacy, First Affiliated Hospital of Harbin Medical University, Harbin, 150001, China.

10   [5]Department of Student Affairs, Second Affiliated Hospital of Harbin Medical University, Harbin 150086,

11   China.

12   [6]Department of Pharmacy, Second Affiliated Hospital of Harbin Medical University, Harbin 150086,

13   China.

14   *To whom correspondence should be addressed. Yunbo Shi: Address: 52 Xuefu Road, Harbin 150080,

15   China; Tel/Fax: +86-451-8639-0840; E-mail: shiyunbo@hrbust.edu.cn. Xin Hai: Address: 23 Youzheng

16   Road, Harbin 150001, China; Tel/Fax: +86-451-8555-5345; E-mail: hai_xin@163.com.

ABSTRACT

Prolonged life expectancy in humans has been accompanied by an increase in the prevalence of cancers.

Breast cancer (BC) is the leading cause of cancer-related deaths. It accounts for one fourth of all diagnosed

cancers and affects one in eight females worldwide. Given the high BC prevalence, there is a practical need

for demographic screening of the disease. In the present study, we re-analyzed a large microRNA (miRNA)

expression dataset (GSE73002), with the goal of optimizing miRNA biomarker selection using neural

network cascade (NNC) modeling. Our results identified numerous candidate miRNA biomarkers that are

technically suitable for BC detection. We combined three miRNAs (miR-1246, miR-6756-5p, and miR-

8073) into a single panel to generate a NNC model, which successfully detected BC with 97.1% accuracy

in an independent validation cohort comprising 429 BC patients and 895 healthy controls. In contrast, at

least seven miRNAs were merged in a multiple linear regression model to obtain equivalent diagnostic

performance (96.4% accuracy in the independent validation set). Our findings suggested that suitable

modeling can effectively reduce the number of miRNAs required in a biomarker panel without

compromising prediction accuracy, thereby, increasing the technical possibility of early detection of BC.

INTRODUCTION

Breast cancer (BC) is one of the most common cancers that accounts for one in four diagnosed cancers and affects one in eight females worldwide (Torre et al., 2015). Approximately 1.5 million new BC cases are reported per year (Siegel et al., 2015), which is close to the existing 1.7 million BC cases reported in 2012. Conservative estimates suggested higher morbidity rates associated with BC though only prolonged life expectancy of females was considered. Therefore, early demographic screening is necessary to manage the unprecedented increase in the malignant disease (Myers et al., 2015). However, currently employed BC screening methods have relatively low sensitivity and insufficient identification power, leading to a high false positive rate of 20.5% in women aged 40–49 years (van den Ende et al., 2017). Therefore, there is a need for the development of novel biomarkers for early detection of BC.

MicroRNAs (miRNAs) are a class of single-stranded small non-coding RNA molecules of ~22 nucleotides. miRNAs act as post-transcriptional gene expression regulators via complementary binding to the 3′ -untranslated regions of mRNAs (Bartel, 2009). The miRBase database currently includes data on more than 2800 mature human miRNAs (Kozomara and Griffiths-Jones, 2014). Of these, some miRNAs, such as miR-21 and miR-155, have demonstrated potential value for the early diagnosis of BC (Hamam et al., 2017). Despite significant progress in research on the use of circulating miRNAs as diagnostic BC biomarkers, one major limitation is that most studies have small sample sizes, which results in poor inter-study reproducibility (Nassar et al., 2017). Thus, there is a need for a systematic review of candidate biomarkers reported in previous clinical studies. BC is considered a collection of mammary gland-related heterogeneous diseases (Bertos and Park, 2011). In addition, the high BC prevalence requires large sample sizes so that multiple types of BC can be investigated in a single circulating miRNA biomarker study. So

54   far, only one study has met this requirement. In a study comprising approximately 4000 patients and healthy

55   subjects, Shimomura and his colleagues performed a microarray-based circulating miRNA biomarker assay

56   for early detection of BC in the Japanese population (Shimomura et al., 2016). The authors validated the

57   effectiveness of a biomarker panel comprising five miRNAs (miR-1246, miR-1307-3p, miR-4634, miR-

58   6861-5p, and miR-6875-5p) for BC diagnosis with 89.7% accuracy. Surprisingly, the aforementioned five

59   miRNAs were not reported by other studies with small sample sizes (Nassar et al., 2017). Therefore, larger

60   sample sizes can facilitate the discovery of miRNA biomarkers, while smaller sample sizes can introduce

61   more sampling error and inconsistencies in miRNA biomarkers among different studies.

62   Although the authors provided a valuable data resource for expression levels of circulating miRNAs in

63   BC (GSE73002), no optimization was performed for the miRNA biomarker panel, which could potentially

64   increase diagnostic accuracy. The neural network cascade (NNC) modeling has been demonstrated to have

65   high prediction accuracy than the traditional artificial neural network (ANN) modeling (Li et al., 2015; Hou

66   et al., 2016; Qu et al., 2017). In this study, NNC models were generated to re-analyze the dataset comprising

67   circulating miRNAs in BC and to optimize the miRNA biomarker panel for early detection of BC. Our

68   approach was effective in identifying suitable diagnostic biomarkers for demographic screening for BC.

69

70   **MATERIALS AND METHODS**

71   **miRNA expression data**

72   Expression data of circulating miRNAs in BC patients and healthy controls (GSE73002) were retrieved

73   from the Gene Expression Omnibus (GEO) repository (Shimomura et al., 2016). The present study included

74   data from a total of 3974 participants, including 1288 BC patients and 2686 healthy controls. For each

75   participant, the normalized microarray expression values of 2540 miRNAs were downloaded from GEO,

76   and the corresponding disease statuses (healthy: 0 or BC: 1) were obtained for further modeling.

77   **Biomarker evaluation and selection**

78   Data from the 3974 participants were randomly divided into two sets, namely, a training set (n =2650) and

79   a validation set (n =1324). Later, each of the 2540 miRNAs in the training set was independently evaluated

80   as a potential miRNA biomarker for BC. We performed receiver operating characteristic (ROC) curve

81   analysis of the miRNA expression values using MedCalc version 15.8 (MedCalc, Mariakerke, Belgium).

82   We calculated the area under the curve (AUC) to evaluate whether the generated models can reliably

83   distinguish between BC and healthy controls (AUC >0.95 and p <0.0001). Since all expression values

84   corresponding to miRNA models with AUC ≥0.95 did not satisfy the D'Agostino-Pearson omnibus

85   normality test or follow a Gaussian distribution, we conducted Spearman's correlation tests to determine

86   the collinearities among the miRNAs using GraphPad Prism version 6.0 (GraphPad Software, Inc., La Jolla,

87   CA, USA). Collinearity was considered significant at the threshold of $|\rho| > 0.5$. If the expression values of

88   two miRNAs are collinear ($|\rho| > 0.5$), only the miRNA with a larger AUC was retained for further modeling.

89   A collinearity network of the miRNAs with AUC ≥0.95 was generated using the network visualization

90   software Cytoscape v3.6.0 (Institute of Systems Biology, Seattle, WA, USA) (Shannon et al., 2003).

91   **NNC and multiple linear regression (MLR) models**

92   The expression values of the miRNAs with AUC ≥0.95 were normalized to a value between 0 and 1 before

93   further model building, as previously described (Zhu and Kan, 2014). The Intelligent Problem Solver (IPS)

94   tool in the Statistica Neural Networks (SNN, Release 4.0E; Statsoft, Tulsa, OK, USA) software was used

95   to build a radial basis function (RBF)-ANN model with 1-11-1 network architecture to investigate the

96   associations of individual miRNAs with the disease status. Afterwards, an NNC model was built following

97   a step-by-step procedure as previously described (Li et al., 2015). For comparison, a MLR model was also

98   built based on the miRNAs with AUC ≥0.95 using the SPSS statistical software version 19.0 (IBM Corp.,

99   New York city, NY, USA).

100   **Model validation**

101   A hold-out cross-validation method was used for internal validation of each of the RBF-ANNs in the NNC

102   model. Briefly, IPS divided the modeling set into three subsets (training subset, verification subset, and

103   testing subset) at a 2:1:1 ratio. Data on participants included in the testing subset were not used for model

104   building but were used for model validation. The correlation coefficients given by IPS were compared to

105   those from the training subset (R_{Tr}); the testing subset (R_{Te}) measured the linear relationship between the

106   model output values and the normalized miRNA expression values. Similar R_{Te} and R_{Tr} values indicated

107   good generalizability of the corresponding RBF-ANN. Furthermore, a tenfold cross-validation method was

108   used to validate the NNC model. The entire training set (n =2650) was randomly divided into ten mutually

109   exclusive groups of nearly equal size. Nine of the groups were selected for model training, while the

110   remaining group was used for model validation. The above procedure was repeated ten times, as previously

111   described (Li et al., 2015). Furthermore, an independent validation set (n =1324) was used for external

112   validation of the NNC and MLR models. Three parameters, namely, sensitivity, specificity, and accuracy

113   rate, were used for model evaluation and validation. Sensitivity was calculated as the percentage of the

114   number of true positives divided by the sum of true positives and false negatives. Specificity was calculated

115   as the percentage of number of true negatives divided by the sum of true negatives and false positives.

116   Accuracy rate (ACC) was calculated as the number of successfully identified BC patients and healthy

117   controls divided by the sum of all the participants.

118   **Data statistics**

119   Spearman's correlation test was conducted using Graphpad Prism v6.0. ROC curve analysis was performed

120   using MedCalc v13.0. Statistically significant differences were considered at $p < 0.0001$.

121

122   **RESULTS**

123   **High AUCs revealed the redundancy of technically suitable miRNA biomarkers for BC**

124   In the present study, data of 3974 participants were obtained from the GSE73002 dataset. All participants

125   were randomly assigned into two sets, namely, the training set and validation set, at a 2:1 ratio. The training

126   set consisted of 859 BC patients and 1791 healthy controls. The validation set comprised 429 BC patients

127   and 895 healthy controls. We then investigated the technical feasibility of each of the 2540 miRNAs for

128   BC identification in each of the training sets. Figure 1A shows the frequency distribution of the AUC values

129   calculated from ROC curve analyses. Approximately 74% of all miRNAs showed high AUC values (AUC

130   >0.9), which indicated the strong reliability of the generated models for BC detection. A total of 82 miRNAs

131   with high AUC values are highlighted (AUC >0.95). Furthermore, we observed consistent collinearity

132   among the 82 miRNAs (Figure 1B), which implied very high redundancy of candidate miRNA biomarkers

133   used for BC detection. Finally, we identified eight non-collinear miRNAs that satisfied AUC >0.95. These

134   miRNAs are listed in Table 1. Compared with the seven other miRNAs, miR-8073 showed the highest

135   AUC value (AUC=0.991) and the highest accuracy for identifying BC in the training set (ACC=97.0%).

136   **NNC model integrating three miRNAs for BC detection**

137   An NNC model was built to generate a miRNA biomarker panel for BC diagnosis using the eight miRNAs

138 listed in Table 1. Finally, three miRNAs, namely, miR-1246, miR-6756-5p, and miR-8073, were used to

139 effectively extend the cascade (Figure 2A). The NCC consisted of three 1-11-1 RBF-ANN units and two

140 2-11-1 RBF-ANN units. Each of the five RBF-ANNs showed similar R_{Te} and R_{Tr} values. Furthermore, an

141 MLC model was built by considering the same eight miRNAs as candidate model inputs. Except for miR-

142 6748-5p, seven miRNAs were automatically selected into an MLR model using the SPSS software.

143 Although only three miRNAs were included in the NNC model and four additional miRNAs were used in

144 the MLR model, significant differences were not observed between the NNC and MLR models. Table 2

145 lists the core evaluation parameters for the two models. Although both models had the same AUC, the NNC

146 model showed better performance for BC identification. The accuracy rate of the NNC model was 98.5%,

147 while that of the MLR model was 97.4%.

148 **NNC successfully identified BC in the validation set**

149 To validate the effectiveness of the NNC model for BC identification, we performed a tenfold cross-

150 validation. The NNC model had an AUC of 0.995, which demonstrated its effectiveness for BC detection

151 (Figure 3A). An independent validation set consisting of data from 1324 participants was used to further

152 validate the NNC model. The NNC model was found to have an AUC similar with those of the single

153 miRNA (miR-8073) model and the MLR model (Figure 3B). However, the NNC model actually showed

154 the highest accuracy for BC identification (ACC=97.1%; Figure 3C). The sensitivity and specificity of the

155 NNC model were 96.7% and 97.2%, respectively.

156

157 **DISCUSSION**

158 The GSE73002 breast cancer (BC) dataset, comprising data from a Japanese population, is the largest

159   miRNA dataset published in GEO. China has also witnessed an increase in the number of BC cases in recent

160   years (Jiang et al., 2018). BC has become the most prevalent malignant disease in Chinese females, with

161   nearly 270,000 new BC cases reported in 2015. Consequently, BC has become a serious and widespread

162   social issue that cannot be solved by treatment alone. Early detection of BC in the population represents an

163   optimal strategy for improving the survival rates of BC patients (Sun et al., 2017).

164   Considerable evidence has demonstrated the technical reliability of miRNAs as early diagnostic markers

165   for BC because of their relatively simple molecular structure and stability (Bahrami et al., 2018). However,

166   multiple studies have not agreed upon a consensus set of miRNAs that are useful as biomarkers, which

167   could be attributed to inadequate sample sizes in the majority of studies (Nassar et al., 2017). The above

168   findings indicated that cohorts with small samples sizes within the range of dozens to hundreds are of little

169   or no value for the identification of potential miRNAs as early diagnostic biomarkers. Therefore, simply

170   counting the number of times a miRNA was validated as a BC biomarker in different studies is not reliable.

171   The heterogeneity of different BC subtypes is a major consideration for the initial research design (Yeo and

172   Guan, 2017). The sample size is a crucial design parameter for a clinical study. Inadequate samples do not

173   fully represent the whole population (Freiman et al., 1978). However, large sample sizes are difficult to

174   obtain because of certain criteria that limit the number of available samples. For example, research funding

175   may not be sufficient to support a large sample size, and different research groups may have limited staff

176   for implementation of the research protocols. Moreover, research groups are usually relatively isolated from

177   each other and most studies tend to address problems in the regions in which the studies were performed.

178   Re-analysis of the GSE73002 dataset facilitated the selection and optimization of biomarkers from the

179   human miRNAome, a sample set that is most representative of the population. One of our main findings is

180 that circulating miRNAs serve as highly useful markers for BC detection. A single miRNA can be used as

181 a biomarker without the need for data reprocessing. For example, our models achieved close to 96%

182 accuracy in the independent validation set using miR-8073 as the biomarker. Our results were consistent

183 with those of previously reported miRNA biomarkers (Nassar et al., 2017).

184   In addition, our findings showed that appropriate data modeling is necessary for optimization of miRNA

185 biomarkers. The currently developed NNC model showed significantly higher accuracy ranging from

186 95.8% to 97.1% after integration of single miRNA, miR-8073, with the two miRNAs miR-1246 and miR-

187 6756-5p. However, six additional miRNAs were needed to achieve the same accuracy rate when using the

188 MLR model. The MLR model is a widely used mathematical model that can be used to construct miRNA

189 biomarker panels for the detection of various human diseases (Ding et al., 2017). The NNC is a tandem

190 mode of multiple small ANNs that generates a gradual gain in target information. In the present study, we

191 confirmed that the NNC models achieved higher prediction accuracy with a lower number of input

192 biomarkers than traditional modeling methods, including MLR and ANN (Li et al., 2015; Hou et al., 2016;

193 Qu et al., 2017).

194

195 **CONCLUSION**

196 In conclusion, we constructed and validated a NNC-based biomarker panel comprising three miRNAs for

197 early detection of BC. The models were generated using data from a miRNA microarray database

198 comprising nearly 4000 participants. Compared with a single miRNA biomarker, the NNC-based miRNA

199 biomarker panel showed significantly higher accuracy. Our developed biomarker panel is superior to the

200 MLR-based miRNA biomarker panel because the same detection accuracy was obtained using a smaller

201   number of input miRNAs. However, given that the dataset used for constructing the biomarker panel was

202   derived from a Japanese population, further studies are required to confirm the generalizability of the

203   developed panel to other Asian populations, such as the Chinese population. Future research studies should

204   be conducted using a large cohort of female participants.

205

206   **REFERENCES**

207   **Bahrami A, Aledavood A, Anvari K, Hassanian SM, Maftouh M, Yaghobzade A, Salarzaee O,**

208   **ShahidSales S, Avan A. 2018.** The prognostic and therapeutic application of microRNAs in breast cancer:

209   Tissue and circulating microRNAs. *Journal of Cellular Physiology* **233(2):**774-786 DOI

210   10.1002/jcp.25813.

211   **Bartel DP. 2009.** MicroRNAs: target recognition and regulatory functions. *Cell* **136(2):**215-233 DOI

212   10.1016/j.cell.2009.01.002.

213   **Bertos NR, Park M. 2011.** Breast cancer - one term, many entities? *Journal of Clinical Investigation*

214   **121(10):**3789-3796 DOI 10.1172/JCI57100.

215   **Ding H, Meng J, Zhang W, Li Z, Li W, Zhang M, Fan Y, Wang Q, Zhang Y, Jiang L, Zhu W. 2017.**

216   Medical examination powers miR-194-5p as a biomarker for postmenopausal osteoporosis. *Scientific*

217   *Reports* **7(1):**16726 DOI 10.1038/s41598-017-17075-w.

218   **Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. 1978.** The importance of beta, the type II error

219   and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative"

220   trials. *The New England Journal of Medicine* **299(13):**690-694 DOI 10.1056/NEJM197809282991304.

221   **Hamam R, Hamam D, Alsaleh KA, Kassem M, Zaher W, Alfayez M, Aldahmash A, Alajez NM. 2017.**

222 Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death &*

223 *Disease* **8(9):**e3045 DOI 10.1038/cddis.2017.440.

224 **Hou S, Wang J, Li Z, Wang Y, Wang Y, Yang S, Xu J, Zhu W. 2016.** Five-descriptor model to predict

225 the chromatographic sequence of natural compounds. *Journal of Separation Science* **39(5):**864-872 DOI

226 10.1002/jssc.201501016.

227 **Jiang X, Tang H, Chen T. 2018.** Epidemiology of gynecologic cancers in China. *Journal of Gynecologic*

228 *Oncology* **29(1):**e7 DOI 10.3802/jgo.2018.29.e7.

229 **Kozomara A, Griffiths-Jones S. 2014.** miRBase: annotating high confidence microRNAs using deep

230 sequencing data. *Nucleic Acids Research* **42(Database issue):**D68-D73 DOI 10.1093/nar/gkt1181.

231 **Li Z, Li Y, Sun L, Tang Y, Liu L, Zhu W. 2015.** Artificial neural network cascade identifies multi-P450

232 inhibitors in natural compounds. *PeerJ* **3:**e1524 DOI 10.7717/peerj.1524.

233 **Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, Davidson B, Mongtomery**

234 **RC, Crowley MJ, McCrory DC, Kendrick A, Sanders GD. 2**015. Benefits and Harms of Breast Cancer

235 Screening: A Systematic Review. *JAMA* 314(15):1615-1634. DOI 10.1001/jama.2015.13183.

236 **Nassar FJ, Nasr R, Talhouk R. 2017.** MicroRNAs as biomarkers for early breast cancer diagnosis,

237 prognosis and therapy prediction. *Pharmacology & Therapeutics* **172:**34-49 DOI

238 10.1016/j.pharmthera.2016.11.012.

239 **Qu Y, He Y, Li Z, Chen X, Liu Q, Zou S, Kong C, Liu Y, Gao C, Zhang G, Zhu W. 2017.** Constructing

240 an ovarian cancer metastasis index by dissecting medical records. Oncotarget Advance Publications DOI

241 10.18632/oncotarget.22336.

242 **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker**

243   **T. 2003.** Cytoscape: a software environment for integrated models of biomolecular interaction networks.

244   *Genome Research* **13(11):**2498-2504 DOI 10.1101/gr.1239303.

245   **Shimomura A, Shiino S, Kawauchi J, Takizawa S, Sakamoto H, Matsuzaki J, Ono M, Takeshita F,**

246   **Niida S, Shimizu C, Fujiwara Y, Kinoshita T, Tamura K, Ochiya T. 2016.** Novel combination of serum

247   microRNA for detecting breast cancer in the early stage. *Cancer Science* **107(3):**326-334 DOI

248   10.1111/cas.12880.

249   **Siegel RL, Miller KD, Jemal A. 2015.** Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians*

250   **65(1):**5-29 DOI 10.3322/caac.21254.

251   **Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, Shi W, Jiang J, Yao PP, Zhu HP. 2017.** Risk Factors

252   and Preventions of Breast Cancer. *International Journal of Biological Sciences* **13(11):**1387-1397 DOI

253   10.7150/ijbs.21635.

254   **Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. 2012.** Global cancer statistics, 2012.

255   *CA: A Cancer Journal for Clinicians* **65(2):**87-108 DOI 10.3322/caac.21262.

256   **van den Ende C, Oordt-Speets AM, Vroling H, van Agt HME. 2017.** Benefits and harms of breast cancer

257   screening with mammography in women aged 40-49 years: A systematic review. *International Journal of*

258   *Cancer* **141(7):**1295-1306 DOI 10.1002/ijc.30794.

259   **Yeo SK, Guan JL. 2017.** Breast Cancer: Multiple Subtypes within a Tumor? *Trends in Cancer* **3(11):**753-

260   760 DOI 10.1016/j.trecan.2017.09.001.

261   **Zhu W, Kan X. 2014.** Neural network cascade optimizes microRNA biomarker selection for

262   nasopharyngeal cancer prognosis. *PLoS One* **9(10):**e110537 DOI 10.1371/journal.pone.0110537.

263

264 **FIGURE LEGENDS**

265 **Figure 1. AUC distribution and collinearity of miRNA expression. A.** Frequency distribution of AUCs.

266 **B.** Collinearity network of the 82 miRNAs with AUC ≥0.95. An edge represents collinear expression

267 between the two miRNAs ($\rho^2 > 0.5$).

268 **Figure 2. Establishment of the NNC model. A.** Illustration of the NNC model. L1-L3: Layers 1- 3 of the

269 NNC model; R1246: miR-1246; R6756-5p: miR-6756-5p; R8073: miR-8073; AUC values are shown above

270 the layers. **B.** ROC curve diagrams of the NNC and MLR models (training set).

271 **Figure 3. Model validation. A.** ROC curve diagram of the tenfold cross-validation of the NNC model. **B.**

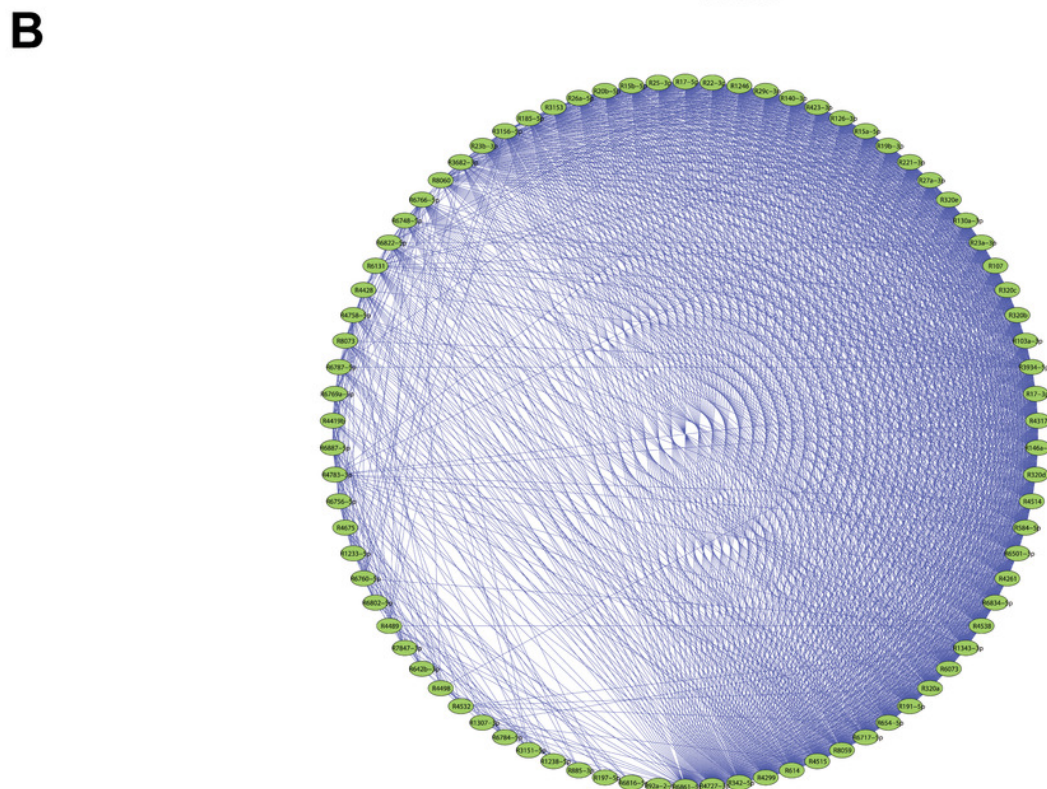
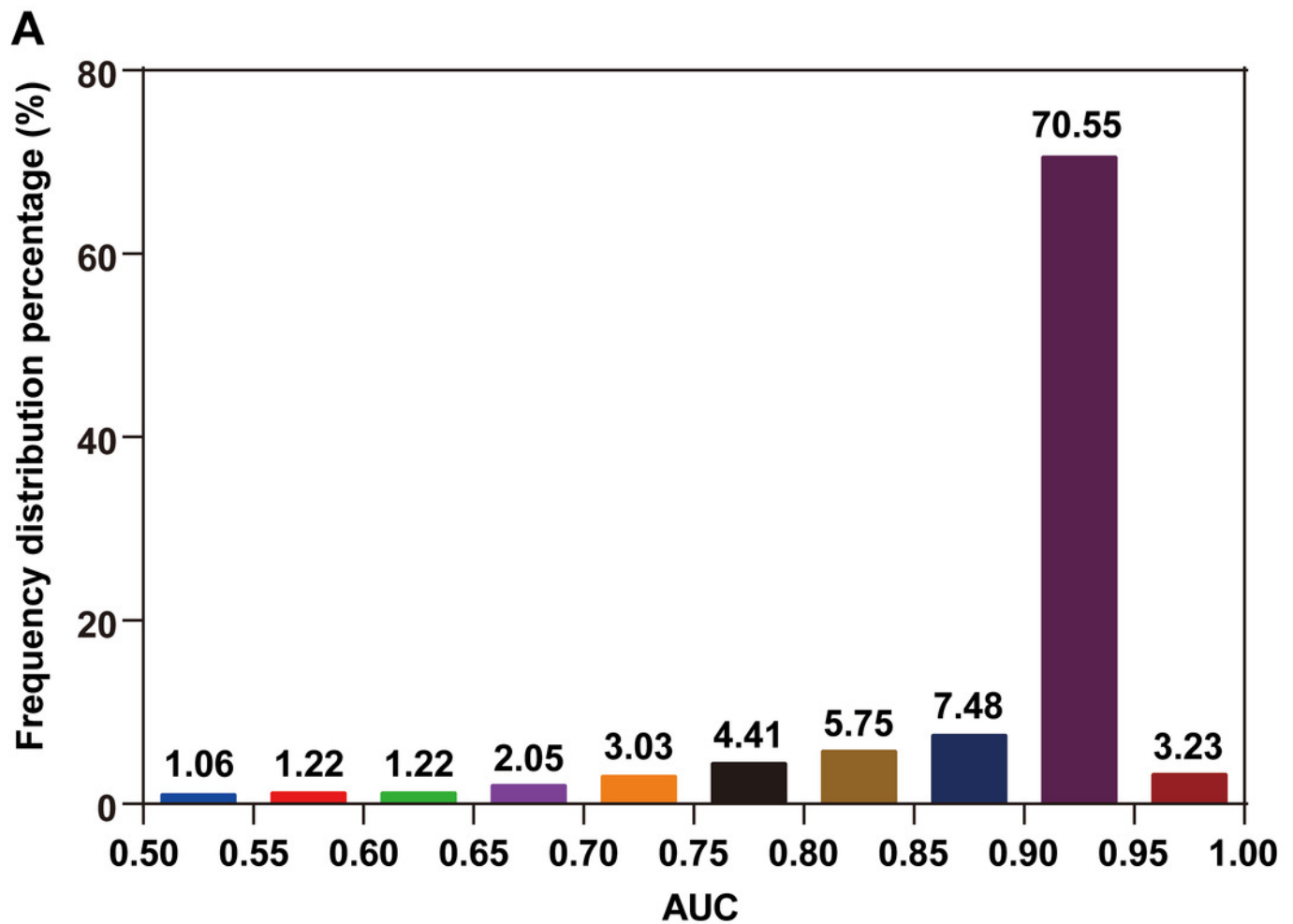
272 ROC curve diagram of the NNC and MLR models (validation set). 10FCV: Tenfold cross-validation. **C.**

273 Accuracy evaluation of miR-8073, MLR, and NNC BC detection (validation set). R8073: miR-8073. ACC:

274 accuracy rate; Se: sensitivity; Sp: specificity.

Figure 1

AUC distribution and collinearity of miRNA expression

A. Frequency distribution of AUCs. **B.** Collinearity network of the 82 miRNAs with AUC ≥0.95. An edge represents collinear expression between the two miRNAs ($\rho^2 > 0.5$).

A

B

Figure 2

Establishment of the NNC model

A. Illustration of the NNC model. L1-L3: Layers 1- 3 of the NNC model; R1246: miR-1246; R6756-5p: miR-6756-5p; R8073: miR-8073; AUC values are shown above the layers. **B.** ROC curve diagrams of the NNC and MLR models (training set).

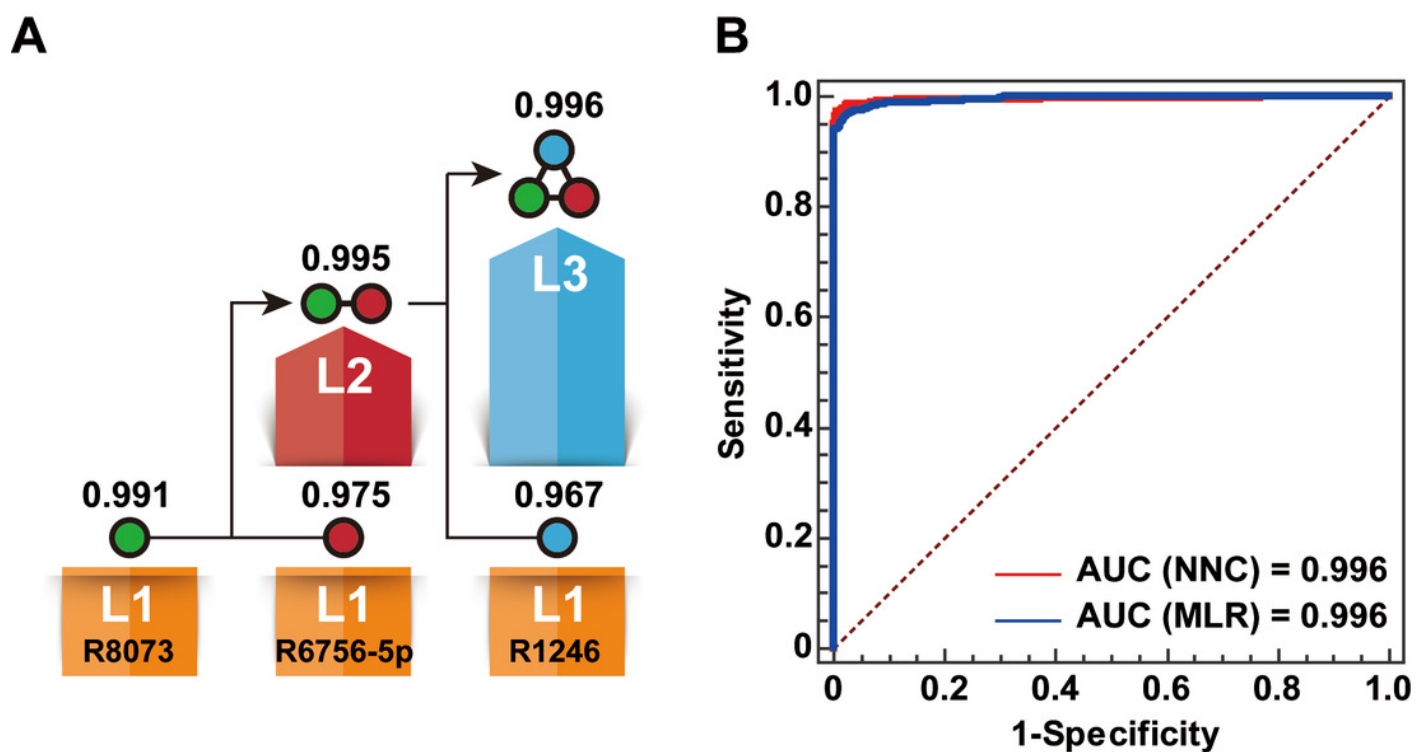
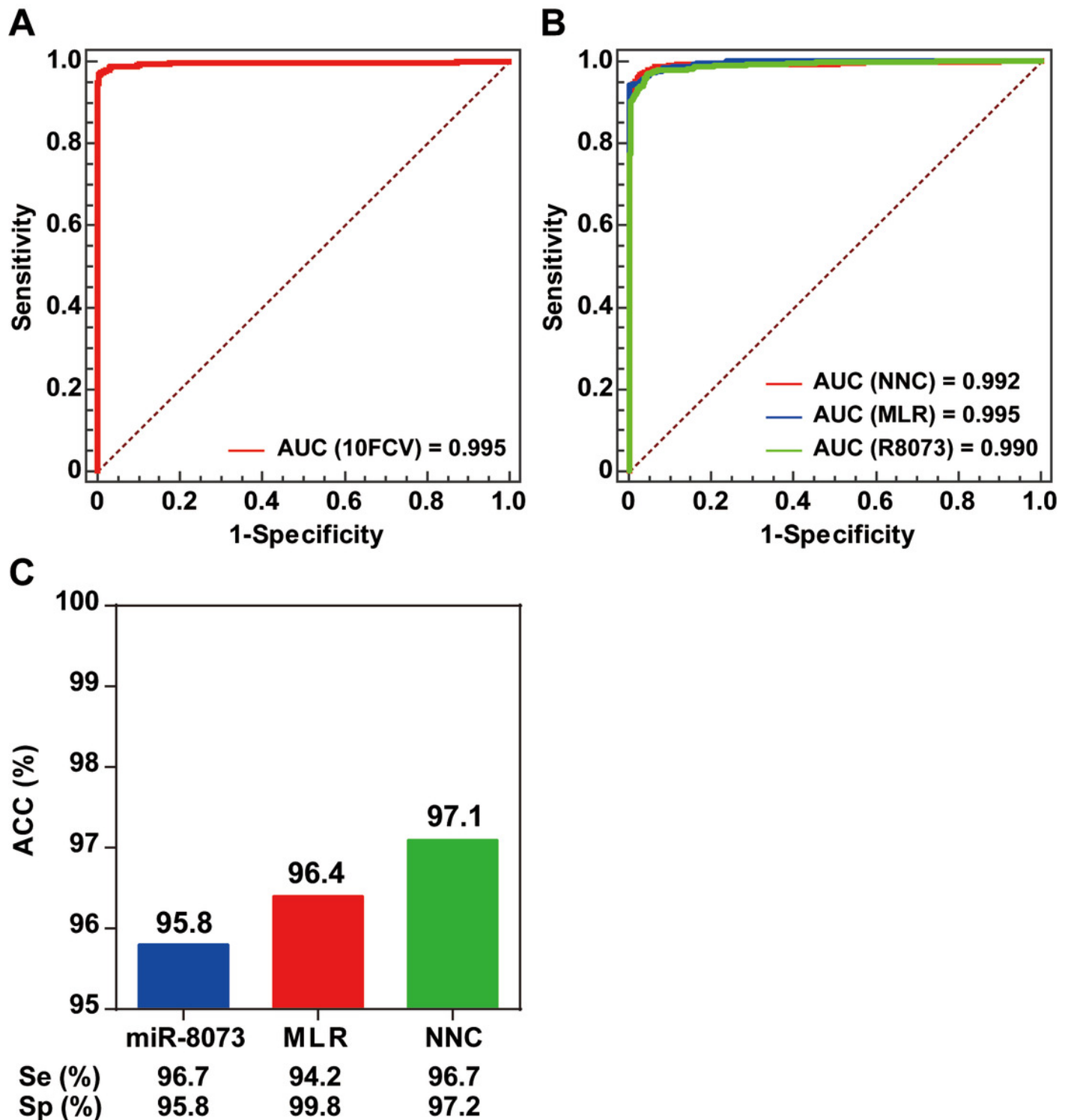



Figure 3

Model validation

A. ROC curve diagram of the tenfold cross-validation of the NNC model. **B.** ROC curve diagram of the NNC and MLR models (validation set). 10FCV: Tenfold cross-validation. **C.** Accuracy evaluation of miR-8073, MLR, and NNC BC detection (validation set). R8073: miR-8073. ACC: accuracy rate; Se: sensitivity; Sp: specificity.

ROC curve analysis of individual miRNAs (training set)

1   **Table 1. ROC curve analysis of individual miRNAs (training set).**

miRNA ID	AUC	Sensitivity (%)	Specificity (%)	ACC (%)
miR-197-5p	0.961	90.7	95.9	94.2
miR-1238-5p	0.964	90.3	97.1	94.9
miR-1246	0.967	89.8	91.7	91.1
miR-3156-5p	0.976	89.8	96.1	94.0
miR-4532	0.968	89.8	98.7	95.8
miR-6748-5p	0.962	90.2	90.3	90.3
miR-6756-5p	0.975	92.7	97.2	95.7
miR-8073	0.991	95.7	97.6	97.0

Comparison between NNC and MLR models (training set)

1 **Table 2. Comparison between NNC and MLR models (training set).**

Model	AUC	Sensitivity (%)	Specificity (%)	ACC (%)
Layer 1 of NNC	0.991	95.7	97.6	96.9
Layer 2 of NNC	0.995	95.8	98.5	97.6
Layer 3 of NNC	0.996	97.3	99.1	98.5
MLR	0.996	96.5	97.9	97.4