Predicting lymph node metastasis and prognosis of individual cancer patients based on miRNA-mediated RNA interactions

Shulei Ren, Wook Lee and Kyungsook Han*

From The 20th International Conference on Bioinformatics (InCoB 2021) Kunming, China. 6-8 November 2021

Abstract
Background: Lymph node metastasis is usually detected based on the images obtained from clinical examinations. Detecting lymph node metastasis from clinical examinations is a direct way of diagnosing metastasis, but the diagnosis is done after lymph node metastasis occurs.

Results: We developed a new method for predicting lymph node metastasis based on differential correlations of miRNA-mediated RNA interactions in cancer. The types of RNAs considered in this study include mRNAs, IncRNAs, miRNAs, and pseudogenes. We constructed cancer patient-specific networks of miRNA mediated RNA interactions and identified key miRNA–RNA pairs from the network. A prediction model using differential correlations of the miRNA–RNA pairs of a patient as features showed a much higher performance than other methods which use gene expression data. The key miRNA–RNA pairs were also powerful in predicting prognosis of an individual patient in several types of cancer.

Conclusions: Differential correlations of miRNA–RNA pairs identified from patient-specific networks of miRNA mediated RNA interactions are powerful in predicting lymph node metastasis in cancer patients. The key miRNA–RNA pairs were also powerful in predicting prognosis of an individual patient of solid cancer.

Keywords: Lymph node metastasis, miRNA–mediated RNA interaction, Prognosis, Competitive endogenous RNA

Background
The spread of cancer cells from the original (primary) tumor to another part of the body is called metastasis. During metastasis, cancer cells travel to other areas through either the bloodstream or the lymph system. As one of the steps of tumor metastasis, lymph node metastasis is commonly observed in cancer patients. Lymph node metastasis itself does not directly endanger the life of patients, but malignant tumors can metastasize to other parts of the body through lymph node metastasis [1]. Many studies have reported that the prognosis of patients with lymph node metastasis is worse than that of patients without lymph node metastasis [2]. Lymph node metastasis is also an important factor in determining effective treatment options for cancer patients.

Lymph node metastasis is usually detected based on the images obtained from clinical examinations. Recently deep learning methods such as convolutional neural networks (CNN) have been used to help clinicians detect lymph node metastasis in ultrasound images...
were also powerful in predicting prognosis of individual
use gene expressions alone. The key miRNA–RNA pairs
much higher performance than other methods which
predicting lymph node metastasis. The model showed a
the correlations of the miRNA–RNA pairs as features for
interactions from the networks. We built a model using
RNA interactions, and identified key miRNA–RNA
cancer patient-specific networks of miRNA mediated
lncRNAs, miRNAs, and pseudogenes. We constructed
of miRNA-mediated RNA interactions in cancer. The
occurs.

Several studies have reported abnormal gene expres-
sion in the process of lymph node metastasis [6]. For
example, the study of Okugawa et al. [7] suggested that
the expression of KISS1 is closely related to lymph node
metastasis in colorectal cancer. Zhang et al. [8] predicted
lymph node metastasis using differentially expressed
mRNAs and noncoding RNAs. Dihge et al. predicted
lymph node metastasis using gene expressions combined
with clinicopathological characteristics [9].

Expression data of mRNAs and noncoding RNAs are
valuable resources for studying and predicting lymph
node metastasis. But, cancer is a complex and hetero-
geneous disease, so abnormal expression of individual
genes cannot fully explain the development and meta-
tasis of cancer. The development and metastasis of cancer
better explained by the dysregulation of gene interac-
tions rather than by individual genes alone. For exam-
ple, AKT1 is abnormally expressed in many types of
cancer and the up-regulation of AKT1 has been known
to be related to lymph node metastasis. But, recent stud-
ies found that miR-138 binding to AKT1 regulates the
expression of AKT1 in tongue squamous cell carcinoma
[10]. miR-519d inhibits lymph node metastasis by regu-
lating MMP3 in oral squamous cell carcinoma and breast
cancer [11, 12].

Salmena et al. [13] proposed a new gene regula-
tion known as competitive endogenous RNA (ceRNA)
hypothesis. The ceRNA hypothesis suggests that RNAs
with similar miRNA response elements compete to bind
to the same miRNA, thereby regulate each other indi-
rectly. Motivated by the increasing evidence support-
the hypothesis, several computational methods have
been developed to construct a network of ceRNAs [14,
15]. Most of the methods focused on mRNAs or IncR-
NAs only as ceRNAs and did not consider pseudogenes
when constructing ceRNA networks.

In this study, we propose a new method for predicting
lymph node metastasis based on differential correlations
of miRNA-mediated RNA interactions in cancer. The
types of RNAs considered in this study include mRNAs,
IncRNAs, miRNAs, and pseudogenes. We constructed
cancer patient-specific networks of miRNA mediated
RNA interactions, and identified key miRNA–RNA
interactions from the networks. We built a model using
the correlations of the miRNA–RNA pairs as features for
predicting lymph node metastasis. The model showed a
much higher performance than other methods which
use gene expressions alone. The key miRNA–RNA pairs
were also powerful in predicting prognosis of individual
patients in several types of cancer. The rest of this paper
presents the method and the experimental results in sev-
eral types of cancer.

Results
Prediction of lymph node metastasis
Using the ΔPCCs of miRNA–RNA pairs obtained in
our study, we predicted lymph node metastasis using
the stacking model and base models (SVM and logistic
regression) in seven types of cancer. As expected, the
stacking model showed the better performance than the
other models both in cross-validation and in independent
testing (Additional file 1).

We compared the performance of stacking models
using two different types of features: ΔPCC of miRNA–
RNA pairs and RNA expression. ΔPCC of miRNA–RNA
pairs was computed by equation 4 in the Methods sec-
tion. For RNA expression feature, we used the RNAs with
a p-value < 0.01 both in differential analysis between
normal samples and tumor samples and in additional dif-
ferential analysis between lymph node metastasis sam-
plies and non-metastatic samples. The performance of
the stacking models was evaluated by fivefold cross-val-
idation and independent testing using several measures:
sensitivity, specificity, accuracy, positive predictive value
(PPV), negative predictive value (NPV) and area under
curve (AUC).

Tables 1 and 2 show the performance of two stacking
models in the fivefold cross validation and in the inde-
dependent testing, respectively. The stacking models with
ΔPCCs as features showed a better performance than
those with RNA expressions both in the fivefold cross
validation and in independent testing, except for thyroid
cancer (THCA) in independent testing. These results
indicate that ΔPCC of miRNA–RNA pairs is a more pow-
ner feature than the gene expression level in predict-
ing lymph node metastasis, which in turn supports that
lymph node metastasis is associated with dysregulation
of gene interactions rather than individual genes, as men-
tioned in the Background section.

We also compared the performance of our method
with that of Zhang’s method [8] using the same data
sets and the same SVM model. Among the seven types
of cancer used in our study, comparison was made for
four types of cancer because the four cancer types are
common to both studies. The train_score and test_score
in Table 3 were obtained using the scikit-learn package,
which was used by Zhang’s study. In all cancer types used
in comparison, our model which used ΔPCCs as features
was better than the four SVM models of Zhang’s method,
which used the expression levels of mRNAs, miRNAs and
IncRNAs separately. These results also demonstrate that
ΔPCCs of miRNA–RNA pairs are much more powerful
features than expression data of RNAs when predicting lymph node metastasis.

Overall survival of cancer patients
We analyzed the overall survival of cancer patients by performing a log-rank test with respect to ΔPCCs of miRNA–RNA pairs obtained in this study. Table 4 shows top three miRNA–RNA pairs with the smallest p-value from the log-rank test in each type of cancer. The remaining miRNA–RNA pairs with p-value less than 0.01 are available in Additional file 2.
Table 3  Comparison of the performance of our SVM model with that of Zhang’s SVM model [8]

| Cancer   | Method_feature | Train_score | Test_score |
|----------|----------------|-------------|------------|
| BRCA     | Our model\_ΔPCC | 0.972       | 0.787      |
|          | Zhang\_mRNA    | 0.798       | 0.680      |
|          | Zhang\_miRNA   | 0.764       | 0.757      |
|          | Zhang\_lncRNA  | 0.793       | 0.696      |
| COAD     | Our model\_ΔPCC | 0.984       | 0.905      |
|          | Zhang\_mRNA    | 0.849       | 0.871      |
|          | Zhang\_miRNA   | 0.902       | 0.886      |
|          | Zhang\_lncRNA  | 0.869       | 0.871      |
| LUAD     | Our model\_ΔPCC | 0.996       | 0.850      |
|          | Zhang\_mRNA    | 0.808       | 0.849      |
|          | Zhang\_miRNA   | 0.885       | 0.795      |
|          | Zhang\_lncRNA  | 0.798       | 0.849      |
| LUSC     | Our model\_ΔPCC | 0.999       | 0.904      |
|          | Zhang\_mRNA    | 0.871       | 0.900      |
|          | Zhang\_miRNA   | 0.939       | 0.847      |
|          | Zhang\_lncRNA  | 0.861       | 0.900      |

In comparison of two types of features (RNA expression vs. deltaPCC), the better performances are shown in bold

Among the seven types of cancer used in our study, comparison was made in four types of cancer because they are the only common cancer types in both studies. The train_score and test_score were obtained using the scikit-learn package, which was used by Zhang’s study. In four cancer types, our model showed the better performance in both training and testing. our model\_ΔPCC: SVM model using ΔPCCs as features. Zhang\_X: SVM model using the expression levels of RNA type X as features

As shown in Table 4, the p-values from the log-rank test with ΔPCC are much smaller than those with individual RNAs involved in the miRNA–RNA pairs. Three pseudogenes (RPL26P29, PNLIPRP2, and CSAG4) are included in the top three miRNA–RNA pairs with the smallest p-value (Table 4), and several miRNA-pseudogene pairs were found as potential prognostic pairs for all 7 types of cancer (Additional file 2).

Figure 1 compares the overall survival rates of two groups of patients with respect to ΔPCC of miRNA–RNA pairs in 7 types of cancer. In all 7 types of cancer, ΔPCCs of miRNA–RNA pairs were powerful in predicting the survival rates of patients. For comparative purposes, Fig. 2 shows the overall survival rates of patients of BRCA, COAD and LUAD with respect to RNA expressions instead of ΔPCC of miRNA–RNA pairs. The RNAs involved in the miRNA–RNA pairs of Fig. 1 (miR-26b\_AC079414.1 pair for BRCA, miR-604\_AL162426.1 pair for COAD, and miR-581\_LINC00628 for LUAD) were selected for the comparison. None of the individual RNAs involved in the pairs showed predictive power of the survival rates of cancer patients, whereas the miRNA–RNA pairs were very powerful in predicting the survival rates of patients as demonstrated in Fig. 1.

ceRNA networks

For every tumor sample in Table 5, we constructed a ceRNA network and derived ΔPCC of miRNA–RNA pairs from the network. We then constructed ceRNA networks with the miRNA–RNA pairs. Figure 3 shows a ceRNA network composed of all miRNA–RNA pairs for breast invasive carcinoma (BRCA). The network includes 1563 miRNA–RNA interactions among 119 miRNAs, 423 lncRNAs, 380 mRNAs and 252 pseudogenes. The small network centered at mir-149 is a blowup of the subnetwork enclosed by a red box.

miR-149 is a miRNA that interacts with ceRNAs most frequently in the ceRNA network. miR-149 is known to promote metastasis in breast cancer when it is down regulated [16]. The ceRNA network also contains several genes associated with breast cancer. For instance, mutations in ERBB4 have been known to be associated with breast cancer [17]. Overexpression of YWHAE increases the proliferation, migration and invasion ability of breast cancer cells [18]. KAT6A promotes SMAD3 binding to oncogenic chromatin modifier TRIM24 and disrupts its interaction with the tumor suppressor TRIM33, which lead to the tumor cell metastasis in breast cancer [19].

As an example of patient-specific networks, Fig. 4 shows the ceRNA networks specific to two LUAD patients with different ΔPCCs of the miR-581\_LINC00628 pair. Figure 4A is a ceRNA network for a LUAD patient (sample ID: TCGA-44-7670) with a high ΔPCC group of the pair, whereas Fig. 4B is a ceRNA network for a LUAD patient (sample ID: TCGA-NJ-A55O) with a low ΔPCC group of the pair. The network in Fig. 4A is composed of 210 miRNA–RNA pairs among 29 miRNAs, 77 lncRNAs, 47 mRNAs and 38 pseudogenes, and the network in Fig. 4B is composed of 111 miRNA–RNA pairs among 5 miRNAs, 53 lncRNAs, 30 mRNAs and 19 pseudogenes.

Apparently, the network in Fig. 4A includes more RNAs and interactions among them than that in Fig. 4B. As shown earlier in Fig. 1, patients with a high ΔPCC of the miR-581\_LINC00628 pair have a much lower survival rate than those with a low ΔPCC of the pair. Similar observations were made in the other types of cancer.

Discussion

The result of our work showed that ΔPCCs of miRNA–RNA pairs derived from patient-specific ceRNA networks are more powerful than the expression levels of individual RNAs in predicting lymph node metastasis. This is related with dysregulated ceRNA interactions in cancer [20]. For instance, miR-125b may induce breast cancer metastasis by binding to STARD13 [21]. HOXD-AS1 prevents miR-130a-3p mediated degradation of...
SOX4 through competitive binding to miR-130a-3p, thereby promoting hepatocellular carcinoma transfer [22]. MT1JP regulates gastric cancer progression by binding to miR-92a-3p competitively with FBXW7 [23].

Unlike other studies on ceRNA interactions, our study considered pseudogenes as well as mRNAs and IncRNAs as ceRNAs. Pseudogenes were previously considered as genomic junk and neglected in the studies on ceRNA interactions as well. However, several experimental evidences suggested that pseudogenes can act as ceRNAs in the development of disease [24–26]. For instance, Karreth et al. [27] demonstrated that the pseudogene BRAFP1 functions as a ceRNA and induces lymphoma in vivo. Overexpression of the oncogenic pseudogene BRAFP1 promotes the formation of human B-cell lymphomas through serving as a ceRNA of the parental gene BRAF [28]. In prostate cancer, the pseudogene PTENP1 functions as a ceRNA to regulate PTEN expression by sponging miR-499-5p [29]. Straniero et al. [30] demonstrated that the pseudogene GBAP1 can function as a ceRNA for the glucocerebrosidase gene GBA by sponging miR-22-3p, thus revealing a new regulatory network in the pathogenesis of Parkinson's Disease.

There are limitations in our current work. A patient-specific ceRNA network consists of miRNA–RNA pairs with significant changes from other patients by including miRNA–RNA pairs whose $\Delta PCC$ is larger than the median $|\Delta PCC|$ of all tumor samples of the same type. Since we used $|\Delta PCC|$ instead of $\Delta PCC$, a patient-specific network does not show the direction of change (i.e., increase or decrease) in PCC. In the future, we plan to come up with a better way of presenting such information in a patient-specific network. Another direction of future work is to improve the performance of the prediction model further, in particular for thyroid carcinoma.
Fig. 2 Overall survival rates of patients with respect to expressions of individual RNAs in Fig. 1. In contrast to the miRNA–RNA pairs, none of the individual RNAs showed predictive power of the survival rates of cancer patients.
Conclusion

The spread of tumors has always been a difficulty in tumor treatment, especially large-scale spread, which greatly reduces the survival rate of patients. Lymph node metastasis is the first step in the spread of many tumors. Therefore, predicting lymph node metastasis can make medical interventions in advance and reduce the risk of large-scale spread.

In this study, we constructed ceRNA networks for 7 types of solid cancer. Unlike other ceRNA networks, our ceRNA networks include pseudogenes as well as mRNA and lncRNAs. From the miRNA–RNA pairs in the ceRNA networks, we built a prediction model of lymph node metastasis in tumor samples.

Experimental results of the prediction model showed that ΔPCCs of miRNA–RNA pairs from ceRNA networks are powerful for predicting lymph node metastasis in tumor samples. Comparison of our method with the features of other methods using the same data sets showed that ΔPCCs of miRNA–RNA pairs are much more powerful than gene expression levels in predicting lymph node metastasis of cancer patients. Some miRNA–RNA pairs were also powerful in predicting prognosis of individual patients. Our work is preliminary and requires further investigation for clinical use. However, this approach will help characterize individual cancer patients and predict the occurrence of lymph node metastasis in advance.

Methods

The overall workflow of our method is shown in Fig. 5. It shows data collection, data filtering, data processing, generation of miRNA–RNA gene pairs, and construction of a prediction model and patient-specific ceRNA network.

Data collection

We collected gene expression profiles of lncRNAs, mRNAs, pseudogenes, and miRNAs and clinical profiles from The
Fig. 3  ceRNA network for breast invasive carcinoma (BRCA). The network is composed of 1563 miRNA–RNA interactions among 119 miRNAs, 423 lncRNAs, 380 mRNAs and 252 pseudogenes. The small network centered at miR-149 is a blowup of the subnetwork enclosed by a red box.
Cancer Genome Atlas (TCGA) data portal [31] for primary tumor samples of all solid cancer types. Normal samples of each type of cancer were also obtained from the TCGA data portal. All the gene expression profiles used in this study were obtained by RNA-sequencing (RNA-seq).

In TCGA, there were 18 types of solid cancer which have at least 200 samples. Among the 18 types, 6 types were excluded due to insufficient data on lymph node metastasis in their tumor samples. In the remaining 12 types of solid cancer, we selected the types which have at least 30 normal samples and 50 tumor samples with lymph node metastasis. Only 7 types of solid cancer satisfied such criteria: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), stomach adenocarcinoma (STAD) and thyroid carcinoma (THCA).

The clinical profiles of the TCGA data includes the Tumor, Node, Metastasis (TNM) stage of samples. Samples with an M stage of 1 were excluded because distant organ metastasis often coexists with lymph node metastasis and makes the evaluation of prediction difficult. Based on the TNM staging system, we clustered the tumor samples into those with lymph node metastasis and those without lymph node metastasis.

- Samples with lymph node metastasis: tumor samples with T stage of 1–4, N stage of 1–3, and M stage of 0
- Samples without lymph node metastasis: tumor samples with T stage of 1–4, N stage of 0, and M stage of 0

Table 5 shows the number of normal samples, tumor samples, tumor samples with lymph node metastasis, and tumor samples without lymph node metastasis in 7 types of cancer. The TCGA barcodes of all normal samples and tumor samples of Table 5 are provided as Additional file 3. The TCGA barcode is the primary identifier of biospecimen data in the TCGA project.
Fig. 5: The overview of the overall workflow. There are three types of samples: normal samples (gray), tumor samples without lymph node metastasis (sky blue) and tumor samples with lymph node metastasis (pink). In our prediction model, tumor samples with lymph node metastasis (pink) and tumor samples without lymph node metastasis (sky blue) are treated as positive and negative instances, respectively.
Table 6 The number of RNAs of four biotypes in each cancer type studied in this study

| Cancer | #miRNAs | #mRNAs | #lncRNAs | #pseudogenes |
|--------|---------|--------|----------|--------------|
| BRCA   | 165     | 18,084 | 8553     | 5528         |
| COAD   | 157     | 17,573 | 7284     | 5304         |
| HNSC   | 95      | 18,018 | 7427     | 4643         |
| LUAD   | 197     | 18,084 | 8553     | 5528         |
| LUSC   | 161     | 18,227 | 8706     | 5680         |
| STAD   | 379     | 18,617 | 10,354   | 9039         |
| THCA   | 153     | 17,568 | 7342     | 4753         |

Gene filtering
The gene names of the TCGA data are represented by Ensembl ID. Thus, we obtained the annotation files from the Ensembl project [32] and determined the names and biotypes of the genes (mRNAs, lncRNAs, pseudogenes and miRNAs). Table 6 shows the number of genes and their types.

We filtered out genes with an average count below 1. In RNA-seq data, counts are non-negative integer values. The count of unexpressed genes is 0, so the count of expressed genes is at least 1. Since the genes with the average count < 1 are unexpressed genes in most samples, we removed them. We then normalized the RNA-seq data of the genes using the trimmed mean of M values (TMM) method [33].

Deriving miRNA–RNA pairs and feature selection
As mentioned earlier, any of lncRNAs, mRNAs, and pseudogenes with common miRNA response elements compete to bind to the same miRNA, so can act as competitive endogenous RNAs (ceRNAs). To obtain initial miRNA–RNA pairs we computed the maximal information coefficient (MIC) [34] of each miRNA with ceRNA miRNA–RNA pairs we computed the maximal information coefficient (MIC). Pearson correlation coefficient (PCC) is the most commonly used for gene association. However, we used MIC instead of PCC to select potential miRNA–RNA pairs for a few reasons: (1) PCC can measure linear association only, but MIC measures linear or non-linear association between two variables. (2) MIC is less susceptible to outliers than PCC.

RNAs of the miRNA–RNA pairs are scattered into the two-dimensional space, which is divided into $n_X \times n_Y$ bins in the X and Y axes, Here X denotes the expression level of miRNA and Y denotes the expression level of any one of mRNA, lncRNA, or pseudogene in the pairs. Based on the number of scattered points in each bin, we calculate the mutual information $I(X, Y)$ by Eq. 1. This process is repeated until the largest mutual information is obtained as the MIC (Eq. 2).

$$I(X, Y) = \sum_{X,Y} p(X,Y) \log_2 \left( \frac{p(X,Y)}{p(X)p(Y)} \right)$$ (1)

$$\text{MIC}(X,Y) = \max_{n_X,n_Y < B} \frac{I(X, Y)}{\log_2 \min(n_X, n_Y)}$$ (2)

The parameter B of MIC controls how much of the characteristic matrix we search over. Setting B too high can lead to non-zero scores even for random data, while setting B too low results in searching only for simple patterns [34]. we used the default setting for B, the 0.6th power of the number of samples, because the default setting is known to work well in practice [34].

Unlike the parameter B, there is no default setting for MIC. When selecting miRNA–RNA pairs for analysis, the threshold for MIC was set to 0.5, which is the median of its range [0, 1]. Setting the threshold of MIC smaller than 0.5 results in more miRNA–RNA pairs, which will contain a large number of spurious pairs. In contrast, with a larger threshold, we may miss potential prognostic gene pairs.
MICs of miRNA–RNA pairs were computed separately in tumor samples and normal samples because the association strength of miRNAs and ceRNAs are different between tumor and normal samples. Those miRNA–RNA pairs MIC < 0.5 in normal samples and tumor samples were removed because their association is not strong enough to be included in a ceRNA network.

We constructed a ceRNA network by subtracting a reference network for a group of normal samples from a perturbed network with a single tumor sample. Thus, each edge in the patient-specific network represents a differential PCC (ΔPCC) of miRNA–RNA pair between a single tumor sample and a group of normal samples. MIC was not used at this stage because ΔMIC does not make sense by its definition and ΔPCC is more suitable for quantifying the perturbation by a single sample.

\[
PCC(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}
\]

(3)

where n: number of samples; X: miRNA; Y: mRNA, lncRNA, or pseudogene

\[
\Delta PCC(X, Y) = PCC_{n+1}(X, Y) - PCC_n(X, Y)
\]

(4)

Every edge of a ceRNA network represents ΔPCC of a miRNA–RNA pair, which is obtained by the following procedure:

1. Compute PCC of every miRNA–RNA pair in n normal samples.
2. Recompute PCC in n + 1 samples which include n normal samples and a single tumor sample.
3. Compute differential PCCs (ΔPCCs) between normal samples and the tumor sample.

We divided the ΔPCCs of miRNA–RNA pairs into 2 groups, lymph node metastasis and non-metastasis, and performed the Wilcoxon test [35] in the two groups. We selected miRNA–RNA pairs with a p-value less than 0.01 in the Wilcoxon test. We reduced the number of miRNA–RNA pairs further by PCA. Table 7 shows the number of miRNA–RNA pairs left after each filtering process.

### Table 7

| Cancer  | #Features after MIC filtering | #Features after Wilcoxon test | #Features after PCA |
|---------|-----------------------------|-------------------------------|---------------------|
| BRCA    | 90,837                      | 1563                          | 480                 |
| COAD    | 178,983                     | 1969                          | 80                  |
| HNSC    | 67,020                      | 800                           | 100                 |
| LUAD    | 341,146                     | 12,981                        | 200                 |
| LUSC    | 165,765                     | 2436                          | 200                 |
| STAD    | 976,763                     | 17,445                        | 60                  |
| THCA    | 38,077                      | 3397                          | 150                 |

The miRNA–RNA pairs with a p-value ≥ 0.01 were removed by the Wilcoxon test. The number of features was further reduced after dimension reduction by PCA of ΔPCCs. In both MIC filtering and the Wilcoxon test, each feature represents a miRNA–RNA pair. In PCA, the number of features indicates the dimension of a feature vector.

Stacking ensemble learning in the scikit-learn library [41]. We first trained the SVM model and LR model (base learners) with the original training set. We then used their prediction results as features to train a secondary learner. We used LR as the secondary classifier, which is the default classifier in the library. Stacking integrates the prediction results of the base learners in the best way through the secondary learner.

The tumor samples obtained from TCGA were divided into training and test sets with the ratio of 7:3. The parameters of the prediction model were determined by the grid search in the training set. When training and validating the prediction model, tumor samples with lymph node metastasis were considered as positive instances, and tumor samples without lymph node metastasis were considered as negative instances.

### Construction of a ceRNA network

For each type of cancer, we constructed a ceRNA network with the miRNA–RNA pairs obtained by the Wilcoxon test. A node of the ceRNA network represents one of miRNA, mRNA, lncRNA or pseudogene, and an edge represents the interaction of miRNA with other RNAs.

The patient-specific ceRNA network is a sub-network of the ceRNA network. For each miRNA–RNA pair, we computed the median of the absolute value of ΔPCC (i.e., |ΔPCC|) of the pair in all tumor samples of the same cancer type. A patient-specific ceRNA network was constructed by selecting the miRNA–RNA pairs whose |ΔPCC| is larger than the median |ΔPCC|. Thus, the edges in a patient-specific ceRNA network represent the miRNA–RNA interactions which show a significant change from other patients of the same cancer type.
Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-022-01231-x. All additional files are available at http://bclab.inha.ac.kr/LNM/.

Additional file 1. Performance of two base models (logistic regression and SVM) and the ensemble model by stacking the base models in predicting lymph node metastasis.

Additional file 2. Potential prognostic miRNA–RNA pairs in seven types of cancer.

Additional file 3. TCGA barcodes of all normal samples and tumor samples studied in our work.

Acknowledgements

The authors thank the editor and the referees for their valuable comments and suggestions.

About this supplement

This article has been published as part of BMC Medical Genomics Volume 15 Supplement 1, 2022: The 20th International Conference on Bioinformatics (InCoB 2021): medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-15-supplement-1.

Author contributions

SR worked on correlations among RNAs, derived triplets and prepared the initial manuscript. WL helped the initial manuscript. KH supervised the work and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Ministry of Science and ICT (2020R1A2B801096299) and INHA UNIVERSITY Research Grant. The funders did not play any role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript. Publication costs are funded by the NRF Grant.

Availability of data and materials

The TCGA barcodes of all normal samples and tumor samples studied in our work are available in Additional file 3. The source code for generating miRNA–RNA pairs is available at https://github.com/rslrl/LNM.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Sleeman JP, Thiele W. Tumor metastasis and the lymphatic vasculature. Int J Cancer. 2009;125(2):2747–56.
2. Jones D, Pereira ER, Padera TP. Growth and immune evasion of lymph node metastasis. Front Oncol. 2018;8:36.
3. Zhou LQ, Wu XL, Huang SY, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology. 2020;294(1):19–28.
4. Nguyen S, Polat D, Karbassi P, et al. Preoperative prediction of lymph node metastasis from Clinical DCE MRI of the primary breast tumor using a 4D CNN. Med Image Comput Assist Interv. 2020;1262:326–34.
5. Sun Q, Lin X, Zhao Y, et al. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don’t forget the peritumoral region. Front Oncol. 2020;10:53.
6. Kawada K, Taketo MM. Significance and mechanism of lymph node metastasis in cancer progression. Cancer Res. 2011;71(4):1214–8.
7. Okugawa Y, Inoue Y, Tanaka K, et al. Loss of the metastasis suppressor gene K551 is associated with lymph node metastasis and poor prognosis in human colorectal cancer. Oncol Rep. 2013;30(3):1449–54.
8. Zhang S, Zhang C, Du J, et al. Prediction of lymph-node metastasis in cancers using differentially expressed miRNA and non-coding RNA signatures. Front Cell Dev Biol. 2021;9:605977.
9. Diéghé L, Vollon-Chriesterson J, Hegardt C, et al. Prediction of lymph node metastasis in breast cancer by gene expression and clinicopathological models: development and validation within a population-based cohort. Clin Cancer Res. 2019;25(21):6968–81.
10. Ji M, Wang W, Yan W, Chen D, Ding X, Wang A. Dysregulation of AKT1, a miR-138 target gene, is involved in the migration and invasion of tongue squamous cell carcinoma. J Oral Pathol Med. 2017;46(9):731–7.
11. Jin Y, Li Y, Wang X, Yang Y. Dysregulation of MiR-519d affects oral squamous cell carcinoma invasion and metastasis by targeting MMP3. J Cancer. 2019;10(12):2720–34.
12. Chu C, Liu X, Bai X, et al. MiR-519d suppresses breast cancer tumorigenesis and metastasis via targeting MMP3. Int J Biol Sci. 2018;14(2):228–36.
13. Salmena L, Poliseno L, Tay Y, Kats L. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011;146(3):533–8.
14. Park B, Lee W, Park I, Han K. Finding prognostic gene pairs for cancer from patient-specific gene networks. BMC Med Genomics. 2019;12(Suppl 8):179.
15. Zhang G, Han C, Shen Z, et al. Identification of cancer-related miRNA–lncRNA biomarkers using a basic miRNA-lncRNA network. PLoS ONE. 2018;13(5):e0196681.
16. Sánchez-González I, Bobien A, Molnar C, et al. miR-149 suppresses breast cancer metastasis by blocking paracrine interactions with macrophages. Cancer Res. 2020;80(6):1330–41.
17. Segers VFM, Dugasquercuer L, Feyen E, Shakeri H, De Keulenaer GW. The role of ErbB4 in cancer. Cell Oncol (Dordr). 2020;43(3):335–52.
18. Yang YF, Lee YC, Wang YY, Wang CH, Hou MF, Yuan SF. YWHAE promotes proliferation, metastasis, and chemoresistance in breast cancer cells. Kaohsiung J Med Sci. 2019;35(7):408–16.
19. Yu B, Luo F, Sun B, et al. KAT6A acetylation of SMAD3 regulates myeloid-derived suppressor cell recruitment, metastasis, and immunotherapy in triple-negative breast cancer. Adv Sci. 2021;8(20):2100014.
20. Chiu HS, Martinez MR, Bansal M, et al. High-throughput validation of ceRNA regulatory networks. BMC Genomics. 2017;18(1):1–11.
21. Tang F, Zhang R, He Y, et al. MicroRNA-125b induces metastasis by targeting STARD13 in MCF-7 and MDA-MB-231 breast cancer cells. PLoS ONE. 2012;7(5):e35435.
22. Wang H, Huo X, Yang XR, et al. STAT3-mediated upregulation of IncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOK. Mol Cancer. 2017;16(1):1–15.
23. Zhang G, Li S, Lu J, et al. LncRNA MT1JP functions as a ceRNA in regulating FBXW7 through competitively binding to miR-92a-3p in gastric cancer. Mol Cancer. 2018;17(1):1–11.
24. Poliseno L, Salmena L, Zhang J, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465(7301):1033–8.
25. Welch JD, Baran-Gale J, Perou CM, et al. Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. BMC Genomics. 2015;16(1):1–16.
26. An Y, Furber KL, Ji S. Pseudogenes regulate parental gene expression via ceRNA network. J Cell Mol Med. 2017;21(1):185–92.
27. Karreth FA, Reschke M, Ruocco A, et al. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell. 2015;161(2):319–32.
28. Chan JJ, Kwok ZH, Chew XH, et al. A FTH1 gene pseudogene: microRNA network regulates tumorigenesis in prostate cancer. Nucl Acids Res. 2018;46(4):1998–2011.
29. Wang L, Zhang N, Wang Z, et al. Pseudogene PTENP1 functions as a competing endogenous RNA (ceRNA) to regulate PTEN expression by sponging miR-499-5p. Biochem Mosc. 2016;81(7):739–47.
30. Straniero L, Rimoldi V, Samaran M, et al. The GRAP1 pseudogene acts as a ceRNA for the glucocerebrosidase gene GBA by sponging miR-22-3p. Sci Rep. 2017;7(1):1–13.
31. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20.
32. Hove KL, Achuthan P, Allen J, et al. Ensembl 2021. Nucl Acids Res. 2021;49(D1):D884–91.
33. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
34. Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. Science. 2011;334(6062):1518–24.
35. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;16(6):80–3.
36. Syarf I, Zaluska E, Prugel-Bennett A, et al. Application of bagging, boosting and stacking to intrusion detection. In: Proceedings of the 8th international conference on machine learning and data mining in pattern recognition, vol. 7376. 2012. p. 593–602.
37. Ribeiro MHDM, dos Santos Coelho L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agriculture time series. Appl Soft Comput. 2020;86:105837.
38. Ting KM, Witten IH. Stacking bagged and dagged models. 1997.
39. Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res. 1999;10:271–89.
40. Mahendran N, Vincent PMDR, Srinivasan K, et al. Realizing a stacking generalization model to improve the prediction accuracy of major depressive disorder in adults. IEEE Access. 2020;8:49509–22.
41. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.
Learn more biomedcentral.com/submissions