Bioinformatics, 35(17), 2019, 3127–3132
doi: 10.1093/bioinformatics/btz042
Advance Access Publication Date: 22 January 2019
Original Paper

OXFORD

Databases and ontologies

CMEP: a database for circulating microRNA expression profiling

Jian-Rong Li[1,2,†], Chun-Yip Tong[1,†], Tsai-Jung Sung[1], Ting-Yu Kang[1], Xianghong Jasmine Zhou[3] and Chun-Chi Liu[1,2,*]

[1]Institute of Genomics and Bioinformatics and [2]Advanced Plant Biotechnology Center, National Chung Hsing University, Taichung City 402, Taiwan and [3]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed.
†These authors contributed equally to the paper as first authors.

Associate Editor: Janet Kelso
Received on August 22, 2018; revised on December 29, 2018; editorial decision on January 14, 2019; accepted on January 18, 2019

Abstract

Motivation: In recent years, several experimental studies have revealed that the microRNAs (miRNAs) in serum, plasma, exosome and whole blood are dysregulated in various types of diseases, indicating that the circulating miRNAs may serve as potential noninvasive biomarkers for disease diagnosis and prognosis. However, no database has been constructed to integrate the large-scale circulating miRNA profiles, explore the functional pathways involved and predict the potential biomarkers using feature selection between the disease conditions. Although there have been several studies attempting to generate a circulating miRNA database, they have not yet integrated the large-scale circulating miRNA profiles or provided the biomarker-selection function using machine learning methods.

Results: To fill this gap, we constructed the Circulating MicroRNA Expression Profiling (CMEP) database for integrating, analyzing and visualizing the large-scale expression profiles of phenotype-specific circulating miRNAs. The CMEP database contains massive datasets that were manually curated from NCBI GEO and the exRNA Atlas, including 66 datasets, 228 subsets and 10 419 samples. The CMEP provides the differential expression circulating miRNAs analysis and the KEGG functional pathway enrichment analysis. Furthermore, to provide the function of noninvasive biomarker discovery, we implemented several feature-selection methods, including ridge regression, lasso regression, support vector machine and random forests. Finally, we implemented a user-friendly web interface to improve the user experience and to visualize the data and results of CMEP.

Availability and implementation: CMEP is accessible at http://syslab5.nchu.edu.tw/CMEP.
Contact: chunchiliu@gmail.com

1 Introduction

Many types of diseases, especially cancer, are associated with disease-specific biomarkers that serve as diagnosis, prognosis and monitoring tools, and provide a better understanding of disease pathogenesis (Weiland *et al.*, 2012). However, some of the current diagnostic procedures have limitations in the application of routine health checkups, since they are invasive and inconvenient (Chen *et al.*, 2008; Duffy, 2007). Hence, minimally invasive biomarkers of human disease, such as a diagnosis using blood-based liquid biopsies, can significantly improve the disease prognosis by facilitating early diagnosis and routine clinical monitoring (Lieben, 2015). MicroRNAs (miRNAs) are small endogenous noncoding RNAs with approximately 22 nucleotides that can modulate up to 60% of the protein-coding genes in the human genome at the

posttranscriptional level (Bartel, 2004; Friedman *et al.*, 2009). Additionally, large amounts of miRNAs were derived from various tissues/organs and present in stable forms in the serum, plasma, exosome and whole blood (Alhasan *et al.*, 2014; Chen *et al.*, 2008). Thus, circulating miRNAs have emerged as promising potential in noninvasive biomarkers for human disease diagnosis and surveillance using blood-based liquid biopsies (Kawaguchi *et al.*, 2016; Ma *et al.*, 2012; O'Brien *et al.*, 2017; Singh *et al.*, 2016; Weiland *et al.*, 2012). Recently, several studies have identified some disease-specific circulating miRNA signatures for various diseases (Alhasan *et al.*, 2016; De Rosa *et al.*, 2018; Liu *et al.*, 2016; Motawi *et al.*, 2015; Zhang *et al.*, 2017).

Although there have been several studies that attempt to construct circulating miRNA databases, e.g. miRandola (Russo *et al.*, 2018) and exRNA Atlas (http://exrna-atlas.org/) (Ainsztein *et al.*, 2015), they have not yet integrated large-scale circulating miRNA profiles or provided the biomarker-selection function using machine learning methods. The miRandola database is a manually curated database based on literature for extracellular circulating noncoding RNAs, which contains the relations between circulating miRNAs and diseases from published articles. However, miRandola has not provided miRNA expression profiles and profile analyses. On the other hand, exRNA Atlas contains numerous circulating small RNA datasets with expression profiles of various types of diseases, but it does not provide further analysis, such as the functional pathway enrichment with differentially expressed circulating miRNAs or feature selection functions for finding the circulating biomarkers of noninvasive diagnosis and prognosis.

To fill this gap, we have developed the *Circulating MicroRNA Expression Profiling* (CMEP) database (http://syslab5.nchu.edu.tw/CMEP), which is a public database that not only contains large-scale circulating miRNA datasets from diverse platforms (e.g. small RNA sequencing, miRNA microarray and qRT-PCR, etc.) but also provides miRNA expression profiling, pathway enrichment analysis with miRNA target genes and feature-selection methods. Figure 1 shows the framework of the CMEP construction. In the CMEP database, we systematically collected 169 circulating miRNA expression-profile datasets with specific disease conditions. To provide the comprehensive miRNA resource for noninvasive diagnosis, we collected wide-ranging sample types, including serum, plasma, exosome, microvesicle, urine, peripheral blood mononuclear cells, red blood cells and platelets, etc. Each dataset contains several groups of samples with different phenotypes. We manually performed data curation for these circulating miRNA datasets to create phenotype-specific subsets and to assign samples to subsets according to the experimental description of the samples. Afterwards, we comprehensively categorized subsets according to disease state, disease subtypes, mutations, cancer stages and sample types, etc. This resulted in 66 miRNA datasets, including 228 subsets and 10 419 samples. To identify phenotype-specific differentially expressed miRNAs in each dataset, we selected subsets with at least three samples, normalized the expression profiles and then performed a t-test between the two subsets without overlapping samples, which resulted in 194 subset pairs.

To demonstrate the functionality to biologists, the CMEP visualizes the expression profiles of all differentially expressed circulating miRNAs with the significance level and expression values between two subsets, and provides the filtering function based on the *P*-value threshold, up/downregulation, or autocomplete search field. Furthermore, the CMEP provides an enrichment analysis function,

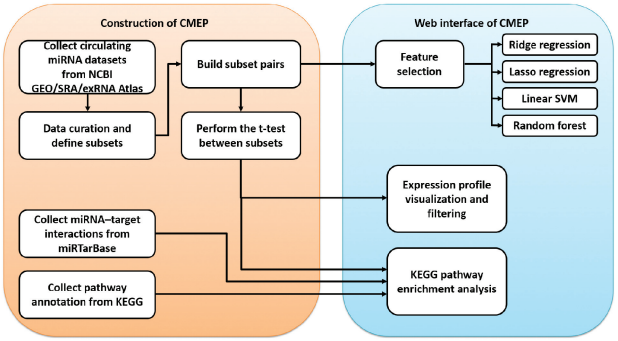


Fig 1. Framework for constructing the CMEP database. Circulating miRNA datasets were collected from NCBI GEO, SRA and the exRNA Atlas, and then all samples were classified into phenotype-specific subsets. To identify phenotype-specific differentially expressed circulating miRNAs in each dataset, we performed t-tests between each pair of subsets with no overlapping samples from the same dataset. For each subset pair, we constructed a KEGG functional pathway enrichment analysis that integrated the information of miRNA–target interactions from the miRTarBase and the functional pathway annotations of miRNA target genes from the KEGG. Furthermore, four feature-selection pipelines, such as ridge regression, lasso regression, linear support vector classification (SVC) and random forests, were constructed with recursive feature elimination (RFE) to identify the important circulating miRNAs as potential biomarkers. Finally, all data and analysis functions were integrated into a user-friendly web interface

which integrated the miRNA–target interactions from the miRTarBase (Chou *et al.*, 2016) and the functional pathway annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017), for a better understanding of the differentially expressed circulating miRNAs within specific disease conditions. For noninvasive disease biomarker identification, we implemented several feature-selection methods, such as ridge regression (Hoerl and Kennard, 1970), lasso regression (Tibshirani, 1996), linear support vector classification (SVC) (Chang and Lin, 2011) and random forests (Genuer *et al.*, 2010; Strobl *et al.*, 2008), into the CMEP database to provide users with the ability to identify the crucial circulating miRNAs.

In summary, the CMEP database characterizes differentially expressed circulating miRNAs, analyzes the biological pathways that involve the circulating miRNAs, and recognizes the relevant miRNAs across various types of diseases in different organ systems. The CMEP database serves as a resource to enable biological and clinical researchers to develop new noninvasive biomarkers for disease diagnosis and routine monitoring using blood-based liquid biopsies.

2 Materials and methods

2.1 Circulating miRNA datasets collections

We systematically collected 66 human circulating miRNA datasets regarding diseases and the annotation data of the platforms used in the datasets from the NCBI Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002), the Sequence Read Archive (SRA) (Leinonen *et al.*, 2011) and the exRNA Atlas. Each dataset consists of several subsets, where each subset is a group of samples associated with a specific phenotype. Thus, we manually created several phenotype-specific subsets within each dataset and then assigned samples to subsets according to the textual description of the datasets and samples. Consequently, each subset contained a group of samples with specific phenotypic traits or disease conditions, e.g. cancer state,

disease progression, a type of tissue and genotype. Finally, a total of 228 subsets with at least 3 samples were created, containing a total of 10 419 samples. We categorized these datasets and subsets into specific disease structures referring to the Unified Medical Language System (UMLS) (Bodenreider, 2004). To obtain the expression profiles of each dataset, we converted the probe IDs into human miRNA names based on the platform annotation. If an miRNA maps to multiple probe IDs, we calculate the average expression values.

2.2 Phenotype-specific differentially expressed circulating miRNAs identification and pathway enrichment analysis

To identify phenotype-specific differentially expressed circulating miRNAs in each dataset, we selected the subsets with at least three samples and then performed t-tests between two subsets without overlapping samples. The comparison of two subsets is referred to as a subset pair. There were 194 subset pairs with differentially expressed circulating miRNAs (P-value < 0.05). In turn, to investigate the phenotype-specific regulation of circulating miRNAs, which may reveal the potential biological functional insight into circulating miRNAs, we integrated the information on miRNA–target interactions that had been downloaded from the miRTarBase (Chou *et al.*, 2016) release 7.0, and the functional pathway annotations of miRNA target genes that had been collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017) to construct a KEGG functional pathway enrichment analysis for the target genes of differentially expressed circulating miRNAs within a subset pair. To perform the KEGG enrichment analysis, the hypergeometric distribution was used to calculate the statistical significance. The calculating formula of hypergeometric test P-value is:

$$p = \sum_{i=x}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Here N is the number of genes with any KEGG annotation; M is the number of genes involved in the specific KEGG pathway; n is the number of target genes of a given differentially expressed circulating miRNA; and x is the number of n belonging to M, that is, the intersection of M and n. Taking the calculated P-value ≤ 0.05 as a threshold, CMEP will dynamically analyze to identify the significantly enriched KEGG pathways for the target genes of a given differentially expressed circulating miRNA selected by the user. In the web interface, users can select differentially expressed miRNAs using t-test P-value thresholds. When users type a miRNA name, an autocomplete search function provides suggestions for miRNAs after quickly searching and displaying partially matched terms. Furthermore, we linked the circulating miRNA IDs to miRandola (Russo *et al.*, 2018) on the CMEP web interface.

2.3 Feature-selection pipeline construction for circulating miRNAs

To identify the circulating miRNAs as potential biomarkers for the detection or monitoring of various diseases, we implemented four feature-selection methods, such as ridge regression (Hoerl and Kennard, 1970), lasso regression (Tibshirani, 1996), linear support vector classification (SVC) (Chang and Lin, 2011) and random forests (Genuer *et al.*, 2010; Strobl *et al.*, 2008), to calculate the weight associated with each feature (i.e. circulating miRNAs) and to reflect the importance of the circulating miRNAs for phenotype-specific subset pair classification. The scikit-learn library (http://scikit-learn.org/) was used to implement all feature-selection pipelines. In the web interface, CMEP performs the feature selection procedure and visualizes the profile of features when users choose a feature-selection method.

Since the goal is to identify the subset of features that can classify the subset pair, the ridge regression was implemented by logistic regression (Fan *et al.*, 2008) with *L2* norm regularization, and lasso regression was implemented by logistic regression with *L1* norm regularization. Moreover, every forest contained 100 decision trees in random forests. All feature-selection algorithms were applied with recursive feature elimination (RFE) (Guyon *et al.*, 2002) to extract the specific numbers of relevant circulating miRNAs.

To validate the feature-selection methods and to compare the performance of each feather-selection method, we selected the subset pair 'Prostate cancer versus Normal', which have 86 samples from dataset GSE71008 for demonstration. The samples in dataset GSE71008 were randomly equally split into training and test sets. The training set was used to select features using REF. The test set was used to calculate the performance using the linear SVM classifier.

3 Results

3.1 The web interface

The web interface of CMEP comprises four distinctive panels (Fig. 2) as follows: Disease panel (upper left), Dataset panel (upper right), Subset pair panel (lower left) and Expression profiling panel (lower right). The disease panel consists of two components, the sample type selection box and the disease tree. Users will first select the desired sample type (serum, plasma, urine, etc.), and the disease tree will interactively display all corresponding diseases using UMLS classification, such as breast cancer and colon cancer. After selecting a disease type, the dataset panel displays all related datasets and information including the dataset's GSE numbers, sample type and title. Once a particular dataset is chosen, the subset pair panel shows all subset pairs, such as normal versus disease, within the dataset. After selecting a subset pair, the expression profiling panel shows the differential expression profile of the subset pair's miRNAs. Within the expression profile panel, the dataset title with GSE number, subset titles and number of samples are provided. Users can choose up/down regulation and P-value threshold to filter the results. The miRNA search box with autocomplete function is provided to access the miRNAs that are particularly of interest. For each miRNA, the t-test P-value and subset mean values are calculated, and the expression profile is visualized by a heat map to denote the expression levels.

3.2 KEGG enrichment analysis and feature selection

For each miRNA in the expression profiling panel, by interacting with miRTarBase to search for miRNA's target genes, pathway enrichment analysis calculates the overlapped genes to discover significantly enriched biological pathways (Fig. 3).

To unveil more relevant miRNAs in the dataset, we applied machine learning algorithms, such as the linear support vector classifier, ridge regression, lasso regression and random forest classification for feature selection. Within each of the feature-selection methods, the desired number of features can be freely chosen by the user's interest (Fig. 4).

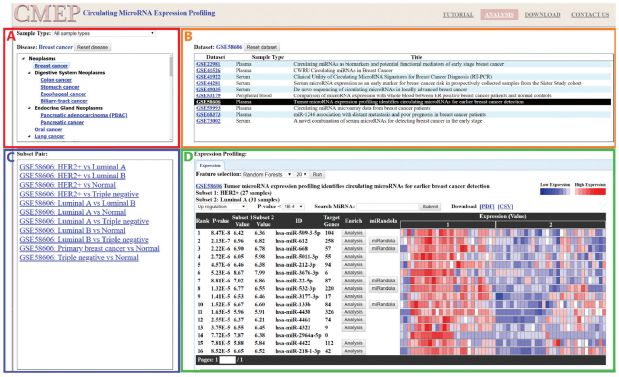
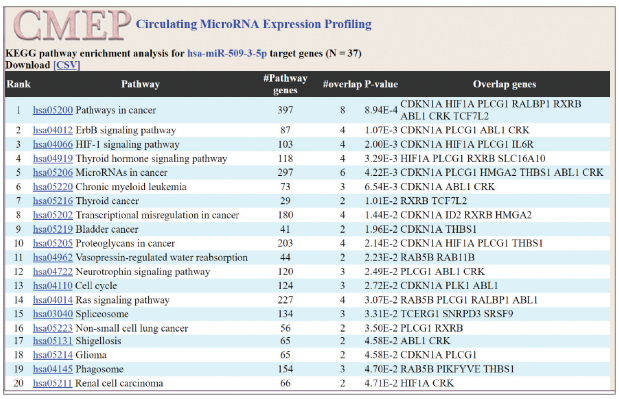


Fig 2. The CMEP web interface provides four major panels as follows. (**A**) Disease panel (upper left): The sample type box shows all available sample types. The diseases tree lists all diseases according to UMLS classification in a hierarchical manner. (**B**) Dataset panel (upper right): All datasets of the selected disease are listed with sample type and title. (**C**) Subset pair panel (lower left): All subset pairs within datasets are listed. (**D**) Expression profiling panel (lower right): The differential expression profile of miRNAs is graphically presented, with corresponding *P*-values calculated by t-test. The feature selection functions apply machine learning algorithms to search for important miRNAs



CMEP Circulating MicroRNA Expression Profiling

KEGG pathway enrichment analysis for hsa-miR-509-3-5p target genes (N = 37)
Download [CSV]

Rank	Pathway	#Pathway genes	#overlap	P-value	Overlap genes
1	hsa05200 Pathways in cancer	397	8	8.94E-4	CDKN1A HIF1A PLCG1 RALBP1 RXRB ABL1 CRK TCF7L2
2	hsa04012 ErbB signaling pathway	87	4	1.07E-3	CDKN1A PLCG1 ABL1 CRK
3	hsa04066 HIF-1 signaling pathway	103	4	2.00E-3	CDKN1A HIF1A PLCG1 IL6R
4	hsa04919 Thyroid hormone signaling pathway	118	4	3.29E-3	HIF1A PLCG1 RXRB SLC16A10
5	hsa05206 MicroRNAs in cancer	297	6	4.22E-3	CDKN1A PLCG1 HMGA2 THBS1 ABL1 CRK
6	hsa05220 Chronic myeloid leukemia	73	3	6.54E-3	CDKN1A ABL1 CRK
7	hsa05216 Thyroid cancer	29	2	1.01E-2	RXRB TCF7L2
8	hsa05202 Transcriptional misregulation in cancer	180	4	1.44E-2	CDKN1A ID2 RXRB HMGA2
9	hsa05219 Bladder cancer	41	2	1.96E-2	CDKN1A THBS1
10	hsa05205 Proteoglycans in cancer	203	4	2.14E-2	CDKN1A HIF1A PLCG1 THBS1
11	hsa04962 Vasopressin-regulated water reabsorption	44	2	2.23E-2	RAB5B RAB11B
12	hsa04722 Neurotrophin signaling pathway	120	3	2.49E-2	PLCG1 ABL1 CRK
13	hsa04110 Cell cycle	124	3	2.72E-2	CDKN1A PLK1 ABL1
14	hsa04014 Ras signaling pathway	227	4	3.07E-2	RAB5B PLCG1 RALBP1 ABL1
15	hsa03040 Spliceosome	134	3	3.31E-2	TCERG1 SNRPD3 SRSF9
16	hsa05223 Non-small cell lung cancer	56	2	3.50E-2	PLCG1 RXRB
17	hsa05131 Shigellosis	65	2	4.58E-2	ABL1 CRK
18	hsa05214 Glioma	65	2	4.58E-2	CDKN1A PLCG1
19	hsa04145 Phagosome	154	3	4.70E-2	RAB5B PIKFYVE THBS1
20	hsa05211 Renal cell carcinoma	66	2	4.71E-2	HIF1A CRK

Fig 3. Web interface of KEGG pathway enrichment analysis. The target genes of the miRNA are overlapped with pathway genes, in order to calculate the *P*-value. The pathway with a smaller *P*-value has a higher ranking

3.3 Example applications

To demonstrate the biological applications of CMEP, we used dataset GSE71008 as an example to show the differentially expressed circulating miRNA profile, feature selection function and KEGG pathway enrichment analysis. Within the dataset, we used a subset pair (prostate cancer versus normal) containing 86 samples (36 prostate cancer samples and 50 normal samples) in total.

3.4 Differentially expressed circulating miRNAs

To demonstrate how differentially expressed circulating miRNAs assisted users in evaluating the statistically significant miRNAs within a large amount of the rest, hsa-mir-146b was taken as an example (Fig. 5). Users can intuitively notice that the 5 most statistically significant miRNAs within the subset pair are hsa-mir-146a (*P*-value 2.73E−8), hsa-mir-150-5p (*P*-value 2.33E−7), hsa-mir-144-5p (*P*-value 3.16E−7), hsa-mir-223-3p (*P*-value 6.29E−7) and hsa-mir-146b (*P*-value 8.10E−7). The subset-pair mean values were also calculated to provide more in-depth insight into the miRNA variation between normal samples and prostate cancer samples. The heat map

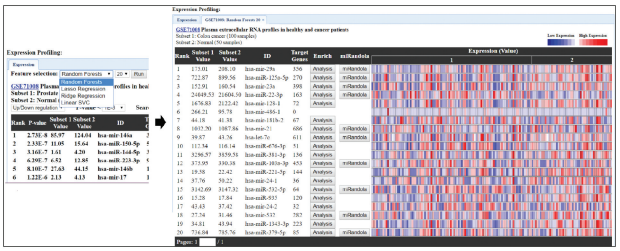
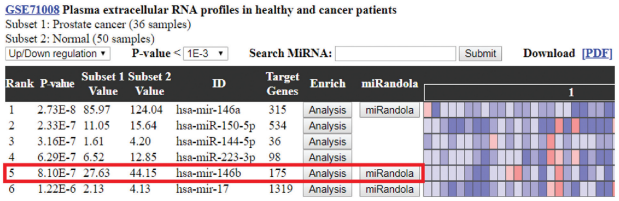


Fig 4. Web interface of feature selection methods. The four feature-selection methods, random forest classification, lasso regression, ridge regression, linear SVC allow users to select any numbers of features



GSE71008 Plasma extracellular RNA profiles in healthy and cancer patients
Subset 1: Prostate cancer (36 samples)
Subset 2: Normal (50 samples)
Up/Down regulation | P-value < 1E-3 | Search MiRNA: | Submit | Download [PDF]

Rank	P-value	Subset 1 Value	Subset 2 Value	ID	Target Genes	Enrich	miRandola
1	2.73E-8	85.97	124.04	hsa-mir-146a	315	Analysis	miRandola
2	2.33E-7	11.05	15.64	hsa-miR-150-5p	534	Analysis	
3	3.16E-7	1.61	4.20	hsa-miR-144-5p	36	Analysis	
4	6.29E-7	6.52	12.85	hsa-miR-223-3p	98	Analysis	
5	8.10E-7	27.63	44.15	hsa-mir-146b	175	Analysis	miRandola
6	1.22E-6	2.13	4.13	hsa-mir-17	1319	Analysis	miRandola

Fig 5. The miRNA expression profile of GSE71008 (Prostate cancer versus Normal). The miRNA expression profile calculates the *P*-values by t-test and the miRNAs are ranked by the *P*-value. The hsa-mir-146b is the fifth miRNA in the expression profile, and its expression is shown down-regulated in prostate cancer samples. This finding is proven by a previous study that mir-146b expression in prostate tumor cells is significantly decreased compared to that in normal prostate cells

graphically represents the miRNA expression of each of the samples, with red representing high expression and blue representing low expression. In the miRNA expression profile, users can easily observe that the fifth miRNA, hsa-mir-146b, is obviously down-regulated among the prostate cancer samples. Interestingly, a previous study showed that hsa-mir-146, which was a potential tumor suppressor, was significantly reduced in prostate cancer tissues (Ding *et al.*, 2016).

3.5 Feature selection

CMEP provides four feature selection methods, which are lasso regression, ridge regression, linear support vector classifier and random forest classifier, which are all integrated with a recursive feature elimination algorithm to examine the significance of each of the miRNAs in the dataset. In the GSE71008 subset (prostate cancer versus normal), we performed the four different feature selection methods and selected 20 miRNAs using each of the methods. The four feature selection methods obtained 45 unique miRNAs, and 15 miRNAs were documented in different studies related to prostate cancer previously. For example, Okato et al. reported that hsa-mir-150-5p expressed the antitumor property in prostate cancer by targeting SPOCK1 (Okato *et al.*, 2017). Interestedly, the expression profile of hsa-miR-150-5p also showed that it was down-regulated among the prostate cancer samples, which consolidated the validity of the feature selection methods. hsa-mir-150-5p was selected by all four feature selection methods, suggesting that it may be one of the most important circulating miRNA biomarkers for prostate cancer.

To further investigate the predictive capacity of the feature-selection methods, we split the original prostate cancer versus normal subset into a training subset and testing subset in a ratio of 1: 1. We performed the four methods of the training subset, selecting 1 feature to 50 features and using the same training subset to fit the model using different classifiers such as the support vector classifier or logistic regression. Then, we used the testing set to validate the

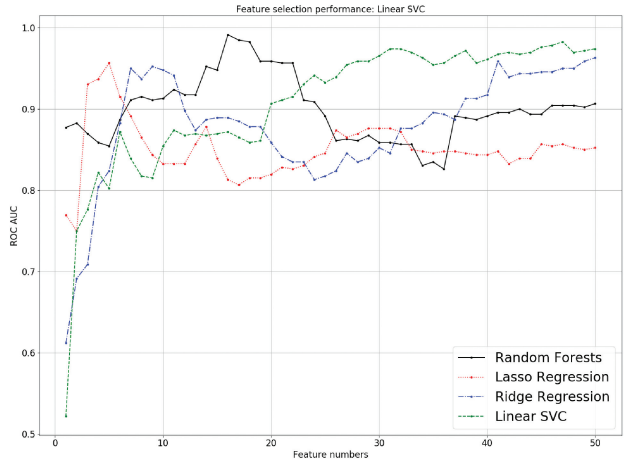


Fig 6. The performance of four feature selection methods in subset pair 'Prostate cancer versus Normal' of GSE71008 using linear SVM with different numbers of features. All four feature-selection methods had a similar performance ranging from 1 to 50 features. Ridge regression and linear SVC performed poorly when less than 5 features were selected (AUCs < 0.80), while the random forest method ($0.85 <$ AUCs < 0.90) and lasso regression method ($0.75 <$ AUCs < 0.95) performed pretty well. The random forest method performed better when fewer features were selected (14–22 features, AUCs > 0.95), while other methods had increasing AUCs (AUCs > 0.80) when more features were selected

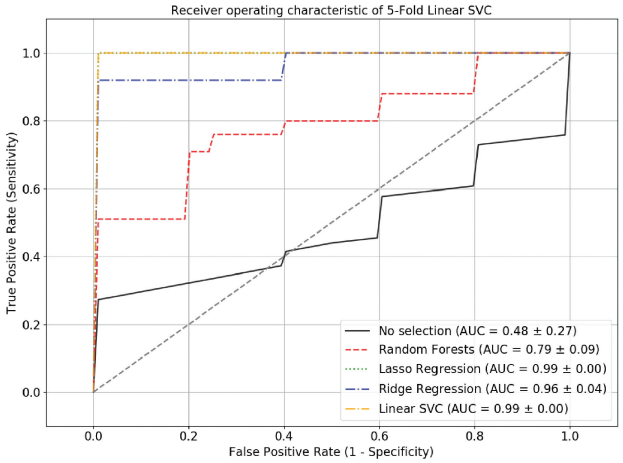


Fig 7. ROC curve of the 5-fold cross-validation linear SVC model. The ROC curve showed that the linear SVC model performed better when feature-selection methods were included, as their AUCs ranged from 0.79 to 0.99, which were significantly higher than those without feature selection (AUC = 0.48)

area under the curve (AUC) of the receiver operating characteristic curve (ROC Curve) (Fig. 6).

By implementing the features selected by four different methods and performing validation, we observed that under the linear SVC, the AUCs of the four feature-selection methods obtained an overall similar performance, but random forest-selected features appeared to perform better when fewer numbers of features were selected (14-22 features, with AUCs > 0.95), while other feature-selection methods, such as lasso regression, ridge regression and SVC had increasing AUCs (AUCs > 0.80) when more features were selected. From a statistics point of view, we preferred to use fewer, more important features to prevent overfitting. From the above results, the four feature-selection methods were all effective in predicting important miRNAs, which assisted biologists in discovering potential candidates for further biological experiments.

To demonstrate whether feature selection followed by machine learning was outperformed by solely using machine learning, we tested another dataset, GSE98181, to perform linear SVC 5-fold cross-validations, with and without feature selection (Fig. 7). To examine their performance, ROC analysis was implemented. The results showed that all feature-selection methods followed by linear SVC had AUCs ranging from 0.79 to 0.99, which significantly outperformed linear SVC without feature selection (AUC = 0.48), suggesting that the feature selection methods substantially improved performance. Interestingly, miR-29a-3p, which was simultaneously selected by three feature-selection methods, was found to be documented in the previous published research (Pei *et al.*, 2016). The research found that miR-29a promoted breast cancer cell proliferation and was upregulated in the breast cancer cell line, which was consistent with our expression profile on breast cancer samples and normal samples with mean values of 0.58 and 0.19, respectively.

3.6 KEGG pathway enrichment analysis

To further explore how miRNAs were involved in different biological pathways, 15 miRNAs from GSE71008, which were selected from the previously mentioned feature-selection methods and were proven to be related to prostate cancer, underwent KEGG pathway enrichment analysis. Six miRNAs (mir-223-3p, mir-451a, mir-125a-5p, mir-233, mir-143 and mir-221) showed that their target genes were significantly enriched to the prostate cancer pathway. This gave users a more thorough understanding of the miRNA's attributes.

4 Conclusions

In this study, we processed large-scale datasets collected from NCBI GEO, SRA and exRNA Atlas to extract the differentially expressed circulating miRNAs, implemented the KEGG pathway analysis for the target genes of the circulating miRNAs, and constructed four feature-selection pipelines to identify the crucial circulating miRNAs as potential noninvasive biomarkers for diagnosis or prognosis in various types of diseases. We integrated and visualized all data and functions into the CMEP database. Although CMEP was developed with the goal of collecting, processing, analyzing and visualizing all publicly circulating miRNA data regarding various types of disease, it still has limitations that arise from the characteristics of the data. First, the circulating miRNA datasets were generated using different array or sequencer platforms, which caused most of datasets to have different numbers of circulating miRNAs and different scales of expression values. Second, the circulating data was generated from various research teams with diverse experiment back-grounds, designs and protocols, which resulted in inevitable batch effects and potential quality issues between different datasets in CMEP. To summarize, we have therefore applied the strategy of only comparing subsets within the same dataset and have demonstrated the performance of feature selection across different datasets in this study.

Despite these limitations, the CMEP characterizes differentially expressed circulating miRNAs, analyzes the bio-logical pathways they involve, and provides the feature-selection methods for identifying the crucial circulating miRNAs. Moreover, the systematic and user-friendly web interface can assist users in accessing the information on CMEP efficiently. We anticipate that CMEP can facilitate biological and clinical researchers in better studying the biological insight of the circulating miRNAs and develop new noninvasive biomarkers for diagnosis and routine monitoring. Finally, as more and more circulating miRNA datasets are generated and provided, we

will keep the data in our CMEP database updated to ensure that it can provide comprehensive information on circulating miRNAs.

Funding

This work was supported in part by the Advanced Plant Biotechnology Center from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. Moreover, this study was supported by the grant NIH1U01CA230705 of the National Institutes of Health (NIH).

Conflict of Interest: none declared.

References

Ainsztein,A.M. *et al*. (2015) The NIH extracellular RNA communication consortium. *J. Extracell. Vesicles*, **4**, 27493.

Alhasan,A.H. *et al*. (2014) Exosome encased spherical nucleic acid gold nanoparticle conjugates as potent microRNA regulation agents. *Small*, **10**, 186–192.

Alhasan,A.H. *et al*. (2016) Circulating microRNA signature for the diagnosis of very high-risk prostate cancer. *Proc. Natl. Acad. Sci. USA*, **113**, 10655–10660.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*., **32**, D267–D270.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol*., **2**, 1–27.

Chen,X. *et al*. (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res*., **18**, 997–1006.

Chou,C.H. *et al*. (2016) miRTarBase 2016: updates to the experimentally validated miRNA–target interactions database. *Nucleic Acids Res*., **44**, D239–D247.

De Rosa,S. *et al*. (2018) Transcoronary concentration gradients of circulating microRNAs in heart failure. *Eur. J. Heart Fail*, **20**,1000–1010.

Ding,H.Y. *et al*. (2016) MicroRNA-146b acts as a potential tumor suppressor in human prostate cancer. *J BUON*, **21**, 434–443.

Duffy,M.J. (2007) Role of tumor markers in patients with solid cancers: a critical review. *Eur. J. Intern. Med*., **18**, 175–184.

Edgar,R. *et al*. (2002) Gene Expression Omnibus: nCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.

Fan,R.E. *et al*. (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res*., **9**, 1871–1874.

Friedman,R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*., **19**, 92–105.

Genuer,R. *et al*. (2010) Variable selection using random forests. *Pattern Recogn. Lett*., **31**, 2225–2236.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn*., **46**, 389–422.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression—biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55.

Kanehisa,M. *et al*. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*., **45**, D353–D361.

Kawaguchi,T. *et al*. (2016) Circulating MicroRNAs: a next-generation clinical biomarker for digestive system cancers. *Int. J. Mol. Sci*., **17**, 1459.

Leinonen,R. *et al*. (2011) The sequence read archive. *Nucleic Acids Res*., **39**, D19–D21.

Lieben,L. and Diagnosis, (2015) RNA-seq for blood-based pan-cancer diagnostics. *Nat. Rev. Cancer*, **15**, 696–697.

Liu,X.L. *et al*. (2016) Disease-specific miR-34a as diagnostic marker of non-alcoholic steatohepatitis in a Chinese population. *World J. Gastroenterol*., **22**, 9844–9852.

Ma,R. *et al*. (2012) Circulating microRNAs in cancer: origin, function and application. *J. Exp. Clin. Cancer Res*., **31**, 38.

Motawi,T.K. *et al*. (2015) Serum MicroRNAs as potential biomarkers for early diagnosis of hepatitis C virus-related hepatocellular carcinoma in egyptian patients. *PLoS One*, **10**, e0137706.

O'Brien,K.P. *et al*. (2017) Circulating MicroRNAs in cancer. *Methods Mol. Biol*., **1509**, 123–139.

Okato,A. *et al*. (2017) Dual strands of pre-miR150 (miR1505p and miR1503p) act as antitumor miRNAs targeting SPOCK1 in naive and castration-resistant prostate cancer. *Int. J. Oncol*., **51**, 245–256.

Pei,Y.F. *et al*. (2016) MiR-29a promotes cell proliferation and EMT in breast cancer by targeting ten eleven translocation 1. *Biochim. Biophys. Acta*, **1862**, 2177–2185.

Russo,F. *et al*. (2018) miRandola 2017: a curated knowledge base of non-invasive biomarkers. *Nucleic Acids Res*., **46**, D354–D359.

Singh,R. *et al*. (2016) Circulating microRNAs in cancer: hope or hype? *Cancer Lett*., **381**, 113–121.

Strobl,C. *et al*. (2008) Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B Met*., **58**, 267–288.

Weiland,M. *et al*. (2012) Small RNAs have a large impact: circulating microRNAs as biomarkers for human diseases. *RNA Biol*., **9**, 850–859.

Zhang,Y. *et al*. (2017) Serum microRNA panel for early diagnosis of the onset of hepatocellular carcinoma. *Medicine (Baltimore)*, **96**, e5642.